Networks: September 12

October 12, 2016

1 Agenda

1. Random Graphs
   (a) Giant component and small world
   (b) Statistics
   (c) Problems: degree distribution, triangles
2. Centralities and display

A Graph $G = (V, E)$ has edges $E \subseteq V \times V$
Adjacency matrix $A_{ij} = 1$ if $(i, j) \in E$ else $A_{ij} = 0$ if $(i, j) \notin E$

2 Random Graphon Model

$G(n, p)$ where $n = \#$ nodes, $p =$ probability of an edge
All edges form independently with probability p.
In the directed version $A_{ij} \overset{iid}{\sim} Bern(p)$
In the undirected version $A_{ji} = A_{ij} \overset{iid}{\sim} Bern(p)$
All edges are independent and equi-probable.
$Degree(i) \sim Binomial(n-1, p)$ so degree $\to \infty$ as $n \to \infty$ and p constant: "dense graph (sequence) limit"

Alternate Parametrization
$\lambda = (n-1)p =$ mean degree (so $p = \frac{\lambda}{n-1}$)
This gives an alternate limit: $n \to \infty$ but λ constant: "sparse graph (sequence) limit"
In this limit, $Degree(i) \rightsquigarrow Pois(\lambda)$: "Poisson Random Graphs"

3 Large Scale Connectivity

How many connected components?
At $p = 1$, obviously only 1 CC (Connected Component)
At $p = 0$, no CC's
Claim: at some magic intermediate value, 1 CC has size $\mathcal{O}(n)$, all others are smaller. Below that, all CC's will be small and won't scale with the size of the graph.

Self-Consistency Argument Suppose such a "giant" CC exists
Say that a fraction $S > 0$ of all nodes are in the giant CC (call it GC)

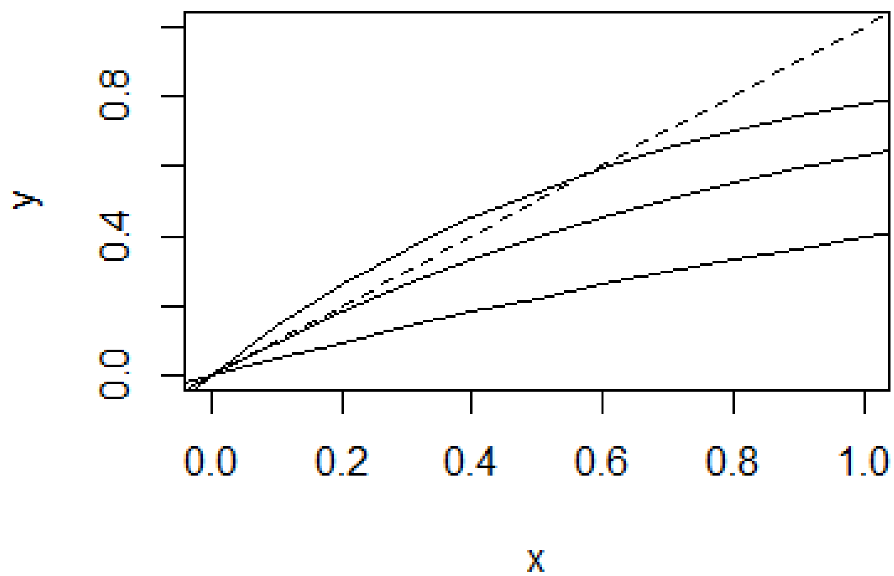

Figure 1: Self-consistent argument solutions

Pick a favorite node, the probability it is in GC is S. So the probability it is not in GC is:

$$\begin{aligned}
1-S &= (1-p+p(1-S))^{n-1} \\
&= (1-pS)^{n-1} \\
\Rightarrow 1-S &= \left(1-S\frac{\lambda}{n-1}\right)^{n-1} \\
\Rightarrow log(1-S) &= (n-1)log\left(1-S\frac{\lambda}{n-1}\right) \\
\Rightarrow log(1-S) &\approx -(n-1)\frac{S\lambda}{n-1} \\
\Rightarrow log(1-S) &\approx -S\lambda \\
\Rightarrow 1-S &\approx e^{-S\lambda} \\
\Rightarrow S &\approx 1-e^{-S\lambda}
\end{aligned}$$

If $\lambda < 1$, no self-consistent way to get a GC
If $\lambda > 1$ there is a solution $s*$ which increases as λ increases

"Epidemic" Picture Start at the favorite node i

Z_i first neighbors, $Z_i \sim Pois(\lambda)$
Each first neighbor has a random number of neighbors.

$Z_1 \sim Pois(\lambda) \Rightarrow E(Z_i) = \lambda$

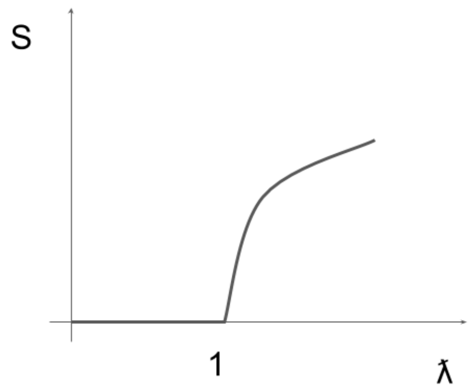

Figure 2: Phase transition Process

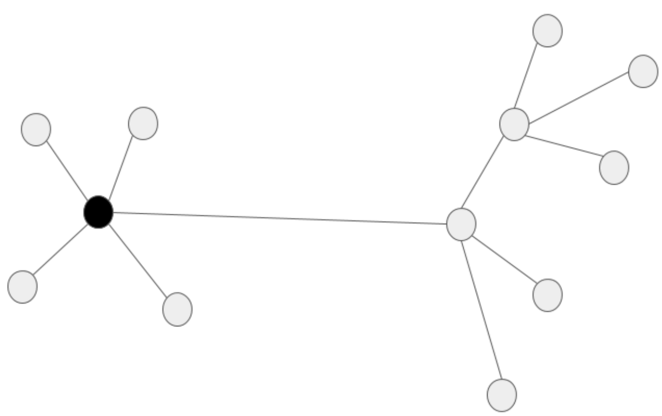

Figure 3: Epidemic Process

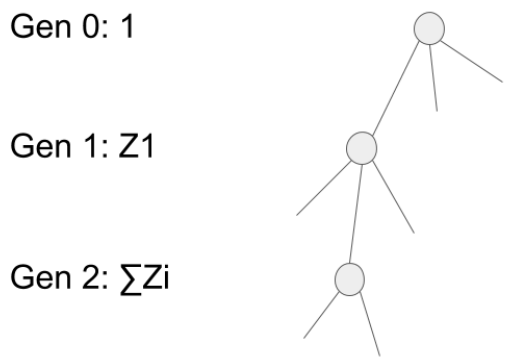

Figure 4: Branching Process

$Z_2 = \sum_{j=1}^{Z_i}(Pois(\lambda) - 1)$ Where we are ignoring overlapping neighbors. As λ fixed, $n \to \infty$.
Expectation of each term in the sum is $\lambda - 1$

$$\begin{aligned} E(Z_2) &= E(Z_1)(\lambda - 1) \\ &= \lambda(\lambda - 1) \\ \Rightarrow E(Z_1 + Z_2) &= \lambda^2 \end{aligned}$$

After k steps, $E(Z_k) = \lambda^k$
If $\lambda < 1$, λ^k tends to a finite number $(\frac{1}{1-\lambda})$
If $\lambda > 1$, λ^k tends to ∞

We have mapped the connected component onto a *branching process*.
Each object i produces Z_i branches to new objects, independently.

Solve for the probability of extinction within a finite time: must have finite total number of nodes in the tree.
"subcritical branching" $E(Z) < 1$
"supercritical branching" $E(Z) > 1$: will live forever with positive probability.
If $E(Z) = 1$, the tree will also die out but only after a very long time.
This can be used to model processes like nuclear reactions.

Takeaway: $\lambda > 1$ means we get a giant component that will keep going.

The number of nodes reached after k steps $\approx \mathcal{O}(\lambda^k)$. The size of the giant component $\approx \mathcal{O}(n)$. How big can you make k? The diameter of GC.

$$\begin{aligned} \mathcal{O}(\lambda^k) &= \mathcal{O}(n) \\ \Rightarrow klog(\lambda) &= log(n) \\ \Rightarrow k &= \mathcal{O}\left(\frac{logn}{log\lambda}\right) \end{aligned}$$

4 Small-world phenomenon

Diameter is $\mathcal{O}(logn)$ in a random graph as long as $\lambda > 1$
Cool: Giant component, small world property
Easy: Poisson degree distribution
Stats:

$$\begin{aligned} L(p) &= P(A; n, p) \\ &= \prod_{i<j} p^{A_{ij}}(1-p)^{1-A_{ij}} \\ \Rightarrow log(L(p)) &= \sum_{i<j} l(p) \\ &= \sum_{i<j} log(1-p) + A_{ij} log(\frac{p}{1-p}) \\ \Rightarrow \hat{p}_{MLE} &= \frac{\sum_{i<j} A_{ij}}{\binom{n}{2}} \end{aligned}$$

This is an exponential family so the sufficient statistic is the total number of edges. $\hat{p}_{MLE}$ is strongly consistent and efficient.

5 Problems

This is a horrible model of any real network know to science. There are 2 main issues:

1. Degree distributions are not Binomial or Poisson: much too light-tailed and unskewed
2. Hardly any triangles
   (a) Proportion of triples forming triangles $= p^3$
   (b) Probability we have a triangle given two nodes are connected to a third is p. In a random graph, this is $\frac{\lambda}{n-1}$
   (c) In lots of networks, triangles are very common
   (d) In some networks, like routers, they are rare

Lesson for model criticism: find things where

1. the model makes predictions
2. you didn't fit them to
3. check them against data

Random graph model: all edges are *independent* with *equal probabilities* and all nodes are *exchangeable.*

1. Dependent edges: stochastic block and latent-space, and exponential family models
2. Unequal probabilities: inhomogeneous random graphs; block models; p_1 models; configuration model; β model
3. Differentiating nodes: covariate-based link prediction; block model; some sorts of "degree correction"

Next time: simplest extensions