Identifying Twin Travelers Using Ridesourcing Trip Data

Nicolas Chiabaut, Cyril Veve

Articles

Identifying Twin Travelers Using Ridesourcing Trip Data

Nicolas Chiabaut[*], Cyril Veve[†]

Keywords: transportation network companies, ridesourcing, similarity, data clustering, shared mobility demand

DOI: 10.32866/9223

Transport Findings

Shared mobility services are announced as a game changer in transportation and a promising solution to reduce congestion and improve performance of urban mobility. However, only few studies aim at identifying if it really exists a significant demand for such services. This paper propose a first attempt toward this objective by finding twin travelers into ridesourcing trips data. Using the Didi Chuxing's dataset, a general methodology is defined to assess the similarity between trips and to cluster analogous travelers. The study reveals, among others, that at least 18% of the trips can be paired with introducing tolerable delays.

RESEARCH QUESTION AND HYPOTHESIS

This study is devoted to identifying twin trips in a city, i.e., pairs of travelers who make almost the same trips. Such travelers demonstrate the potential demand for shared mobility systems, especially possible trip-sharing services such as ridesourcing, shared taxis, ridesharing, etc. A major hypothesis of this study is to limit consideration to spatiotemporal features of the trips to assess their similarities and their potential for matching (Rayle et al. 2015). Other attributes such as cost, comfort, additional behavioral variables, or the characteristics of the transportation service are not yet accounted for (Zhan, Qian, and Ukkusuri 2016; Vazifeh et al. 2018).

METHODS AND DATA

The Chinese transport network company DiDi Chuxing has released two months' worth of data consisting of more than 6 million trips performed by their drivers (Xu et al. 2019). For each trip i, this dataset gives access to the following information: departure time t_i^{PU} and location $p_i^{PU} = (x_i^{PU}, y_i^{PU})$ of the passenger(s) pick-up; arrival time t_{DO}^i and location $p_i^{DO} = (x_i^{DO}, y_i^{DO})$ of the drop-off. For this study, we only used a subset of the dataset by focusing on the peak hours of a regular day: approximately 10,000 trips from $8h$ to $11h$ on November 18, 2016. Moreover, we consider that these observations correspond to the desired departure/arrival times and origins/destinations of the travelers.

* **Institution:** Université de Lyon **Department:** ENTPE / IFSTTAR, LICIT **ORCID iD:** 0000-0003-0450-4890 **Link:** https://nicolaschiabaut.weebly.com

† **Institution:** Université de Lyon **Department:** ENTPE / IFSTTAR, LICIT

To identify the trips that can be made with the same vehicle, we use the following method. First, we define a function $S(i,j)$ to express the similarity between two trips i and j. This function must encompass the different spatiotemporal attributes of the trips. It should reproduce the trip information that two travelers can share if their origins and locations and also their departure and arrival times are close enough. To the authors' best knowledge, this kind of similarity index is almost nonexistent in the literature (Ketabi, Alipour, and Helmy 2018). Consequently, we propose the following function:

$$\overline{S(i,j)} = \sum_{l \in [PU,DO]} \alpha_l e^{|f^l(i,j)|}$$

where $f^l(i,j)$ is a feasibility function and α_l is a coefficient.

Function f describes the service's potential to operate the shared trips, i.e., the ability to pick up (or drop off) the two travelers before both of their desired departure times:

$$f^l(i,j) = |t_i^l - t_j^l| - \gamma d(p_i^l, p_j^l)$$

where d is the geodesic distance and γ is the average duration pace to connect travelers who wish to share a trip. This parameter is a general and synthetic formula to describe the operation of the service and the way in which this service gathers two demand requests into the same vehicle: defining a meeting point, successive pick-ups, etc. For example, if the first traveler must walk to the second traveler's pick-up point, then γ is the inverse of the walking speed. If this distance is traveled by car, meaning that the service offers door-to-door service, then γ is the inverse of the vehicle speed. Consequently, f is positive if the match is realized before the two desired departure times t_i^l and t_j^l, whereas f is negative if travelers must experience delays to make the match possible. Moreover, α_l is equal to $1/2$ if $f^l(i,j) > 0$ and to $3/2$ otherwise because it is more disadvantageous to be delayed.

In addition to this measure of similarity $\overline{S(i,j)}$, excessive distances/durations for rendezvous are penalized. Thus, penalties θ_x^l and θ_t^l are added when, respectively, the distances between pick-up (or drop-off) locations and departure (or arrival) times of trips i and j exceed, respectively, specific thresholds δ_x^l and δ_t^l :

$$\theta_x^l = e^{d(p_i^l, p_j^l) - \delta_x^l} \quad \forall l \ / \ d(p_i^l, p_j^l) > \delta_x^l$$

$$\theta_t^l = e^{|t_i^l - t_j^l| . \frac{\delta_x^l}{\delta_t^l} - \delta_t^l} \quad \forall l \ / \ |t_i^l, t_j^l| > \delta_t^l$$

Otherwise, these penalties are null. In this manner, $S(i,j) = S(i,j) + \theta_x^l + \theta_t^l$ defines a sharp function that enhances the differences between trips and facilitates identification of twin travelers in the dataset.

Next, trips are gathered using a clustering method. It is important to note that a cluster is not a region of the city but a set of trips that are similar based on their pick-up and drop-off attributes. These trips are related to travelers, i.e., demand, who may share a vehicle according to their origin/destination and departure/arrival time. For this study, a DB-SCAN approach with S as the distance function is used. This makes it possible to fix the minimum number of points requested by cluster (Ester et al. 1996). Here, this minimal number is fixed at two, and we only select clusters with two elements because the study aims to determine pairs of similar trips. DB-SCAN also requires a threshold ϵ on the similarity function that is the radius of a neighborhood with respect to some point, i.e., the maximal dissimilarity authorized to determine if two trips can be paired. The parameters used to obtain the different figures in this article are summarized in Table 1.

Table 1: Parameters Used To Obtain the Different Figures

	Parameter	Value	Significance
DB-Scan	MinPts	2	Nb of trips per cluster
	ϵ	4	Radius of a neighborhood
Similarity	γ	0.1 h/km	Average pace to connect travelers
	δ_t^{PU}	0.1 h	Threshold of departure times
	δ_t^{DO}	0.25 h	Threshold of arrival times
	δ_x^{PU}	0.25 km	Threshold of PU locations
	δ_x^{DO}	0.25 km	Threshold of DO locations

FINDINGS

Figure 1 shows the trips of 7 different pairs of twin travelers projected on the roadmap of Chengdu, China. Visual inspection reveals that these results are very promising. Pick-up and drop-off locations are close (less than 1 km, geodesic distance) while the differences in departure and arrival times remain low (less than 10 min). Moreover, $\rho = 18.3\%$ of the trips can be paired for the studied period. This is very interesting because the fleet size of DiDi, and, by extrapolation, the number of cars flowing in the network can be significantly reduced if vehicles are shared. This reduction can even be higher if more than two travelers share the same vehicle. The methodology can be extended to such cases by changing the minimal number of points in the clustering process. Even if the DiDi data is not fully representative of the complete traffic flow, these results highlight the fact that shared mobility may be a promising strategy to improve the transportation system's performance.

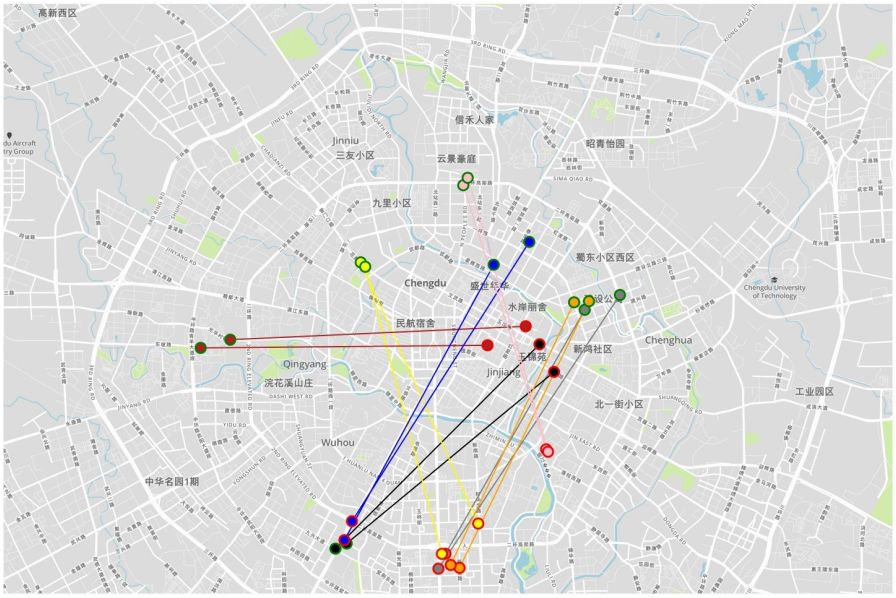

Figure 1: Similar trips for Six Different Pairs; Pick-up Locations Are Circled in Green Whereas Drop-off Locations Are Circled in Red

Visual observations are confirmed by Figure 2.a, which depicts the distribution of the average length $\bar{l}_k$ of the trips for each pair k, whereas Figure 2.b shows the distributions of the average travel times $\bar{\tau}_k$. In addition, Figures 2.c and 2.d present the distributions of the absolute difference in departure times $|t_i^{PU} - t_j^{PU}|_k$ and the absolute difference in the two arrival times $|t_i^{DO} - t_j^{DO}|_k$. It appears that all these values are entirely consistent with the natural idea of what the characteristics of similar trips should be:

- The average length $\bar{l}_k$ of the twin trips is equal to 6.2 km/h (road distance). Notice that the dataset focuses on a subpart of Chengdu's network (a circle with a 5.5 km radius). The associated average travel time is around 17.3 min, leading to an average speed of 21.6 km/h. Consequently, trips are long enough to allow for the delay caused by sharing the vehicle with another traveler.
- Consequently, the difference in the two departure times is on average equal to 4.9 min and lower than 6.6 min for 80% of the trips.
- The average estimated delay is equal to 7.2 min and more than 80% of the trips experience a delay of less than 10 min.
- Finally, it means that a traveler may find their twin to share a vehicle with an increase of only 30% in travel time. This extra time could be drastically reduced by optimizing dispatch of the transportation

supply (Mourad, Puchinger, and Chu 2019).

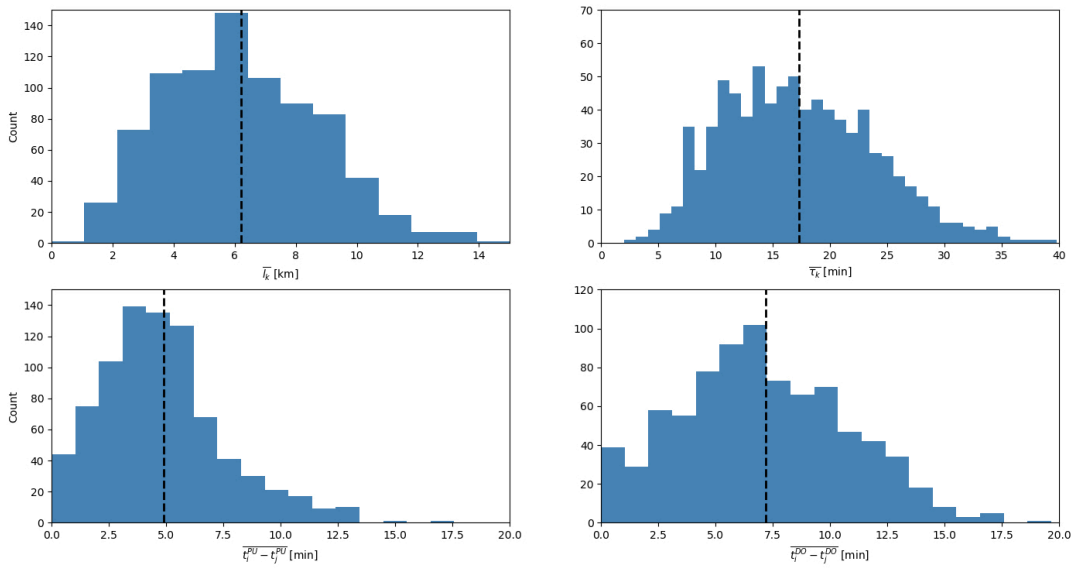

Figure 2: (a) Distribution of the Average Length $\bar{l}_k$ of Trips Within the Cluster k; (b) Distribution of Average Travel Times $\bar{\tau}_k$, (c) Absolute Difference in Departure Times $\overline{|t_i^{PU} - t_j^{PU}|}_k$ and (d) Absolute Difference in Arrival Times $\overline{|t_i^{DO} - t_j^{DO}|}_k$ Among the Pairs; Dotted Lines Show the Mean of the Distributions

ACKNOWLEDGMENTS

The authors thank Dr. "MFD" Guilhem Mariotte for his valuable comments. Data source: DiDi Chuxing GAIA Open Dataset Initiative, available at: https://gaia.didichuxing.com

REFERENCES

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters a Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *The Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, edited by Evangelos Simoudis, Jiawei Han, and Usama Fayyad, 226–31. AAAI Press.

Ketabi, Roozbeh, Babak Alipour, and Ahmed Helmy. 2018. "Playing with Matches: Vehicular Mobility through Analysis of Trip Similarity and Matching." In *The 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '18)*, 544–47. New York, NY: ACM. https://doi.org/10.1145/3274895.3274992.

Mourad, Abood, Jakob Puchinger, and Chengbin Chu. 2019. "A Survey of Models and Algorithms for Optimizing Shared Mobility." *Transportation Research Part B: Methodological* 123: 323–46. https://doi.org/10.1016/j.trb.2019.02.003.

Rayle, Lisa, Susan Shaheen, Nelson Chan, Danielle Dai, and Robert Cervero. 2015. "App-Based, On-Demand Ride Services: Comparing Taxi and Ridesourcing Trips and User Characteristics in San Francisco." 94th Transportation Research Board Annual Meeting. Washington, D.C. https://www.its.dot.gov/itspac/dec2014/ridesourcingwhitepaper_nov2014.pdf.

Vazifeh, M.M., P. Santi, G. Resta, S.H. Strogatz, and C. Ratti. 2018. "Addressing the Minimum Fleet Problem in On-Demand Urban Mobility." *Nature* 557: 534–38.

Xu, Chuan, Jingqin Gao, Fan Zuo, Kaan Ozbay, Hong Yang, and Haipeng Cui. 2019. "Understanding Spatial-Temporal Impacts on Mode Preference between Taxi and E-Hailing Service." 19–05001. 98th Annual Meeting of the Transportation Research Board.

Zhan, Xianyuan, Xinwu Qian, and Satish V. Ukkusuri. 2016. "A Graph-Based Approach to Measuring the Efficiency of an Urban Taxi Service System." *IEEE Transactions on Intelligent Transportation Systems* 17 (9): 2479–89. https://doi.org/10.1109/tits.2016.2521862.

FIGURES, TABLES, AND SUPPLEMENTARY MATERIALS

Figure 3: (a) Evolution of ρ the ratio of twin travelers over the total number of trips; (b) To be changed

Download: http://app.scholasticahq.com/api/v1/attachments/22780/download