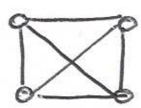
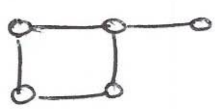
LECTURE 5

In many networks one can identify sub-groups of vertices sharing similar properties. One of such properties is connectivity. Namely, there are groups of connected vertices forming "communities".
Many algorithms have been developed to identify them automatically. Before describing these algorithms we first introduce some notions of connectivity and provide some intuition on how these components arise using RANDOM NETWORK MODELS

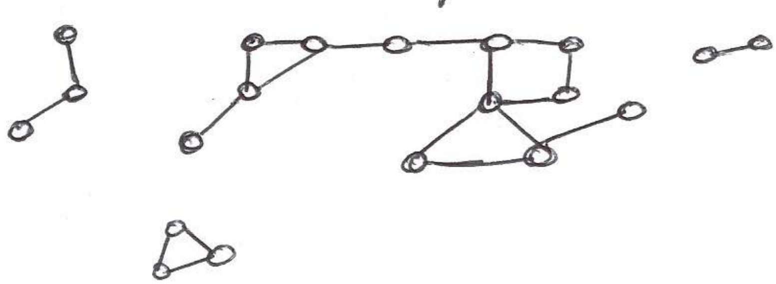
(1) Clique or strongly connected network.
There is an edge between any two vertices



(2) Connected component There is a path between any two vertices



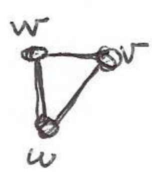
(3) Giant connected component. A connected component containing a positive fraction of all vertices



α-connected component contains an α-fraction of the total number of nodes.
If a network contains an α-connected component we say it is α-connected.

Many networks that we observe in practice are α-connected for some $0<\alpha<1$ and we are interested in the properties of the giant component. For example:

- k-connectivity — there are k indep paths that connect any two vertices
- shortest paths — length of shortest path between any two vertices
- diameter — largest among shortest paths between any two nodes
- clustering



a path u-v-w that is closed forms a closed triad.

clustering coeff. is the fraction of paths of length 2 edges that are closed

$$C = \frac{\#\text{ closed triads}}{\#\text{ paths of length } 2} = \frac{(\#\text{ triangles})\cdot 6}{\#\text{ paths of length } 2}$$

$$= \frac{(\#\text{ triangles})\cdot 3}{\#\text{ connected triples}}$$

In general social networks tend to have: giant component, small diameter, small (average) shortest paths, a larger number of independent paths, and high clustering. In other words they exhibit

COMPLEXITY

Our generative models must try to reflect these properties as well as centrality, degree distributions, navigability, optimization of flows etc. that we can discover / observe from data.
Coming-up with the "right" generative model is difficult.

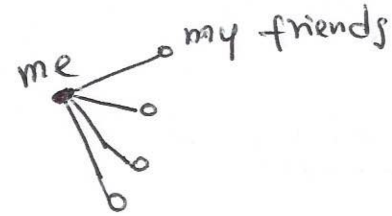
EXAMPLE

Typical values of clustering coefficients observed in practice

$C = 0.10,\ 0.20$ for social networks

This indicates (not surprisingly) that friends are not chosen at random from a population but follow some more complicated rules for connectivity. Consider a not very realistic random model:

Assume everybody has K friends. There are a total of n nodes.
A completely random selection of friends implies



each of my friends will create a closed triad with probability $\frac{K-1}{n} \simeq K/n$ because they select at random w. prob $1/n$ any one of my $(K-1)$ friends.

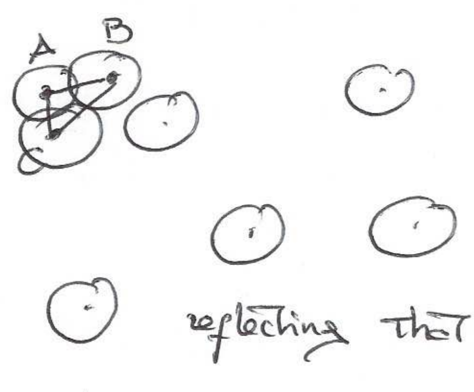
Hence $P(\text{any two my friends are friends})$

$\simeq K/n \ll 1$ for large n

$\Rightarrow$ for large random networks the clustering coeff. would be much smaller than 0.10 or 0.20. Clearly, a completely random model is not realistic.

In practice, many other factor influence the selection of friends which does not occur independently at random, namely if $A \leftrightarrow B$ & $B \leftrightarrow C$, then more likely $A \leftrightarrow C$.

Consider for example a Random Geometric Graph where two nodes are connected if their distance is d_{ij}: $\|x_i - x_j\| < r$ and nodes are distributed at random.



In this model although the node positions are random iid, $P(A \leftrightarrow B)$ is not indep from $P(A \leftrightarrow C)$ reflecting that it is more likely for people in the same geographic region to know each other.

GIANT COMPONENT

How can we make sure to generate models that have giant connected components?

Fortunately random graph theory & percolation theory can help us with this.

EXAMPLE 1. Random Graphs

\# nodes N

\# edges $p_{ij} = p$

Theorem If $p = \frac{c}{n}$, we have:

- $\boxed{c < 1}$ $\Rightarrow$ largest connected component is $O(\log N)$ vertices, meaning $P(|G| \leq K \log N) \to 1$ for $N \to \infty$ and $K < \infty$
- $\boxed{c > 1}$ $\Rightarrow$ $\exists\ K(c) > 0$: $P(|G| \geq K n) \to 1$ for $N \to \infty$

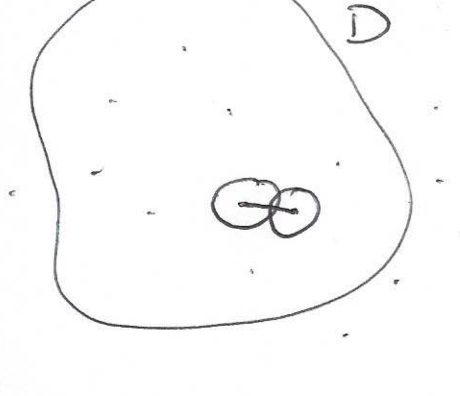
EXAMPLE 2 Random geometric graphs

Consider a Poisson-distribution of nodes in $\mathbb{R}^2$

This is defined by the two properties:

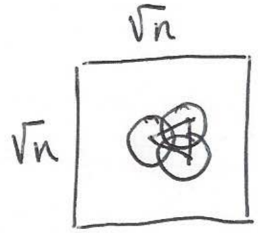
(1) $D_1\ D_2 \dots D_k$ disjoint $\Rightarrow$ $X(D_1)\ X(D_2)\ \ X(D_k)$ are indep. where $X(D) = \#$ points inside D

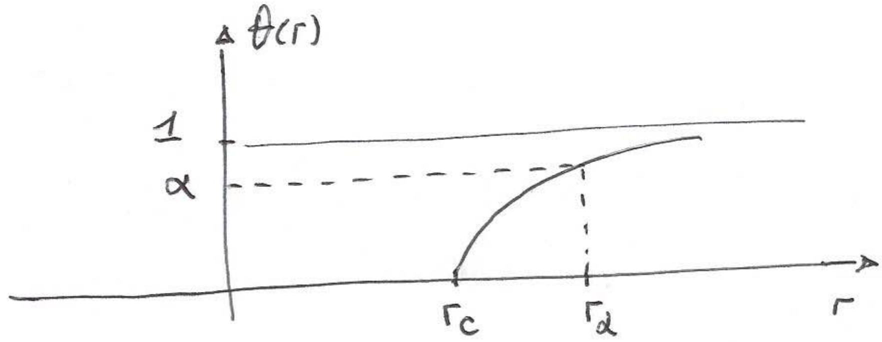
(2) $D \subset \mathbb{R}^2 \Rightarrow \forall k > 0 \quad P(X(D) = k) = e^{-\lambda |D|} \frac{(\lambda |D|)^k}{k!}$



Consider $\lambda = 1$ & add edge between any two points at distance $\leq 2r$

Focus on a box $\sqrt{n} \times \sqrt{n} = B_n$





$G_n(r)$ is α-connected if it contains a component of at least αn vertices.

__Thm__ $\forall\ 0<\alpha<1$ if $r > r_\alpha \Rightarrow P(G_n(r) \text{ is } \alpha\text{-connected}) \longrightarrow 1$
$r < r_\alpha \Rightarrow P(G_n(r) \text{ is } \alpha\text{-connected}) \longrightarrow 0$

Note that here r_α plays the role of $K(c)$ in the random graph model. The larger the (constant) radius, the larger the fraction of connected nodes. Similarly the larger the expected node degree the larger the fraction of connected nodes

$$\mathbb{E}(\text{node degree}) = \pi \lambda (2r)^2 = \pi 4 r^2$$

Moreover there is a __critical__ expected node degree (viz radius) for which we have α-connectivity and this value is

$$r_\alpha = \inf\{r : \theta(r) > \alpha\}$$

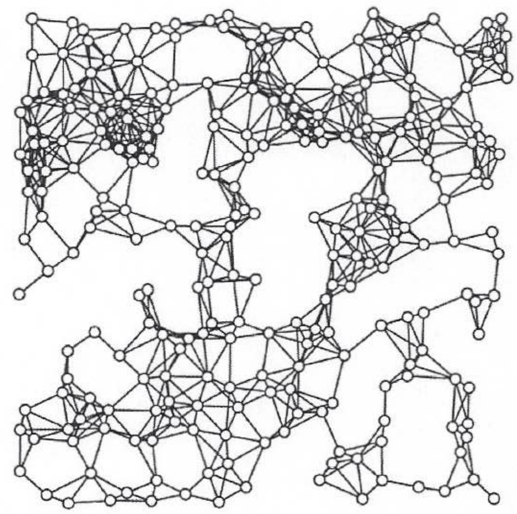
The function $\theta(r)$ is called the __percolation function__ for the random geometric graph model.

__Question__ what is the expected node degree for having $\alpha > 0$? This corresponds to $r > r_c$. The precise value is unknown but by simulations we have $\mathbb{E}(\text{node degree}) > 4.5$

__In contrast__ for random graphs $\mathbb{E}(\text{node degree}) > 1$ suffices to have a giant component with $\alpha > 0$.

Random geometric graph

In graph theory, a **random geometric graph** (**RGG**) is the mathematically simplest spatial network, namely an undirected graph constructed by randomly placing N nodes in some metric space (according to a specified probability distribution) and connecting two nodes by a link if and only if their distance is in a given range, e.g. smaller than a certain neighborhood radius, r.



Example of Random Geometric Graph on a flat 2-d closed square [0, 1] with $N=256$ vertices and connectivity threshold $r=0.1$.

Random geometric graphs resemble real human social networks in a number of ways. For instance, they spontaneously demonstrate community structure - clusters of nodes with high modularity. Other random graph generation algorithms, such as those generated using the Erdős–Rényi model or Barabási–Albert (BA) model do not create this type of structure. Additionally, random geometric graphs display degree assortativity: "popular" nodes (those with many links) are particularly likely to be linked to other popular nodes.

A real-world application of RGGs is the modeling of ad hoc networks.[1]

Examples

- In 1 dimension, one can study RGGs on a line of unit length (open boundary condition) or on a circle of unit circumference.
- In 2 dimensions, an RGG can be constructed by choosing a flat unit square [0, 1] (see figure) or a torus of unit circumferences $[0, 1)^2$ as the embedding space.

The simplest choice for the node distribution is to sprinkle them uniformly and independently in the embedding space.

References

1. Nekovee, Maziar (28 June 2007). "Worm epidemics in wireless ad hoc networks". *New Journal of Physics*. **9** (6): 189–189. arXiv:0707.2293 (https://arxiv.org/abs/0707.2293). doi:10.1088/1367-2630/9/6/189 (https://doi.org/10.1088%2F1367-2630%2F9%2F6%2F189).

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