Spatial Networks and Probability

David Aldous

April 17, 2014

This talk has a different style.

Not start with a probability model for a random network and then study its mathematical properties

instead start with some data, then try to find some simplified math model that illuminates the data.

Advice to young researchers: the Socratic question: if you wish to make an impact with new theory, should you
(i) choose a field with extensive theory and no data
(ii) choose a field with extensive data and no theory?

Also: I teach an undergraduate "probability in the real world" course – 20 lectures on different topics, each "anchored" by (ideally) new real-world data that students could obtain themselves.

(22) Psychology of probability: predictable irrationality
(18) Global economic risks
(17) Everyday perception of chance
(16) Luck
(16) Science fiction meets science
(14) Risk to individuals: perception and reality
(13) Probability and algorithms.
(13) Game theory
(13) Coincidences, near misses and paradoxes.
(11) So what do I do in my own research? (spatial networks)
(10) Stock Market investment, as gambling on a favorable game
(10) Mixing and sorting
(9) Tipping points and phase transitions
(9) Size-biasing, regression effect and dust-to-dust phenomena
(6) Prediction markets, fair games and martingales
(6) Branching processes, advantageous mutations and epidemics
(5) Toy models of social networks
(4) The local uniformity principle
(2) Coding and entropy
(-5) From neutral alleles to diversity statistics

A 2012 paper "Evolution of subway networks" concludes

subway systems of very large cities consist of a highly-connected core with branches radiating outwards.

You might regard this as

- a remarkable observation; or
- breathtakingly obvious [cf. MPB]

but it suggests a math project: make a model

- model where people want to travel in a generic city
- model costs/benefit of every possible subway network
- find optimal networks (1-parameter family)
- compare with data

In this field, one finds optimal networks via numerical optimization; actual **proofs** about their detailed nature seem very hard.

This project not done (possible Masters thesis?); here's a similar project which has been done.

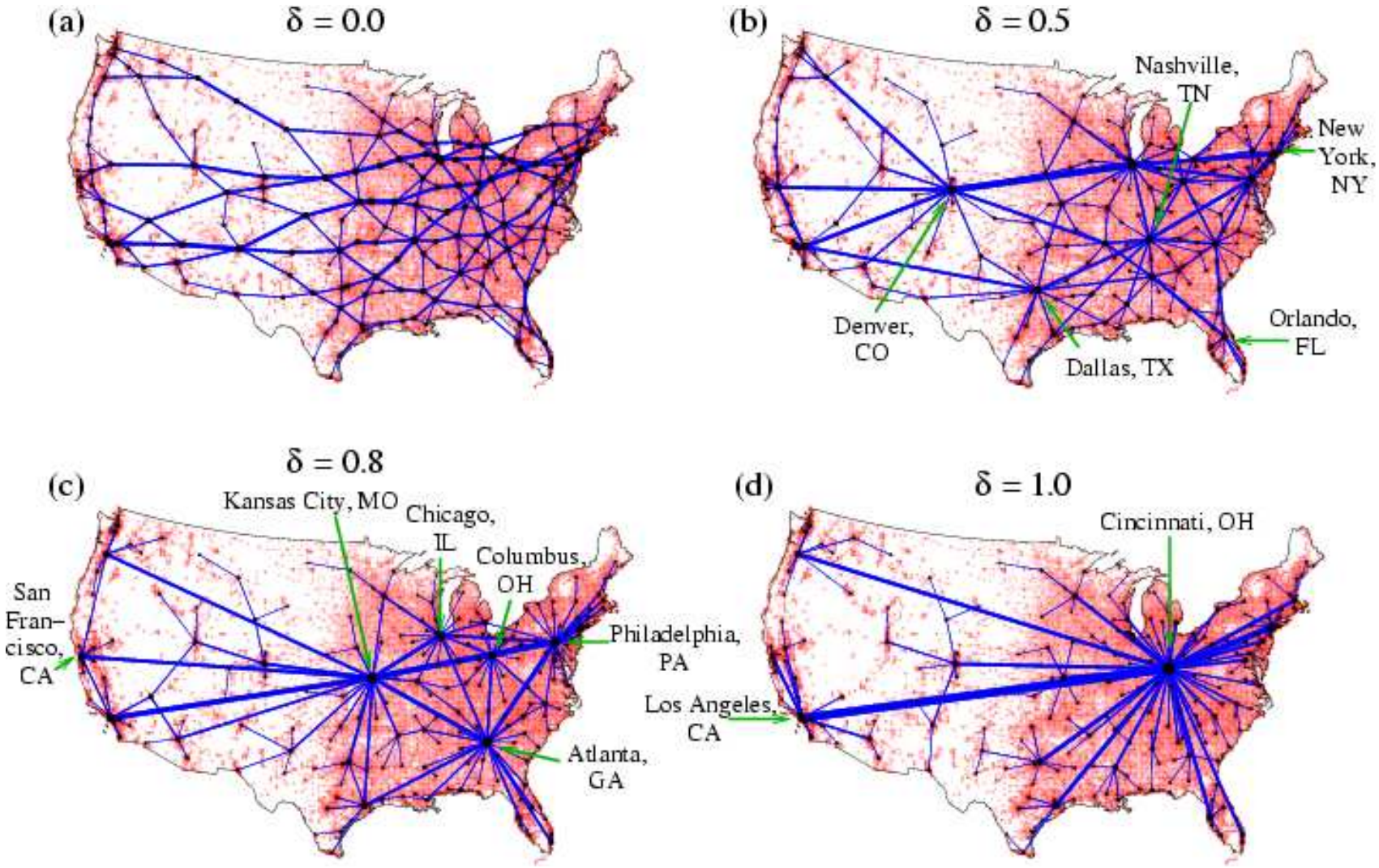


FIG. 5: Optimal networks for the population distribution of the United States with $p = 200$ vertices for different values of δ and with $\gamma = 10^{-14}$.

Figure from 2006 Gastner-Newman paper "Optimal design of spatial distribution networks".

These examples part of a broad "statistical physics" literature surveyed in 2010 Barthélemy "Spatial Networks"; cites 338 papers, most without proofs. So many **open problems** are implicit.

Can we invent much oversimplified math models in which we can **prove** something?

My attempt: *Spatial transportation networks with transfer costs* (2008).

[Where does probability enter the story? – we model city positions as random.]

Hub-and-spoke networks (passenger air travel; package delivery)

Setting: the time to travel a route depends on route length and number of hops/transfers. A weighting parameter Δ controls relative cost of transfers.

For a network $\mathcal{G}_n$ linking n cities $\mathbf{x}_n$ in square of area n, define

$$\begin{gathered}\text{time to traverse a given route from } x_i \text{ to } x_j \\ = n^{-1/2}(\text{ route length}) + \Delta(\text{ number of transfers }).\end{gathered}$$

$$\mathrm{time}(x_i, x_j) = \text{ min. time, over all routes}$$

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

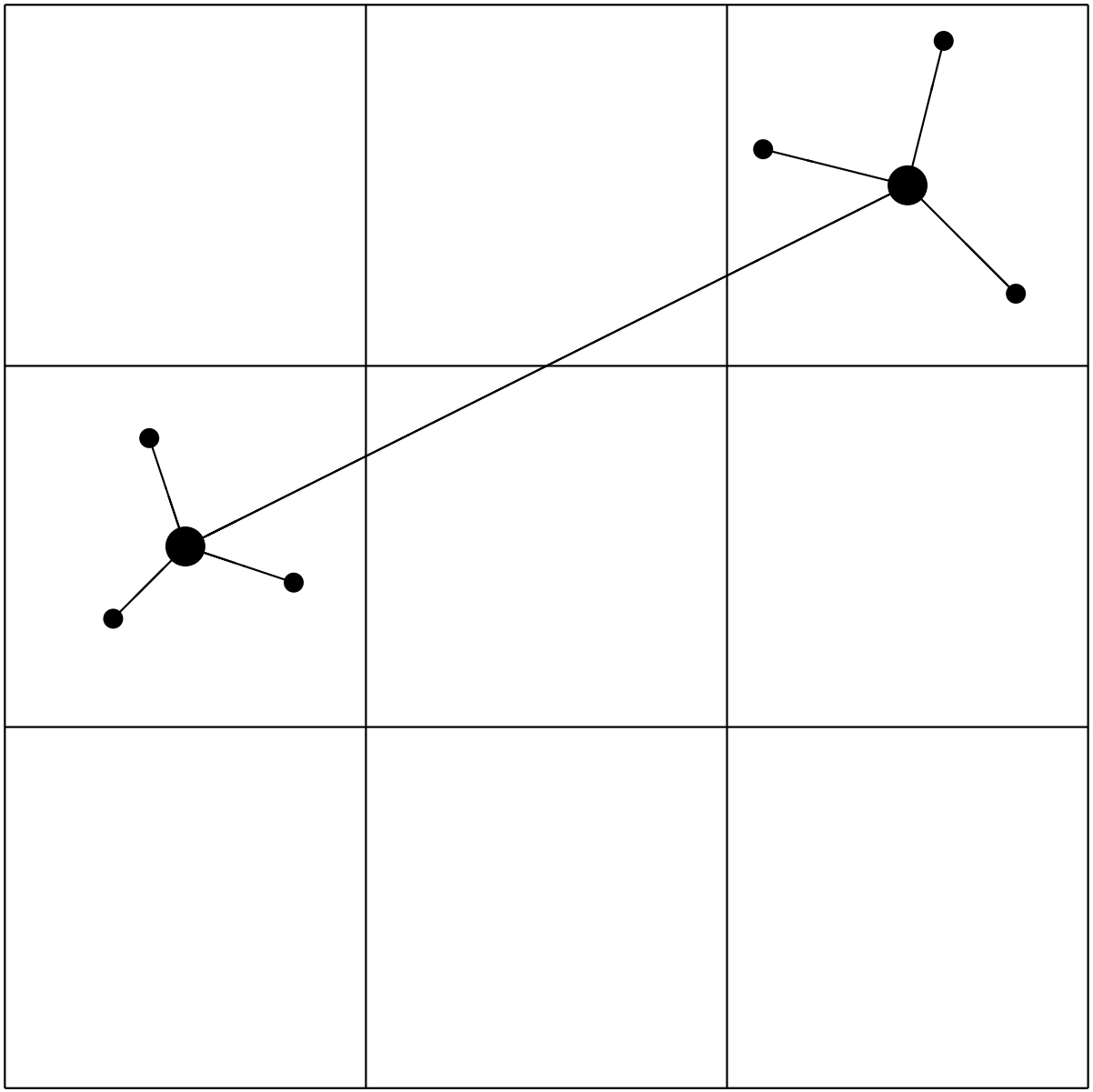
This set-up leads to a 2-parameter question. What network $\mathcal{G}_n$ over cities $\mathbf{x}_n$ minimizes $\mathrm{time}(\mathcal{G}_n)$ for a given value of total length and Δ?

Let's think about designing a network where routes typically involve 3 hops (2 transfers). Here's one scheme.

- Divide area-n square into subsquares of side L.
- Put a **hub** in center of 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 $n^{13/10}$.



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

$$(i) \qquad \text{time}(\mathcal{G}_n) - \text{dist}(\mathbf{x}_n) \to 2\Delta$$

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

Note (i) says mean number of transfers $\to 2$.

Theorem

For "really 2-dimensional" $\mathbf{x}_n$, no networks satisfying (i) can satisfy

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

So the hub-and-spoke network is minimal length in an order-of-magnitude sense. But proving that the minimal-length networks actually "look like" the hub-and-spoke network in some quantifiable way seems very hard.

For **road networks** we have easy access to data (online maps), so it's a useful topic for working with undergraduates. But what to do?

Much literature on spatial networks assumes the **graph network** setting – edges can only be line segments linking specified vertices. We are thinking of edges as physical links and allow junctions – **Steiner networks**.

The *stretch* or *spanning ratio* of a network is the maximum (over vertex-pairs) of the ratio (route-length)/(Euclidean distance). Studied as part of *geometric spanner networks* literature emphasizing algorithms. Consider a network linking the points of a rate-1 Poisson point process on the plane. Write $\Psi^{\text{ave}}(s)$ for the minimum possible mean length per unit area of such a network, subject to the constraint "stretch $\leq s$".

Well-known results for the Delaunay triangulation:

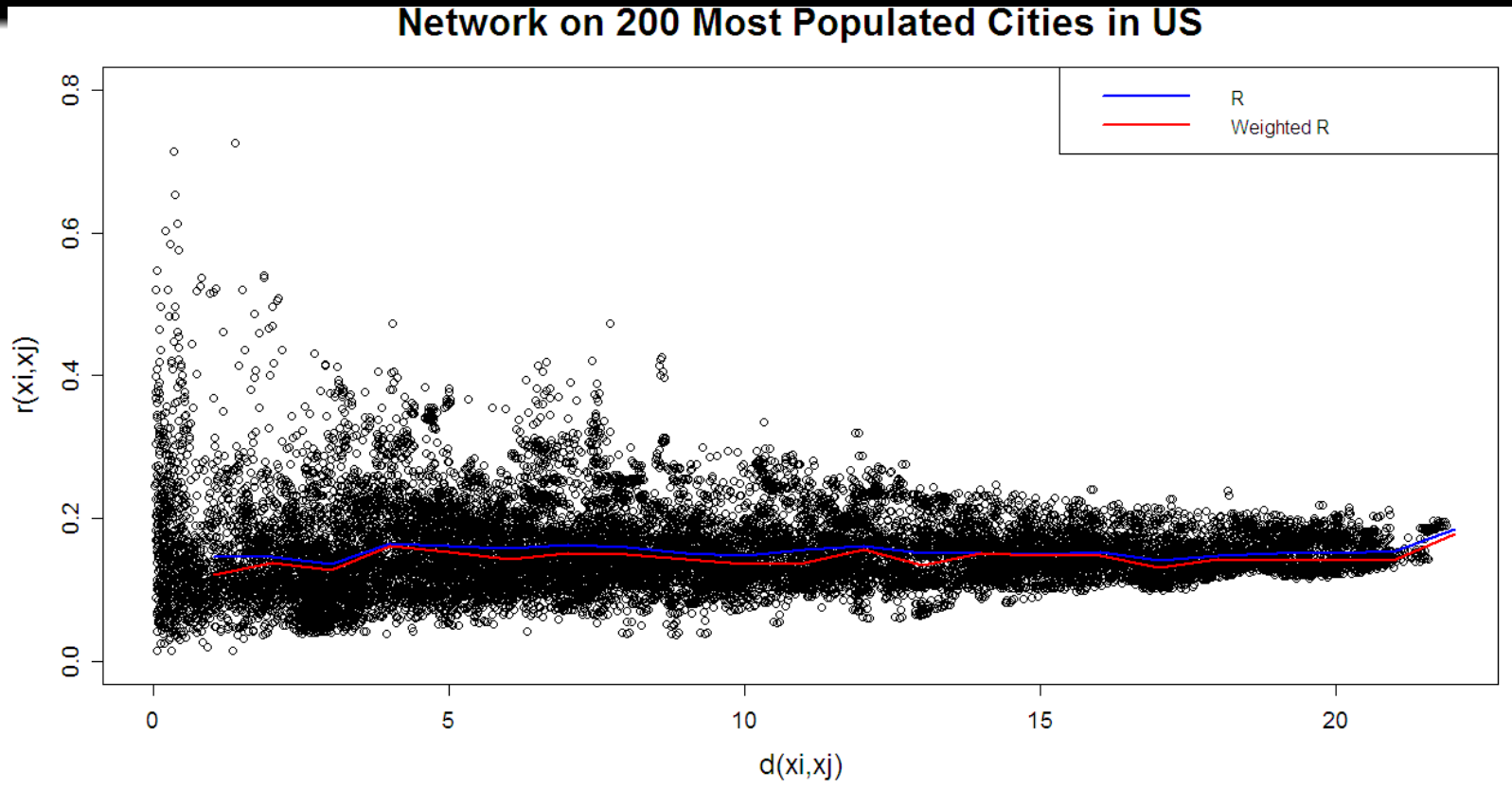
Keil-Gutwin (1992): worst-case stretch ≤ 2.42

Miles (1970): length-per-unit-area for Poisson points $= 3.40$

imply $\Psi^{\text{ave}}(2.42) \leq 3.40$.

In preprint *Stretch - Length Tradeoff in Geometric Networks*. we study upper and lower bounds on $\Psi^{\text{ave}}(s)$ in the Steiner network setting but our bounds are embarassingly crude .....

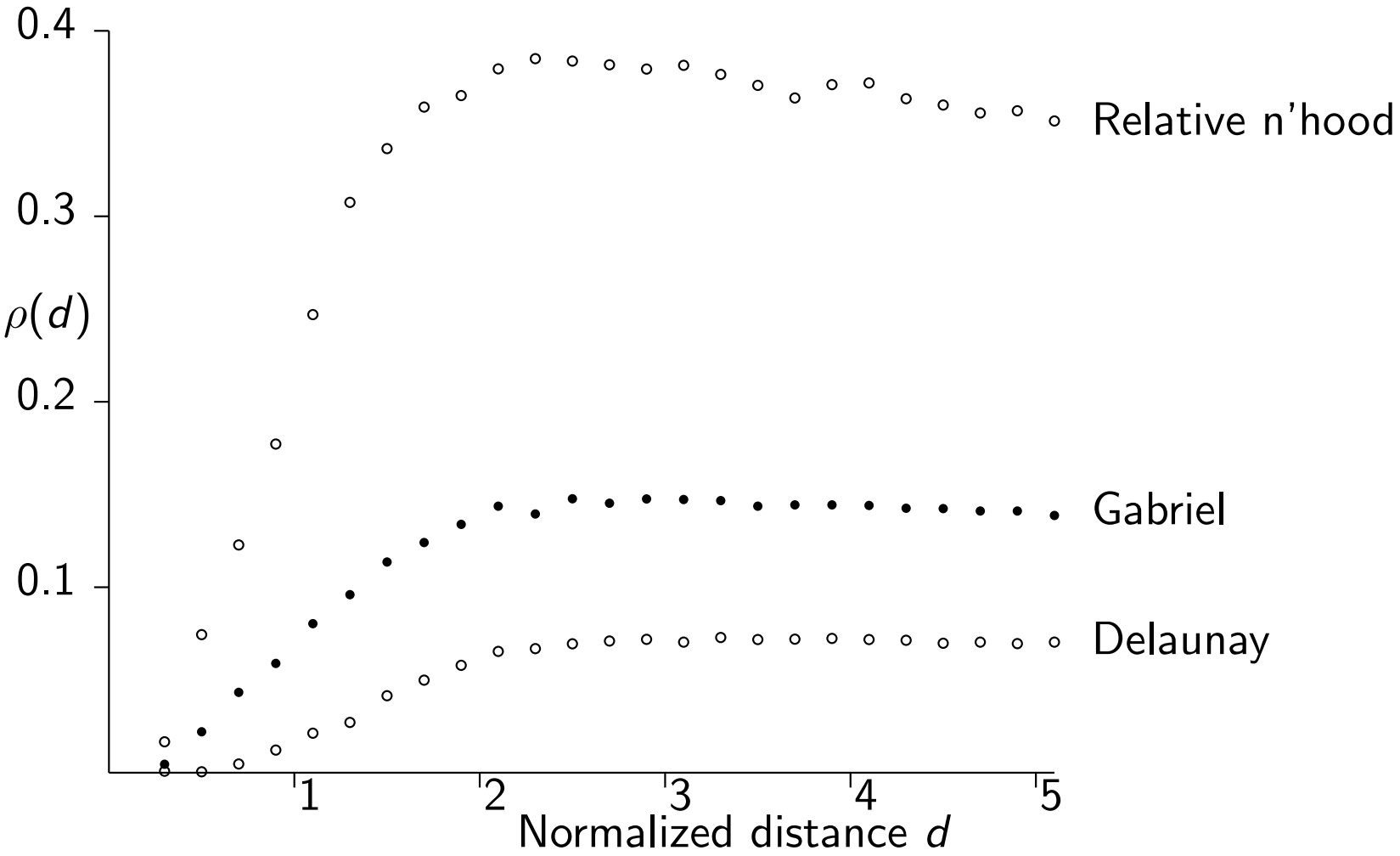
- scope for clever ad hoc constructions to improve upper bounds.
- lower bounds rely on stochastic geometry – seem very hard to improve.



A point for each **pair** of cities
horizontal axis: straight-line distance
vertical axis: 0.3 means route-length 30% longer than straight-line distance

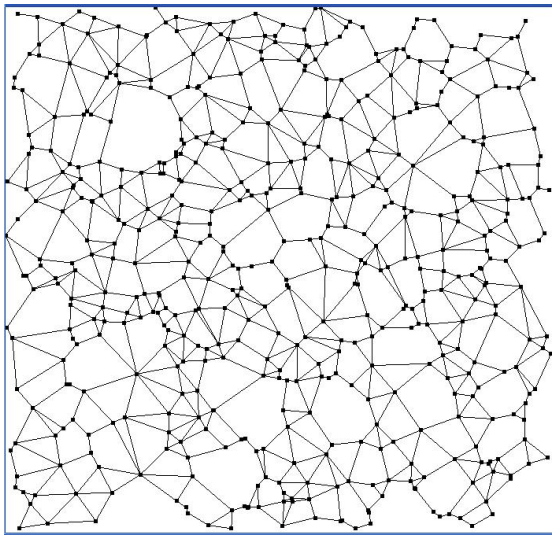
- Average route-length between city centers is $\approx$ 18% longer that straight-line, on all scales.
- "Law of large numbers" for route-lengths.

What do we see in toy models of connected networks over random points?



The function $\rho(d)$ = average excess relative route-length, for cities at normalized distance d, for three theoretical networks on random cities.

But the actual networks look very different



Good news or bad news?

Reasonable to speculate that ‘Law of large numbers” for route-lengths holds very generally. What can we prove?

First imagine a completely general deterministic rule to create edges (city - city roads) for an arbitrary configuration of vertices ("cities") in $\mathbb{R}^2$.
Then require

- rule is translation- and rotation-invariant
- rule always produces a connected network.

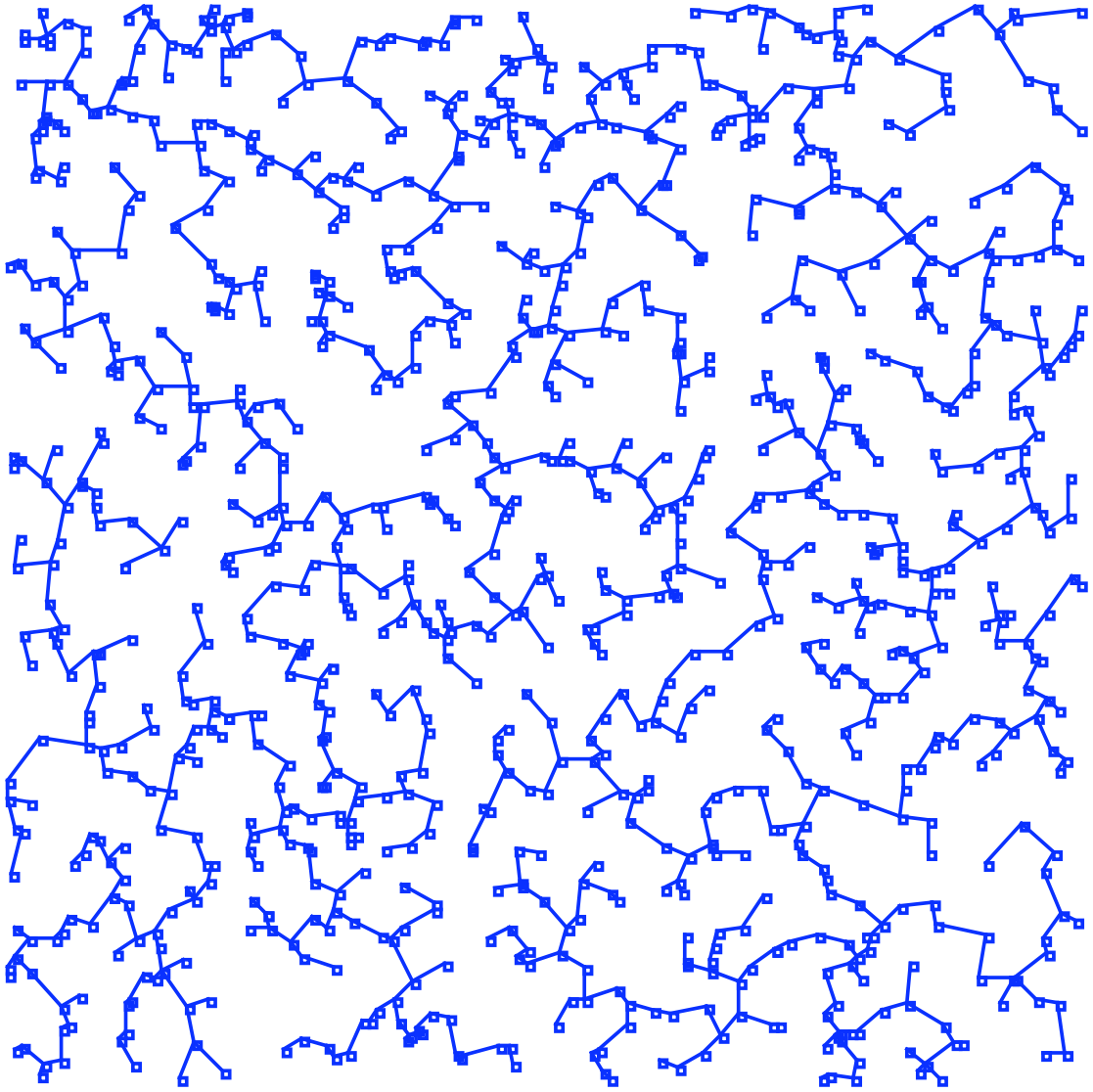
Apply to a Poisson point process of "cities" on $\mathbb{R}^2$ and consider
$R(d) =$ route-length between cities at distance d apart.
If we can prove

$$\mathbb{E}R(d) = O(d) \tag{1}$$

then subadditivity implies (not quite trivially) $d^{-1}R(d) \to c \in [1, \infty)$.

However . . . (1) is not true for MST (or any other tree-network – cf. BLPS (2001) for subtrees of lattice – minor research project to prove?)

Vague Conjecture: (1) true for every network which is not "essentially a tree".



It turns out that one simple condition is sufficient. Consider the $L \times L$ square $[0, L]^2$. Then consider the subnetwork $\mathcal{G}_L$ defined in words as

$$\text{the cities in } [0, L]^2, \text{ with the roads that are present} \\ \text{regardless of the configuration of cities outside } [0, L]^2. \tag{2}$$

The subnetwork $\mathcal{G}_L$ need not be connected, so write N_L^0 for the number of cities inside $[0, L]^2$ that are not in the largest component of $\mathcal{G}_L$.

Theorem

The condition

$$L^{-2}\ \mathbb{E}N_L^0 \to 0 \text{ as } L \to \infty. \tag{3}$$

implies $\mathbb{E}R(d) = O(d)$ *and then* $d^{-1}R(d) \to c \in [1, \infty)$.

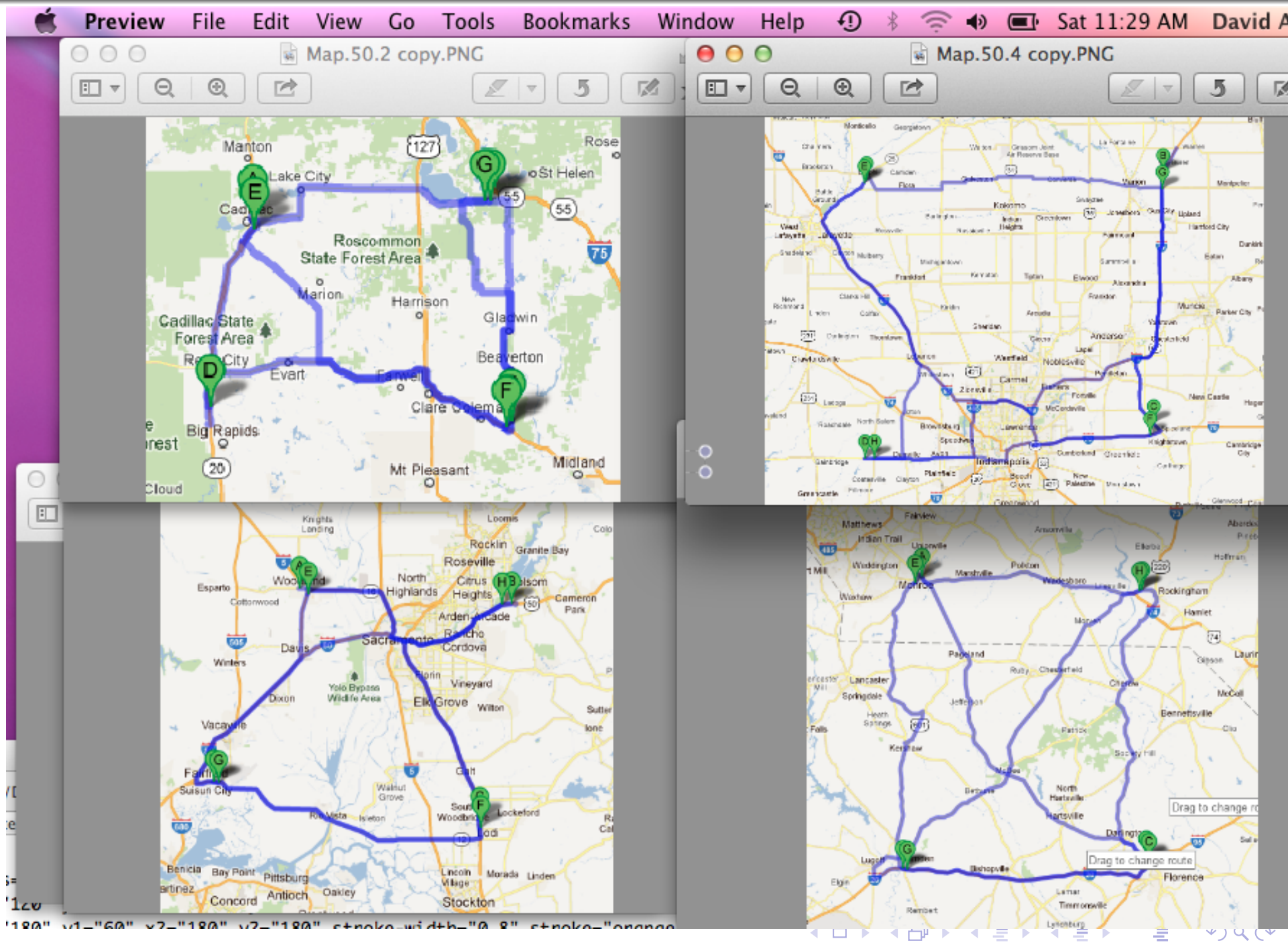
Hack proof in arXiv preprint *Which Connected Spatial Networks on Random Points have Linear Route-Lengths?* – better paper soon . . . Condition (3) is a "no long-range dependence" condition; how does it relate to others: " stability" (Penrose, Yukich . . .); Stein's method?

For **road networks** we have easy access to data (online maps), so it's a useful topic for working with undergraduates. But what to do?

In many science fields (e.g. gene regulatory networks), a large network is studied by looking at frequencies of small subgraphs to see which are most common – **motifs**.

And counting triangles etc is a classic topic within random graph theory.

What about road networks?


Preview File Edit View Go Tools Bookmarks Window Help Sat 11:29 AM David A
Map.50.2 copy.PNG
Map.50.4 copy.PNG
Manton
Lake City
Cadillac
Roscommon State Forest Area
Marion
Harrison
Cadillac State Forest Area
Evart
Beaverton
Big Rapids
Mt Pleasant
Midland
St Helen
Rose
Roseville
Granite Bay
North Highlands
Citrus Heights
Cameron Park
Esparto
Davis
Winters
Vacaville
Dixon
Vineyard
Elk Grove
Suisun City
Benicia
Bay Point
Pittsburg
Concord
Antioch
Oakley
Stockton
Lockeford
Lodi
Matthews
Indian Trail
Waxhaw
Monroe
Rockingham
Hamlet
Pageland
Lancaster
Springdale
Bennettsville
Florence
Drag to change route

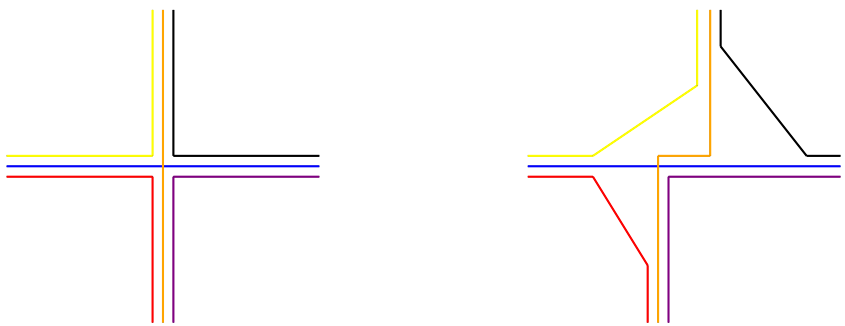
What to do with this data?
Consider the statistics of the sub-network we see in the window; do they depend on real-world side-length of square?

For "proximity graph" type model; routes in large squares would stay close to straight lines.

One idea is to consider topological shapes of subnetworks. Does the distribution over *shapes* vary with scale?

About 70 shapes on 4 points – hard to count carefully – need to make an atlas. Another undergraduate project . . . . . .

file:///Users/davidaldous/Meetings/2014/Leide

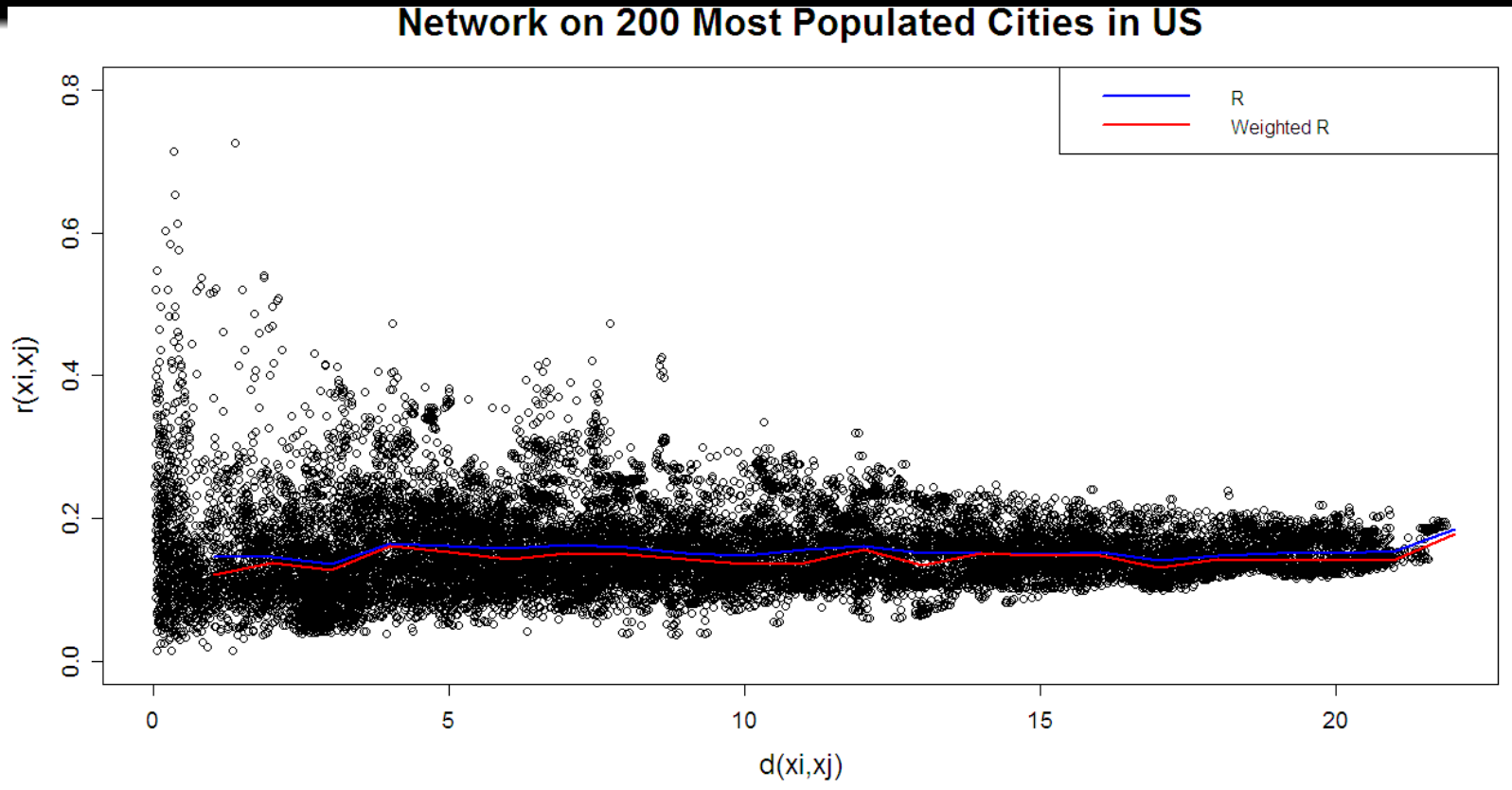


Call a network *scale-invariant* if distribution of subnetworks on sampled points does not depend on the scale.

As first sight this seems contradicted by our previous "law of large numbers" graphic.

But major cities are not "randomly sampled points" in a road network. Conjecture that sampling individual addresses would show the "spread" decreasing more slowly with distance (undergraduate project – how exactly to sample).

As the only relevant data by other people, Kalapala et al (2006) studied proportions of route-length spent on the i'th longest road segment in the route (identifying roads by their highway number designation) and observe that in the U.S. the averages of these ordered proportions for $1 \leq i \leq 5$ are around $(0.40, 0.20, 0.13, 0.08, 0.05)$ over a range of medium to large distances. This is consistent with scale-invariance.



A point for each **pair** of cities
horizontal axis: straight-line distance
vertical axis: 0.3 means route-length 30% longer than straight-line distance

- Average route-length between city centers is $\approx$ 18% longer that straight-line, on all scales.
- "Law of large numbers" for route-lengths.

What do we see in toy models of connected networks over random points?

With rather flimsy support from such data, we have started to study scale-invariant network models. This is a new aspect of a broad technique I call "exchangeable representations of $n \to \infty$ limits of discrete random structures" exemplified by

- continuum random tree as limit of uniform random n-trees
- graphons as limits of dense graphs.

Key idea is to consider induced substructure on k randomly sampled points; first let $n \to \infty$ for fixed k to get a limit continuous structure over k points; these have consistent distributions as k increases, and so define some random structure over infinitely many points.

In context of spatial networks, to be exactly scale-invariant we need to work in the 2-dimensional continuum (cf. random walk and Brownian motion) – a network specifies a route between arbitrary $z_1, z_2 \in \mathbb{R}^2$. Formalizing this idea requires some work – 6 page overview paper *True scale-invariant random spatial networks* – but starts with thinking of subnetworks on finitely many points as analog of "finite-dimensional distributions" of stochastic processes.

Final open problem.

There are two standard 1-parameter families of graph networks on arbitrary configurations.
• (well-known) β-skeleton family
• (less well known) "power law cost" family. Parameter $1 < p < \infty$ cost of jump x_i to x_j is $||x_j - x_i||^p$. Include the edge (x_i, x_j) iff cheapest route from x_i to x_j is the direct jump.

Open Problem: Give corresponding schemes – algorithmically simple and mathematically natural – for defining Steiner networks on arbitrary configurations

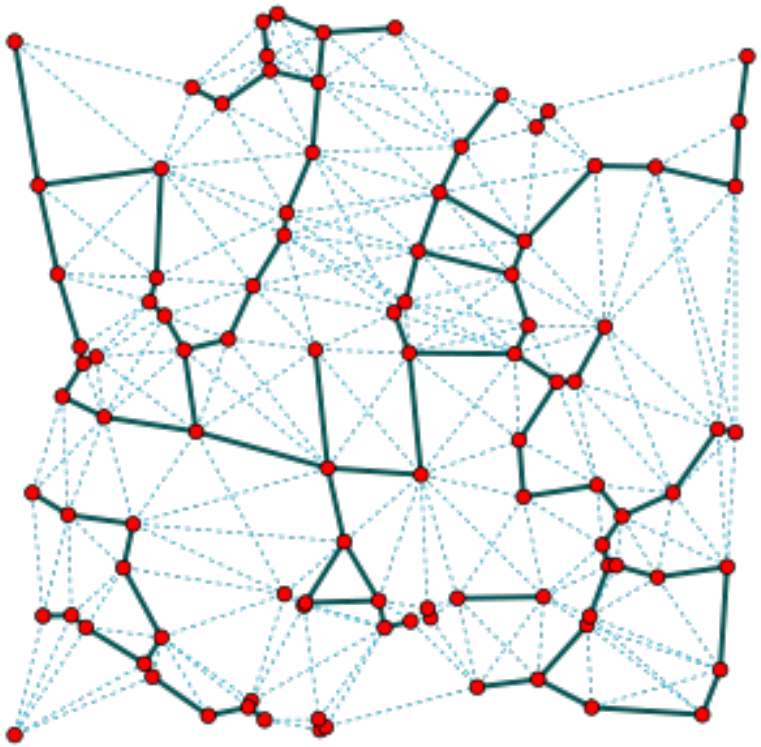
In particular, do any such schemes look like real-world road networks?

Beta skeleton

From Wikipedia, the free encyclopedia

In computational geometry and geometric graph theory, a ***β*-skeleton** or **beta skeleton** is an undirected graph defined from a set of points in the Euclidean plane. Two points p and q are connected by an edge whenever all the angles prq are sharper than a threshold determined from the numerical parameter β.

Contents



The circle-based 1.1-skeleton (heavy dark edges) and 0.9-skeleton (light dashed blue edges) of a set of 100 random points in a square.

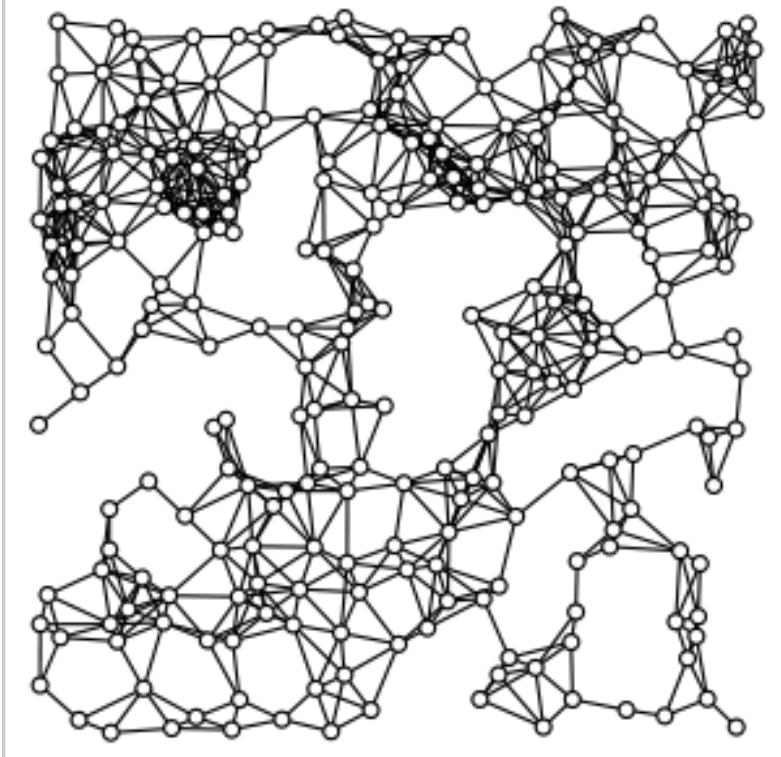
Random geometric graph

From Wikipedia, the free encyclopedia

In graph theory, a **random geometric graph** is a random undirected graph drawn on a bounded region, e.g. the unit torus $[0, 1)^2$. It is generated by

1. Placing vertices at random uniformly and independently on the region
2. Connecting two vertices, u, v if and only if the distance between them is at most a threshold r, i.e. $d(u, v) \leq r$.

Several probabilistic results are known about the number of components in the graph as a function of the threshold r and the number of vertices n.



Example of Random Geometric Graph with 256 vertices and distance=0.1.

References

- Penrose, Mathew: *Random Geometric Graphs* (Oxford Studies in Probability, 5), 2003.