Multimodal Transportation 2 (2023) 100080

Contents lists available at ScienceDirect

Multimodal Transportation

journal homepage: www.elsevier.com/locate/multra

Full Length Article

An analytical framework for modeling ride pooling efficiency and minimum fleet size

Steffen Mühle

Max-Planck Institute for Dynamics and Self-Organization (MPIDS), Am Faßberg 17, Göttingen 37077, Germany

ARTICLE INFO

Keywords:
Ride pooling
Ride sharing
Demand-responsive transport
Mobility as a service
Sustainability
Street network

ABSTRACT

Ride pooling (RP) is a transport mode using on-demand buses to combine the trips of multiple users into one vehicle. Its required fleet size and carbon emissions are quantified by the system's efficiency. Due to the complex interplay between street network, buses, users and dispatch algorithm, efficiency case studies are available but bottom-up predictions are not. Here we close this gap using probabilistic and geometric arguments in an analytical model framework. Its modular design allows for adaptation to specific usage scenarios and provides an over-arching view of them. In a showcase on Euclidean spaces, our model quantifies how RP outperforms private cars as user demand increases. Its predicted power-law scaling is verified using a custom simulation framework, which further reveals improved scaling on real street networks and graphs with hierarchical structures. Henceforth, our work may help to identify street networks well-suited for RP, and predict key performance indicators analytically.

1. Introduction

Humankind collectively needs to find a sustainable way to manage its resources. One major area with potential for optimization is human travel where, currently, our infrastructure focuses on the paradigm of private car ownership (Bureau of Transportation Statistics, 2011). As a consequence, most households require their own vehicle(s), which are financially costly in terms of purchase and maintenance (Edmonds, 2019), but are idly parked most of the time (Bates and Leibling, 2012) and transport only 1.3 passengers on average when driving (European Environment Agency, 2020). This leads to an unnecessary amount of traffic congestion (Arnott and Small, 1994), carbon emissions (Agency, 2021; Hensher, 2008), valuable urban space dedicated to their convenient parking (Manville and Shoup, 2005), as well as noise (Zwick et al., 2021a) and air pollution (Agency, 2020). All of this begs the urgent question of whether this historically grown status quo can be improved.

1.1. What is ride pooling?

Ride pooling (RP) (or *ride sharing*), classifiable as a stochastic dynamic vehicle routing problem (Agatz et al., 2012; Berbeglia et al., 2010; Psaraftis et al., 2016), is an on-demand shared shuttle bus service that works on short notice. The idea is that users send requests for transportation to a service provider who uses its buses to bring them from their desired pickup to their dropoff locations. Users are not necessarily served one after another, but can be on board at the same time, see Fig. 1. RP thereby opens the potential to offer service quality similar to private cars while drastically reducing traffic (Herminghaus, 2019), and thereby even decrease travel times for RP and private car users alike (Ke et al., 2020b). The number of seats in one bus, its *capacity*, naturally limits the system's attainable efficiency (Merlin, 2019), with buses of capacity 1 being dubbed *taxis*. The decision process of which bus serves which

E-mail addresses: steffen.muehle@ds.mpg.de, steffen.muehle@moia.io, steffen_muehle@gmx.de

https://doi.org/10.1016/j.multra.2023.100080
Received 27 August 2022; Received in revised form 28 November 2022; Accepted 31 December 2022

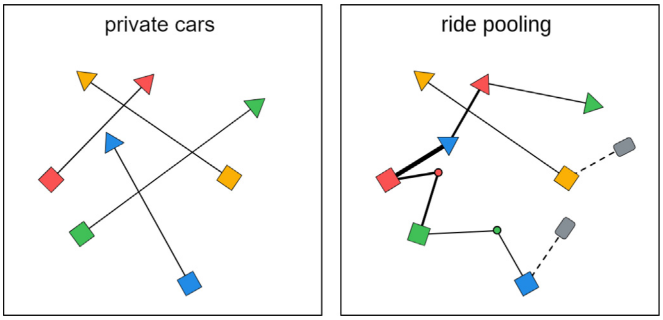

Fig. 1. A comparison of transport modes. Left: Users (colors) travel from their desired pickup (square) to their dropoff (triangle) by private car individually and without detours. Right: Users send their requested journeys to a central ride pooling service provider. The provider sends on-demand shuttle buses (gray rectangles) which may transport several users at once, who may consequently experience a detour. The width of the bus trajectories indicates their occupancy (dashed being empty) and the small circles represent times when the respective user submits their transport request: New information is coming in live while buses are already driving.

user, and in what order, is automated by a dispatch algorithm. It solves a live problem (requests are submitted while buses are already driving) by minimizing some cost function and takes into account user (e.g. waiting time and detour) and bus (e.g. capacity, driver shifts) constraints.

Only in the past decade has large-scale RP become feasible mainly due to the availability of mobile internet (Furuhata et al., 2013; Psaraftis et al., 2016). At the end of 2020, more than 400 RP service providers have been active (Foljanty, 2020a; 2020b; 2021), with more than 100,000 daily shared trips made in 2018 in New York City alone (Schneider, 2022). While it is crucial to understand the users' incentives for opting for RP (de Ruijter et al., 2020; Ke et al., 2020a; Kostorz et al., 2021; Storch et al., 2021; Wang et al., 2018; Wolf et al., 2022), here we are not concerned with their decision process regarding the chosen mode of transport. Instead, given a fixed demand for RP, together with user-given constraints on service quality, we ask how efficient its operation is and how many buses are required.

What is ride pooling efficiency? The definition of RP efficiency

$$\eta := \frac{\sum_{u\in\text{users}} t_u^{\text{direct}}}{\sum_{b\in\text{buses}} t_b^{\text{driven}}} =: \frac{T^{\text{direct}}}{T^{\text{driven}}} \tag{1}$$

employed here is that from the bus company's perspective: the inverse of the total time t_b^{driven} driven by all buses b, measured in units of the total direct trip duration t_u^{direct} requested by all served users u. For visual intuition, when assuming uniform vehicle velocities, η compares the total length of trajectories in the panels of Fig. 1. In Chapter 3, it is demonstrated that η further equals (i) the user demand servable per bus, and (ii) the ratio of bus occupancy and detours (also shown in ref Liebchen et al., 2020). Interpretation (i) allows us to solve the minimum fleet size problem and (ii) to model η analytically. This interplay between detours and occupancy quantifies η and is at the heart of this paper. It should be noted that our definition of efficiency differs from that in refs. Molkenthin et al. (2020), Zech et al. (2022).

What is the minimum fleet problem? Let us consider three quantities: The fleet size (number of RP buses), the frequency with which users send requests for transportation, and the served percentage of incoming requests (later on called N, x and p_{accept} respectively). The minimum fleet problem asks, when given two of them, what the third one is. For example, a service provider may aim to serve 90 % of a given user demand, and ask how many buses are required.

1.2. Objective

Currently, there is a missing link between purely analytical treatments of RP and performance indicators that are important for RP service providers, namely RP efficiency and minimum fleet size. The objective of this paper is to close this gap in a generic manner, i.e. without relying on simulation results or optimization procedures applied to specific scenarios. This serves the purposes of analyzing the causal relation between the input (user demand, fleet size, dispatch algorithm, street network) and output (performance indicators) of RP systems systematically, and revealing scaling laws that are valid for (classes of) RP systems in general. Hence, we shall adopt the mind set of breaking RP dynamics down to its most essential features rather than mimicking particular real-world implementations with great detail.

1.3. Contribution

We present a modular framework which achieves this task. It can be adapted to specific usage scenarios, providing a common language and an over-arching view for them. As a proof of principle, we apply our model to a class of street networks with Euclidean topology. There it is found that, for sufficient user demand x, efficiency grows logarithmically with x and that the relationship between fleet size N and servable demand $x(N)$ closely follows a power-law with exponent 1.15. The model's results are then verified via extensive simulations using a custom framework on multiple (real and synthetic) street networks. Among the networks used for simulations, the exponent of 1.15 can be considered as a worst-case scenario due to the Euclidean plane's complete lack of hierarchical structures, with a real street network having exponent 1.24 and a tree-like network surpassing them both. These results confirm the intuition that RP inevitably outperforms private cars ($x \propto N$) as user demand increases. In the future, our model may help to interpret studies on RP efficiency and fleet size, and to identify street networks well-suited for RP.

The paper is structured as follows: In Chapter 2, literature related to this paper is reviewed. In Chapter 3, the basic quantities of RP are defined and their general relationships discussed. It is demonstrated how efficiency is related to the minimum fleet problem. In Chapter 4, a generic analytical model for RP efficiency is introduced. Chapter 5 showcases the model and its predictions by applying it to RP on a square map. Chapter 6 shows results obtained from simulating RP in a custom simulation framework. Chapter 7 constitutes a discussion and summary of our results, as well as an outlook for future work.

2. Related work

A recent review of theoretical studies on RP is available in Zwick et al. (2022), and an overview of RP efficiency values reported in literature in Ruch et al. (2020). In the following, we summarize literature on the dependence of RP efficiency on bus capacity, geometry, dispatcher and demand, and then discuss analytical treatments related to ours.

2.1. What does ride pooling efficiency depend on?

- While more available vehicles (Moreno et al., 2018) and increased user demand (Boesch et al., 2016; Vazifeh et al., 2018) improve the efficiency of on-demand services with capacity 1 (taxis), it remains below 1 due to empty trips (Henao and Marshall, 2019). For capacity greater than 1 (RP) users can be transported together and, consequently, taxis are outperformed in terms of efficiency (Bischoff et al., 2017; Ma et al., 2013; Pernestål and Kristoffersson, 2019; Qian et al., 2017; Zhu and Mo, 2022) and required fleet size (Alonso-Mora et al., 2017; Lokhandwala and Cai, 2018; Pernestål and Kristoffersson, 2019). In realistic settings, however, it is found that a capacity of 3 (Bischoff et al., 2017) or 4 (Fagnant and Kockelman, 2018) suffices.
- Different street networks are disparately well-suited for RP (Manik and Molkenthin, 2020; Zwick and Axhausen, 2022) and other transport modes (Aldous and Barthelemy, 2019; Folco et al., 2022; Szell et al., 2022). On any network, the question of from where to where users wish to be transported is also crucial: Directionality decreases efficiency due to empty trips (de Ruijter et al., 2020), while introducing meeting points (Fielbaum et al., 2021; Lotze et al., 2022; Stiglic et al., 2015) increases it due to reduced detours.
- Naturally, the utilized dispatch algorithm influences the achieved efficiency as well (Agatz et al., 2011; Hörl et al., 2018; Horn, 2002; Jung et al., 2016; Zwick and Axhausen, 2020). Ref. Ruch et al. (2020) compares the performance of RP using various algorithms from literature, finding that the all-time optimal dispatcher Alonso-Mora et al. (2017) came out on top but is computationally expensive (which was improved in Engelhardt et al., 2020).
- As user demand increases, more users can be pooled together, increasing RP efficiency (Agatz et al., 2011; Fagnant and Kockelman, 2018), as each of them can expect to find more other users with similar itineraries (Santi et al., 2014; Tachet et al., 2017). Two simulation studies of particular interest for the present paper are the following. Firstly, Kaddoura and Schlenther (2021) systematically varied demand and adopted the fleet size such that a certain level of service quality is maintained, finding that with rising demand, the system's efficiency increases, while the necessary increase in fleet size per unit demand decreases. Secondly, Zwick et al. (2021b) simulated RP with 12 different levels of demand, empirically showing that bus occupancy, closely linked to efficiency, grows logarithmically with demand.

2.2. Analytical approaches to ride pooling

Since RP dynamics are subject to the complex interplay between details of the street network, request submissions, the dispatch algorithm etc., simplifying assumptions are necessary to identify aspects common to RP in general. Simplifications are often of geometric nature (for instance, Lobel and Martin (2020) considers a 2-dimensional grid, detecting fundamental constraints on the trade-off between detours and saved driven distance due to pooling) or focus on the special case of bus capacity 2 (Daganzo et al., 2020; Ke et al., 2020a; 2020b; Lobel and Martin, 2020; 2020; Zech et al., 2022).

Closely related to our work, two mean-field approaches to RP on the Euclidean plane must be pointed out. Firstly, Herminghaus (2019) estimates detours, waiting times and efficiency by making use of generic geometric arguments and dimensional analysis. However, the average bus occupancy is taken to be a given input parameter. Secondly, Tachet et al. (2017) analytically found a universal scaling law between demand and the probability (”shareability”) that a user finds at least one other user to potentially share their ride with (first introduced in Santi et al., 2014 and elaborated on in Bilali et al., 2020). Shareability can be regarded as an important step towards quantifying RP efficiency in a bottom-up manner. The fact that its scaling law is monotonously growing indicates that efficiency should increase with demand, which is supported by the logarithmic growth of bus occupancy found in Zwick et al. (2021b). However, to the best of our knowledge, no quantitative link between shareability and efficiency is currently known, and no theoretical bottom-up study is available that can model RP efficiency without relying on simulation results.

Our main objective is thus to predict RP efficiency in an analytical manner. This requires simultaneous models of both detours and the average bus occupancy in a given RP setup.

3. Basics

In this chapter, we introduce quantities relevant to our efficiency model and establish their general relationships. Modeling them will be done in Chapter 4.

3.1. Paradigm and notation

Throughout this paper, the RP paradigm is this: Users send requests for transportation to a central service provider. Requests consist of a pickup location, a dropoff location, which are both located on a given transport space (dubbed the *map*), a submission time, and a maximally allowed detour $\delta_{\max}$. The provider commands a fleet of N buses, receives new requests with a frequency f and decides which requests are served by which bus and in which order. For this purpose, a dispatch algorithm is employed that operates according to a cost function while fulfilling the constraints of all users simultaneously. If the latter is impossible when accepting an incoming request, the dispatcher will reject it instead. Rejections are only permitted upon submission. Due to this paradigm, none of the served users' quality constraints are violated, the logic being that they would otherwise not opt for RP.

In the following, averages over served (i.e. not rejected) users are denoted as $\langle\cdot\rangle_u$, and time-averages as $\langle\cdot\rangle$.

3.2. Bus status

At any time, each bus is either driving or idling, and it is either occupied or empty. Considering a random bus at a random time, the probability to find it driving is denoted as p_{driving} and the probability to find at least one user on board as p_{occupied}. Buses only idle when empty, but driving does not imply being occupied: $p(\text{occupied}|\text{driving})$ is the conditional probability that a driving bus is occupied.

3.3. Bus occupancy

Let us denote the number of users on a single bus as the bus occupancy $b = 0, 1, 2, \ldots$. Then $\langle b\rangle$ is its time-average as experienced by the bus drivers. Excluding time intervals from the time-average when the bus is idling ($\langle b\rangle_{\text{driving}}$), or has no passengers on board ($\langle b\rangle_{\text{occupied}}$), the following identity holds:

$$\langle b\rangle = p_{\text{driving}}\langle b\rangle_{\text{driving}} = p_{\text{occupied}}\langle b\rangle_{\text{occupied}} \tag{2}$$

Next, let us consider a sufficiently long time interval of length T during which requests are accepted with probability p_{accept} such that $U \approx T\, f\, p_{\text{accept}}$ users $u = 1, \ldots, U$ are served and the rest rejected. Denoting the times these users spend on the bus as t_u^{bus}, the occupancy of all buses combined is

$$N\langle b\rangle = \frac{\sum_u t_u^{\text{bus}}}{T} = \frac{\sum_u t_u^{\text{bus}}}{U} \cdot \frac{U}{T} = \langle t^{\text{bus}}\rangle_u \cdot f\, p_{\text{accept}} \tag{3}$$

3.4. Detour factor

Let us denote the direct requested travel times of served users as t_u^{direct}. Then the ratio $t_u^{\text{bus}}/t_u^{\text{direct}} \geq 1$ defines a user-specific detour factor, comparing the travel time via RP with that of a private car. Likewise, the system-wide detour factor is defined as

$$\delta^{\text{system}} := \frac{T^{\text{bus}}}{T^{\text{direct}}} = \frac{\langle t^{\text{bus}}\rangle_u}{\langle t^{\text{direct}}\rangle_u} \tag{4}$$

where $T^{\text{bus, direct}} = \sum_u t_u^{\text{bus, direct}}$.

3.5. Demand

The average requested direct trip duration of *all* users (served or rejected), t_0, does not depend on the actual RP dynamics, but only on the map. It makes for a natural time scale that allows us to introduce the *demand*

$$x := f \cdot t_0 \tag{5}$$

which is a dimensionless request frequency that renders the dynamics on differently sized maps comparable. Its numeric value is best interpreted as the average number of private cars driving on the map at any given time if all users opted for that transport mode instead[1]. In realistic scenarios where most incoming requests are served, t_0 and $\langle t_u^{\text{direct}}\rangle_u$ can be expected to coincide well – an approximation we immediately employ.

[1] If you do not find this obvious, consider Eq. (8) with $\eta = p_{\text{accept}} = 1$ after finishing this chapter.

3.6. Supply-demand balance

Inserting Eqs. (2), (4) and (5) into Eq. (3) reveals the balance between supply (buses driving with some transport efficiency) and demand (user requests to be served):

$$\underbrace{N\,\frac{\langle b\rangle_{\text{driving}}}{\delta^{\text{system}}}}_{\text{supply}}\,p_{\text{driving}} = \underbrace{x}_{\text{demand}}\,p_{\text{accept}} \tag{6}$$

An excess of demand (supply) is balanced by a decrease in p_{accept} (p_{driving}) to match the left (right) hand side. This equation hence also captures scenarios in which demand exceeds supply ($p_{\text{driving}} \approx 1$ and $f \cdot p_{\text{accept}}$ is the frequency with which requests are served) or vice versa ($p_{\text{accept}} \approx 1$ and $N \cdot p_{\text{driving}}$ is the effective number of driving buses).

3.7. Efficiency

Reconsidering the RP efficiency defined in Eq. (1), it equals the ratio of bus occupancy and thus entailed detours:

$$\eta = \frac{T_{\text{direct}}}{T_{\text{driven}}} = \frac{T_{\text{bus}}/T_{\text{driven}}}{T_{\text{bus}}/T_{\text{direct}}} = \frac{\langle b\rangle_{\text{driving}}}{\delta^{\text{system}}} \tag{7}$$

Here, we have used Eqs. (2), (3), and that the bus fleet's driving total time equals $T_{\text{driven}} = N\,T\,p_{\text{driving}}$ for identifying the nominator as $\langle b\rangle_{\text{driving}}$.

Other than its definition in terms of driven distance, η thus also quantifies the trade-off between users being pooled together and experiencing detours (Liebchen et al., 2020). A third interpretation of it, which is shown in the following, is the ratio of served demand and the number of driving buses.

3.8. Minimum fleet problem

Inserting Eq. (7) into Eq. (6), one obtains

$$x = \eta(x, N) \cdot N \cdot \frac{p_{\text{driving}}}{p_{\text{accept}}} \tag{8}$$

This constitutes a single equation linking together x, p_{accept} and N as long as the following assumption holds:

$$\boxed{\text{Assumption 1:}\quad p_{\text{driving}} = 1} \tag{9}$$

The assumption is justified like this: In economically realistic scenarios, supply and demand are well balanced and it is reasonable to assume that buses are driving most of the time while serving as much demand as they can. It will be numerically tested in Chapter 6, and turns out to hold true for sufficient demand.

Making use of it, solving the minimum fleet problem via Eq. (8) is reduced to having at hand the functional form of η. Fixing a target percentage of served requests, say $p_{\text{accept}} = 0.8$, η thus relates driving buses to served transport requests: one RP bus replaces η private cars and saving driven distance is thus equivalent to serving more customers.

4. Efficiency model

Next we present a generic, modular modeling framework with the goal of obtaining an analytical expression for η. It does not rely on simulation results, but only on ingredients required to start one: the setup $\mathcal{S}$, which is the collection of x, N, a map, spatial request pattern and user constraints. It is agnostic of the dispatch algorithm and adaptable to different $\mathcal{S}$.

4.1. Single-user view

The model's core idea is to consider the efficiency as it is experienced by a single RP user. Consider for example the blue user's journey in Fig. 1, who can report a bus occupancy b^{user} and a detour δ^{user} based on the experienced trip. Their ratio serves as a measure of the system's overall efficiency $\eta = \langle b\rangle_{\text{driving}}/\delta^{\text{system}}$. The representative single-user's (from here on called *our*) view yields the advantage that certain probabilistic statements can be made about the bus trajectory between our pickup and dropoff. Let us assume our trip to have the fixed duration t_0 at the moment of request submission, but that b^{user} and δ^{user} are realizations of a stochastic process. An ensemble average over their ratio, $\eta^{\text{user}} = b^{\text{user}}/\delta^{\text{user}}$, then yields

$$\eta(\mathcal{S}) \approx \int \mathrm{d}\eta^{\text{user}}\eta^{\text{user}}p(\eta^{\text{user}}|\mathcal{S}) \tag{10}$$

This replacement inflicts the bias that the bus can never be empty as at least the representative user is on board. As η is only affected by driving buses, this bias is inconsequential if buses were not driving empty in the first place, i.e. when

$$\boxed{\text{Assumption 2:}\quad p(\text{occupied}|\text{driving}) = 1} \tag{11}$$

holds.[2] The single-user view's purpose is that it reduces finding η and solving the minimum fleet problem to modeling the probability distribution $p(\eta^{\text{user}}|\mathcal{S})$, which is our next task. To this end, we must first introduce two auxiliary variables whose purpose will become clear shortly after.

4.2. Auxiliary variables M and k

M denotes the number of other requests being submitted during a time interval t_0 representing our trip. Not all M other users are necessarily being served – any number of them may get rejected. M represents the number of users we can *potentially* meet on the bus, simply because they submit their requests at similar times as we do. Its expected value is $\langle M\rangle = f \cdot t_0 = x$.

k denotes the number of users we in fact meet on the bus. For each of them, we either experience one or two additional stops on our journey due to those users' pickup and/or dropoffs, or zero extra stops in case they enter before and exit after us. For simplicity, we assume that each of the k met users causes one extra stop on our journey, and is on the bus for a percentage χ of its duration. Hence the bus occupancy we experience is

$$b^{\text{user}} = 1 + \chi \cdot k =: b_k \tag{12}$$

where the 1 reflects our own presence on the bus. Typically, $\chi \approx 0.5$ but in general it may depend on $\mathcal{S}$.

4.3. Bayesian chain

Let us now expand $p(\eta^{\text{user}}|\mathcal{S})$ into a chain of conditional probabilities. Without approximation, it reads

$$\begin{aligned} p(\eta^{\text{user}}|\mathcal{S}) &= \sum_{k,M} p(\eta^{\text{user}}, k, M|\mathcal{S}) \\ &= \sum_{k,M} p(\eta^{\text{user}}|k, M, \mathcal{S}) p(k, M|\mathcal{S}) \\ &= \sum_{k,M} p(\eta^{\text{user}}|k, M, \mathcal{S}) p(k|M, \mathcal{S}) p(M|\mathcal{S}) \end{aligned} \tag{13}$$

Next $\mathcal{S}$ is dropped from our notation for readability and the assumption is made that the conditional probability of η^{user} may depend on k but not on M:

$$\boxed{\text{Assumption 3:} \quad p(\eta^{\text{user}}|k, M) = p(\eta^{\text{user}}|k)} \tag{14}$$

It means that for our journey's features, it matters only how many other users we meet on the bus, not how many others exist in the overall system. This simplifies Eq. (13) to

$$p(\eta^{\text{user}}) = \sum_{k,M} p(\eta^{\text{user}}|k) p(k|M) p(M) \tag{15}$$

Expression (15) should be understood as a systematic way of breaking down the task of modeling $p(\eta^{\text{user}})$ into the series of subtasks of modeling conditional probabilities. The reason for the particular choices made in defining k and M is that they make these subtasks feasible and assumption 3 reasonable. Let us now go through the three terms in their hierarchical order.

- $p(M)$ quantifies the temporal statistics of request submissions. It can simply be modeled as a Poisson process in time. Since its mean value is x,

  $$p(M) = e^{-x} \frac{x^M}{M!} \tag{16}$$

- $p(k|M)$ encodes how many users we are likely to meet on the bus, given that we could meet up to M. For this purpose, we developed what we call an *insertion model*, which is the subject of the next subsection.
- $p(\eta^{\text{user}}|k)$ assesses the experienced efficiency, given that k users are met on the bus. Since $b_k = 1 + \chi \cdot k$, the remaining question is this: What detour δ_k can be expected on a route with k extra stops? Its answer will depend on $\mathcal{S}$ and is thus not part of the general framework, but treated in the next chapter. Once found, we use

  $$p(\eta^{\text{user}}|k) = \delta(\eta^{\text{user}} - \frac{b_k}{\delta_k}) \tag{17}$$

  where the non-indexed δ denotes the delta distribution.

The last point can be made use of immediately and the delta distribution be resolved. Combining this with Eqs. (10) and (15), one may write

$$\eta \approx \sum_{k,M} \frac{1 + \chi \cdot k}{\delta_k} p(k|M) p(M) \tag{18}$$

[2] In order for the single-user view to accurately solve the minimum fleet problem, assumptions 1 and 2 both need to hold true. Since occupied buses are never idling, they combine into $p_{\text{occupied}} = 1$.

4.4. Insertion model for $p(k|M)$

If $M = 0$, there is no other user we could meet on the bus, meaning $k = 0$, and thus $p(k|0) = \delta_{k,0}$ where $\delta_{k,M}$ is the Kronecker delta. If $M = 1$, there is one other user whose request temporally overlaps with ours and the question of whether we meet on the bus can be considered a geometric one: If that user's requested journey is close and aligned to ours, any sensible dispatcher would pool us together[3] and therefore

$$p(k|1) = \delta_{k,0} \cdot (1 - R_0) + \delta_{k,1} \cdot R_0 \tag{19}$$

where R_k is the probability that a new incoming request can be inserted into our route when it already has k extra stops. The idea behind the insertion model is to iterate this process: considering all M other users one at a time, with each of them the probability distribution $p(k|M)$ flows towards greater values of k. Following this logic, its evolution (M playing the role of time) is quantified by the following Master equation[4]:

$$p(k|M+1) = p(k|M) + \overbrace{p(k-1|M)R_{k-1}}^{\text{inflow from k-1}} - \overbrace{p(k|M)R_k}^{\text{outflow to k+1}} \tag{20}$$

Given the functional form of R_k, Eq. (20) is easily solved in an iterative manner. It is the final piece of the puzzle to obtain a completely bottom-up expression for η. Since it depends on $\mathcal{S}$, it will only be further specified in the next chapter.

4.5. Application recipe

In summary, the workflow for applying the presented model framework is the following:

1. Specify map and spatial request pattern. Estimate overlap of shared routes χ. Obtain mean route duration t_0.
2. Specify user constraints. Geometrically model trajectory length growth δ_k and insertion probability R_k.
3. Choose dimensionless demand x. Together with temporal request submission statistics, obtain p(M).
4. Solve Master Eq. (20) iteratively and obtain p(k|M).
5. Evaluate efficiency according to Eq. (18).
6. Using steps 3-5, find pairs (N,x) that satisfy the supply-demand balance (8) to solve the minimum fleet problem.

5. Application of model framework

As a proof-of-principle, the model framework will now be applied to RP on a Euclidean square map. Its size is chosen such that the mean distance between two randomly chosen points on it is 1, and vehicles move with unit speed such that $t_0 = 1$. The former property is attained by setting the map's side length s to $s = (2 + \sqrt{2} + 5\log(1 + \sqrt{2}))/15 \approx 1.92$ and, consequently, the map's area is $A = s^2 \approx 3.68$. Requests are submitted according to the Poisson process in Eq. (16) with mean x, and a user-given maximum allowed detour $\delta_{\max} = 2.0$ must be adhered to. Pickup and dropoff locations are independently and uniformly distributed on the map. In the spirit of a reasonable average, we assume that each other traveller met on the bus is staying on that bus for half the duration of our own trip, i.e. $\chi = 0.5$. Boundary effects are neglected.

5.1. Detour model

Denoting our pickup as A and dropoff as B, the journey for $k = 0$ is a straight line from A to B with duration $t_0 = |A - B| = \delta_0 = 1$. Our maximally allowed journey duration is $\delta_{\max}$ and the set of all points X for which $\delta_1 = |A - X| + |X - B| < \delta_{\max}$ defines the inside of an ellipse $(\delta_0, \delta_{\max})$, which is shown in Fig. 2a. Here (a, b) denotes an ellipse with distance a between its focal points and distance b from focal point to border to focal point. Its area is $\frac{\pi}{4} b \cdot \sqrt{b^2 - a^2}$. As shown in Herminghaus (2019), the mean detour factor caused by adding a random point inside an ellipse $(1, \delta_{\max})$ is $\bar{\delta} = 2\delta_{\max}/3 + 1/(3\delta_{\max})$, which we thus employ as $\delta_1 = \bar{\delta}$.

For simplicity, we assume that in subsequent steps the remaining detour $\delta_{\max} - \delta_k$ shrinks by the same factor, i.e.

$$\delta_k = \delta_{\max} - (\delta_{\max} - 1) \cdot \left(\frac{\delta_{\max} - \bar{\delta}}{\delta_{\max} - 1} \right)^k \tag{21}$$

The presented ellipse-based model is inspired by the one presented in the mean-field theory in Herminghaus (2019) with the difference that here, the threshold $\delta_{\max}$ can never be exceeded.

[3] This is assumed to be true for any sensible optimization objectives of the dispatch algorithm, of which we intentionally remain agnostic here.

[4] As a side note, a finite bus capacity $b_{\max} = 1 + \chi \cdot k_{\max}$ could be implemented via $R_{k_{\max}} = 0$.

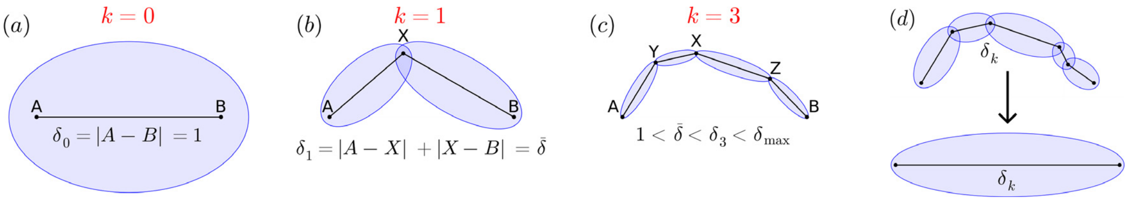
(a)
k = 0
A
B
$\delta_0 = |A - B| = 1$
(b)
k = 1
X
A
B
$\delta_1 = |A - X| + |X - B| = \bar{\delta}$
(c)
k = 3
Y
X
Z
A
B
$1 < \bar{\delta} < \delta_3 < \delta_{\max}$
(d)
δ_k
δ_k

Fig. 2. The ellipse-based trajectory model. Stops are added to the journey from our pickup A to dropoff B. Ellipses define sets of all allowed extra stop sites keeping total journey duration below a threshold $\delta_{\max}$. **(a)** Our originally requested trip. With $k = 0$ other users met, there is no detour and thus $\delta_0 = 1$. **(b)** Experiencing one extra stop at X incurs a mean detour $\bar{\delta}$. After that insertion, the allowed additional detour defines the width of the two ellipses between A and X or X and B respectively. **(c)** As more stops are added, the journey length δ_k approaches $\delta_{\max}$ and the allowed ellipses for the next stop insertion become increasingly slim. **(d)** A simple heuristic for estimating the area of the union of overlapping ellipses around our trajectory (top). Our route is flattened, preserving its length, such that a single ellipse around it can be considered (bottom).

5.2. Insertion probabilities

As each request consists of an independently chosen pickup and a dropoff site, we take

$$R_k = p_k^{\text{pickup}} \cdot p_k^{\text{dropoff}} \tag{22}$$

to be the product of two factors. The first factor, p_k^{pickup}, represents the probability that a randomly chosen point on the map lies within the set of allowed points for extra stops around our current route. It equals this set's area divided by A. Neglecting boundary effects, for $k = 0$ the set is the ellipse $(\delta_0, \delta_{\max})$ and in subsequent steps, it is a union of overlapping ellipses as seen in Fig. 2b,c. For analytical tractability, let us replace the area of this union with the area of the ellipse $(\delta_k, \delta_{\max})$. Intuitively, the current (winded) trajectory with length δ_k is flattened to a straight line, preserving its length, such that a single ellipse around it can be considered, see Fig. 2d. This replacement overestimates the union's area and was thus heuristically divided by a factor of 2[5]:

$$p_k^{\text{pickup}} = \frac{\pi}{8\,A}\delta_{\max} \cdot \sqrt{\delta_{\max}^2 - \delta_k^2} \tag{23}$$

The second factor, p_k^{dropoff}, represents the probability that the dropoff location can be inserted onto our route as well. Comparing p_k^{dropoff} to p_k^{pickup}, there is an effect increasing it and an effect decreasing it: It is increased because the dropoff may lie behind ours where the bus' future trajectory is not yet determined and the dispatcher is relatively free to schedule as desired. However, it is decreased due to the constraint that the dropoff cannot occur *before* the pickup. For simplicity, we assume that the two effects roughly cancel each other out such that $p_k^{\text{dropoff}} = p_k^{\text{pickup}}$ and thus

$$R_k = \left(\frac{\pi\,\delta_{\max}}{8\,A}\right)^2 \cdot \left(\delta_{\max} + \delta_k\right) \cdot \left(\delta_{\max} - \delta_k\right) \tag{24}$$

Note that by inserting Eq. (21) into Eq. (24) one finds that R_k decays exponentially as k increases.

5.3. Model results

Inserting Eq. (24) into Eq. (20) yields $p(k|M)$. Together with the Poisson distribution (16) and the trajectory model (21) it can be used to evaluate η via Eq. (18). This was then used to obtain the function $x(N)$ (using p_{accept}=0.8) from Eq. (8). Figure 3 shows the results, where they are also described.

For low demand, the model is strongly affected by the single-user view's assumption that at least one user is on board. This assumption will be numerically tested in Chapter 6, and amending its bias will be discussed in Chapter 7. In the following, the model results for high demand are interpreted.

For increasing demand x, three effects take place. Firstly, the number of sent requests M grows. In fact, the Poisson distribution 16 becomes sharply peaked around its mean, which is x, such that we may replace $M \approx x$. Secondly, the distribution $p(k|M)$ drifts towards larger values of k, and thus $\partial_x \langle k \rangle > 0$. This shift becomes less pronounced as x increases, hence $\partial_x^2 \langle k \rangle < 0$. This is due to the monotonic decay of the transition rates R_k and thus ultimately to the shrinking of ellipse areas in Fig. 2. Thirdly, the efficiency in Eq. (18) becomes proportional to k since the denominator is bound while the nominator is not. Taken together, it is for these reasons that $\eta(x)$ has the general features $\partial_x \eta > 0$ and $\partial_x^2 \eta < 0$. Its functional form will, however, depend upon that of R_k.

For our choice of modeling R_k, $\eta(x)$ is well approximated by a logarithm (x-axis in Fig. 3b is logarithmic), a feature also empirically found in Zwick et al. (2021b). Here, we can explain this logarithmic growth as follows: A unit increase Δk is caused by $\Delta M \approx R_k^{-1}$, which, together with $x \approx M$ and $\eta \propto k$, gives rise to the following scaling:

$$\frac{\mathrm{d}\eta}{\mathrm{d}x} \propto \frac{\Delta k}{\Delta M} \approx R_k \propto \exp(-k/\text{const.}) \propto \exp(-\eta/\text{const.}) \tag{25}$$

[5] For uniformity, we also divide p_0^{pickup} by 2, which can be partly justified by neglected boundary effects. In any case, R_0 does not affect the model for large x where $p(k = 0) \to 0$ and, according to Eq. (20), rescaling *all* R_k uniformly may be considered rescaling M and thus x.

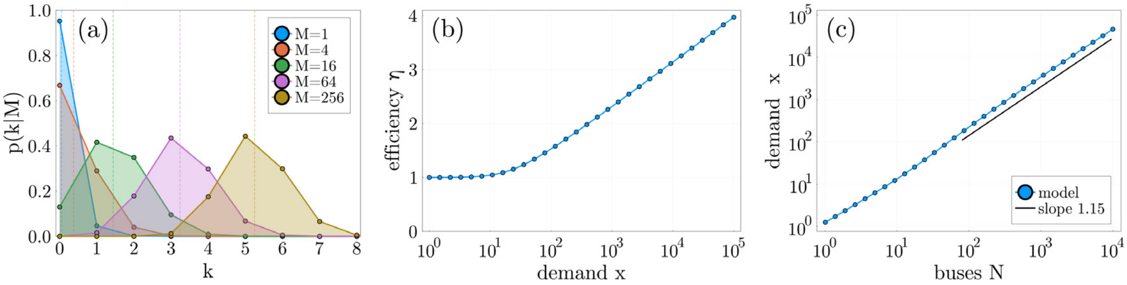

Fig. 3. Model results. Purely analytical model results for RP on a square map with $\delta_{\max} = 2.0$. **(a)** The evolution of $p(k|M)$ with varying M. The distribution drifts strictly towards the right, but does so increasingly slowly with growing M. Mean values are shown as vertical dashed lines. The corresponding bus occupancies are $b = 1 + k/2$. **(b)** The evaluated model expression in Eq. (18) for increasing demand. The fact that η never drops below 1 is an artifact from the biased single-user view in which there is always at least one user on board, and should thus not be over-interpreted for small x. The curve is increases monotonically and shows logarithmic growth beyond $x \approx 100$. **(c)** The corresponding solution to the minimum fleet size problem in Eq. (8) with a target served percentage of incoming requests of 80 %. After an initial linear increase due to η being constant in panel (b), the demand servable by N buses appears to be reasonably well-approximated by a power-law with exponent ≈ 1.15 as indicated by the black line.

This can be read as an ordinary differential equation for $\eta(x)$ which is indeed solved by $\eta \propto \ln(\text{const.} + \text{const.} \cdot x)$.

Turning our attention to the minimum fleet problem, the function $x(N)$ is thus the solution to Eq. (8) with η being a logarithm of x. This solution is not a power-law, but is reasonably well approximated by one as displayed in Fig. 3c. The fact that its exponent exceeds 1 ($x(N)$ is concave) fits together with the finding in Kaddoura and Schlenther (2021) that $N(x)$ is concave, i.e. that the necessary increase in fleet size decreases with demand.

The only free parameter chosen here is the detour threshold $\delta_{\max} = 2.0$. It does not have a qualitative influence on the shown results. This is because increasing $\delta_{\max}$ triggers two opposing effects: Increased ellipse areas lead to higher k, thus improving efficiency, but at the same time increased average detours decrease efficiency. This is strongly related to the findings in Lobel and Martin (2020), but is not a subject of the present paper.

5.4. Outlook: Hierarchical networks

It is clear that the scaling behavior just described does not apply to all transport spaces. Many finite graphs and street networks cannot be treated as an area, but as having a finite total length. The geometric arguments (ellipse areas) from above then no longer apply, which affects the functional form of R_k. Instead, the following effect comes in: There is a non-zero probability that a new request lies exactly on a bus' route and thus does not incur any additional detour at all. From our model's perspective, this leads to R_k no longer decaying exponentially but remaining finite. Following the line of reasoning above for large x, $\eta \propto x + \text{const.}$ then grows linearly with demand, not logarithmically. The corresponding solution to the minimum fleet problem, $x \propto N/(1 - N \cdot \text{const.})$, has the remarkable interpretation that a finite number of buses can serve arbitrary demand, but obviously breaks down when finite stop durations and bus capacities are no longer neglected.

6. Simulations

Our custom simulation framework follows the RP paradigm discussed in Chapter 3. Its spirit is to reduce RP dynamics to its essential features rather than aim to be as realistic as possible, and its purpose is to test the model's assumptions and predictions. An overview of simulation studies available in literature can be found in Markov et al. (2021). In relation to other dispatch algorithms, ours can be classified as follows. Idle buses are not relocated (no *rebalancing* Lu et al., 2021; Wen et al., 2017) but stay put, the bus capacity is unlimited, stop times are neglected (i.e. pickups and dropoffs are instantaneous), and demand data is synthetic. An insertion heuristic similar to that in Bischoff et al. (2017), in which the best possible insertion of a pickup and dropoff job are selected via a cost function (minimizing total driven distance) and then written onto a bus' schedule, is used. The user constraints are (i) a maximum allowed detour $\delta_{\max} = 2.0$ and (ii) a maximum allowed waiting time of $2\,t_0$. Rejections of requests are possible if any user's constraints would otherwise need to be violated. Unless stated otherwise, the map resembles the square map from the previous chapter. Further details can be found in the methods section.

We focus on scenarios in which the number of buses N and the demand x are well-balanced. This balance is quantified by a prescribed percentage p_{accept} of incoming requests that are served. For a fixed map, dispatcher and number of buses, the function $p_{\text{accept}}(x)$ is monotonically decreasing. The critical demand $x(N)$ at which e.g. 80 % of users can be served is thus the unique root of the equation $0.8 = p_{\text{accept}}(x|N)$. The numerical root-finding process is described in the SI. Its bottleneck is evaluating the right-hand side, for which a sufficiently long simulation must be run. In total, around 5000 simulations were run that resulted in over 1000 tuples $(N, x, p_{\text{accept}})$.

The simulation code, as well as animations of the simulated dynamics are made available in Mühle (2022b).

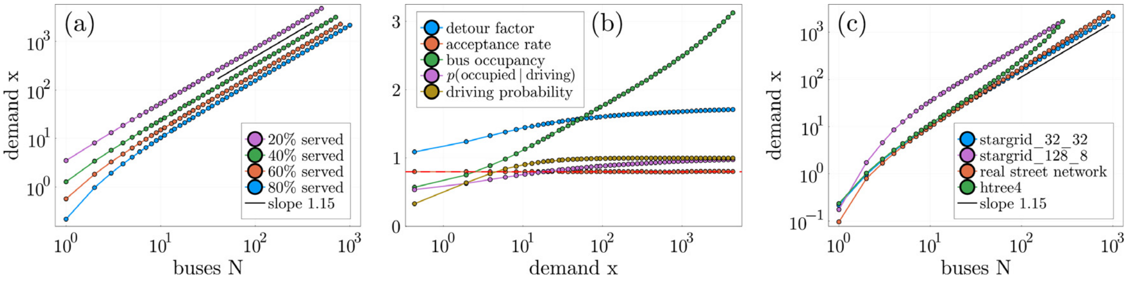

Fig. 4. Simulation results. All data points use a dispatcher which minimizes the total length of bus trajectories. **(a)** Solutions to the minimum fleet problem on the map *stargrid_32_32*. Shown are demands at which a given fleet size can serve 20, 40, 60 and 80% of all incoming requests respectively. After an initial regime that ends at around 10 buses, all curves are well approximated by a power-law with exponent ≈ 1.15. **(b)** Various simulation observables evaluated along the blue curve from Fig. (a). Shown here in blue is the system detour δ^{system} which has the lower bound 1.0 and upper bound $\delta_{\max} = 2.0$. As a consistency check, the served percentage of requests p_{accept} (orange) indeed coincides with the target value 0.8 (red horizontal line). The occupancy of driving buses $\langle b \rangle_{\text{driving}}$ (green) grows unbounded and is ultimately responsible for the power-law behavior in Fig. (a). The conditional probability $p(\text{occupied}|\text{driving})$ that a driving bus is occupied is shown in purple and approaches 1. Lastly, the fraction of time p_{driving} that buses are driving, not idling, is shown in yellow and also approaches 1, but faster than $p(\text{occupied}|\text{driving})$ does. **(c)** Analogous curves to those in Fig. (a) for $p_{\text{accept}} = 0.8$, but on different maps. The two *stargrids* have aspect ratios 1 (blue) and 16 (purple), but are identical otherwise and show power-laws with the same exponent. A higher aspect ratio of the map leads to a higher prefactor of the power-law. The green curve belongs to simulations on a real street network (Göttingen, orange) having an improved power-law exponent of 1.24, and the green curve refers to simulations on the map *htree4* – a strongly hierarchical structure on which the other maps' power-law is outperformed. The blue curves in figures (a) and (c) are identical. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.1. Comparing targets

For a fixed fleet size N, we determined the corresponding demand x at which $p_{\text{accept}} \in \{0.2, 0.4, 0.6, 0.8\}$ of all requests can be served. This was repeated for varying N, and the corresponding four curves $x(N)$ are shown in Fig. 4a. Since more buses can serve more users, all curves are monotonically growing while, naturally, the curves never intersect. Less obviously, they are well-approximated by a power-law with exponent 1.15[6], which is why a line of that slope was displayed in Fig. 3c for comparing model and simulations. The fact that this exponent is greater than 1 means that the ratio between served demand and buses increases. This indicates that η increases with demand and that RP inevitably outperforms non-shared transport modes, for which $x \propto N$. For small fleet sizes, the curves with a higher target percentage lie below the power-law. This can be attributed to idling buses (see Eq. (8)), which will be elaborated on in what follows.

6.2. Observables in supply-demand balance

Next, five observables are measured along the $x(N)$ curve such that both x and N grow simultaneously while keeping $p_{\text{accept}} = 0.8$ fixed. In Fig. 4b they are plotted against their corresponding value of x.

1. By construction, the served percentage p_{accept} is very close to 80 %, which is a simple benchmark for our method.
2. The fraction of time that buses are driving (not idling), p_{driving}, is close to 0.35 for $x \approx 0.5$ (1 bus), then quickly grows to exceed 0.9 for $x \approx 10$ (6 buses) and remains very close to 1.0 for $x > 20$ (10 buses). This can be understood as follows. At the time of submission, for large fleet sizes, one of the 20 % of rejected users will typically have several buses close to their pickup location, but still get rejected. If a considerable percentage of these buses would be idling, this would be a contradiction as any nearby idling bus could serve said request. Therefore, for large x and large N, buses are unlikely to idle when users get rejected. For small fleet sizes, this is not the case: A request may get rejected even though most buses are idling – they are simply too far away from the pickup location.
3. The detour factor δ^{system} has a lower bound of 1.0 by definition, and an upper bound of $\delta_{\max} = 2.0$ due to user constraints. It grows monotonically, which is not trivial as x alone can be expected to increase detours and N to reduce them. This confirms the intuition that our trajectory model captures in Eq. (21), with the difference that here the detour factor seemingly converges towards a value around 1.7 instead of its upper bound.
4. The bus occupancy $\langle b \rangle_{\text{driving}}$ grows monotonically and is not bound. Its growth appears logarithmic over a wide range of demands, as was also found in Zwick et al. (2021b), and predicted by our model in Eq. (25). Because δ^{system} is bound, this functional form is ultimately responsible for that of η and thus that of the minimum fleet size. The intersection of $\langle b \rangle_{\text{driving}}$ and δ^{system} marks an interesting critical demand (at around $x \approx 50$), at which the system's efficiency exceeds that of private cars.
5. Monitoring the conditional probability that a driving bus is occupied, $p(\text{occupied}|\text{driving})$, serves the purpose of testing the assumption our model framework makes when relying on the single-user view; namely that there is at least one user on board

[6] Power-law exponents stem from least-squares fits of a straight line to the last 15 shown data points.

when a bus is driving. Like p_{driving}, it converges towards 1, which is consistent with the fact that the overall bus occupancy b grows.

6.3. Comparing map topologies

Next we continue to focus on the $p_{\text{accept}} = 0.8$ case in which the majority of requests are served, but consider different maps, which are explained in the methods section and displayed in the SI. In brief, *stargrid_n_m* is a rectangle with side lengths n and m, the real street network is a rural region near Göttingen, Germany, and *htree4* is a strongly hierarchical tree-like network. It should be noted that each map has its own value for the time scale t_0 which was used to define x according to Eq. (5), yet each curve individually can be compared with that for private cars, for which $x(N) = N$ regardless of t_0. The real street network has non-uniform bus velocities, as explained in the methods. Figure 4c shows the simulated $x(N)$ curves.

The emergent power-law behavior for sufficiently high demand is universal across the three tested *stargrid* maps (having aspect ratios 1, 16, and 4, the latter only being shown in the SI). We link this universality to the geometric arguments (shrinking ellipse areas around user routes) presented in the previous chapter, which apply also to routes within narrow Euclidean rectangles. Fleets on more rectangular stargrids can serve a higher demand, thus increasingly outperforming non-shared transport modes. We attribute this to the geometric fact that shortest routes, and thus the directions of requested trips, naturally align on a rectangle with high aspect ratio.

Interestingly, the $x(N)$ curve for Göttingen also follows a power-law, but with an improved exponent of 1.24, which we associate with the presence of hierarchical structures in form of connected villages and the presence of a fast highway. Lastly, the curve for the map *htree4* shares the initial concave feature for low demand, but clearly outperforms all power-laws. This can be qualitatively explained with the conjecture made at the end of Chapter 5: With increasing fleet size, the union of all currently scheduled bus-routes covers most of the map such that there is a non-zero probability that incoming requests can be accepted without incurring any additional detour.

The performances observed on different topologies may also be qualitatively linked to the concept of *route volumes* introduced in Manik and Molkenthin (2020).

6.4. Comparing dispatchers

The simulation results discussed above stem from a single used dispatch algorithm, i.e. implemented user constraints and cost function. We have also tested different dispatch strategies without finding them to have a major impact (thus these additional simulation results are only shown in the SI), indicating that the geometric scaling arguments, which were made in the previous chapter and are agnostic of dispatch strategies, capture the simulation results quite well.

7. Conclusion

7.1. Interpretation

Combining all our results, we conjecture three major regimes of RP efficiency and fleet size scaling. Firstly, for low demand, buses are idling part of the time despite having to reject requests, which results in the initial concave regime in Fig. 4c. Secondly, once buses are driving at virtually all times, the power-law scaling takes over. This regime only exists as long as the probability of a new request's direct route to lie *exactly* on a bus' current route is negligible, because only then the ellipse-based geometry of stop-insertions is valid. Thirdly comes the regime where the bus routes can no longer be considered to cover a negligible fraction of the overall graph. Here, the ellipse-based picture of stop insertions overestimates detours and underestimates insertion probabilities, and hence the power-law associated with it is outperformed.

As real world applications will be unlikely to operate on regions small enough to enter the third regime, or to be profitable in the first one, the power-law regime may be expected to be the one with the most practical significance. The exponent of 1.15 found on the used maps with Euclidean topology can then be considered as a worst-case value due to a complete lack of hierarchical structures and thus incurred spatial correlations between user routes. Indeed, the exponent on real street networks was found to be greater.

Lastly, it should not be forgotten that RP efficiency is fundamentally limited by the bus capacity. Fully occupied buses may be considered as a fourth regime in which $x \propto N$, but its limitations on efficiency would likely be circumvented by adapting the choice of vehicle.

7.2. Summary

We presented a purely analytical model for RP efficiency, which is built upon a representative user's view point. Its central paradigm is the number of other users one may expect to meet on the bus, and how this affects the trade-off between bus occupancy and incurred detours. The framework is modular by design and its individual components (conditional probabilities) can be improved and adapted for specific applications, with hybrid approaches of feeding simulation or real-world data into the model also being feasible. As a proof of principle, the model was applied to RP on a Euclidean square map by making use of generic geometric arguments. To validate the model results and test its assumptions, simulations were run for a variety of fleet sizes, demands and street networks, which yielded further insights into model assumptions and relevant observables. The model's predictions meet the simulated results well when demand is sufficiently high such that buses can be considered to be driving, not idling. To report an order of magnitude for when this assumption is met, demand should roughly exceed 10 user requests being submitted in an area of an average trip length squared during the time one request takes to be served. Further simulations on street networks with hierarchical structures showed an improved performance compared to the Euclidean case.

7.3. Outlook

Finally, we suggest three directions for future work.

Firstly, the model framework should be supplemented with models for the probabilities $p(\text{occupied}|\text{driving})$ and p_{driving}. For low demands, $p(\text{occupied}|\text{driving})$ would improve the validity of the model's efficiency based on the single-user view, and an expression for p_{driving} would allow for a more accurate solution to the minimum fleet problem. Finite stop times for picking up and dropping off users have been neglected throughout this paper, but could be incorporated into p_{driving}.

Secondly, explicit expressions for the trajectory length δ_k and insertion probability R_k should be sought for maps other than areas on the Euclidean plane, linking the efficiency's scaling behavior to the map's topology directly. This would allow us to quantitatively predict and explain the simulated $x(N)$ curves on a variety of street networks without running simulations, and thereby identify regions well-suited for RP.

Thirdly, systematic simulations on simple rectangular maps should be run with state-of-the-art optimization techniques in order to investigate whether or not our model's predicted scaling behavior can be improved upon by means of the dispatch algorithm alone.

Materials and methods

Our simulation framework is written in *Julia* (Bezanson et al., 2017). It is made available in Mühle (2022b), where animations (described in the SI) of the simulated dynamics and all data displayed as simulation results can also be found. Code for evaluating the model framework is made available in Mühle (2022a).

Maps A map is a weighted, directed graph, populated with N buses. The weight of an edge denotes the time required to travel from the one node to another and buses are only allowed to u-turn at nodes (having in mind one-way lanes), i.e. the process of travelling through an edge must be completed. For letting buses travel across the map, existing code from the Julia package *Agents.jl* (Datseris et al., 2022) is made use of. Simulations were run on three types of maps. Images of them can be found in the SI.

- *stargrid_X_Y* is an X by Y grid with diagonal connections such that there are $X \cdot Y$ nodes total and the map is a rectangle with aspect ratio Y/X. This map is meant to resemble a finite region with Euclidean distances being a decent proxy for travel times between nodes.
- *OpenStreetMaps*: One real street networks is used. It is a rural region west of Göttingen, Germany, which consists of 6 connected villages and one fast highway. Bus velocities are chosen according to the respective street class' speed limit according to German law.
- *htree_4* consists of the first four layers of a fractal tree structure. From the ends of a central line, perpendicular lines branch out resembling the letter an H (thus the name). This process is repeated 3 more times at all ends of the previous iteration's perpendicular lines. The resulting graph has a strong hierarchical structure which mimics that of a rural environment.

Unless stated otherwise, the map *stargrid_32_32* is used as a default.

Requests Requests have a spatial and temporal component. They come with constraints regarding service quality which are discussed below as rejection criteria. Temporally, each request has a submission time. The arrival of requests follows a Poisson process such that the time intervals between two subsequent request submissions are exponentially distributed. The Poisson process' mean value is the inverse frequency f^{-1}. Spatially, requests consist of two points on the map – the pickup and the dropoff location. Both are drawn uniformly and independently from the set of all points on all edges of the map.

This request pattern can immediately be made use of to obtain the (map-dependent) average requested time $t_0 = \langle t_u^{\text{direct}} \rangle_u$ and define the demand x as in Eq. (5).

To-do list insertions Each bus has an ordered todo-list which consists of jobs of picking up or dropping off one user. Without new incoming requests, each bus works through its to-do list and then becomes idle. Once written onto a todo-list, the relative order in which jobs are carried out is fixed. Both jobs must be carried out by the same bus with the pickup occurring before the dropoff, but jobs associated with other users can lie between them. Upon arrival of a new request, the dispatcher writes the pickup and the dropoff jobs concerning this user onto one bus' to-do-list – a process called an *insertion*.

Rejection criteria For each served user, two constraints are simultaneously implemented:

- *Maximum waiting time*: The time between request submission and being picked up by a bus, t_w, has an upper bound $t_w^{\max} = 2\, t_0$ (the time required for one taxi to transport one user, then drive to the next pickup).
- *Maximum detour*: The detour δ_u for each individual user must be below $\delta_{\max} = 2.0$.

A particular insertion is feasible if and only if it violates no constraints. When determining an insertion's feasibility, all rejection criteria are taken into account for all users in the system. If all available insertions are not feasible, the incoming request is rejected. Hence the rejection criteria encode a minimal service quality that is always guaranteed for all served users. This follows the logic that users are willing to endure a certain inconvenience and would otherwise not opt for RP.

Cost functions From the set of all feasible insertions, the dispatcher chooses the insertion which is yields the smallest-possible increase in a cost function. In the main text, this cost function's value equals the sum of all bus' scheduled trajectory length. In the SI two other cost functions are considered: One which sums over all users' dropoff times, and one which returns random values.

Declaration of Competing Interest

No competing interests are declared.

Acknowledgement

I thank Stephan Herminghaus, Prakhar Godara and Matthias Häring for valuable discussions and comments on the structure of this manuscript.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.multra.2023.100080.

References

Agatz, N., Erera, A., Savelsbergh, M., Wang, X., 2012. Optimization for dynamic ride-sharing: a review. Eur. J. Oper. Res. 223 (2), 295–303.

Agatz, N., Erera, A.L., Savelsbergh, M.W., Wang, X., 2011. Dynamic ride-sharing: a simulation study in metro atlanta. Procedia-Social Behav. Sci. 17, 532–550.

Agency, E. E., 2020. Air quality in Europe - 2020 report. https://www.eea.europa.eu/publications/air-quality-in-europe-2020-report, Accessed: 2022-05-18.

Agency, I. E., 2021. Global energy-related CO2 emissions by sector. https://www.iea.org/data-and-statistics/charts/global-energy-related-co2-emissions-by-sector, Accessed: 2022-05-18.

Aldous, D., Barthelemy, M., 2019. Optimal geometry of transportation networks. Phys. Rev. E 99 (5), 052303.

Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E., Rus, D., 2017. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. Proc. Natl. Acad. Sci. 114 (3), 462–467.

Arnott, R., Small, K., 1994. The economics of traffic congestion. Am. Sci. 82 (5), 446–455.

Bates, J., Leibling, D., 2012. Spaced out: perspectives on parking policy. In: The Royal Automobile Club Foundation for Motoring, pp. 9–11.

Berbeglia, G., Cordeau, J.-F., Laporte, G., 2010. Dynamic pickup and delivery problems. Eur. J. Oper. Res. 202 (1), 8–15.

Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B., 2017. Julia: a fresh approach to numerical computing. SIAM Rev. 59 (1), 65–98. doi:10.1137/141000671.

Bilali, A., Engelhardt, R., Dandl, F., Fastenrath, U., Bogenberger, K., 2020. Analytical and agent-based model to evaluate ride-pooling impact factors. Transp. Res. Record 2674 (6), 1–12.

Bischoff, J., Maciejewski, M., Nagel, K., 2017. City-wide shared taxis: a simulation study in Berlin. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 275–280.

Boesch, P.M., Ciari, F., Axhausen, K.W., 2016. Autonomous vehicle fleet sizes required to serve different levels of demand. Transp. Res. Record 2542 (1), 111–119.

Bureau of Transportation Statistics, Household, individual, and vehicle characteristics, 2011. https://www.bts.gov/archive/publications/highlights_of_the_2001_national_household_travel_survey/section_01, Accessed: 2022-05-18.

Daganzo, C.F., Ouyang, Y., Yang, H., 2020. Analysis of ride-sharing with service time and detour guarantees. Transp. Res. Part B Methodol. 140, 130–150.

Datseris, G., Vahdati, A.R., DuBois, T.C., 2022. Agents.jl: A Performant and Feature-Full Agent-Based Modeling Software of Minimal Code Complexity. Sage Publications Sage UK, London, England. 00375497211068820

de Ruijter, A., Cats, O., Alonso-Mora, J., Hoogendoorn, S., 2020. Ride-sharing efficiency and level of service under alternative demand, behavioral and pricing settings. Transportation Research Board 2020 Annual Meeting.

Edmonds, E., 2019. Your driving costs: spike in finance costs drives increase. https://newsroom.aaa.com/2019/09/your-driving-costs-spike-in-finance-costs-drives-increase/, Accessed: 2022-05-18.

Engelhardt, R., Dandl, F., Bogenberger, K., 2020. Speed-up heuristic for an on-demand ride-pooling algorithm. arXiv preprint arXiv:2007.14877.

European Environment Agency, 2020Are we moving in the right direction? Indicators on transport and environmental integration in the EU. https://www.eea.europa.eu/ds_resolveuid/0c1c4a6acf289ffdefa1876ea5d60f07, Accessed: 2022-07-14.

Fagnant, D.J., Kockelman, K.M., 2018. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in austin, texas. Transportation 45 (1), 143–158.

Fielbaum, A., Bai, X., Alonso-Mora, J., 2021. On-demand ridesharing with optimized pick-up and drop-off walking locations. Transp. Res. Part C Emerg.Technol. 126, 103061.

Folco, P., Gauvin, L., Tizzoni, M., Szell, M., 2022. Data-driven bicycle network planning for demand and safety. arXiv preprint arXiv:2203.14619.

Foljanty, L., 2020a. Mapping the global on-demand ridepooling market. https://lukas-foljanty.medium.com/mapping-the-global-on-demand-ridepooling-market-f8318de1c030, Accessed: 2022-05-18.

Foljanty, L., 2020b. On-demand ridepooling market: 2020 recap. https://lukas-foljanty.medium.com/on-demand-ridepooling-market-2020-recap-71a229f2e7b9, Accessed: 2022-05-18.

Foljanty, L., 2021. On-demand ridepooling market size. https://lukas-foljanty.medium.com/on-demand-ridepooling-market-size-f3ff93845c5c, Accessed: 2022-05-18.

Furuhata, M., Dessouky, M., Ordóñez, F., Brunet, M.-E., Wang, X., Koenig, S., 2013. Ridesharing: the state-of-the-art and future directions. Transp. Res. Part B Methodol. 57, 28–46.

Henao, A., Marshall, W.E., 2019. The impact of ride-hailing on vehicle miles traveled. Transportation 46 (6), 2173–2194.

Hensher, D.A., 2008. Climate change, enhanced greenhouse gas emissions and passenger transport–what can we do to make a difference? Transp. Res. Part D Transp. Environ. 13 (2), 95–111.

Herminghaus, S., 2019. Mean field theory of demand responsive ride pooling systems. Transp. Res. Part A PolicyPract. 119, 15–28.

Hörl, S., Ruch, C., Becker, F., Frazzoli, E., Axhausen, K.W., 2018. Fleet control algorithms for automated mobility: a simulation assessment for zurich. In: 2018 TRB Annual Meeting Online. Transportation Research Board, pp. 18–02171.

Horn, M.E., 2002. Fleet scheduling and dispatching for demand-responsive passenger services. Transp. Res. Part C Emerg.Technol. 10 (1), 35–63.

Jung, J., Jayakrishnan, R., Park, J.Y., 2016. Dynamic shared-taxi dispatch algorithm with hybrid-simulated annealing. Comput.-Aided Civ. Infrastruct. Eng. 31 (4), 275–291.

Kaddoura, I., Schlenther, T., 2021. The impact of trip density on the fleet size and pooling rate of ride-hailing services: a simulation study. Procedia Comput. Sci. 184, 674–679.

Ke, J., Yang, H., Li, X., Wang, H., Ye, J., 2020. Pricing and equilibrium in on-demand ride-pooling markets. Transp. Res. Part B Methodol. 139, 411–431.

Ke, J., Yang, H., Zheng, Z., 2020. On ride-pooling and traffic congestion. Transp. Res. Part B Methodol. 142, 213–231.

Kostorz, N., Fraedrich, E., Kagerbauer, M., 2021. Usage and user characteristics–insights from MOIA, Europe's largest ridepooling service. Sustainability 13 (2), 958.

Liebchen, C., Lehnert, M., Mehlert, C., Schiefelbusch, M., 2020. Ridepooling-effizienz messbar machen. Der Nahverkehr 9 (2020), 18–21.

Lobel, I., Martin, S., 2020. Detours in Shared Rides. Available at SSRN 3711072.

Lokhandwala, M., Cai, H., 2018. Dynamic ride sharing using traditional taxis and shared autonomous taxis: a case study of nyc. Transp. Res. Part C Emerg.Technol. 97, 45–60.

Lotze, C., Marszal, P., Schröder, M., Timme, M., 2022. Dynamic stop pooling for flexible and sustainable ride sharing. N. J. Phys. 24 (2), 023034.

Lu, C., Maciejewski, M., Nagel, K., 2021. Effective operation of demand-responsive transport (DRT): implementation and evaluation of various rebalancing strategies. In: Proceedings of the 27th ITS World Congress. Hamburg

Ma, S., Zheng, Y., Wolfson, O., 2013. T-share: a large-scale dynamic taxi ridesharing service. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE). IEEE, pp. 410–421.

Manik, D., Molkenthin, N., 2020. Topology dependence of on-demand ride-sharing. Appl. Netw. Sci. 5 (1), 1–16.

Manville, M., Shoup, D., 2005. Parking, people, and cities. J. Urban Plann. Dev. 131 (4), 233–245.

Markov, I., Guglielmetti, R., Laumanns, M., Fernández-Antolín, A., de Souza, R., 2021. Simulation-based design and analysis of on-demand mobility services. Transp. Res. Part A PolicyPract. 149, 170–205.

Merlin, L.A., 2019. Transportation sustainability follows from more people in fewer vehicles, not necessarily automation. J. Am. Plann. Assoc. 85 (4), 501–510.

Molkenthin, N., Schröder, M., Timme, M., 2020. Scaling laws of collective ride-sharing dynamics. Phys. Rev. Lett. 125, 248302.

Moreno, A.T., Michalski, A., Llorca, C., Moeckel, R., 2018. Shared autonomous vehicles effect on vehicle-km traveled and average trip duration. J. Adv. Transp. 2018, 1–10.

Mühle, S., 2022a. The code for evaluating the analytical model. https://github.com/SteffenMuehle/RidePoolingAnalytics.

Mühle, S., 2022b. The code of our custom simulation framework, animations, and simulated data sets. https://github.com/SteffenMuehle/RidePoolingSimulations.

Pernestål, A., Kristoffersson, I., 2019. Effects of driverless vehicles: comparing simulations to get a broader picture. Eur. J. Transp. Infrastruct.Res. 19 (1), 1–23.

Psaraftis, H.N., Wen, M., Kontovas, C.A., 2016. Dynamic vehicle routing problems: three decades and counting. Networks 67 (1), 3–31.

Qian, X., Zhang, W., Ukkusuri, S.V., Yang, C., 2017. Optimal assignment and incentive design in the taxi group ride problem. Transp. Res. Part B Methodol. 103, 208–226.

Ruch, C., Lu, C., Sieber, L., Frazzoli, E., 2020. Quantifying the efficiency of ride sharing. IEEE Trans. Intell. Transp. Syst. 22 (9), 5811–5816.

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S.H., Ratti, C., 2014. Quantifying the benefits of vehicle pooling with shareability networks. Proc. Natl. Acad. Sci. 111 (37), 13290–13294.

Schneider, T., 2022. Taxi and Ridehailing Usage in New York city. https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/, Accessed: 2022-05-18.

Stiglic, M., Agatz, N., Savelsbergh, M., Gradisar, M., 2015. The benefits of meeting points in ride-sharing systems. Transp. Res. Part B Methodol. 82, 36–53.

Storch, D.-M., Timme, M., Schröder, M., 2021. Incentive-driven transition to high ride-sharing adoption. Nat. Commun. 12 (1), 1–10.

Szell, M., Mimar, S., Perlman, T., Ghoshal, G., Sinatra, R., 2022. Growing urban bicycle networks. Sci. Rep. 12 (1), 1–14.

Tachet, R., Sagarra, O., Santi, P., Resta, G., Szell, M., Strogatz, S.H., Ratti, C., 2017. Scaling law of urban ride sharing. Sci. Rep. 7 (1), 1–6.

Vazifeh, M.M., Santi, P., Resta, G., Strogatz, S.H., Ratti, C., 2018. Addressing the minimum fleet problem in on-demand urban mobility. Nature 557, 534–538.

Wang, X., Yang, H., Zhu, D., 2018. Driver-rider cost-sharing strategies and equilibria in a ridesharing program. Transp. Sci. 52 (4), 868–881.

Wen, J., Zhao, J., Jaillet, P., 2017. Rebalancing shared mobility-on-demand systems: a reinforcement learning approach. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). Ieee, pp. 220–225.

Wolf, H., Storch, D.-M., Timme, M., Schröder, M., 2022. Spontaneous symmetry breaking in ride-sharing adoption dynamics. Phys. Rev. E 105 (4), 044309.

Zech, R.M., Molkenthin, N., Timme, M., Schröder, M., 2022. Collective dynamics of capacity-constrained ride-pooling fleets. Sci. Rep. 12 (1), 1–9.

Zhu, P., Mo, H., 2022. The potential of ride-pooling in VKT reduction and its environmental implications. Transp. Res. Part D Transp. Environ. 103, 103155.

Zwick, F., Axhausen, K.W., 2020. Analysis of ridepooling strategies with MATSim. In: 20th Swiss Transport Research Conference (STRC 2020) (Virtual). IVT, ETH Zurich.

Zwick, F., Axhausen, K.W., 2022. Ride-pooling demand prediction: a spatiotemporal assessment in germany. J. Transp. Geogr. 100, 103307.

Zwick, F., Kuehnel, N., Axhausen, K.W., 2022. Review on theoretical assessments and practical implementations of ride-pooling. In: 22nd Swiss Transport Research Conference (STRC 2022). STRC.

Zwick, F., Kuehnel, N., Moeckel, R., Axhausen, K.W., 2021. Agent-based simulation of city-wide autonomous ride-pooling and the impact on traffic noise. Transp. Res. Part D Transp.Environ. 90, 102673.

Zwick, F., Kuehnel, N., Moeckel, R., Axhausen, K.W., 2021. Ride-pooling efficiency in large, medium-sized and small towns-simulation assessment in the munich metropolitan region. Procedia Comput. Sci. 184, 662–667.