A macroscopic comparison approach for substitutionary on-demand flexible micro-transit services

Lynn Fayed^{*1}, Gustav Nilsson¹, and Nikolas Geroliminis¹

¹Urban Transport Systems Laboratory (LUTS) École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. {lynn.fayed,gustav.nilsson,nikolas.geroliminis}@epfl.ch

SHORT SUMMARY

On-demand travel alternatives offer a myriad of advantages for both service users and service operators. With on-demand trip-sharing, the derived benefits are more accentuated, but little is known about the macroscopic user performance and efficiency of high-capacity on-demand services as a replacement for bus services. In this work, we develop a macroscopic comparison scheme between flexible-route and flexible-schedule micro-transit services, and fixed-route and fixed-schedule public transportation. To analyze the replaceability potential between the two types of services under different graph structures and various spatial demand distributions and intensities, we develop two simulation frameworks to replicate station-based micro-transit and fixed-schedule bus. By examining the values of detour, waiting time, and occupancy under various graph and demand configurations, we infer the performance and efficiency of each service. The results show that the difference in aggregate waiting times is insignificant under relatively low or high demand rates. **Keywords**: Bus services, Dynamic matching and dispatching, On-demand micro-transit.

1 INTRODUCTION

On-demand micro-transit service structure functions as an intermediate between fixed buses and flexible ride-hailing: while its relatively high capacity is similar to that of public transportation, its fully flexible operation and schedule are in line with ride-hailing. Given these characteristics, it has been recently proposed to use on-demand micro-transit services as a substitute for fixed-line buses. However, it remains debatable if high-capacity micro-transit can match the efficiency of more established public transportation options. Efforts in the area aim at providing a classification of the different demand-responsive buses based on service flexibility, service functionality as a feeder or self-standing, and service optimization framework with vehicle-passenger matching and vehicle dispatching Vansteenwegen et al. (2022). Researchers focused on designing fast and efficient real-time exact methods Alonso-Mora et al. (2017), heuristics Santos & Xavier (2013); Jung et al. (2016), or meta-heuristic Ali et al. (2019), to achieve a high-quality matching and routing. Clearly, the key feature of this service is trip-sharing, where pooling efficiency is trip-dependent Soza-Parra et al. (2023); Ke et al. (2021), and is influenced by many factors, including network configuration, demand intensity, and demand distribution. The authors in Molkenthin et al. (2020); Zech et al. (2022) provided scaling laws that output the efficiency of high-capacity ride-sharing services under different network topologies and spatiotemporal demand. In Daganzo et al. (2020), the authors advanced closed-form formulas for performance metrics and detour upper bound guarantees but their approach was restricted to trips with two passengers only. In this work, we address the replaceability of fixed buses with on-demand micro-transit services. By replicating each service independently, we focus on the substitution potential by comparing macroscopic performance metrics. To achieve so, we first develop a comparative framework between the two services and then we run them for various network structures and demand configurations. Finally, we delineate the difference in performance and level of service. The summary of the comparative framework we construct in this work is found in Figure 1.

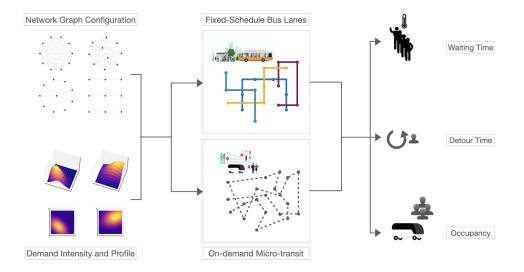


Figure 1: Summary sketch of the approach we adopt.

2 Methodology

Consider a complete network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where \mathcal{V} is the set of network station nodes, and \mathcal{E} is the set of edges between the nodes. The boarding and alighting of passengers happen exclusively at stations \mathcal{V} . We denote by $T \in \mathbb{R}_{\geq 0}^{\mathcal{V} \times \mathcal{V}}$ the matrix defining the travel time on the edge e_{ij} , where $e_{ij} \in \mathcal{E}$ is the edge between two stations i and $j, i, j \in \mathcal{V}$. The travel time between any two stations is strictly positive. As a consequence, together with the completeness assumption of the graph, for all $i, j \in \mathcal{V}$, $T_{ij} = 0$ if and only if i = j. We also make the natural assumption that the matrix T satisfies triangular inequality such that $T_{ij} \leq T_{ik} + T_{kj}$ for $i, j, k \in \mathcal{V}$. The passenger arrival covers a time period $\overline{T} > 0$, and the average hourly passenger demand rate is given by the matrix $Q \in \mathbb{R}_{\geq 0}^{\mathcal{V} \times \mathcal{V}}$, where Q_{ij} is the demand rate between any two nodes $i, j \in \mathcal{V}$. We will throughout the work assume that matrix Q is zero on the diagonal, i.e., $Q_{ii} = 0$ for all $i \in \mathcal{V}$. The total demand for all potential origins and destinations is given by $\lambda = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} Q_{ij}$.

Next, we introduce the set of potential passengers \mathcal{P} , in which each element is the tuple (p^o, p^d, t^p) consisting of the passenger's origin $p^o \in \mathcal{V}$, destination $p^d \in \mathcal{V}$ and arrival at the pick-up location $t^p \in [0, \overline{T}]$. We model the set \mathcal{P} as a set generated by a Poisson process with rates Q. For ever origin $o \in \mathcal{V}$ and every destination $d \in \mathcal{V}$, let $N_{od}(t_a, t_b)$ be the number of passenger arrivals during the time-interval (t_a, t_b) , i.e.,

$$N_{od}(t_a, t_b) = |\{(p^o, p^d, t^p) \in \mathcal{P} \mid p^o = o \land p^d = d \land t_a < t^p \le t_b\}|,$$
(1)

then

$$\mathbb{P}\{N_{od}(t_a, t_b) = n\} = \frac{[Q_{od}(t_b - t_a)]^n}{n!} e^{-Q_{od}(t_b - t_a)}, \quad \forall o, d \in \mathcal{V}, 0 \le t_a < t_b \le \overline{T}.$$
 (2)

Each served passenger $p \in \mathcal{P}$ incurs a waiting time before pick-up that we refer to as w_p , and a detour time that we refer to as Δt_p defined as the additional time traveled by passenger p compared to the direct trip between p^o and p^d .

Irrespective of the type of operating service, the network under consideration has a set of \mathcal{N} in-service vehicles with fleet size $N = |\mathcal{N}|$, each with a capacity $C \in \mathbb{Z}_{>0}$. To every vehicle $n \in \mathcal{N}$ we associate a route $R_n = \{r_1, r_2, ..., r_m\}$ consisting of an ordered sequence of station nodes $r_1, r_2, ..., r_m \in \mathcal{V}$ to be visited by vehicle n, with $m \in \mathbb{Z}^+$ being the length of the route R_n . Additionally, we relate to every route R_n an ordered sequence of station arrival time $A_n = \{a_1, a_2, ..., a_m\}$, where $a_i \in \mathbb{R}_{\geq 0}$ represents the arrival time of vehicle n at station $r_i, i \in [1, m]$. For a given passenger $p \in \mathcal{P}$, we let $\nu(p)$, where $\nu : \mathcal{P} \to \mathcal{N}$, denote the assignment of the passenger to a vehicle $n \in \mathcal{N}$. Moreover, once a vehicle is assigned, let $\sigma_n(p)$, where $\sigma_n : \mathcal{P} \to A_n$ denote the pick-up time of the passenger. Similarly, we let $\tau_n(p), \tau_n : \mathcal{P} \to A_n$, denote the drop-off time of the passenger has not been assigned to a vehicle, and subsequently has no pick-up or drop-off time, the functions will take the value $+\infty$.

The occupancy of vehicle $n \in \mathcal{N}$ at any point in time is given by the function $o_n : [0, \overline{T}] \to \mathbb{Z}$ where

$$o_n(t) = \sum_{p \in \mathcal{P}: \nu(p) = n} \mathbb{1}_{\sigma_n(p) \le t} - \mathbb{1}_{\tau_n(p) \le t} .$$
(3)

Here 1 denotes the indicator function. Once the vehicle routing and passenger assignment are done, the waiting time to be picked up for a passenger $p \in \mathcal{P}$ is given by

$$w_p = \sigma_{\nu(p)}(p) - t^p \,. \tag{4}$$

Similarly, the detour time is given by

$$\Delta t_p = \tau_{\nu(p)}(p) - \sigma_{\nu(p)}(p) - T_{p^o, p^d}, \qquad (5)$$

where, with slight abuse of notation, we let $\Delta t_p = +\infty$ if either $\tau_{\nu(p)}(p)$ or $\sigma_{\nu(p)}(p)$ is infinity. Next, we elaborate on how the route of every operating vehicle R_n and its respective schedule A_n are built and updated, their construction being dependent on the type of service under consideration.

On-demand micro-transit services

On-demand micro-transit services are characterized by a flexible route R_n and flexible schedule A_n that are updated dynamically with the arrival of new potential passengers. A station-based service structure requires a newly arriving request $z \in \mathcal{P}$ to input into the micro-transit platform their origin station z^o , their destination station z^d , and their time of arrival at station t^z . The vehiclepassenger assignment, ν , is based on a first come first served. As soon as the request arrives on the platform, we compute the operational incremental cost of assigning passenger $z \in \mathcal{P}$ to vehicle $n \in \mathcal{N}, C(z, n)$ as the solution to the following optimization problem:

where $\mathcal{P}_n^{t^z} = \{p \in \mathcal{P} \mid t^p < t^z \land \nu(p) = n\}$ is a subset of \mathcal{P} including all the requests that arrived before request z and are matched by the platform to vehicle n. The optimization problem in (6) aims at assigning the newly arriving request with the decision variables being the updated vehicle route R_n and the updated arrival times $A_n, n \in \mathcal{N}$. The objective is to minimize the total passenger delays that is composed of a combination of the waiting time and detour. Furthermore, the first constraint in (6) guarantees that request z is assigned to vehicle n, and the second constraint ensures the waiting time of all the received requests assigned to vehicle n does not exceed the threshold w_{max} . Likewise, the third constraint ensures that the detour time of all passengers assigned to vehicle n does not exceed the threshold Δt_{max} . Finally, the fourth constraint checks if the request assignment does not violate vehicle capacity limits.

Once the cost is computed for every vehicle, the passenger is assigned to the vehicle with the lowest cost, i.e.,

$$\nu(z) = \operatorname*{arg\,min}_{n \in \mathcal{N}} C(z, n) \,. \tag{7}$$

Fixed-schedule bus services

In this part, we elaborate on the approach that we adopt to construct the bus routes R_n , and the bus schedules A_n . Different from the previous on-demand approach, R_n and A_n will be predefined and hence computed offline. The objective is to establish a general approach to design fixed-schedule bus lines for different graph structures \mathcal{G} to eventually compare the two types of flexible and fixed services. For this reason, the design of both services must aim to achieve the same objective, i.e., minimizing the total passenger delays.

Given a set of station nodes \mathcal{V} , we follow a two-step approach to illustrate the design of bus services on the strategic level. First, we start by grouping the different station nodes into various clusters or lines, and then we utilize the travel time between nodes to determine which station nodes belong to which line.

Bus station clustering: Let \mathcal{K} be the set of bus lines grouping all the stations \mathcal{V} of the network \mathcal{G} . We will in this work limit our comparison to a setting where the passengers will not have to perform any bus transfers to reach their destinations. To formulate this problem, we let $X \in \{0, 1\}^{\mathcal{V} \times \mathcal{K}}$ be a binary matrix with element X_{ik} , $i \in \mathcal{V}$, $k \in \mathcal{K}$, equal to 1 if station node *i* belongs to bus line (cluster) *k*, and 0 otherwise. Additionally, it will turn out to be useful to define y_{ij}^k , $i, j \in \mathcal{V}$, $k \in \mathcal{K}$ to be another binary variable equal to the product $X_{ik}X_{jk}$, i.e., $y_{ij}^k = 1$ if both nodes *i* and *j* belongs to bus line *k*. Given the objective of constructing bus lines while minimizing the overlap between clusters, we formulate the bus line station clustering problem as follows

$$\begin{array}{ll}
\underset{X \in \{0,1\}^{\mathcal{V} \times \mathcal{K}}, s \in \geq 0}{\text{minimize}} & s \\ \text{subject to} & X \mathbb{1} > \mathbf{0} \\ & X^T \mathbb{1} \leq s \mathbb{1} \\ & \sum_{k \in \mathcal{K}} y_{ij}^k \geq 1 \\ & y_{ij}^k = X_{ik} X_{jk} \end{array} \quad \forall i, j \in \mathcal{V} \end{array} \tag{8}$$

where s is a decision variable referring to the maximum number of stations per cluster. The first constraint in (8) makes sure that each station node in \mathcal{V} belongs to at least one line. The second constraint ensures that the number of stations per line or cluster is no more than s. The third constraint makes sure that each origin and destination station is covered by at least one cluster. Finally, the fourth constraint ensures that an origin-destination pair is covered by cluster $k \in \mathcal{K}$ only when both the origin and destination belong to k.

Bus line design: Next, we expound on the approach we follow to construct the fixed routes for buses. The objective of constructing bus lines is to eventually determine the set of routes \mathcal{R} , where each route will be an ordered set of the nodes in a cluster $k \in \mathcal{K}$, corresponding to each bus line k, with $R_k \in \mathcal{R}$ being the route that buses operating on line k follows. Note that any vehicle n has a route $R_n \in \mathcal{R}$, and all the vehicles operating on the same bus line have the same route. Therefore, we start by selecting one permutation matrix $P, P \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}}$ that we use to permute the rows of the matrix X^* , X^* being the solution of the optimization problem defined in the previous part. Therefore, by multiplying the two matrices, the output PX^* is one possible permutation of the stations \mathcal{V} over the clusters \mathcal{K} while still abiding by the constraints that we set in the problem formulation in (8). Once the belonging of each node is defined according to the output of the permuted matrix PX^* , we proceed next in determining the sequence of nodes in each cluster to eventually obtain the route R_n for the subset of vehicles $n \in \mathcal{N}$ operating in cluster k. To do so, we resort to a simulated annealing approach to solve the Traveling Salesman Problem (TSP) for each cluster k independently. Therefore, we define $\mathcal{V}_k = \{v \in \mathcal{V} \mid (PX^*e_k)_v = 1\}$ to be the set of nodes belonging to cluster k after the permutation of nodes over the different clusters, with e_k being a $1 \times |\mathcal{K}|$ column vector with entry 1 on the kth element, and 0 otherwise. Once all the sequence of routes for all the clusters are found, the arrival times for the nodes among a route $R_k = \{v_1, v_2, \dots, v_\ell\} \in \mathcal{R}$, is given by

$$a_{v_i}^k = \begin{cases} 0 & \text{if } i = 1, \\ T_{v_{i-1}, v_i} + a_{v_{i-1}}^k & \text{if } i > 1. \end{cases}$$
(9)

The total detour δ for all the possible OD pairs, weighted by the demand intensity between these pairs, can then be computed by

$$\delta = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} \frac{Q_{ij}}{\lambda} \min_{R_k \in \mathcal{R} \mid \{i, j\} \in R_k} ((a_j^k - a_i^k) - T_{i, j})$$
(10)

where a_i^k and a_j^k indicate the arrival time at stations *i* and *j* respectively in route $R_k \in \mathcal{R}$. Moreover, even if more than one bus line serves the same origin and destination, we assume under this scope that passengers utilize the line with the lowest delay. We repeat this procedure for a fixed number of permutations and we retain the best solution found.

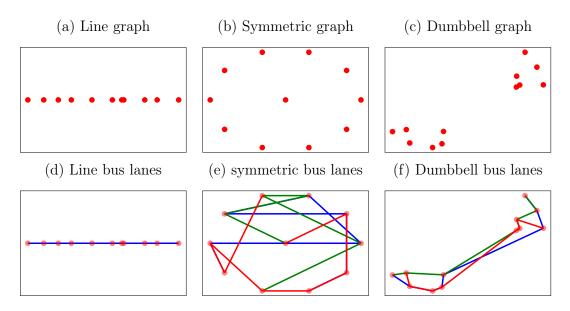


Figure 2: Graph structure and bus lines.



Figure 3: Graph structure and bus lines for Glyfada.

3 Results and discussion

We start by describing the network configurations that we adopt for comparison. Both simulation frameworks run over a period $\overline{T} = 3$ hr. Moving to the network graph structure \mathcal{G} , we choose to run the two services on a line graph, a symmetric graph, and a dumbbell graph as shown in Figures 2(a), 2(b), and 2(c) respectively. We additionally envision the two services running in a real network graph representing the city of Glyfada in Athens with the station locations shown in Figure 3(a). We also define the different bus lines assuming $|\mathcal{K}| = 3$ except for the line graph where we set $|\mathcal{K}| = 1$, such that only one line serves all the stations. The different bus lines for the various graph structures are presented in Figures 2(d), 2(e), and 2(f) for the line graph, symmetric graph, and dumbbell graph, respectively, and for the Glyfada network in Figure 3(b).

In all simulations, we display the results under uniform demand for intensities ranging from 50 to 1500 pax/hr, where for each simulation, we extract the measured values after the occupancy has reached stationarity. The total overall service capacity is always kept at 150 seats. The results for the line graph in Figure 4 show identical waiting and detour time performance for both services. When closely examining the detour times in Figure 4(a), their values are equal to 0 for bus services and fairly non-negligible for micro-transit services. The non-zero values are an artifact of the matching algorithm we use, and therefore the vehicle-request assignment might imply picking up the request at the expense of an inevitable detour. With respect to the waiting time in Figure 4(b), the performance of flexible services with N = 30 vehicles always yields the lowest waiting time, yet the gap between all other scenarios remains insignificant. For the same

fleet and capacity configuration of N = 15 and C = 10, the average user waiting time is slightly better for bus services, but this changes when the average occupancy gets nearly equal to 10 as observed in Figure 4(e). After this point, the micro-transit level of service exceeds that of buses for medium-level demand, and this is justified by a better operational efficiency for micro-transit where vehicles move with higher flexibility in the line graph network, except for high demand rates where fixed-schedule bus services result in lower waiting times.

Regarding the symmetric graph results in Figure 5, we first note the micro-transit configuration with N = 30 vehicles and a capacity of C = 5 seats continues to perform best. Looking at the detour values in Figure 5(a), bus user detour remains constant while for micro-transit, it increases first with the demand rate before stagnating when vehicles reach their capacities as displayed in Figure 5(e). As for the waiting time, micro-transit produces a better service quality with lower waiting times with both the configurations of N = 15, C = 10, and N = 30, C = 5 despite some comparability observed with the bus scenario under N = 15 and C = 10. This justifies the long queue length and the low service rate for bus scenarios in Figure 5(c) and 5(d) respectively.

Moving to the dumbbell graph in Figure 6, we note here that we are comparing on-demand and bus services with similar fleet and capacity configurations, i.e., N = 15, C = 10, but the only difference between the scenarios is the travel time between the two agglomerations in the dumbbell graph, where the term large is added in the legend of Figure 6 to represent instances where the span between the clusters is larger than the default one. When looking at Figure 6(a), we note that for every network graph, the detour is higher for fixed bus service under low demand rates, but when the demand rate increases, the bus trip becomes shorter than micro-transit trips. This observation also applies to the average waiting time graph in Figure 6(b) where micro-transit users incur a larger wait than fixed bus users, despite the service quality of the two being fairly similar for very low demand rates. This means that under saturation when the vehicle occupancy is at capacity as per Figure 6(e), the micro-transit users wait longer for an available vehicle with empty capacity compared to bus users, and their detour is higher where micro-transit trips become less and less efficient under a dumbbell graph structure. To further understand these results, we provide in Figure 6(f) an insight into the ratio of time that vehicles spend moving in between agglomerations for both types of services and different network graphs. For low demand rates, micro-transit vehicles spend shorter times commuting between agglomerations due to the low number of requests, which implies that vehicles are stopped or serving requests within agglomerations for the majority of the time. On the other hand, fixed-schedule buses spend an equal amount of time between agglomerations irrespective of the demand rate, and this amount is dependent on the line configurations. However, when the demand rises, micro-transit vehicles go back and forth between agglomerations more frequently, which justifies the relatively lower waiting time compared to bus services in 6(f). Finally, the last set of results we provide under uniform demand distribution is for the city of Glyfada displayed in Figure 7. With respect to the level of the service, the average detour is the lowest for the micro-transit services with N = 30 due to the low vehicle capacity which causes the trip length to be relatively low despite pooling, as shown in Figure 7(a). Regardless of the relatively significant differences in detour under a low demand rate, the detour values become almost identical for flexible and fixed services under a high demand rate and a fleet-capacity configuration of N = 15, C = 10. This is because of the relatively low occupancy of micro-transit for low demand rate observed in Figure 7(e), which mostly implies direct origin-destination trip types whereas for high demand rate, and subsequently high-occupancy, the detour of micro-transit users increases. For the waiting time values in Figure 7(b), the higher availabilities of micro-transit vehicles cause the lowest values of waiting times, despite their low vehicle capacity.

4 CONCLUSIONS

In this work, we set forth a comparative scheme between on-demand micro-transit and public transit services. We design two simulation-based frameworks to evaluate the performance and efficiency of each service under a line, symmetric, dumbbell, and real case graph. We additionally generate the assessment metric for different demand intensities and non-uniform spatial demand distribution. The results show that user waiting time and detour are the lowest for low capacity large fleet size micro-transit and the highest for high capacity low fleet size public transit. Nevertheless, under the same fleet-capacity configurations, the service performance and user quality of service are fairly identical.

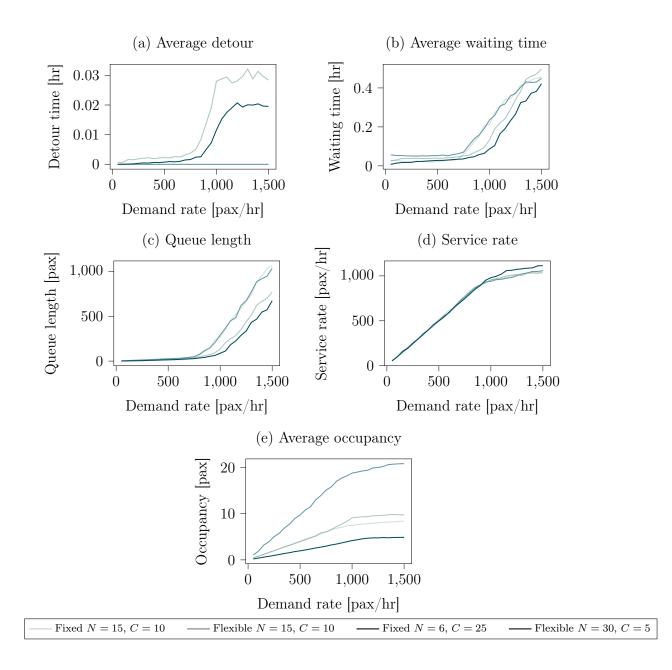


Figure 4: Waiting time, detour time, queue length, service rate, and average occupancy for the line graph network.

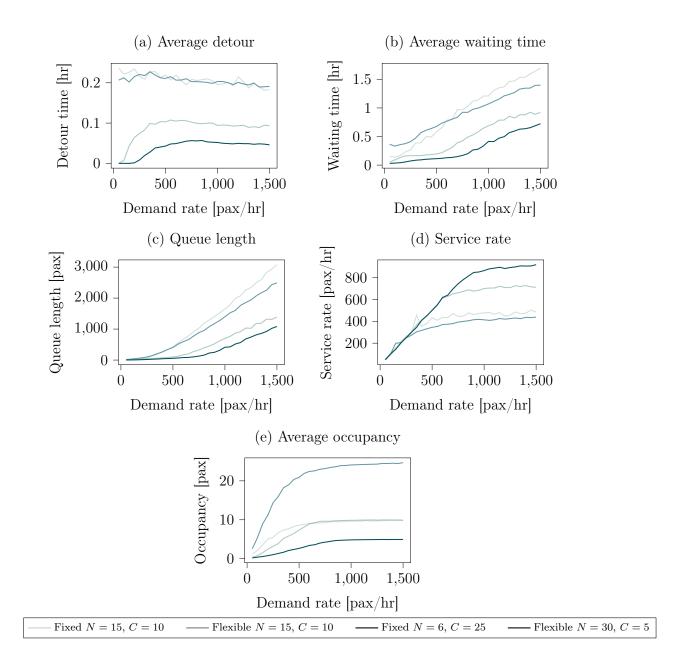


Figure 5: Waiting time, detour time, queue length, service rate, and average occupancy for the symmetric graph network.

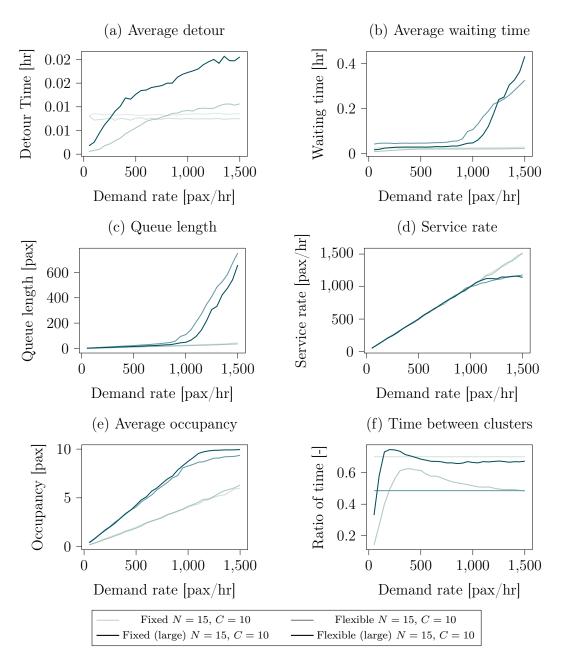


Figure 6: Waiting time, detour time, queue length, service rate, and average occupancy for the dumbbell graph network.

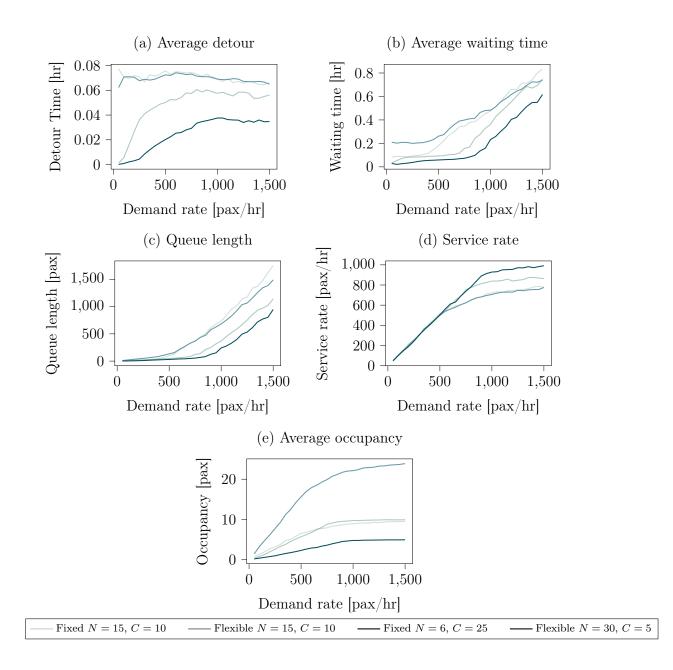


Figure 7: Waiting time, detour time, queue length, service rate, and average occupancy for the Glyfada network

ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40 180545.

References

- Ali, A., Kaltenhäuser, B., Gaponova, I., & Bogenberger, K. (2019, 10). Asynchronous adaptive large neighborhood search algorithm for dynamic matching problem in ride hailing services. In (p. 3006-3012). doi: 10.1109/ITSC.2019.8916943
- Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E., & Rus, D. (2017). On-demand highcapacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy* of Sciences, 114, 201611675. doi: 10.1073/pnas.1611675114
- Daganzo, C. F., Ouyang, Y., & Yang, H. (2020). Analysis of ride-sharing with service time and detour guarantees. Transportation Research Part B: Methodological, 140, 130-150. doi: 10.1016/j.trb.2020.07.005
- Jung, J., Jayakrishnan, R., & Park, J. Y. (2016). Dynamic shared-taxi dispatch algorithm with hybrid-simulated annealing. Computer-Aided Civil and Infrastructure Engineering, 31(4), 275-291. doi: 10.1111/mice.12157
- Ke, J., Zheng, Z., Yang, H., & Ye, J. (2021). Data-driven analysis on matching probability, routing distance and detour distance in ride-pooling services. *Transportation Research Part C: Emerging Technologies*, 124, 102922. doi: 10.1016/j.trc.2020.102922
- Molkenthin, N., Schröder, M., & Timme, M. (2020). Scaling laws of collective ride-sharing dynamics. Phys. Rev. Lett., 125. doi: 10.1103/PhysRevLett.125.248302
- Santos, D. O., & Xavier, E. C. (2013). Dynamic taxi and ridesharing: A framework and heuristics for the optimization problem. In *Proceedings of the twenty-third international joint conference* on artificial intelligence (p. 2885–2891). AAAI Press.
- Soza-Parra, J., Kucharski, R., & Cats, O. (2023). The shareability potential of ride-pooling under alternative spatial demand patterns. *Transportmetrica A Transport Science*. doi: 10.1080/ 23249935.2022.2140022
- Vansteenwegen, P., Melis, L., Aktaş, D., Montenegro, B. D. G., Sartori Vieira, F., & Sörensen, K. (2022). A survey on demand-responsive public bus systems. *Transportation Research Part C: Emerging Technologies*, 137, 103573. doi: 10.1016/j.trc.2022.103573
- Zech, R., Molkenthin, N., Timme, M., & Schröder, M. (2022, 06). Collective dynamics of capacityconstrained ride-pooling fleets. *Scientific Reports*, 12, 10880. doi: 10.1038/s41598-022-14960-x