A Tractable Stochastic "Complex Network" Model

David Aldous, U.C. Berkeley.

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

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

Slides on my home page.

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 $c +$ 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.

Outline of slides

7 − 8	half description of model
9 - 16	gallery of explicit formulas
17 - 28	complete description of model
29	now see why we can calculate things
30 +	the Yule process, local weak convergence,

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

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 $p(d(v, w))$.

So (i) is reminiscent of lattice-based small worlds, and (iii) of proportional attachment/copying.

Our specific model uses in (i) a model of "random points in infinite-dimensional space" – details later. Has property: number of vertices within distance r of new vertex grows as e^r. This "geometry" model has no dimensionless parameters.

In (iii) we somewhat arbitrarily take

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

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

Model gives evolving random graph $\mathcal{G}_n$ on n vertices. **Designed so that** there is a $n \to \infty$ limit random graph $\mathcal{G}^*_\infty$ representing "the limit as seen from a typical vertex". By doing calculations within the limit graph, we get exact formulas in the $n \to \infty$ 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.

GALLERY OF EXPLICIT FORMULAS
exact in $n \to \infty$ limit.
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

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

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

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

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

$$\begin{array}{ll} 0 < \bar{\partial} < \infty, \ \ 0 < \kappa_{\text{cluster}} \le \frac{1}{\bar{\partial}+2} & \text{[low]} \\ 0 < \bar{\partial} < \infty, \ \ \frac{1}{\bar{\partial}+2} < \kappa_{\text{cluster}} < 1. & \text{[high]} \end{array}$$

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

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

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

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

where $D,\ D_i,\ i \geq 1$ are independent with the distribution of D_{out} and where $0 < \xi_1 < \xi_2 < \dots$ 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_{in} and D_{out} 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 $d_{\mathrm{metric}}(v,w)$ 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 $d_{\mathrm{metric}}(v,w)$, 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:

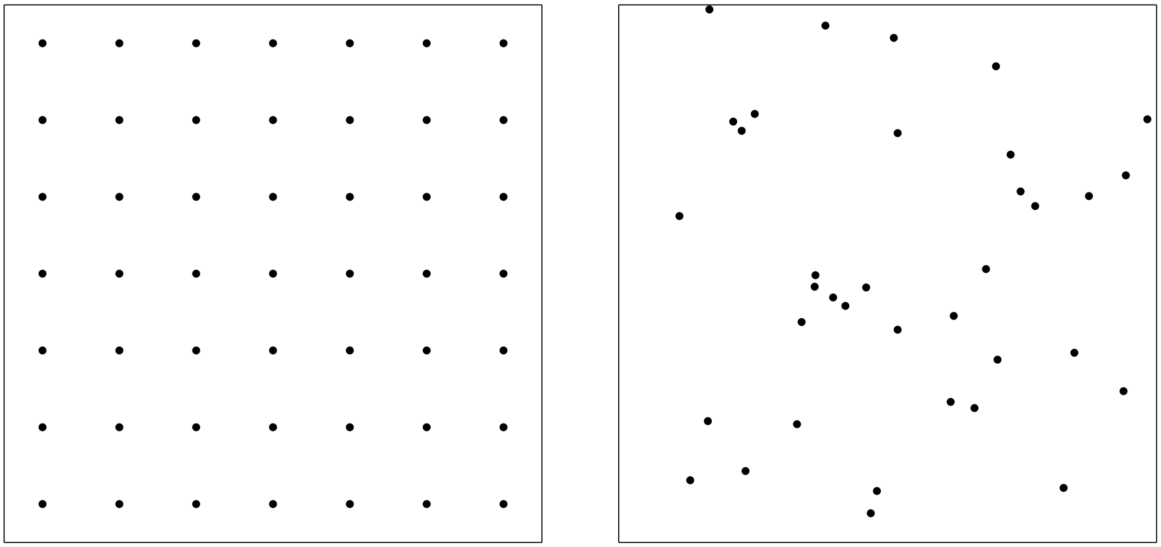
- it has the three qualitative features desired in a complex networks model (power-law degree distribution, clustering, small diameter)
- it fits the complete possible range of mean degree (or scaling exponent) and clustering parameters
- it permits a broad range of explicit calculations.

****

- $\mathcal{G}_n$ 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.

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.

There are various names (random link; randomly-weighted complete graph; mean field model of distance) for a **static model of n points**: take the $\binom{n}{2}$ inter-point link lengths to be independent r.v.'s with Exp (mean n) distribution. Then set distance = length of shortest path.

Turns out (2 minutes thought!) there is $n \to \infty$ limit geometry, as seen from typical point $\emptyset$. Limit geometry is the PWIT. One property: E(number points within distance r of $\emptyset$) $= e^r$.

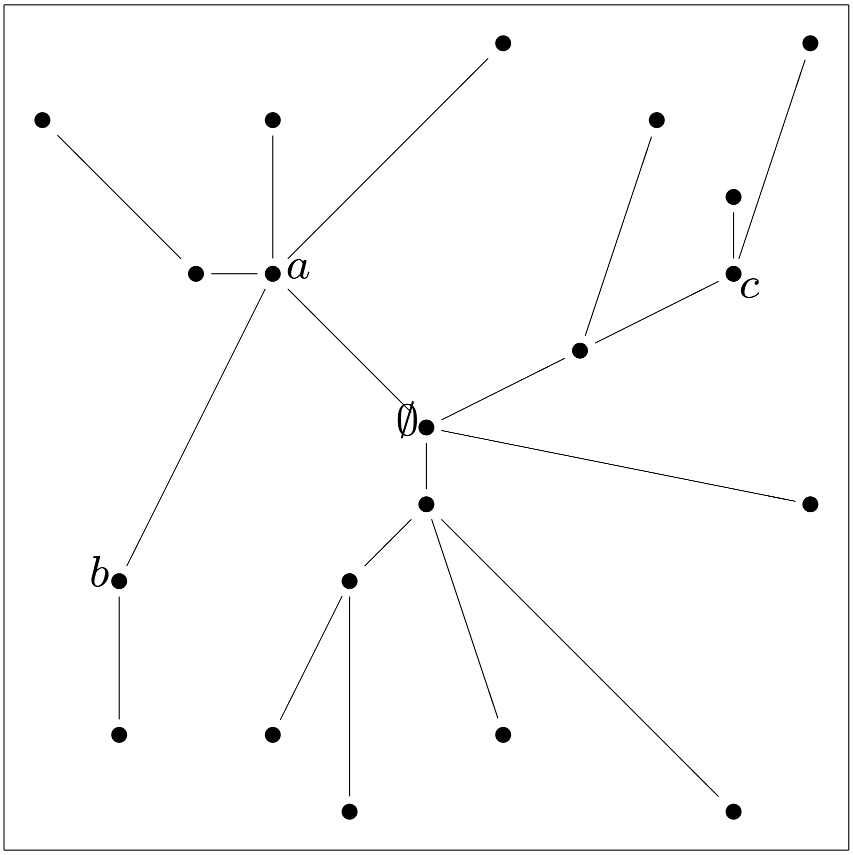
Moreover, PWIT is time-invariant distribution of the space-time "steady-state Universe" process, as in 2 dimensions.

Finally, we formulate the finite-n model of arriving vertices and evolving geometry which has the space-time PWIT as its $n \to \infty$ limit.

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

Over this model of geometry we build our complex network model as described earlier, but with a restriction to near neighbors.

- When vertex v arrives, for each existing near neighbor w and each existing edge (w, x), new edges (v, w) and (v, x) appear independently with probability $p(d(v, w))$.

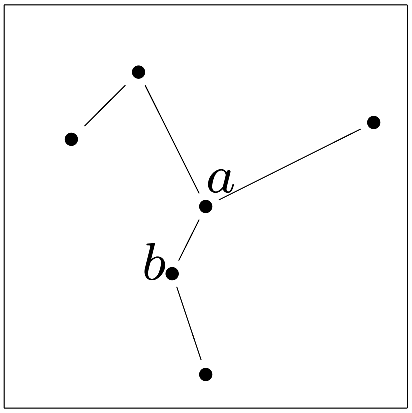

a
c
∅
b

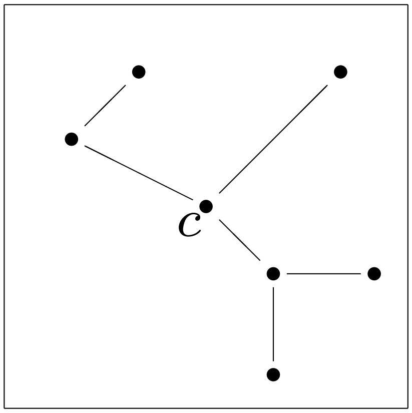
The PWIT – a window centered on point $\emptyset$.

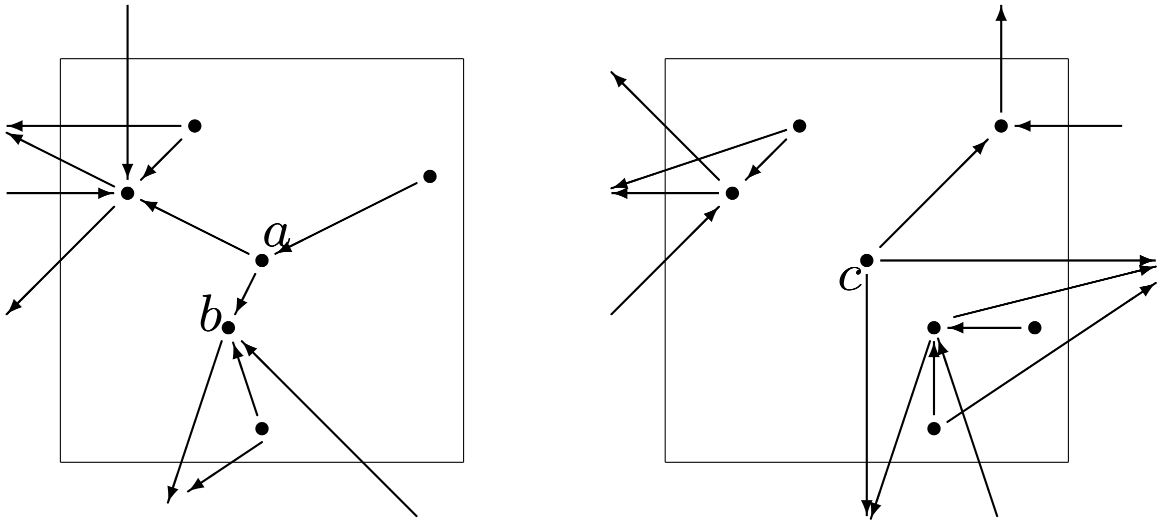
Distances $0 < \xi_1 < \xi_2 < \xi_3 < \ldots$ from a vertex to its **near neighbors** (indicated by lines) are successive points of a Poisson (rate 1) process on $(0, \infty)$. Continue recursively.

An earlier time, when only the three vertices a, b, c from current-time window had arrived.

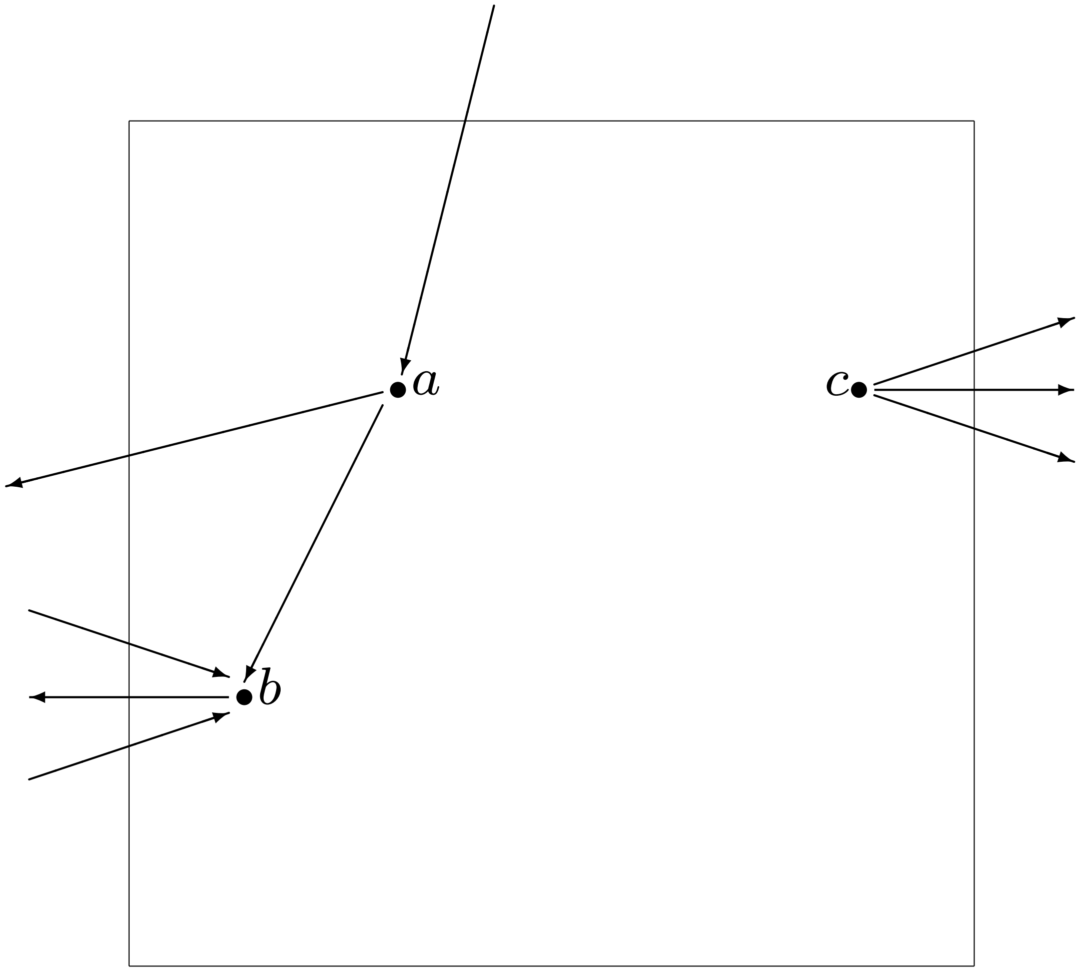
.

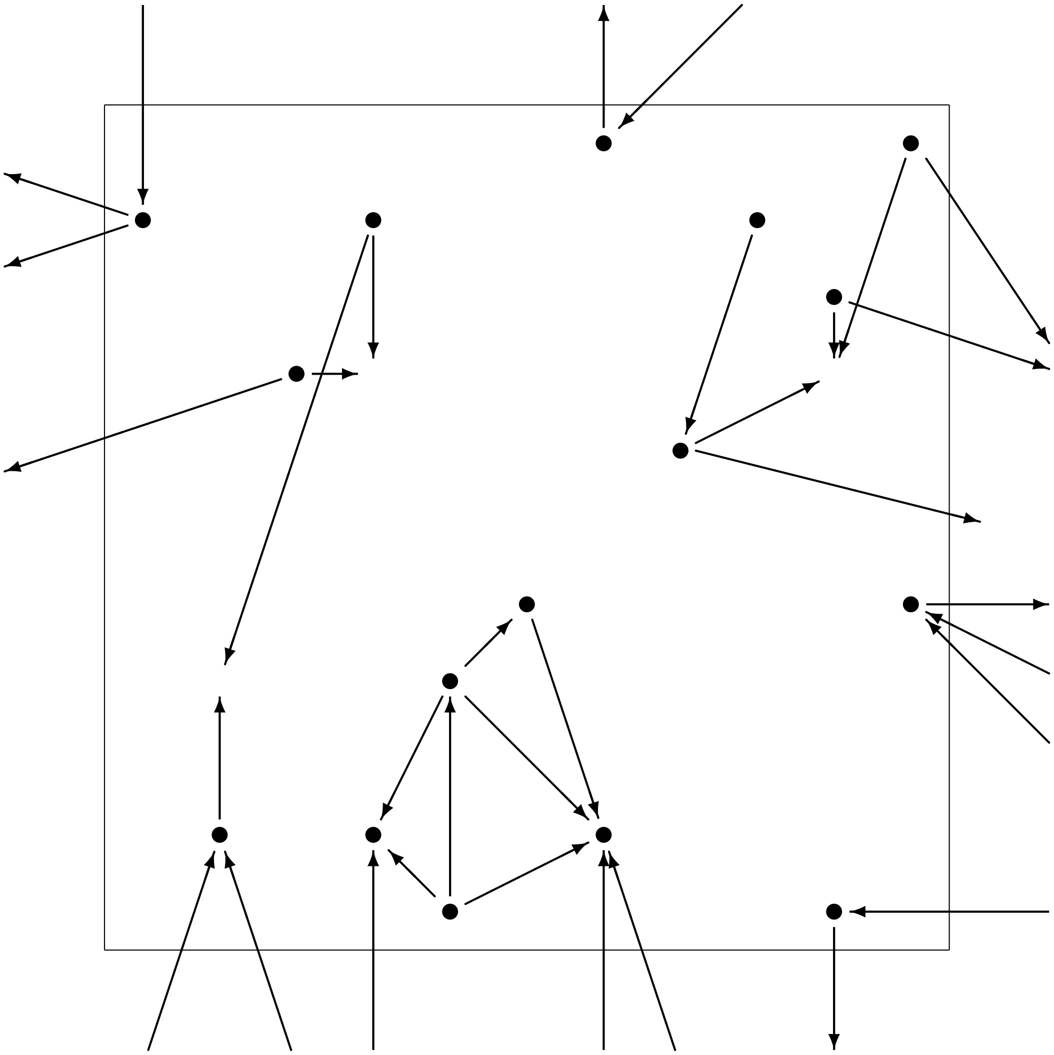

a
b

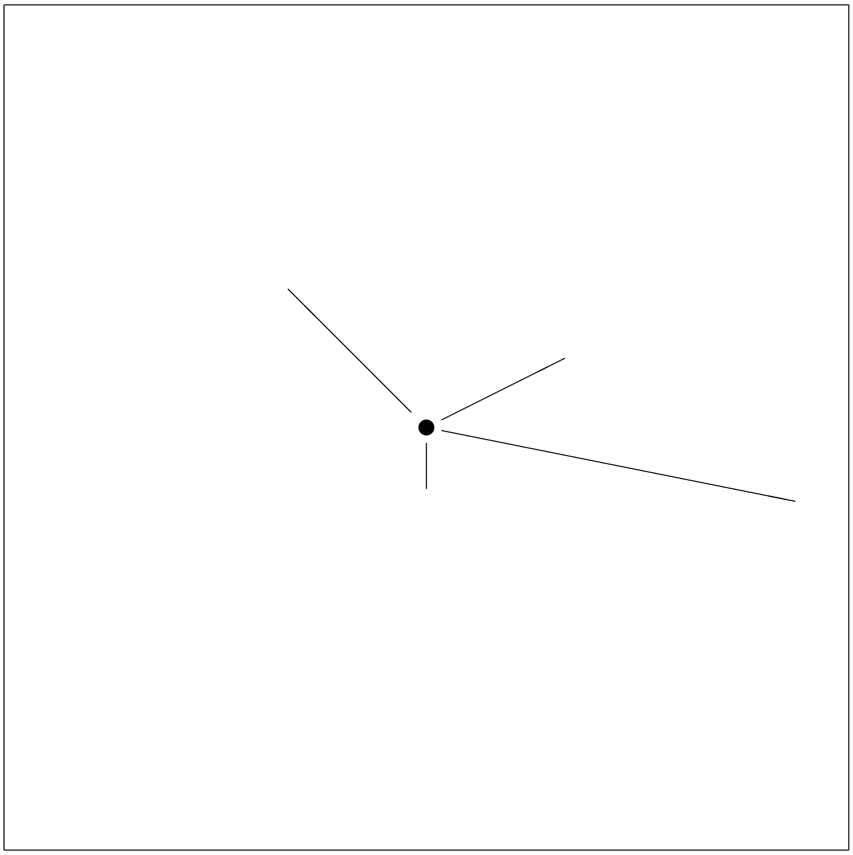

c

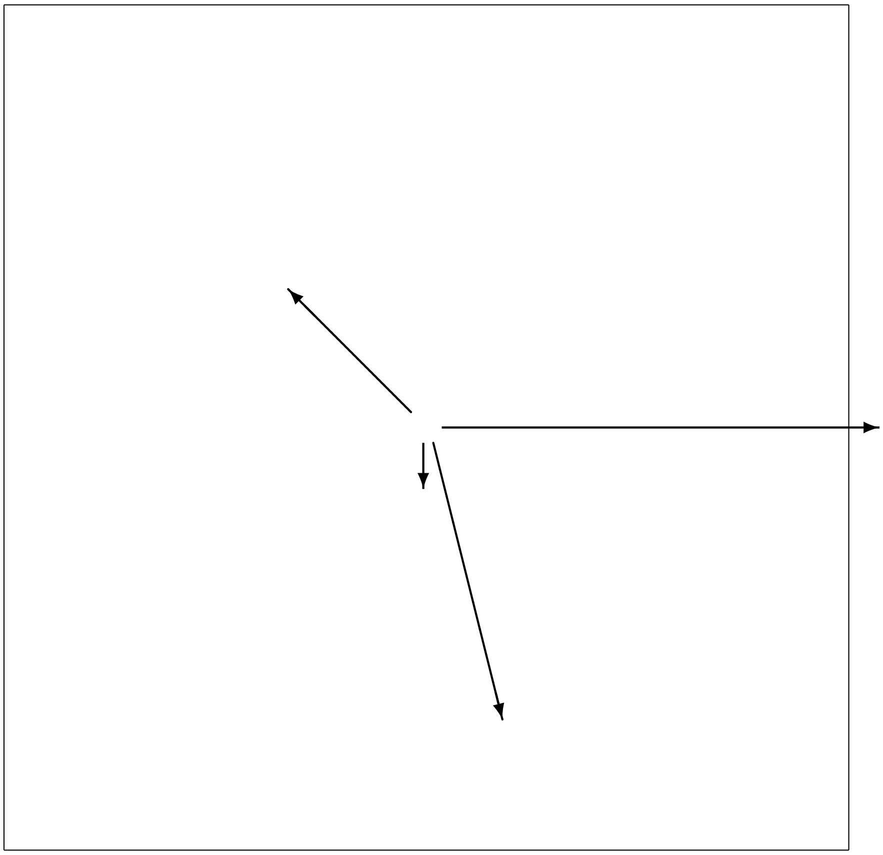


a
b
c


a
b
c







Bottom line: we get a non-explicit description of the $n \to \infty$ limit complex network model as "this process at the current time". Tractable because

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

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

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

———————————————————————————

The model's underlying geometry – random points in infinite dimensional space – may seem arbitrary but in fact is less arbitrary than the usual "extreme" alternates

- vertices are points in d-dimensional space
- no geometry, which is tantamount to assuming all vertex-pairs are at equal intrinsic distance.

Yule process of rate μ

Each individual at time t has chance $\mu\ dt$ to have daughter during $[t,\ t+dt]$.

$N(t) =$ number individuals at time t ($N(0) = 1$).

Yule (1924) used as model for species within a genus and proved two results.

$$N(t) \ \overset{d}{=}\ \text{Geo}(e^{-\mu t}).$$

Yule extended model by assuming that from within a genus, a new species founding a new genus arises at constant stochastic rate θ. In long run, age of typical genus has law T/θ for $T\ \overset{d}{=}\ \text{Exp}(1)$ and so size N of typical genus has

$$N\ \overset{d}{=}\ \text{Geo}(e^{-\mu T/\theta})$$

which has a power law tail.

Yule processes appear within our model in two different ways. First consider the PWIT, the “geometry” of random points in infinite-dimensional space. Consider

$$N(r) = \text{number of points within distance } r \text{ of } \emptyset$$

and include $\emptyset$ itself, then $(N(r),\ 0 \le r)$ is the Yule process of rate 1. So

$$EN(r) = e^r.$$

Next consider our space-time random graph process. For a typical vertex w consider

$$N(t) = 1 + \text{ in-degree(w) at time } t$$

starting time when w arrives. Using time-invariance of space-time process of arriving points, and the underlying stochastic dynamics

> 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 $p(d(v, w))$

easy to see that $N(t)$ is Yule process of rate β for

$$\beta = \int_0^\infty p(x)\ dx.$$