Submitted to *Manufacturing & Service Operations Management*
manuscript (Please, provide the manuscript number!)

Demand Equilibria in Spatial Service Systems

John G. Carlsson
Industrial and Systems Engineering, University of Southern California, jcarlsso@usc.edu, https://viterbi-web.usc.edu/ jcarlsso/

Xiaoshan Peng
Operations & Decision Technologies, Kelley School of Business, Indiana University, xp1@iu.edu, http://www.xiaoshanpeng.com/

Ilya O. Ryzhov
Decision, Operations & Information Technologies, Robert H. Smith School of Business, University of Maryland, iryzhov@umd.edu, https://sites.google.com/umd.edu/iryzhov/

Problem definition: A service is offered at certain locations ("facilities") in a geographical region. Customers can appear anywhere in the region, and each customer chooses a facility based on travel distance as well as expected waiting time. Customer decisions affect waiting times by increasing the load on a facility, and thus impact other customers' decisions. The service provider can also influence service quality by adjusting service rates at each facility. **Methodology/results**: Using a combination of queueing models and computational geometry, we characterize demand equilibria in such spatial service systems. An equilibrium can be visualized as a partition of the region into service zones that form as a result of customer decisions. Service rates can be set in a way that achieves the best possible social welfare purely through decentralized customer behavior. **Managerial implications**: We provide techniques for computing and visualizing demand equilibria, as well as calculating optimal service rates. Our analytical and numerical results indicate that, in many situations, resource allocation is a far more significant source of inefficiency than decentralized behavior.

Key words: spatial service systems; geographical partitioning; Voronoi diagrams; equilibrium analysis
History:

1. Introduction

Consider a service system in which customers are served by facilities at different locations within a geographical region. For example, the facilities could represent mobile medical stations or locations of a department of motor vehicles. Customers, located anywhere in the geographical region, can choose one of the facilities where they wish to be served. They will not necessarily choose the closest facility: the decision is based, not only on the travel

distance, but also on the expected waiting time at the facility, which depends on how many other customers have chosen to receive service there. Real-time information about the waiting times is not available to the customer, but they can infer the average.

Thus, customer decisions influence one another through the waiting time. A facility located in the middle of a densely populated area may appear to be convenient, but the low travel time may attract a large number of customers, driving up the waiting time and making the facility less appealing. Some customers may then prefer to travel farther in order to take advantage of shorter lines. These tradeoffs can be studied using a notion of demand equilibrium, which consists of customer choices (for every possible location in the region) that are individually optimal for each customer and, in aggregate, lead to a stable set of loads on the facilities.

We study the equilibrium using a queueing model in which customers arrive according to a spatio-temporal Poisson process (not only at random times, but at random locations), and each facility is represented by an M/M/1 queue. As long as there is enough capacity in the system to handle aggregate demand, a unique equilibrium exists and gives rise to a geometric partition of the region, with each facility serving all customers whose arrival locations are in a particular "zone" whose shape and size are jointly determined by all customer choices. The partition belongs to a class of spatial structures called "additively weighted Voronoi diagrams" (Devulapalli et al. 2015), but satisfies an additional equilibrium condition not previously studied in that literature. Thus, our paper brings together computational geometry and queueing in a novel way. We present comparative statics of the equilibrium partition, and show how it can be computed and visualized.

We then study the social welfare of the equilibrium, measured in terms of the expected total cost (travel plus waiting) per customer. It is well-known (Ghosh and Hassin 2021) that decentralized behavior in queueing systems generally does not lead to socially optimal outcomes. This "price of anarchy" can also arise in our problem: if the service provider had the ability to design the service zones and impose them on customers in a centralized manner (i.e., assign each customer to a facility), instead of allowing the zones to form as a result of customer decisions, then, in general, the social welfare could be improved. However, we find that, for certain values of the service rates, the decentralized equilibrium partition coincides with the centralized one – and, what is more striking, these service rates are *optimal* for the social welfare.

More specifically, suppose that the service provider can control mean service times at the facilities through allocation of resources such as money or staff. Our single-server model enables a simple and natural formulation of this decision: the service rate itself can be viewed as the "resource," with the total service capacity across all facilities satisfying a budget constraint. This interpretation of service rates as a resource to be allocated has previously been used to study problems in healthcare systems (Chao et al. 2003), service systems (Shanthikumar and Xu 1997), and business process management (Dieker et al. 2017). In our context, by changing the service rates, one also changes the geometry of the centralized partition, as well as the social welfare that it achieves. One can then set service rates to optimize social welfare. But, under those same optimal service rates, the centralized partition is also identical to the decentralized partition formed through individual choice. In other words, it is possible to recover the *absolute best possible* social cost, purely through decentralized behavior, simply by setting the service rates correctly.

We further show that this result continues to hold when customer choice is subject to random shocks, which allows for differences in perception between customers of the waiting times or of the inconvenience of waiting. In that setting, the equilibrium can no longer be visualized as a geographical partition, because two customers arriving at the same location can now choose different facilities, but it still exists as a set of choice probabilities that are individually optimal and induce stable loads on each facility. The optimal service rates are different from those computed under deterministic choice, but they still recover the best possible equilibrium under decentralized customer behavior.

The managerial value of our work is threefold. First, we provide a clean geometric interpretation of the decentralized equilibrium, where each facility is associated with a service zone. Second, we provide techniques for computing these zones, allowing for easy visualization and comparison of various "what-if" scenarios. (These same techniques can also be used to compute optimal service rates.) Third, our analysis shows that social welfare optimization is entirely a matter of resource allocation. As long as the service rates are set optimally, there is no loss in efficiency resulting from decentralized behavior, and socially optimal outcomes can be achieved without the need for any other manipulation of the system. Our numerical results also indicate that, even under suboptimal allocations, resource allocation is a far more significant source of inefficiency than decentralized behavior when the aggregate demand is high.

2. Literature Review

In the following, we discuss the connections between our work and the literature on queueing and computational geometry, respectively.

Queueing. There are numerous papers on equilibrium analysis of strategic customer behavior in queueing systems; see Hassin and Haviv (2003) and Hassin (2016) for a comprehensive review. More specifically, there is a significant literature on strategic joining decisions, where customers decide between queues in a way that balances expected waiting time against some other consideration, such as reward from the service. (Empirical evidence indicates that customers do consider congestion levels when making such decisions; see Dong et al. 2019.) For example, in Armony and Maglaras (2004) the choice is between two service modes representing real-time and call-back options at a call center; in Pender et al. (2020) it is between two identical queues whose states are observable; in Zhou and Ryzhov (2021) it is between a slow, but free queue and a fast, but expensive one.

Most of these studies do not incorporate spatial structure. Travel cost (distance or time) is included in some models, but in a fundamentally one-dimensional manner: Rajan et al. (2019), Hassin and Roet-Green (2020) and Wang et al. (2023) all model such costs as i.i.d. random variables, which suffices for their purposes because they all assume that service is delivered by a single M/M/1 queue. Thus, these studies capture the effect of travel cost on congestion, but not the tradeoffs that arise when comparing multiple queues at different locations. In such situations, it becomes necessary to model the spatial position of the arrival rather than only the distance to a queue.

Very few studies have done so. Heinhold (1978), Lee and Cohen (1985) and Grossman and Brandeau (2002) consider customer choice between facilities, but restrict the customers to a finite, pre-specified set of locations, in effect reducing the problem to matching between supply and demand nodes, as in the classic transportation network model. (Along those lines, there are also some network optimization models with queueing elements, for example the work by Kullman et al. 2021 on electric vehicle routing, but they have a more algorithmic focus and do not study demand equilibria.) Alptekinoglu and Corbett (2010) and Xu et al. (2016) use a Hotelling-type location model where customers can appear anywhere, but only on a one-dimensional interval. A recent work by Ding et al. (2022) considers a general abstract setting with uncountably many customer types (thus, potentially, each possible arrival location could be a "type"). This work focuses on the theoretical

characterization of fluid and diffusion limits, whereas our paper studies a more explicitly spatial setting and focuses on geometric characterization and interpretation of customer choice.

There is an extensive literature on inefficiencies caused by strategic customer behavior, and various mechanisms for improving the social welfare, such as fee structures (Gavirneni and Kulkarni 2016), imposition of delays (Baron et al. 2022), specialized priority rules (Haviv and Oz 2018), and partial information structures (Economou 2021). One could perhaps think of our resource allocation problem (dividing service capacity between queues) as another such mechanism; in our setting, it turns out to be entirely sufficient for eliminating the inefficiency.

Computational geometry. Geographical partitioning problems have been extensively studied in computational geometry, with facility logistics being a well-known application area (Carlsson and Devulapalli 2013). The additively weighted Voronoi diagram (Aurenhammer 1991) is a class of partitions where each customer minimizes the sum of travel distance and a facility-specific "weight." This literature assumes that the weights (or, equivalently, the areas of the service zones) are fixed constants and primarily focuses on computation of the diagram; see, e.g., Aurenhammer and Klein (2000), Pavone et al. (2011), and Hartmann and Schuhmacher (2020). The novelty of our paper, relative to this literature, is that the weights in our model are endogenized: they represent expected waiting times at each queue, and thus are impacted by customer decisions.

In the operations literature, geographical partitioning has been viewed primarily as a central planning problem. For example, in Haugland et al. (2007), a logistics firm designs delivery districts for vehicles, while Ricca et al. (2008) proposes to use weighted Voronoi diagrams for political districting. In these and other applications, the central planner can impose any desired partition on the region. This is not the case in our paper: the partition is determined in a decentralized manner through customer choice, and the service provider can at most influence it indirectly through the service rates. Thus, while Voronoi diagrams have been used to achieve various fairness and social welfare objectives (Aronov et al. 2009, Yushimito et al. 2012), we appear to be the first to study decentralized partitioning and whether it can achieve social optimality.

For computing and visualizing demand equilibria, we develop a mathematical programming formulation that generalizes Carlsson et al. (2016) to a broader class of partitioning

problems. The formulation can be viewed as a two-stage model where the second (inner) stage is based on Carlsson et al. (2016), and the first (outer) stage is new to our paper. Our analysis presents some stand-alone interest for computational geometry as it relates the weights of the Voronoi diagram to the gradient of the cost function (e.g., the waiting time). It also accommodates a high level of generality: in particular, to find the socially optimal equilibrium, it is necessary to use a nonconvex cost.

3. Demand Equilibria and Their Properties

This section proposes a model to characterize the equilibrium of the spatial queueing problem and presents properties of the equilibrium. Section 3.1 describes our spatial queueing problem and defines the equilibrium. Section 3.2 proves existence and uniqueness of the equilibrium, and Section 3.3 analyzes its comparative statics.

3.1. Definitions

We consider a metric space $(\mathbb{R}^n, d)$ and a region $\mathcal{S} \subseteq \mathbb{R}^n$, which is closed and convex. Customers arrive according to an exogenous spatio-temporal Poisson process in $\mathcal{S}$. The arrival rate is spatially heterogeneous with a positive intensity function $\lambda(x)$ for $x \in \mathcal{S}$. In the simplest case, let $n = 2$, d be the Euclidean distance, and $\mathcal{S}$ be a geographical region; then, $\lambda(x)$ is the arrival rate of demand at a particular physical location x on the map, with the dependence on x reflecting the population density and demographics. However, in Section 5, we consider a more general setting where $\mathcal{S}$ includes geography as well as other attributes. We denote by $m(x) = \lambda(x)/\lambda$ the normalized arrival intensity in the region, where $\lambda = \iint_{\mathcal{S}} \lambda(x)\, dx$ is the aggregate arrival rate. For any measurable set $\mathcal{A} \subseteq \mathcal{S}$ on this region, we define $A = \iint_{\mathcal{A}} m(x)\, dx$ to be the fraction of total demand arising in the subregion. We refer to A as the "area" of subregion $\mathcal{A}$ because, when arrivals are uniformly distributed on the region, A reduces to the Lebesgue measure of $\mathcal{A}$. Thus, the number of customers arriving in the time interval $[0, t]$, inside a measurable set $\mathcal{A} \subseteq \mathcal{S}$, follows a Poisson distribution with mean $\lambda A t$.

The exogenous Poisson arrival assumption implicitly assumes that the demand arises at location x independently of other locations; roughly speaking, the arrival process in each infinitesimal subregion is an independent renewal process. Under the mild assumption that the arrival rates of these processes scale similarly, the Palm-Khintchine Theorem (ch. 5.9 of Karlin and Taylor 1975, ch. 5.8 of Heyman and Sobel 2003) guarantees that their

superposition, which is the aggregated arrival process of any $\mathcal{A} \subseteq \mathcal{S}$, is a Poisson process. The assumption of independent demand is a good fit for, e.g., public services or healthcare services, where it is reasonable to suppose that the need of a particular household for a new driver's license or a visit to the emergency room is independent of its neighbors. In these applications, every arrival needs the service and has to choose a facility to complete the service.

When an arrival occurs, the customer chooses from one of the K facilities with predetermined locations $x_1, ..., x_K \in \mathcal{S}$.The travel distance between a customer and facility k is $d_k(x) = d(x, x_k)$ for $x \in \mathcal{S}$. A common choice for the distance metric is $d(x, x_k) = \|x - x_k\|_2$, but any other continuous metric can also be used. The kth facility operates an M/M/1 queue with service rate μ_k. We assume that $\sum_k \mu_k > \lambda$, that is, the aggregate capacity of the system is sufficient to serve all of the demand. Thus, the arrival rate to each queue will be determined endogenously by the set of customers that prefer that facility to others.

A geographical partition of the region is a collection $\{\mathcal{A}_k\}_{k=1}^K$ of subsets of $\mathcal{S}$ such that $\bigcup_k \mathcal{A}_k = \mathcal{S}$ and $\mathcal{A}_j \cap \mathcal{A}_k$ has measure zero for any $j \neq k$. Let $A_k = \iint_{\mathcal{A}_k} m(x)\, dx$ denote the area of the kth subset. A demand equilibrium is represented by a specific partition that satisfies the condition

$$d_k(x) + f_k(A_k) \leq \min_{j \neq k} d_j(x) + f_j(A_j) \qquad \forall x \in \mathcal{A}_k,\ k = 1, ..., K, \tag{1}$$

where

$$f_k(A) = \begin{cases} \frac{c}{\mu_k - \lambda A} & A < \frac{\mu_k}{\lambda}, \\ \infty & \text{otherwise} \end{cases}$$

is the expected waiting time at facility k, given that the facility serves a proportion A of arrivals (i.e., the arrival rate to the kth queue is λA), scaled by a constant $c > 0$. Equation (1) ensures that a customer arriving at a point $x \in \mathcal{A}_k$ prefers facility k to the others. Thus, each customer minimizes the cost of traveling to a facility (represented by the distance) plus the expected waiting time at that facility. The constant c is used to weigh these two types of costs; one may think of c as the slope of a linear utility function used to convert between distance and time. The sets $\mathcal{A}_k$ are determined by customer choice according to (1), so the space $\mathcal{S}$ is divided between facilities in a decentralized manner.

There is an important connection between this model and the framework of additively weighted Voronoi diagrams (Devulapalli et al. 2015). Such a diagram partitions a plane into regions around the locations $x_1, ..., x_K$. The kth region consists of all points x satisfying

$$d_k(x) + w_k \leq \min_{j \neq k} d_j(x) + w_j, \tag{2}$$

where $w_1, ..., w_K$ are fixed weights. We abuse notation slightly by letting $\mathcal{A}_k(w)$ be the set of all x such that (2) holds for the given vector $w = (w_1, ..., w_K)$, with $A_k(w) = \iint_{\mathcal{A}_k(w)} m(x)\, dx$ being the corresponding area. In computational geometry, the weights w_k are pre-specified constants, which then determine the regions (alternately, one can fix the areas A_k first, which then determines the weights; see Hartmann and Schuhmacher 2020). In our setting, however, the region served by facility k is the set $\mathcal{A}_k(w^*)$, where w^* is a vector of weights that satisfy the *equilibrium condition*

$$w_k^* = f_k(A_k(w^*)). \tag{3}$$

In other words, (1) describes a particular additively weighted Voronoi diagram whose weights are chosen to satisfy (3).

We now state several properties of the area A_k that will be used (in Sections 3.2-3.3) to prove existence, uniqueness, and structural properties of the equilibrium. These properties will apply to any area function that satisfies the conditions in Lemma 1.

LEMMA 1. *For a weight vector w, the areas of the partition satisfy the following:*

1. *$A_k(w)$ is continuous in w;*
2. *For weights w and $\tilde{w}$ such that $w_k - \tilde{w}_k \geq w_j - \tilde{w}_j$ for all j, $A_k(w) \leq A_k(\tilde{w})$;*
3. *$\frac{\partial A_k}{\partial w_k} \leq 0$ and $\frac{\partial A_k}{\partial w_j} \geq 0$ for $j \neq k$;*
4. *$\sum_j \frac{\partial A_k}{\partial w_j} = 0$.*

Proof. The continuity of A_k in w follows from the properties of additively weighted Voronoi diagrams (see, e.g., Lemma 3.3 in Hartmann 2016). The second property follows from (2), which implies $\mathcal{A}_k(w) \subseteq \mathcal{A}_k(\tilde{w})$, whence $A_k(w) \leq A_k(\tilde{w})$. The last two properties also follow from (2). Increasing the weight of any one region reduces the value of that region, relative to all others, to a customer originating at any given x. On the other hand, adding the same constant to all the weights will not affect the areas. □

3.2. Existence and Uniqueness of the Equilibrium

The main result of this section is that (3) always has a unique solution. Existence of the equilibrium weights w^* is shown using Brouwer's fixed point theorem, but some care is required because the scaled waiting time functions f_k are not bounded and the area functions $A_k(w)$ in (3) do not have a closed form. Because the weights determine the partition in an additively weighted Voronoi diagram, uniqueness of w^* implies the uniqueness of the equilibrium partition (1).

We begin by noting several useful properties of the functions f_k. First, each f_k maps $\mathbb{R}_+$ into the extended positive numbers $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$. Moreover, each f_k is increasing with $f_k(0) > 0$ and $f_k(t) = \infty$ for $t \geq t_k$ and some known $t_k > 0$ that satisfy $\sum_k t_k > 1$. Each f_k is finite and continuous on $[0, t_k)$ with $\lim_{t \nearrow t_k} f_k(t) = \infty$.

LEMMA 2. *Let $\bar{w}$ be the optimal value of the problem*

$$\max_{\mathbf{a}} \min_k f_k(a_k) \tag{4}$$

subject to $\mathbf{a} = (a_1, \cdots, a_K) \geq 0$ and $\sum_k a_k = 1$. Then, $\bar{w} < \infty$.

Proof. Suppose that $\bar{w} = \infty$. The feasible region of problem (4) is closed, so there must be some feasible a for which $f_k(a_k) = \infty$ for all k. Therefore, by the structure of f_k, we have $a_k \geq t_k$ for all k, whence $1 = \sum_k a_k \geq \sum_k t_k$. This leads to a contradiction. □

Define the compact set $\mathcal{W} = \left\{w \in \mathbb{R}^K : 0 \leq w_k \leq \bar{w}^*, \quad k = 1, ..., K\right\}$, where $\bar{w}^* = \max\left\{1 + \bar{w} + \sup_{x \in \mathcal{S}, j, k} |d_j(x) - d_k(x)|, \max_k f_k(0)\right\}$. Also define the function $\phi : \mathcal{W} \to \mathcal{W}$ as

$$\phi_k(w) = \min\left\{f_k(A_k(w)), \bar{w}^*\right\}.$$

To help us solve the original equilibrium problem, we first show that ϕ has a fixed point.

LEMMA 3. *The fixed-point problem $w = \phi(w)$ has a solution in $\mathcal{W}$.*

Proof. The set $\mathcal{W}$ is compact and convex, and A_k is continuous in w by Lemma 1. The result then follows by Brouwer's fixed point theorem. □

THEOREM 1. *Let $w^* = \phi(w^*)$ be a fixed point of ϕ. Then, w^* also solves the equilibrium problem (3).*

Proof. It is sufficient to show that $f_k(A_k(w^*)) \leq \bar{w}^*$ for all k. Suppose the contrary, i.e., that there exists some k with $f_k(A_k(w^*)) > \bar{w}^*$. Thus, $w_k^* = \phi_k(w^*) = \min\{f_k(A_k(w^*)), \bar{w}^*\} = \bar{w}^*$. At the same time, by the definition of (4), there exists $j \neq k$

such that $f_j\left(A_j\left(w^*\right)\right) \leq \bar{w} < \bar{w}^*$, which implies that $w_j^* = f_j\left(A_j\left(w^*\right)\right)$. Then, for any $x \in \mathcal{S}$, we have

$$d_j(x) + w_j^* \leq d_j(x) + \bar{w} \leq d_k(x) + \bar{w} + |d_j(x) - d_k(x)| < d_k(x) + \bar{w}^*.$$

Consequently, j is always preferred to k, whence $A_k\left(w^*\right) = 0$. Then, we have $f_k\left(A_k\left(w^*\right)\right) = f_k\left(0\right)$, which leads to a contradiction since $f_k\left(0\right) \leq \bar{w}^*$. $\square$

THEOREM 2. *The solution w^* of (3) is unique.*

Proof. We proceed by contradiction: suppose that $w^{(1)} \neq w^{(2)}$ are two solutions of (3). Without loss of generality, suppose that

$$w_1^{(1)} - w_1^{(2)} \geq w_2^{(1)} - w_2^{(2)} \geq ... \geq w_K^{(1)} - w_K^{(2)}, \tag{5}$$

with at least one of the inequalities being strict.

We first argue that $w_1^{(1)} - w_1^{(2)} \geq 0$. To show this, suppose that $w_1^{(1)} - w_1^{(2)} < 0$. Then, by (5), we have $w_k^{(1)} - w_k^{(2)} < 0$ for all k. Since $f_k\left(A\right)$ is a strictly increasing function of A, we have $A_k\left(w^{(1)}\right) < A_k\left(w^{(2)}\right)$ for all k. Then,

$$1 = \sum_k A_k\left(w^{(1)}\right) < \sum_k A_k\left(w^{(2)}\right), \tag{6}$$

which is impossible because $\sum_k A_k\left(w\right) = 1$ for any w. Thus, we have shown $w_1^{(1)} - w_1^{(2)} \geq 0$.

Next, we argue that $w_1^{(1)} - w_1^{(2)} > 0$. Suppose that $w_1^{(1)} - w_1^{(2)} = 0$ (we have already handled the case where the difference is strictly negative). Then, there exists $k < K$ such that $w_j^{(1)} - w_j^{(2)} = 0$ for $j \leq k$ and $w_j^{(1)} - w_j^{(2)} < 0$ for $j > k$. We then obtain (6) again, which is impossible.

Finally, we argue that $w_1^{(1)} - w_1^{(2)} \leq 0$, which will contradict the preceding statement and complete the proof. It follows from (5) that $w_1^{(1)} - w_k^{(1)} \geq w_1^{(2)} - w_k^{(2)}$ for all $k > 1$. By Lemma 1, we have $A_1\left(w^{(1)}\right) \leq A_1\left(w^{(2)}\right)$. By (3), we have $w_1^{(1)} \leq w_1^{(2)}$, as desired. $\square$

3.3. Comparative Statics of the Equilibrium Solution

In this section, we fix the normalized arrival intensity function $m(x)$ and investigate how the expected waiting times change as the aggregate arrival rate λ and the service rates μ_k change. Through (3), the weights of the Voronoi diagram also give the expected waiting times (scaled by c) at each facility. We show that 1) increasing the aggregate arrival rate λ will increase the expected waiting time at every facility, and 2) increasing the service rate μ_k at *any* facility will reduce *all* of the waiting times, with the kth facility seeing the most improvement.

Some technical preliminaries are needed for these results. In linear algebra, a matrix C is said to be an "M-matrix" (Young 1971) if its off-diagonal entries are negative (i.e., $C_{ij} \leq 0$ for $i \neq j$) and the real parts of its eigenvalues are all positive. It is known that all M-matrices are monotone, meaning that $Cv \geq 0$ implies $v \geq 0$ for all v. This notion is used in the following technical lemma.

LEMMA 4. *Fix an arbitrary w and define the matrix D by*

$$D_{jk} = \frac{\lambda c}{\left(\mu_j - \lambda A_j(w)\right)^2} \cdot \frac{\partial A_j}{\partial w_k}.$$

Then, the matrix $I - D$ is an M-matrix.

Proof. By Lemma 1, we have $\frac{\partial A_j}{\partial w_j} \leq 0$ and $\frac{\partial A_j}{\partial w_k} \geq 0$ for $j \neq k$. Consequently, $(I-D)_{jk} \leq 0$ for $j \neq k$, one of the properties required of an M-matrix.

Note from Lemma 1 that $\sum_k \frac{\partial A_j}{\partial w_k} = 0$. Therefore, for any j, we also have

$$\sum_k D_{jk} = \frac{\lambda c}{\left(\mu_j - \lambda A_j(w)\right)^2} \sum_k \frac{\partial A_j}{\partial w_k} = 0.$$

Consequently,

$$\left|(I-D)_{jj}\right| = (I-D)_{jj} > (-D)_{jj} = \sum_{k \neq j} D_{jk} = \sum_{k \neq j} \left|(I-D)_{jk}\right|,$$

whence it follows that the matrix $I - D$ is strictly diagonally dominant. Then, by Thm. 6.1.10 of Horn and Johnson (2013), every eigenvalue of $I - D$ has positive real part, completing the proof. $\square$

We now show that both of the desired structural properties follow by considering equations of the form $(I-D)\,w' = v$, where w' is the vector of partial derivatives of w^* with respect to the parameter of interest, and applying the monotone property of M-matrices.

THEOREM 3. *Let w^* be the unique solution of (3). Then, we have $\frac{\partial w_j^*}{\partial \lambda} \geq 0$ for all j, and $\frac{\partial w_j^*}{\partial \mu_k} \leq 0$ for all j, k. Furthermore, $\frac{\partial w_k^*}{\partial \mu_k} < \frac{\partial w_j^*}{\partial \mu_k}$.*

Proof. To obtain the first result, we differentiate both sides of the equation $w_j = f_j\left(A_j(w)\right)$ with respect to λ and obtain

$$\frac{\partial w_j}{\partial \lambda} = -\frac{c}{\left(\mu_j - \lambda A_j(w)\right)^2}\left(-A_j(w) - \lambda \sum_k \frac{\partial A_j}{\partial w_k} \cdot \frac{\partial w_k}{\partial \lambda}\right),$$

which can be rewritten as the linear system $(I-D)\,w' = v$, where $w'_j = \frac{\partial w_j}{\partial \lambda}$, and

$$v_j = \frac{cA_j(w)}{\left(\mu_j - \lambda A_j(w)\right)^2}.$$

It is clear that $v \geq 0$, so we conclude $w' \geq 0$ because $I - D$ is an M-matrix by Lemma 4.

To obtain the second result, we differentiate both sides of $w_j = f_j(A_j(w))$ with respect to μ_k. Similarly to the first case, we obtain the linear system $(I-D)w' = v$ where $w'_j = \frac{\partial w_j}{\partial \mu_k}$ and

$$v_j = \begin{cases} -\frac{c}{(\mu_j - \lambda A_j(w))^2} & j = k, \\ 0 & j \neq k. \end{cases} \tag{7}$$

Because $v \leq 0$ and $I - D$ is an M-matrix, we obtain $w' \leq 0$.

Finally, we recall that $\sum_k (I-D)_{jk} > 0$ for all j. It then follows by Lemma 3.14 in Chapter 9 of Berman and Plemmons (1994) that $(I-D)^{-1}_{kk} > (I-D)^{-1}_{jk}$ for all $j \neq k$. Returning to the system $(I-D)w' = v$ where v is as in (7), we find

$$w'_k = -\frac{c}{(\mu_k - \lambda A_k(w))^2}(I-D)^{-1}_{kk} < -\frac{c}{(\mu_k - \lambda A_k(w))^2}(I-D)^{-1}_{jk} = w'_j,$$

as desired. □

Theorem 3 allows us to characterize the comparative statics of the equilibrium areas $A_j(w^*)$ with respect to the service rates. As expected, increasing μ_j will make the jth facility more attractive to customers, increasing A_j. Increasing μ_k for $k \neq j$ will reduce A_j, a property that is less obvious than may seem at first: if the kth facility becomes more attractive, some customers may switch to the kth facility from the jth, but this will also have the effect of reducing the load on the jth facility, making it more attractive for other customers. The next result shows that the net effect on A_j will be negative.

PROPOSITION 1. *Let w^* be the unique solution of (3). Then, $\frac{\partial A_j(w^*)}{\partial \mu_k} \leq 0$ for $k \neq j$, and $\frac{\partial A_j(w^*)}{\partial \mu_j} \geq 0$.*

Proof. The first statement follows from the relation $\frac{\partial w_j}{\partial \mu_k} = \frac{\lambda c}{(\mu_j - \lambda A_j(w))^2}\frac{\partial A_j}{\partial \mu_k}$ together with Theorem 3. To obtain the second statement, we write

$$\begin{aligned} \frac{\partial A_j}{\partial \mu_j} &= \frac{\partial A_j}{\partial w_j}\cdot\frac{\partial w_j}{\partial \mu_j} + \sum_{k\neq j}\frac{\partial A_j}{\partial w_k}\cdot\frac{\partial w_k}{\partial \mu_j} \\ &\geq \frac{\partial A_j}{\partial w_j}\cdot\frac{\partial w_j}{\partial \mu_j} + \sum_{k\neq j}\frac{\partial A_j}{\partial w_k}\cdot\frac{\partial w_j}{\partial \mu_j} \qquad (8) \\ &= \frac{\partial w_j}{\partial \mu_j}\sum_k \frac{\partial A_j}{\partial w_k} = 0, \end{aligned}$$

where (8) is due to Theorem 3 and the fact that $\frac{\partial A_j}{\partial w_k} \geq 0$ for $k \neq j$ by Lemma 1. □

4. Social Welfare and the Equilibrium

In this section, we introduce the notion of a "centralized" partition that optimizes the social welfare, and compare it to the decentralized partition studied in Section 3. These results complement, but are not directly connected to the structural properties derived in Section 3. The primary analytical tool for this comparison is an abstract partitioning problem formulated in Section 4.1. The decentralized equilibrium in Section 3 and the two centralized partitions introduced in this section can be viewed as special cases of this general problem. We show that any locally optimal solution of the general problem is described by a particular kind of Voronoi diagram, allowing us to compare different types of partitions visually.

In Section 4.2, we consider a hypothetical situation where the service provider can design the partition and impose it on customers. In other words, the partition is no longer required to satisfy (1). One can then formulate an instance of the abstract partitioning problem to optimize the social welfare under a fixed set of service rates. We characterize the geometry of such a socially optimal partition and compares it with the decentralized partition in equilibrium studied in Section 3. Then, in Section 4.3, we suppose that the service provider can also set the service rates μ subject to a budget constraint. We then show that, if the rates are set in a certain way, the *decentralized* equilibrium partition is identical to the centralized one that achieves the local optimum. In particular, this is true for the *globally* optimal service rates, meaning that the absolute best possible social welfare can be achieved purely through decentralized customer behavior.

Lastly, Section 4.4 discusses how the abstract partitioning problem can be solved computationally. In particular, when the service rates are variable, the objective function of this problem becomes nonconvex, and branch-and-bound methods have to be used to find a global optimum.

4.1. A General Partitioning Problem

Let ν be a probability density on $\mathcal{S}$ that is absolutely continuous with respect to Lebesgue measure. Suppose that we are given continuous and differentiable functions $g_k : \mathbb{R}^n \to \mathbb{R}$, $k = 1, ..., K$, and $G : \mathbb{R}^K \to \mathbb{R}$. We assume that the functions g_k satisfy the following regularity condition: if Y is a random vector with density ν, then the random variable $g_j(Y) - g_k(Y)$, for any $j \neq k$, has a density. This condition holds in all of the specific instances of this problem that we will consider.

We now formulate the problem

$$\min_{\mathcal{A}_1,...,\mathcal{A}_K} G\left(A_1,...,A_K\right)+\sum_k \iint_{\mathcal{A}_k} g_k\left(x\right)\, d\nu, \tag{9}$$

subject to the constraints

$$A_k = \iint_{\mathcal{A}_k} d\nu, \qquad k=1,...,K \tag{10}$$

$$\nu\left(\mathcal{A}_j \cap \mathcal{A}_k\right) = 0, \qquad j \neq k, \tag{11}$$

$$\bigcup_k \mathcal{A}_k = \mathcal{S}. \tag{12}$$

The following analytical result shows that any locally optimal solution of (9)-(12) is described by an additively weighted Voronoi diagram whose weights are found by evaluating the gradient of G at the corresponding areas. We specifically mention *local* optima because G is not assumed to be convex in $(A_1,...,A_K)$, and in fact one of our major results in this section will use a nonconvex instance. However, if G is convex, every local optimum will also be a global optimum, with no additional conditions on g_k required. The proof is highly technical and deferred to the Appendix.

THEOREM 4. *For all k, any local optimum $\bar{\mathcal{A}}$ of (9)-(12) satisfies*

$$g_k\left(x\right)+\frac{\partial G\left(\bar{A}\right)}{\partial A_k} \leq \min_{j\neq k} g_j\left(x\right)+\frac{\partial G\left(\bar{A}\right)}{\partial A_j}, \qquad x \in \bar{\mathcal{A}}_k.$$

Problem (9)-(12) can be viewed as a generalization of the setting of Carlsson et al. (2016), with several important distinctions. First, Carlsson et al. (2016) does not include G in the objective function, and only considers the linear (integral) term. Second, in Carlsson et al. (2016), the areas A_k are fixed ahead of time, whereas in our formulation they are decision variables together with the sets $\mathcal{A}_k$. Our computational approach in Section 4.4 interprets (9)-(12) as an optimization problem with two layers: an inner layer that optimizes the linear part of the cost subject to fixed areas, and an outer layer that optimizes $G(A)$ and the optimal value of the inner problem over A_k. The inner layer is handled in a manner similar to Carlsson et al. (2016), but the outer layer is completely new to our paper and requires additional careful analysis (particularly when G is nonconvex).

4.2. Socially Optimal Partitions

First, we consider the case where the service provider can design the partition and impose it on customers with fixed service rates μ_k. Given a particular set of facility locations and

service rates, and letting $\mathcal{A}_k$ be a partition with corresponding areas A_k, we can define the social welfare as

$$W = \sum_k A_k \frac{c}{\mu_k - \lambda A_k} + \iint_{\mathcal{A}_k} d_k(x)\, m(x) dx, \tag{13}$$

the expected total cost (travel plus waiting) per customer, taken over the spatial distribution of arrivals. Note that the area A_k is also the probability that a new arrival will be served by facility k, hence the expected waiting times in the first term of (13) are weighted by the areas. Thus, the partition that maximizes the social welfare can be solved using the partition problem (9)-(12) by letting $\nu = m$, $g_k(x) = d_k(x)$ and

$$G(A) = \sum_k A_k \frac{c}{\mu_k - \lambda A_k},$$

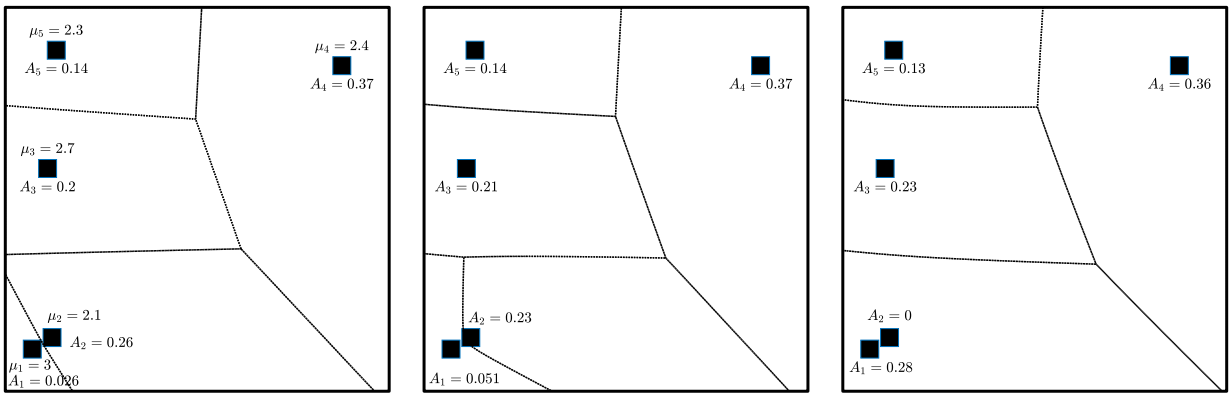
the scaled expected waiting time of a single customer. With these specifications, (9)-(12) can be viewed as a centralized social welfare optimization problem, in which a customer arriving at location $x \in \mathcal{A}_k$ must receive service at the kth facility and cannot choose another one. The objective is precisely the social welfare in (13). Since G is convex in A, any partition satisfying the condition of Theorem 4 is globally optimal.

By Theorem 4, it is described by an additively weighted Voronoi diagram whose weight vector $\bar{w}$ satisfies the *centralized* optimality condition

$$\bar{w}_k = \frac{\partial G(\bar{A})}{\partial A_k} = \frac{c}{\mu_k - \lambda \bar{A}_k} + \frac{\lambda c \bar{A}_k}{\left(\mu_k - \lambda \bar{A}_k\right)^2} = \frac{\mu_k c}{\left(\mu_k - \lambda \bar{A}_k\right)^2}. \tag{14}$$

Clearly, this does not yield the same partition as (3), meaning that, for arbitrary μ, the decentralized demand equilibrium is not socially optimal. The weight $\bar{w}_k$ in the centralized case consists of two parts. The first term is identical to the decentralized case, where customers are self-interested and care about their own waiting times. The second term captures the total externality, i.e., the marginal disutility, that one choosing facility k imposes on others who also choose facility k. The externality term increases as the area associated with the facility increases. It is well-known (Naor 1969 and references in Hassin 2016) that self-interested customers may over-congest a system; equation (14) shows that the centralized partition reduces the area assigned to a facility if it has a high externality term, thus partially mitigating this effect.

Figures 1- 2 illustrate the difference between the decentralized equilibrium whose weights are given by (3), and the centralized socially optimal partition whose weights are given by



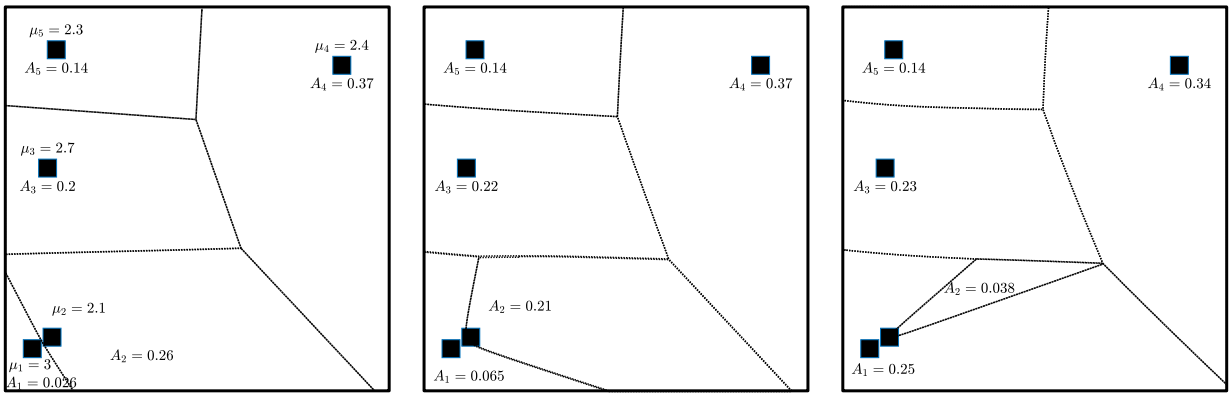
(a) Small c. (b) Medium c. (c) High c.

Figure 1 **Illustration of equilibrium partitions with $\lambda = 1$, fixed service rates, and different c.**

(14), on a simple instance with five facilities. The example instance uses $\mathcal{S} = [0,1]^2$ with Euclidean distance and uniformly distributed demand (i.e., $m(x) \equiv 1$).

Of particular interest is the relative load between facilities 1 and 2, which are located very close together. For small c, the decentralized equilibrium and the centralized partition are mainly determined by travel distance, so the centralized and decentralized partitions are virtually identical. As c increases, Facility 1 (which has a higher service rate) begins to pull some of the load away from facility 2 as c increases as shown in cases (b) and (c).

For large c, the economies of scale at facility 1 become so great that facility 2 becomes totally idle in the decentralized equilibrium, even though it has the ability to serve customers. Thus, we may observe resources going to waste under decentralized customer behavior, an example of the well-known phenomenon of inefficiency in systems with strate-



(a) Small c. (b) Medium c. (c) High c.

Figure 2 **Illustration of centralized partitions with $\lambda = 1$, fixed service rates, and different c.**

gic customers (Ghosh and Hassin 2021). In contrast, the centralized partition in Figure 2(b) creates a more balanced partition by making some use of the resources at facility 2, which is underutilized and has a very small exernality term. In other words, the centralized approach asks 3.8% of customers to accept a higher waiting time at facility 2 in order to improve the conditions for 25% of customers at facility 1. Again, however, this necessity arises because the service capacity at facility 1 was simply too high from the beginning.

However, decentralized behavior is not necessarily the root cause of the inefficiency. The true problem is that a large amount of service capacity has been divided between two facilities in close proximity. For a customer, switching from facility 2 to facility 1 produces only a negligible increase in travel cost, but a significant reduction in waiting time, further amplified by the fact that the total capacity of the system is much higher than the aggregate demand. In Section 4.3, we show that that the inefficiency can be eliminated by jointly optimizing the partition and service rates.

4.3. Socially Optimal Resource Allocation

Different values of μ will lead to different optimal solutions of (9)-(12). One can then define the notion of a globally optimal value for the social welfare, by formulating the problem

$$W^{**} = \min_{\mathcal{A},\mu} \sum_k A_k \frac{c}{\mu_k - \lambda A_k} + \iint_{\mathcal{A}_k} d_k(x) m(x)\, dx, \tag{15}$$

subject to (10)-(12) as well as the budget constraint $\sum_k \mu_k = B$. This is a simple and natural model for resource allocation in spatial service systems. Clearly, W^{**} is a lower bound on the social welfare for any given value of μ.

The objective function in (15) is convex in A for *fixed* μ, but not jointly convex in (A, μ). Therefore, (15) may admit multiple local optima $(\mu^{**}, \mathcal{A}^{**})$. Every such local optimum corresponds to an additively weighted Voronoi diagram whose weights solve yet another type of equilibrium equation, and the service rates μ^{**} have a closed-form dependence on the areas A^{**}. The following theorem gives the characterization. The proof is essentially an application of Theorem 4 to a particular instance of the abstract partitioning problem (9)-(12) in which the function G in (9) is nonconvex.

THEOREM 5. *Any local optimum of (15) is described by an additively weighted Voronoi diagram whose weights w^{**} and areas A^{**} satisfy the equilibrium condition*

$$w_k^{**} = \frac{c}{B - \lambda} \left(\sum_j \sqrt{A_j^{**}} \right) \frac{1}{\sqrt{A_k^{**}}}. \tag{16}$$

Furthermore, the service rates μ^{**} are given by

$$\mu_k^{**} = \lambda A_k^{**} + (B-\lambda)\frac{\sqrt{A_k^{**}}}{\sum_j \sqrt{A_j^{**}}}. \tag{17}$$

Proof. First, let us fix a partition $\mathcal{A}_k$ and solve the problem

$$\min_{\mu} \sum_k A_k \frac{c}{\mu_k - \lambda A_k} + \iint_{\mathcal{A}_k} d_k(x)\,dx \tag{18}$$

subject to $\sum_k \mu_k = B$. Note that, once $\mathcal{A}_k$ is fixed, the integral term in (18) has no dependence on μ and can be omitted. Letting ζ be the Lagrange multiplier of the budget constraint, we write the Lagrangian

$$L(\mu,\zeta) = \sum_k A_k \frac{c}{\mu_k - \lambda A_k} + \zeta\left(\sum_k \mu_k - B\right). \tag{19}$$

Setting $\nabla_\mu L = 0$ yields

$$\zeta = A_k \frac{c}{(\mu_k - \lambda A_k)^2}, \qquad k = 1, ..., K.$$

Equivalently, $(\mu_k - \lambda A_k)\,\zeta = A_k \frac{c}{\mu_k - \lambda A_k}$. Adding up both sides over k yields

$$(B-\lambda)\,\zeta = \sum_k A_k \frac{c}{\mu_k - \lambda A_k}. \tag{20}$$

The right-hand side of (20) is precisely the objective function to be minimized. From (19), we also have

$$\mu_k = \lambda A_k + \frac{\sqrt{cA_k}}{\sqrt{\zeta}}. \tag{21}$$

Adding up both sides of (21) over k and solving for ζ, we obtain

$$\zeta = \frac{c}{(B-\lambda)^2}\left(\sum_k \sqrt{A_k}\right)^2. \tag{22}$$

Substituting (22) into (20), we find that the optimal objective value is

$$\frac{c}{B-\lambda}\left(\sum_k \sqrt{A_k}\right)^2 = \frac{c}{B-\lambda}\|A\|_{\frac{1}{2}},$$

where $\|\cdot\|_{\frac{1}{2}}$ is the $\frac{1}{2}$-quasinorm.

Now, let us return to problem (15). If, for every feasible partition, we set μ optimally according to (21), the problem can be reformulated as

$$\min_{\mathcal{A}} \frac{c}{B-\lambda}\|A\|_{\frac{1}{2}} + \sum_k \iint_{\mathcal{A}_k} d_k(x) m(x)\,dx \tag{23}$$

subject to (10)-(12). This is a special case of Theorem 4 with $G(A) = \frac{c}{B-\lambda}\|A\|_{\frac{1}{2}}$. This function is not convex, but any locally optimal solution is described by an additively

weighted Voronoi diagram with weights

$$w_k^{**} = \frac{\partial G\left(A^{**}\right)}{\partial A_k} = \frac{c}{B-\lambda}\left(\sum_j \sqrt{A_j^{**}}\right)\frac{1}{\sqrt{A_k^{**}}},$$

as required. We then obtain (17) by substituting (22) into (21). □

The local optima described by Theorem 5 are special cases of centralized equilibria. In other words, if we were to fix the service rates to a set of locally optimal values μ^{**}, then the framework of Section 4.2 would yield the same partition as Theorem 5. What is much more surprising, however, is that the local optima described by Theorem 5 are also special cases of *decentralized* equilibria. In other words, setting the service rates to μ^{**} will recover the same partition in both the centralized and decentralized setting. Furthermore, since the *global* optimum is also a local optimum, it follows that we can recover the absolute best possible social welfare, purely through decentralized customer behavior, as long as the service rates are set correctly.

THEOREM 6. *Let $(\mu^{**}, \mathcal{A}^{**})$ be a local optimum of (15). Let $\mathcal{A}^*$ be the equilibrium partition attained in the decentralized model under the service rates μ^{**}. Then, $\mathcal{A}^* = \mathcal{A}^{**}$.*

Proof. By the definition of an additively weighted Voronoi diagram, we have

$$d_k(x) + \frac{c}{B-\lambda}\left(\sum_j \sqrt{A_j^{**}}\right)\frac{1}{\sqrt{A_k^{**}}} \leq \min_{j\neq k} d_j(x) + \frac{c}{B-\lambda}\left(\sum_j \sqrt{A_j^{**}}\right)\frac{1}{\sqrt{A_j^{**}}} \tag{24}$$

for any $x \in \mathcal{A}_k^{**}$. Recall (17) and observe that

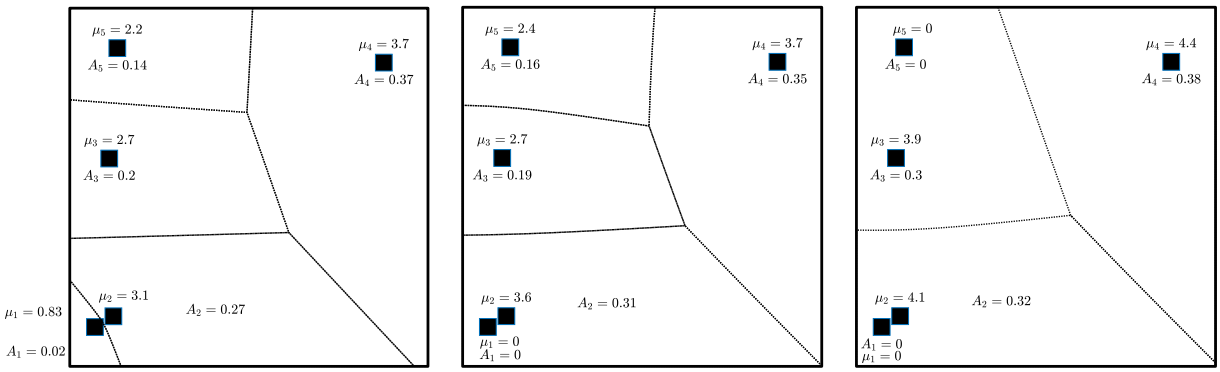
$$\frac{1}{\mu_k^{**} - \lambda A_k^{**}} = \frac{1}{B-\lambda}\left(\sum_j \sqrt{A_j^{**}}\right)\frac{1}{\sqrt{A_k^{**}}}.$$

Consequently, (24) can be rewritten as

$$d_k(x) + \frac{c}{\mu_k^{**} - \lambda A_k^{**}} \leq \min_{j\neq k} d_j(x) + \frac{c}{\mu_k^{**} - \lambda A_k^{**}}.$$

But this is exactly (1). Thus, $\mathcal{A}^{**}$ must be the unique partition $\mathcal{A}^*$ that satisfies the decentralized equilibrium condition (3). □

To understand this striking result, recall from (14) that, for any centralized partition, the Voronoi weight can be decomposed into two parts: the waiting cost incurred by a customer (same as in the decentralized case), plus an additional externality term. When the service rates are chosen optimally, this externality term becomes constant across all facilities. This is because optimizing the service rates forces the marginal value of the resource at each facility to be identical. Since adding a constant to the Voronoi weights does not change



(a) Small c. (b) Medium c. (c) High c.

Figure 3 Illustration of globally optimal partitions with $\lambda = 1$, fixed service rates, and different c.

the partition, the externality no longer has any effect, and the same partition is obtained in both the centralized and decentralized settings.

Figure 3 revisits the instance from Figures 1-2, with the same facility locations, but reallocates the service rates in a globally optimal manner. Earlier, a large amount of capacity was divided between facilities 1 and 2, which are situated very close together. Customers choosing these facilities thus imposed a high externality on each other, and centralization (Figure 2) mitigated this effect only partially. The global optimum balances the externality across facilities by significantly reducing the allocation to facility 1 for small c, and setting that allocation to zero (essentially closing the facility) for larger c. In other words, optimization eliminates the root cause of the inefficiency in Figures 1-2, namely, the fact that facility 1 should never have been assigned so much capacity to begin with. Once the allocation is optimized, the price of anarchy is totally eliminated.

When the waiting time has more impact on the social welfare, the optimal allocation focuses more on economies of scale. In the extreme case $c \to \infty$, where travel distance has no impact at all, it would be optimal to put the entire budget into a single facility. This is why we see more consolidation in Figure 3 as c increases.

4.4. Computation of Equilibrium Partitions

Thus far, we have considered three types of partitions: the decentralized equilibrium from Section 3, the centralized equilibrium from Section 4.2, and the optimal equilibrium from Section 4.3. Each type of partition is the optimal solution to a particular instance of problem (9)-(12) with a different definition of G:

- Decentralized equilibrium: $G(A) = -\frac{c}{\lambda}\sum_k \log(\mu_k - \lambda A_k)$.

- Centralized equilibrium: $G(A) = \sum_k A_k \frac{c}{\mu_k - \lambda A_k}$.
- Optimal equilibrium: $G(A) = \frac{c}{B-\lambda}\|A\|_{\frac{1}{2}}$.

For the decentralized equilibrium, G does not have a clear economic meaning in and of itself, but by Theorem 4, differentiating it will produce the Voronoi weights that satisfy (3). For both centralized and decentralized equilibria, G is separable and convex, but it is neither of these in the third case. Thus, the computational challenge is to solve the general partition problem (9)-(12) for a general nonconvex G.

We use a two-stage reformulation of (9)-(12) in which the first stage has the form

$$\min_{A_1,...,A_K} G(A_1,...,A_K) + \Pi(A_1,...,A_K)$$

subject to $\sum_k A_k = 1$ and $A_k \geq 0$ for $k = 1,...,K$. The function Π is the optimal value of the second-stage problem

$$\Pi(A_1,...,A_K) = \min_{\mathcal{A}_1,...,\mathcal{A}_K} \sum_k \iint_{\mathcal{A}_k} g_k(x)\, d\nu$$

subject to (10)-(12). Thus, the inner problem finds the optimal partition under fixed areas A_k, while the outer problem optimizes over A. The inner problem is thus infinite-dimensional, while the outer problem is finite-dimensional.

Using the theory of semidiscrete optimal transport (Hartmann and Schuhmacher 2020), we show in the Appendix (the proof of Theorem 4) that the gradient $\nabla_A \Pi$ is the vector of optimal Lagrange multipliers for the constraint (10) in the inner problem. For any fixed A satisfying $A_k > 0$ for all k, we can compute $\nabla_A \Pi$ with reasonable accuracy by discretizing $\mathcal{S}$, which is a viable strategy for a two-dimensional geographical region. The integral in (10) is then approximated by a sum, and the entire inner problem reduces to an instance of the well-known "transportation problem" (Ford and Fulkerson 1956), which can be solved using linear programming. We then use the dual variables of the constraint (10) as proxies for $\nabla_A \Pi$.

Since G has a closed form, we can easily evaluate its gradient. Given a fixed A, we thus have an approximation of the gradient $\nabla_A G(A) + \nabla_A \Pi(A)$ of the outer problem. We may now use any standard continuous optimization algorithm that uses gradient information, e.g., a first-order barrier method (Boyd and Vandenberghe 2004). Such a method will find a locally optimal A^*, and by Theorem 4, the Voronoi weights that characterize the corresponding partition are simply $\nabla_A G(A^*)$. A minor technical complication in the implementation of the method is the presence of the equality constraint $\sum_k A_k = 1$, but

this can be handled by fixing $A_K = 1 - \sum_{k=1}^{K-1} A_k$ and optimizing over the remaining $K-1$ areas with inequality constraints only.

For the first two types of equilibria, this approach is sufficient because G is convex in those cases, and therefore any locally optimal solution is globally optimal. For the third type of equilibrium, where G is nonconvex, it is possible to find the global optimum by using the branch-and-bound method (Horst and Tuy 2013). This well-known approach imposes additional constraints of the form $A_k^\ell \leq A_k \leq A_k^u$ for all k, thus dividing the feasible region of the outer problem (the set of feasible A) into "blocks." The global optimum can be provably obtained as long as we have a tractable lower bound on the objective value $G(A) + \Pi(A)$ that becomes tight as the size of the block vanishes to zero. We use the previous method to optimize over each block separately. Then, we eliminate those blocks whose lower bounds are worse than the best solution found thus far (the "bound" in branch-and-bound), and "branch" on the remaining blocks by splitting the intervals $\left[A_k^\ell, A_k^u\right]$.

Thus, all that is needed to find the global optimum is a lower bound on the objective $G(A) + \Pi(A)$. Carlsson et al. (2016) provides a lower bound on the second term $\Pi(A)$. For the first term, we recall from (15) and Theorem 5 that $G(A) = \sum_k A_k \frac{c}{\mu_k - \lambda A_k}$ with the service rates set to $\mu_k = \lambda A_k + (B - \lambda) \frac{\sqrt{A_k}}{\sum_j \sqrt{A_j}}$. Each term $A_k \frac{c}{\mu_k - \lambda A_k}$ is decreasing in μ_k and increasing in A_k, and so we may write

$$G(A) \geq \sum_k A_k^\ell \frac{c}{\mu_k^u - \lambda A_k^\ell},$$

where $\mu_k^u = \lambda A_k^u + (B - \lambda) \frac{\sqrt{A_k^u}}{\sum_j \sqrt{A_j^\ell}}$. It is routine to verify that this bound becomes tight as $A_k^u - A_k^\ell \to 0$.

5. Equilibria and Social Optimality Under Random Shocks

Suppose now that customer decisions are subject to additional random shocks. That is, given loads A_k on the facilities, a new customer arriving at location x will prefer the kth facility if

$$d_k(x) + f_k(A_k) + \tau_k \leq \min_{j \neq k} d_j(x) + f_j(A_j) + \tau_j, \tag{25}$$

where the random variables τ_k are identically distributed and independent of the arrival process, service times, and each other. The presence of such random shocks in the model can be viewed as a form of customer heterogeneity, reflecting differences in perception

between individual customers of the nominal utilities of the choices. If we assume that each τ_k follows a Gumbel distribution, (25) becomes an instance of the well-known multinomial logit (MNL) choice model, used by, e.g., Armony and Maglaras (2004) to represent customer decisions in an unobservable queue. Other alternatives are also possible: for example, if τ_k are exponentially distributed, (25) will be an instance of the exponomial choice model (Alptekinoglu and Semple 2016).

Such models are adopted in part because they provide tractable expressions for the fractions of customers that choose each option, provided that the nominal utilities of the options are fixed. For example, under the MNL model, the fraction of customers choosing an option with nominal disutility a_k is proportional to e^{-a_k}. In our setting, however, this is no longer straightforward, as the nominal disutility of visiting the kth facility now depends on the load on that facility, which is determined by customer choices. The equilibrium demand is no longer a purely geometric partition of $\mathcal{S}$ since customer choice is now probabilistic. Thus, we can no longer visualize it as a Voronoi diagram on a plane. Instead, the equilibrium is a partition of a higher-dimensional space encompassing both the geographical location and the idiosyncratic preferences of a customer.

We denote this location-preference space by $\tilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}^K$. The "location" at which a customer arrives is now described by a vector of the form $(x, \tau_1, ..., \tau_K)$. Thus, demand still follows a spatio-temporal Poisson process, but the intensity function is now $m \times h^K$, where h is the common density of the random shocks. The "distance" between a customer at $(x, \tau_1, ..., \tau_K)$ and facility k is replaced by

$$\tilde{d}_k\left(x, \tau_1, ..., \tau_K\right) = d_k(x) + \tau_k.$$

Since the random shocks can be negative, the "distance" function may take negative values as well. While this is not a typical setting for Voronoi diagrams, it does not affect the theory because, in (25), customer choices depend on the differences between $\tilde{d}_k$ values. Thus, we may simply repeat the setup of Section 3. Given a fixed vector w, we define a partition $\{\mathcal{A}_k\}_{k=1}^K$ of $\tilde{\mathcal{S}}$ where

$$\mathcal{A}_k = \left\{ (x, \tau) : \tilde{d}_k\left(x, \tau\right) + w_k \leq \min_{j \neq k} \tilde{d}_j\left(x, \tau\right) + w_j \right\}.$$

The area $A_k\left(w\right)$ of the set $\mathcal{A}_k$ is the proportion of customers who prefer facility k, given by

$$A_k\left(w\right) = \iint_{\mathcal{A}_k} \left(\prod_{j=1}^{K} h(\tau_j) d\tau_j \right) m(x)\, dx$$

$$= \iint_{\mathcal{S}} P\left(d_k(x) + w_k + \tau_k \leq \min_{j\neq k} d_j(x) + w_j + \tau_j\right) m(x)\,dx$$
$$= \iint_{\mathcal{S}} \left(\int \prod_{j\neq k} \bar{H}\left(d_k(x) - d_j(x) + w_k - w_j + \tau\right) h(\tau)\,d\tau\right) m(x)\,dx, \qquad (26)$$

where $\bar{H}$ is the tail of the common distribution of the random shocks. The equilibrium weights w_k^* are, again, the solution to (3) with this new definition of A_k. The values $A_k(w^*)$ are precisely the proportions of customers that choose each option. Recall from Section 3 that existence, uniqueness, and structural properties of the equilibrium all followed from properties of the area function established in Lemma 1. Therefore, the same results will hold in the present setting as long as (26) can be shown to have these same properties. This is fairly straightforward to show.

LEMMA 5. *The area $A_k(w)$ defined in* (26) *has all of the properties listed in Lemma 1.*

Proof. The continuity of A_k in w follows directly from (26) as it is an integral. To show the second property, note $w_k - \tilde{w}_k \geq w_j - \tilde{w}_j$, then $w_k - w_j \geq \tilde{w}_j - \tilde{w}_k$. Since the tail $\bar{H}$ decreases in $w_k - w_j$, we have $A_k(w) \leq A_k(\tilde{w})$ by (26). The third property also follows from (26) as $\bar{H}$ is a decreasing function. By taking the derivatives of (26), we have

$$\frac{\partial A_k(w)}{\partial w_k} = \sum_{j\neq k} \iint_{\mathcal{S}} -\left(\int \prod_{l\neq j,k} \tilde{H}_l h_j(\tau) h(\tau)\,d\tau\right) m(x)\,dx = -\sum_{j\neq k} \frac{\partial A_k(w)}{\partial w_j},$$

where $\bar{H}_l = \bar{H}\left(d_k(x) - d_l(x) + w_k - w_l + \tau\right)$ and $h_j(\tau) = h\left(d_k(x) - d_j(x) + w_k - w_j + \tau\right)$. This proves the last property. □

What is perhaps more surprising is that our analysis of social optimality from Section 4 also continues to hold. For the sake of argument, let us imagine a central planner with the ability to observe the precise values of the random shocks for every customer. In other words, the planner knows exactly where each customer arrives in the location-preference space. The planner then partitions the space, assigning customers to facilities in a manner that optimizes the social welfare, given by

$$W = \sum_k A_k \frac{c}{\mu_k - \lambda A_k} + \iint_{\mathcal{A}_k} \tilde{d}_k\left(x, \tau_1, ..., \tau_K\right) \left(\prod_{j=1}^{K} h(\tau_j) d\tau_j\right) m(x)\,dx.$$

In essence, the planner is solving the same social welfare optimization problem as in Section 4.2, but on a different space $\tilde{S}$ and with different distance functions $\tilde{d}_k$. The term $G(A)$ in the objective function remains unchanged from Section 4.2 because the expected waiting time still has the same dependence on the proportions A_k; we have only changed the

manner in which the proportions are computed. Applying Theorem 4, we find that the optimal partition $\bar{\mathcal{A}}$ (not of the plane this time, but of $\mathcal{S} \times \mathbb{R}^K$) is described by

$$d_k(x) + \frac{\partial G\left(\bar{A}\right)}{\partial A_k} + \tau_k \leq \min_{j \neq k} d_j(x) + \frac{\partial G\left(\bar{A}\right)}{\partial A_j} + \tau_j, \qquad (x, \tau_1, ..., \tau_K) \in \bar{\mathcal{A}}_k,$$

with A_k computed according to (26). Since the analysis in Section 4 does not depend on the specific choice of distance function, all of the results continue to hold. It is important to note that the *numerical values* of the areas A_k, or the optimal service rates μ_k^{**}, will not be the same as under deterministic choice, because the areas are now computed in a different way based on (26).

In sum, we still have the key result that any local optimum is achievable in the decentralized setting. Arguably the result has become even stronger: all we need to compute μ_k^{**} is the *distribution* of the random shocks, but the social welfare that can be attained under locally optimal values of the service rates will be as good as if we had known the precise realizations of the shocks.

Lastly, we comment on the computation of the optimal partitions. Essentially, the technique is the same as in Section 4.4. In the context of random choice models, it was shown by Anderson et al. (1989) that the second-stage objective $\Pi(A)$ admits a lower-dimensional reformulation; one can apply Algorithm 3 of Carlsson et al. (2016) to obtain the dual variables of this problem efficiently. However, the solution becomes more cumbersome to visualize, as the spatial region can no longer be neatly divided between facilities. Figure 4 presents a simple example with four facilities, which are located symmetrically in $[0, 1]^2$ with uniformly distributed demand. Any facility can be chosen by any customer, but the choice probabilities depend on the location. The probability of choosing each facility is visualized a separate 3D plot. The probability of choosing a facility is highest in a neighborhood immediately surrounding it, but due to the differing service rates, the maximum value attained by each probability is quite different. For the facility with the highest service rate, the choice probability is above 0.5 in some parts of the region, but the facility with the lowest service rate is never chosen with probability above 0.2.

6. Case Study: Hospital Beds in LA County

In this section, we present an in-depth illustration of the insights that our framework can provide on a problem instance based on realistic population and resource allocation data. Specifically, we use the geographical locations of the seven largest hospitals in Los

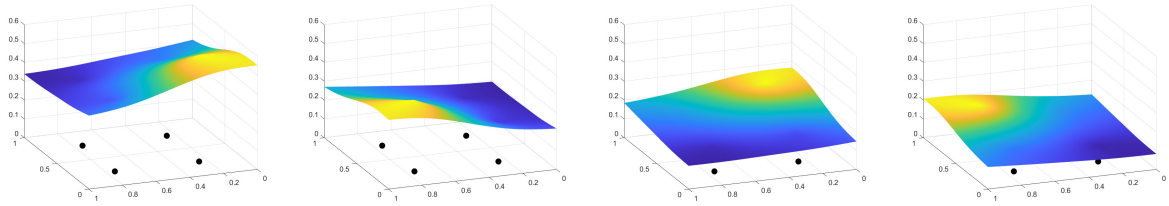


Figure 4 **Decentralized equilibrium for four facilities with service rates $1/2$, $1/3$, $1/4$, and $1/5$ under MNL choice.**

Angeles County as our facilities, and the number of beds in each hospital as a proxy for the service rates μ_k. Using the procedures in Section 4.4, we compute both centralized and decentralized equilibria under these rates and compare them against the global optimum. In addition, we use our framework to perform "what-if" analysis on the possible addition (and geographical placement) of a hypothetical eighth facility.

We extracted a dataset of 90,855 census blocks comprising Los Angeles County, as well as their populations, from the United States Census. Distances were rescaled so that all blocks are located in the unit square. The population data provides an atomic intensity $m(x)$ for the arrival process. The total arrival rate λ was set to 1, as the presence of the constant c in our model allows us to scale the arrival and service rates according to our convenience. The total budget was set to $B = 1.25$, corresponding to an aggregate occupancy rate of 80%, which is consistent with national standards (Phillip et al. 1984) and expert understanding of hospital effectiveness (Keegan 2010). Letting n_k be the (publicly available) number of beds in hospital k, we let $\mu_k = \frac{n_k}{\sum_j n_j} B$ be the proportion of the budget currently assigned to the corresponding facility. We normalize the space, letting $\mathcal{S} = [0,1]^2$, and use Euclidean distance to compute travel cost. Since both travel distances and arrival/service rates have been normalized, we pick a small $c = 0.001$ to achieve a reasonable balance between the two types of costs. We will briefly discuss the effect of varying this parameter later on.

For this fixed resource allocation, we compute the decentralized equilibrium from Section 3 and the centralized equilibrium from Section 4.2. Thus, we can test whether, and how much, centralization would help to increase the efficiency of the current allocation. We also compute the globally optimal allocation, described in Theorem 5, by running the procedures in Section 4.4 with different starting points to avoid getting stuck in a local optimum. This allows us to compare the current allocation against the most efficient one.

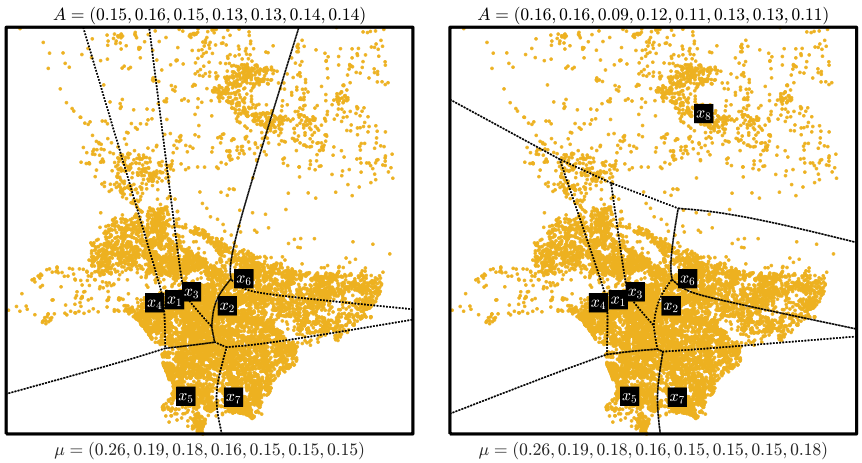
Additionally, we consider a hypothetical scenario in which an eighth facility is added. We assume that the number of beds in this facility is given by $n_8 = \frac{1}{7}\sum_{k=1}^{7} n_k$, i.e., the new

facility has average capacity relative to the others. The budget is increased accordingly to reflect this new capacity. Under this modified allocation, we again compute centralized and decentralized equilibria, and compare these against the global optimum in which all eight service rates are variable. These computations are performed for 64 candidate locations for the new facility, and the best placement is reported for each type of equilibrium. In this way, we can test how the optimal placement and resource allocation depend on the equilibrium type.

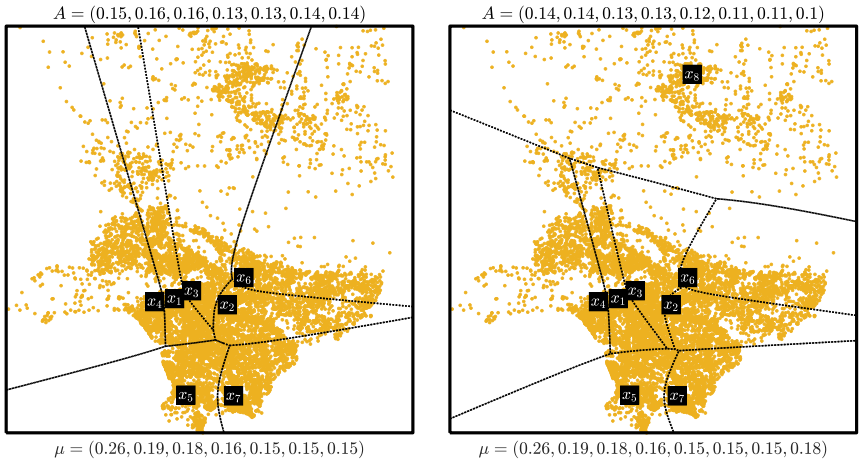
Figure 5 shows each type of equilibrium partition under both 7 and 8 facilities, with the population density shown using yellow dots. Several insights can be obtained from this comparison:

First, although we have seen in Figures 1-2 that the centralized and decentralized partitions can be quite different, the aggregate demand in that stylized instance was much lower than the total service capacity. In marked contrast, centralization offers little benefit in a more realistic setting with high demand. Even under a suboptimal resource allocation, where the centralized and decentralized partitions do not coincide, they remain quite similar in the 7-hospital scenario. This is because, when λ is close to B, every facility will experience high load and there is little room for improvement by transferring part of the load from one facility to another. There is more of a difference in the 8-hospital scenario: the presence of additional capacity allows the centralized partition to shift some demand away from more congested facilities (e.g., facilities 2, 5, and 6) and make the load more balanced. However, in high-demand situations, decentralization is not a significant source of inefficiency.

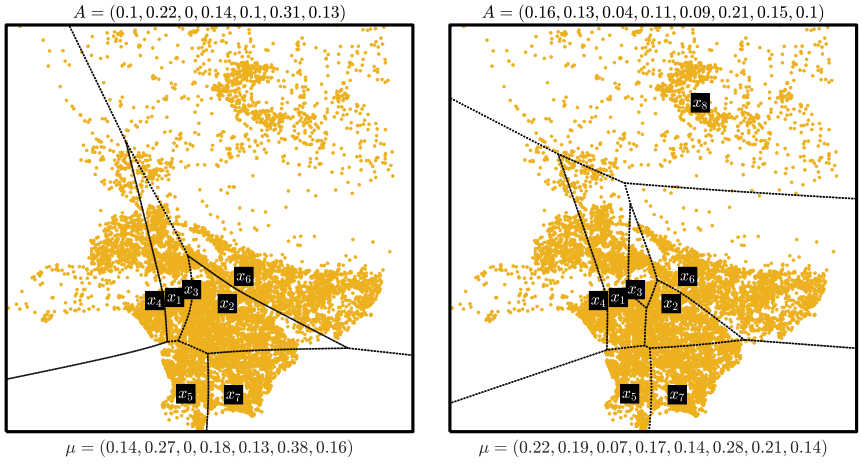
On the other hand, even when the demand is high, there is significant improvement to be had by optimizing the *resource allocation*. As we saw earlier in Figure 3, the globally optimal allocation tends to consolidate demand and make use of economies of scale. We see this tendency in Figure 5 as well. For example, in Figure 5(a), the relatively sparsely populated northern part of the map is divided between four facilities. The global optimum, in Figure 5(e), reallocates capacity so that most of this demand is served by facility 6, which is given a much higher service rate than the others. Additionally, from a purely spatial standpoint, the placement of the seven existing hospitals is inefficient: facilities 1, 3 and 4 are very close together. This encourages the global optimum to close facility 3,



(a) Decentralized partition, 7 facilities. (b) Decentralized partition, 8 facilities.



(c) Centralized partition, 7 facilities. (d) Centralized partition, 8 facilities.



(e) Global optimum, 7 facilities. (f) Global optimum, 8 facilities.

Figure 5 Optimal partitions with 7 and 8 facilities under different equilibrium types.

as it offers virtually no savings in travel cost over its neighbors (however, this should be contrasted with the 8-facility scenario as discussed below).

Next, we observe that the equilibrium type may not necessarily affect the optimal placement of a new facility: perhaps unsurprisingly, in all three cases, the best location is in the isolated population cluster in the northern part of the map. However, the presence of the new facility significantly changes the optimal *partition*. One particularly interesting observation is that facility 3, which was closed in Figure 5(e), is actually open again (albeit with a small service rate) when a new facility is added. This is because facility 6 no longer has to handle such a large portion of the demand, as part of it is now served by the new facility. The service rate of facility 6 can then be reduced, and as a result, facility 3 again becomes more attractive to the population in its immediate vicinity. Thus, somewhat surprisingly, the inclusion of a new facility with additional capacity can actually reduce the degree of consolidation and spread the capacity out more evenly. This behavior is due to the complex spatial relationships between facilities, and can only be observed when the spatial position of arrivals is modeled explicitly.

We also considered other values of c, but they did not appreciably change these insights. As observed previously in Figure 3, higher c leads to a more consolidated global optimum, eventually closing all but one of the facilities in extreme cases. This will also happen in the present setting. In our opinion, however, it is more informative to examine situations where there is a non-trivial tradeoff between travel cost and waiting time, as these are the cases where we see the most interesting distinctions between equilibrium types and partitions.

7. Concluding Remarks

We have presented a novel framework for describing, studying, computing, and visualizing different types of demand equilibria in spatial service systems, where customers incur cost based on both travel and waiting time, and the load on each queue is endogenized by arrival locations. Our work is the first to provide a geometric characterization and interpretation of demand equilibria in such systems. We formulate a general mathematical program that can be used to compute both centralized and decentralized equilibria, and we prove that these two types of equilibria coincide when the service capacity is allocated optimally between facilities.

A consistent message of our work is that the "price of anarchy" in spatial service systems is, relatively, of less importance than the efficiency of the resource allocation. Even if the allocation is suboptimal, a high aggregate load on the system will produce very similar partitions for centralized vs. decentralized paradigms. On the other hand, an optimal resource allocation can change the partition dramatically, and moreover, this optimal partition is achievable purely through decentralized behavior.

There are many possible avenues for future work. One possible direction is to endogenize the arrival rate, so that the frequency of service is also determined by customers (e.g., in the leisure or hospitality industry). Interdependence and interaction between arrivals at different locations, using, e.g., self-exciting spatio-temporal demand processes, could be another way to model endogenous arrival rates. Since the waiting time function in such models depends on the region rather than the area, the equilibrium equation would become infinite-dimensional, introducing substantial challenges. Yet another direction would be to introduce a facility location decision into the model, so that, for example, the new hospital in our case study could be placed algorithmically rather than through exhaustive search. Such an algorithm would require considerable new developments in optimization theory.

References

Alptekinoglu A, Corbett CJ (2010) Leadtime-variety tradeoff in product differentiation. *Manufacturing & Service Operations Management* 12(4):569–582.

Alptekinoglu A, Semple JH (2016) The exponomial choice model: A new alternative for assortment and price optimization. *Operations Research* 64(1):79–93.

Anderson SP, De Palma A, Thisse JF (1989) Demand for differentiated products, discrete choice models, and the characteristics approach. *The Review of Economic Studies* 56(1):21–35.

Armony M, Maglaras C (2004) Contact centers with a call-back option and real-time delay information. *Operations Research* 52(4):527–545.

Aronov B, Carmi P, Katz MJ (2009) Minimum-cost load-balancing partitions. *Algorithmica* 54(3):318–336.

Aurenhammer F (1991) Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Computing Surveys* 23(3):345–405.

Aurenhammer F, Klein R (2000) Voronoi diagrams. Sack JR, Urrutia J, eds., *Handbook of Computational Geometry*, 201–290 (Elsevier Science B. V.).

Baron O, Economou A, Manou A (2022) Increasing social welfare with delays: Strategic customers in the M/G/1 orbit queue. *Production and Operations Management* 31(7):2907–2924.

Berman A, Plemmons RJ (1994) *Nonnegative matrices in the mathematical sciences* (SIAM).

Boyd S, Vandenberghe L (2004) *Convex optimization* (Cambridge University Press).

Carlsson JG, Carlsson E, Devulapalli R (2016) Shadow prices in territory division. *Networks and Spatial Economics* 16(3):893–931.

Carlsson JG, Devulapalli R (2013) Dividing a territory among several facilities. *INFORMS Journal on Computing* 25(4):730–742.

Chao X, Liu L, Zheng S (2003) Resource allocation in multisite service systems with intersite customer flows. *Management Science* 49(12):1739–1752.

Devulapalli R, Peterson N, Carlsson JG (2015) Data visualization using weighted Voronoi diagrams. Bozkaya B, Singh VK, eds., *Geo-Intelligence and Visualization through Big Data Trends*, 181–204 (IGI Global).

Dieker AB, Ghosh S, Squillante MS (2017) Optimal resource capacity management for stochastic networks. *Operations Research* 65(1):221–241.

Ding Y, Nagarajan M, Zhang Z (2022) Parallel queues with discrete-choice arrival pattern: Empirical evidence and asymptotic characterization. *Available at SSRN 3584880* .

Dong J, Yom-Tov E, Yom-Tov GB (2019) The impact of delay announcements on hospital network coordination and waiting times. *Management Science* 65(5):1969–1994.

Economou A (2021) The impact of information structure on strategic behavior in queueing systems. Anisimov V, Limnios N, eds., *Queueing Theory*, volume 2, 137–169 (John Wiley and Sons, New York).

Ford LR, Fulkerson DR (1956) Solving the transportation problem. *Management Science* 3(1):24–32.

Gavirneni S, Kulkarni VG (2016) Self-selecting priority queues with Burr distributed waiting costs. *Production and Operations Management* 25(6):979–992.

Ghosh S, Hassin R (2021) Inefficiency in stochastic queueing systems with strategic customers. *European Journal of Operational Research* 295(1):1–11.

Grossman TA, Brandeau ML (2002) Optimal pricing for service facilities with self-optimizing customers. *European Journal of Operational Research* 141(1):39–57.

Hartmann V (2016) A geometry-based approach for solving the transportation problem with Euclidean cost. Technical report, Georg August University of Göttingen.

Hartmann V, Schuhmacher D (2020) Semi-discrete optimal transport: a solution procedure for the unsquared Euclidean distance case. *Mathematical Methods of Operations Research* 92(1):133–163.

Hassin R (2016) *Rational queueing* (Chapman and Hall/CRC).

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Springer New York, NY).

Hassin R, Roet-Green R (2020) On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management* 23(4):989–1004.

Haugland D, Ho SC, Laporte G (2007) Designing delivery districts for the vehicle routing problem with stochastic demands. *European Journal of Operational Research* 180(3):997–1010.

Haviv M, Oz B (2018) Self-regulation of an unobservable queue. *Management Science* 64(5):2380–2389.

Heinhold M (1978) An operational research approach to allocation of clients to a certain class of service institutions. *Journal of the Operational Research Society* 29(3):273–276.

Heyman DP, Sobel MJ (2003) *Stochastic Models in Operations Research: Stochastic Processes and Operating Characteristics* (Dover Publications).

Horn RA, Johnson CJ (2013) *Matrix Analysis (2nd ed.)* (Cambridge University Press).

Horst R, Tuy H (2013) *Global optimization: deterministic approaches* (Springer Science & Business Media).

Karlin S, Taylor HM (1975) *A first course in stochastic processes (2nd ed.)* (Academic Press, Inc.).

Keegan AD (2010) Hospital bed occupancy: more than queuing for a bed. *Medical Journal of Australia* 193(5):291–293.

Kullman ND, Goodson JC, Mendoza JE (2021) Electric vehicle routing with public charging stations. *Transportation Science* 55(3):637–659.

Lee HL, Cohen MA (1985) Equilibrium analysis of disaggregate facility choice systems subject to congestion-elastic demand. *Operations Research* 33(2):293–311.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Pavone M, Arsie A, Frazzoli E, Bullo F (2011) Distributed algorithms for environment partitioning in mobile robotic networks. *IEEE Transactions on Automatic Control* 56(8):1834–1848.

Pender J, Rand R, Wesson E (2020) A stochastic analysis of queues with customer choice and delayed information. *Mathematics of Operations Research* 45(3):1104–1126.

Phillip PJ, Mullner R, Andes S (1984) Toward a better understanding of hospital occupancy rates. *Health Care Financing Review* 5(4):53–61.

Rajan B, Tezcan T, Seidmann A (2019) Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science* 65(3):1236–1267.

Ricca F, Scozzari A, Simeone B (2008) Weighted Voronoi region algorithms for political districting. *Mathematical and Computer Modelling* 48(9-10):1468–1477.

Shanthikumar JG, Xu SH (1997) Asymptotically optimal routing and service rate allocation in a multiserver queueing system. *Operations Research* 45(3):464–469.

Wang Z, Cui S, Fang L (2023) Distance-based service priority: An innovative mechanism to increase system throughput and social welfare. *Manufacturing & Service Operations Management* 25(1):353–369.

Xu X, Lian Z, Li X, Guo P (2016) A Hotelling queue model with probabilistic service. *Operations Research Letters* 44(5):592–597.

Young DM (1971) *Iterative solution of large linear systems* (Academic Press, Inc.).

Yushimito WF, Jaller M, Ukkusuri S (2012) A Voronoi-based heuristic algorithm for locating distribution centers in disasters. *Networks and Spatial Economics* 12(1):21–39.

Zhou J, Ryzhov IO (2021) Equilibrium analysis of observable express service with customer choice. *Queueing Systems* 99(3-4):243–281.

8. Appendix: Proofs

In the following, we give proofs for all results that were stated in the main text.

8.1. Proof of Theorem 4

We first observe that (9) admits an equivalent representation

$$\min_{A_1,...,A_K} G(A_1,...,A_K) + \inf_{\substack{\nu(\mathcal{A}_k)=A_k \\ \mathcal{A}_j \cap \mathcal{A}_k = \emptyset}} \sum_k \iint_{\mathcal{A}_k} g_k(x)\, d\nu, \tag{27}$$

in which (9) is expressed as the sum of an "outer" problem over $(A_1,...,A_K) \in \mathbb{R}_+^K$ and an "inner" problem over the space of all partitions $(\mathcal{A}_1,...,\mathcal{A}_K)$ such that $\nu(\mathcal{A}_k) = A_k$ for all k. We state an auxiliary result that describes the optimal solution of the inner problem; the proof is given in a separate section of the online supplement.

LEMMA 6. *The optimal solution $\mathcal{A}'_k$ of the inner problem*

$$\min_{\mathcal{A}_1,...,\mathcal{A}_K} \sum_k \iint_{\mathcal{A}_k} g_k(x)\, d\nu, \tag{28}$$

subject to the constraints

$$\begin{aligned} A_k &= \nu(\mathcal{A}_k), \\ \nu(\mathcal{A}_j \cap \mathcal{A}_k) &= 0, \qquad j \neq k, \\ \bigcup_k \mathcal{A}_k &= \mathcal{S}, \end{aligned}$$

satisfies

$$\mathcal{A}'_k = \{x \in \mathcal{S} : g_k(x) - \eta_k \leq g_j(x) - \eta_j\} \tag{29}$$

for all k and some constants η_k.

By (29), we may assume that $\eta_K = 0$ without loss of generality. This has the effect of reducing the dimension of the problem by 1. We can restate Theorem 4 so that the objective function depends only on $\mathcal{A}_1,...,\mathcal{A}_{K-1}$. Problem (9)-(12) becomes

$$\min_{\mathcal{A}_1,...,\mathcal{A}_{K-1}} G(A_1,...,A_{K-1}) + \sum_{k=1}^K \iint_{\mathcal{A}_k} g_k(x)\, d\nu \tag{30}$$

subject to $\nu(\mathcal{A}_k) = A_k$ for $k = 1,...,K-1$ as well as (11)-(12). The final region $\mathcal{A}_K$ is uniquely determined by the others. Then, the desired result can be equivalently stated as follows: the optimal partition $\bar{\mathcal{A}}$ for problem (30) is of the form (29) with weights $\bar{\eta}$ satisfying $-\bar{\eta} = \nabla G$.

Using a Kantorovich duality argument, as in the proof of Lemma 6, allows us to rewrite (30) as

$$\min_{A_1,...,A_{K-1}} G(A_1,...,A_{K-1}) + \sup_{\eta\in\mathbb{R}^{K-1}} \left(\iint_{\mathcal{S}} \min_k \left\{ g_K(x), \min_k g_k(x) - \eta_k \right\} d\nu + \sum_k^{K-1} A_k\eta_k \right), \tag{31}$$

subject to the constraints $A_k \geq 0$ and $\sum_{k=1}^{K-1} A_k \leq 1$. This formulation now depends entirely on $A = (A_1,...,A_{K-1})$. Given any fixed value for this vector, Lemma 6 can be used to recover the partition. The function

$$h^*(A) = \sup_{\eta\in\mathbb{R}^{K-1}} \left(\iint_{\mathcal{S}} \min_k \left\{ g_K(x), \min_k g_k(x) - \eta_k \right\} d\nu + \sum_k^{K-1} A_k\eta_k \right)$$

is, by definition, the convex conjugate of the function

$$h(\eta) = -\iint_{\mathcal{S}} \min_k \left\{ g_K(x), \min_k g_k(x) - \eta_k \right\} d\nu.$$

Furthermore, $\frac{\partial h}{\partial \eta_k} = \nu(\mathcal{A}'_k)$ where the sets $\mathcal{A}'_k$ are as in (29). Moreover, $h(\eta)$ induces a weighted Voronoi partition of the form

$$\mathcal{A}_k(\eta) = \left\{ x\in\mathcal{S} : g_k(x) - \eta_k \leq \min_{j\neq k} g_j(x) - \eta_j \right\}$$

with $\eta_K = 0$.

By assumption, G is differentiable. The KKT conditions for problem (31), which are necessary and sufficient for local optimality of a vector $\bar{A}$, are given by

$$0 \in \nabla G\left(\bar{A}\right) + \partial h^* - \zeta + \zeta_K \cdot e, \tag{32}$$

$$\zeta_k \bar{A}_k = 0, \qquad k = 1,...,K-1, \tag{33}$$

$$\zeta_K \left(1 - \sum_{k=1}^{K-1} \bar{A}_k \right) = 0, \tag{34}$$

$$\zeta_k \geq 0, \qquad k = 1,...,K-1. \tag{35}$$

Equations (33)-(34) are the complementary slackness conditions for the constraints on A. In (32), e denotes a vector in $\mathbb{R}^{K-1}$ whose elements are all equal to 1, and ∂h^* denotes the Clarke generalized subdifferential of h^*. When $\bar{A}_k > 0$, the kth element of ∂h^* is simply the partial derivative $\frac{\partial h^*\left(\bar{A}\right)}{\partial A_k}$. However, when $\bar{A}_k = 0$, the partial derivative is not defined, requiring the more general notion of a subdifferential.

Because there is always at least one nonzero element of $\bar{A}$, we may assume without loss of generality that $\zeta_K = 0$. Recall from convex analysis that $A \in \partial h(\eta)$ if and only if

$\eta \in \partial h^*(A)$, a basic fact about conjugacy and subdifferentials (Prop. 1.4.3 of Hiriart-Urruty and Lemaréchal 2004). Since h is differentiable, we have

$$\partial h^*(A) = \{\eta : \nabla h(\eta) = A\},$$

which tells us that the subdifferential ∂h^* consists precisely of those vectors η that induce a weighted Voronoi partition whose cells have masses equal to A. It therefore follows that there exists some $\bar{\eta} \in \partial h^*\left(\bar{A}\right)$ such that

$$\begin{aligned} 0 &= \nabla G\left(\bar{A}\right) + \bar{\eta} - \zeta, \\ \zeta_k \bar{A}_k &= 0, \qquad k = 1, ..., K-1, \\ \zeta_k &\geq 0, \qquad k = 1, ..., K-1, \end{aligned}$$

or, equivalently,

$$\begin{aligned} \frac{\partial G\left(\bar{A}\right)}{\partial A_k} &= -\bar{\eta}_k^* \quad \Leftrightarrow \quad \bar{A}_k > 0, \\ \frac{\partial G\left(\bar{A}\right)}{\partial A_k} &\geq -\bar{\eta}_k^* \quad \Leftrightarrow \quad \bar{A}_k = 0. \end{aligned}$$

Let $\bar{\mathcal{A}}$ be the partition corresponding to the vector $\bar{A}$, and consider $x \in \bar{\mathcal{A}}_k$. If $\bar{A}_j > 0$, we have

$$g_k(x) + \frac{\partial G\left(\bar{A}\right)}{\partial A_k} = g_k - \bar{\eta}_k \leq g_j(x) - \bar{\eta}_j = g_j(x) + \frac{\partial G\left(\bar{A}\right)}{\partial A_j},$$

and, if $\bar{A}_j = 0$, we have

$$g_k(x) + \frac{\partial G\left(\bar{A}\right)}{\partial A_k} = g_k - \bar{\eta}_k \leq g_j(x) - \bar{\eta}_j \leq g_j(x) + \frac{\partial G\left(\bar{A}\right)}{\partial A_j},$$

which completes the proof.

8.2. Proof of Lemma 6

As in Section 4 above, the quantities A_k are fixed. We rewrite problem (28)-(29) equivalently as the infinite-dimensional integer program

$$\min_{J_k(\cdot)} \sum_k \iint_{\mathcal{S}} J_k(x)\, g_k(x)\, d\nu, \tag{36}$$

subject to the constraints

$$\sum_{k=1}^{K} J_k(x) = 1, \qquad \forall x \in \mathcal{S}, \tag{37}$$

$$\iint_{\mathcal{S}} J_k(x)\, d\nu = A_k, \qquad k = 1, ..., K, \tag{38}$$

$$J_k(x) \in \{0, 1\}, \qquad \forall x \in \mathcal{S}, k = 1, ..., K. \tag{39}$$

We shall make use of this formulation at the end of the proof. For the moment, however, let us relax the problem by dropping the binary constraint (39). Then, problem (36)-(38) is an instance of semidiscrete optimal transport (Hartmann and Schuhmacher 2020), and therefore has the same optimal value as the Kantorovich dual

$$\max_{\eta\in\mathbb{R}^K} \iint_{\mathcal{S}} \left(\min_k g_k(x) - \eta_k\right) d\nu + \sum_{k=1}^{K} A_k \eta_k, \tag{40}$$

which can be straightforwardly derived using Thm. 1.3 of Villani (2021). Because the cost functions g_k are continuous, both the primal and dual have optimal solutions (Thm. 1.3 and Exercise 2.36 of Villani 2021).

We now construct a partition whose objective value matches that of (40) while additionally satisfying (39). Letting η' be the optimal dual solution, define a partition $\mathcal{B}_1, ..., \mathcal{B}_K$ by setting

$$\mathcal{B}_k = \left\{ x \in \mathcal{S} : g_k(x) - \eta'_k \leq \min_{j\neq k} g_j(x) - \eta'_j \right\}. \tag{41}$$

The regularity conditions on the cost functions g_k ensure that $\nu(\mathcal{B}_j \cap \mathcal{B}_k) = 0$ for all $j \neq k$. If we let $J_k(x) = 1_{\{x\in\mathcal{B}_k\}}$ and plug this into (36), the objective value is

$$\begin{aligned}
\sum_k \iint_{\mathcal{S}} J_k(x)\, g_k(x)\, d\nu &= \sum_k \iint_{\mathcal{B}_k} g_k(x)\, d\nu \\
&= \sum_k \iint_{\mathcal{B}_k} (g_k(x) - \eta'_k + \eta'_k)\, d\nu \\
&= \sum_k \iint_{\mathcal{B}_k} (g_k(x) - \eta'_k)\, d\nu + \sum_k \iint_{\mathcal{B}_k} \eta'_k\, d\nu \\
&= \iint_{\mathcal{S}} \min_k (g_k(x) - \eta'_k)\, d\nu + \sum_k \eta_k \nu(\mathcal{B}_k),
\end{aligned} \tag{42}$$

where (42) follows from (41). We see that (42) differs from the dual objective in (40) only in that we have $\nu(\mathcal{B}_k)$ where we wish to see A_k. Thus, it remains to show that $\nu(\mathcal{B}_k) = A_k$. This equality, however, is precisely the first-order optimality condition of (40), derived by differentiating the dual objective with respect to η.

References

Boyd, S., L. Vandenberghe 2004. *Convex optimization*. Cambridge University Press.

Hartmann, V., D. Schuhmacher. 2020. Semi-discrete optimal transport: a solution procedure for the unsquared Euclidean distance case. *Mathematical Methods of Operations Research* **92**(1) 133–163.

Hiriart-Urruty, J.-B., C. Lemaréchal. 2004. *Fundamentals of convex analysis*. Springer Science & Business Media.

Villani, C. 2021. *Topics in optimal transportation*. American Mathematical Society.