Spatial Networks

David Aldous

August 30, 2013

Course web site: Google "David Aldous". [show]

I deliberately chose a topic where there is not any clean definition-theorem-proof account – but such accounts exist for various related topics.

A **metaphor** for research styles: external DLA and internal DLA. [show 2 figs].

So start by looking outside the math literature. What does Google Scholar find?

Journal *Networks and spatial economics* [show]

This is O.R. style – not quite what we will focus on. Note student requirements: give a talk, either on a research paper you read (easy option), or on some project you do (harder option). [show lists]

"Big Picture" on slides. When we get to non-trivial math, I'll do chalk.

Consider points in 2-dimensional space. Mathematicians understand regular patterns. [show Lattice-wikipedia].

Mathematicians also understand completely random points, as modeled by the Poisson point process.
[show poisson.jpg] [show Complete-wikipedia]. [show list projects]

In mathematics, **graph** has a fairly precise meaning – vertices (nodes) and edges (links) – I will call this an “abstract graph”. [draw on board]. But **network** does not have a standard meaning – I say “a network is a graph with context-dependent extra structure”.

In our topic, **spatial networks**, vertices have positions in the plane $\mathbb{R}^2$, and the edges (usually made up of line segments) are point-sets in $\mathbb{R}^2$. [introduce coordinates].

Familiar “regular pattern networks” [show 9networks] and the kind of questions we will study are (usually) elementary for regular patterns.

What can we do with arbitrary configurations of points in the plane? Two classical structures are the **Voronoi tessellation** and the **Delaunay triangulation**. [show Voronoi and Delaunay]

We can think of the Delaunay triangulation as the prototype example of a spatial network.

The Delaunay triangulation (DT) is at the intersection of several academic fields. In the background are

- "graph theory" in Math
- much of "theory of algorithms" in C.S. involves algorithms on abstract graphs

and in the foreground are

- Euclidean geometry (the DT itself)
- **computational geometry** studies algorithms for problems such as finding the DT
- **stochastic geometry** ($\approx$ **geometric probability**) studies quantitative aspects (expectations etc) of structures such as the DT built over random points.

[show book list]

One starting point for this course is to think about random analogs of regular networks. There are several ways to make probability models. I will talk first about

- schemes for defining a network over arbitrary points (e.g. the DT), applied to random (Poisson) points in the plane

but other ways start with a regular network (e.g. square grid) and then

- randomly delete edges (percolation)
- randomly add long edges (small worlds)
- assign random lengths to edges (first passage percolation).

Side comments. There is extensive work on
(i) random processes built over regular networks – percolation, interacting particle systems – see Grimmett *Probability on Graphs* for a start.
(ii) random models for abstract graphs, generalizing the Erdős-Rényi model – see van der Hofstad *Random Graphs and Complex Networks*.

Something potentially confusing is the notion of a **planar graph**. This is defined as an abstract graph which can be drawn in the plane with non-crossing edges. This verbal definition looks ambiguous – are edges required to be straight lines? – but it's a fact that the two definitions are equivalent. [draw picture on board]

This differs from a spatial network in several ways. A spatial network is drawn in the plane – vertices and edges are at specified coordinates. So there are several possibilities for a spatial network:

1. edges are line segments and do not cross (e,g, the DT)
2. edges are line segments; may cross but you cannot switch (e.g. airplane routes)
3. edges can meet at positions (Steiner points or junctions) that are not the given vertices.

Most literature concerns networks satisfying (1) or (2), but we (envisaging inter-city road networks) will often allow (3).
[show US-road-map.pdf]

Side comment. There is literature on "random planar graphs": see Benjamini (2010) for a brief overview. There the model is the uniform distribution on the finite set of *n*-vertex planar graphs, but this model does not seem realistic for any data.

Another fact about connected planar graphs is **Euler's formula**

$$N - E + F = 2$$

$N =$ number of vertices
$E =$ number of edges
$F =$ number of faces ($=$ cells $=$ components of plane).
[true for tree; induct on extra edges]

Taking $E \geq 3$, a face has at least 3 edges and each edge separates at most 2 faces, so $E \geq (3/2)F$; substitute into Euler's formula:

$$E \leq 3N - 6.$$

So the average degree $\bar{d}$ in a connected planar graph

$$\bar{d} = 2E/N \leq 6 - \tfrac{12}{N} < 6.$$

[show 9networks again]

Finite vs Infinite networks

- Real networks are finite, and algorithmic questions typically ask how many steps are required to solve some problem over n points in the worst case.
- Mathematicians like me often think in terms of infinite numbers of points in the infinite plane.

The latter is perhaps an example of mathematicians putting a lot of effort into being lazy; in particular one can ignore boundary effects. For instance

- On the infinite square grid the average degree is 4 and the average length per unit area is 2.
- The Voronoi tessellation on a (reasonable) infinite set of points has only finite-area cells.
- In any triangulation of a (reasonable) infinite set of points, the average degree is 6. [informal argument on board].

The infinite setting is not so relevant for algorithmic or worst-case studies, but is useful for average-case studies. The natural "average-case" model for n points is to take them IID uniform in a square; to avoid rescaling later we take the **scaling convention** that the square has **area** $= n$. The corresponding infinite model is the rate-1 Poisson point process (PPP) on $\mathbb{R}^2$.

Most explicit constructions of networks on n points can be used as constructions of networks on the infinite PPP. Precise quantitative connections between the finite and infinite networks are one of the technical topics we'll discuss later in the course, involving

- subadditivity
- local weak convergence
- stationarity as the formalization of "reasonable" for infinite networks.

Comments on triangulations (define: = point set triangulation).

- not unique [draw n-gon]
- given a planar graph spatial network which is not a triangulation, we can add edges until it becomes a triangulation
- in a triangulation, [board]

  $$3(F-1) = 2E - C$$

  where C = number of edges (or vertices) in convex hull. Euler's formula then gives

  $$E = 3N - 3 - C$$

  and so the average degree is

  $$\bar{d} = \tfrac{2E}{N} = 6 - \tfrac{6}{N} - \tfrac{2C}{N}$$

  which does not depend on the triangulation
- The DT is in general not the minimum-length triangulation; computing the latter is NP-hard.

A planar graph has a **(planar) dual** graph
[show Dual] [show 9networks: multigraph]
This duality is the relationship between Voronoi tesselation (considered as a 3-regular graph) and Delaunay triangulation (DT).

Mention 4-color theorem. Duality relates this to vertex-coloring of the dual graph.

2 things to note.

- What is the DT for the square-lattice configuration?
- An edge AB of the DT does not necessarily pass through the edge separating the cells of A and B; the subgraph of edge that **do** defines the **Gabriel graph**.

[draw fig 2] [show Gabriel graph = Wiki]

- AB is edge of Gabriel graph iff disc with diameter AB contains no other points
- AB is edge of DT iff there exists some disc, with A and B on boundary, containing no other points.

Average degree of the Gabriel graph depends on configuration; for instance it is 6 for the triangular configuation. We shall soon do a calculation to show that for Poisson points,

$$\text{ave degree } := \bar{d} = 4; \quad \text{length-per-unit-area} := \ell = 2.$$

These of course are the values for the usual grid network. Is there any more conceptual explanation of this coincidence?

Some useful calculus formulas

$$\int_0^\infty r \cdot \exp(-Ar^2)\ dr = \frac{1}{2A}$$

$$\int_0^\infty \exp(-Ar^2)\ dr = \frac{\pi^{1/2}}{2A^{1/2}}$$

Higher moment formulas most easily found by differentiating w.r.t. A

$$\int_0^\infty r^2 \cdot \exp(-Ar^2)\ dr = \frac{\pi^{1/2}}{4A^{3/2}}$$

Some calculations with PPP [on board]

- Definitions of spatial PPP and finite-n model.
- Link via "random position" and "random point".
- PPP as seen from "typical point" is just the PPP with an extra point planted at origin.
- Use in calculations that are exact for PPP model, and are $n \to \infty$ limits for finite-n model.

Basic example. In PPP,

$$R = \text{distance from typical point } \xi \text{ to nearest neighbor } \xi'$$

has $\mathbb{P}(R > r) = \exp(-\pi r^2)$ and so $\mathbb{E}R = 1/2$.

In finite-n model, consider
R_n = distance from random position to nearest point
R_n^* = distance from random point to nearest other point.
Then $R_n \overset{d}{\to} R, \quad R_n^* \overset{d}{\to} R.$

More elaborate example.

In PPP, take ξ as typical point, ξ' as nearest neighbor of ξ, and ξ'' as nearest neighbor of ξ'. Then [board]

$$q := \mathbb{P}(\xi'' = \xi) = \frac{\pi}{\frac{4\pi}{3} + \sqrt{\frac{3}{4}}}.$$

In finite-n model, let $M_n =$ number of points which are the nearest neighbor of their nearest neighbor. Then

$$n^{-1}\mathbb{E}M_n \to q.$$

To re-interpret as a "network", put an edge from each point of a PPP to its nearest neighbor, the mean number of edges per unit area equals $\frac{q}{2} + (1-q)$. And [exercise] can also calculate mean length-per-unit-area of this network.

Almost all the networks we will consider, build over the PPP, have distributions invariant under translation and rotation [as will be discussed more carefully later]. For such a network, given the PPP has points at z_1 and z_2, the chance $p(r)$ that there is an edge (z_1, z_2) depends only on the distance r between z_1 and z_2. Then we can calculate

$$\text{ave degree } := \bar{d} = \int_0^\infty p(r) \cdot 2\pi r \ dr$$

$$\text{length-per-unit-area} := \ell = \tfrac{1}{2} \int_0^\infty r \cdot p(r) \cdot 2\pi r \ dr.$$

For the Gabriel graph,

$$p(r) = \exp(-Ar^2)$$

where $A =$ area of unit-diameter disc. So

$$\bar{d} = \int_0^\infty \exp(-Ar^2) \ \cdot \ 2\pi r \ dr = \pi/A$$

$$\ell = \tfrac{1}{2} \int_0^\infty \exp(-Ar^2) \ \cdot r \cdot \ 2\pi r \ dr = \tfrac{\pi^{3/2}}{4A^{3/2}}$$

Because $A = \pi/4$ we get $\bar{d} = 4$; $\ell = 2$.

As noted before, these values for the Gabriel graph on the PPP equal the values for the usual grid network. Is there any more conceptual explanation of this coincidence?

Similar arguments [board] give more complicated integral expressions for

1. number of triangles per unit area
2. proportion of $\mathbb{R}^2$ covered by triangles.

Do these have simple explicit values?

For other literature on the random Gabriel network see Bose et al (2006), Devroye - Gudmundsson (2009):

Given n points in general position, the literature contains many ways to define edges to make a spatial network – on the web site I will maintain a list of those mentioned in the course.
It is a creative exercise to try to invent other (interesting) ways.

Next topic is **minimum spanning tree**. The definition and the following properties hold for an abstract connected graph G with distinct edge-lengths; in our "spatial" context we will be taking the case of n points in the plane, with the complete graph and Euclidean distance.

Most useful is the following characterization.
An edge AB of G is an edge of the MST if there does not exist $m \geq 2$ and a path $A = A_0, A_1, A_2, \ldots, A_m = B$ in G with each edge having length $<$ the length of AB.

Taking this as a definition of **MST**, it is easy to show [board]

1. the MST has no cycles
2. the MST contains the shortest edge at each point
3. the MST has only one component
4. the MST is the spanning connected subgraph with minimum total length.

By considering 2-edge paths, the MST is a subgraph of the **relative neighborhood graph** (RNG) defined by

AB is an edge of the RNG if there does not exist any point C such that max$(\mathrm{length}(CB), \mathrm{length}(CA)) < \mathrm{length}(AB)$.

Easy to see RNG is a subgraph of the Gabriel graph, so we have an ordering

$$MST \subseteq RNG \subseteq \mathrm{Gabriel} \subseteq \mathrm{DT}$$

which implies they are all connected. Note that for MST (or any tree) we have average degree $\to 2$.

Project: Draw figures of all the networks we will discuss, on a realization of random points.

(Exercise: what definitions/results remain true for abstract graph with edge-lengths?)
(Exercise: for an infinitesimal perturbation of the square grid configuration, what networks do these procedures give?)

If we allow junctions, then the shortest connected network on a finite set of points in $\mathbb{R}^2$ is called the **Steiner tree**. [draw square on board]. Note:

1. A junction must have exactly 3 edges at 120^o angles.
2. The MST can be found by a simple greedy algorithm; but finding the Steiner tree is a famous example of a NP problem.
3. You can become famous by proving or disproving the conjecture that the worst-case ratio $\mathrm{length(MST)/length(Steiner\ tree)}$ equals $2/\sqrt{3}$.

This course does not focus on MSTs or Steiner trees, but rather on networks that are "efficient" for some purpose. One expects a trade-off between some measure of efficiency and some measure of cost, and the simplest measure of cost is just the length of the network. In such an analysis, the MST or Steiner tree appears as one end of the range of possible costs.
Project. It is curious there are (apparently) no standard ways to define networks with junctions, analogous to Gabriel etc for networks without junctions. Can you invent one? In particular, one for which calculations over a PPP are possible?

As example of an “extreme” configuration, take $n-1$ points evenly spaced around circle, n’th point in center; then perturb slightly.

[work on board]

In our scaling, length of MST is order $n^{1/2}$, length of others is order $n^{3/2}$.

Re previous project, would like scheme which does have order-n length in worst case.

The natural "average-case" model for n points is to take them IID uniform in a square; recall our **scaling convention** is take the square to have **area** $= n$, so this model can be compared with the corresponding infinite model, which is the rate-1 Poisson point process (PPP) on $\mathbb{R}^2$. In both models the typical distance from a vertex to its nearest neighbor is order 1.

Useful **general principle**: in many settings, for "optimal" network over n points, worst-case has same order of magnitude as average-case.

Illustrate with MST. The length of the MST grows as order n in both cases.

[work on board]

Exercise: Show that the length of the Steiner tree on random points grows as (not slower than) order n.

As $n \to \infty$ we expect the MST on the finite-n model to look like the MST on the PPP. For our purposes, define the latter by the criterion used before:

AB is an edge of **MST** if there does not exist $2 \leq m < \infty$ and a path $A = A_0, A_1, A_2, \ldots, A_m = B$ with each edge having length $<$ the length of AB.

The previous arguments show this **MST** is a forest, whose components are infinite trees. In fact this **MST** is a single tree, but proving this is much harder – see remark at end of section 2 of Alexander (1995). For the $\mathbb{Z}^2$-with-random-edges-lengths analog see Lyons-Peres (2013) section 11.5. However, the forest-or-tree issue does not affect the calculation

$$\text{ave degree } := \bar{d} = \int_0^\infty p(r) \cdot 2\pi r \; dr$$

$$\text{length-per-unit-area} := \ell = \tfrac{1}{2} \int_0^\infty r \cdot p(r) \cdot 2\pi r \; dr$$

where $p(r)$ is the probability of an edge between two planted points at distance r. Being a tree we know $\bar{d} = 2$. But there is no explicit formula for $p(r)$, because one would need to consider paths of each length $m \geq 2$.

To get a “big picture”, in the next few lectures we will look at the 100 page survey “Spatial Networks” by Marc Barthélemy. [show page 1] This is written in statistical physics style and shows some data. What I’ve selected reflects my own taste – useful to you to browse the survey yourself.

By coincidence when preparing this lecture last month I was browsing the BBC web site and found the following [show streets_of_paris].
Their paper is Barthelemy et al (2013).

Extensive statistical physics literature on (non-spatial) networks since 2000 rather obsessed with power law degree distribution:

$$d(i) := \text{proportion vertices with degree } i \approx i^{-\gamma} \text{ for large } i$$

This almost never holds for real-world **spatial** networks. But (analogy: classical statistics and Normal distributions and regression) one can use a trick (first non-textbook math from Barthélemy). In a real-world road network (and some other physical networks) the roads are assigned names or numbers. [show map.pdf]

So by formalizing "a specific road" as "a specific path of edges in a spatial network", with each edge in exactly one road, we can define a "road-dual" graph in which

- a "vertex" is a road in the original network
- there is an "edge" between two vertices iff the two roads intersect in the original network.

The point of the trick is that degree distribution $\hat{d}(i)$ in the road-dual network becomes

$$\hat{d}(i) = \text{proportion of roads with } i \text{ intersections.}$$

From real-world experience there are a few long major roads and many more short minor roads. So some power-law-like distribution is at least possible in this dual. For analysis in e.g. “unplanned” European cities see Porta - Crucitti - Latora (2006) and citations thereto.

Making some math model seems difficult – on the course **projects list**.

Barthélemy survey gives some discussion of "abstract graph" statistics based on hop distance. In particular, "betweenness centrality" is the function $q(e)$ on edges defined by:

for each pair (v, w) let $q(e; v, w) =$ probability that a uniform random min-hop-length path from v to w contains e.

then $q(e) = \sum_{v,w} q(e; v, w)$ and analogously for vertices.

This has been used often in study of "streets of Paris" type examples. But using hop-length seems rather unnatural as measure of traffic flow on urban road networks. (ignores lengths/capacity of streets). In one sense more natural to use real-valued edge-lengths (= times), implying **unique** shortest paths. But . . . . . . does using shortest paths lead to congestion?

Hop-length is more relevant when there are transfer costs – passenger air travel, FedEx package delivery. We will (soon) study a math model for this setting.

Barthélemy survey makes the following observation.

In a spatial network, for vertex v write

$d(v) =$ degree of v,
$D(v) =$ sum-of-edge-lengths at v.

If the graph structure ignored the spatial structure, (e.g. assign IID edge-lengths to abstract graph) then the function $\mathbb{E}(D(v)|d(v) = d)$ is linear. In a real-world spatial network (no junctions) we expect edges to nearby points, implying faster-than-linear (maybe $d^{3/2}$).

Is there a math question here?

Another aspect of statistical analysis of (abstract) graph data is called "motifs", which simply means counting the number of occurences of fixed small graphs as subgraphs of a given large graph. (This is related to "local weak convergence" of sparse graphs, discussed much later). For simple random models of abstract graphs (Erdős-Rényi etc) one can calculate the expectation (and more) of such numbers.

Re spatial network models, perhaps the simplest "motif" one could consider would be the number of triangles per unit area in the random Gabriel network – we asked this question earlier.

When we allow junctions (e.g. large-scale road networks) the number of possible motifs on 4 points becomes surprisingly large. We have some data on routes between 4 addresses at corners of a square

[show maps100-567]

Project: to get good data; and a systematic classification of the different topologies in the "leaf-labeled" planar model of Aldous (2014).

A recurring theme within the “spatial networks” field is to relate Euclidean distance between two points to either hop distance (number of edges) or route-length; one can then define statistics to measure “efficiency” of a given network, and then one can consider optimal networks.

Graphic from Gastner - Newman (2006) shows a certain notion of “optimal network” for different parameters δ representing relative weight of hop-distance and route-length; these look quite realistic. In particular,

simulated $\delta = 0$ network $\approx$ real-world road network,
simulated $\delta = 0.5$ network $\approx$ airline hub-spoke network.
simulated $\delta = 1$ $\approx$ UPS (Louisville) and FedEx (Memphis) networks.

[show newman-figure.pdf] [show NHS-pdf] [show hub.svg]

As the first non-textbook“honest math” in this course, I will outline results from Aldous (2008a) involving (as a conclusion) hub-spoke models. Not deep or fundamental, but an **example** of looking at non-math literature and inventing some math.

Let's think about designing a network where routes involve 3 hops (2 transfers). Take n arbitrary points in area-n square.

- Divide area-n square into subsquares of side L.
- Choose a **hub** in each subsquare.
- Link each pair of hubs.
- Link each city to the hub in its subsquare (a **spoke**).

Cute freshman calculus exercise: what total network length do we get by optimizing over L?

[length of short edges]: order nL
[length of long edges]: order $(n/L^2)^2 n^{1/2}$.

Sum is minimized by $L =$ order $n^{3/10}$ and total length is order $n^{13/10}$. Note that the total length of all short edges, and of all long edges, have the same order.

Math project; how does this network compare to some "theoretically optimum" network?
Seek to model the situation where the time to travel a route depends on route length and number of hops/transfers, each term contributing sae order of magnitude. Introduce a weighting parameter Δ and define (for a network $\mathcal{G}_n$ linking n cities $\mathbf{x}_n$ in square of area n)

$$\text{time to traverse a given route from } x_i \text{ to } x_j$$
$$= n^{-1/2} \times (\text{ route length}) + \Delta \times (\text{ number of transfers }).$$
$$\textbf{time}(x_i, x_j) = \text{ min. time, over all routes}$$

$$\begin{aligned} \textbf{ave\_time}(\mathcal{G}_n) &= \mathrm{ave}_{i,j}\textbf{time}(x_i, x_j) \\ &\geq n^{-1/2}\mathrm{ave}_{i,j}\|x_i - x_j\| := \textbf{ave\_dist}(\mathbf{x}_n). \end{aligned}$$

Our construction gave a network such that (even for worst-case configuration $\mathbf{x}_n$)

$$(*) \qquad \textbf{ave\_time}(\mathcal{G}_n) - \textbf{ave\_dist}(\mathbf{x}_n) \to 2\Delta$$

$$\mathrm{length}(\mathcal{G}_n) = O(n^{13/10}).$$

Interpret (*) as saying that average number of transfers $\to 2$.

Our construction gave a network $\mathcal{G}_n$ such that (even for worst-case configuration $\mathbf{x}_n$)

$$(*) \qquad \textbf{ave\_time}(\mathcal{G}_n) - \textbf{ave\_dist}(\mathbf{x}_n) \to 2\Delta$$

$$\mathrm{length}(\mathcal{G}_n) = O(n^{13/10}).$$

Interpret (*) as saying that average number of transfers → 2.
One can do analogous constructions for other (integer) values of "number of transfers" which give different power exponents.

I will outline proof of

Theorem (Aldous (2008a))

In the random model, for any network $\mathcal{G}_n$ satisfying (*), its length grows at least as fast as order $n^{13/10}$.

Central idea: pick two of the points at random – call them x_I and x_J – and study the quantity

$$p_n(a,b) = \text{ probability route } x_I\text{-to-}x_J \text{ has exactly three edges,}$$

$$\text{the middle edge-length} > b \text{ and each end-length} < a.$$

We first do calculations based on assuming x_I and x_J are independent uniform, then go back and address this approximation.

$$p_n(a,b) \le C \frac{a^4}{n^2 b} \operatorname{length}(\mathcal{G}_n).$$

$$1 - p_n(a,b) \le \frac{\pi(2a+b)^2}{n} + \frac{4 \operatorname{length}(\mathcal{G}_n)}{an} + \mathbb{P}(H_n \ge 4)$$

where H_n = number of hops on route.
[outline on board]

Here I discuss a technical point: in fact the Theorem holds for non-random configurations $\mathbf{x}_n$ satisfying a certain "quantitative equidistribution" property.

Take integers L_n and partition $[0, n^{1/2}]^2$ into L_n^2 subsquares σ of side-length $s_n = n^{1/2}/L_n$. Define the smoothed empirical distribution ψ_n

$$\psi_n = \sum_{\sigma} \sum_{i} \tfrac{1}{n} 1_{(x_i \in \sigma)} \mu_\sigma$$

where μ_σ is the uniform distribution on a subsquare σ.

Theorem

For any network $\mathcal{G}_n$ satisfying (*) over a configuration $\mathbf{x}_n$, its length grows at least as fast as order $n^{13/10}$, provided the configurations satisfy

$$||\psi_n - \bar{\mu}_n||_{VD} \to 0 \textit{ for } s_n \sim n^{3/10}$$

$\bar{\mu}_n$ is the uniform distribution on $[0, n^{1/2}]^2$.

Remarks on this setup.

1. The classical "equidistribution" property is that (after scaling to $[0,1]^2$) the empirical distribution of $\mathbf{x}_n$ converges weakly to the uniform distribution. This is equivalent (exercise) to saying that, for $L_n \to \infty$ sufficiently slowly,

$$||\psi_n - \bar{\mu}_n||_{VD} \to 0$$

where convergence is in **variation distance** and $\bar{\mu}_n$ is the uniform distribution on $[0, n^{1/2}]^2$.

2. In the random model, for any $L_n = o(n^{1/2})$ we have

$$\mathbb{P}\left(||\psi_n - \bar{\mu}_n||_{VD} > \varepsilon\right) \to 0, \quad \forall \varepsilon > 0.$$

3. For general $\mathbf{x}_n$, the property

$$||\psi_n - \bar{\mu}_n||_{VD} \to 0$$

becomes a stronger property for larger L_n.

[outline proof on board]
Hypothesis implies

$$\mathbb{P}(||x_I - U_n|| > s_n\sqrt{2}) \to 0$$

for U_n uniform on $[0, n^{1/2}]^2$.

Repeat earlier arguments with this correction: the major change is to the term

$$p_n(a,b) \leq C\frac{(a+s_n)^4}{n^2 b}\ \mathrm{length}(\mathcal{G}_n).$$

Because we took $a_n \sim n^{3/10}$ we need also $s_n \sim n^{3/10}$ or smaller.

We will often consider route-length $R(x_i, x_j)$ between two points in a given network. It's natural to compare this with Euclidean distance $||x_j - x_i||$.

Digression in weird direction to get to a research problem. Intuitively, using a **tree** as a network is not good for having short routes. Is this correct?

An easy construction [outline on board] shows

Proposition

Consider either
(a) the $n = m \times m$ grid of points; or
(b) n points in square of area n.
Then there exists, in case (a) a spanning tree in $\mathbb{Z}_m^2$, in case (b) a spanning tree-network with junctions, such that

$$\text{ave } R(x_i, x_j) = O(n^{1/2}).$$

This is the optimal order of magnitude, so in this sense trees are "not bad". Is there another sense in which trees are indeed bad?

The following project would make a little research paper. I will outline why I think it is true.

Conjecture (Route-length in tree-networks)

Consider either
(a) the $n = m \times m$ lattice of points; or
(b) n points in square of area n.
Consider in case (a) a spanning tree in $\mathbb{Z}^2_m$, in case (b) an arbitrary spanning tree-network with junctions. In either case let L_n be the minimum over trees of the quantity

$$\max_j \quad \frac{\mathrm{ave}\{R(x_i,x_j)\ :\ 2^j \le ||x_j - x_i|| < 2^{j+1}\}}{2^j}.$$

Then (under some equidistribution assumption in case (b)) L_n grows as order $\log n$.

Note that in case (a) every tree has the same length, whereas in case (b) the length could be chosen arbitrarily large.
[outline on board: upper bound from previous construction, lower bound from isoperimetry and tree-centroid].

The previous Proposition suggests that statistics based on averaging route-lengths over all pairs are not very helpful. Here is a more dramatic illustration.

The construction is from Aldous-Kendall (2008); see also Dujmovic - Morin - Smid (2013).

[show AK-lines]

Construction: arbitrary configuration $\mathbf{x}_n$ of n points in square of area n. Take the Steiner tree; this has length $ST(\mathbf{x}_n)$, say. Superimpose a sparse Poisson line process, length-per-unit-area $= w_n \downarrow 0$ slowly.

This gives a network for which

$$\text{length } = ST(\mathbf{x}_n) + o(n); \qquad \mathrm{ave}_{i,j} \frac{R(x_i, x_j)}{||x_i - x_j||} \to 1.$$

So the network is "optimal" in both respects, but not sensible as a road network! By analogy with the previous Conjecture, instead of the $\mathrm{ave}_{i,j}$ above, a more sensible statistic to measure "route-length efficiency" of a network might be

$$(*) \quad \max_j \frac{\mathrm{ave}\{R(x_i, x_j) \ : \ 2^j \leq ||x_j - x_i|| < 2^{j+1}\}}{2^j}.$$

Later, I will discuss some heuristics/simulations for optimal networks under this criterion.

Summary. There are 3 reasons why you might want a statistic to measure “route-length efficiency” of a network:

- as a descriptive statistic of a real-world network
- as a statistic of a given mathematical model of a network
- as a criterion for designing optimal networks.

We now have 3 statistics “r” one might use to measure “route-length efficiency”.

$$r_{ave} := \mathrm{ave}_{i,j} \ \frac{R(x_i, x_j)}{||x_i - x_j||}$$

What has been extensively discussed in algorithms literature is the statistic **stretch** defined by

$$r^* := \max_{i,j} \ \frac{R(x_i, x_j)}{||x_i - x_j||}.$$

A statistic like (*) above is intermediate between these; it is awkward to formalize in worst-case, but in our average-case model we can define it as

$$r_{\sqcup} := \sup_r \ \mathbb{E}\left(\tfrac{R(x_i, x_j)}{r} \ | \ ||x_i - x_j|| = r\right).$$

Conceptual comments:

- r_{ave} not useful as design criterion (examples above).
- r^* not satisfactory as real-world descriptive statistic (e.g. comparing railway networks in two countries).

Technical comments:

- In $n \to \infty$ limit, if the worst-case value of r^* is finite in the worst-case model, then it has the same value in the average-case model, for any network in which edge are determined by some local rule. This is because the worst-case configuration for given n will appear somewhere in the random model for $N \gg n$. However, when $r_n^* \to \infty$ then the order of magnitude will typically be different in the two models.

We will try to fill in a table showing $n \to \infty$ limits of these statistics, and

$$\ell := \text{ length-per-unit-area}$$

for some "mathematically natural" networks.

This may appear to be a boring exercise, but it illustrates different techniques and open problems.

Stretch in the worst case. For a given network over arbitrary points consider

$$r_n^* := \max_{\{x_1,\ldots,x_n\}} \max_{i,j} \frac{R(x_i,x_j)}{||x_i - x_j||}.$$

Recall the ordering

$$MST \subseteq RNG \subseteq \mathrm{Gabriel} \subseteq \mathrm{DT}$$

so r_n^* can only decrease through this sequence.

For MST, by considering n-cycle we have $r_n^* = n - 1$.

For RNG, a simple construction [on board] from Bose et al (2006) shows $r_n^* = \Omega(n)$ and so

$$r_n^* = \Theta(n)$$

For Gabriel, another simple construction [on board] from Bose et al (2006) shows $r_n^* \geq (\frac{1}{2} - o(1))n^{1/2}$. They also show the corresponding upper bound, so

$$r_n^* = \Theta(n^{1/2})$$

For Delaunay triangulation, it is known that r^*_∞ is finite, with bounds

$$1.58 < r^*_\infty < 2.42.$$

Upper bound given in Keil - Gutwin (1999) and lower bound in Bose et al (2009).

Now consider **stretch** in the **random** model. As noted earlier, for the Delaunay triangulation, where r^*_∞ is finite, we must get the same value as in the worst-case. But for the other networks we expect different orders of magnitude for r^*_n. Bose et al (2006) show that for the random Gabriel network

$$r^*_n = \Omega\left(\sqrt{\frac{\log n}{\log \log n}}\right).$$

[brief outline on board]
For the MST in the random model, the "subtrees at centroid" argument shows that r^*_n must grow at least as fast as $n^{1/2}$, and this holds for any tree. In fact we expect (xxx later; MST exponents – project literature).

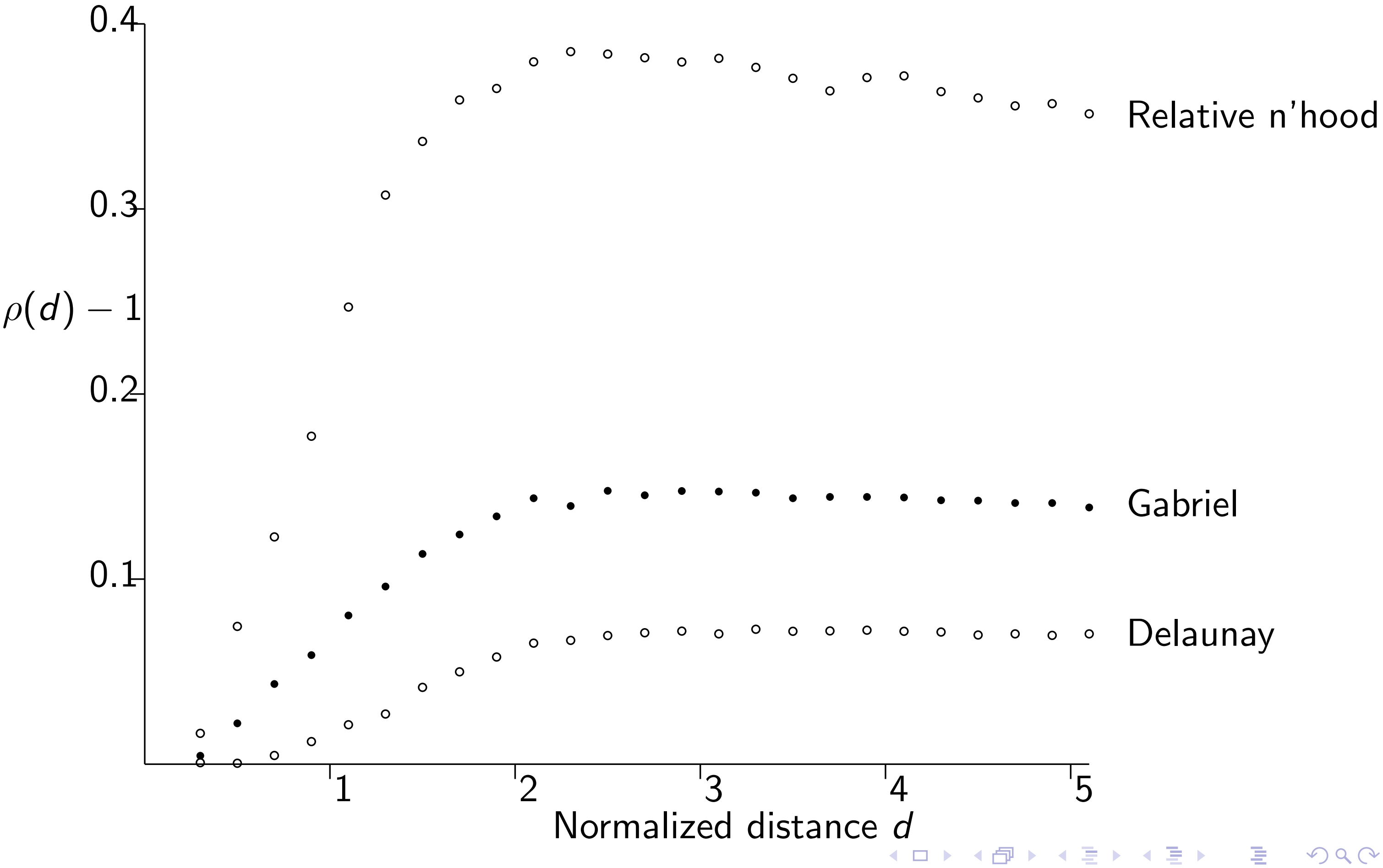
$$r_{ave} := \mathrm{ave}_{i,j} \ \frac{R(x_i, x_j)}{||x_i - x_j||}$$

$$(\textbf{stretch}) \ r^* := \max_{i,j} \ \frac{R(x_i, x_j)}{||x_i - x_j||}.$$

Intermediate is

$$\begin{aligned} r_{\sqcup} \quad &:= \quad \sup_d \ \mathbb{E}(\tfrac{R(x_i,x_j)}{r} \mid ||x_i - x_j|| = d). \\ &= \quad \sup_d \rho(d), \quad \text{say.} \end{aligned}$$

Alas this is too difficult to study analytically. Here are plots of the function $\rho(d)$ in the random model, from Aldous - Shun (2010) simulations.


0.4
0.3
0.2
0.1
$\rho(d) - 1$
1
2
3
4
5
Normalized distance d
Relative n'hood
Gabriel
Delaunay

Note the shape – maximum around $d = 2$.
[board]
Note also r_{ave} is asymptotic value.

The beta-skeleton family of networks

[show old proximity_slides.pdf]

(i) For $0 < \beta < 1$ let A_β be the intersection of the two open discs of radius $1/(2\beta)$ passing through v_- and v_+.
(i) For $1 \leq \beta \leq 2$ let A_β be the intersection of the two open discs of radius $\beta/2$ centered at $(\pm(\beta - 1)/2, 0)$.

For the random model, we can calculate length-per-unit-area, for any such proximity network, as we did for the Gabriel network.

[next slide repeats earlier]

Almost all the networks we will consider, build over the PPP, have distributions invariant under translation and rotation [as will be discussed more carefully later]. For such a network, given the PPP has points at z_1 and z_2, the chance $p(r)$ that there is an edge (z_1, z_2) depends only on the distance r between z_1 and z_2. Then we can calculate

$$\text{ave degree } := \bar{d} = \int_0^\infty p(r) \cdot 2\pi r \ dr$$

$$\text{length-per-unit-area} := \ell = \tfrac{1}{2} \int_0^\infty r \cdot p(r) \cdot 2\pi r \ dr.$$

For the Gabriel graph,

$$p(r) = \exp(-Ar^2)$$

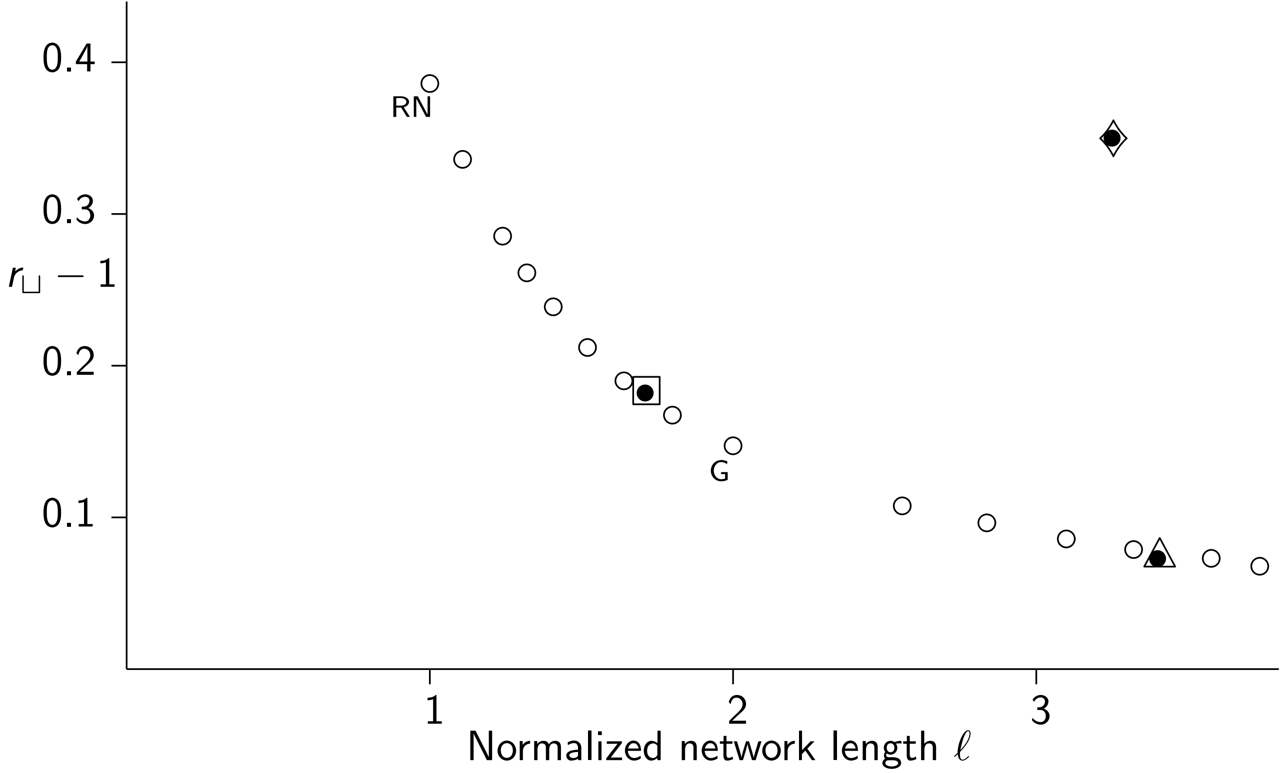
where $A =$ area of unit-diameter disc. So

$$\bar{d} = \int_0^\infty \exp(-Ar^2) \ \cdot \ 2\pi r \ dr = \pi / A$$

$$\ell = \tfrac{1}{2} \int_0^\infty \exp(-Ar^2) \ \cdot r \cdot \ 2\pi r \ dr = \tfrac{\pi^{3/2}}{4A^{3/2}}.$$

So we can calculate the values of ℓ for the beta-skeleton family, just from the area of the excluded region.

Here are plots of ℓ versus $r_{\sqcup}$ for the beta-skeleton family (random model); Aldous - Shun (2010) simulations.



Networks based on powers of edge-lengths

. Here is one scheme used in, for example, Narayanaswamy et al (2002).

Fix $1 \leq p < \infty$. Given a configuration $\mathbf{x}$ and a route (sequence of vertices) $x_0, x_1, \ldots x_k$ say, define the cost of the route to be the sum $\sum_i ||x_i - x_{i-1}||^p$ of pth powers of the step lengths. Now say that a pair (x, y) is an edge of the network $\mathcal{G}_p$ if the cheapest route from x to y is the one-step route. Easy to see

[board]

- as p increases from 1 to ∞, the networks $\mathcal{G}_p$ decrease from the complete graph to the MST.
- $\mathcal{G}_2$, and so $\mathcal{G}_p$ for $p \leq 2$, is a subgraph of the Gabriel graph.

There are several projects here.

- Literature on this model?
- Simulations of ℓ versus $r_\sqcup$ for this family, as done for the beta-skeleton family ($\mathcal{G}_2$ was included in graph as $\Box$).
- Can you define other "interesting" one-parameter families of networks.

Miscellaneous comments

1. Regarding worst-case values of $\ell =$ length-per-unit-area.
For MST we know this is $O(1)$.
For RNG (and hence others) we know it is $\Omega(n^{3/2})$ by example of $n-1$ points on cycle, n'th in center. I guess this is worst-case (literature search needed).

2: Project: Here is a perhaps more interesting and less-studied question. For (say) Gabriel or DT, can one give sharp conditions for configurations $\mathbf{x}_n$ to have the property $\ell = O(1)$?

3. Folklore and simulations (literature search needed) say that for the MST on the infinite Poisson PP, the route-length betwwen point as distance d should grow as d^γ for some $\gamma > 1$. This would imply

$$r_{ave} \approx r_{\sqcup} \approx n^{(\gamma-1)/2}, \quad r^* \approx n^{\gamma/2}.$$

	ℓ	r_{ave}	$r_{\sqcup}$	r^*
MST-ave	0.633 MC	$\approx n^{(\gamma-1)/2}$?	$\approx n^{(\gamma-1)/2}$?	$\approx n^{\gamma/2}$?
MST-worst	c		–	$1 \cdot n$
RNG-ave	1.02	1.33 MC	1.38 MC	as implied by below
RNG-worst	$\Omega(n^{1/2})$		–	$\Theta(n)$
Gabriel-ave	2	1.12 MC	1.15 MC	$\Omega\left(\sqrt{\frac{\log n}{\log\log n}}\right)$
Gabriel-worst	$\Omega(n^{1/2})$		–	$\Theta(n^{1/2})$
DT-ave	3.40	1.05 MC	1.07 MC	as below
DT-worst	$\Omega(n^{1/2})$		–	$1.58 < c < 2.42$

xxx needs explanations

xxx indicates what has perhaps not been studied – worst-case for r_{ave}.

For MST, easy to see $r_{ave} = \Theta(n)$.

[board]

Project: is it true for RNG and Gabriel that worst-case r_{ave} is same order as worst-case r^*?

Back to the Barthélemy survey "Spatial Networks" .

Summary statistics of data (a real-world graph); definitions for abstract (non-spatial) graphs can of course be used for spatial networks.

$$n = \text{ number of vertices.}$$

Most basic is

$$d(i) := \text{proportion vertices with degree } i$$

Common "top-down" classification of graphs starts with qualitative properties of $(d(i), i \geq 1)$; in particular **scale-free** means

$$d(i) \approx i^{-\gamma} \text{ for large } i.$$

This is not so natural or common for spatial networks.

A fun elementary fact – should be the first theorem in a graph theory course!

Your friends have more friends than you do (on average).

[board]

Distinguish two cases.

Case (a). Pick uniform random edge; write as (V_1, V_2) ordering end-vertices randomly. Then $V_1 \overset{d}{=} V_2$ has the degree-biased distribution with $\mathbb{P}(V = v) \propto d(v)$, so

$$\mathbb{P}(V = v) = \tfrac{d(v)}{2m}$$

where $d(v) =$ degree of v and $m =$ number of edges.

Case (b). Pick uniform random vertex U; then pick uniform random neighbor F of U. (this is model: $U =$ you and $F =$ friend). Here

$$\mathbb{P}(U = u) = 1/n$$

but the distribution of F depends on more detailed structure – dependence between degrees across edges.

A natural idea is to consider "correlation between vertex-degrees of adjacent vertices" but the best way to do this isn't quite obvious. What appears in the literature, is an **assortivity coefficient**. This is (roughly) the covariance (or correlation ρ, depending on normalization) between $d(V_1)$ and $d(V_2)$, the end-vertices of a random edge.

Recall freshman statistics: the slope of the regression line for predicting $d(F)$ from $d(U)$ is

$$\text{slope} = \frac{\text{cov}(d(U), d(F))}{\text{var } d(U)}$$

and this has a more concrete interpretation.

[board]

How is this related to ρ?

For abstract graphs one can have “independence” of degrees across edges, e.g. the **configuration model**.

[board]

But there do not seem to be “natural” spatial models with this independence. Deijfen et al (2012) start with a Poisson PP, assign IID degrees to vertices (as in the configuration model) and then show it is *possible* to join the stubs to make a translation-invariant random network.

Simulations of ρ and “slope” for Gabriel (etc) in the random model?

Betweenness centrality.

For an edge e and a pair of vertices (v, w) let $q(e; v, w) =$ probability that a uniform random min-hop-length path from v to w contains e.

Then define $q(e) = \sum_{(v,w)} q(e; v, w)$. This function $q(e)$ (or a normalized form) gives "relative traffic flow" across different edges e, assuming uniform sources and destinations of traffic. Can interpret as "order of importance" for edges.

There is a natural analog $q'(v)$ for vertices.

data from air transportation network.
xxx show Figure 10

Previously mentioned one “alternative representation” for e.g. a road network:

by formalizing “a specific road” as “a specific path of edges in a spatial network”, with each edge in exactly one road, we can define a “road-dual” graph in which

- a “vertex” is a road in the original network
- there is an “edge” between two vertices iff the two roads intersect in the original network.

Here are more alternatives, for e.g. a railway network.

[show Figure 6]
Suppose there are different “routes” or “lines”, e.g. Richmond-Fremont BART route. Can define

P-space: edge (a,b) if some route includes a and b (can travel without changing trains).
L-space: edge (a,b) if these are successive stops on some route.

data from cargo ship network

[show Figure 17]

Power-law relation between degree and “betweenness centrality”, in P-space and L-space.

The spirit of this course is to look for research problems that fall **between** established math theories.

But of course we need to know something about what's in the established math theories . . . . . .

Relating finite-n random models to infinite models

Recall our finite-n random model for vertex positions is to take them IID uniform in a square of **area** $= n$, so this model can be compared with the corresponding infinite model, which is the rate-1 Poisson point process (PPP) on $\mathbb{R}^2$. In this part of the course we consider technical aspects of this relationship, starting with this overview lecture.

For models such as the Gabriel network which are defined by explicit local rules, the relationship is fairly simple; we can use the same rules to define a network over the PPP and then relate the finite and infinite models based on local weak convergence of the point processes. The theory we will develop does apply, but isn't really needed, for such models. Instead, it is intended for study of **optimal** (according to some criterion) networks, where we don't have any simple explicit construction of the network.

4 examples; in each we study the total length L_n of the network, in the finite-n random model.

1. Travelling salesman problem (TSP): the network consisting of the edges of the shortest possible route that visits each point exactly once and returns to the starting point.

2. Steiner tree: the network (necessarily a tree) with junctions which minimizes, over all networks connecting the n cities, the total network length.

3. Optimal w.r.t. route-length: for fixed $r_0 > 1$, the minimum-length network subject to $r_\sqcup \leq r_0$, in previous notation.

In discussing existing theory we will start with TSP, the classic example, though examples like (3) are the focus of this course. A somewhat different example which can also be handled by these methods is

4. The shortest path in PPP which starts at origin and goes through some n different points.

Bottom line: in each of these examples there is a constant $0 < \beta < \infty$ such that $n^{-1}\mathbb{E}L_n \to \beta$.

To prove this, there are 2 techniques, which at first look quite different, but are in fact related.

- **Subadditivity:** study the numbers $a_n = \mathbb{E}L_n$ and relate a_{m+n} to $a_n + a_m$.
- **Local weak convergence:** show that the network itself, centered at a random position or a random point of the PPP, converges in distribution (within any fixed window width) to a limit network on the PPP, interpretable as "the same network" on the PPP.

[board: explain connection]

- Subadditivity is more elementary, in that one doesn't need to consider explicitly an infinite network.
- LWC is more powerful, in that it identifies the limit constant in terms of edge-lengths at the root in the planted PPP, and gives extra information, such as distribution of edge-lengths.

Comparisons with other techniques

1. The advantage of these techniques is that they apply to general models; the disadvantage is that they only give information about the first moment.

2. For networks given by an explicit local rule; more precisely, under the assumption that adding one point to the PPP changes the network in only some a.s. finite window; there is a general CLT for L_n due to Penrose - Yukich (2001). The technique ultimately rests on the martingale CLT.

Whether the CLT holds for the TSP is a hard open problem; studying whether the Penrose - Yukich technique works for our kind of "optimal networks" is a **project**.

3. Yet another technique is **concentration inequalities**, giving bounds on large deviations (LD) $\mathbb{P}(|L_n - \mathbb{E}L_n| > x)$.

4. In our spatial setting, the two techniques we study (Subadditivity and Local weak convergence) tend to be applicable to the same examples. But more broadly, the various techniques above are useful in different contexts.

For instance LWC is useful for (abstract) random graphs models which are locally tree-like. A toy example was invented in Aldous-Steele (2003) as the simplest where need this technique

[explain on board]

and one of the highlights is the TSP for the randomly-weighted complete graph

[show java simulation]

5. Breadth of other techniques [xxx not written].

6. In our spatial setting, the techniques for proving CLTs and for proving LDs each depend explicitly on bounding the effect of local changes of vertex-positions; and the same issue arises implicitly in LWC methodology, to show uniqueness of a structure on the infinite PPP defined via some optimization criterion. These three techniques are treated independently in the literature; a (literature survey) **project** is to compare the technical assumptions used, in our spatial setting.