Lecture 2 described an example "Maximum Partial Matchings on Random Trees" and gave

Theorem $n^{-1}EM(T_n) \to c = E[\xi\ 1(\xi > Y+Z)]$ where Y and Z are solutions of

$$Y \stackrel{d}{=} \max(0, \xi_i - Y_i,\ 1 \le i \le \text{Pois}(1))$$

$$Z \stackrel{d}{=} \max(\xi - Z', Y).$$

Each equality comes out of a simple picture.

Amazingly, there is a large class of problems – combinatorial optimization over locally tree-like random structures – where one can just write down the (non-rigorous) solution in the same format; 2 or 3 equalities from simple pictures. This can be viewed as an aspect of the **cavity method**.

Most studied example is the "random assignment problem", though this can now be analyzed combinatorially exploiting special structure. I will describe 2 examples in the "mean-field model of distance". Recall this means:

Take the complete graph on n vertices. To each of the $\binom{n}{2}$ edges e assign i.i.d. lengths ℓ_e with Exponential(mean n) distribution. Call this network G_n. Recall

$$G_n \to_{LWC} PWIT.$$

Write $TSP(G_n)$ for the shortest cycle through all n vertices of G_n. Mézard - Parisi (1986, non-rigorous) and Wästlund (2006, rigorous) gave limit behavior, equivalent to the following.

$$\lim_n n^{-1} E \text{ len}(TSP(G_n)) = c = 2.04...$$

where c is specified by

$$c = \tfrac{1}{2} \int_0^\infty \; x P(X_1 + X_2 > x) \; dx$$

where X_1, X_2 are independent with the distribution determined as the solution of the RDE (recursive distributional equation)

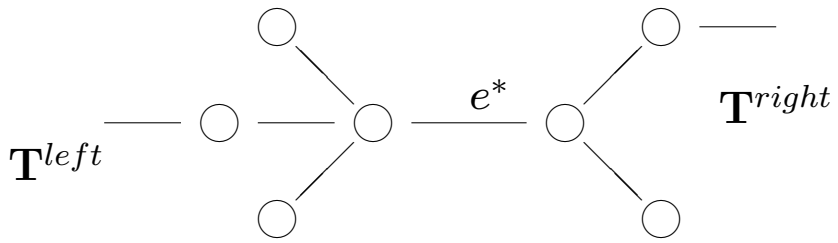
$$X \stackrel{d}{=} \min_i{}^{[2]} (\xi_i - X_i)$$

where $0 < \xi_1 < \xi_2 < \ldots$ are the points of a Poisson(rate 1) process, and $\min^{[2]}$ denotes second-minimum.

[show Java of solution]

Jump into heuristics [not easy to formalize]; consider optimal solution to TSP on the PWIT, without enquiring what this means precisely .... Note feasible solutions are collections of infinite paths.

Distinguished edge e^* of PWIT specifies two rooted subtrees $\mathbf{T}^{right}, \mathbf{T}^{left}$.



Optimal solution either uses/doesn't use e^*; in restriction of solution to $\mathbf{T}^{right}$ the root has either degree 1 or degree 2. So on $\mathbf{T}^{right}$ define heuristically

$$\begin{aligned} X^{right} &= \text{total length opt. solution, rootdegree } = 2 \\ &- \text{total length opt. solution, rootdegree } = 1. \end{aligned}$$

Optimal solution on $\mathbf{T}$ will use e^* iff

$$\text{len}(e^*) \leq X^{right} + X^{left}. \quad \text{[inclusion criterion].}$$

At a given vertex, the mean density of edges with lengths $\in [x, x + dx]$ equals dx; so contribution to mean length of TSP equals

$$\tfrac{1}{2} \int_0^\infty x \; P(x \leq X^{right} + X^{left}) \; dx.$$

Natural recursive structure of $\mathbf{T}^{right}$ leads to the RDE for distribution of X.

[blackboard]

Outline of general rigorous proof technique

Start with a combinatorial optimization problem over some size-n random structure.

- Find the $n \to \infty$ limit in the sense of local weak convergence; the "size-∞" random structure,

- Formulate a corresponding combinatorial optimization problem on the size-∞ structure – that is, specify "feasible solution" and "cost per vertex", requiring "stationarity".

- [start heuristics] Define relevant quantities on the size-∞ structure via additive renormalization.

- If the size-∞ structure is tree-like (the only case where one expects exact asymptotic solutions), observe that the relevant quantities satisfy a problem-dependent RDE.

- Solve the RDE. Use the unique solution and in "inclusion criterion" to calculate the value of the optimization problem on the size-∞ structure. [end heuristics].

- Check that the **recursive tree process** associated with the solution on the size-∞ structure, and the inclusion criterion, define a feasible solution. This has some cost c_∞, which we want to show is the limit.

- Show that the recursive tree process is **endogenous**, that is the root value is is a measurable function of the data. Since a measurable function is almost continuous, we can pull back to define almost-feasible solutions of the size-n problem with almost the same cost.

- Show that in the size-n problem one can patch an almost-feasible solution into a feasible solution for asymptotically negligible cost ($\Rightarrow$ upper bound).

- A weak limit of optimal solutions to the finite-n problems is a "stationary" feasible solution $\mathcal{E}^*$ on the size-∞ structure; endogeny couples this to the claimed optimal solution $\mathcal{E}$. Recycle the heuristic argument to show $\mathcal{E}^*$ cannot have lower cost than $\mathcal{E}$ ($\Rightarrow$ lower bound).

The 2 examples [Maximum Partial Matchings on Random Trees; TSP in mean-field model of distance] have a simplifying feature; there are only 2 options for a vertex at a cut edge, so the cost difference between the two optimal solutions is a single random variable X. There's one more interesting example with this feature [later: a challenge to make rigorous proof] but let me first indicate how heuristics apply in many cases where there are more than two options. In the next example we get a RDE for the distribution of a infinite sequence $(X_k, -\infty < k < \infty)$.

A new variant. Continue within the same framework of $G_n \to_{LWC} PWIT$. Note that distance $d(v, w)$ in G_n is shortest path length; for random V, W we have $d(V, W) \approx \log n$.

Fix $0 < q < 1/2$. In G_n take $\lfloor qn \rfloor$ random vertices to be **sources** and another $\lfloor qn \rfloor$ to be **destinations**. For any matching of sources v and destinations w there is a total transportation cost

$$\sum_{(v,w) \text{ in matching}} d(v, w)$$

and we study the minimum cost, say $M_{n,q}$, over all matchings of sources and destinations. As usual we expect a limit constant

$$n^{-1} E M_{n,q} \to c(q).$$

We first think a little about structure of optimal matching – view as a set of paths – in G_n. A given edge e might be in 0 or 1 or several paths. Orient paths from source to destination. If there are several paths through e they must all be oriented the same way through e [blackboard].

So, giving each edge e of G_n an arbitrary direction, associated with the optimal matching is a collection $-\infty < K(e) < \infty$ of random integers counting how many paths go through e.

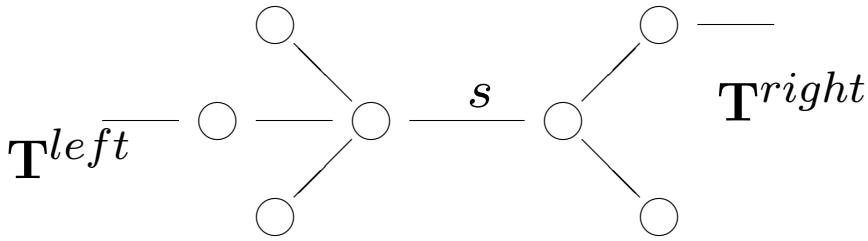
Now consider the analogous optimization problem on the PWIT – random sources and sinks at density q. Consider the optimal solution, relative to an edge of length s at the root. There are some random number $-\infty < K(s) < \infty$ of paths through the edge, and the limit constant is

$$c(q) = \tfrac{1}{2} \int_0^\infty E|K(s)| \; s \; ds.$$

Take a realization of the PWIT and the source and destination vertices, consider for each k the solution of the optimization problem modified by requiring k paths into the root from external sources (or $-k$ out from the root to external destinations); write

$$X_k = \text{optimal cost of } k\text{-modified problem}$$

$$- \text{ optimal cost of un-modified problem.}$$



$$K(s) = \arg\min_k (s|k| + X'_k + X''_{-k})$$

where (X'_k) and (X''_k) are independent copies of (X_k).

So heuristics give formula for optimal cost in terms of dist of (X_k).

To write out the RDE, there are several cases.

In case root is not source or destination

[blackboard]

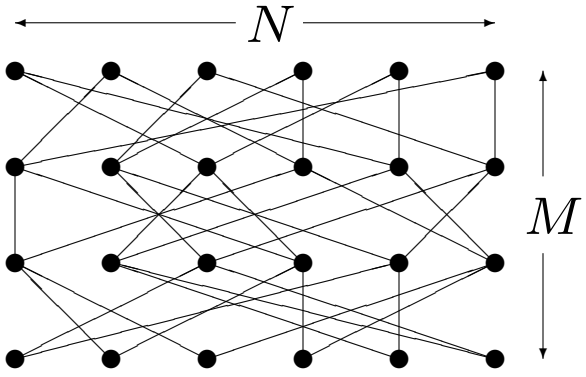
$$X_k = \min_{(k_i):\sum_i k_i = k} \sum_i (X^i_{k_i} + \xi_i |k_i|)$$

Any volunteers to write code to simulate solution?

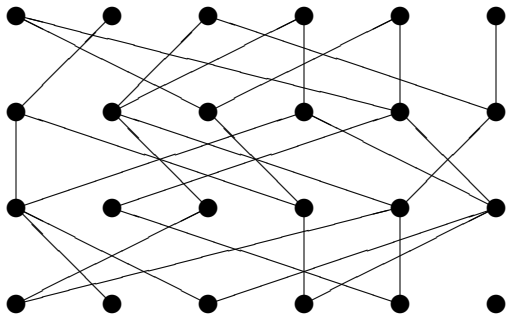
Example: flow through a disordered network. A challenge to make rigorous . . . . . .

Consider a network with

- M layers
- N vertices per layer
- directed edges upwards from one layer to next
- edges between successive layers are placed randomly subject to each vertex having in-degree = out-degree = 2.



$p = 1$



$p = 0.8$

Special problem. Suppose
• edges have capacity = 1.
• retain each edge with probability p, delete with probability $1 - p$.
Study maximum flow from bottom to top layers; same as maximum number of edge-disjoint paths from bottom to top layers. Clearly for $p = 1$ the maximum flow $= 2N$, so for general p we consider the relative flow

$$F_{N,M}(p) = \tfrac{1}{2N} \times (\text{max flow through network}).$$

We anticipate a limit function

$$EF_{N,N}(p) \to v^*(p) \text{ as } n \to \infty.$$

Cavity method tells you how to write down an equation whose solution determines $v^*(p)$.

Lecture 3 part 2 moves away from CO to

A Tractable Stochastic "Complex Network" Model

short paper: *AMS Electronic Research Announcements*, 2003.

longer paper: `arXiv:cond-mat/0304701` published in Springer Lect. Notes Physics 650.

Slides only minimally updated since 2003, though much subsequent work.

1. Recent literature.

Three popular science books, a dozen articles in *Science* and *Nature*, and 154 preprints at `xxx.arXiv.org/cond-mat` deal with *complex networks*, which in this context means the empirical and theoretical study of large graphs, focusing in particular on those possessing the following three qualitative properties, asserted to hold in many interesting real-world examples.

- the degree distribution has power-law tail
- local clustering of edges: graph is not locally tree-like
- small diameter – $O(\log(\text{number of vertices}))$.

The nature of that subject – typically not presented as rigorous mathematics – is most easily seen from the long survey papers

R. Albert and A.-L. Barabási, *Statistical mechanics of complex networks*, Rev. Mod. Phys. **74** (2002), 47–97.

S.N. Dorogovtsev and J.F.F. Mendes, *Evolution of networks*, Adv. Phys. **51** (2002), 1079–1187.

M.E.J. Newman, *The structure and function of complex networks*, SIAM Review **45** (2003), 167–256.

A shorter survey emphasizes rigorous mathematical results

B. Bollobás and O. Riordan, *Mathematical results on scale-free random graphs*, Handbook of Graphs and Networks (S. Bornholdt and H.G. Schuster, eds.), Wiley, 2002.

See also Durrett lecture notes (Fall 2004).

Almost all this literature concerns variants of two modelling ideas.

Small worlds.

- Take n-vertex lattice-neighborhood graph
- Add long edges in some random way.

Proportional attachment.

- Vertices arrive sequentially ($n = 1, 2, 3, \ldots$);
- each vertex attaches to k existing vertices v chosen with probabilities proportional to $c + \text{degree}(v)$.

Some natural summary statistics for a complex network.

• $\bar{\partial}$ = average vertex-degree

• an exponent γ indicating power-law tail behavior of degree distribution

• a "clustering coefficient" κ measuring relative density of triangles

• the average distance $\bar{\ell}$ between vertex-pairs.

Desiderata for a stochastic model

• *mathematical tractability*: one can find reasonably explicit formulas for a variety of quantities of interest

• *fitting flexibility*: by varying model parameters one can vary summary statistics (like the 4 listed above) broadly through their possible ranges

• *naturalness*: the qualitative properties emerge from some simple underlying mathematical structure rather than being forced by fiat.

No ideal model known. I will describe a specific two-parameter model, and implicitly a class of models, which satisfy many of these desiderata.

Comment: we are accustomed to models (percolation, interacting particle systems) which are simple to state but complicated to analyze. In contrast, this model is conceptually sophisticated to state but easy (in some respects) to analyze and get explicit formulas.

Platform for model: directed graphs

(i) Vertices $v, w, x, \ldots$ arrive sequentially; some intrinsic "geometry" given by distances $d(v, w)$.
(ii) Given $1 \geq p(r) \downarrow 0$ fast as $r \uparrow \infty$.
(iii) When vertex v arrives, for each existing vertex w and each existing edge (w, x), new edges (v, w) and (v, x) appear independently with probability $p(d(v, w))$.

We will build this over a version of our "mean-field model of distance". Loosely, the model is between the extremes of
(a) lattice-based small world models
(b) proportional attachment/copying models.

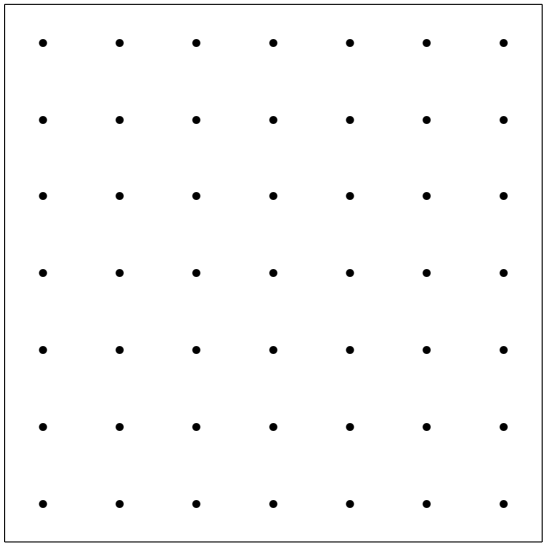
In (iii) we somewhat arbitrarily take

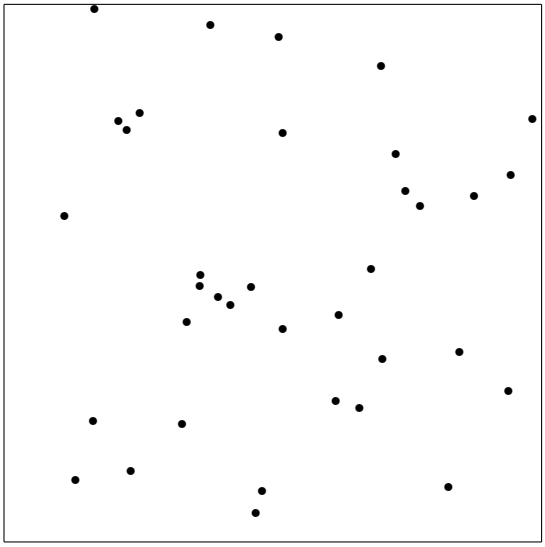
$$p(r) = \min(1, \alpha\lambda e^{-\lambda r})$$

with two parameters $\alpha, \lambda > 0$.

Geometry of n points as $n \to \infty$.

In d dimensions, pictured for $d = 2$. Could take the points ordered or random, in region of area n. In either case there is a $n \to \infty$ limit: the infinite lattice, or the Poisson point process.





Draw attention to one feature of each.

• On lattice, point has $2d$ "near neighbors".

• Poisson process is time-equilibrium of a certain space-time process, in which points move to infinity as deterministic exponentials $x(t_2) = x(t_1)e^{(t_2-t_1)/d}$ and new random points arrive at space-time rate 1. "Enterprise under warp drive", or Hoyle's 1950s steady-state model of Universe.

Conceptually new ingredient : we can also regard the PWIT as the equilibrium spatial distribution of a space-time process. Existing edge-lengths $\ell(t)$ increase as determinstic exponentials

$$\ell(t_2) = \ell(t_1)e^{t_2 - t_1}.$$

New vertices appear and by fiat are at Poisson(rate 1) distances $0 < \xi_1 < \xi_2 < \ldots$ from existing points (changing the geometry, cf. "wormholes").

We will do calculations within this space-time PWIT model. So we are "cheating" in first devising a limit model; here's how we "reverse engineer" to represent it as a limit of finite models.

The finite model.

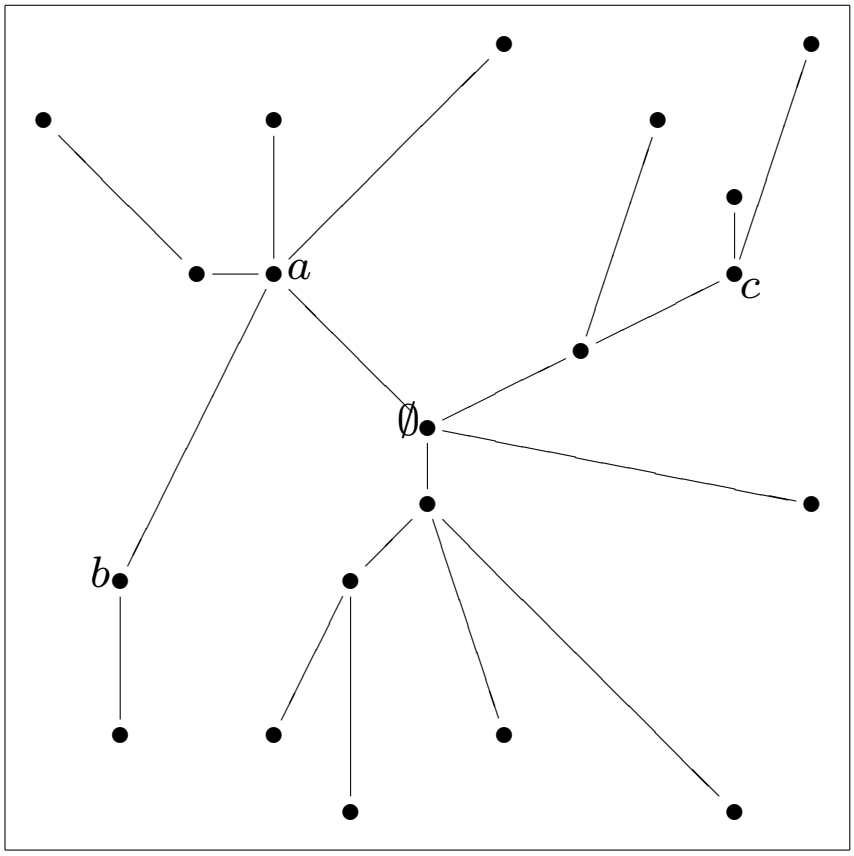
• vertex n arrives at time $\log n$
• the link lengths $\ell(\cdot)$ from n to previous $n-1$ vertices are independent Exp (mean n) r.v.'s
• lengths increase deterministically with time, at exponential rate 1.

Over this model of geometry we build our complex network model as described earlier (but using link lengths rather than shortest-path distances).

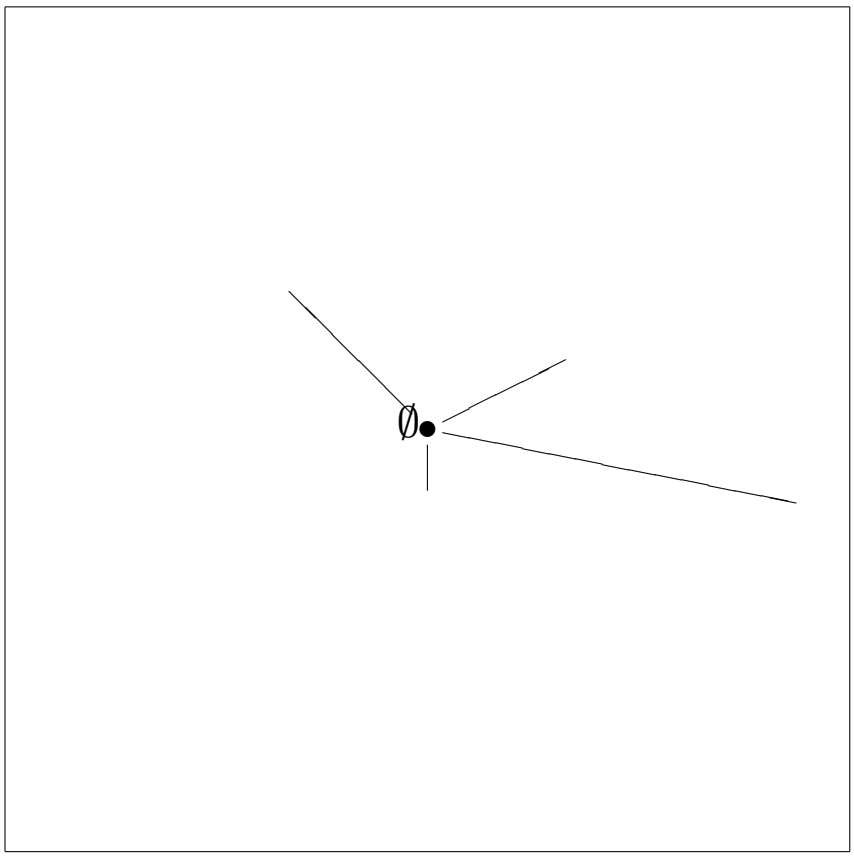
• When vertex v arrives, for each existing w and each existing edge (w, x), new edges (v, w) and (v, x) appear independently with probability $p(\ell(v, w))$. Note: only the "near neighbors" w with small $\ell(v, w)$ matter.

Next 9 slides provide a low-tech simulation – volunteers to write a high-tech simulation?

The space-time geometry process, conditioned on a particular vertex $\emptyset$ arriving at a particular time 0. By PASTA, the spatial geometry is the PWIT.

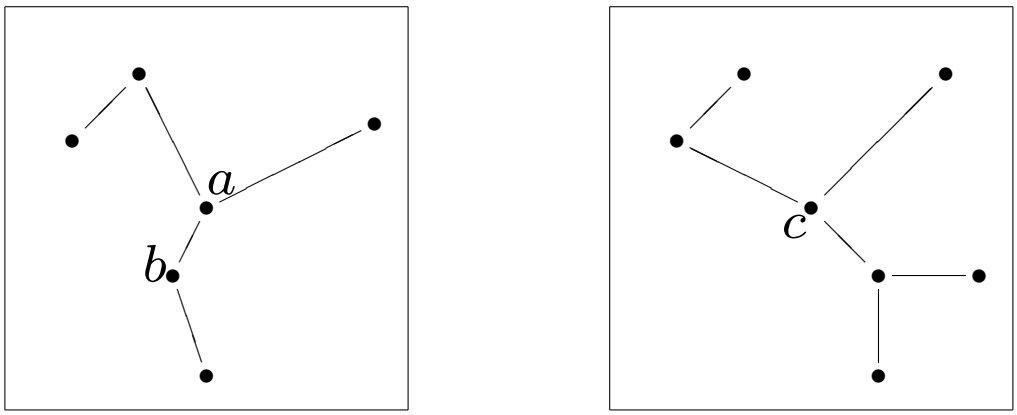


What just happened at time 0− was that ∅ arrived, was assigned random link-lengths and random neighbors.



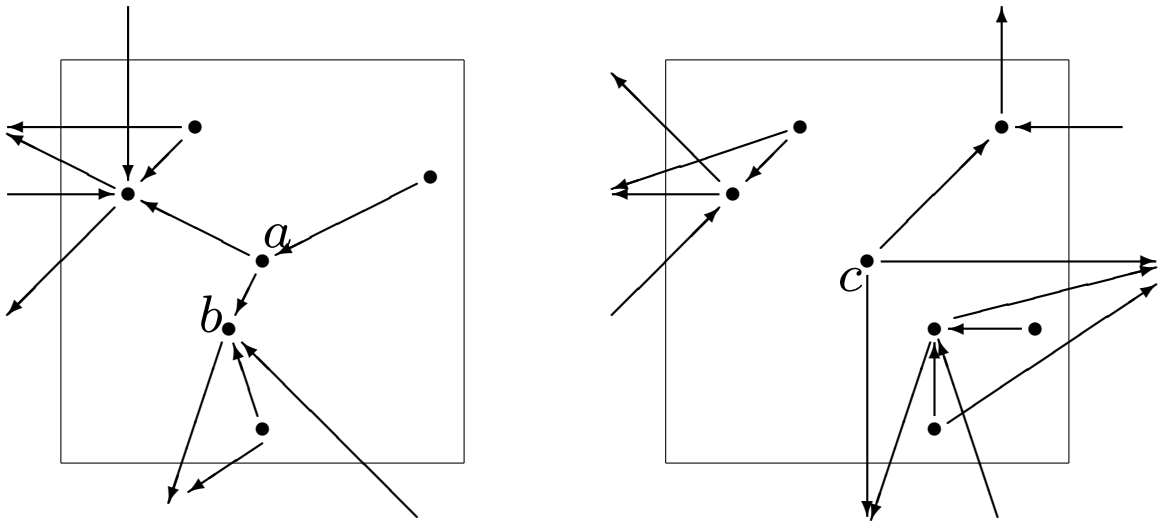
This causes previously disconnected (in the PWIT; far apart in the finite model) components to become connected via ∅.

Here are two of the previously disconnected components at an earlier time; edges were shorter.

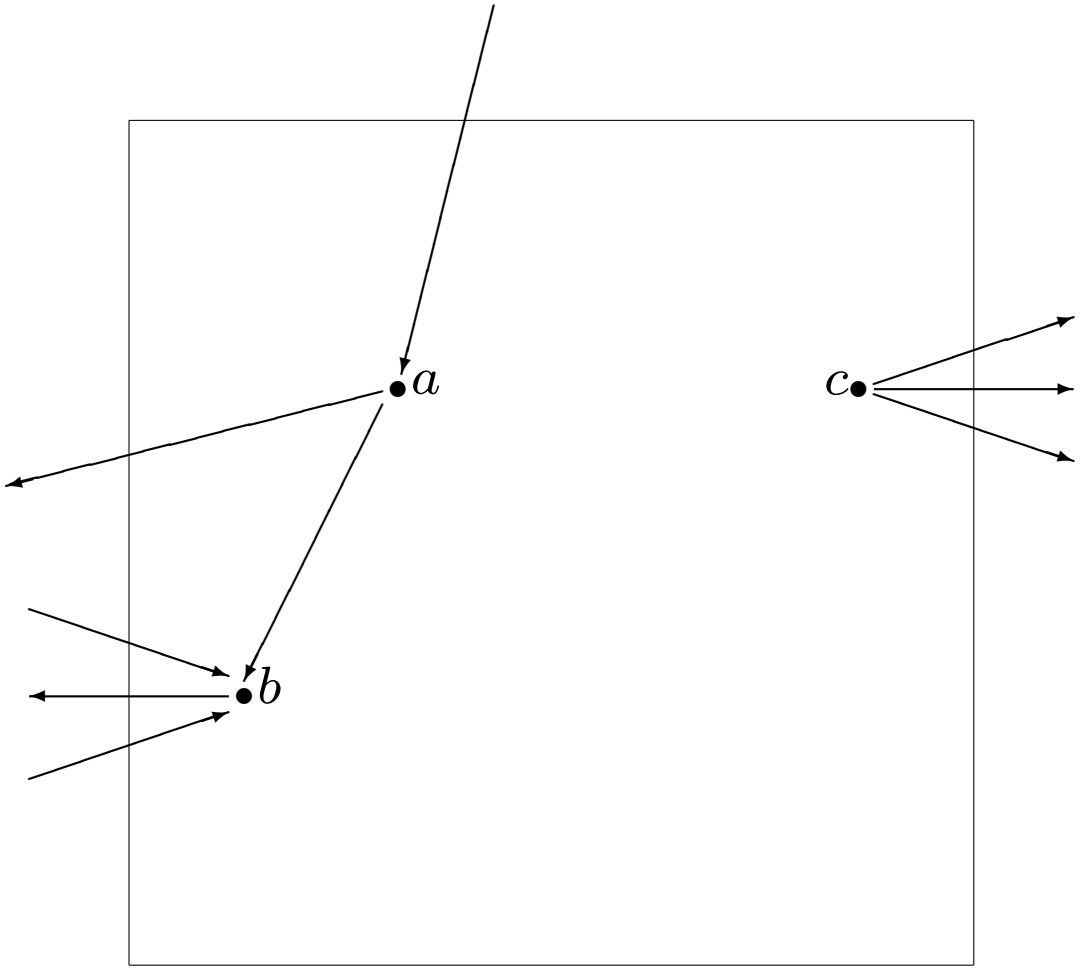


So far we're just showing the geometry of our "steady-state expanding universe". Now we superimpose the random directed graph structure.

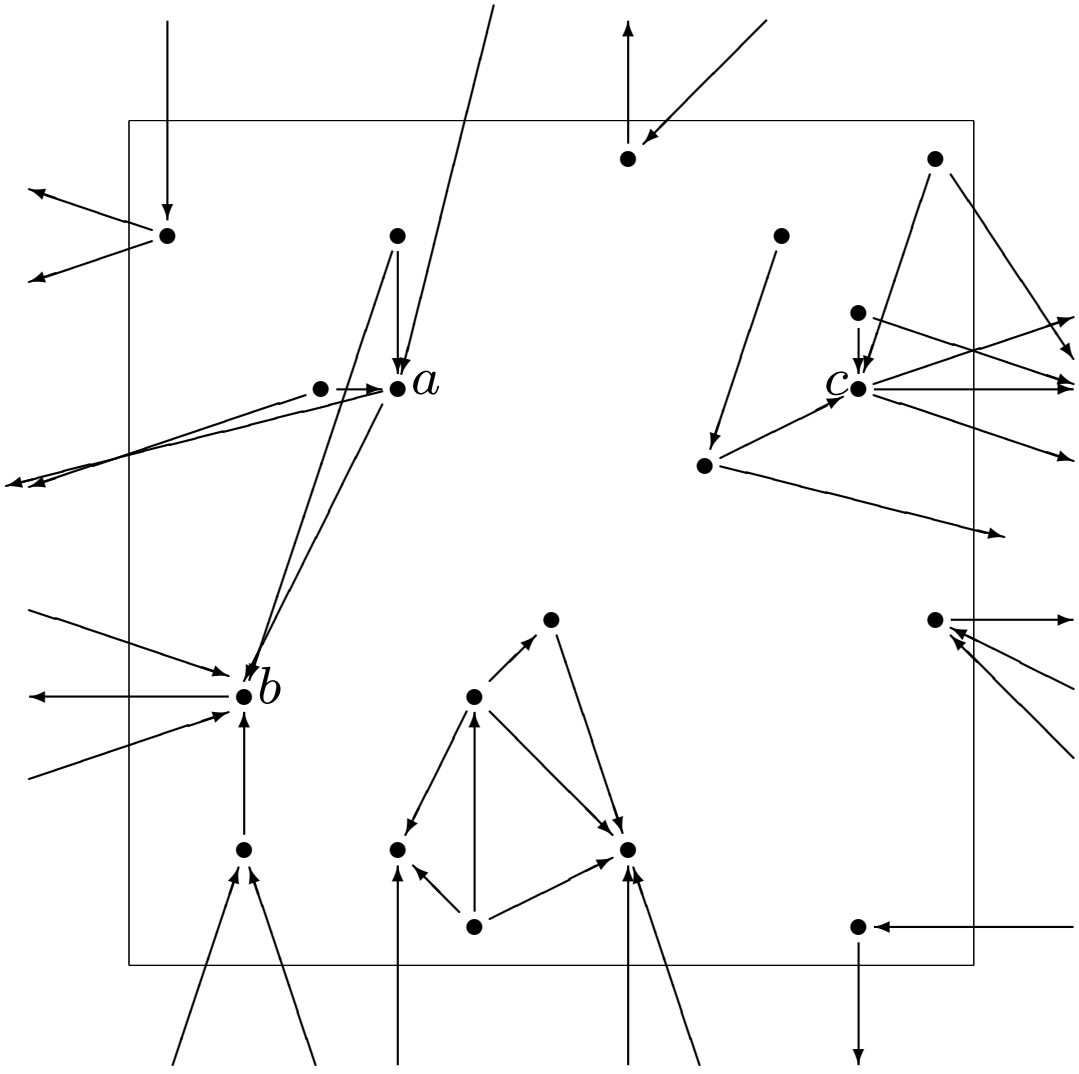
The random graph structure on the two previously disconnected components at an earlier time.



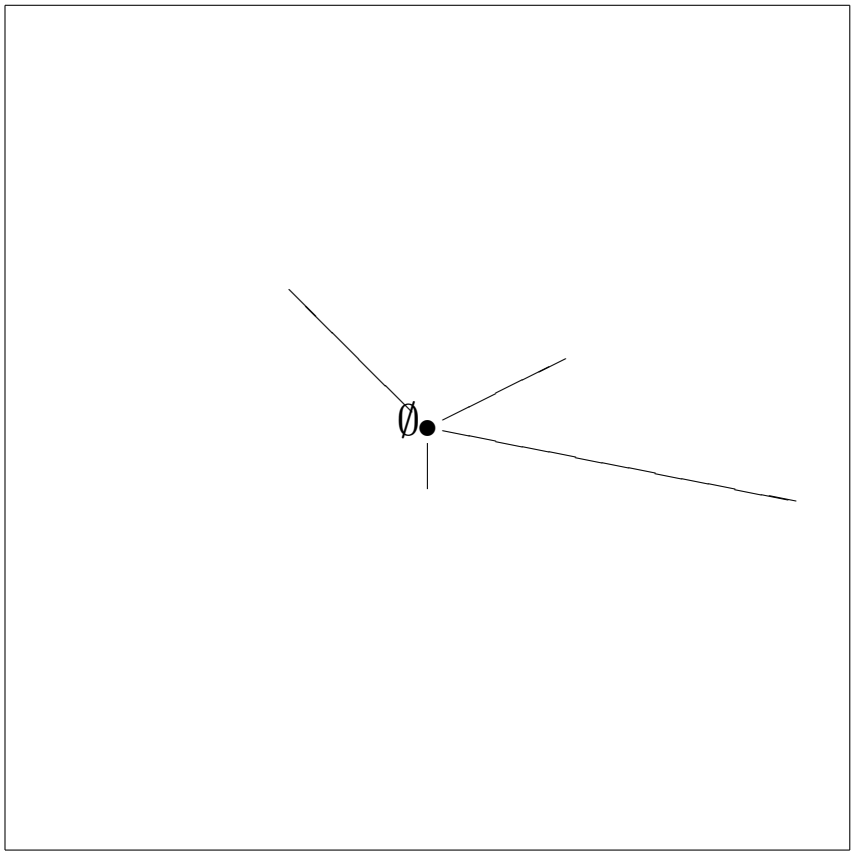
. . . . . these edges have expanded at time 0−



and more vertices have joined these components by time 0−

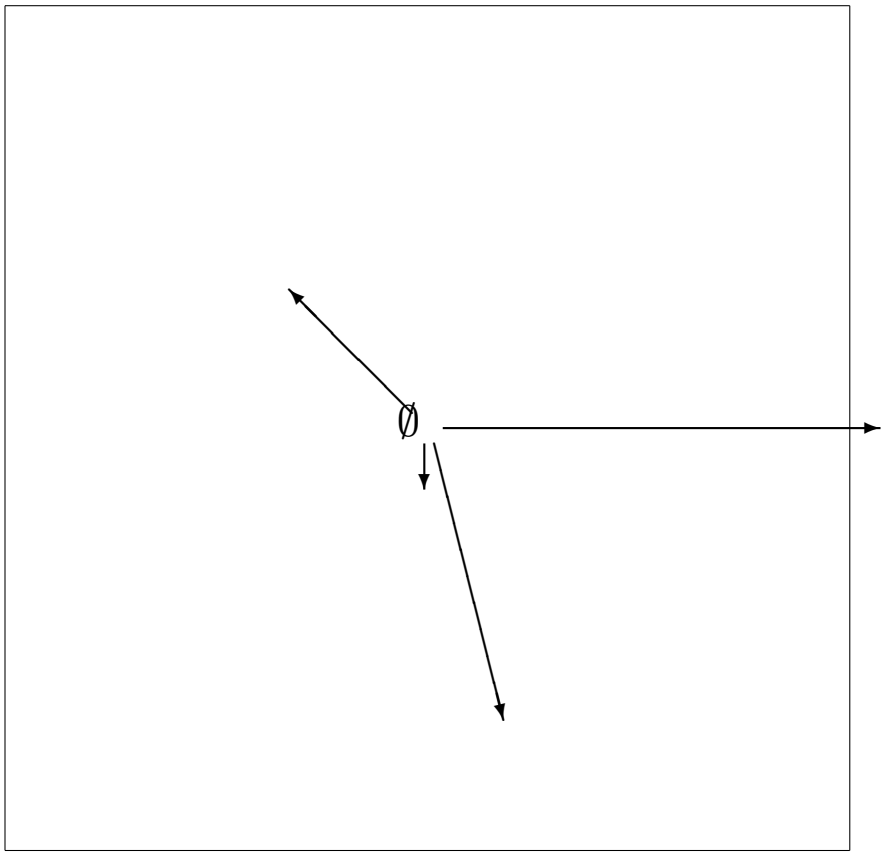


Now $\emptyset$ arrived and is assigned random link-lengths and random neighbors.

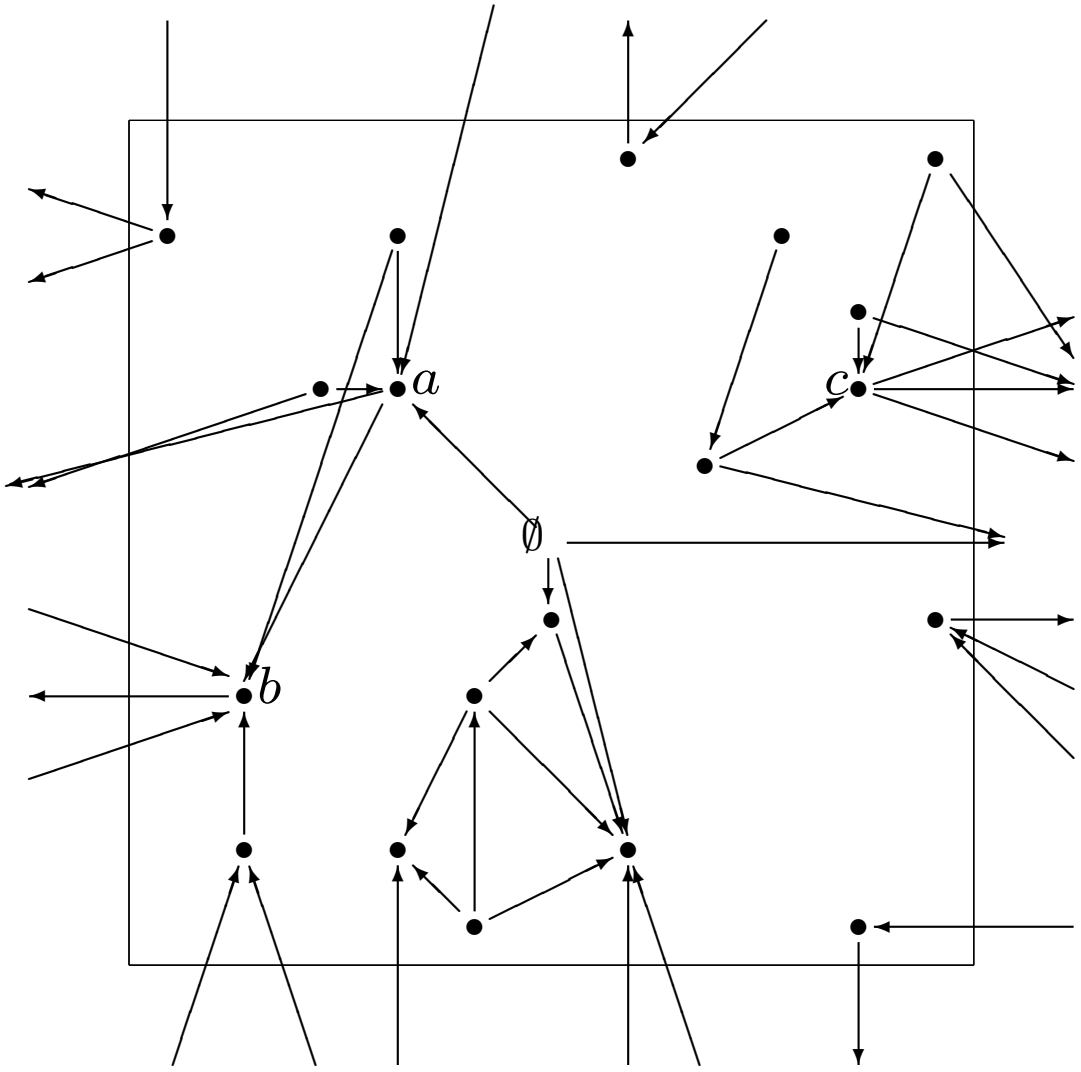


For each neighbor w there are separate random choices; whether or not to put an edge of our random graph from $\emptyset$ to w, and whether or not to copy existing edges (w, y) to new edges $(\emptyset, y)$ . . .

. . . and here are the choices made.



So here is the state of the random graph after $\emptyset$ arrives.



This may seem insanely complicated, but as said before:

We are accustomed to models which are simple to state but complicated to analyze. In contrast, this model is conceptually sophisticated to state but easy (in some respects) to analyze and get explicit formulas. That is, explicit formulas in the PWIT model, representing $n \to \infty$ asymptotics in the finite model.

Tractable because

• everything is time-invariant, so can immediately write down various equations, e.g. for out-degree D

$$D \stackrel{d}{=} \sum_{i=1}^{\infty} \text{Bin}(1 + D_i, \alpha\lambda e^{-\lambda\xi_i}) \quad \text{[low]}$$

• the process "1 + in-degree(v) at time t" is precisely a Yule process.

End talk with

GALLERY OF EXPLICIT FORMULAS

exact in $n \to \infty$ limit.

Recall the two parameters enter via the function

$$p(r) = \min(1, \alpha\lambda e^{-\lambda r}), \quad 0 \le r < \infty.$$

We will need to distinguish between a *low clustering region* with parameter ranges

$$0 < \alpha < 1, \quad 0 < \lambda \le 1/\alpha \qquad \text{[low]}.$$

and the complementary *high clustering region* where $\alpha\lambda > 1$. In the latter case
$p(r) = 1, \ r \le \eta := \lambda^{-1} \log(\alpha\lambda)$
and it is convenient to reparametrize using η in place of α, making the parameter range

$$0 < \eta < 1, \quad \eta + 1/\lambda < 1. \qquad \text{[high]}.$$

Greek letters denote quantities computable from parameters.

The two parameters control mean degree and clustering.

First consider D_{in} and D_{out}, the random in-degree and out-degree of a typical vertex. Then

$$ED_{\text{in}} = ED_{\text{out}} = \bar{\partial} = \begin{cases} \frac{\alpha}{1-\alpha} & \text{[low]} \\ \frac{\eta+1/\lambda}{1-\eta-1/\lambda} & \text{[high]} \end{cases} \quad (1)$$

Second, define a normalized *clustering coefficient* κ_{cluster} as

> The proportion of directed 2-paths $v_1 \to v_2 \to v_3$ for which $v_1 \to v_3$ is also an edge.

Then

$$\kappa_{\text{cluster}} = \begin{cases} \frac{\alpha(1-\alpha)\lambda}{2-\alpha^2\lambda} & \text{[low]} \\ \frac{(\eta+\frac{1}{2\lambda})(1-\eta-\frac{1}{\lambda})}{(\eta+\frac{1}{\lambda})(1-\eta-\frac{1}{2\lambda})} & \text{[high]} \end{cases} \quad (2)$$

By solving (1,2) we find that every pair of values of $\bar{\partial}, \kappa_{\text{cluster}}$ in the complete range

$$0 < \bar{\partial} < \infty, \quad 0 < \kappa_{\text{cluster}} < 1$$

occurs for a unique parameter pair (α, λ) or (η, λ). Moreover the two regions can be specified as

$$0 < \bar{\partial} < \infty, \quad 0 < \kappa_{\text{cluster}} \leq \frac{1}{\bar{\partial}+2} \quad \text{[low]}$$
$$0 < \bar{\partial} < \infty, \quad \frac{1}{\bar{\partial}+2} < \kappa_{\text{cluster}} < 1. \quad \text{[high]}$$

So the two model parameters α, λ have fairly direct interpretations in terms of mean degree and clustering; of course we could re-parametrize the model in terms of $\bar{\partial}$ and κ_{cluster}, but the internal mathematical structure is more conveniently expressed using the given parameters.

Distribution of in-degree

The distribution of D_{in} is specified as

$$1 + D_{\text{in}} \stackrel{d}{=} \text{Geo}(e^{-\beta T}) \text{ where } T \stackrel{d}{=} \text{Exp}(1)$$

and where

$$\beta = \begin{cases} \alpha & \text{[low]} \\ \eta + 1/\lambda & \text{[high]} \end{cases}$$

This works out explicitly as

$$P(D_{\text{in}} = d) = \frac{\Gamma(d+1)\Gamma(1/\beta)}{\beta^2 \Gamma(d+2+\frac{1}{\beta})}, \quad d \geq 0 \quad (3)$$

with asymptotics

$$P(D_{\text{in}} = d) \sim \beta^{-2}\Gamma(1/\beta)\ d^{-1-\frac{1}{\beta}}.$$

Formula (3) appears in recent proportional attachment models, but in fact is a famous 80-year old calculation.

Distribution of out-degree

Distribution of D_{out} determined by the identity

$$D \stackrel{d}{=} \sum_{i=1}^{\infty} \text{Bin}(1 + D_i, \alpha\lambda e^{-\lambda\xi_i}) \quad \text{[low]}$$

where $D,\ D_i,\ i \geq 1$ are independent with the distribution of D_{out} and where $0 < \xi_1 < \xi_2 < \ldots$ are the points of a rate-1 Poisson point process on $(0, \infty)$.

We do not know how to extract a useful explicit formula from the identity, but we can compute moments. For instance

$$\text{var } D_{\text{out}} = \begin{cases} \dfrac{\alpha(1-\alpha+\alpha^2\lambda/2)}{(1-\alpha)^2(1-\frac{1}{2}\alpha^2\lambda)} & \text{[low]} \\ \dfrac{(\eta+\frac{1}{2\lambda})(2-\eta-\frac{1}{\lambda})}{(1-\eta-\frac{1}{2\lambda})(1-\eta-\frac{1}{\lambda})^2} & \text{[high]} \end{cases}$$

Note also

$$D_{\text{in}} \text{ and } D_{\text{out}} \text{ are independent.}$$

Densities of induced subgraphs

Let G be a finite directed acyclic graph. We expect a limit

$$\chi(G) = \lim_n \frac{\#\text{subgraphs of } \mathcal{G}_n \text{ isomorphic to } G}{n}.$$

For the complete directed acyclic graph K_r on $r \geq 2$ vertices,

$$\chi(K_r) = \prod_{u=1}^{r-1} \frac{\beta_u}{1-\beta_u}.$$

$$\beta_u := \begin{cases} u^{-1}\alpha^u\lambda^{u-1} & \text{[low]} \\ \eta + \frac{1}{u\lambda} & \text{[high]} \end{cases}$$

For the complete bipartite directed graph $K_{2,2}$, for $\beta_2 < \frac{1}{2}$ (which always holds in the low density case), the corresponding limit for "subgraphs including $K_{2,2}$" is

$$\tfrac{1}{2}\bar{\chi}(K_{2,2}) = \frac{\bar{\partial}\beta_2(\beta_2 + \frac{1}{2}\bar{\partial}\beta)}{(1-2\beta_2)(1-\beta_2)}.$$

Triangle density as a function of degree

The parameter κ_{cluster} gives an overall measure of triangle density. A more detailed description is provided by statistics $C(k),\ k \geq 2$ defined by

$$C(k) = \frac{E(\#\text{ triangles contain. random degree-}k\text{ vertex})}{\binom{k}{2}}.$$

In principle could obtain exact formula for $C(k)$, but easier to get the tail property

$$C(k) \sim \frac{2\beta_2}{\beta - \beta_2} \times \frac{1}{k} \text{ as } k \to \infty.$$

Relates to suggestion that property $C(k) \sim c/k$ indicates "hierarchical structure" in complex networks.

Edge-lengths

Our model has a "metric structure", meaning that there is a distance $d_{\text{metric}}(v, w)$ between any two vertices which does not involve the realization of edges in the random graph. So each edge (v, w) of the graph has a real-valued length $d_{\text{metric}}(v, w)$, and so a typical edge has a random length L. The probability density function for L is given by the formula

$$\frac{1-\alpha}{\alpha} \sum_{i=0}^{\infty} \frac{(i+1)\Gamma(\alpha+3)\ (-\lambda x)^i}{\Gamma(i+\alpha+3)},\ 0 < x < \infty \quad \text{[low]}.$$

and $f(x) \approx \exp(-(\lambda \pm o(1))x)$ as $x \to \infty$. In the underlying metric space, the number of vertices within distance k of a typical vertex grows as e^k. So one can give a rough reinterpretation of the tail behavior of $f(x)$ as

the chance that a vertex has an edge to its k'th nearest neighbor should scale as $k^{-\lambda-1}$.

Note this property appears without being explicitly built into the model.

Advantages/disadvantages of the model:

• it has the three qualitative features desired in a complex networks model (power-law degree distribution, clustering, small diameter)
• it fits the complete possible range of mean degree (or scaling exponent) and clustering parameters
• it permits a broad range of explicit calculations.

****

• $\mathcal{G}_n$ is not connected (for large n).
• There is no power law for distribution of out-degree.
• in-degree and out-degree are independent.
• The scaling exponent for in-degree is determined by the mean degree; one might prefer a model where these could be specified separately.
• In the $n \to \infty$ limit not every finite graph is possible as an induced subgraph.