UC Berkeley

UC Berkeley Previously Published Works

Title

THE NEAREST UNVISITED VERTEX WALK ON RANDOM GRAPHS

Permalink

https://escholarship.org/uc/item/90j8z5bk

Journal

Probability in the Engineering and Informational Sciences, 36(3)

ISSN

0269-9648

Author

Aldous, David J

Publication Date

2022-07-01

DOI

10.1017/s0269964821000115

Copyright Information

This work is made available under the terms of a Creative Commons
Attribution License, availalbe at
https://creativecommons.org/licenses/by/4.0/

Peer reviewed

THE NEAREST UNVISITED VERTEX WALK ON RANDOM GRAPHS

DAVID J. ALDOUS

Department of Statistics, University of California, Berkeley, California 94720, USA
E-mail: aldous@stat.berkeley.edu

We revisit an old topic in algorithms, the deterministic walk on a finite graph which always moves toward the nearest unvisited vertex until every vertex is visited. There is an elementary connection between this cover time and ball-covering (metric entropy) measures. For some familiar models of random graphs, this connection allows the order of magnitude of the cover time to be deduced from first passage percolation estimates. Establishing sharper results seems a challenging problem.

Keywords: deterministic walk, metric entropy, nearest neighbor, random graph

1. INTRODUCTION

Consider a connected undirected graph G on n vertices, where the edges e have positive real lengths $\ell(e)$. Consider an entity—let us call it a robot—that can move at speed 1 along edges. There are many different rules one might specify for how the robot chooses which edge to take after reaching a vertex—for instance, the "random walk" rule, to choose edge e with probability proportional to $\ell(e)$ or $1/\ell(e)$. One well-studied aspect of the random walk is the *cover time*, the time until every vertex has been visited—see Ding *et al.* [7] for references to special examples and surprisingly deep connections with other fields. This article instead concerns what we will call[1] the *nearest unvisited vertex* (NUV) walk, defined as follows. A path of edges has a length, the sum of edge-lengths, and the distance $d(v, v^*)$ between vertices is the length of the shortest path. For simplicity, assume that all such distances are distinct, so the shortest path is unique. Now, the NUV walk is the deterministic walk defined in words by

> after arriving at a vertex, next move at speed 1 along the path to the closest unvisited vertex

---

[1] Confusingly previously called *nearest neighbor*, inconsistent with the usual terminology that neighbors are linked by a single edge, but justifiable by the artifice of extending the given graph to a complete graph via defining each edge (v, v^*) to have length $d(v, v^*)$. But the phrase *nearest neighbor* is used in many other contexts, so the more precise name NUV seems preferable.

851

and continue until every vertex has been visited.[2] In symbols, from the initial vertex v_0, the vertices can be written as $v_0, v_1, v_2, \ldots, v_{n-1}$ in order of first visit:

$$v_i = \arg \min_{v \notin \{v_0, \ldots, v_{i-1}\}} d(v_{i-1}, v), \quad 1 \leq i \leq n - 1 \tag{1}$$

and this walk has length $L = L_{\mathrm{NUV}} = L_{\mathrm{NUV}}(G, v_0) = \sum_{i=1}^{n-1} d(v_{i-1}, v_i)$.

There are several types of question that one can ask about NUV walks.

- The order of magnitude of L for a general graph?
- Sharper estimates of L for specific models of random graphs?
- The structural properties of the NUV path in different contexts?

The first question has been studied in the context of TSP (traveling salesman problem) heuristics and robot motion, and a 2012 survey of the general area, under the name online graph exploration, is given in Megow *et al.* [16].

1.1. Outline of Results

Our first purpose is to record a formalization (Proposition 1) of the basic general relationship between L_{NUV} and ball-covering. This is implicit in two now-classical results: Corollary 2, which compares L_{NUV} to the length L_{TSP} of the shortest path through all n vertices, and Corollary 3, which upper bounds L_{NUV} for n arbitrary points in the unit square with the Euclidean distance. As shown in Section 2, each follows easily from our formalization.

Our main purpose is to point out that the relation with ball-covering enables (in some simple probability models) the order of magnitude of L to be deduced easily from known first passage percolation (FPP) estimates. In Section 4, we study two specific models.

- For the $m \times m$ grid with i.i.d. edge-lengths, Corollary 6 shows that L is indeed $O(m^2)$ rather than larger order.
- For the complete graph on n vertices, with i.i.d. edge-lengths normalized so that the shortest edge at a vertex is order 1, Corollary 7 shows that L is indeed $O(n)$ rather than larger order.

In both of those models, the (first-order) behavior of FFP is well understood, via the *shape theorem* on the two-dimensional grid, and the Yule process approximation on the complete graph model.

A final purpose is to point out that the second and third questions above have apparently never been studied. The NUV rule on a deterministic graph is "fragile" in the sense that small changes in the length of an edge might affect a large proportion of the walk. But it is possible that introducing random edge-lengths might "smooth" the typical properties of the walk on a random graph. We defer further general discussion to Section 5.

2. BASICS

2.1. Relation with Ball-Covering

A basic mathematical observation is that L_{NUV} is related to ball-covering.[3] Given $r > 0$ define $N(r) = N(G, r)$ to be the minimal size of a set $\mathcal{S}$ of vertices such that every vertex

---

[2] This *walk* convention is consistent with random walk cover times; one could alternatively use the *tour* convention that the walk finally returns to its start, consistent with TSP.

[3] And thereby to *metric entropy*—see Section 2.3.

is within distance r from some element of $\mathcal{S}$. In other words, the union over $s \in \mathcal{S}$ of the balls of radii r centered at s covers the entire graph.

PROPOSITION 1:

(i) $N(r) \leq 1 + L_{NUV}/r, \ 0 < r < \infty$.

(ii) $L_{NUV} \leq 2 \int_0^{\Delta/2} N(r) \, dr$ where $\Delta = \max_{v,w} d(v,w)$ is the diameter of the graph.

PROOF: Inequality (i) is almost obvious. As at (1), write the vertices as $v_0, v_1, v_2, \ldots, v_{n-1}$ in order of first visit by the NUV walk, and say v_i has rank i. Write $\zeta(v_i) = \sum_{j=0}^{i-1} d(v_j, v_{j+1})$ for the length of the walk up to v_i. Select vertices $(z(k), 0 \leq k \leq k^* - 1)$ along the walk by selecting the first vertex at distance $> r$ along the walk after the previous selected vertex. That is, $z(k) = v_{I(k)}$ where $I(0) = 0$ and for $k \geq 0$

$$I(k+1) = \min\{i > I(k) : \zeta(v_i) - \zeta(v_{I(k)}) > r\}$$

until no such i exists. By construction, every vertex is within distance r of some z, and the number k^* of selected vertices is at most $1 + L_{\text{NUV}}/r$. This establishes (i).

For inequality (ii), write $D(v_i) = d(v_i, v_{i+1})$ for the length of the *path* (which may encompass several edges) from the rank-i vertex to the rank-$(i+1)$ vertex, and $D(v_{n-1}) = 0$. The argument rests upon the following simple observation, as illustrated in Figure 1. Fix a vertex v^* and a real $r > 0$, and consider the set of vertices within distance r from v^*:

$$B(v^*, r) := \{v : d(v, v^*) \leq r\}.$$

Consider the vertex $\bar{v}$ of highest NUV-rank within $B(v^*, r)$. When the NUV walk first visits $v_i \in B(v^*, r)$ with $v_i \neq \bar{v}$, there is then some first unvisited vertex $\tilde{v}$ on the minimum-length path from v_i to $\bar{v}$, and so

$$D(v_i) \leq d(v_i, \tilde{v}) \leq d(v_i, \bar{v}) \leq 2r$$

the final inequality using the triangle inequality via v^*. We conclude that

$$D(v) \leq 2r \text{ for all } v \in B(v^*, r) \text{ except perhaps one vertex.} \tag{2}$$

Now by considering a set, say $S(r)$, containing $N(r)$ vertices, such that every vertex is within distance r from some element of $S(r)$, inequality (2) implies

$$\text{the number of vertices } w \text{ with } D(w) > 2r \text{ is at most } N(r). \tag{3}$$

Because $D(w)$ is bounded by the graph diameter Δ, for a uniformly random vertex J we have

$$\begin{aligned} L_{\text{NUV}} &= n\mathbb{E}[D(J)] \\ &= n\int_0^{\Delta} P(D(J) > r) \, dr \\ &= \int_0^{\Delta} (\text{number of vertices } w \text{ with } D(w) > r) \, dr \\ &\leq \int_0^{\Delta} N(r/2) \, dr \end{aligned}$$
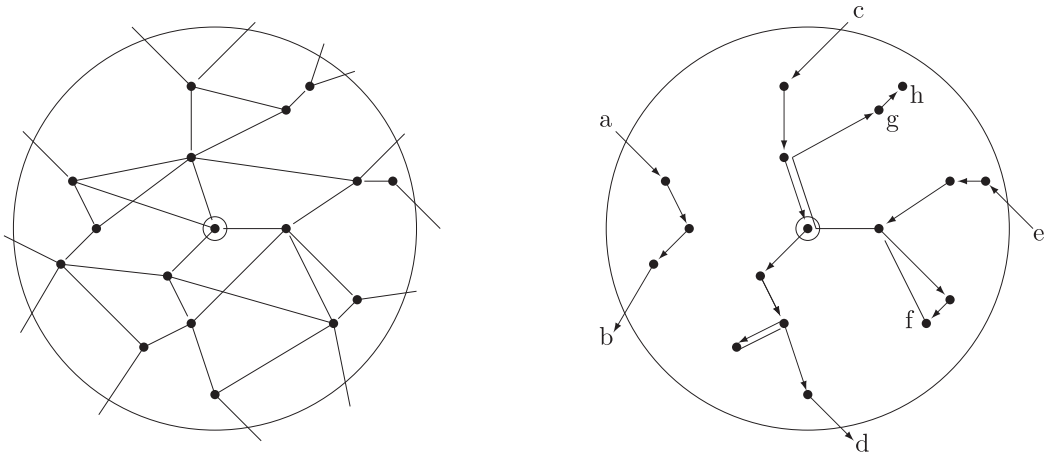
which is equivalent to (ii). ■

FIGURE **1.** Illustration of the proof of (2). The left panel shows the subgraph within a radius-r ball. The NUV walk must consist of one or several excursions within the ball. These excursions depend on the configuration outside the ball, and the right side shows one possibility. The first excursion enters via edge a and exits via edge b. The second excursion enters via edge c and exits via edge d, en route backtracking across one edge. The third excursion enters via edge e and proceeds to vertex f; at that time, only vertices g, h within the ball are unvisited, and the next step of the walk is a path going via three previously visited vertices to reach g and then h. The next step from h, not shown, might be very long, depending on whether nearby vertices outside the ball have all been visited. Arrowheads indicate the end of a step of the NUV walk, that is, the edge by which the vertex is first entered.

REMARKS: *The simple formulation of Proposition 1 is more implicit than explicit in the literature we have found. Part (i) is a less sharp version of a more complex lemma used in Rosenkrantz et al. [19] to prove Corollary 2 below. In the context of TSP or robot exploration heuristics, the NUV algorithm is typically (e.g. [11,13]) mentioned only briefly before continuing to better algorithms. From an algorithmic viewpoint, calculating $N(r)$ on a general graph is not simple, so part (ii) of Proposition 1 is not so relevant, but as we see in Section 4, it is very helpful in providing order-of-magnitude bounds for familiar models of random networks.*

2.2. Two Classical Results

Two classical results follow readily from the formulation of Proposition 1. Write $L_{\text{TSP}} = L_{\text{TSP}}(G, v_0)$ for the length of the shortest *walk* starting from v_0 and visiting every vertex.[4] So $L_{\text{NUV}} \geq L_{\text{TSP}}$ and it is natural to ask how large the ratio can be. This was answered in Rosenkrantz *et al.* [19].

COROLLARY 2: *Let $a(n)$ be the maximum, over all connected n-vertex graphs with edge-lengths and all initial vertices, of the ratio L_{NUV}/L_{TSP}. Then, $a(n) = O(\log n)$.*

---

[4] The convention that TSP refers to a *tour* has the virtue that the length is independent of the starting vertex. But the latter is not true for the NUV tour.

PROOF: The argument for Proposition 1(i) is unchanged if we use the TSP path instead of the NUV path, so in fact gives the stronger result $N(r) \leq 1 + L_{\mathrm{TSP}}/r$, $0 < r < \infty$. Now apply Proposition 1(ii) and note that $\Delta \leq L_{\mathrm{TSP}}$, so

$$L_{\mathrm{NUV}} \leq 2 \int_0^{L_{\mathrm{TSP}}/2} \min(n, 1 + L_{\mathrm{TSP}}/r) \, dr \leq 2L_{\mathrm{TSP}} + 2L_{\mathrm{TSP}} \log n$$

the second inequality by splitting the integral at $r = L_{\mathrm{TSP}}/n$.   ∎

There are examples to show that the $O(\log n)$ bound cannot be improved—see Johnson and Papadimitriou [13], Hurkens and Woeginger [11], Hougardy and Wilde [10], and Rosenkrantz *et al.* [19]. As noted in the elementary expository article Aldous [2], in constructing such an example the key point is to make the bound in (2) be tight, in the sense

for appropriate values of r with $1 \ll L_{\mathrm{TSP}}/r \ll n$, there are distinguished vertices separated by distance r along the TSP path such that the NUV path from one to the next is order r.

Hurkens and Woeginger [11] show that one can make such examples be planar, embedded in the plane with edge-lengths as Euclidean length, and edge-lengths constrained to a neighborhood of 1. But such constructions seem very artificial.

Here is the second classical result. See Steele [20] for one proof and the early history of this result.

COROLLARY 3: *There is a constant A such that, for the complete graph on n arbitrary points in the unit square, with Euclidean lengths,*

$$L_{NUV} \leq An^{1/2}.$$

Note this implies the well-known corresponding result $L_{\mathrm{TSP}} \leq An^{1/2}$.

PROOF: By ball-covering in the continuum unit square, there is a numerical constant C such that $N(r) \leq C/r^2$, and so Proposition 1(ii) gives

$$L_{\mathrm{NUV}} \leq 2 \int_0^{\sqrt{1/2}} \min(n, C/r^2) \, dr \leq 4C^{1/2}n^{1/2}.$$

∎

2.3. The Order of Magnitude Question

What is the size of L_{NUV} for a *typical* graph? That is a very vague question, but let us attempt a discussion anyway. For this informal discussion, it is convenient to scale distances so that the typical distance from a vertex to its closest neighbor is order 1, and therefore, L_{NUV} is at least order n. Examples mentioned above show that L_{NUV} can still be as large as order $n \log n$, but intuition suggests that, for natural examples, L_{NUV} is of order n rather than larger order. For this, it is certainly necessary, but not sufficient, that the length L_{MST} of the minimum spanning tree (MST)[5] is $O(n)$. Proposition 1(ii) provides a quantitative criterion: it is sufficient that $N(r)/n$ is order $r^{-\alpha}$ for some $\alpha > 1$ over $1 \ll r \ll \Delta$. Intuitively,

---

[5] Recall $L_{\mathrm{MST}} \leq L_{\mathrm{TSP}} \leq 2L_{\mathrm{MST}}$.

this corresponds to "dimension >1", where dimension is measured by metric entropy,[6] as illustrated in the examples in Section 4.

2.4. Other Questions in the Deterministic Setting

It is not clear what other results might hold for general graphs G. One can ask about the variability of $L_{\mathrm{NUV}}(G, v)$ as v varies. Clearly, it can be arbitrarily concentrated, e.g., on the complete graph with edge-lengths arbitrarily close to 1. On the other hand, consider the linear graph G_n on vertices $\{0, 1, \ldots, n-1\}$ with slowly decreasing edge-lengths $\ell(i-1, i) = 1 - i/n^2$. Here, there is a factor of 2 variability in $L_{\mathrm{NUV}}(G, v)$ as v varies. We do not see any easy example with large variability, prompting the following question.

OPEN PROBLEM 4: *Is $\max_v L_{NUV}(G, v)/\min_v L_{NUV}(G, v)$ bounded over all finite graphs G?*

In this context, it is perhaps more natural to extend the NUV walk to a *tour* which finally returns to its start. Note that in the linear graph example above, $|L_{\mathrm{NUV}}(G, v) - L_{\mathrm{NUV}}(G, v')|$ is small for adjacent vertices (v, v'), so one can ask whether there is a general bound for some average of $|L_{\mathrm{NUV}}(G, v) - L_{\mathrm{NUV}}(G, v')|$ over nearby vertex-pairs (v, v').

One can also consider the overlap of edges used in walks from different starts. Note that if two vertices are each other's nearest neighbor, then every NUV walk uses their linking edge. One can ask, for the two walks started at arbitrary different vertices, how small can be the proportion of time spent on edges used by both walks, though we hesitate to formulate a conjecture.

2.5. The Three Levels of Randomness

Introducing randomness leads to different questions. There are three ways that one can introduce randomness. One can simply randomize the starting vertex. This suggests the following conjecture, modifying Open Problem 4.

CONJECTURE 5: *The ratio s.d.$(L_{NUV}(G, V))/\mathbb{E}L_{NUV}(G, V)$, where the initial vertex V is uniform random, is bounded over all finite graphs.*

A second level of randomness is to start with a given deterministic G, but then consider the random graph $\mathcal{G}$ in which the edge-lengths $\ell(e)$ are replaced by independent random lengths $\ell^*(e)$ with the Exponential(mean $\ell(e)$) distribution. So here, we have a random variable $\mathcal{L}^*(G) = L_{\mathrm{NUV}}(\mathcal{G}, V)$ where again the initial vertex V is uniform random. In this model of random graphs $\mathcal{G}$, results of Aldous [1] for FFP say that the percolation time is weakly concentrated[7] around its mean provided no single edge contributes non-negligibly to the total time. So one can ask whether a similar result holds for $\mathcal{L}^*(G)$.

The third level of randomness involves more specific models of random graphs, which we will consider in the next sections.

---

[6] The reader may be more familiar with metric entropy involving *small* balls for continuous spaces, but it is equally relevant in our context of large balls, as used, for instance, in defining the fractal dimension of subsets of $\mathbb{Z}^d$.

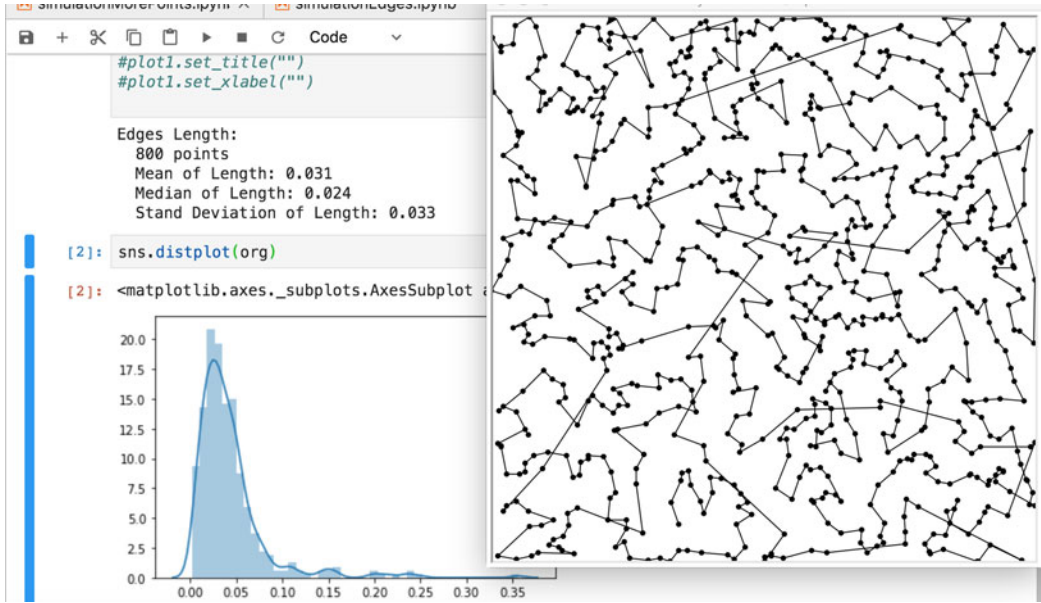[7] As in the weak law of large numbers.

FIGURE 2. A NUV walk through 800 random points in the unit square and the histogram of step lengths.

TABLE 1. Simulation data for lengths L_n^* in the random points in the unit square model. Simulations and data in this model by Yechen Wang

n	$\mathbb{E}L_n^*$	$n^{-1/2}\mathbb{E}L_n^*$	s.d.(L_n^*)
100	9.05	0.91	0.41
200	12.78	0.90	0.54
400	18.06	0.90	0.54
800	25.54	0.90	0.49

3. RANDOM POINTS IN THE SQUARE

One very special model of random graph is to take the complete graph on n random (i.i.d. uniform) points in the unit square, with Euclidean edge-lengths. Figure 2 shows a realization of the corresponding NUV walk with $n = 800$ random points, and Table 1 shows some simulation data for the lengths L_n^* of the NUV walk (see discussion below). The qualitative behavior seen in simulations corresponds to intuition: the walk starts to traverse through most (but not all) vertices in any small region, goes through different regions as some discrete analog of a space-filling curve, and near the end has to capture missed patches and the remaining isolated unvisited vertices via longer steps across already-explored regions. Indeed in Figure 2, we see that the actual behavior of the walk within a medium-sized ball is like the sketch in Figure 1, with several different excursions.

   The lack of scaling for the s.d. may seem surprising but is understandable as follows. To adhere to our scaling convention (the distance to nearest neighbor is order 1), we should take the square to have area n and write $L_n = n^{1/2}L_n^*$ for the length of the NUV walk. Intuition, thinking of L_n as the sum of n order-1 lengths, suggests that there are limit

constants

$$c := \lim_n n^{-1} L_n = \lim_n n^{-1/2} L_n^*; \quad \sigma := \lim_n n^{-1/2} \text{s.d.}(L_n) = \lim_n \text{s.d.}(L_n^*). \qquad \textbf{(4)}$$

Our small-scale simulation data suggests that this holds in the present model with $c \approx 0.9$ and $\sigma \approx 0.5$. How generally this holds is a natural question, and we defer further discussion to Section 5.

Corollary 3 implies $\mathbb{E} L_n \leq An$, which is all that we know rigorously. But there are many questions that one can ask. As well as the limits (4) one might conjecture that there are concentration bounds and a Gaussian limit for $n^{-1/2}(L_n - \mathbb{E} L_n)$. For TSP length, the existence of a limit constant is known via subadditivity arguments [21,23] and concentration via now-classical Talagrand arguments, and for MST length, the Gaussian limit is also known by martingale arguments [15]. Alas, it seems hard to find any rigorous such arguments for the NUV walk. One might also bear in mind that, for the *random walk* cover time problem, the two-dimensional case is the hardest to analyze sharply, so this might also hold for the NUV walk.

In any of our models, by considering the length as $L_n(G_n, V_n)$ for a uniform random starting vertex V_n, we can consider the variance decomposition

$$\text{var } L_n = \text{var } \mathbb{E}(L_n | G_n) + \mathbb{E} \text{var}(L_n | G_n)$$

where the first term represents the variability due to the random graph and the second term represents the variability due to the starting vertex. In simulations of the present model, for $n = 100$ the two terms are roughly equal. Figure 3 superimposes the NUV walks from three different starts, in a realization of the present model, giving some impression of the extent of overlap.

4. RELATION WITH FPP

For graphs with i.i.d. random edge-lengths, one can seek to find the correct order of magnitude of L_{NUV} by combining Proposition 1(ii) with known FPP results. Here is the basic example.

4.1. The Two-Dimensional Grid

Consider the $m \times m$ grid, that is, the subgraph of the Euclidean lattice $\mathbb{Z}^2$, and assign i.i.d. edge-lengths $\ell(e) > 0$ to make a random graph G_m. Because the shortest edge-length at a given vertex is $\Omega(1)$, clearly L_{NUV} is $\Omega(m^2)$.

COROLLARY 6: *For the two-dimensional grid model G_m above, the sequence $(m^{-2} L_{NUV}(G_m),$ $m \geq 2)$ is tight.*

We conjecture that, in fact, $m^{-2} L_{\text{NUV}}(G_m)$ converges in probability to a constant, but we do not see any simple argument. Table 2 shows simulation data, where $\ell(e)$ has the Exponential(1) distribution.

PROOF: For a vertex v of G_m, write $B(v, r)$ for the random set of vertices v' with $d(v, v') \leq r$, and write $D(v, r)$ for the nonrandom set of vertices v' with the Euclidean distance $\|v - v'\| \leq r$. Standard results for FPP on $\mathbb{Z}^2$ going back to Kesten [14] (see [4] Thm. 3.41 for
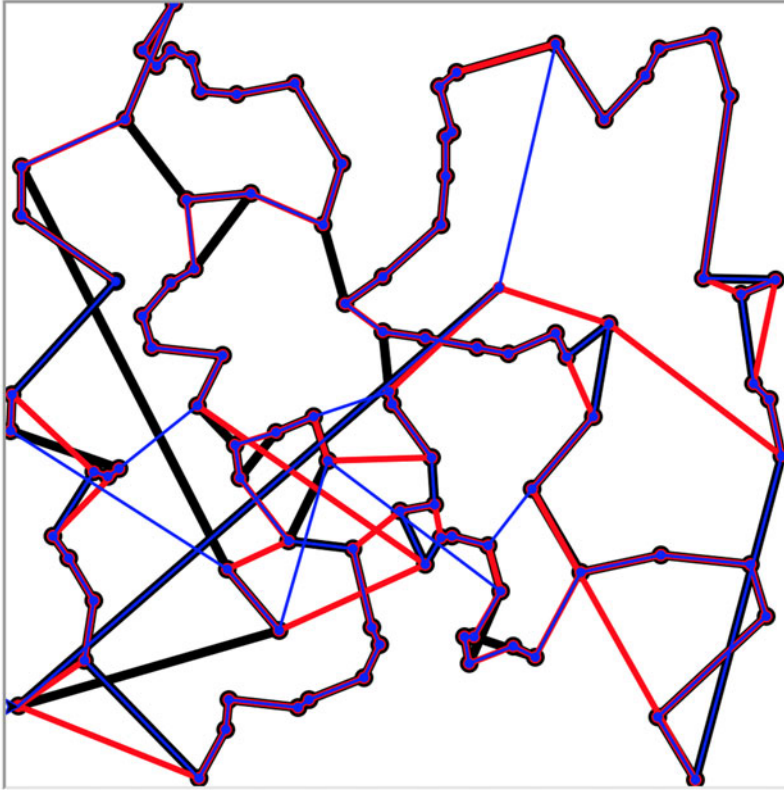
FIGURE **3.** Three different starts for the NUV walk on 100 points in the square.

recent discussion) imply that there exist constants c_1, c_2, c_3 (depending on the distribution of $\ell(e)$) such that

$$\mathbb{P}(D(v,r) \nsubseteq B(v, c_1 r)) \le c_2 \exp(-c_3 r), \quad 0 < r < \infty. \tag{5}$$

The remainder of the proof is conceptually straightforward. Given large m and r, there is a set $S(m,r)$ of at most $a_1 m^2/r^2$ vertices of G_m such that $\cup_{v \in S(m,r)} D(v,r)$ covers G_m, and note that $D(v,r)$ contains at most $a_2 r^2$ vertices; here, a_1 and a_2 are absolute constants. By Markov's inequality and (5), the probability of the event

$$\text{the number of } v \text{ in } S(m,r) \text{ such that } D(v,r) \nsubseteq B(v, c_1 r)$$

$$\text{exceeds a given } s > 0 \tag{6}$$

is at most $a_1 m^2 r^{-2} c_2 \exp(-c_3 r)/s$. Apply this with $s = m^2 r^{-2} \exp(-c_3 r/2)$. Now, define a vertex-set $S^+(m,r)$ as

the union of $S(m,r)$ and all the vertices in all the discs $D(v,r)$ with $v \in S(m,r)$ and $D(v,r) \nsubseteq B(v, c_1 r)$.

Outside the event (6), we have that $\cup_{v \in S^+(m,r)} D(v,r)$ covers G_m, and $S^+(m,r)$ has cardinality at most

$$n_m(r) := a_1 m^2/r^2 + s a_2 r^2 = a_1 m^2/r^2 + a_2 m^2 \exp(-c_3 r/2).$$

TABLE **2.** Simulation data for lengths $L(G_m)$ in the grid model

$n = m^2$	$\mathbb{E}L(G_m)$	$n^{-1}\mathbb{E}L(G_m)$	s.d.$(L(G_m))$	$n^{-1/2}$ s.d.$(L(G_m))$
100	66.2	0.66	7.67	0.77
400	259	0.65	14.8	0.74
900	576	0.64	17.0	0.57

So, we have shown

$$\mathbb{P}(N(G_m, r) > n_m(r)) \leq a_1 c_2 \exp(-c_3 r/2). \tag{7}$$

This holds for fixed r, but because $N(G_m, r)$ and $n_m(r)$ are decreasing in r, we have the inclusion of events, for $j = 1, 2, \ldots$

$$\{N(G_m, r) > n_m(r - 1) \text{ for some } j \leq r \leq j + 1\} \subseteq \{N(G_m, j) > n_m(j)\}$$

Applying (7) and summing over j,

$$\mathbb{P}(N(G_m, r) > n_m(r - 1) \text{ for some } r > r_0) \leq \Phi(r_0)$$

where Φ depends on the distribution of $\ell(e)$ but not on m, and

$$\Phi(r_0) \downarrow 0 \quad \text{as } r_0 \to \infty. \tag{8}$$

Noting that $n_m(r)/m^2$ does not depend on m and

$$\psi(r_0) := \int_{r_0}^{\infty} n_m(r - 1)/m^2 \, dr \to 0 \quad \text{as } r_0 \to \infty$$

and $N(G_m, r) \leq m^2$ we have, for all $r_0 > 0$,

$$\mathbb{P}\left(\int_0^{\infty} m^{-2} N(G_m, r) \, dr > r_0 + \psi(r_0)\right) \leq \Phi(r_0)$$

which, together with (8) and Proposition 1(ii), implies tightness of the sequence $(m^{-2}L_{\text{NUV}}(G_m), \; m \geq 2)$. ∎

The central point is that the argument depends only on some bound like (5), which one expects to hold very generally in FPP-like settings in dimension >1. For instance, FPP on a large family of connected random geometric graphs is studied in Hirsch *et al.* [9] and it seems plausible that results from that topic can be used to prove that L_{NUV} is $O(n)$ on such n-vertex graphs.

The next example is infinite dimensional, and the bound (10) below will be the analog of the bound (5) above.

4.2. The Mean-Field Model of Distance

Take the complete graph on n vertices and assign to edges i.i.d. random weights with Exponential (mean n) lengths. This "mean-field model of distance" G_n turns out to be surprisingly tractable, because the smallest edge-lengths $0 < \ell_1 < \ell_2 < \cdots$ at a given vertex are distributed (in the $n \to \infty$ limit) as the points of a rate-1 Poisson point process on $(0, \infty)$, and as regards short edges, the graph is locally tree-like. A now classical result of Frieze

TABLE **3.** Simulation data for lengths L_n in the mean-field model

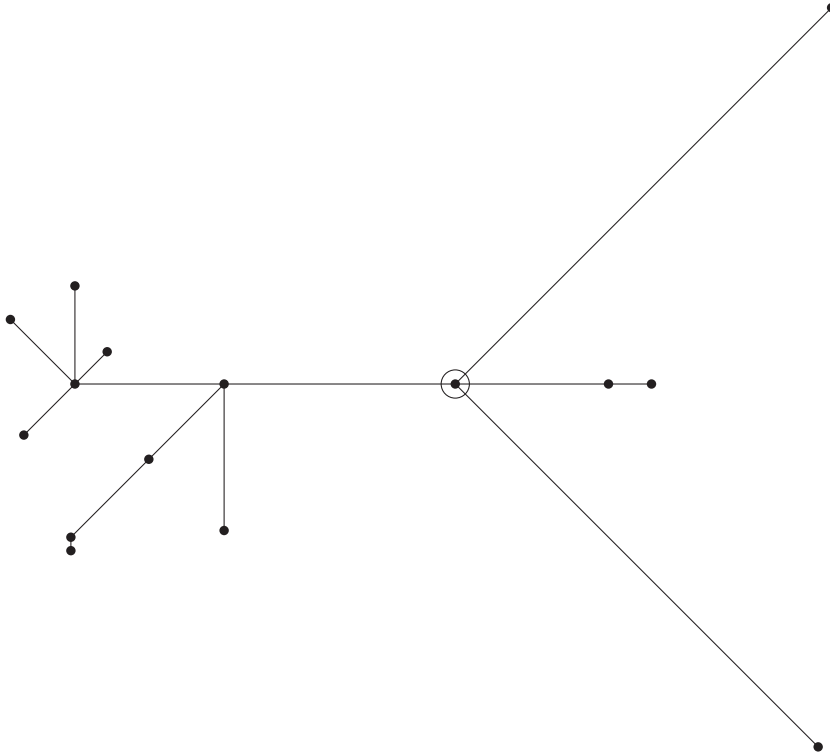| n	$\mathbb{E}L_n$	$n^{-1}\mathbb{E}L_n$	s.d.(L_n)	$n^{-1/2}$ s.d.(L_n)
100	209	2.09	22	2.2
400	865	2.14	41	2.1
900	1954	2.17	57	1.9

FIGURE **4.** Mean-field model: vertices and edges within a ball of radius 4 in a realization, illustrating the local tree-like property. Edges to vertices outside the ball not shown.

[8] proves that the length $L_{\mathrm{MST}}^{(n)}$ of the MST in this model satisfies $\mathbb{E}L_{\mathrm{MST}}^{(n)} \sim \zeta(3)n$. A later remarkable result of Wästlund [22], formalizing the ideas of Mézard and Parisi [17], shows that the expected length of the TSP path in this model is asymptotically cn for an explicit constant $c = 2.04\ldots$. Might it be possible to get a similar explicit result for the NUV length? Corollary 7 below gives the correct order of magnitude by essentially the same method as above for Corollary 6. Table 3 gives some simulation results.

As in the previous models, we expect limits of the form

$$c := \lim_n n^{-1}\mathbb{E}L_n, \quad \sigma := \lim_n n^{-1/2}\,\mathrm{s.d.}(L_n)$$

and Table 3 is loosely consistent with that (Figures 4 and 5).

As in Section 3, by considering the length as $L_n(G_n, V_n)$ for a uniform random starting vertex V_n, we can consider the variance decomposition

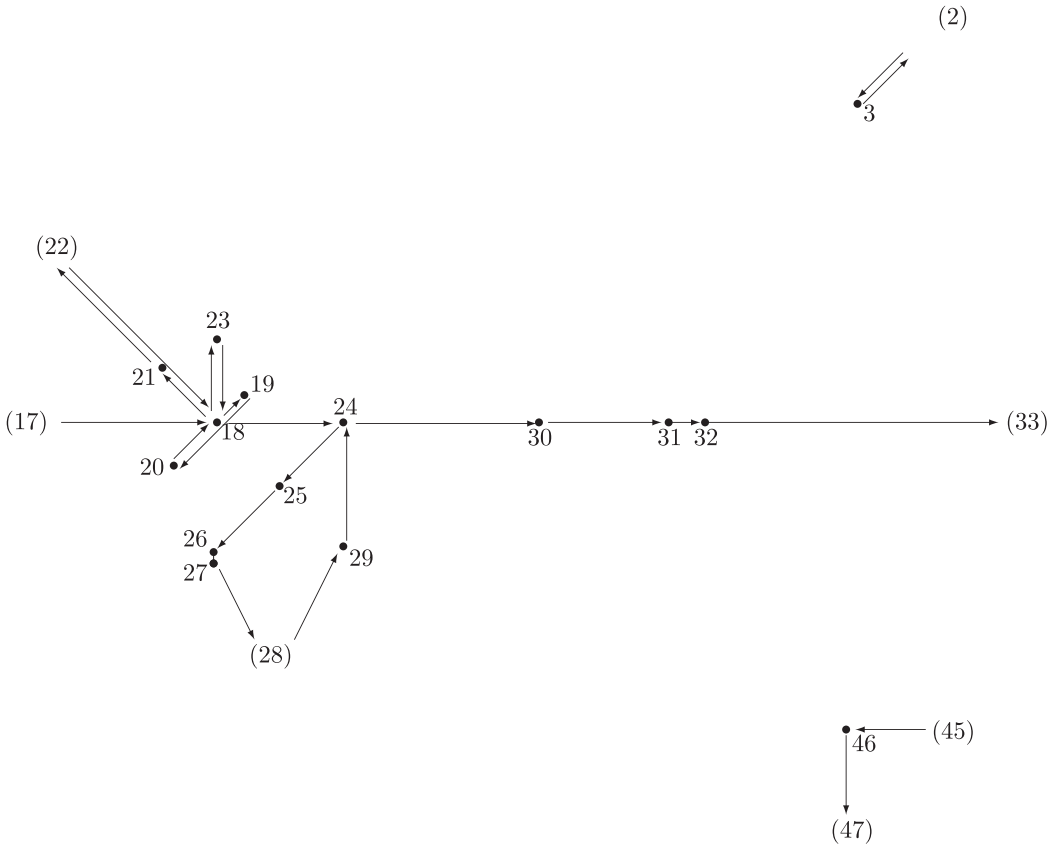$$\mathrm{var}\, L_n = \mathrm{var}\,\mathbb{E}(L_n|G_n) + \mathbb{E}\,\mathrm{var}(L_n|G_n)$$

FIGURE **5.** Mean-field model: in the Figure 4 realization, the NUV walk within the ball and entrance-exit edges. Vertices numbered according to the order in an NUV walk started outside the ball, with vertices outside the ball in parentheses.

where the first term represents the variability due to the random graph and the second term represents the variability due to the starting vertex. In simulations with $n = 100$, the former variance term is around 30 times larger than the second term, consistent with the general conjectures (Section 2.5) that the initial state v typically has little influence on $L_{\text{NUV}}(G, v)$.

We now prove the $O(n)$ upper bound in this model.

COROLLARY 7: *For the mean-field model of distance G_n, the sequence $(n^{-1}L_{NUV}(G_n), \; n \geq 2)$ is tight.*

To prove this, we first record a simple estimate.

LEMMA 8: *Let Z_p have the Geometric(p) distribution. Let Z_p^* coincide with $Z_p - 1$ outside an event A. Let H be a random subset of $[n] = \{1, 2, \ldots, n\}$ distributed uniformly on size Z_p^* subsets of $[n]$. Then,*

$$\mathbb{P}(A^c \text{ and } H \cap [s] = \emptyset) \leq \frac{p}{1 - e^{-s/n}}.$$

PROOF: It is standard (by comparing sampling with and without replacement) that

$$\mathbb{P}(H \cap [s] = \emptyset | Z_p^* = i) \leq \exp(-si/n).$$

So

$$\mathbb{P}(A^c \text{ and } H \cap [s] = \emptyset) \leq \sum_{i \geq 0} p(1-p)^i \exp(-si/n)$$

$$= \frac{p}{1 - (1-p)e^{-s/n}}$$

$$\leq \frac{p}{1 - e^{-s/n}}.$$

$\blacksquare$

As before, for a vertex $v \in [n] = \{1, 2, \ldots, n\}$, write $B_n(v, r) = \{v' : d(v, v') \leq r\}$ for the ball of radius r in G_n. Conceptually, we want to consider balls around s randomly chosen vertices, but by symmetry, this is equivalent to using the first s vertices, which is notationally simpler. So define the vertex-set

$$C_n(s, r) = \text{complement of} \bigcup_{i \leq s} B(i, r)$$

and then by appending to $[s]$ every vertex in $C_n(s, r)$,

$$N(G_n, r) \leq s + |C_n(s, r)|, \quad 1 \leq s \leq n. \tag{9}$$

Recall (see, e.g., [18] Sec. 6.1.3) the *standard Yule process* $(Y(r), 0 \leq r < \infty)$ for which $Y(r)$ has exactly the Geometric(e^{-r}) distribution. The $n \to \infty$ limit distribution of the process $(|B_n(v, r)|, 0 \leq r < \infty)$ over a fixed r-interval is well known to be this standard Yule process (This is part of the theory in Aldous and Steele [3] surrounding the PWIT.[8]). Choosing $r_1 = \frac{1}{3} \log n$, so that $\exp(r_1) = n^{1/3}$, it is not difficult to use the natural coupling of the two processes to quantify this convergence to show

the distribution of $(|B_n(v, r)|, 0 \leq r \leq r_1)$ agrees with the distribution of $(Y(r), 0 \leq r \leq r_1)$ outside an event $A_n(v)$ of probability $\delta_n = O(n^{-1/4}) \to 0$ as $n \to \infty$.

For a vertex $v \in [s + 1, n]$, and for $r \leq r_1$,

$$\mathbb{P}(A_n^c(v) \text{ and } v \in C_n(s, r)) = \mathbb{P}(A_n^c(v) \text{ and } B_n(v, r) \cap [s] = \emptyset)$$

$$\leq \frac{e^{-r}}{1 - e^{-s/(n-1)}} \tag{10}$$

the inequality from Lemma 8 applied to $[n] \setminus \{v\}$. Apply this with

$$s = s_n(r) := -(n-1) \log(1 - e^{-r/2})$$

which is the solution of $e^{-r/2} = 1 - e^{-s/(n-1)}$, so

$$\mathbb{P}(A_n^c(v) \text{ and } v \in C_n(s_n(r), r)) \leq e^{-r/2}.$$

_____

[8] Poisson Weighted Infinite Tree.

Summing over v, from (9) we can write, for $r \le r_1$,

$$N(G_n, r) \le s_n(r) + X_n + Y_n(r) \quad \text{where } \mathbb{E}X_n \le n\delta_n \text{ and } \mathbb{E}Y_n(r) \le ne^{-r/2}.$$

Applying Markov's inequality separately to the two terms on the right side of the first inequality above,

$$\mathbb{P}(N(G_n, r) > s_n(r) + n\delta_n^{1/2} + ne^{-r/4}) \le \delta_n^{1/2} + e^{-r/4}, \quad r \le r_1.$$

As in the proof of Corollary 6, we can use monotonicity to convert this fixed-r bound to a uniform bound over a "medium" interval $r_0 \le r \le r_1$:

$$\mathbb{P}(N(G_n, r) > s_n(r-1) + n\delta_n^{1/2} + ne^{-(r-1)/4} \quad \text{for some } r_0 \le r \le \lfloor r_1 \rfloor) \le \delta_n^{1/2} \log n + 5e^{-r_0/4}.$$

Because $s_n(r) \approx ne^{-r/2}$ over the interval of interest,

$$n^{-1} \int_{r_0}^{r_1} (s_n(r-1) + n\delta_n^{1/2} + ne^{-(r-1)/4}) \, dr \le Ke^{-r_0/4} + \delta_n^{1/2} \log n$$

for some constant K, and so

$$\mathbb{P}\left(n^{-1} \int_{r_0}^{r_1} N(G_n, r) \, dr > Ke^{-r_0/4} + \delta_n^{1/2} \log n\right) \le \delta_n^{1/2} \log n + 5e^{-r_0/4}.$$

For the tail of the integral, the diameter Δ of G_n is known [12] to be asymptotically $3 \log n$ and so by monotonicity of $N(r)$

$$n^{-1} \int_{r_1}^{\Delta} N(G_n, r) \, dr = O(n^{-1} \cdot N(G_n, r_1) \cdot \log n) \to 0 \text{ in probability.}$$

We will show below that

$$\mathbb{E}N(G_n, r_1) = O(n^{11/12}). \tag{11}$$

Because $\delta_n^{1/2} \log n \to 0$ and $n^{-1}N(G_n, r) \le 1$ for $r \le r_0$, these bounds establish tightness of the sequence

$$n^{-1} \int_0^{\Delta/2} N(G_n, r) \, dr, \quad n \ge 2$$

which by Proposition 1(ii) implies that the sequence $(n^{-1}L_{\text{NUV}}(G_n), n \ge 2)$ is tight.

To outline a proof of (11), take expectation in (9) to get

$$\mathbb{E}N(G_n, r_1) \le s + n\mathbb{P}(v \in C_n(s, r_1)), \ 1 \le s \le n \tag{12}$$

for a vertex $v \in [s+1, n]$. We will use this with $s = n^{3/4}$. Conditional on $|B_n(v, r_1)| = \beta$ we have, in order of magnitude,

$$\mathbb{P}(v \in C_n(s, r_1)) \asymp (1 - \beta/n)^s \asymp \exp(-\beta s/n).$$

Now, the distribution of β is asymptotically Exponential with mean $e^{r_1} = n^{1/3}$, so by integrating over β, the unconditional probability becomes

$$\mathbb{P}(v \in C_n(s, r_1)) \asymp \frac{n^{-1/3}}{n^{-1/3} + s/n} \asymp n^{-1/12}.$$

Combining with (12) gives (11).

5. FINAL REMARKS

5.1. Analogy with the MST

As an algorithm, the NUV walk is somewhat similar to the greedy (Prim's) algorithm for the MST (minimum spanning tree), in that both grow a connected graph one edge at a time. Recall that for the MST, there is an intrinsic criterion for whether a given edge e is in the MST

> e is in the MST if and only if there is no alternative path between the endpoints of e, all of whose edges are shorter than $\ell(e)$.

This enables a martingale proof [15] of the central limit theorem for the length L_{MST} within the Euclidean model (complete graph on random points in the square) which we will discuss in Section 3. There is no such intrinsic criterion for the NUV walk, so to improve the order-of-magnitude result (Corollary 3 below) for L_{NUV} in that model one would need some other kind of control over the geometry of the set of points visited before each step. Also, as noted in Section 4.2, in the "mean-field model of distance", the exact asymptotic constants for the lengths of the TSP tour and the MST are known: can they also be calculated for the NUV walk?

5.2. Local Weak Convergence

Our results are conceptually merely consequences of Proposition 1, and further progress would require some other technique. One possible general approach is via local weak convergence [3,5]. Our three specific models each have local weak convergence limits (complete graph on a Poisson point process on the infinite plane with the Euclidean distance; i.i.d. edge-lengths on the infinite lattice; and the PWIT), and intuitively, the conjectured limits $\lim_n n^{-1}\mathbb{E}L_n$ are the mean step-lengths in an appropriately defined NUV walk on the limit infinite graph. Can this intuition be made rigorous?

In fact, one expects the limits in our models to be *collections* of disjoint doubly infinite walks which cover the infinite graph. This relates to a longstanding folklore problem: for the NUV walk on the complete-graph Poisson point process on the infinite plane, estimate the number of never-visited vertices in the radius-r ball, as $r \to \infty$ (see Bordenave *et al.* [6] for discussion).

5.3. Restrictions on Local Behavior of Paths

For another possible direction of analysis, consider the Figure 1 sketch of one possible trajectory for the NUV path through a given ball. In general, there will be many possible trajectories, depending on the graph outside the ball, but can one find restrictions on the possibilities, extending the obvious restriction:

> if two vertices are each other's nearest neighbor, then every NUV walk, after visiting the first, immediately visits the second.

Intuitively, for $1 \ll r_1 \ll r_2$, given the subgraph in the ball $B(v^*, r_2)$, in a random graph, there will typically be only a few possibilities for the NUV trajectory within $B(v^*, r_1)$.

5.4. Variance of L_{NUV}?

A final issue involves the variance of L_{NUV} in random graph models. We expect order n "each other's nearest neighbor" pairs, and then, the randomness of edge-lengths suggests

that the contribution to variance of L_{NUV} from these edges alone must be at least order n (in our conventional scaling). However, our small-scale simulation results in Tables 2 and 3 cast some doubt on this conjectured lower bound.

Competing Interests

The authors declare no conflict of interest.

References

1. Aldous, D.J. (2016). Weak concentration for first passage percolation times on graphs and general increasing set-valued processes. *ALEA. Latin American Journal of Probability and Mathematical Statistics* 13(2): 925–940.
2. Aldous, D.J. (2021). Exploring endless space. In preparation.
3. Aldous, D.J. & Steele, J.M. (2004). The objective method: Probabilistic combinatorial optimization and local weak convergence. In H. Kesten (ed.), *Probability on Discrete Structures. Encyclopaedia of Mathematical Sciences*, vol. 110. Berlin: Springer, pp. 1–72.
4. Auffinger, A., Damron, M., & Hanson, J. (2017). *50 years of first-passage percolation*, vol. 68. Providence, RI: American Mathematical Society University Lecture Series.
5. Benjamini, I. & Schramm, O. (2001). Recurrence of distributional limits of finite planar graphs. *Electronic Journal of Probability* 6: 1–13.
6. Bordenave, C., Foss, S., & Last, G. (2011). On the greedy walk problem. *Queueing Systems* 68: 333–338.
7. Ding, J., Lee, J.R., & Peres, Y. (2012). Cover times, blanket times, and majorizing measures. *Annals of Mathematics* 175(3): 1409–1471.
8. Frieze, A.M. (1985). On the value of a random minimum spanning tree problem. *Discrete Applied Mathematics* 10(1): 47–56.
9. Hirsch, C., Neuhäuser, D., Gloaguen, C., & Schmidt, V. (2015). First passage percolation on random geometric graphs and an application to shortest-path trees. *Advances in Applied Probability* 47(2): 328–354.
10. Hougardy, S. & Wilde, M. (2015). On the nearest neighbor rule for the metric traveling salesman problem. *Discrete Applied Mathematics* 195: 101–103.
11. Hurkens, C.A.J. & Woeginger, G.J. (2004). On the nearest neighbor rule for the traveling salesman problem. *Operations Research Letters* 32(1): 1–4.
12. Janson, S. (1999). One, two and three times $\log n/n$ for paths in a complete graph with random weights. *Combinatorics, Probability and Computing* 8(4): 347–361.
13. Johnson, D.S. & Papadimitriou, C.H. (1985). Performance guarantees for heuristics. In E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, & D. B. Shmoys (eds), *The traveling salesman problem. Wiley-Interscience Series in Discrete Mathematics.* Chichester: Wiley, pp. 145–180.
14. Kesten, H. (1986). Aspects of first passage percolation. In *École d'été de probabilités de Saint-Flour, XIV—1984*, vol. 1180. Lecture Notes in Mathematics. Berlin: Springer, pp. 125–264.
15. Kesten, H. & Lee, S. (1996). The central limit theorem for weighted minimal spanning trees on random points. *Annals of Applied Probability* 6(2): 495–527.
16. Megow, N., Mehlhorn, K., & Schweitzer, P. (2012). Online graph exploration: New results on old and new algorithms. *Theoretical Computer Science* 463: 62–72.
17. Mézard, M. & Parisi, G. (1986). A replica analysis of the travelling salesman problem. *Journal de Physique* 47: 1285–1296.
18. Pinsky, M.A. & Karlin, S (2011). *An introduction to stochastic modeling*. Amsterdam: Elsevier/Academic Press.
19. Rosenkrantz, D.J., Stearns, R.E., & Lewis II, P.M. (1977). An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing* 6(3): 563–581.
20. Steele, J.M. (1989). Cost of sequential connection for points in space. *Operations Research Letters* 8(3): 137–142.

21. Steele, J.M (1997). Probability theory and combinatorial optimization. In *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 69. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
22. Wästlund, J. (2010). The mean field traveling salesman and related problems. *Acta Mathematica* 204(1): 91–150.
23. Yukich, J.E. (1998). *Probability theory of classical Euclidean optimization problems*, vol. 1675. Lecture Notes in Mathematics. Berlin: Springer-Verlag.