Network Analysis and Modeling, CSCI 5352 **Prof. Aaron Clauset**
Lecture 10 **2016**

1 Spatial networks

Many real-world networks exist physically in space. That is, with each vertex we can associate a location in some $\mathbb{R}^d$ space. Spatial networks are the general term we apply to describe a class of networks that includes transportation networks (road, rail, subway, air travel, walking, shipping, etc.), distribution networks (oil and gas, power, etc.), and more. What makes spatial networks different from many other types of non-spatial networks is that the probability that two vertices are connected depends in some way on their distance, e.g., vertices separated by a smaller distance may be more likely to be connected. This simple difference can have a dramatic impact on the structure of networks.

In addition to the more common questions of the structure and dynamics of spatial networks, another property that spatial networks sometimes exhibit is that of *navigability.* Time allowing, we will also examine the kinds of general patterns found in real-world networks that are embedded in 2- or 3-dimensions, as most of our spatial networks are.

2 Navigable networks

Stanley Milgram's famous small-world experiment demonstrated two interesting properties of social networks. First, it demonstrated that short paths exist between arbitrary pairs of individuals. This property is what we typically mean when we say a network is a small world (although formally, we mean that the diameter of a network model grows like $O(\log n)$.)

The experiment also demonstrated that these paths can often be found using only local information, i.e., without resorting to a global search algorithm (like Dijkstra or BFS) in order to know how to route a packet from one vertex to another. Instead, each vertex can make a reasonably accurate guess about which of its neighbors is "closer" to an arbitrary destination in the network. Repeating this process at each vertex serves to route the packet across the network.

This property is, in fact, somewhat remarkable. In a simple random graph model, the only way to find a short path between some i and j is to run a full global search algorithm, i.e., to explore the entire network first. But Milgram's experiment demonstrated that social networks are embedded within a kind of latent space, and that each individual stores in their mind a rough map of this latent space and their approximate position on that map. Then, by working together, individuals can effectively route information across the social network without either knowing the entire social network structure or running an expensive global search algorithm in their mind.

This is what we mean by *navigability*, that when each vertex only stores local information about the network, i.e., information about its local portion of the graph, there exists a protocol or algorithm that can use this local information to find short paths between pairs of individuals.

2.1 The Kleinberg model

To formalize the notion of navigability, Jon Kleinberg took the Watts-Strogatz model and modified the way the rewired links were rewired so that they improved the navigability of the network.

Unlike the Watts-Strogatz model, this model assumes that the underlying lattice network was fixed in place, and that these "local" links cannot be rewired. On top of this lattice (whose presence guarantees that some path, albeit likely a long one, always exists between some pair of vertices) is layered a set of "long-range links" that provided routing short cuts. In the simplest version of the model, we assume a k-dimensional lattice, in which each vertex connects to all of its nearest neighbors. Thus, a vertex has $2k$ local connections, and these connections are bidirectional: if u connects to some v, then v also connects back to u. In addition, each vertex u has a single long-range link (u, x), where x is chosen uniformly among the vertices some distance d away (see Figure 1 on the next page).[1] If we choose d uniformly at random, then this model is very similar to the Watts-Strogatz model.

Kleinberg showed that when the probability distribution for link lengths follows a power law with exponent $r = k$, i.e., the scaling exponent r equals the dimensionality of the lattice, then a *greedy routing* algorithm will deliver packets in $O(\log^k n)$ steps. (While this time is *not* $O(\log n)$, it is still quite fast, even for large networks.) The greedy routing algorithm is simply the same procedure Milgram gave his participants: examine all of your neighbors (the $2k$ local neighbors and the 1 long-range neighbor) and forward the packet to the neighbor whose remaining distance to the target is smallest; repeat at the new location until the destination is reached.

The general outline of this result can be seen as follows. For some source u and target v, we want to route a packet between the two. At each intermediate vertex x, we will make the greedy choice, choosing to forward the packet to the vertex among x's neighbors that minimizes the remaining distance to v. We divide the total routing time into a sequence of "phases," where the jth phase ends when the packet is within 2^j steps of the target. Thus, there can be at most $\log_2 n$ phases in the routing. If the probability is $1/\log n$ that some x in the current phase has a long-range link to a vertex in the next phase, then the expected total number of steps will be $O(\log^2 n)$.

To begin, we set the probability distribution for the length of a long-range connection to follow a power law form:

$$\Pr(u \to v) = \frac{d(u,v)^{-r}}{\sum_{u \neq v} d(u,v)^{-r}} \ , \tag{1}$$

[1] This figure reprinted from J. Kleinberg, "The Small-World Phenomenon: An Algorithmic Perspective." *Proc. 32nd ACM Symposium on Theory of Computing* (2000).

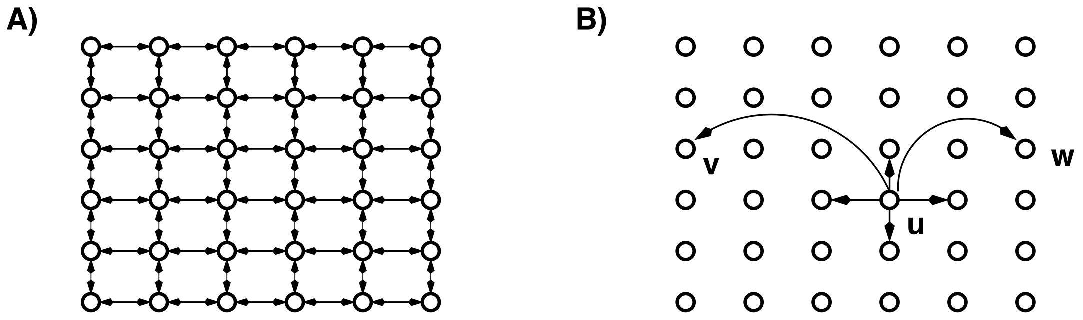


Figure 1: The Kleinberg model, showing (A) the fixed local, bidirectional connections and (B) the neighbors for a particular vertex u, which includes (directed) long-range connections.

where we define the distance measure $d(u, v)$ to be the Manhattan distance on the lattice between u and v, i.e., $d(u, v) = \sum_{i=1}^{k} |u_i - v_i|$, and r is the exponent (called α in previous lectures). For the remainder of the analysis, we will assume a $k = 2$ dimensional lattice, and thus the optimal routing occurs when $r = k = 2$.

This choice fixes Eq. (1) and allows us to simplify its denominator:

$$\sum_{u \neq v} d(u, v)^{-2} \leq \sum_{j=1}^{2n-2} (4j)(j^{-2}) \ , \tag{2}$$

where the first term counts the number of vertices at a distance j in a $k = 2$ lattice, and the second term is the probability that u links to a vertex at a distance j. Simplifying further yields

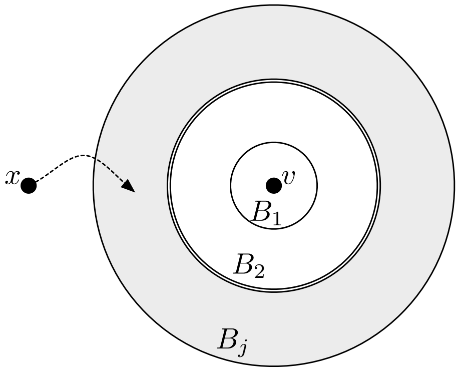
$$\begin{aligned} \sum_{u \neq v} d(u, v)^{-2} &= 4 \sum_{j=1}^{2n-2} j^{-1} \\ &\leq 4\,(1 + \ln(2n - 2)) \\ &\leq 4\,(\ln 3 + \ln 2n)) \\ &\leq 4 \ln 6n \ . \end{aligned}$$

This expression then allows us to rewrite Eq. (1) as

$$\Pr(u \to v) \geq \frac{d(u, v)^{-r}}{4 \ln 6n} \ , \tag{3}$$

which provides a normalized distribution.

Now consider a packet traveling from some u to some v, which we divide into a set of "phases," where phase j is defined as the packet being at some vertex x such that $2^j < d(x,v) \leq 2^{j+1}$. Thus, the 0th phase begins when $d(x,v) \leq 2$, and the packet is at most two steps away from the target. In general, a phase ends when the distance between the packet and the destination has been halved. In this way, we are modeling the routing as a kind of binary search, and $j \leq \log n$.



When does the jth phase end? From the above definition, the jth phase ends when $d(x,v) < 2^j$. Each time the packet is passed along a local connection, it gains a new chance to find a long-range link that will end the phase, i.e., link to a vertex within a distance 2^j of the target v. Each such long-range link is independent of any other, and thus the probability that such an event happens is the probability that x connects to some $w \in B_j$, where B_j is the set of vertices within distance 2^j of the target v. The vertex x could connect to any of those vertices, of which there are

$$\begin{aligned} 1 + \sum_{i=1}^{2^j} i &= 2^{2j-1} + 2^j + 1 \\ &= \frac{1}{2} 2^{2j} + \frac{1}{2} 2^j + 1 \\ &> 2^{2j-1} \quad . \end{aligned} \tag{4}$$

Furthermore, each of these vertices is within a distance $2^{j+1} + 2^j < 2^{j+2}$ of x. Thus, the probability that x links to some $w \in B_j$ is

$$\begin{aligned} \Pr(x \to w) &= \frac{d(x,w)^{-2}}{\sum d(x,w)^{-2}} \\ &\geq \left[\left(2^{2j+4}\right)\left(4 \ln 6n\right)\right]^{-1} \quad . \end{aligned}$$

This therefore implies that the probability the phase comes to an end at vertex x is

$$\begin{aligned}\Pr(j\text{th phase ends at } x) &\geq \left(2^{2j-1}\right)\left[(2^{2j+4})(4\ln 6n)\right]^{-1} \\ &= \frac{1}{128\ln 6n} \ . \end{aligned} \tag{5}$$

How many steps are there in the jth phase? Let X_j count the number of such steps in the jth phase. The expected value of X_j is thus

$$\begin{aligned}\mathrm{E}[X_j] &= \sum_{i=1}^{\infty}\Pr(X_j \geq i) \\ &\leq \sum_{i=1}^{\infty}\left(1-\frac{1}{128\ln 6n}\right)^{-1} \\ &= 128\ln 6n \ . \end{aligned}$$

which is exactly equal to $1/\Pr(j\text{th phase ends at } x)$.

Finally, the total number of steps to deliver the packet is the sum of the lengths of each of the phases. Recall that because each phase halves the remaining distance to the target, there can be at most $\log_2 n$ phases. Thus, expected total time is

$$\begin{aligned}\mathrm{E}[X] &= \mathrm{E}\left[\sum_{j=0}^{\log_2 n} X_j\right] \\ &= \sum_{j=0}^{\log_2 n}\mathrm{E}\left[X_j\right] \\ &\leq (1+\log_2 n)(128\ln 6n) \\ &\leq \alpha_2 \log^2 n = O(\log^2 n) \ . \end{aligned}$$

Kleinberg's analysis treated the more general case of unspecified r, and he showed that $r = k$ is a special value. If $r < k$, i.e., if the power-law link-length distribution is *more* heavy tailed than is optimal, then routing slows down, because most of the links at x overshoot the target area B_j. In the other direction, if $r > k$, i.e., if the power-law link-length distribution is *less* heavy tailed than necessary, then routing also slows down, not because most of the links at x undershoot the target area. Figure 2 shows the general pattern.

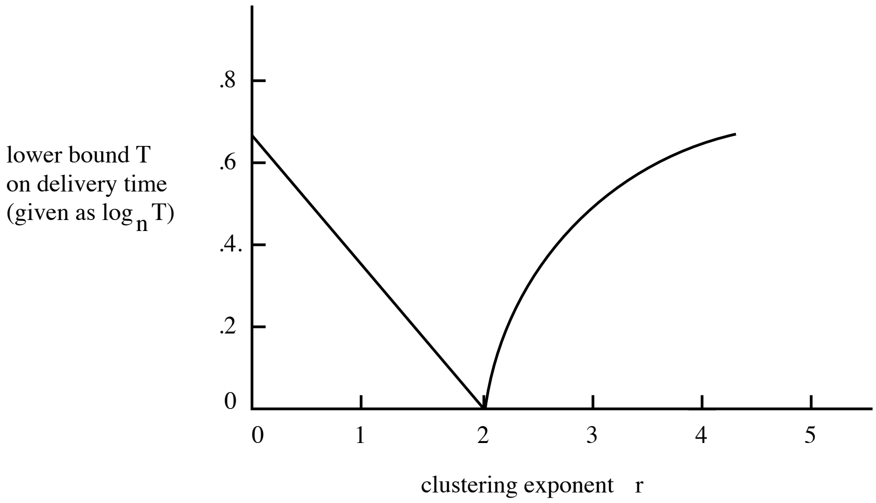


Figure 2: Kleinberg's general result on the routing time, as a function of the link-length distribution exponent r, showing that only for $r = k$ do we achieve optimal routing.

2.2 What about real networks?

Although Kleinberg's result is for a toy model with highly unrealistic structure, its central assumption, that the link-length distribution in social networks should follow a power-law distribution with a particular structure in order to produce the efficient routing observed by Milgram, holds up when we examine real social networks.

David Liben-Nowell and a number of colleagues tested this idea using data from a large and public social network from the early 2000s called LiveJournal. In this social network, each node is a kind of blog, which links to other blogs in the LiveJournal network. Crucially, many blogs list a zip code, which gives us an estimate of the author's physical location. For each pair of blogs connected by an edge, if both listed a zip code, we may estimate the physical "length" of the link between them. One difference between the LiveJournal network and the Kleinberg model is that each vertex in this network can have many long-range links, which thereby increases the probability that some x has a link that takes us closer to some destination.[2]

[2] This difference may be important. If the long-range out-degree distribution follows a certain pattern, we can still achieve efficient routing even with a non-nice link-length distribution. I don't think this has been worked out exactly, but perhaps it should be. It does imply that even if the link-length distribution doesn't follow exactly the pattern Kleinberg showed is necessary, we may be able to compensate with the out-degree distribution.

Taking this information, the empirical link-length distribution does indeed very roughly follow a power-law distribution (Figure 3, left).[3] However, the approximate exponent for this distribution is closer to 1.2, which is significantly heavier-tailed than expected from Kleinberg's result, which predicts $\alpha \approx k = 2$.

The reason is that the physical locations of individuals are not arranged in a lattice or distributed uniformly across the 2-dimensional surface of the Earth. Instead, they tend to clump together in cites and other urban/suburban areas, and along the coasts of the US. This non-uniform distribution implies that a different number of individuals are within some distance δ from u, depending on where on the map u is located.

Liben-Nowell et al. showed that this difference leads to efficient routing occurring at a slightly different value of the scaling exponent.[4] In Kleinberg's model, the number of individuals a distance δ away grows linearly. In a non-uniform density situation, however, the number of individuals within a distance δ can vary considerably. However, for every vertex, we may rank other individuals by their distance to u. By redefining $\Pr(u \to v)$ to be inversely proportional to the rank of v in u's list (raised to some power), Liben-Nowell et al. recover the same kind of model that Kleinberg studied, but adapted to the non-uniform case. When population is uniform, this model reduces to Kleinberg's model exactly, while for a non-uniform population, the probability that u connects to v depends on the number of people within some distance $d(u, v)$.

The result is that efficient routing is the result of combining a rank-distance relationship (how far away is the jth closest neighbor) and the link-length distribution. In the LiveJournal network, the former is roughly linear and the latter is roughly inversely related, which gives us something close to the value of 2 expected for routing in a social network embedded on a 2-dimensional surface.

3 At home

1. Reread Chapter 8.2 (pages 241–242)
2. Read Chapter 15.1 (pages 552–564) in *Networks*

[3] Both figures reprinted from D. Liben-Nowell et al., "Geographic routing in social networks." *PNAS* **102**(33), 11623–11628 (2005).

[4] With some differences: there is a connection here with fractals, and Liben-Nowell et al. estimate the fractal dimension of the LiveJournal network to be close to 0.8, which means the underlying geography is non-linear, in contrast to Kleinberg's model.

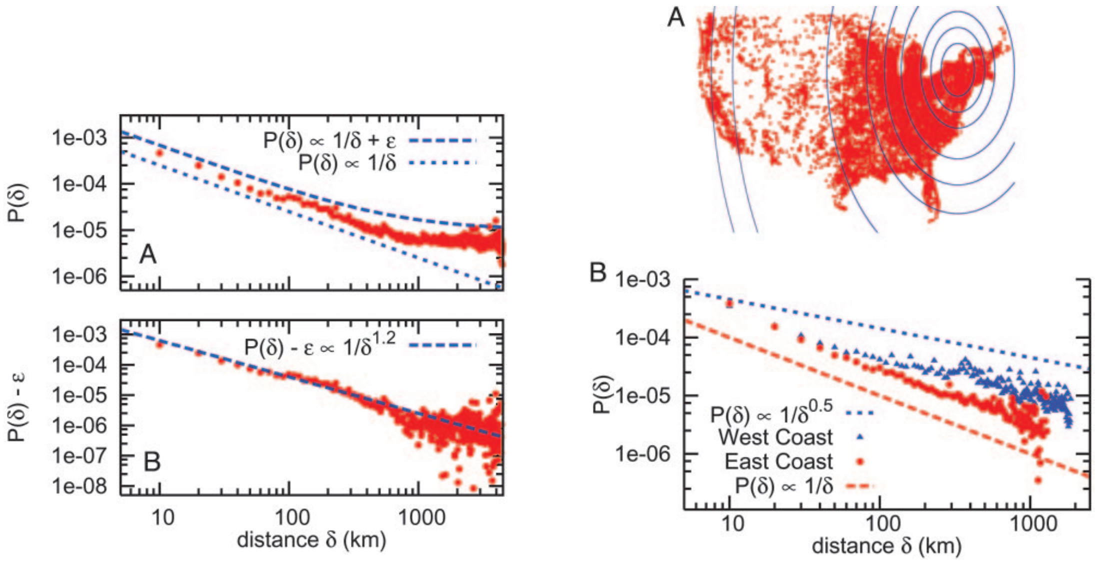


Figure 3: (left) The empirical link-length distribution for vertices in the LiveJournal blog network, when both ends of an edge can be approximately geolocated. (right) The same, but now with a correction for the non-uniform population density.