

# A Route-Length Efficiency Statistic for Road Networks

David J. Aldous\*      Alan Choi

June 4, 2009

## Abstract

This note compares some current theoretical mathematical work on spatial networks with data on inter-city road networks within States. In designing a network, a natural constraint is the total length, and a natural objective is to want the route length  $\ell(i, j)$  between typical cities  $i, j$  to be not much longer than straight line distance  $d(i, j)$ . Write  $r(i, j) = \frac{\ell(i, j)}{d(i, j)} - 1$  for the relative excess route-length. With an  $n$ -city network there are  $\binom{n}{2}$  such numbers  $r(i, j)$ . A recent theoretical insight is that, from a mathematical viewpoint, a good way to combine these into a single “route-length efficiency” statistic  $R^*$  is to define  $R^*$  as the maximum over  $d$  of the average of  $r(i, j)$  over city-pairs with  $d(i, j) \approx d$ . The optimal tradeoff between total network length and  $R^*$  is discussed elsewhere in a theoretical model of randomly-placed cities. What about real networks? For each of 10 U.S. States we studied the road network linking the 20 largest cities. In this note we present the data and discuss the relationship between data and the predictions from theory.

---

\*Department of Statistics, 367 Evans Hall # 3860, U.C. Berkeley CA 94720; aldous@stat.berkeley.edu; www.stat.berkeley.edu/users/aldous. Aldous’s research supported by N.S.F Grant DMS-0704159.

# 1 Introduction, theory and data

Quantitative aspects of road networks have been studied from many different viewpoints – see [6] for a brief recent review and section 2.4 for previous work most closely related to ours. This note is a facet of an ongoing mathematical project with somewhat different focus (see [1] for an overview) to which we would like to draw the attention of those interested in mathematical aspects of transportation research. If a goal of a transportation network linking cities is to provide short routes, with a constraint on the total length of the network, then one can compare a real network on given cities with a hypothetical optimal network of the same total length, choosing the hypothetical network to optimize some statistic  $R$  measuring how efficient a network is in providing short routes. For reasons discussed in [1] we use the following non-obvious statistic, specified more carefully in section 2.3 below. For cities  $i, j$  write  $\ell(i, j)$  for route-length and  $d(i, j)$  for straight line distance between  $i$  and  $j$ . Then write  $r(i, j) = \frac{\ell(i, j)}{d(i, j)} - 1$  for the relative excess route-length. Now define a function

$$\rho(d) = \text{average of } r(i, j) \text{ over city-pairs with } d(i, j) \approx d \quad (1)$$

and finally define

$$R^* = \max_d \rho(d). \quad (2)$$

The mathematical theory [1] assumes a model in which cities are at random positions and follows a methodology based on taking limits as the number  $n$  of cities tends to infinity; within that framework it studies the optimal relationship between  $R^*$  and a normalized total network length statistic  $L$  (*normalizations* are described in section 2.1). Of course neither “random positions of cities” nor “number of cities  $\rightarrow \infty$ ” seems very realistic. Can there be any connection between this theory and real networks?

For each of 10 U.S. States we studied the road network linking the 20 largest cities. Figure 1 shows the scatter diagrams of all values of  $r(i, j)$  and normalized distance  $d(i, j)$  for 4 of the States (the entire set of 10 diagrams and complete underlying data is available<sup>1</sup>), together with estimates of the curve  $\rho(d)$ , with and without population-weighting (see section 2.3). Here are two aspects of the data that can be compared with theoretical predictions for optimal networks.

---

<sup>1</sup><http://www.stat.berkeley.edu/users/aldous/Spatial/>

**Qualitative features of the curve  $\rho(d)$ .** Nearby cities are more likely than distant cities to have a direct road link, so one might naively suppose that  $\rho(d)$  would increase smoothly with  $d$ . But in fact theory predicts (see Figure 2 and its legend) that  $\rho$  should increase rapidly to a maximum at normalized distance 2 - 3, then very slowly decrease as  $d$  increases further. Our data is not extensive enough to indicate small  $d$  behavior, but it does show that  $\rho$  is roughly constant over  $1 \leq d \leq 5$ , consistent with the part of the theoretical prediction that  $\rho$  should be almost constant over  $2 \leq d \leq 5$ .

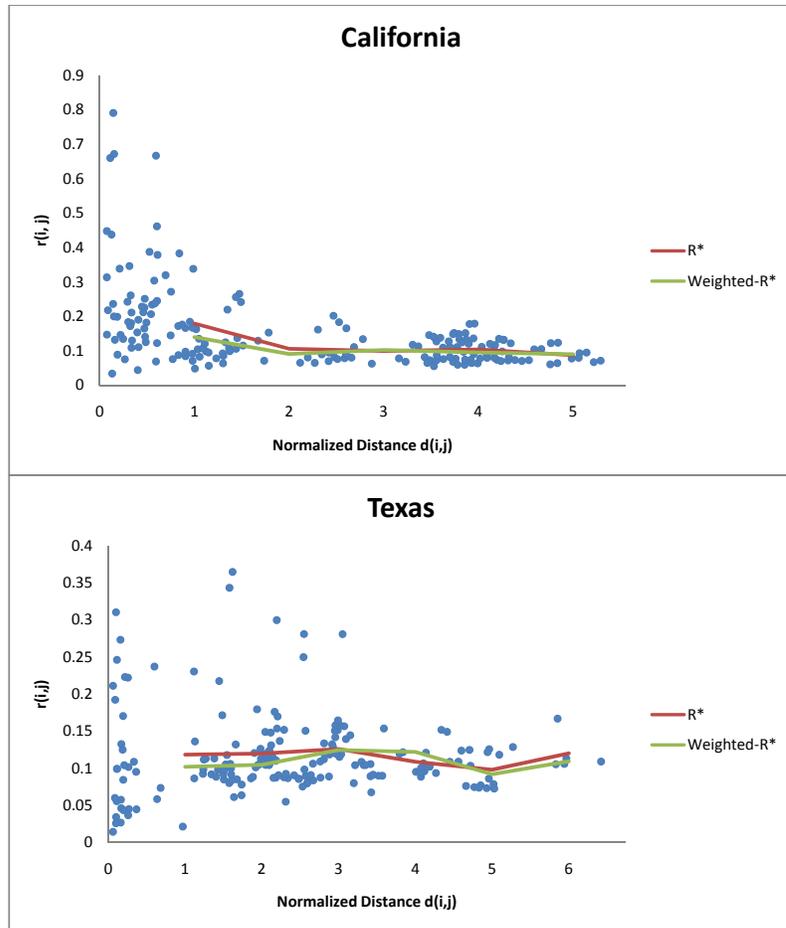
**Tradeoff between  $R^*$  and normalized network length  $L$ .** There is a theoretical prediction for the tradeoff between  $R^*$  and normalized network length  $L$  in near-optimal networks, shown in Figure 3, where it is compared with the values of the pair  $(L, R^*)$  of statistics for each of the 10 States data. At first sight there seems almost no relation between theory and this raw data. However there is an issue not yet addressed. Theory envisages cities positioned uniformly randomly in a State of area  $A$ , and the value of  $A$  enters into the normalization. The left outliers in Figure 3 are Michigan and California, in which the 20 largest cities are concentrated in relatively small-area regions of the State. In comparing data with theory it would be better to replace the full area  $A$  (in normalization formula (3)) by a smaller area  $A'$  representing the “effective area” in which the largest cities lie. This would have the effect of increasing the normalized network length  $L$ . But any way of actually implementing this replacement seems rather arbitrary, so we have not done so.

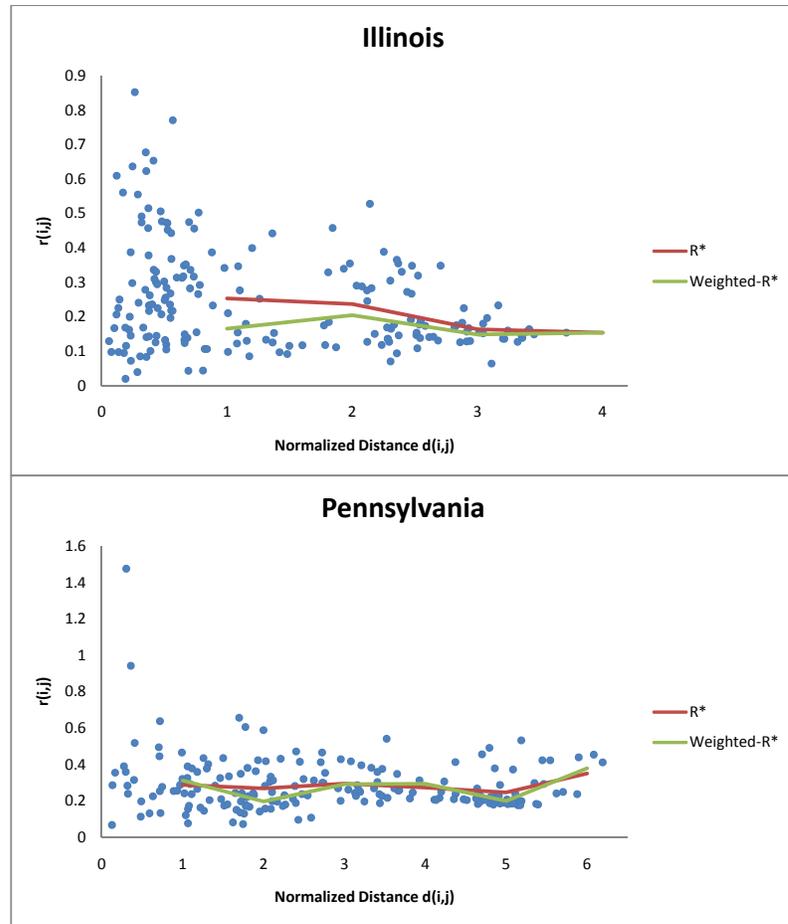
## 1.1 Conclusions.

One can only draw tentative conclusions from such a small set of data, and indeed our main goal is to encourage professionals with better experience of GIS databases to examine larger data-sets from this viewpoint, for instance for the road networks of large European countries and the rail networks of India and China.

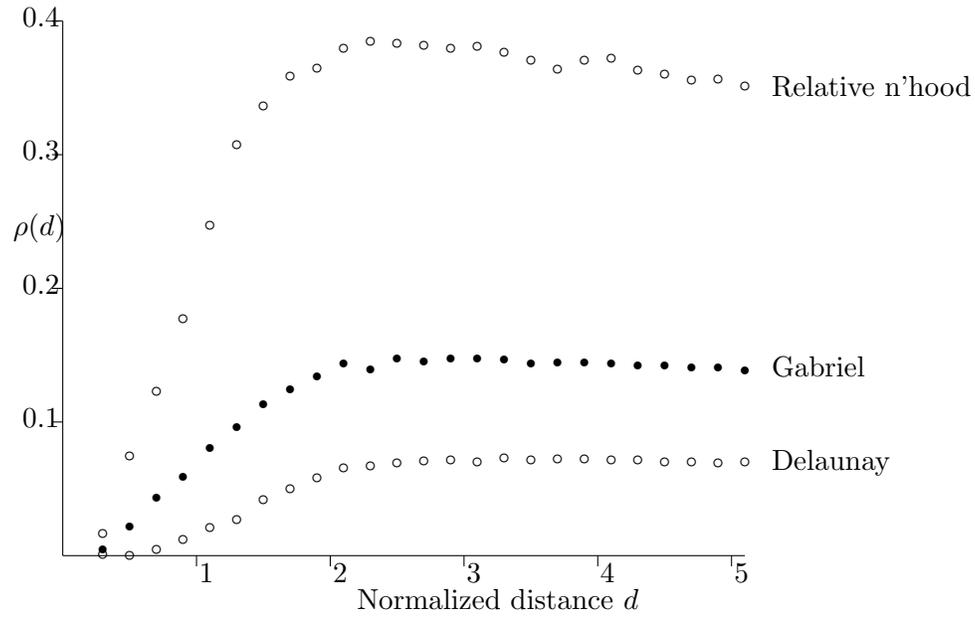
The conceptual point of the underlying theory was to try to quantify the obvious qualitative “law of diminishing returns”, that to improve route-length efficiency beyond a certain point requires a substantial increase in total network length. The theoretical results in Figure 3 suggest values of around 2 for normalized length and around 0.25 for  $R^*$  as a reasonable trade-off. We boldly conjecture that for real-world networks in which short route-lengths are a major desideratum, and that are perceived as efficient in actually having short routes, their summary statistics  $(L, R^*)$  will typically

be near those values, after some correction for the “effective area” issue above.

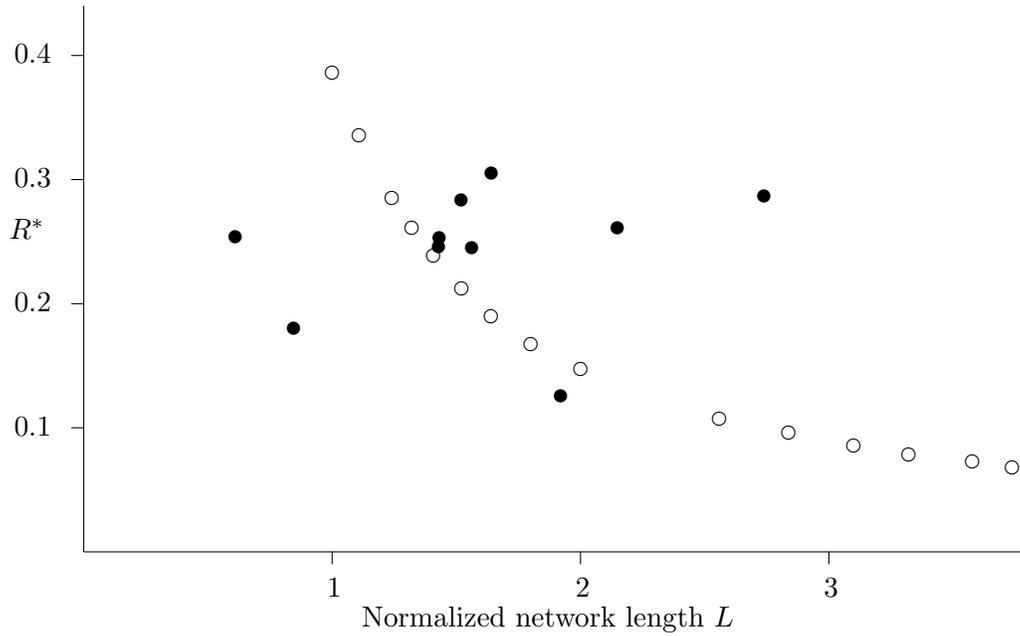




**Figure 1.** Scatter diagrams of normalized distance and relative excess route-length for city-pairs, together with estimates of the nonlinear regression function  $\rho(d)$  as predictor for  $r(i, j)$  given  $d(i, j) = d$ .



**Figure 2.** The function  $\rho(d)$  for three theoretical networks on random cities (copied from Figure 6 of [1]). Irregularities are Monte Carlo random variation. The numerical values differ because the total network lengths differ; but for every reasonable-looking theoretical network we have studied, the curve  $\rho(d)$  shows the same characteristic shape, increasing rapidly to a maximum at normalized distance 2 - 3, then very slowly decreasing.



**Figure 3.** Tradeoff, for different networks, between the normalized network length  $L$  and the route length efficiency statistic  $R^*$ . The  $\circ$  show the beta-skeleton family [3], a theoretical family believed to be near-optimal for uniform random points (this part copied from Figure 7 of [1]). The  $\bullet$  show the values for our data on the road network on 20 cities in each of 10 States.

## 2 Some technical details

### 2.1 Normalizations

To compare networks with different numbers  $n$  of cities and in spatial regions of different areas  $A$ , it is convenient to normalize by setting

$$\text{standard unit of length} = \sqrt{A/n} \quad (3)$$

so that the density of cities is 1 per unit area. Then, measuring distances in standard units, we can define

$$L := n^{-1} \times (\text{total length of network})$$

as normalized network length per city, and we regard  $L$  as the “normalized cost” of the network.

### 2.2 Data collection

The States studied were CA, TX, NY, FL, IL, PA, OH, MI, GA, NC. Data on inter-city distances, route-lengths, city populations and States areas is readily accessible (we used Google Earth). To find the total network length is harder – we need the total length of the subnetwork of the State road network that is used in the shortest route between some two of the largest 20 cities – and we could not completely automate this process, instead needing to manually derive and examine the subnetwork. This is the practical reason we were unable to analyze larger networks.

### 2.3 Estimating the curve $\rho(d)$ from data

Conceptually,  $\rho(d)$  is a nonlinear regression predictor for  $r(i, j)$  given  $d(i, j) = d$ . One could derive it from data using whatever algorithm is built into one’s statistical package; instead we did the more naive procedure of averaging over city-pairs whose distance apart (in standard units) was the same when rounded to the closest integer. This gives averages  $\tilde{\rho}(1), \tilde{\rho}(2), \dots$  plotted in Figure 1, and the reported value of  $R^*$  was the maximum over  $1 \leq d \leq 5$  (because there were very few city-pairs at larger distance).

Such *unweighted* averaging does not take city populations into account. Intuitively we expect that larger cities are more likely to have direct road links, so we also calculated *weighted* averages, where each city-pair was weighted by the product of population sizes. Confirming intuition, the Figure 1 data shows the weighted version of  $\rho$  to typically be slightly smaller than the unweighted version.

Trying to estimate  $\rho(d)$  from data for small  $d$  (say,  $d < 1/2$ ) isn't very sensible for obvious reasons; a city is not really a single point, and (representing a city by some arbitrary central point) freeways are typically not designed to run through city centers, so the model isn't realistic on such a small scale.

## 2.4 Related work

Spatial networks arise in many different disciplines, and the idea of comparing route length and straight line length also arises naturally. Road networks in different countries were studied in [2] who calculated (in our notation) the average of  $r(i, j)$  over all pairs, finding a wide variation between countries (from 0.12 to 1.1) and in particular obtaining 0.2 for the network on 299 U.S. cities. Theoretical work in computer science studies the *maximum* value of  $r(i, j)$  over all pairs, under the name *stretch* [5]. The only occurrence we know of a curve analogous to  $\rho(d)$  is in [4] Figure 3, in the context of home-work commutes in an intra-metropolitan network. There  $\rho$  decreases sharply, from 0.42 in the 5-10 km range to 0.2 in the 45-50 km range. However there are so many real-world factors here – house-building along roads with good access to city center – that one expects such networks to be statistically quite different from inter-city networks, for which we advocate using  $R$  as summary statistic.

*Acknowledgements.* Preliminary data collection not shown here was done by Yanjiao Cheng, Jesse Friedman, Yu-Jay Huoh, Wayne Lee and Harrison Liu and supported by an N.S.F. VIGRE Grant and by the U.C.B. Undergraduate Research Apprentice Program.

## References

- [1] D.J. Aldous and J. Shun. Models for connected networks over random points and a route-length statistic. In preparation. Draft available at <http://www.stat.berkeley.edu/users/aldous/Spatial/poisson.pdf>, 2009.
- [2] R.H. Ballou, H. Rahardja, and N. Sakai. Selected country circuitry factors for road travel distance estimation. *Transportation Research A*, 36:843–848, 2002.

- [3] D.G. Kirkpatrick and J.D. Radke. A framework for computational morphology. In G.T. Toussaint, editor, *Computational Geometry*, pages 217–248. Elsevier, 1985.
- [4] D. Levinson and A. El-Geneidy. The minimum circuitry frontier and the journey to work. Unpublished, 2009.
- [5] G. Narasimhan and M. Smid. *Geometric spanner networks*. Cambridge University Press, Cambridge, 2007.
- [6] F. Xie and D. Levinson. Measuring the structure of road networks. *Geographical Analysis*, 39:336–356, 2007.