Percolation-like Scaling Exponents for Minimal Paths and Trees in the Stochastic Mean Field Model

David J. Aldous
Department of Statistics
367 Evans Hall # 3860
U.C. Berkeley CA 94720
aldous@stat.berkeley.edu

August 6, 2004

Abstract

In the mean field (or random link) model there are n points and inter-point distances are independent random variables. For $0 < \ell < \infty$ and in the $n \to \infty$ limit, let $\delta(\ell) = 1/n \times$ (maximum number of steps in a path whose average step-length is $\leq \ell$). The function $\delta(\ell)$ is analogous to the *percolation function* in percolation theory: there is a critical value $\ell_* = e^{-1}$ at which $\delta(\cdot)$ becomes non-zero, and (presumably) a scaling exponent β in the sense $\delta(\ell) \asymp (\ell - \ell_*)^\beta$. Recently developed probabilistic methodology (in some sense a rephrasing of the cavity method developed in the 1980s by Mézard and Parisi) provides a simple albeit non-rigorous way of writing down such functions in terms of solutions of fixed-point equations for probability distributions. Solving numerically gives convincing evidence that $\beta = 3$. A parallel study with *trees* and *connected edge-sets* in place of paths gives scaling exponent 2, while the analog for classical percolation has scaling exponent 1. The new exponents coincide with those recently found in a different context (comparing optimal and near-optimal solutions of the mean-field TSP and MST problems), and reinforce the suggestion that scaling exponents determine universality classes for optimization problems on random points.

Key words and phrases. Combinatorial optimization, mean field model, percolation, probabilistic analysis of algorithms, scaling exponent,

1 Introduction

1.1 Paths

Consider n points with inter-point distances $(d(v,w) = d(w,v), 1 \leq v,w \leq n)$. A *path* $\pi = (v_0, v_1, \ldots, v_m)$ visits a set of points, distinct except that maybe $v_m = v_0$. Associated with a path π is its length (number of steps) $\mathrm{len}(\pi)$ and the average step-distance $A(\pi)$:

$$\begin{aligned} \mathrm{len}(\pi) &= m \\ A(\pi) &= m^{-1} \sum_{i=1}^{m} d(v_{i-1}, v_i). \end{aligned}$$

The celebrated *Traveling Salesman Problem* (TSP) concerns minimizing $A(\pi)$ subject to $\mathrm{len}(\pi) = n$. One can also consider, for given $m < n$, the question of the minimum value of $A(\pi)$ subject to $\mathrm{len}(\pi) \geq m$. This has also been studied as an algorithmic question [7, 8]; but instead we take a "statistical physics" viewpoint of studying the values $\min_\pi A(\pi)$ under a probability model for random points. The most natural probability model is n independent uniform random points in the unit square, and study of the TSP in this model goes back 45 years to Beardwood et al [9]. See Steele [21] for a recent survey of the general area. Unfortunately the kind of questions we study seem far out of reach of analytic methods in this two-dimensional model. Instead we use a more tractable model with several names (we say *stochastic mean-field* (SMF_n) but also called *random link* or *complete graph with random (exponentially distributed) edge-lengths*) which we imagine roughly as random points in *infinite*-dimensional space. Section 2 provides details of the SMF_n model. In the mid 1980s Mézard and Parisi [16] studied the TSP (and other optimization problems [17, 19]) in the SMF_n model, using the non-rigorous *cavity method* from statistical physics: see [18] for a recent survey of the cavity method. Recent work of the author [4, 1, 6] develops a methodology based on (additive) renormalization within an infinite-point random model of distance. This methodology, in some sense just a rephrasing of the cavity method, provides a consistent framework for a wide variety of different calculations for different optimization problems in the context of SMF_n.

In this paper we study a deterministic function $(\varepsilon(\delta), 0 < \delta \leq 1)$ arising as the limit

$$\varepsilon(\delta) = \lim_n E \min\{A(\pi) : \ \mathrm{len}(\pi) \geq \delta n, \ \pi \text{ a path in } \mathrm{SMF}_n\}. \tag{1}$$

(Limits asserted here and later are presumed, but not rigorously proved, to exist – see section 1.5.) The value $\varepsilon(1) \approx 2.04$ (obtained by numerically solving a fixed-point equation) goes back to Mézard and Parisi [16], while the value $\varepsilon(0+) = e^{-1} \approx 0.368$ is given in Aldous [3] Proposition 7 (other aspects of paths are treated by Janson [13]). Our purpose is to show how the recent methodology enables one to determine numerically the whole function $\varepsilon(\delta)$. A plot of the whole function is given in Figure 1 (left side). Of particular interest is the scaling behavior as $\delta \downarrow 0$. The numerical evidence (right side of Figure 1, and Table 1) strongly suggests a scaling exponent

$$\varepsilon(\delta) - \varepsilon(0+) \asymp \delta^\alpha \text{ with } \alpha = 1/3. \tag{2}$$

This kind of scaling exponent is precisely analogous to scaling exponents around the critical value in percolation theory, as explained in section 1.3.

1.2 Trees

There are parallel questions using trees in place of paths. Consider a complete graph on n vertices whose edges e have lengths $d(e)$. For any tree $\mathbf{t}$ in the graph, with edges $e_1, \ldots, e_m$, write size($\mathbf{t}$) for the number of edges of $\mathbf{t}$ and $A(\mathbf{t})$ for the average edge-length:

$$\begin{aligned} \text{size}(\mathbf{t}) &= m \\ A(\mathbf{t}) &= m^{-1} \sum_{e \in \mathbf{t}} d(e). \end{aligned}$$

The *Minimum Spanning Tree* (MST) problem asks for the minimum of $A(\mathbf{t})$ subject to size$(\mathbf{t}) = n - 1$. Take n random points in our stochastic mean field model SMF_n. Analogously to the results for paths, we anticipate a deterministic function $(\varepsilon^*(\delta), 0 < \delta \leq 1)$ arising as the limit

$$\varepsilon^*(\delta) = \lim_n E \min\{A(\mathbf{t}) : \text{ size}(\mathbf{t}) \geq \delta n, \ \mathbf{t} \text{ a tree in } \text{SMF}_n\}. \tag{3}$$

A well known result of Frieze [11] for the MST says that $\varepsilon^*(1) = \zeta(3) \approx 1.202$, whereas Aldous [3] argued $\varepsilon^*(0+) \approx 0.263$ by numerics with fixed point equations. Parallel to the study of paths, our methodology tells how in principle to determine numerically the whole function $\varepsilon^*(\delta)$. In practice we have not be able to carry this through (see section 5) but instead have analyzed the following related question. Instead of trees we consider *connected edge-sets* $\mathbf{e} = (e_1, \ldots, e_m)$.

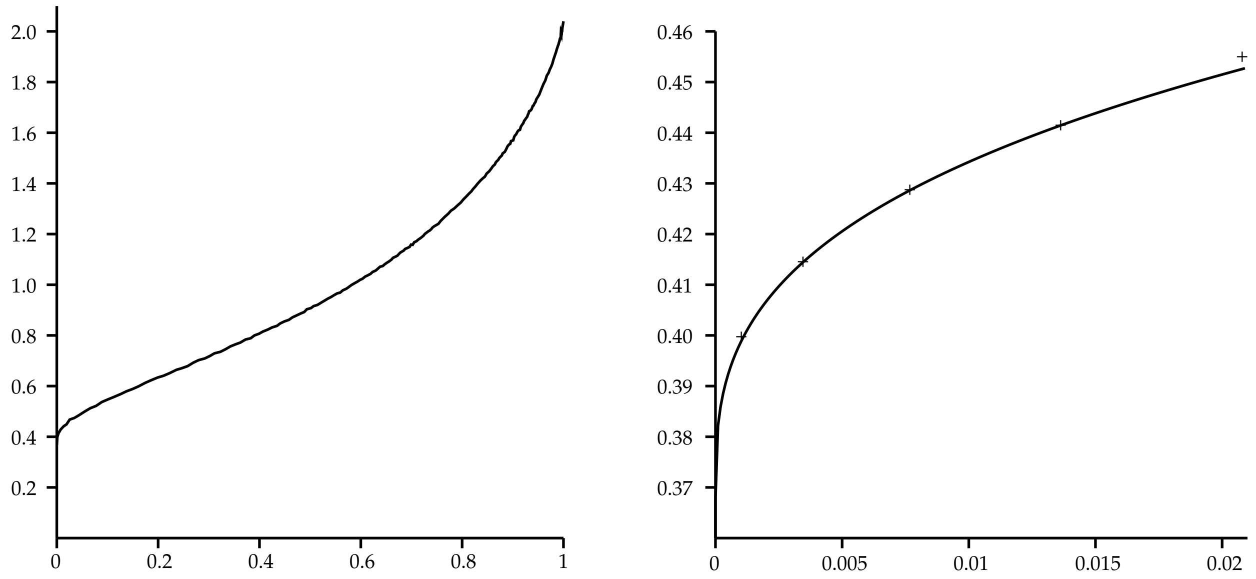

Figure 1. The limit function for paths. On the left is the function $\varepsilon(\delta)$ defined at (1). The horizontal axis is δ, the vertical axis is ε. The right side gives a close-up of the behavior for small δ: the points $+$ are the values estimated numerically in Table 1, and the curve is $\varepsilon - e^{-1} = 0.308\ \delta^{1/3}$.

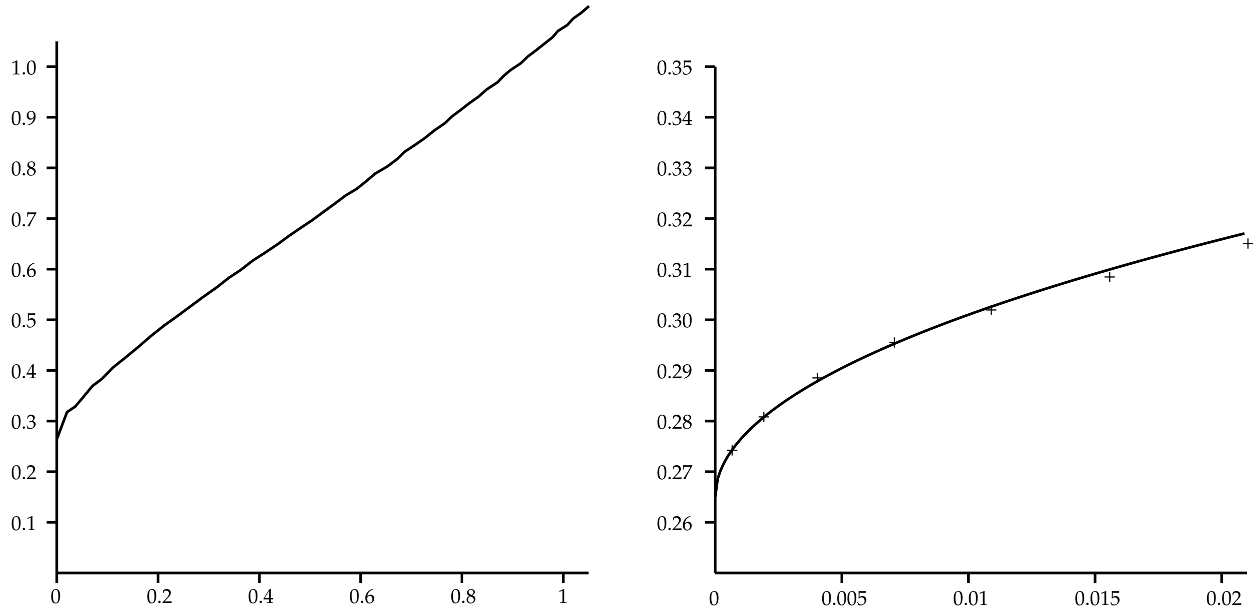

Figure 2. The limit function for connected edge-sets. On the left is the function $\tilde{\varepsilon}(\delta)$ defined at (4). The right side gives a close-up of the behavior for small δ: the points $+$ are the values estimated numerically in Table 1, and the curve is $\tilde{\varepsilon} - 0.265 = 0.360\ \delta^{1/2}$.

Define size($\mathbf{e}$) and $A(\mathbf{e})$ as above, and as at (3) we anticipate a deterministic function $(\tilde{\varepsilon}(\delta), 0 < \delta < \infty)$ arising as the limit

$$\tilde{\varepsilon}(\delta) = \lim_n E \min\{A(\mathbf{e}) : \ \mathrm{size}(\mathbf{e}) \geq \delta n, \ \mathbf{e} \text{ a connected edge-set in } \mathrm{SMF}_n\}. \tag{4}$$

A plot of the whole function is given in Figure 2 (left side). Again, scaling as $\delta \downarrow 0$ is of interest. The numerical evidence (right side of Figure 2, and Table 1) gives an estimate $\tilde{\varepsilon}(0+) \approx 0.265$ and strongly suggests a scaling exponent

$$\tilde{\varepsilon}(\delta) - \tilde{\varepsilon}(0+) \asymp \delta^{\alpha^*} \text{ with } \alpha^* = 1/2. \tag{5}$$

As explained in section 5, we must have the same $\delta \downarrow 0$ behavior for the "tree" function $\varepsilon^*(\cdot)$ as for the "connected edge-set" function $\tilde{\varepsilon}(\cdot)$.

$\varepsilon(\delta)$		paths		$\tilde{\varepsilon}(\delta) = \varepsilon^*(\delta)$		trees	
λ	δ	$\varepsilon - e^{-1}$	$\frac{\varepsilon - e^{-1}}{\delta^{1/3}}$	λ	δ	$\varepsilon - 0.265$	$\frac{\varepsilon - 0.265}{\delta^{1/2}}$
0.53	.0386	.112	.332	0.34	.0211	.0502	.346 $\pm$.001
0.51	.0293	.100	.324	0.33	.0156	.0436	.349 $\pm$.001
0.49	.0209	.0872	.317	0.32	.0110	.0371	.355 $\pm$.002
0.47	.0137	.0737	.308	0.31	.00714	.0307	.364 $\pm$.003
0.45	.00773	.0610	.308	0.30	.00411	.0237	.370 $\pm$.005
0.43	.00352	.0468	.308	0.29	.00198	.0160	.360 $\pm$.010
0.41	.00108	.0321	.313	0.28	.000730	.0094	.348 $\pm$.027

Table 1. Scaling behavior near the critical point, for $\varepsilon(\delta)$ (left side) and $\tilde{\varepsilon}(\delta)$ (right side). In each case the function is defined implicitly via functions $\varepsilon(\lambda)$ and $\delta(\lambda)$, as explained below (7). See section 6.1 for discussion of the $\pm$ sampling error.

1.3 The analogy with percolation functions

Instead of the functions $\varepsilon(\delta)$ and $\tilde{\varepsilon}(\delta)$ at (1,4), we could equivalently study their inverse functions $\delta(\ell)$ and $\tilde{\delta}(\ell)$ whose interpretations are

$$\delta(\ell) = \lim_n E \max\{n^{-1} \,\mathrm{len}(\pi) : \ A(\pi) \leq \ell, \ \pi \text{ a path in } \mathrm{SMF}_n\}.$$

$$\tilde{\delta}(\ell) = \lim_n E \max\{n^{-1} \,\#\mathbf{e} : \ A(\mathbf{e}) \leq \ell, \ \mathbf{e} \text{ a connected edge-set in } \mathrm{SMF}_n\}.$$

Of course the scaling exponent for trees at (5) can be rewritten as

$$\tilde{\delta}(\ell) \asymp (\ell - \ell_*)^{\beta} \text{ with } \beta_{\mathrm{tree}} = 2$$

for $\ell > \ell_* = \tilde{\varepsilon}(0+)$. Similarly the scaling exponent for paths at (2) can be rewritten as

$$\delta(\ell) \asymp (\ell - \ell_*)^\beta \text{ with } \beta_{\text{path}} = 3$$

for $\ell > \ell_* = \varepsilon(0+) = e^{-1}$. To make the analogy with percolation, for $0 < t < \infty$ consider the maximal size connected edge-subset $\mathrm{perc}_n(t)$ such that

$$\max_{e \in \mathrm{perc}_n(t)} d(e) \leq t.$$

So $\mathrm{perc}_n(t)$ is the largest percolation cluster, that is the largest connected component of the subgraph of SMF_n consisting of edges of length $\leq t$. Well known theory concerning giant components in the random graph process implies

$$\lim_n n^{-1} E \# \mathrm{perc}_n(t) = p(t)$$

where $p(t)$ has the properties

$$p(t) = 0, \ 0 \leq t \leq 1; \quad p(t) \sim 2(t-1) \text{ as } t \downarrow 1.$$

Thus the scaling exponent for ordinary percolation in SMF_n is $\beta_{\text{perc}} = 1$. Note we can rewrite $p(\cdot)$ as

$$p(\ell) = \lim_n E \max\{n^{-1} \# \mathbf{e} : \max_{e \in \mathbf{e}} d(e) \leq \ell, \ \mathbf{e} \text{ a connected edge-set in } \mathrm{SMF}_n\}.$$

This differs from $\tilde{\delta}(\ell)$ only in the use of $\max_{e \in \mathbf{e}} d(e)$ in place of $\mathrm{ave}_{e \in \mathbf{e}} d(e)$. So we have a rather precise analogy between our function and the usual percolation function.

1.4 The big picture

This paper provides some pieces of a big picture. Time is not yet ripe for a complete survey, but let us provide some glimpses of other pieces. Our main results here are the scaling exponents $\beta_{\text{tree}} = 2$, $\beta_{\text{path}} = 3$ near the "percolative critical values" $\varepsilon^*(0+)$, $\varepsilon(0+)$. In Aldous and Percus [1] we study a different notion of "scaling exponent" dealing with behavior near the "spanning constants", i.e. near the MST and TSP constants $\varepsilon^*(1)$, $\varepsilon(1)$. These exponents are based on comparing near-optimal solutions to the optimal solution, and turn out to take the values 2 and 3. These values hold in the SMF_n model by the methodology used here, and there is evidence (from Monte Carlo simulations) they hold for random points in real $d \geq 2$ dimensional space. That the "percolative" scaling exponents in this paper coincide with the "spanning" exponents of [1] is remarkable, and reinforces the idea put forward in [1] that these scaling exponents provide a natural way of defining "universality classes" of optimization problems on random points. A natural next project is to study via Monte Carlo these percolative scaling exponents for random points in $d \geq 2$ dimensions, although this seems algorithmically difficult. At the time of writing, the only one of the four exponents we understand non-computationally is the tree/spanning exponent "2", which is easily explained [1] using the greedy algorithm for finding the MST. See section 6.3 for further remarks.

1.5 Methodology

Here is our methodology, in brief.

- The stochastic mean field model for n points has a $n \to \infty$ limit, the PWIT (section 2).

- Introducing Lagrange multipliers turns the constrained maximization problem into an unconstrained maximization problem. One can formulate the corresponding maximization problem for the PWIT, and define random variables (X, Y) measuring the relative effect on the maximized value of including or excluding a reference edge in the solution.

- The recursive structure of the PWIT enables one to write down equations (11,12) satisfied by (X, Y), which can be numerically solved. The limit optimal values of length and $A(\cdot)$ are determined from the definitions of (X, Y).

The arguments are not mathematically rigorous, for two main reasons. First, the central idea of identifying limits of solutions of finite-n optimization problems with solutions of infinite-n optimization problems requires justification, which has been given only in the case (related to but slightly different from those considered here) of mean-field *minimal matching* [4] and the less closely related case of some random graph problems [12]. Second, the scaling exponents are found by numerically solving equations with a parameter and examining numerical behavior as the parameter goes to a limit, and this falls short of analyzing the parameter-limit behavior rigorously.

2 The stochastic mean field model and its infinite-point limit

For fixed n, the SMF_n model is defined as follows. There are n points. For each of the $\binom{n}{2}$ pairs of points, there is a "link" whose length is random with exponential (mean n) distributions, these random lengths being independent. The distance between two points is then the length of the shortest path of links between them. The assumption of *exponential* distribution is convenient but not essential; results are unchanged if the link lengths are nL where $L > 0$ has a density with $f_L(0+) = 1$.

The scaling of link lengths is set up so that, as $n \to \infty$, the mean distance from a typical point to its nearest neighbor converges to 1. But much more is true, as we now outline briefly (see [6] for detailed survey). There is an infinite-point model, the *PWIT*, defined as follows. There is a root $\emptyset$. The root has an infinite number of links to points labeled $(1, 2, 3, \ldots)$, and these link lengths $0 < \xi_1^\emptyset < \xi_2^\emptyset < \ldots$ are the successive points of a Poisson process of rate 1 on $(0, \infty)$. Recursively, each point i has an infinite number of further links to points $(i1, i2, i3, \ldots)$ whose lengths $0 < \xi_1^i < \xi_2^i < \ldots$ are independent copies of the Poisson process. The PWIT is illustrated in Figure 3, and the web site [10] enables one to explore its structure via genuine simulations.

The PWIT is the $n \to \infty$ limit of SMF_n in a precise sense called *local weak convergence* [6]. Choose a random point of SMF_n to be a root. Then as $n \to \infty$, for any fixed "window size" r the configuration of points in SMF_n within a window of radius r centered at the root converges in distribution to the configuration of points in the PWIT within a window of radius r centered at the root.

Two properties of the PWIT enter into our calculations later.

(a). For each "child" i linked to the root, there is a subtree $\mathbf{T}_i$ consisting of i and its descendants. The *recursive structure of the PWIT*, built into the definition, says that the subtrees $\mathbf{T}_i$ are independent as i varies and are distributed as the PWIT itself.

(b). The fact that we choose a (uniform) *random* vertex of SMF_n to be the root leads to a *stationarity* property of the PWIT. Roughly, this says that the root is a "typical" vertex of the PWIT and therefore, by the ergodic principle, we can compute averages over all vertices of the PWIT by computing expectations at the root. As a more explicit instance, given a random vertex subset A_n of SMF_n, suppose we have joint local weak convergence of (SMF_n, A_n) to $(\mathrm{PWIT}, A_\infty)$ for a random vertex subset A_∞ of the PWIT. Then $n^{-1}E\#A_n \to P(\text{root} \in A_\infty)$, where # denotes cardinality. Note that here A_n is dependent on SMF_n, but the root of SMF_n is then chosen independently of A_n.

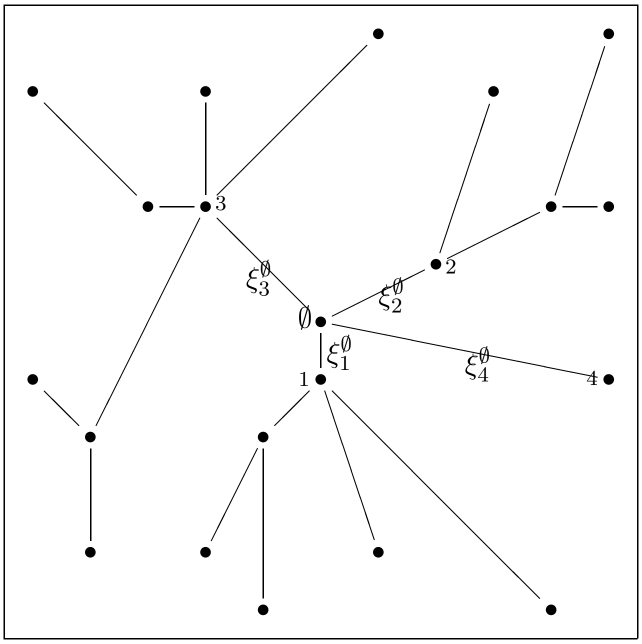

Figure 3. The PWIT. Illustration of the vertices of the PWIT within a window of radius 3 centered on the root $\emptyset$. Lines indicate the links, but are drawn only when both end-vertices are within the window. Thus the four links at $\emptyset$ shown are at distances $0 < \xi_1^\emptyset < \xi_2^\emptyset < \xi_3^\emptyset < \xi_4^\emptyset < 3$ from $\emptyset$, while there are an infinite number of links at $\emptyset$ of lengths greater than 3. Orientation of lines in pictures is arbitrary.

3 The recursive distributional equation: the path case

By introducing a Lagrange multiplier $\lambda > 0$, the finite-n problem of minimizing $A(\pi)$ subject to $\mathrm{len}(\pi)$ can be reformulated as

$$\begin{array}{lcl} \texttt{maximize} & : & \lambda \frac{\mathrm{len}(\pi)}{n} - A(\pi) \\ \texttt{subject to} & : & \pi \text{ a path in } \mathrm{SMF}_n. \end{array}$$

This has a random solution $\pi_n(\lambda)$. We expect that as $n \to \infty$

$$\begin{aligned} n^{-1} E \mathrm{len}(\pi_n(\lambda)) &\to \delta(\lambda) && (6)\\ EA(\pi_n(\lambda)) &\to \varepsilon(\lambda) && (7)\end{aligned}$$

and that the function $\varepsilon(\delta)$ at (1) is determined implicitly via the two functions $\delta(\lambda), \varepsilon(\lambda)$.

To set up the analogous optimization problem on the PWIT, we first define what will be seen to be sets of feasible solutions. Write $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ for a family of vertex-disjoint doubly-infinite paths in the PWIT. Define

- $\mathcal{E}_0$ is the set of such families for which no path goes through the root;
- $\mathcal{E}_2$ is the set of such families for which some path goes through the root;
- $\mathcal{E}_1$ is the set of such families, where in addition to the doubly-infinite paths there exists exactly one singly-infinite path, and this path starts at the root.

Note the subscript indicates degree of root in the family. For $\boldsymbol{\pi} = (\pi_u) \in \mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2$ consider the objective function

$$b(\boldsymbol{\pi}) = \lambda \times \#\{v : v \text{ a vertex of some } \pi_u\} - \sum_{e:e \text{ edge of some } \pi_u} \xi_e$$

with the convention that, for $\boldsymbol{\pi} \in \mathcal{E}_1$, the root vertex counts as $1/2$. (Recall that ξ_e is the length of edge e in the PWIT.) In the limit procedure which takes SMF_n to the PWIT, the limits of "paths of length order n" is exactly the set $\mathcal{E}_0 \cup \mathcal{E}_2$ of families of doubly-infinite paths. Thus the optimization problem on the PWIT can be written symbolically as

$$\text{maximize } b(\boldsymbol{\pi}) \text{ over } \boldsymbol{\pi} \in \mathcal{E}_0 \cup \mathcal{E}_2. \tag{8}$$

We seek to study the $\boldsymbol{\pi}$ that attains the maximum. But we can't work directly with definition (8), because $b(\boldsymbol{\pi})$ is the difference of two sums, each sum having value $+\infty$. Instead we can consider *differences* between maximized $b(\cdot)$ values. Specifically, given a realization of the PWIT we define realizations of two random variables via

$$\begin{aligned} X &= \max_{\boldsymbol{\pi} \in \mathcal{E}_1} b(\boldsymbol{\pi}) - \max_{\boldsymbol{\pi} \in \mathcal{E}_0} b(\boldsymbol{\pi}) && (9)\\ Z &= \max_{\boldsymbol{\pi} \in \mathcal{E}_2} b(\boldsymbol{\pi}) - \max_{\boldsymbol{\pi} \in \mathcal{E}_0} b(\boldsymbol{\pi}). && (10)\end{aligned}$$

To see why such definitions are useful, note that the solution $\boldsymbol{\pi}$ to (8) will have a path through the root if and only if

$$\max_{\boldsymbol{\pi}\in\mathcal{E}_2} b(\boldsymbol{\pi}) > \max_{\boldsymbol{\pi}\in\mathcal{E}_0} b(\boldsymbol{\pi}),$$

that is if and only if $Z > 0$.

We now set up the recursion that determines the joint distribution of (X, Z). We remark that X is introduced only because it arises in the recursion for Z – it would obviously be preferable to find a recursion involving only a single quantity like Z, but that seems impossible to find. Figure 4 may be helpful in visualizing the argument below.

By the recursive structure of the PWIT, for each subtree $(\mathbf{T}_i,\ i = 1, 2, 3, \ldots)$ defined by the children of the root, the random pairs (X_i, Z_i) defined as at (9,10) on $\mathbf{T}_i$ are distributed as (X, Z) and are independent as i varies. We will first show

$$X = \max_i(\lambda - \xi_i + X_i - Z_i^+) \tag{11}$$

where $Z^+ = \max(0, Z)$ and where ξ_i are the edge-lengths at the root.

Consider the families $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_0$ attaining the maxima over $\mathcal{E}_1$ and $\mathcal{E}_0$ in the definition (9) of X. So $\boldsymbol{\pi}_1$ contains an edge from the root to child i, say. On the subtrees $(\mathbf{T}_j, j \neq i)$ the maximal families must be identical, so we only need compare $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_0$ on the root-edges and the subtree $\mathbf{T}_i$. There is a contribution $\lambda - \xi_i$ to $b(\cdot)$ from the edge (root, i). In the subtree $\mathbf{T}_i$, we have

$$\begin{aligned} X_i &= \max_{\boldsymbol{\pi}\in\mathcal{E}_1(i)} b(\boldsymbol{\pi}) - \max_{\boldsymbol{\pi}\in\mathcal{E}_0(i)} b(\boldsymbol{\pi}) \\ Z_i^+ &= \max_{\boldsymbol{\pi}\in\mathcal{E}_2(i)\cup\mathcal{E}_0(i)} b(\boldsymbol{\pi}) - \max_{\boldsymbol{\pi}\in\mathcal{E}_0(i)} b(\boldsymbol{\pi}). \end{aligned}$$

The family $\boldsymbol{\pi}_1$ contains the first-term maximizing family $\boldsymbol{\pi} \in \mathcal{E}_1(i)$ in this definition of X_i, while the family $\boldsymbol{\pi}_0$ contains the first-term maximizing family $\boldsymbol{\pi} \in \mathcal{E}_2(i) \cup \mathcal{E}_0(i)$ in this definition of Z_i^+. So the contribution to $b(\boldsymbol{\pi}_1)$ from $\mathbf{T}_i$ equals $X_i - Z_i^+$. This establishes (11), since we can choose the maximizing value of i to be the edge at the root.

A similar argument leads to a recursion for Z. A family $\boldsymbol{\pi}_2$ containing a path through the root must contain two edges (root, i) and (root, j), say. The contribution to $b(\cdot)$, relative to using no edges at the root, of using (root, i) equals $\lambda - \xi_i + X_i - Z_i^+$. Hence we get

$$Z = \max_i(\lambda - \xi_i + X_i - Z_i^+) + \max_i{}^{[2]}(\lambda - \xi_i + X_i - Z_i^+) \tag{12}$$

where $\max_i^{[2]}$ denotes second maximum. Equations (11,12) together give a formula for (X, Z) in terms of $(X_i, Z_i), i \geq 1$ and $(\xi_i, i \geq 1)$. By the recursive structure of the PWIT, the $(X_i, Z_i), i \geq 1$ are independent copies of (X, Z). Thus (11,12) constitute a *recursive distributional equation* (RDE) for the "unknown" joint distribution (X, Z).

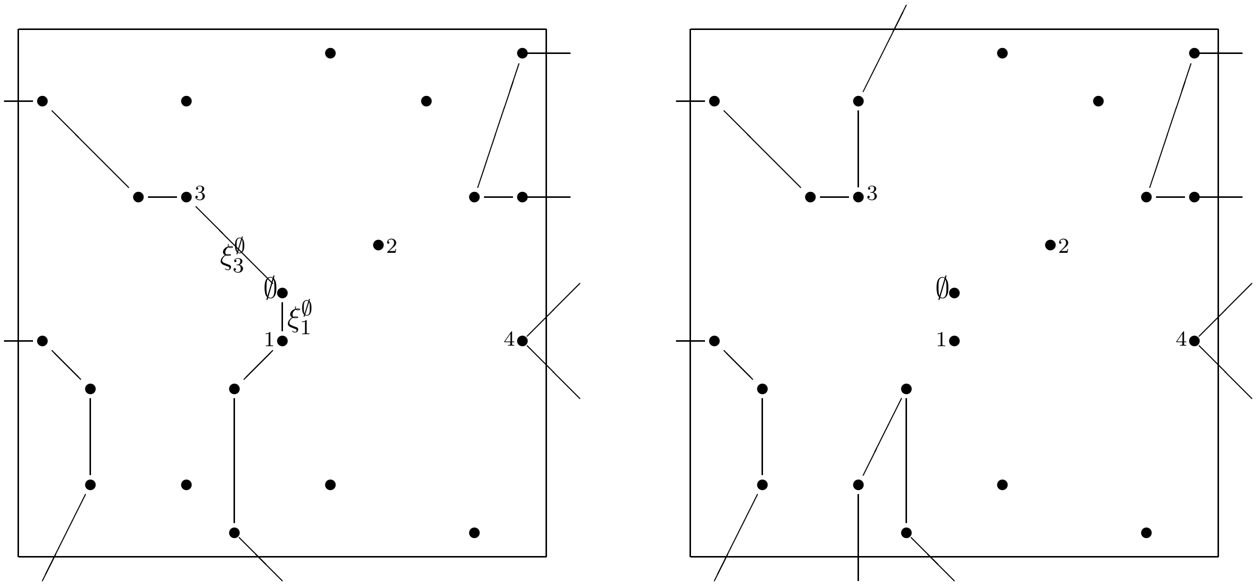

Figure 4. On the realization of the PWIT from Figure 3, the left side illustrates the optimal $\boldsymbol{\pi} \in \mathcal{E}_2$ which does pass through the root (which happens to use the edges from the root to 1 and to 3), and the right side illustrates the optimal $\boldsymbol{\pi} \in \mathcal{E}_0$ which does not pass through the root. These path-families coincide on the subtrees of children except $\{1, 3\}$. On the subtree $\mathbf{T}_3$, the optimal family on the right side has a path through the root 3, whereas on the subtree $\mathbf{T}_1$ it does not.

We next show how the desired quantities $\delta(\lambda)$ and $\varepsilon(\lambda)$ at (6,7) can be obtained from the distribution of (X, Z). The quantity $\delta(\lambda)$ represents the proportion of vertices in the optimal solution to (8). By the stationarity property of the root of the PWIT, $\delta(\lambda)$ is just the probability that the optimal family contains a path through the root. As observed above, this happens if and only if $Z > 0$, so

$$\delta(\lambda) = P(Z > 0). \tag{13}$$

When $Z > 0$, the lengths of the two edges in the path at the root are ξ_I and

ξ_J, where in the notation of (12)

$$\begin{aligned} I &= \arg\max_i (\lambda - \xi_i + X_i - Z_i^+) \\ J &= \arg\max_i{}^{[2]} (\lambda - \xi_i + X_i - Z_i^+). \end{aligned}$$

Again by stationarity, the mean edge-lengths over <u>all</u> edges in the optimal family equals the mean edge-length in the edges <u>at the root</u> in the optimal family, conditioned on the root being used, and so

$$\varepsilon(\lambda) = \frac{E\left[(\frac{\xi_I + \xi_J}{2}) \ 1_{(Z>0)} \right]}{\delta(\lambda)}. \tag{14}$$

As mentioned before, equations (11,12) together form a *recursive distributional equation* (RDE) for the joint distribution of (X, Z). Such RDEs are pervasive not only in problems within SMF_n but also in many other areas of applied probability: see [5] for a survey. They rarely allow explicit solutions, but there is a standard *bootstrap Monte Carlo* method ([5] section 8.1) which is very easy to implement and which gives, in principle, arbitrarily-accurate approximate solutions of RDEs. This method was used to solve the RDE for (X, Z) and then estimate $\delta(\lambda)$ and $\varepsilon(\lambda)$ via (13,14). Numerical values were shown in Table 1 and Figure 1.

4 The connected edge-set case

The conceptual ideas behind the analysis of $\tilde{\varepsilon}(\delta)$ at (4) are very similar to the analysis of $\varepsilon(\delta)$ in the previous section, so we will only detail the differences.

Consider a forest $\mathbf{f} = (\mathbf{t}_1, \mathbf{t}_2, \ldots)$ in the PWIT, each of whose tree-components $\mathbf{t}_i$ is infinite. Define
$\mathcal{F}$ is the set of such forests $\mathbf{f}$;
$\mathcal{F}_0$ is the set of such forests such that the root <u>is not</u> in any component;
$\mathcal{F}_1$ is the set of such forests such that the root <u>is</u> in some component;
$\mathcal{F}_2$ is the set of such forests, where in addition to the infinite tree-components we allow the tree-component containing the root to be either empty, or finite, or infinite.

In the limit procedure which takes SMF_n to the PWIT, the limits of "connected edge-sets of size order n" is exactly the set $\mathcal{F}$ of forests whose tree-components are all infinite. For $\mathbf{f} = (\mathbf{t}_i) \in \mathcal{F}_2 \supset \mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$, consider

$$b(\mathbf{f}) = \lambda \times \#\{e : e \text{ an edge of some } \mathbf{t}_i\} - \sum_{e:e \text{ edge of some } \mathbf{t}_i} \xi_e.$$

The optimization problem on the PWIT is

$$\text{maximize } b(\mathbf{f}) \text{ over } \mathbf{f} \in \mathcal{F}. \tag{15}$$

To study this we define

$$\begin{aligned} Y &= \max_{\mathbf{f}\in\mathcal{F}} b(\mathbf{f}) - \max_{\mathbf{f}\in\mathcal{F}_0} b(\mathbf{f}) && (16)\\ Z &= \max_{\mathbf{f}\in\mathcal{F}_1} b(\mathbf{f}) - \max_{\mathbf{f}\in\mathcal{F}_0} b(\mathbf{f}) && (17)\\ X &= \max_{\mathbf{f}\in\mathcal{F}_2} b(\mathbf{f}) - \max_{\mathbf{f}\in\mathcal{F}_0} b(\mathbf{f}). && (18) \end{aligned}$$

Because $\mathcal{F}_2 \supset \mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$ we have

$$X \geq Y = Z^+.$$

The recursion for X, analogous to (11), is

$$X = \sum_i (\lambda - \xi_i + X_i - Y_i)^+. \tag{19}$$

The argument is the same as for (11): the contribution to $b(\cdot)$ by using edge (root, i), as compared to not using it, equals $(\lambda - \xi_i + X_i - Y_i)$, and we may use any number, or zero, such edges. The recursion for Z is

$$Z = \max_I \left(\sum_{i\in I} (\lambda - \xi_i + Z_i - Y_i) + \sum_{i\notin I} (\lambda - \xi_i + X_i - Y_i)^+ \right) \tag{20}$$

where I denotes a *non-empty* subset of $\{1, 2, 3, \ldots\}$. Here the first sum represents the contribution from the set I of children i such that, in the optimal $\mathbf{f} \in \mathcal{F}_1$, in the subtree $\mathbf{T}_i$ the root i is in an infinite component. The set I must be non-empty in order for $\mathbf{f} \in \mathcal{F}_1$. Now the fact $X_i \geq Z_i$ implies

$$\lambda - \xi_i + Z_i - Y_i \leq (\lambda - \xi_i + X_i - Y_i)^+$$

which implies there is an optimal I with only one element, and we can rearrange (20) to become

$$Z = X + \max_i \left((\lambda - \xi_i + Z_i - Y_i) - (\lambda - \xi_i + X_i - Y_i)^+ \right).$$

Finally, since $Y_i = Z_i^+$ we obtain the following RDE for the joint distribution of (X, Z).

$$\begin{aligned} X &= \sum_i (\lambda - \xi_i + X_i - Z_i^+)^+ && (21)\\ Z &= X + \max_i \left((\lambda - \xi_i + Z_i - Z_i^+) - (\lambda - \xi_i + X_i - Z_i^+)^+ \right). && (22) \end{aligned}$$

We next show how the desired quantities $\tilde{\delta}(\lambda)$ and $\tilde{\varepsilon}(\lambda)$ can be obtained from (X, Z). Consider the optimal $\mathbf{f}$ in (15). This $\mathbf{f}$ contains the root if and only if $\mathbf{f} \in \mathcal{F}_1$, that is if and only if $Z > 0$, so

$$\tilde{\delta}(\lambda) = P(Z > 0). \tag{23}$$

When $Z > 0$, the set $\mathcal{I}$ of edges at the root used in $\mathbf{f}$ is the set of i for which the contribution $(\lambda - \xi_i + X_i - Z_i^+)$ is strictly positive, plus (if distinct) the maximizing i in (22). This leads to

$$\tilde{\varepsilon}(\lambda) = \frac{E\left[\left(\frac{1}{2}\sum_{i \in \mathcal{I}} \xi_i\right) 1_{(Z>0)}\right]}{\tilde{\delta}(\lambda)} \tag{24}$$

for $\mathcal{I}$ as above.

5 Trees

Studying trees $\mathbf{t}$ in order to study the limit function $\varepsilon^*(\delta)$ at (3) is a little more subtle. What are the feasible solutions on the PWIT corresponding to the limits of trees in SMF_n? At first sight they are just the set $\mathcal{F}$ of forests $\mathbf{f} = (\mathbf{t}_i)$ in section 4. But this is wrong; instead, by analogy with many other examples of limits of infinite trees [2, 15] the relevant feasible solutions are forests $\mathbf{f} = (\mathbf{t}_i)$ with the extra property that each of whose tree-components $\mathbf{t}_i$ have *one end*; that is, from each vertex of $\mathbf{t}_i$ there is exactly one infinite path in $\mathbf{t}_i$.

To mimic the analysis of the previous section with this family of forests, it turns out we need, in place of $\mathcal{F}_2$ before, the family defined as

$\mathcal{F}_2$ is the set of such forests, modified so that the tree-component containing the root may be either empty or finite, but not infinite.

But now the analog of X at (18) cannot be represented recursively, since (roughly speaking) there is no recursive criterion for finiteness. Instead we need to consider, separately for $m = 0, 1, 2, \ldots$, a definition such as

$\mathcal{F}_{(m)}$ is the set of such forests, modified so that the tree-component containing the root has exactly m edges.

Defining X_m in terms of a maximum over $\mathcal{F}_{(m)}$ leads to a RDE for the infinite family $(X_0, X_1, X_2, \ldots, Z)$. But we have not attempted to solve this numerically.

Fortunately, this detailed analysis is unnecessary for investigating the scaling exponent because

$$\varepsilon^*(\delta) = \tilde{\varepsilon}(\delta) \text{ when } \tilde{\varepsilon}(\delta) < e^{-1}.$$

To outline the argument, consider the minimizing edge-set $\mathbf{e}$ for $\tilde{\varepsilon}(\delta)$ in this range. Suppose $\mathbf{e}$ contains a cycle of length order n. By the fact (for *paths*) $\varepsilon(0+) = e^{-1}$, this cycle has average edge-length $> e^{-1}$ and hence has some edge of length $> e^{-1}$. Removing this edge would reduce $A(\mathbf{e})$ without essentially affecting the constraint on $\mathrm{len}(\mathbf{e})$, contradicting minimality. So $\mathbf{e}$ can have no cycles of length order n. As for short cycles, fix $a < e^{-1}$ and consider a typical point v of SMF_n. By the arguments of [3, 13] (comparison with the Yule process),

$$P(v \text{ in any cycle } \mathbf{c} \text{ with } A(\mathbf{c}) < a) \to 0 \text{ as } n \to \infty$$

and it follows that the contribution to $A(\mathbf{e})$ from short cycles $\to 0$ as $n \to \infty$.

6 Final remarks

6.1 Sampling errors in Table 1

We treat the case of trees; the case of paths could be treated similarly. To obtain the numerical values in Table 1, we represented the distribution (X, Z) via 10^6 points and iterated the RDE 1000 times, truncating the Poisson process $(\xi_i, 1 \le i < \infty)$ at $i = 20$. This necessitated, for each value of λ, a total of 2×10^{10} calls to the random number generator. We calculated ε and δ using the final 200 generations, that is using 2×10^8 points. There are various possible errors in this way of estimating scaling exponents, of which the only one which can be quantified is "sampling error". Clearly

$$\text{s.d. (estimate of } \delta) \approx \delta^{1/2}/\sqrt{2 \times 10^8} \approx 0.7 \times 10^{-4}\ \delta^{1/2}$$

which is negligible. But the error for ε is not negligible, since it is based on only a proportion δ of the samples, giving

$$\text{s.d. (estimate of } \varepsilon) \approx \frac{\text{s.d.}(\xi)}{\sqrt{2 \times 10^8\ \delta}}$$

where $\text{s.d.}(\xi) \approx 0.3$ is the s.d. of the ξ-values used to estimate ε via (24). This leads to

$$\text{s.d. (estimate of } \varepsilon/\delta^{1/2}) \approx 2 \times 10^{-5}\ \delta^{-1}$$

which are the $\pm$ values shown in Table 1.

6.2 Rigorous bounds on scaling exponents

Because the limit $\varepsilon(0+) = e^{-1}$ in the paths setting is essentially just a first moment calculation, a referee suggests that similar first moment methods should establish rigorously some bound on $\varepsilon(\delta)$ and hence some bound of the scaling exponent in the paths case. We concur, but have not attempted a detailed calculation.

6.3 Scope of scaling exponents

It seems difficult to specify precise the range of settings in which a definition of *percolation-like scaling exponent* makes sense and is interesting. Within the stochastic mean field model there is a well studied *minimum matching* problem (see [14, 20] for recent proofs of the Parisi conjecture) in which context one could define

$$\varepsilon^{\text{match}}(\delta) = \lim_n E \frac{\text{length min matching of some } \delta n \text{ vertices}}{\frac{1}{2}\delta n}.$$

But here it is clear that

$$\varepsilon^{\text{match}}(\delta) \sim \delta \text{ as } \delta \downarrow 0$$

so that the critical value equals 0 and the scaling exponent equals 1. However, since the critical value equals zero we are inclined to regard this case as "not percolation-like". A referee suggests the example (again, within the stochastic mean field model) of the path through δn points chosen greedily by choosing the shortest available edge at each successive vertex, but this also seems "not percolation-like".

Acknowledgement. I thank two anonymous referees for helpful comments.

References

[1] David Aldous and Allon G. Percus. Scaling and universality in continuous length combinatorial optimization. *Proc. Natl. Acad. Sci. USA*, 100:11211–11215, 2003.

[2] D.J. Aldous. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.*, 1:228–266, 1991.

[3] D.J. Aldous. On the critical value for percolation of minimum-weight trees in the mean-field distance model. *Combin. Probab. Comput.*, 7:1–10, 1998.

[4] D.J. Aldous. The $\zeta(2)$ limit in the random assignment problem. *Random Structures Algorithms*, 18:381–418, 2001.

[5] D.J. Aldous and A. Bandyopadhyay. A survey of max-type recursive distributional equations. arXiv:math.PR/0401388, 2004.

[6] D.J. Aldous and J.M. Steele. The objective method: Probabilistic combinatorial optimization and local weak convergence. In H. Kesten, editor, *Probability on Discrete Structures*, volume 110 of *Encyclopaedia of Mathematical Sciences*, pages 1–72. Springer-Verlag, 2003.

[7] S. Arora. Nearly linear time approximation schemes for Euclidean TSP and other geometric problems. In *38th Annual Symp. Foundations Computer Science*, pages 554–565. IEEE, 1997.

[8] E. Balas. The prize collecting traveling salesman problem. *Extremes*, 19:621–636, 1989.

[9] J. Beardwood, H.J. Halton, and J.M. Hammersley. The shortest path through many points. *Proc. Cambridge Phil. Soc.*, 55:299–327, 1959.

[10] D. Brightly. A Java Visualization of the PWIT, 2004. http://www.stat.berkeley.edu/users/aldous/PWIT/PWITDemo.html.

[11] A.M. Frieze. On the value of a random minimum spanning tree problem. *Discrete Appl. Math.*, 10:47–56, 1985.

[12] D. Gamarnik, T. Nowicki, and G. Swirscsz. Maximum weight independent sets and matchings in sparse random graphs: Exact results using the local weak convergence method. arXiv:math.PR/0309441, 2003.

[13] S. Janson. One, two and three times $\log n/n$ for paths in a complete graph with random weights. *Combin. Probab. Comput.*, 8:347–361, 1999.

[14] S. Linusson and J. Wästlund. A proof of Parisi's conjecture on the random assignment problem. *Probab. Th. Rel. Fields*, 128:419–440, 2004.

[15] R. Lyons, Y. Peres, and O. Schramm. Minimal spanning forests. In preparation, 2004.

[16] M. Mézard and G. Parisi. A replica analysis of the travelling salesman problem. *J. Physique*, 47:1285–1296, 1986.

[17] M. Mézard and G. Parisi. On the solution of the random link matching problem. *J. Physique*, 48:1451–1459, 1987.

[18] M. Mézard and G. Parisi. The cavity method at zero temperature. *J. Statist. Phys.*, 111:1–34, 2003.

[19] M. Mézard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

[20] C. Nair, B. Prabhakar, and M. Sharma. A proof of Parisi's conjecture for the finite random assignment problem. Unpublished, 2004.

[21] J.M. Steele. *Probability Theory and Combinatorial Optimization*. Number 69 in CBMS-NSF Regional Conference Series in Applied Math. SIAM, 1997.

Short title: Percolation-like scaling exponents.