

Household-level economies of scale in transportation

John Gunnar Carlsson*, Mehdi Behroozi†, Xiangfei Meng‡ and Raghuveer Devulapalli§

October 16, 2015

Abstract

The efficiency of centralized delivery services over a backbone network in comparison to that of localized or multiple individual trips from the perspective of the overall carbon footprint is a fundamental subject of the analysis of logistical systems. In this paper, we quantify the changes in overall efficiency due to such services by looking at “household-level” economies of scale in transportation: a person might perform many errands in a day (such as going to the bank, grocery store, and post office), and that person has many choices of locations at which to perform these tasks (e.g., a typical metropolitan region has many banks, grocery stores, and post offices). Thus, the total driving distance (and therefore the overall carbon footprint) that that person traverses is the solution to a *Generalized Travelling Salesman Problem* (GTSP) in which he or she selects both the best locations to visit and the sequence in which to visit them. We perform a probabilistic analysis of the GTSP under the assumption that all relevant locations are independently and identically distributed uniformly in a region and then determine the amount of adoption of centralized delivery services that is necessary, under our model, in order for the overall carbon footprint of the region to decrease.

1 Introduction

One of the fundamental concerns in the analysis of logistical systems is the trade-off between localized, independent provision of goods and services versus provision along a centralized infrastructure such as a backbone network. On the one hand, service executed at a local level features the obvious benefits of proximity and specialization, inasmuch as people and communities obtain things from locations that are close to them. Conversely, by aggregating network

*Epstein Department of Industrial and Systems Engineering, University of Southern California. J.G.C. gratefully acknowledges the support of the Boeing Company, DARPA YFA grant N66001-12-1-4218, NSF award CMMI-1234585, and ONR grant N000141210719.

†Department of Industrial and Systems Engineering, University of Minnesota.

‡Epstein Department of Industrial and Systems Engineering, University of Southern California.

§Department of Industrial and Systems Engineering, University of Minnesota.

flows via a backbone network, individuals and communities are able to reap the benefits of economies of scale, economies of agglomeration, and economies of density.

One phenomenon in which this trade-off has recently been made manifest is the transition of businesses from traditional brick-and-mortar stores to retail sales facilitated via e-commerce [6, 36, 39], such as grocery delivery services. Particular recent examples include Google Shopping Express, Amazon Prime, Instacart, and Walmart To Go [10, 13, 24, 43], among many others. As discussed in [46] for the case of grocery delivery,

There’s still a lot of debate over what works and what doesn’t. Is it a good idea to have a warehouse for food storage, or ask the customers to pick up their food? How much should delivery cost? How often and where should online grocery companies deliver?

In the past, the costs associated with delivery service have been so big – huge warehouses and refrigerators, gas-guzzling trucks traveling door to door ... that the math has never worked out[.]

A major complication in problems of this kind is the difficulty of creating a model that is mathematically tractable enough to give useful insights as well as faithful to the original phenomenon being modelled. The specific topic that we emphasize in this paper is the analysis of *multi-stop trips* made by households: on a given day, a person will often visit multiple locations on one outing (such as running errands on the way to or from one’s place of work), and each of these locations will usually have *alternatives* (e.g. there are usually multiple choices of which grocer or post office to use). Thus, the calculation of the “cost” of a multi-stop trip is more complicated than a mere travelling salesman tour or a sequence of direct trips to and from the various destinations and the household. For example, one might be willing to travel a long distance to visit a bank that is farther away than the nearest available branch if it is located more closely to other businesses that they will also visit (say, by virtue of being located in a central business district or shopping center).

In the current literature, this complication is handled by either simplifying the problem at hand or by introducing additional assumptions into the problem structure. For example, in the recent paper [51], the authors perform a detailed computational study that estimates the changes in net CO₂ emissions that result by introducing grocery delivery services in Seattle, Washington. As suggested in Figure 1 of that paper (which we reproduce here in Figure 1 as well), the authors do not consider the possibility of multi-stop trips and compare the cost of a direct trip between one’s house and back to the marginal cost incurred by adding oneself to a travelling salesman tour that services many households. As another example, the paper [37] acknowledges the importance of multi-stop trips (which they call “trip chaining”) in calculating carbon footprints, and assumes constant values for trip lengths, such as 12.8 miles for an average shopping trip by car and a constant number of stops:

Shopping can be part of a wider combined trip and involve only a minor detour. We assume that

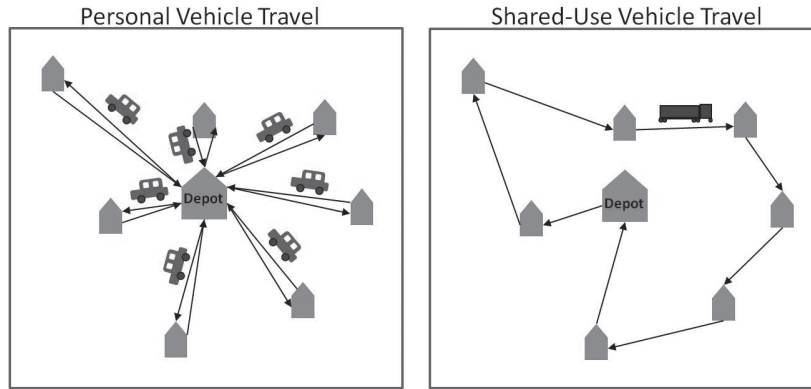


Figure 1: The above image is reproduced from Figure 1 of [51], which compares the cost of direct trips originating from a central location (at left) with travelling salesman tours that visit multiple locations (right).

where a shopper undertakes trip chaining, the shopping component of the trip makes up a quarter of the overall total mileage.

A third example can be found in [50], which considers a closely related problem in which a customer receiving a package can specify *multiple* locations at which the delivery service may drop off the package (e.g. “please drop off my package at my home, my work, my gym, or my friend’s house”). The analysis therein is based on Monte Carlo simulation and is highly sensitive to problem specifics, and the issue of trip chaining at the household level is addressed by assigning a fixed amount of trip chaining to estimate marginal costs (the authors also cite [12], which makes a similar assumption):

Generally, social network members will not participate or choose the burden of pickup if they have to go to a pickup point solely for the purpose of making a pickup for another person. Pickup trips for social network actors can be regarded as a chain event and is a determining variable. We assumed a 100% trip chain to additional mileage for pickup in both PLS and SPLS – in other words, the entire detour distance for pickup is attributed to the package. By contrast, previous research has applied a 0% trip chain effect for pickup. [12]

The objective of this paper is to apply tools from geometric probability theory to derive a more nuanced model of the costs of multi-stop trips taken by a household as a function of the number of destinations that one must visit and the number of alternatives of each of these that is available. We then apply this analysis to predict the change in carbon footprint as households in a geographic region make increased use of delivery services versus trips to brick-and-mortar stores.

Related work

This paper is primarily concerned with continuous approximation models for transportation problems and is therefore in the company of [14, 17, 28, 29, 41, 42], for example. We are particularly interested in the *asymptotic* behavior of a certain Euclidean optimization problem, the *generalized travelling salesman problem* (generalized TSP), and as such our analysis is closely related to such papers as [8, 11, 23, 44, 48]. One particularly related result can be found in [5], which studies the asymptotic behavior of the “TSP on sparse subsets”, which is closely related to our generalized TSP.

The generalized TSP is a reasonably well-studied combinatorial problem and has been studied for almost 50 years [47]. The primary focus of study on this problem has been on the rapid solution of particular problem instances in a combinatorial setting [9, 34, 40, 45, 52], whereas our work is focused more on limiting behavior. For our purposes, the generalized TSP arises organically as a way of studying the impact of multi-stop trips, or “trip chaining”. The seminal paper [3] studies this phenomenon from a theoretical perspective by describing a particular utility function at the household level that is justified with empirical travel data. A subsequent paper, [31], develops a model of destination choice that employs the “prospective utility” of a destination zone as its “attraction measure”. The paper [25] builds a variety of logit models in order to study the potential barrier that trip chaining creates in attracting car users to switch to public transport.

The intent of our work is to better understand the trade-off between centralized and decentralized distribution schemes. Our work could be said to follow naturally from [37, 50, 51], for example, and would also fit in the company of [15, 16, 18], all of which study this same trade-off in one form or another.

Notational conventions

In this paper we will consider a planar region \mathcal{R} with area A , which is usually assumed without loss of generality to be the unit square. We assume that \mathcal{R} has a population of N people, who perform n tasks each day, and there are k locations at which each task can be performed. The cost per mile of delivery trucks along the backbone network is given by ϕ and the cost per mile of passenger vehicles is given by ψ . We will also make use of some standard conventions in asymptotic analysis:

- We say that $f(x) \in \mathcal{O}(g(x))$ if there exists a constant $c > 0$ and a value x_0 such that $f(x) \leq c \cdot g(x)$ for all $x \geq x_0$,
- We say that $f(x) \in \mathcal{\Omega}(g(x))$ if there exists a constant $c > 0$ and a value x_0 such that $f(x) \geq c \cdot g(x)$ for all $x \geq x_0$,

- We say that $f(x) \in \Theta(g(x))$ if there exists constants $c_1 > 0$ and $c_2 > 0$ and a value x_0 such that $c_1 \cdot g(x) \leq f(x) \leq c_2 \cdot g(x)$ for all $x \geq x_0$, and
- We say that $f(x) \sim g(x)$ if $\lim_{x \rightarrow \infty} f(x)/g(x) = 1$.

When necessary, we will clarify this notation in some particular cases because we are interested in limiting behavior that concerns two values, the number of locations n and the number of choices k , and such notation is known to introduce complicated ambiguities [19, 27].

2 Summary of key facts and findings from related work

This brief section introduces two previous geometric results that are closely related to the problem that we will study in the following section.

Lemma 1. *For any set of n' points $x_1, \dots, x_{n'}$ contained in a square of area A' , the length of the optimal TSP tour through $x_1, \dots, x_{n'}$, denoted $\text{TSP}(x_1, \dots, x_{n'})$, satisfies $\text{TSP}(x_1, \dots, x_{n'}) \leq \sqrt{2A'n'} + 7/4\sqrt{A'} < \alpha_1\sqrt{A'n'}$, where $\alpha_1 = 2.7$.*

Proof. See [21, 30]. □

This result can be stated more strongly in a probabilistic fashion as in the celebrated *Beardwood-Halton-Hammersley (BHH) Theorem*:

Theorem 2. *Suppose that $\{X_1, X_2, \dots\}$ is a sequence of random points i.i.d. according to the uniform distribution defined on a compact planar region \mathcal{R} with area A . Then with probability one,*

$$\lim_{N \rightarrow \infty} \frac{\text{TSP}(X_1, \dots, X_N)}{\sqrt{AN}} = \beta,$$

where β is a constant.

Proof. See for example [8, 44, 49]. □

It is additionally known that $0.6250 \leq \beta \leq 0.9204$ and conjectured that $\beta \approx 0.7124$; see [7, 22].

3 Analysis of the generalized TSP tour

The main combinatorial object that we will use to model household trips is the *generalized TSP tour*, defined as follows:

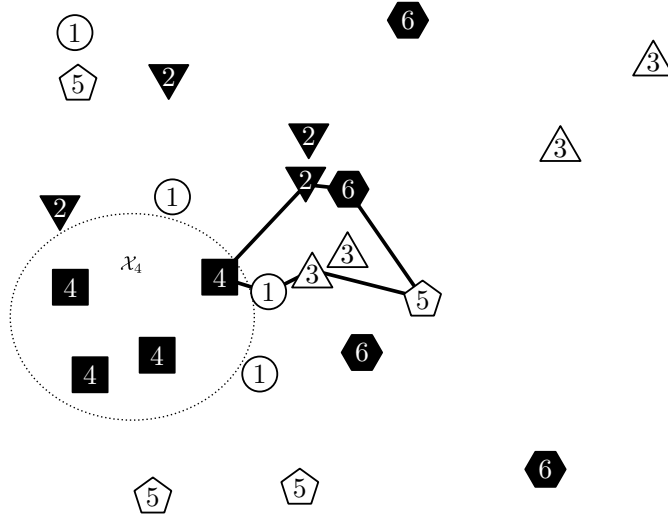


Figure 2: A generalized TSP tour of six sets of points $\mathcal{X}_1, \dots, \mathcal{X}_6$, with \mathcal{X}_4 outlined for purposes of clarity. The optimal tour contains one element from each such set (and is the shortest such tour to do so).

Definition 3. Given n sets of points $\mathcal{X}_1, \dots, \mathcal{X}_n$ in the plane, the *generalized TSP tour* $\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n)$ is the shortest cycle that contains one element from each point set \mathcal{X}_i .

See Figure 2 for an example. Clearly, when each point set is a singleton (i.e. $\mathcal{X}_i = \{x_i\}$ for all i), we see that $\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) = \text{TSP}(x_1, \dots, x_n)$. We will commit a minor abuse of notation throughout this paper by using the term $\text{GTSP}(\cdot)$ to refer both to the tour itself and to its length.

The GTSP is, of course, a generalization of the TSP, which has been analyzed extensively from a geometric probabilistic perspective [44, 48]. However, many of these results for the TSP cannot be generalized in a straightforward way to the GTSP. To give one example, page 30 of [49] establishes a simple nearest-neighbor argument that explains why $\mathbf{E} \text{TSP}(X_1, \dots, X_n) \in \Omega(\sqrt{n})$ for independent, uniformly distributed X_i in the unit square: for any point X_i , it can be shown that $\mathbf{E} \min_{j:j \neq i} \|X_i - X_j\| \in \Omega(1/\sqrt{n})$, from which the desired result follows by summing over all n points. This does not carry over to the GTSP because we have $k_1 + k_2 + \dots + k_n$ points in total and we are summing over only n of them.

It is also worth noting that there are two kinds of limits of interest to us, namely the case where the number of point sets gets large, i.e. $n \rightarrow \infty$, and the case where the cardinality of these point sets gets large, i.e. average of k_i 's goes to infinity, and the latter does not appear to have much connection to previous works on the TSP.

3.1 Analysis of the case $n \rightarrow \infty$

The following theorem describes the behavior of the GTSP when we let $n \rightarrow \infty$. In order to focus on the asymptotic behavior in terms of n we fix $|\mathcal{X}_i| = k$ for all i .

Theorem 4. Let $\mathcal{X}_1, \dots, \mathcal{X}_n$ denote n sets of points, each having cardinality k , and suppose that all nk points are distributed independently and uniformly at random in a region \mathcal{R} having area A . Then the expected length of a generalized TSP tour of $\mathcal{X}_1, \dots, \mathcal{X}_n$ satisfies

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \in \Theta(\sqrt{An/k})$$

as $n \rightarrow \infty$ with k fixed. In particular, there exist constants $\alpha_1 < 2.7$ and $\alpha_2 > 0.0681$ such that, for any k , there exists a threshold \bar{n} such that $\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{An/k}$ and $\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \geq \alpha_2 \sqrt{An/k}$ whenever $n \geq \bar{n}$.

The upper bound that $\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{An/k}$ is a fairly straightforward generalization of the result from [21] and we prove it presently; we will prove the lower bound that $\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \geq \alpha_2 \sqrt{An/k}$ afterwards because we will require an additional combinatorial lemma.

Proof of the upper bound. Assume without loss of generality that $A = 1$, and make an additional further assumption that \mathcal{R} is the unit square (generalizing the argument to arbitrary regions \mathcal{R} is a routine task that we omit for brevity): let m be an even integer and consider the path \mathcal{P} obtained by traversing the width of \mathcal{R} horizontally a total of m times, starting at the upper leftmost corner of \mathcal{R} and moving downward by an amount $1/m-1$, as shown in Figure 3a; it is obvious that the length of \mathcal{P} is simply $m + 2$. Given a collection of point sets $\mathcal{X}_1, \dots, \mathcal{X}_n$ in \mathcal{R} , we can perturb \mathcal{P} to form a new path \mathcal{P}' that visits one point from each point set \mathcal{X}_i by simply inserting a pair of vertical line segments between \mathcal{P} and the nearest point (measured only in the vertical direction) in each \mathcal{X}_i , as shown in Figure 3a. Of course, the vertical distance between \mathcal{P} and any arbitrary point $x \in \mathcal{X}_i$ simply follows a uniform distribution between 0 and $1/2(m-1)$, and therefore we see that the vertical distance d between \mathcal{P} and its nearest neighbor in \mathcal{X}_i (again, in the vertical sense) satisfies

$$\mathbf{E}(d) = \frac{1/2(m-1)}{k+1}.$$

Our objective is therefore to select a value m so as to minimize the total length of \mathcal{P} plus these additional n displacements, i.e. to minimize

$$m + 2 + 2n \cdot \frac{1/2(m-1)}{k+1}$$

where we have an additional multiplier “2” since each vertical displacement consists of an outbound and an inbound

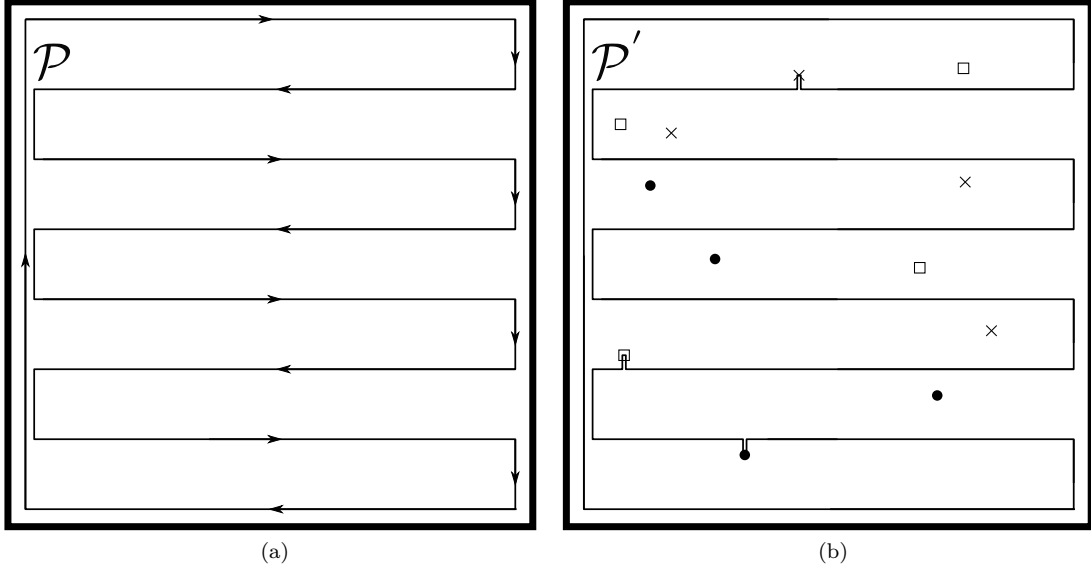


Figure 3: In (3a), we show the path \mathcal{P} , which traverses the region \mathcal{R} horizontally a total of $m = 8$ times. In (3b), we show the perturbed path \mathcal{P}' , where we have $n = 3$ point sets \mathcal{X}_i consisting of $k = 4$ points each.

trip. As n becomes large, we see that the optimal m satisfies

$$m^* \sim \sqrt{\frac{n}{k+1}},$$

which results in a total length that satisfies

$$\text{length}(\mathcal{P}') \sim \frac{2\sqrt{n}}{\sqrt{k+1}} < 2\sqrt{\frac{n}{k}}$$

which proves that $\mathbf{EGTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{An/k}$.

□

To prove the lower bound that $\mathbf{EGTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \geq \alpha_2 \sqrt{An/k}$, we must first introduce a combinatorial lemma:

Lemma 5. *Let $\mathcal{L} \subset \mathbb{Z}^2$ denote an $m \times m$ square integer lattice in the plane, let $n \geq 2$ be an integer, and let $\ell > 0$. Let \mathcal{P} denote the set of all paths of the form $\{x_1, \dots, x_n\}$, with $x_i \in \mathcal{L}$ for each i , and whose length does not exceed ℓ . Then*

$$|\mathcal{P}| \leq m^2 \cdot \binom{\ell+n-1}{n-1} \cdot \left(\frac{8\ell}{n-1}\right)^{n-1}.$$

Proof. We thank Douglas Zare. Note that we have allowed “replacement” in the construction of \mathcal{P} , i.e. we are

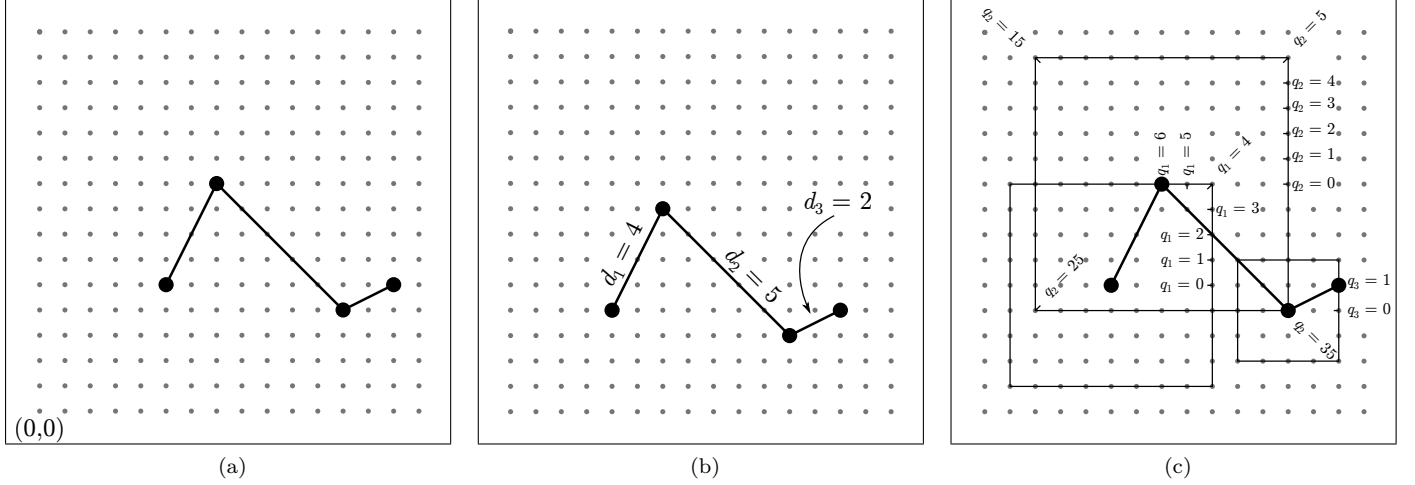


Figure 4: The above path has a length of approximately 13.8, and therefore is a member of \mathcal{P} if we have ℓ equal to (say) 15. We will construct its triplet $(x, \mathbf{d}, \mathbf{q})$ as follows: obviously, we have $x = (5, 5)$ (the first element of the path), as indicated in 4a. Figure 4b shows that the ℓ_∞ distances between the four points are $d_1 = 4$, $d_2 = 5$, and $d_3 = 2$, respectively, which would then imply that $d_4 = \ell - (d_1 + d_2 + d_3) = 4$. Finally, the construction of \mathbf{q} is shown in 4c: given a point x_i on the path and a distance d_i , there are $8d_i$ possible places where the consecutive point x_{i+1} could be located. The path shown has $q_1 = 6$, $q_2 = 35$, and $q_3 = 1$.

also considering paths in which $x_i = x_j$ for some (i, j) pairs. We note that any element $\mathcal{P} \in \mathcal{P}$ can be uniquely described by specifying the triplet $(x, \mathbf{d}, \mathbf{q})$ defined as follows:

- The point $x \in \mathcal{L}$ is simply the first member of \mathcal{P} .
- The n -tuple $\mathbf{d} = \{d_1, \dots, d_n\}$ represents the *distance* travelled from each point to the next, measured in the ℓ_∞ norm. In other words, for a path $\{x_1, \dots, x_n\}$, we have $d_i = \|x_{i+1} - x_i\|_\infty$. Note that d_n is not defined according to this definition; we therefore define $d_n := \ell - \sum_{i=1}^{n-1} d_i$, so that $\sum_{i=1}^n d_i = \ell$ for all valid \mathbf{d} . Obviously, we have $0 \leq d_i \leq \ell$ for all i .
- The $(n - 1)$ -tuple $\mathbf{q} = \{q_1, \dots, q_{n-1}\}$ represents the “angles” between pairs of points. Specifically, given a point x_i and a corresponding distance d_i , we see that there are at most $8d_i$ possible places where x_{i+1} could be located (since x_{i+1} must lie on the boundary of a square of side length $2d_i$ centered at x_i). The element q_i specifies which of these is the correct location of x_{i+1} .

An example of this is shown in Figure 4. We will bound $|\mathcal{P}|$ from above by looking at the set of all triplets $(x, \mathbf{d}, \mathbf{q})$ such that $x \in \mathcal{L}$, $\sum_{i=1}^n d_i = \ell$, and $q_i \leq 8d_i$ for all i . Assume without loss of generality that ℓ is an integer, and let \mathcal{D} denote the set of all permissible n -tuples \mathbf{d} . By construction, of course, \mathcal{D} is simply the set of all integer n -tuples $\{d_1, \dots, d_n\}$ such that $d_i \geq 0$ and $\sum_{i=1}^n d_i = \ell$. Given any $\mathbf{d} \in \mathcal{D}$, let \mathbf{d}' denote the n -tuple defined by setting

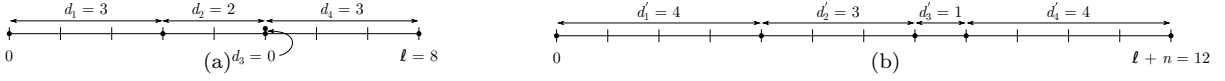


Figure 5: The figures above correspond to the case where $n = 4$ and $\ell = 8$. As 5a shows, any n -tuple \mathbf{d} can be uniquely represented by simply placing n points on the number line from 0 to ℓ , where the first point is placed a distance d_1 from the origin and each subsequent point i is placed a distance d_i to the right of its predecessor. Since d_4 is defined so that the entries of \mathbf{d} sum to ℓ , the last such point that is placed must be precisely at a distance ℓ from the origin. The placement shown corresponds to the case $\mathbf{d} = (3, 2, 0, 3)$. In 5b, we show the same point placement for the n -tuple \mathbf{d}' which is obtained by adding 1 to each of the entries of \mathbf{d} . It is obvious that any such n -tuple \mathbf{d}' can be uniquely constructed by selecting $n - 1 = 3$ points from the $\ell + n - 1 = 11$ valid locations of points in 5b and computing the sequential distances between those points; the diagram corresponds to the case $\mathbf{d}' = (4, 3, 1, 4)$.

$d'_i = d_i + 1$ for all i . We then see that $\sum_{i=1}^n d'_i = \ell + n$ and that $1 \leq d'_i \leq \ell + n$ for all i . As Figure 5 suggests, each \mathbf{d}' corresponds to a selection of $n - 1$ elements out of $\ell + n - 1$ possibilities. Thus, we see that

$$|\mathcal{D}| \leq \binom{\ell + n - 1}{n - 1}.$$

It is simpler to bound the set \mathcal{D} of all valid $(n - 1)$ -tuples \mathbf{q} . For any fixed \mathbf{d} , the number of possible choices of \mathbf{q} is at most $8^{n-1} \prod_{i=1}^{n-1} d_i$. By the AM-GM inequality, using the fact that $\sum_{i=1}^{n-1} d_i \leq \ell$, we see that

$$8^{n-1} \prod_{i=1}^{n-1} d_i \leq 8^{n-1} \left(\sum_{i=1}^{n-1} \frac{d_i}{n-1} \right)^{n-1} = \left(\frac{8\ell}{n-1} \right)^{n-1}$$

and thus

$$|\mathcal{D}| \leq \left(\frac{8\ell}{n-1} \right)^{n-1}.$$

Finally, since there are m^2 choices of the initial point x in the triplet $(x_1, \mathbf{d}, \mathbf{q})$, we conclude that \mathcal{P} satisfies

$$|\mathcal{P}| \leq m^2 \cdot \binom{\ell + n - 1}{n - 1} \cdot \left(\frac{8\ell}{n - 1} \right)^{n-1}$$

as desired. □

We are now ready to prove the lower bound of Theorem 4:

Proof of the lower bound of Theorem 4. We will again assume that the service region \mathcal{R} is the unit square, and we will then let \mathcal{L} denote a lattice within \mathcal{R} of the form

$$\left\{ 0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1 \right\} \times \left\{ 0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1 \right\}.$$

It is immediately clear that for any path $\mathcal{P} := \{x_1, \dots, x_n\} \subset \mathcal{R}$, there exists a path in \mathcal{L} obtained by rounding the terms of \mathcal{P} to their nearest neighbor in \mathcal{L} whose length is within n/m of the original path. Thus, it will suffice to prove the lower bound of Theorem 4 for the special case where all of the elements of the sets \mathcal{X}_i lie in elements of \mathcal{L} , and study the limiting behavior as $m \rightarrow \infty$.

We first see that there are at most k^n total distinct subsets of the form $\{x_1, \dots, x_n\}$, with $x_i \in \mathcal{X}_i$ for each i . For each of these subsets, there are $n!$ different orderings, and we therefore conclude that there are at most $n! \cdot k^n$ valid paths that visit one point from each of the subsets \mathcal{X}_i . By applying the union bound, we see that

$$\begin{aligned} \Pr(\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_2 \sqrt{n/k}) &= \Pr(\text{one of the } n! \cdot k^n \text{ valid paths has length } \leq \alpha_2 \sqrt{n/k}) \\ &\leq n! \cdot k^n \Pr(\text{length}(\mathcal{P}) \leq \alpha_2 \sqrt{n/k}) \end{aligned}$$

where \mathcal{P} is the path obtained by selecting the first element from each set \mathcal{X}_i and visiting these elements in a sequence chosen uniformly at random. By construction, we see that \mathcal{P} is simply a path that is sampled uniformly at random from the collection of *all* possible paths between n points taken from \mathcal{L} . Of course, we can see immediately that there are exactly m^{2n} of these. By scaling the lattice of Lemma 5 by a factor of $1/m$ (i.e. our lattice \mathcal{L}), we see that there are at most

$$m^2 \cdot \binom{m\alpha_2 \sqrt{n/k} + n - 1}{n - 1} \cdot \left(\frac{8m\alpha_2 \sqrt{n/k}}{n - 1} \right)^{n-1} \leq m^2 \cdot \frac{(m\alpha_2 \sqrt{n/k} + n - 1)^{n-1}}{(n - 1)!} \cdot \left(\frac{8m\alpha_2 \sqrt{n/k}}{n - 1} \right)^{n-1}$$

paths of length $\alpha_2 \sqrt{n/k}$ through our lattice \mathcal{L} and therefore

$$\begin{aligned} \limsup_{m \rightarrow \infty} \Pr(\text{length}(\mathcal{P}) \leq \alpha_2 \sqrt{n/k}) &\leq \limsup_{m \rightarrow \infty} \frac{m^2 \cdot \frac{(m\alpha_2 \sqrt{n/k} + n - 1)^{n-1}}{(n - 1)!} \cdot \left(\frac{8m\alpha_2 \sqrt{n/k}}{n - 1} \right)^{n-1}}{m^{2n}} \\ &= (8\alpha_2^2)^{n-1} \cdot \frac{n^{n-1}}{(n - 1)! \cdot (n - 1)^{n-1} \cdot k^{n-1}} \\ \implies \Pr(\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_2 \sqrt{n/k}) &\leq n! \cdot k^n \cdot (8\alpha_2^2)^{n-1} \cdot \frac{n^{n-1}}{(n - 1)! \cdot (n - 1)^{n-1} \cdot k^{n-1}} \\ &= k \cdot (8\alpha_2^2)^{n-1} \cdot \frac{n^n}{(n - 1)^{n-1}}. \end{aligned}$$

The above quantity approaches 0 as $n \rightarrow \infty$ because $8\alpha_2^2 < 1$. Thus,

$$\begin{aligned} \mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) &\geq \left(1 - k \cdot (8\alpha_2^2)^{n-1} \cdot \frac{n^n}{(n-1)^{n-1}}\right) \alpha_2 \sqrt{n/k} \\ &= \alpha_2 \sqrt{n/k} - \sqrt{k} \cdot 8^{n-1} \alpha_2^{2n-1} \cdot \frac{n^{n+1/2}}{(n-1)^{n-1}} \\ &\sim \alpha_2 \sqrt{n/k} \text{ as } n \rightarrow \infty \end{aligned}$$

which proves the desired lower bound as desired. \square

The proof above completes our analysis of the GTSP for the case where the number of point sets, n , becomes large, and the cardinalities $|\mathcal{X}_i|$ are all fixed and equal. The next section describes the case where n is fixed and the cardinalities $|\mathcal{X}_i|$ become large.

3.2 Analysis of the case $|\mathcal{X}_i| \rightarrow \infty$

This section studies the limiting behavior of the GTSP when we assume that the number of sets, n , is fixed, and the cardinalities of each set \mathcal{X}_i become large. In order to describe these n cardinalities in terms of a single parameter, we define the geometric mean of the k_i 's as $k_G = (\prod_{i=1}^n k_i)^{1/n} = t(\prod_{i=1}^n q_i)^{1/n}$, where we assume that $|\mathcal{X}_i| = k_i = tq_i$, with \mathbf{q} being a vector of probabilities, and we let the single parameter t approach infinity.

Theorem 6. *Let $\mathcal{X}_1, \dots, \mathcal{X}_n$ denote n sets of points, each having cardinality $k_i > 0$, and suppose that all $\sum_{i=1}^n k_i$ points are distributed independently and uniformly at random in a region \mathcal{R} having area A . Further assume that $k_i = tq_i$ for all i (where \mathbf{q} is a probability vector), and let $k_G = (\prod_{i=1}^n k_i)^{1/n} = t(\prod_{i=1}^n q_i)^{1/n}$ be the geometric mean of the k_i 's. Then the expected length of a generalized TSP tour of $\mathcal{X}_1, \dots, \mathcal{X}_n$ satisfies*

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \in \mathcal{O} \left(\sqrt{\frac{An}{k_G}} \right)$$

and

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \in \Omega \left(\sqrt{\frac{An}{k_G^{n/(n-1)}}} \right)$$

as $t \rightarrow \infty$ with n and \mathbf{q} fixed. In particular, there exist constants $\alpha_1 < 2.7$ and $\alpha_2 > 0.0681$ such that, the following statements hold:

1. For any n , there exists a threshold \bar{k} (or equivalently \bar{t}) such that

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{\frac{An}{k_G}}$$

whenever $k_G \geq \bar{k}$ (or equivalently $t \geq \bar{t}$).

2. For any $n \geq 2$, there exists a threshold \bar{k} (or equivalently \bar{t}) such that

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \geq \alpha_2 \sqrt{\frac{An}{k_G^{n/(n-1)}}}$$

whenever $k_G \geq \bar{k}$ (or equivalently $t \geq \bar{t}$).

In addition, the upper bound in the first statement can be tightened as follows:

1a. For any n , there exists a threshold \bar{k} (or equivalently \bar{t}) such that

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{\frac{An}{k_G^{n/(n-1)}}} \cdot (n^2 \log k_G + \log n)^{\frac{1}{2(n-1)}}$$

whenever $k_G \geq \bar{k}$ (or equivalently $t \geq \bar{t}$).

Proof. As in the proof of Theorem 4, we will assume that the service region \mathcal{R} is the unit square. The proof of the upper bound, Claim 1, proceeds as follows: since we are examining the limiting behavior of $\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n)$ for large k_G , we can divide the region \mathcal{R} into k_G squares $\square_1, \dots, \square_{k_G}$ of area $1/k_G$. By Lemma 1, we see that if one of the k_G squares \square_i happens to contain an element from each of the n point sets $\mathcal{X}_1, \dots, \mathcal{X}_n$, then $\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{n/k_G}$. If none of the squares have this property, then as a crude upper bound we simply use $\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{n}$, so that

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq p \alpha_1 \sqrt{n/k_G} + (1-p) \alpha_1 \sqrt{n},$$

where p is the probability that one of the k_G squares contains an element from each \mathcal{X}_i . Thus, our goal is now to show that $p \rightarrow 1$ at a sufficiently rapid rate as $k_G \rightarrow \infty$.

Our proof now requires a *Poissonization* argument [49]: for each of the k_G boxes \square_i , let Y_1^i, \dots, Y_n^i denote n point sets uniformly and independently distributed within \square_i , where $|Y_j^i|$ follows a Poisson distribution with mean $\frac{k_j}{k_G} = \frac{q_j}{(\prod_{i=1}^n q_i)^{1/n}}$. If we define point sets $\mathcal{Y}_1, \dots, \mathcal{Y}_n$ by setting

$$\mathcal{Y}_j = \bigcup_{i=1}^{k_G} Y_j^i,$$

then it is immediately obvious that $\mathbf{E}(|\mathcal{Y}_j|) = k_j$ for all j ; it is also easy to verify that the distribution of the point sets $\mathcal{X}_1, \dots, \mathcal{X}_n$ is the same as the distribution of the point sets $\mathcal{Y}_1, \dots, \mathcal{Y}_n$, conditioned on the event that $|\mathcal{Y}_j| = k_j$ for all j (see page 100 of [38], for example). For a particular box \square_i , the probability that \square_i contains at least one

element from each of the sets \mathcal{Y}_j is

$$\begin{aligned} \Pr(\square_i \text{ contains at least one from each } \mathcal{Y}_j) &= \Pr(|Y_1^i| \geq 1) \cdots \Pr(|Y_n^i| \geq 1) \\ &= \prod_{j=1}^n (1 - e^{-k_j/k_G}) \end{aligned}$$

and therefore,

$$\Pr(\underbrace{\text{none of the boxes } \square_i \text{ contains at least one from each } \mathcal{Y}_j}_{=:E}) = \left[1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) \right]^{k_G}.$$

By the law of total probability, we have

$$\Pr(E) = \sum_{k'_1=0}^{\infty} \cdots \sum_{k'_n=0}^{\infty} \Pr(E \mid |\mathcal{Y}_1| = k'_1 \cap \cdots \cap |\mathcal{Y}_n| = k'_n) \Pr(|\mathcal{Y}_1| = k'_1 \cap \cdots \cap |\mathcal{Y}_n| = k'_n)$$

and therefore in particular,

$$\Pr(E) \geq \Pr(E \mid |\mathcal{Y}_1| = k_1 \cap \cdots \cap |\mathcal{Y}_n| = k_n) \Pr(|\mathcal{Y}_1| = k_1 \cap \cdots \cap |\mathcal{Y}_n| = k_n).$$

We next observe that

$$\begin{aligned} \Pr(|\mathcal{Y}_1| = k_1 \cap \cdots \cap |\mathcal{Y}_n| = k_n) &= \prod_{j=1}^n \Pr(|\mathcal{Y}_j| = k_j) \\ &= \prod_{j=1}^n \frac{k_j^{k_j}}{k_j!} e^{-k_j} \\ &> \frac{1}{e^n \sqrt{k_1 \cdots k_n}}, \end{aligned}$$

where the last inequality holds because $k'! < e\sqrt{k'}(k'/e)^{k'}$ for all positive integers k' (see Lemma 5.8 of [38]). We therefore conclude that

$$\left[1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) \right]^{k_G} = \Pr(E) > \underbrace{\Pr(E \mid |\mathcal{Y}_1| = k_1 \cap \cdots \cap |\mathcal{Y}_n| = k_n)}_{1-p} \frac{1}{e^n \sqrt{k_1 \cdots k_n}},$$

or in other words,

$$1 - p < \left[1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) \right]^{k_G} e^n \sqrt{k_1 \cdots k_n}. \quad (1)$$

It then follows that

$$\begin{aligned} \mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) &\leq p \alpha_1 \sqrt{\frac{n}{k_G}} + (1-p) \alpha_1 \sqrt{n} \\ &\leq \alpha_1 \sqrt{\frac{n}{k_G}} + \alpha_1 \sqrt{n} \left[1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) \right]^{k_G} e^n \sqrt{k_1 \cdots k_n}. \end{aligned}$$

Our proof is therefore complete if we can show that, for any fixed n , we have

$$\frac{\alpha_1 \sqrt{\frac{n}{k_G}} + \alpha_1 \sqrt{n} \left[1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) \right]^{k_G} e^n \sqrt{k_1 \cdots k_n}}{\alpha_1 \sqrt{\frac{n}{k_G}}} \rightarrow 1$$

or equivalently that

$$\left[1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) \right]^{k_G} e^n k_G^{(n+1)/2} \rightarrow 0$$

as $k_G \rightarrow \infty$. Taking natural logarithms, this is equivalent to proving that

$$k_G \log \left(1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) \right) + n + \frac{n+1}{2} \log k_G \rightarrow -\infty$$

as $k_G \rightarrow \infty$. Notice that $0 < 1 - \prod_{j=1}^n (1 - e^{-k_j/k_G}) < 1$, since k_j/k_G is constant and $0 < e^{-k_j/k_G} < 1$ for all j . The limit above therefore holds, which completes the proof of Claim 1.

In order to prove the tighter upper bound, Claim 1a, the argument is nearly the same, except that we instead divide the region \mathcal{R} into $b(k_G)$ squares $\square_1, \dots, \square_{b(k_G)}$ of area $1/b(k_G)$, where we set

$$b(k_G) = \frac{k_G^{n/(n-1)}}{(n^2 \log k_G + \log n)^{\frac{1}{n-1}}}.$$

In the interest of brevity we will simply write b instead of $b(k_G)$. Applying precisely the same reasoning as before, the counterpart to inequality (1) is now

$$p > 1 - \left[1 - \prod_{j=1}^n (1 - e^{-k_j/b}) \right]^b e^n \sqrt{k_1 \cdots k_n},$$

so that

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{n/b} + \alpha_1 \sqrt{n} \left[1 - \prod_{j=1}^n (1 - e^{-k_j/b}) \right]^b e^n \sqrt{k_1 \cdots k_n}.$$

Our proof is therefore complete if we can show that, for any fixed n , we have

$$\frac{\alpha_1 \sqrt{n/b} + \alpha_1 \sqrt{n} \left[1 - \prod_{j=1}^n (1 - e^{-k_j/b}) \right]^b e^n \sqrt{k_1 \cdots k_n}}{\alpha_1 \sqrt{n/b}} \rightarrow 1$$

or equivalently that

$$\left[1 - \prod_{j=1}^n (1 - e^{-k_j/b}) \right]^b e^n \sqrt{k_1 \cdots k_n} \sqrt{b} \rightarrow 0$$

as $k_G \rightarrow \infty$. This is straightforward and shown in Section A of the Online Supplement.

To prove the lower bound, Claim 2, we apply Lemma 5 again: as in the proof of Theorem 4, we let \mathcal{L} denote a lattice within \mathcal{R} of the form

$$\left\{ 0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1 \right\} \times \left\{ 0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1 \right\}$$

and we assume that all of the elements of the sets \mathcal{X}_i lie in elements of \mathcal{L} (recall that we are assuming that m is arbitrarily large). By scaling the lattice of Lemma 5 by a factor of $1/m$ (i.e. our lattice \mathcal{L}), we see that the number of paths in \mathcal{L} whose length does not exceed length ℓ is at most

$$m^2 \cdot \binom{m\ell + n - 1}{n - 1} \cdot \left(\frac{8m\ell}{n - 1} \right)^{n-1} \leq m^2 \cdot \frac{(m\ell + n - 1)^{n-1}}{(n - 1)!} \cdot \left(\frac{8m\ell}{n - 1} \right)^{n-1}$$

for all $\ell > 0$. Thus, if \mathcal{P} is the path obtained by selecting the first element from each set \mathcal{X}_i and visiting these elements in a sequence chosen uniformly at random, we see that

$$\begin{aligned} \limsup_{m \rightarrow \infty} \Pr(\text{length}(\mathcal{P}) \leq \ell) &\leq \limsup_{m \rightarrow \infty} \frac{m^2 \cdot \frac{(m\ell + n - 1)^{n-1}}{(n-1)!} \cdot \left(\frac{8m\ell}{n-1} \right)^{n-1}}{m^{2n}} \\ &= \frac{8^{n-1} \ell^{2n-2}}{(n-1)!(n-1)^{n-1}}. \end{aligned}$$

The number of all possible GTSP tours through $\mathcal{X}_1, \dots, \mathcal{X}_n$ is at most $n! \prod_{i=1}^n k_i = n! \cdot k_G^n$, and therefore, it follows

from the union bound that

$$\begin{aligned} \Pr(\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \ell) &\leq n! \cdot k_G^n \cdot \frac{8^{n-1} \ell^{2n-2}}{(n-1)!(n-1)^{n-1}} \\ &= n \cdot k_G^n \cdot \frac{8^{n-1} \ell^{2n-2}}{(n-1)^{n-1}}. \end{aligned}$$

We now set $\ell = c \sqrt{\frac{n}{k_G^{n/(n-1)}}}$ with $c = \sqrt{6}/24$ to obtain

$$\begin{aligned} \mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) &\geq \left(1 - \Pr \left(\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq c \sqrt{\frac{n}{k_G^{n/(n-1)}}} \right) \right) \cdot c \sqrt{\frac{n}{k_G^{n/(n-1)}}} \\ &\geq c \left(1 - (8c^2)^{n-1} \cdot \frac{n^n}{(n-1)^{n-1}} \right) \sqrt{\frac{n}{k_G^{n/(n-1)}}} \\ &\geq 0.0681 \sqrt{\frac{n}{k_G^{n/(n-1)}}} \end{aligned}$$

for $n \geq 2$ as desired, which completes the proof. \square

Remark 7. The problem described in Theorem 4 is closely related to the BHH Theorem (i.e. Theorem 2), which corresponds to the special instance of our problem where $k = 1$. For the purpose of studying the household-level economies of scale obtained by multi-stop trips, we assert that Theorem 6 is more relevant. This is because a typical person rarely visits more than, say, 10 destinations in a given day, whereas there are likely to be much more than 10 banks, grocery stores, and so forth in a given metropolitan region. By performing numerical simulations based on [35] for small n and assuming that $|\mathcal{X}_i| = k$ is the same across all point sets \mathcal{X}_i , we adopt the approximation

$$\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \approx \alpha \sqrt{\frac{n}{k^{n/(n-1)}}},$$

where $\alpha = 0.29$.

3.3 Clustering

Our analysis of the GTSP has assumed that all of the demand points are independently and uniformly distributed in the unit square. Of course, there are many reasons why these assumptions might not hold, such as the presence of spatial competition or economies of density. Fortunately, many phenomena of this kind can already be addressed using our existing models by appropriately selecting the cardinalities $|\mathcal{X}_i|$: one example of this is the classical Hotelling model [26], which predicts that competing stores will often locate themselves immediately next to one

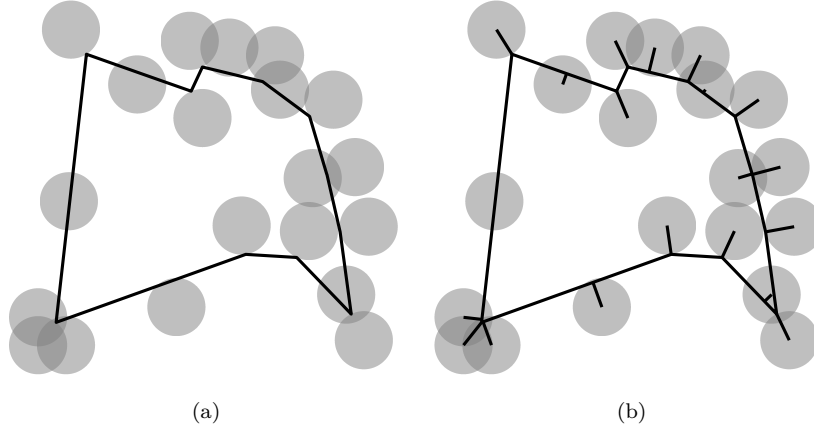


Figure 6: Figure 6a shows a tour of $n = 20$ neighborhoods; the optimal tour intersects each ball and is the shortest such tour to do so. Figure 6b shows that one can always augment an optimal tour in such a way as to touch the centers of each ball.

another. Thus, although there might be a total of k_i stores of type i in \mathcal{R} , the total number of *distinct locations* of these stores would be $k'_i < k_i$, and it would be more realistic to assume that $|\mathcal{X}_i| = k'_i$ instead. Another example is the existence of shopping malls; here, suppose that tasks $1, \dots, i^*$ can be performed at shopping malls (with $i^* \leq n$, obviously), and that $\bar{\mathcal{X}}$ denotes the set of shopping malls in \mathcal{R} . One can then compare the cost of a tour that does not use a mall, $\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n)$, with the cost of a tour that uses a mall, $\text{GTSP}(\bar{\mathcal{X}}, \mathcal{X}_{i^*+1}, \dots, \mathcal{X}_n)$. Which of the two of these is shorter depends on i^* and the cardinality of $|\bar{\mathcal{X}}|$.

If it is truly necessary to explicitly enforce clustering of each point set \mathcal{X}_i (such as a garment district or an enclave such as a Chinatown or Little Italy), then an alternate model is required. One possibility is the *TSP with neighborhoods* [20], which is a special case of the GTSP in which each point set \mathcal{X}_i is a ball \mathcal{B}_i of radius r (as opposed to being a finite set of points) that is centered at a point x_i , and the goal is to find the shortest tour that touches every ball; see Figure 6a for an example. We will derive asymptotic expressions for the TSP with neighborhoods with the help of the lemma below:

Lemma 8. *Let $\mathcal{B}_1, \dots, \mathcal{B}_n$ be a collection of balls of radius r , centered at points x_1, \dots, x_n , all of which are contained in the unit square. We have*

$$\text{TSP}(x_1, \dots, x_n) - 2nr \leq \text{GTSP}(\mathcal{B}_1, \dots, \mathcal{B}_n) \leq \min \left\{ \text{TSP}(x_1, \dots, x_n), \left\lceil \frac{1}{2r} \right\rceil + 3 \right\}.$$

Proof. The leftmost inequality holds because one can always make a tour that touches x_1, \dots, x_n by augmenting the tour $\text{GTSP}(\mathcal{B}_1, \dots, \mathcal{B}_n)$ with n line segments of length at most r , as shown in Figure 6b. The fact that

$\text{GTSP}(\mathcal{B}_1, \dots, \mathcal{B}_n) \leq \text{TSP}(x_1, \dots, x_n)$ is obvious. The fact that $\text{GTSP}(\mathcal{B}_1, \dots, \mathcal{B}_n) \leq \lceil \frac{1}{2r} \rceil + 3$ is due to essentially the same idea as that expressed in Figure 3a; if we construct a tour that traverses the width of \mathcal{R} horizontally a total of $\lceil \frac{1}{2r} \rceil$ times (with the “+3” term added because we also travel one unit down and one unit up as before and because, if $\lceil \frac{1}{2r} \rceil$ is odd, then we must make one additional horizontal traversal), then we must touch each ball at some point. \square

Using the fact that $\text{TSP}(x_1, \dots, x_n) \sim \beta\sqrt{n}$ for uniformly distributed points x_i , we can write an approximation of Lemma 8 for large n as

$$\beta\sqrt{n} - 2nr \lesssim \text{GTSP}(\mathcal{B}_1, \dots, \mathcal{B}_n) \lesssim \min \left\{ \beta\sqrt{n}, \left\lceil \frac{1}{2r} \right\rceil + 3 \right\}$$

where the notation “ \lesssim ” reflects the lower-order terms that we are dropping by introducing the square root approximation. There are two aspects of the inequalities above that need correction: the first is that we can tighten the left-hand inequality by using the fact that $\beta\sqrt{n'} - 2n'r \lesssim \text{GTSP}(\mathcal{B}_1, \dots, \mathcal{B}_n)$ for all $n' \leq n$. The lower bound is maximized when $n' = \frac{\beta^2}{16r^2}$, at which the bound evaluates to $\beta\sqrt{n'} - 2n'r = \frac{\beta^2}{8r}$. Thus, a tighter lower bound is to use $\beta\sqrt{n} - 2nr$ if $n < \frac{\beta^2}{16r^2}$ and to use $\frac{\beta^2}{8r}$ otherwise. The second correction is that we should not require that r be constant, because this would result in both inequalities becoming constant as $n \rightarrow \infty$. Thus, we represent r as a *sequence* indexed by n , $\{r_n\}$, with the assumption that $r_n \rightarrow 0$ and $nr_n \rightarrow 0$ as $n \rightarrow \infty$. In summary, our new bounds are

$$\left. \begin{array}{ll} \beta\sqrt{n} - 2nr_n & \text{if } n < \frac{\beta^2}{16r_n^2} \\ \frac{\beta^2}{8r_n} & \text{otherwise} \end{array} \right\} \lesssim \text{GTSP}(\mathcal{B}_1, \dots, \mathcal{B}_n) \lesssim \min \left\{ \beta\sqrt{n}, \frac{1}{2r_n} \right\},$$

and it is straightforward to verify that the left- and right-hand sides of the inequalities above are always within a factor of $4/\beta^2$ of one another.

4 Models

This section considers several models of increasing complexity. In each model, we assume that each person has n “tasks” that they must complete each day (such as going to work, the grocer, and so forth), and each of these tasks can be performed at k different locations. Thus, it would be sensible to postulate that the distance travelled by each person in the region would be given by the expression $\text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n)$, where the point sets \mathcal{X}_j each denote the locations at which these tasks can be performed. By the analysis in Remark 7, we would approximate this with the expression $\alpha\sqrt{n/k^{n/(n-1)}}$. However, some potential error may result because this expression does not take

into account the additional distance incurred by leaving and returning to each person’s house, which we assume is distributed uniformly at random in the region. Ideally, we would like to approximate this with the expression $\text{GTSP}(\{x_i\}, \mathcal{X}_1, \dots, \mathcal{X}_n)$, where x_i denotes person i ’s home location, and we will do so in Section 6. For now, we simply remark that this expression is somewhat unwieldy because it involves computing a generalized TSP tour of sets of different magnitudes and thus involves a comparison of “inbound-and-outbound” costs of starting and ending a trip as well as the “peddling” cost of moving between destinations on this trip (see [33] for a detailed overview). Since the purpose of this study is to examine the benefits of multi-stop trips at the household level, we instead opt to approximate each person’s distance travelled (from their house to their n destinations) as $\alpha\sqrt{n/k^{n/(n-1)}}$ as justified in Remark 7.

4.1 A simple example: luddites and shut-ins

The scenario we describe in this section is too simple to be of practical use, but is helpful as a “minimum working example” that explains what factors affect the carbon footprint in a region most significantly. Suppose that our city is a square region \mathcal{R} of area 1 and has a population N . Each person has n locations to visit daily ($n - 1$ errands plus their home) and each errand has k possible locations where that errand can be performed (e.g. there are k grocery stores and k banks in the region). Each of the N people in the region corresponds to a point sampled independently and uniformly at random in \mathcal{R} , and each person belongs to one of two classes, either “luddites” or “shut-ins”, distinguished as follows:

- A *luddite* performs all of their tasks by themselves and drives to each of the n locations.
- A *shut-in* shops for everything online and remains stationary while packages are delivered to them.

Let the fraction of shut-ins in the city be p , implying that pN people are to be served by a delivery truck. This truck performs a travelling salesman tour of pN points, whose length is approximately $\beta\sqrt{pN}$ (for large N) by Theorem 2, with $\beta \approx 0.7124$. Therefore, the total carbon footprint due to these shut-ins is $\phi\beta\sqrt{pN}$, where ϕ represents the amount of emissions produced per mile driven by a delivery truck.

A luddite visits n places each day (their house, plus their $n - 1$ tasks) and has k choices for each place to visit. From Theorem 6 and Remark 7, as well as the introductory paragraph to this section, we adopt the expression $\alpha\sqrt{n/k^{n/(n-1)}}$ to model the distance traversed by each luddite, where $\alpha = 0.29$ is the proportionality constant of Remark 7. There are $(1 - p)N$ such people, and their total carbon footprint is therefore $\psi(1 - p)N\alpha\sqrt{n/k^{n/(n-1)}}$, where ψ represents the amount of emissions produced per mile driven by a passenger car. The total carbon footprint

of the region, regarded as a function of p , is then given by

$$f(p) := \phi\beta\sqrt{pN} + \psi(1-p)N\alpha\sqrt{n/k^{n/(n-1)}}, \quad (2)$$

which is concave in p . Note that when $p = 0$ (i.e. there are no shut-ins and everyone does their own driving), the carbon footprint is $\psi N\alpha\sqrt{n/k^{n/(n-1)}}$. We also note that

$$f(p)\Big|_{p=\frac{\phi^2\beta^2k^{n/(n-1)}}{\psi^2\alpha^2nN}} = f(p)\Big|_{p=0} = \psi N\alpha\sqrt{n/k^{n/(n-1)}},$$

which (coupled with the concavity of $f(\cdot)$) tells us that we must have $p \geq \frac{\phi^2\beta^2k^{n/(n-1)}}{\psi^2\alpha^2nN} =: p_0$ in order for the carbon footprint to be reduced as a result of using delivery services. It is also worth pointing out that $f(\cdot)$ is *maximized* when $p = p_0/4$, and attains a maximum value of

$$\phi\beta\sqrt{pN} + \psi(1-p)N\alpha\sqrt{n/k^{n/(n-1)}}\Big|_{p=p_0/4} = \frac{\phi^2\beta^2\sqrt{k^{n/(n-1)}} + 4\psi^2\alpha^2nN/\sqrt{k^{n/(n-1)}}}{4\psi\alpha\sqrt{n}}.$$

4.2 Marginal costs

The preceding model describes an extreme case in which each person in \mathcal{R} either makes no use whatsoever of delivery services or uses delivery services exclusively. One middle ground that is also worth studying is the case where people belong to two classes as before, but the two classes differ by only one task:

- A *luddite* performs all of their tasks by themselves and drives to each of the n locations (this is the same as in the preceding model).
- An *early adopter* visits $n - 1$ locations and uses a delivery service for the remaining task.

A model of this kind is useful when one wants to understand the benefits of implementing a new delivery service for a specific good, such as groceries; one example of this can be found in [51], which discusses the consequences of introducing grocery delivery services in Seattle, Washington. A more nationalized phenomenon would be the recent introduction of “last-mile” services such as Google Shopping Express [13], which offers same-day deliveries facilitated by a specialized fleet of vehicles. If we let p denote the fraction of early adopters in \mathcal{R} , we then see that

the total carbon footprint in the region is given by

$$\begin{aligned}
f(p) : &= \underbrace{\phi\beta\sqrt{pN} + \psi p N \alpha \sqrt{(n-1)/k^{n/(n-1)}}}_{\text{early adopters}} + \underbrace{\psi(1-p)N\alpha\sqrt{n/k^{n/(n-1)}}}_{\text{luddites}} \\
&= \phi\beta\sqrt{pN} + \frac{\psi N \alpha}{\sqrt{k^{n/(n-1)}}} [\sqrt{n} - p(\sqrt{n} - \sqrt{n-1})] \\
&\approx \phi\beta\sqrt{pN} + \frac{\psi N \alpha}{\sqrt{k^{n/(n-1)}}} \left[\sqrt{n} - \frac{p}{2\sqrt{n}} \right]
\end{aligned}$$

where all terms (except for p) are the same as in (2), and we have used the series approximation $\sqrt{n} - \sqrt{n-1} = 1/2\sqrt{n} + \mathcal{O}(n^{-3/2})$ in the last line. In the same manner as in the preceding section, we note that when $p = 0$, the carbon footprint is $\psi N \alpha \sqrt{n/k^{n/(n-1)}}$, and also that

$$f(p)|_{p=\frac{4\phi^2\beta^2nk^{n/(n-1)}}{\psi^2\alpha^2N}} = f(p)|_{p=0} = \psi N \alpha \sqrt{n/k^{n/(n-1)}},$$

which tells us that we must have $p \geq \frac{4\phi^2\beta^2nk^{n/(n-1)}}{\psi^2\alpha^2N}$ in order for the carbon footprint to be reduced as a result of using delivery services. Note that this threshold is greater than that of the preceding section (which had a threshold of $\frac{\phi^2\beta^2k^{n/(n-1)}}{\psi^2\alpha^2nN}$) by a factor of $4n^2$; this is simply a mathematical manifestation of the intuition that larger values of n (i.e. more trip chaining at the household level) lead to significantly greater economies of scale at the household level. This in turn implies that delivery services must be adopted at a larger rate in order for the carbon footprint to decrease.

4.3 Multiple delivery services

The model in Section 4.1 assumes that a single delivery truck serves all of the pN luddites. It is not hard to model the case where there are multiple such services; the main difference is that there is a loss in efficiency because competing delivery services do not consolidate their routes together, thus reducing the benefits of economies of scale. Suppose that there are m delivery services in the region, and that service i delivers goods to a fraction δ_i of the shut-ins in the region, visiting $\delta_i p N$ customers in total (this model assumes that each shut-in is uniquely associated with one delivery service). Therefore, applying Theorem 2, we see that the work done by service i can be approximated as $\beta\sqrt{\delta_i p N}$, and therefore the total carbon footprint due to shut-ins is $\phi\beta\sqrt{pN} \sum_{i=1}^m \sqrt{\delta_i}$. Obviously, since $\sum_{i=1}^m \sqrt{\delta_i} \geq 1$ always holds, we see that the carbon footprint within the region will only increase as a result of employing multiple delivery services (provided that these services do not cooperate to share their loads efficiently).

The total carbon footprint of the region is then given by

$$f(p) := \phi\beta\sqrt{pN} \sum_{i=1}^m \sqrt{\delta_i} + \psi(1-p)N\alpha\sqrt{n/k^{n/(n-1)}} = \bar{\phi}\beta\sqrt{pN} + \psi(1-p)N\alpha\sqrt{n/k^{n/(n-1)}},$$

where we define $\bar{\phi} := \phi \sum_{i=1}^m \sqrt{\delta_i}$, which reduces to the same problem as (2).

4.4 A probabilistic model

The model in this section improves on that of 4.1 by modelling customer behavior in a smoother way than the luddite/shut-in dichotomy. Rather than assigning a set fraction of the population to be shut-ins, we assume that each customer uses a delivery service to do each of their n daily tasks with probability $1 - q$. Thus, the number of locations that the person actually visits is a binomial random variable, X , with parameters n and q , and the expected amount of driving for that person is $\mathbf{E}(\alpha\sqrt{X/k^{n/(n-1)}}) = \frac{\alpha}{\sqrt{k^{n/(n-1)}}}\mathbf{E}(\sqrt{X})$.

If a person chooses to perform a task online, then a delivery truck will visit their house. Note that the only circumstance under which a delivery truck does not visit their house is if that person chooses to complete *all* n activities by driving to n different locations. Thus, the probability that a person is visited by a delivery truck is given by $p := 1 - q^n$, and we see that the number of houses that the truck visits is a binomial random variable Y with parameters N and p , and the expected distance that the delivery truck travels is $\beta\mathbf{E}(\sqrt{Y})$. The total carbon footprint in the region is therefore

$$f(p) := \phi\beta\mathbf{E}(\sqrt{Y}) + \psi\frac{\alpha N}{\sqrt{k^{n/(n-1)}}}\mathbf{E}(\sqrt{X}), \quad (3)$$

where $X \sim B(n, q)$ and $Y \sim B(N, p)$. In order to simplify the above expression, the following lemma is useful:

Lemma 9. *Let $X \sim B(n, p)$ be a binomially distributed random variable. Then as $p \rightarrow 0$ with n fixed, we have $\mathbf{E}(\sqrt{X}) \sim np$, and as $p \rightarrow 1$ with n fixed, we have $\mathbf{E}(\sqrt{X}) \sim \sqrt{np}$.*

Proof. See Section B of the online supplement. □

The preceding lemma allows us to analyze the limiting behavior of the total carbon footprint with respect to p . For values of p near 0 (which implies that q is close to 1, i.e. very little delivery is used), we can approximate (3) by

$$f(p) \approx \phi\beta Np + \psi\alpha N\sqrt{\frac{nq}{k^{n/(n-1)}}} = N \left[\phi\beta p + \psi\alpha\sqrt{\frac{n(1-p)^{1/n}}{k^{n/(n-1)}}} \right].$$

Note that according to this approximation, we have

$$\left. \frac{df}{dp} \right|_{p=0} \approx N \left(\phi\beta - \frac{\psi\alpha}{2\sqrt{nk^{n/(n-1)}}} \right),$$

which we expect to be positive since $\phi\beta$ and $\psi\alpha$ ought to be approximately the same order of magnitude. This tells that initially, as people in \mathcal{R} begin to make more use of delivery services, the total carbon footprint in the region increases since using trucks for a delivering products to a small number of locations is not efficient.

On the contrary, for values of p near 1 (which implies that q is close to 0, i.e. delivery system is used for almost everything), we can approximate (3) by

$$f(p) \approx \phi\beta\sqrt{Np} + \frac{\psi\alpha Nnq}{\sqrt{k^{n/(n-1)}}}.$$

Then according to this approximation, we have

$$\frac{df}{dp} \approx \frac{\phi\beta\sqrt{N}}{2\sqrt{p}} - \frac{\psi\alpha N}{(1-p)^{(n-1)/n}\sqrt{k^{n/(n-1)}}},$$

which goes to $-\infty$ as $p \rightarrow 1$. This means that if everybody in \mathcal{R} uses delivery services, the total carbon footprint in the region decreases rapidly since trucks can serve large number of locations efficiently.

5 A numerical example

In this section, we give a simple example of an instance of the model in Section 4.2 using estimates of the relevant input parameters. This model seems to be the most timely, as evidenced by the prevalence of “last-mile” delivery services such as Google Shopping Express, Amazon Prime, Instacart, and Walmart To Go [10, 13, 24, 43]. Table 1 shows our estimates for parameters ϕ , ψ , α , and β , which we do not expect (for the most part) to vary on the region being served. In order to estimate k for various regions, we used census data obtained from [1] that gives the number of grocery stores in various metropolitan regions; these numbers (as well as N , the populations of these regions) are shown in Table 2. Figure 7 shows plots of the total emissions, $f(p)$ normalized by $f(0)$, for four metropolitan areas. From Table 2, we see that the critical thresholds p^* are quite high, and in many instances, the household-level economies of scale are high enough that even a 100% usage of delivery services is less efficient than leaving drivers to their own devices (this corresponds to the entries in the table that are marked “> 1”). The alternative analysis in the next section is somewhat more optimistic.

Parameter	Estimate	Source
ϕ	1303 $\frac{\text{grams CO}_2}{\text{mile}}$	[32]
ψ	350 $\frac{\text{grams CO}_2}{\text{mile}}$	[4]
α	0.29	Numerical simulations based on [35]
β	0.7124	[7]

(a) Parameter estimates and their sources.

Table 1: Input parameter estimates for our numerical example.

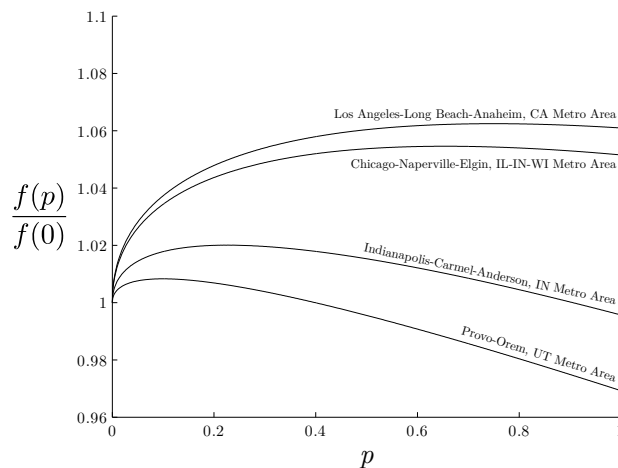


Figure 7: A plot of the total carbon footprint, $f(p)$, for four different cities, for the model described in model in Section 4.2. Here we assume that $n = 6$.

6 Incorporating inbound-and-outbound costs

As described in the beginning of Section 4, the models we have described thus far have not paid special attention to the “inbound-and-outbound” costs associated with leaving and returning to one’s home. This section describes a result that is related to Theorems 4 and 6 in which we take a generalized TSP tour of n point sets, $\mathcal{X}_1, \dots, \mathcal{X}_n$, in addition to a fixed point x_0 (which represents a person’s home). Note that the limiting behavior for fixed k and $n \rightarrow \infty$ is the same as in Theorems 4 because we are merely inserting one additional point, and therefore it will suffice to consider only the limiting behavior for the case where n is fixed and $k \rightarrow \infty$:

Theorem 10. *Let $\mathcal{X}_1, \dots, \mathcal{X}_n$ denote n sets of points, each having cardinality k , and suppose that all nk points are distributed independently and uniformly at random in a region \mathcal{R} having area A . Let x_0 be a point in the interior*

$$p^* = \frac{4\phi^2 \beta^2 n k^{n/(n-1)}}{\psi^2 \alpha^2 N}$$

Region	k	N	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$
Los Angeles-Long Beach-Anaheim, CA Metro Area	3358	13052921	> 1	> 1	> 1	> 1	> 1
Chicago-Naperville-Elgin, IL-IN-WI Metro Area	2889	9522434	> 1	> 1	> 1	> 1	> 1
Indianapolis-Carmel-Anderson, IN Metro Area	295	1928982	> 1	> 1	> 1	0.95	0.92
Salt Lake City, UT Metro Area	192	1123712	> 1	> 1	> 1	0.98	0.96
Tulsa, OK Metro Area	136	951880	> 1	0.98	0.81	0.76	0.75
Albuquerque, NM Metro Area	119	901700	> 1	0.86	0.72	0.68	0.68
El Paso, TX Metro Area	138	830735	> 1	> 1	0.95	0.89	0.88
McAllen-Edinburg-Mission, TX Metro Area	132	806552	> 1	> 1	0.92	0.87	0.86
Little Rock-North Little Rock-Conway, AR Metro Area	124	717666	> 1	> 1	0.96	0.90	0.90
Colorado Springs, CO Metro Area	83	668353	> 1	0.72	0.62	0.60	0.60
Boise City, ID Metro Area	73	637896	0.98	0.64	0.55	0.54	0.54
Provo-Orem, UT Metro Area	50	550845	0.64	0.44	0.40	0.39	0.40
Killeen-Temple, TX Metro Area	84	420375	> 1	> 1	> 1	0.97	0.97
Green Bay, WI Metro Area	43	311098	0.90	0.64	0.59	0.58	0.60
Clarksburg, WV Micro Area	25	94310	> 1	> 1	0.99	> 1	> 1
Elmira, NY Metro Area	24	88911	> 1	> 1	0.99	> 1	> 1
DuBois, PA Micro Area	22	81184	> 1	> 1	0.98	> 1	> 1

Table 2: The number of grocery stores k , the populations N , and the critical thresholds p^* at which emissions decrease due to adoption of delivery services. The first and second columns are obtained from [1]. Here the cells marked “> 1” indicate that, even at 100% adoption of delivery services, the carbon footprint of the region is still larger than the case where $p = 0$, i.e. no delivery services are used.

of \mathcal{R} . Then the expected length of a generalized TSP tour of $\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n$ satisfies

$$\mathbf{E} \text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) \in \mathcal{O}(\sqrt{An/k} \cdot \sqrt{\log k})$$

and

$$\mathbf{E} \text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) \in \Omega\left(\sqrt{An/k}\right)$$

as $k \rightarrow \infty$ with n fixed. Specifically, there exists a constant $\alpha_3 > 0.136$ such that the following statements hold:

1. For any n , there exists a threshold k_0 such that

$$\mathbf{E} \text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{An/k} \cdot \sqrt{\log k}$$

whenever $k \geq k_0$, where $\alpha_1 = 2.7$ is the constant from Lemma 1.

2. For any n , there exists a threshold k_0 such that

$$\mathbf{E} \text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) \geq \alpha_3 \sqrt{\frac{An}{k}}$$

whenever $k \geq k_0$.

Proof. Assume as in the previous proofs that \mathcal{R} is the unit square. To prove Claim 1, consider a square \square_0 of area $a := \log k / 2k$ centered at the point x_0 , and suppose that k is sufficiently large (i.e. that a is sufficiently small) that \square_0 is entirely contained within \mathcal{R} . Let p denote the probability that \square_0 contains at least one element from each point set \mathcal{X}_i , in which case we clearly have $\text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) \leq \alpha_1 \sqrt{a(n+1)}$ from Lemma 1. Since $p = (1 - (1 - a)^k)^n$, we therefore see that as $k \rightarrow \infty$, we have

$$\begin{aligned} \mathbf{E} \text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) &\leq p\alpha_1 \sqrt{a(n+1)} + (1-p)\alpha_1 \sqrt{n+1} \\ &= \frac{\alpha_1}{2} \sqrt{\frac{n+1}{k}} \cdot \left(\left[1 - k^{-k} \left(k - \frac{1}{2} \log k \right)^k \right]^n \left(\sqrt{2 \log k} - 2\sqrt{k} \right) + 2\sqrt{k} \right) \\ &\sim \frac{\alpha_1}{2} \sqrt{\frac{n+1}{k}} \cdot \left[\left(1 - n\sqrt{1/k} \right) \left(\sqrt{2 \log k} - 2\sqrt{k} \right) + 2\sqrt{k} \right] \\ &= \frac{\sqrt{2}\alpha_1}{2} \sqrt{\frac{n+1}{k}} \cdot \sqrt{\log k} \leq \alpha_1 \sqrt{n/k} \cdot \sqrt{\log k} \end{aligned}$$

as desired.

To prove the second claim, we find it useful to revisit Lemma 5, and we again let \mathcal{L} denote a lattice within \mathcal{R}

of the form

$$\left\{0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1\right\} \times \left\{0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1\right\}.$$

We first see that there are at most k^n total distinct subsets of the form $\{x_0, x_1, \dots, x_n\}$, with $x_i \in \mathcal{X}_i$ for each $i \geq 1$. For each of these subsets, there are $n!$ different orderings, and we therefore conclude that there are at most $n! \cdot k^n$ valid paths that originate at x_0 and visit one point from each of the subsets \mathcal{X}_i . By applying the union bound, we see that for any ℓ ,

$$\begin{aligned} \Pr(\text{GTSP}(x_0, \mathcal{X}_1, \dots, \mathcal{X}_n) \leq \ell) &= \Pr(\text{one of the } n! \cdot k^n \text{ valid paths has length } \leq \ell) \\ &\leq n! \cdot k^n \Pr(\text{length}(\mathcal{P}) \leq \ell) \end{aligned}$$

where \mathcal{P} is the path obtained by selecting the first element from each set \mathcal{X}_i and visiting these elements in a sequence chosen uniformly at random. By construction, we see that \mathcal{P} is simply a path that is sampled uniformly at random from \mathcal{P} , the collection of all possible paths originating at x_0 that visit n additional points taken from the lattice \mathcal{L} . Of course, we can again see that $|\mathcal{P}| = m^{2n}$. By scaling the lattice of Lemma 5 by a factor of $1/m$ in the vertical and horizontal directions, there are at most

$$\binom{m\ell + n}{n} \cdot \left(\frac{8m\ell}{n}\right)^n \leq \frac{(m\ell + n)^n}{n!} \cdot \left(\frac{8m\ell}{n}\right)^n$$

paths in \mathcal{P} whose length is at most ℓ . Thus,

$$\begin{aligned} \limsup_{m \rightarrow \infty} \Pr(\text{length}(\mathcal{P}) \leq \ell) &\leq \limsup_{m \rightarrow \infty} \frac{\frac{(m\ell + n)^n}{n!} \cdot \left(\frac{8m\ell}{n}\right)^n}{m^{2n}} \\ &= \frac{\ell^{2n} 8^n n^{-n}}{n!} \\ \implies \Pr(\text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) \leq \ell) &\leq k^n \ell^{2n} 8^n n^{-n} \end{aligned}$$

for all ℓ . We now set $\ell = c\sqrt{n/k}$, with $c = \sqrt{6}/12$, obtaining

$$\begin{aligned} \mathbf{E} \text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) &\geq (1 - k^n \ell^{2n} 8^n n^{-n}) \ell \\ &= \frac{1}{12} \cdot \frac{\sqrt{6n}(1 - 3^{-n})}{\sqrt{k}} > 0.136\sqrt{n/k}, \end{aligned}$$

which completes the proof. □

Remark 11. By performing numerical simulations for small n and large k in the same way as in Remark 7, we adopt

the approximation

$$\text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n) \approx \alpha' \sqrt{n/k},$$

where $\alpha' = 0.47$.

6.1 Revised simulations

In this section we present a revised numerical simulation that is entirely analogous to that of Section 5, only we now incorporate inbound and outbound costs as in Theorem 10 by utilizing Remark 11. Thus, whereas we previously had a critical threshold of

$$p^* = \frac{4\phi^2\beta^2nk^{n/(n-1)}}{\psi^2\alpha^2N},$$

we now see by a straightforward analysis that under the revised model, the critical threshold is instead

$$p^* = \frac{4\phi^2\beta^2nk}{\psi^2(\alpha')^2N}.$$

These thresholds are shown in Table 3. These are somewhat more encouraging than those of Section 5, although we still see that a significant amount of adoption of delivery services is required. A recent survey [2] of 22,000 homes suggests that approximately 13% of shoppers have purchased groceries online in the last 30 days, which happens to be close to the average of the entries of Table 3. This would suggest that, at present, the benefits to carbon footprints due to delivery services are just beginning to be realized, if at all.

7 Conclusions

We have conducted an asymptotic analysis of the generalized TSP tour of n point sets $\mathcal{X}_1, \dots, \mathcal{X}_n$ with cardinalities k_1, \dots, k_n all in the unit square, $\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n)$, in two cases, first where $n \rightarrow \infty$ and k_i 's are fixed and equal to k and second where n is fixed and the average cardinality size gets large, $k_G \rightarrow \infty$. For the case where $n \rightarrow \infty$, our analysis that $\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \in \Theta(\sqrt{An/k})$ is tight inasmuch as the lower and upper bounds have the same order. For the case where n is fixed and average cardinality gets large, $k_G \rightarrow \infty$, our upper and lower bounds differ by a term of order $(n^2 \log k_G + \log n)^{\frac{1}{2(n-1)}}$. We conjecture that, in this case, one actually has

$$\mathbf{E} \text{GTSP}(\mathcal{X}_1, \dots, \mathcal{X}_n) \in \Theta \left(\sqrt{\frac{n}{k_G^{n/(n-1)}}} \right),$$

$$p^* = \frac{4\phi^2\beta^2nk}{\psi^2(\alpha')^2N}$$

Region	k	N	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
Los Angeles-Long Beach-Anaheim, CA Metro Area	3358	13052921	0.10	0.14	0.17	0.20	0.23
Chicago-Naperville-Elgin, IL-IN-WI Metro Area	2889	9522434	0.12	0.16	0.20	0.24	0.28
Indianapolis-Carmel-Anderson, IN Metro Area	295	1928982	0.06	0.08	0.10	0.12	0.14
Salt Lake City, UT Metro Area	192	1123712	0.07	0.09	0.11	0.14	0.16
Tulsa, OK Metro Area	136	951880	0.06	0.08	0.10	0.11	0.13
Albuquerque, NM Metro Area	119	901700	0.06	0.07	0.09	0.11	0.12
El Paso, TX Metro Area	138	830735	0.07	0.09	0.11	0.13	0.15
McAllen-Edinburg-Mission, TX Metro Area	132	806552	0.07	0.09	0.11	0.13	0.15
Little Rock-North Little Rock-Conway, AR Metro Area	124	717666	0.07	0.09	0.12	0.14	0.16
Colorado Springs, CO Metro Area	83	668353	0.05	0.07	0.08	0.10	0.12
Boise City, ID Metro Area	73	637896	0.05	0.06	0.08	0.09	0.11
Provo-Orem, UT Metro Area	50	550845	0.04	0.05	0.06	0.07	0.09
Killeen-Temple, TX Metro Area	84	420375	0.08	0.11	0.13	0.16	0.18
Green Bay, WI Metro Area	43	311098	0.06	0.08	0.09	0.11	0.13
Clarksburg, WV Micro Area	25	94310	0.11	0.14	0.17	0.21	0.24
Elmira, NY Metro Area	24	88911	0.11	0.14	0.18	0.21	0.25
DuBois, PA Micro Area	22	81184	0.11	0.14	0.18	0.21	0.25

Table 3: The number of grocery stores k , the populations N , and the critical thresholds p^* at which emissions decrease as adoption of delivery services increases under the revised model. The first and second columns are obtained from [1].

although this will clearly require further analysis. We have also analyzed the generalized TSP for the case where one also has a singleton x_0 , that is, $\mathbf{E} \text{GTSP}(\{x_0\}, \mathcal{X}_1, \dots, \mathcal{X}_n)$, in which case we have upper and lower bounds that differ by a term of order $\sqrt{\log k}$.

Numerical analyses of the preceding models, when applied to the problem of estimating the change in carbon footprint that results from using delivery services, suggest that a considerable amount of adoption of delivery services is necessary before one begins to see a decrease in carbon footprint. The reason for this is simply that the economy of scale that is realized by delivery services requires a significant initial level of adoption in order to compete with the “household-level” economy of scale that the customers already possess, given their wide choice of locations to visit and the number of daily locations they visit.

Acknowledgments

The authors thank David Aldous and Peter Shor for their helpful insights.

References

- [1] 2012 MSA Business Patterns (NAICS). <http://censtats.census.gov/cgi-bin/msanaic/msasect.pl>. [Online; accessed 22-August-2014].
- [2] 13% bought groceries online: survey, June 2013. [Online; posted 18-June-2013].
- [3] T. Adler and M. Ben-Akiva. A theoretical and empirical model of trip chaining behavior. *Transportation Research Part B: Methodological*, 13(3):243–257, 1979.
- [4] National Highway Traffic Safety Administration. Summary of fuel economy performance. *US Department of Transportation, NHTSA*, 2011.
- [5] D. Aldous and M. Krikun. Percolating paths through random points. *arXiv preprint math/0509492*, 2005.
- [6] W. P. Anderson, L. Chatterjee, and T. R. Lakshmanan. E-commerce, transportation, and economic geography. *Growth and Change*, 34(4):415–432, 2003.
- [7] D. Applegate, W. Cook, D. S. Johnson, and N. J. A. Sloane. Using large-scale computation to estimate the Beardwood-Halton-Hammersley TSP constant. Presentation at 42 SBPO, 2010.
- [8] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

- [9] D. Ben-Arieh, G. Gutin, M. Penn, A. Yeo, and A. Zverovitch. Transformations of generalized atsp into atsp. *Operations Research Letters*, 31(5):357–365, 2003.
- [10] G. Bensinger. Startup grocery deliverer Instacart eyes Amazon. *Wall Street Journal*, August 7 2013.
- [11] D. J. Bertsimas and D. Simchi-Levi. A new generation of vehicle routing research: robust algorithms, addressing uncertainty. *Operations Research*, 44(2):286–304, 1996.
- [12] A. Berube, W. H. Frey, A. Friedhoff, E. Garr, E. Istrate, E. Kneebone, R. Puentes, A. Singer, A. Tomer, H. Wial, et al. State of metropolitan america: On the front lines of demographic transformation, 2010.
- [13] Robert Bowman. Will Google Shopping Express help retailers fend off challenge from Amazon?, June 2014. [Online; posted 17-June-2014].
- [14] L. D. Burns, R. W. Hall, D. E. Blumenfeld, and C. F. Daganzo. Distribution strategies that minimize transportation and inventory costs. *Operations Research*, 33(3):469–490, 1985.
- [15] Gérard P Cachon. Supply chain design and the cost of greenhouse gas emissions. *Working Paper*, 2011.
- [16] Gérard P Cachon. Retail store density and the cost of greenhouse gas emissions. *Management Science*, 60(8):1907–1925, 2014.
- [17] J. F. Campbell. Location and allocation for distribution systems with transshipments and transportation economies of scale. *Annals of Operations Research*, 40(1):77–99, 1992.
- [18] C. Daganzo. *Logistics Systems Analysis*. Springer, 2005.
- [19] E. Demaine, L. W. Sun, and C. E. Leiserson. Introduction to Algorithms, Problem Set 1. <http://courses.csail.mit.edu/6.046/fall101/handouts/ps1.pdf>, September 2001. [Online; accessed 8-Sep-2014].
- [20] A. Dumitrescu and J. S. B. Mitchell. Approximation algorithms for tsp with neighborhoods in the plane. *Journal of Algorithms*, 48(1):135–159, 2003.
- [21] L. Few. The shortest path and the shortest road through n points. *Mathematika*, 2:141–144, 1955.
- [22] S.R. Finch. *Mathematical Constants*. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, 2003.
- [23] M. Haimovich and A. H. G. Rinnooy Kan. Bounds and heuristics for capacitated routing problems. *Mathematics of Operations Research*, 10(4):527–542, 1985.

- [24] C. Harris and J. Cook. Amazon starts grocery delivery service. *Seattle Post-Intelligencer*, August 1 2007.
- [25] D. A. Hensher and A. J. Reyes. Trip chaining as a barrier to the propensity to use public transport. *Transportation*, 27(4):341–361, 2000.
- [26] H. Hotelling. Stability in competition. *The Economic Journal*, 39(153):41–57, 1929.
- [27] R. Howell. On asymptotic notation with multiple variables. Technical Report 2007-4, Department of Computer Science, Kansas State University, 2008.
- [28] M. Huang, K. R. Smilowitz, and B. Balcik. A continuous approximation approach for assessment routing in disaster relief. *Transportation Research Part B: Methodological*, 50:20–41, 2013.
- [29] O. Jabali, M. Gendreau, and G. Laporte. A continuous approximation model for the fleet composition problem. *Transportation Research Part B: Methodological*, 46(10):1591–1606, 2012.
- [30] H. J. Karloff. How long can a euclidean traveling salesman tour be? *SIAM Journal on Discrete Mathematics*, 2(1):91–99, 1989.
- [31] R. Kitamura. Incorporating trip chaining into analysis of destination choice. *Transportation Research Part B: Methodological*, 18(1):67–81, 1984.
- [32] D. Kodjak. Policy discussion – heavy-duty truck fuel economy. Presentation at 10th Diesel Engine Emissions Reduction (DEER) Conference, 2004.
- [33] A. Langevin, P. Mbaraga, and J. F. Campbell. Continuous approximation models in freight distribution: An overview. *Transportation Research Part B: Methodological*, 30(3):163 – 188, 1996.
- [34] G. Laporte, A. Asef-Vaziri, and C. Sriskandarajah. Some applications of the generalized travelling salesman problem. *Journal of the Operational Research Society*, pages 1461–1467, 1996.
- [35] G. Laporte and F. Semet. Computational evaluation of a transformation procedure for the symmetric generalized traveling salesman problem. *INFOR*, 37(2):114–120, 1999.
- [36] H. L. Lee and S. Whang. Winning the last mile of e-commerce. *MIT Sloan Management Review*, 42(4):54–62, 2001.
- [37] A. C. McKinnon and A. Woodburn. Shopping trip or home delivery? – which has the largest carbon footprint. *Logistics and Transport Focus*, 11(7):20–25, 2009.

- [38] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.
- [39] P. L. Mokhtarian. A conceptual analysis of the transportation impacts of b2c e-commerce. *Transportation*, 31(3):257–284, 2004.
- [40] C. E. Noon and J. C. Bean. A lagrangian based approach for the asymmetric generalized traveling salesman problem. *Operations Research*, 39(4):623–632, 1991.
- [41] A. G. N. Novaes, J. E. S. de Cursi, and O. D. Gracioli. A continuous approach to the design of physical distribution systems. *Computers & Operations Research*, 27(9):877–893, 2000.
- [42] Y. Ouyang. Design of vehicle routing zones for large-scale distribution systems. *Transportation Research Part B: Methodological*, 41(10):1079–1093, 2007.
- [43] Associated Press. Got groceries? Wal-Mart testing home delivery. *The Herald Bulletin*, April 23 2011.
- [44] C. Redmond and J. E. Yukich. Limit theorems and rates of convergence for euclidean functionals. *The Annals of Applied Probability*, 4(4):pp. 1057–1073, 1994.
- [45] X. H. Shi, Y. C. Liang, H. P. Lee, C. Lu, and Q. X. Wang. Particle swarm optimization-based algorithms for tsp and generalized tsp. *Information Processing Letters*, 103(5):169–176, 2007.
- [46] H. Somerville. Startups try to find sweet spot in grocery delivery. *San Jose Mercury News*, March 29 2013.
- [47] S. S. Srivastava, S. Kumar, R. C. Garg, and P. Sen. Generalized traveling salesman problem through n sets of nodes. *CORS journal*, 7:97–101, 1969.
- [48] J. M. Steele. Subadditive euclidean functionals and nonlinear growth in geometric probability. *The Annals of Probability*, 9(3):pp. 365–376, 1981.
- [49] J.M. Steele. *Probability Theory and Combinatorial Optimization*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1987.
- [50] K. Suh, T. Smith, and M. Linhoff. Leveraging socially networked mobile ICT platforms for the last-mile delivery problem. *Environmental science & technology*, 46(17):9481–9490, 2012.
- [51] E. Wygonik and A. Goodchild. Evaluating the efficacy of shared-use vehicles for reducing greenhouse gas emissions: a US case study of grocery delivery. In *Journal of the Transportation Research Forum*, volume 51, 2012.

- [52] J. Yang, X. Shi, M. Marchese, and Y. Liang. An ant colony optimization method for generalized tsp problem. *Progress in Natural Science*, 18(11):1417–1422, 2008.

Online supplement to “Household-level economies of scale in transportation”

A Proof of Theorem 6

In this section, we prove that, for any fixed n , we have

$$\left[1 - \prod_{j=1}^n (1 - e^{-k_j/b}) \right]^b e^n \sqrt{k_1 \cdots k_n} \sqrt{b} \rightarrow 0$$

as $k_G \rightarrow \infty$, where

$$b = b(k_G) = \frac{k_G^{n/(n-1)}}{(n^2 \log k_G + \log n)^{\frac{1}{n-1}}}.$$

Taking logarithms, the above statement is equivalent to proving that

$$b \log \left[1 - \prod_{j=1}^n (1 - e^{-k_j/b}) \right] + \frac{n}{2} \log k_G + \frac{1}{2} \log m \rightarrow -\infty,$$

which is accomplished as follows: as $k_G \rightarrow \infty$, we have $k_j/b \rightarrow 0$, this is true regardless of limiting behavior of the particular k_j . A Taylor series expansion shows that

$$\log \left[1 - \prod_{j=1}^n (1 - e^{-k_j/b}) \right] \sim - \prod_{j=1}^n (1 - e^{-k_j/b}) \sim - \frac{k_1 k_2 \cdots k_n}{b^n} = - \left(\frac{k_G}{b} \right)^n$$

and thus we want to prove

$$- \frac{k_G^n}{b^{n-1}} + \frac{n}{2} \log(k_G) + \frac{1}{2} \log b \rightarrow -\infty,$$

as $k_G \rightarrow \infty$. Substituting for $b = \frac{(k_G)^{n/(n-1)}}{(n^2 \log(k_G) + \log n)^{\frac{1}{n-1}}}$, we see that

$$-n^2 \log k_G - \log n + \left[\frac{n}{2} + \frac{n}{2(n-1)} \right] \log k_G - \frac{\log(n^2 \log k_G + \log n)}{2(n-1)}$$

approaches $-\infty$ as $k_G \rightarrow \infty$ as desired.

B Proof of Lemma 9

If $X \sim B(n, p)$, then by definition we have

$$\begin{aligned}
 \mathbf{E}(\sqrt{X}) &= \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i} \sqrt{i} = \sum_{i=1}^n \binom{n}{i} p^i (1-p)^{n-i} \sqrt{i} \\
 \frac{d}{dp} \mathbf{E}(\sqrt{X}) &= \sum_{i=1}^n \binom{n}{i} \left(\frac{i}{p} - \frac{n-i}{1-p} \right) p^i (1-p)^{n-i} \sqrt{i} \\
 &= n \left[1 - \frac{p(n-1)}{1-p} \right] (1-p)^{n-1} + \sum_{i=2}^n \binom{n}{i} \left(\frac{i}{p} - \frac{n-i}{1-p} \right) p^i (1-p)^{n-i} \sqrt{i} \\
 \left. \frac{d}{dp} \mathbf{E}(\sqrt{X}) \right|_{p=0} &= n
 \end{aligned}$$

since the differential terms for $i \geq 2$ are all equal to 0. Thus, the approximation $\mathbf{E}(\sqrt{X}) \approx np$ is nothing more than a first-order approximation evaluated at $p = 0$. To prove that $\mathbf{E}(\sqrt{X}) \sim \sqrt{np}$ as $p \rightarrow 1$, we observe that the series expansion for \sqrt{x} about the point $x = 1$ is given by

$$\sqrt{x} = 1 + \frac{x-1}{2} - \frac{(x-1)^2}{8} + \mathcal{O}(x^3)$$

which says that

$$\mathbf{E}(\sqrt{X}) \approx 1 - \frac{\text{Var}(X)}{8}$$

for any random variable such that $\mathbf{E}(X) = 1$, or equivalently,

$$\mathbf{E}(\sqrt{X}) \approx \sqrt{\mathbf{E}(X)} \left(1 - \frac{\text{Var}(X)}{8} \right) = \sqrt{np} \left[1 - \frac{(1-p)}{8np} \right] \sim \sqrt{np} \text{ as } p \rightarrow 1$$

for any binomial random variable X , which completes the proof.