Veronica Ciocanel

Static Networks I - Reading Group 02/05/2015

I Structural Properties of Networks

- characterize any real-world networks, not just random graphs.

0. Intro $G = (\mathcal{N}, \mathcal{L})$ where $\mathcal{N}$ = nodes

$\mathcal{L}$ = ordered pairs of elements in $\mathcal{N}$

$|\mathcal{N}| = N \Rightarrow$ can have 0 to $\frac{N(N-1)}{2}$ edges.

Component of a graph : maximally connected induced subgraph

Giant component : component with size $O(N)$.

Matricial Representation :

- Adjacency matrix A : $N \times N$ square matrix w/

$$a_{ij} = \begin{cases} 1 & \text{when } l_{ij} \in \mathcal{L} \Leftrightarrow (i,j) \in \mathcal{L}. \\ 0 & \text{else.} \end{cases}$$

(symmetric for undirected graphs)

1. Degree Distributions

node $i \longrightarrow$ degree $k_i = \sum_{j \in \mathcal{N}} a_{ij}$ (from adjacency matrix)

For directed graphs, outgoing links $k_i^{out} = \sum_j a_{ij}$

ingoing —||— $k_i^{in} = \sum_j a_{ji}$

Degree dist'n $p(k)$ = prob. that a node chosen uniformly at random has degree k

= fraction of nodes in the graph having degree k.

Moments of dist'n

m-moment of $P(k)$: $\langle k^m \rangle = \sum_k k^m P(k)$.

$\langle k \rangle$ = mean degree of G

$\langle k^2 \rangle$ = fluctuations of the degree distribution.

Ex: Exponential dist'n : $p_k \sim e^{-k/K}$

Power law : $p_k \sim k^{-\alpha}$ $\rightarrow$ scale-free-network

(1st ex: Price's network of citations between scientific papers, w/ $\alpha = 3.04$)

2. Shortest path, Diameter

matrix D : d_{ij} = geodesic (shortest/optimal path) from node i to node j

$\text{Diam}(G)$ = diameter of graph $G = \max_{i,j \in N} \{d_{ij}\}$

Typical measure : average shortest path length / characteristic path length = mean of geodesic lengths over all couples of nodes.

$$\boxed{L = \frac{1}{N(N-1)} \sum_{\substack{i,j \in N \\ i \neq j}} d_{ij}}.$$

Issue : L diverges if there are disconnected components in the graph.

Alternative : harmonic mean of geodesic lengths (efficiency of G)

$$E = \frac{1}{N(N-1)} \sum_{\substack{i,j \in N \\ i \neq j}} \frac{1}{d_{ij}}.$$

3. Clustering / Transitivity $\rightarrow$ real-world network property - clear deviation from behavior of r. graph.

- vertex A connected to vertex B, B with C $\rightarrow$ higher prob A connected to C.
- heightened # of Δ's in the network.

a) Clustering Coefficient $C = \dfrac{3 \times \#\text{ of } \Delta\text{'s in the network}}{\#\text{ of connected triples of vertices}}$

Ex :  $C = 3 \cdot \frac{1}{8} = \frac{3}{8}$.

- measures how clique-like the friendship network is.

b) Local Clustering Coefficient $c_i = \dfrac{2 e_i}{k_i(k_i-1)} = \dfrac{\sum_{j,m} a_{ij} a_{jm} a_{mi}}{k_i(k_i-1)}$.

where e_i = # of edges in G_i (subgraph of neighbors of node i)

Then $C = \langle c \rangle = \frac{1}{N} \sum_{i \in N} c_i$.

C: easier to compute via a); ε: in b): use numerical methods; efficient algorithms are an area of research.

☆ It is suspected that for many types of networks the probability that a friend of your friend is also a friend should → to a nonzero limit as network gets large.

i.e. $c = O(1)$ as $n \to \infty$.

But for random graphs (we'll see): $c = O\left(\frac{1}{n}\right)$.

Clustering coefficient can be generalised to density of k-loops, etc.

4. Graph Spectra

A = adjacency matrix → eigenvalues form the spectrum of the graph.

G: undirected ⇒ A real and symmetric ⇒ real eigvals $\mu_1 \le \mu_2 \le \dots \le \mu_N$ and eigenvectors corresponding to distinct eigvals are orthogonal.

Perron-Frobenius: ∃ real μ_N s.t. $|\mu| \le \mu_N$ ∀ eigvals μ of A

$-\mu_N$ = spectral radius of $A := \rho(A) = \|A\|$.

where $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$.

Why important? Spectral eigvals and eigenvectors are closely related to topological features such as diameter, # of cycles, connectivity ...

Ex: · Thm $\rho(A) \le \mathrm{Diam}(G)$ = diameter = $\max\{d_{ij}\}$

· $(i,j)^{th}$ entry of A^k: # of walks of length k from node i to node j.

· Eigvals sum to 0 since $\mathrm{Tr}(A) = 0$.

etc

$\rightarrow$ Diam$(G) <$ # of distinct eigvals in a generic graph G.

Other important info on connectivity properties of G:

normal matrix $N = D^{-1}A$, D = diag. matrix; $D_{ii} = \sum_j a_{ij} = k_i$.

Laplacian matrix $\Lambda = D - A$ $\rightarrow$ symmetric positive semi-def. matrix.

(Kirchhoff matrix) $\rightarrow$ all λ's of Λ are real & non-neg.,

full set of n real, orthogonal eigenvectors.

$\rightarrow$ all rows of Λ sum to 0 $\rightarrow$ Λ admits

the lowest eigvalue $\lambda_1 = 0$, w/ eigvector $(1, 1, \ldots 1)$

Corollary: multiplicity of $\lambda_1 = 0$ is # of comps of G.

· Thms for λ_2 $\rightarrow$ the larger it is, the more difficult to cut G into pieces.

5. Small - World Effect.

Milgram experiment : degree 6 connectivity on average between any 2 nodes.

Can define small-world networks as:

- networks whose L = average shortest path length scales as $\log(n)$

Recall $L = \frac{1}{N(N-1)} \sum_{\substack{i,j \in N \\ i \neq j}} d_{ij}$ (mean of optimal paths)

$E = \frac{1}{N(N-1)} \sum_{\substack{i,j \in N \\ i \neq j}} \frac{1}{d_{ij}}$

or

- networks that have small value of L, like r. graphs ($\log n$) and a high clustering coefficient C.

II <u>Random graphs</u> (Particularly Erdos-Renyi r. graph).
↳ initially by Erdos & Renyi in 1959

E-R r. graph: <u>$G_{n,p}$</u> :- n nodes & probability p of connecting each pair of nodes.

- graphs w/ m edges appear w/ probability:

$$p^{m}(1-p)^{M-m}; \quad M = \frac{n(n-1)}{2} = \text{max \# of poss. edges.}$$

- So here have set of vertices $V_{n,p} = \{1, 2, \ldots n\}$
  - & introduce $\{\xi_{xy}\}_{1 \le x < y \le n}$: iid r.v's, Bernoulli(p)

$$\begin{cases} P(\xi_{xy} = 1) = p. \\ P(\xi_{xy} = 0) = 1-p \end{cases}$$

If $\xi_{xy} = 1$: $\exists$ edge between x & y.

- consider undirected graphs: $\xi_{xy} = \xi_{yx}$.

- \# of neighbors $\sim$ Bin $(n-1, p) \Rightarrow E(\text{\# neighbors}) = (n-1) \cdot p$.
  $= \text{average degree} = \langle k \rangle$ from before.

Large n $\Rightarrow$ $p_k \to$ Poisson(λ) ; $\lambda = np$ is more convenient to consider than $(n-1) \cdot p$.

This is why E-R random graphs are sometimes called <u>Poisson random graphs</u>.

<u>Note</u> Many properties of E-R random graphs come from the limit of large graph size $n \to \infty$, but while keeping the mean degree $\langle k \rangle = \lambda$ constant.

Reed-Frost epidemic on an E-R random graph (SIR)

at $t=0$; $\begin{cases} S_0 = \{2, \dots, n\} \\ I_0 = \{1\} \\ R_0 = \phi \end{cases}$

Update Rule: $R_{t+1} = R_t \cup I_t$ (infected → recovered in one timestep)

$I_{t+1} = \{y \in S_t \mid \xi_{xy} = 1 \text{ for some } x \in I_t\}$

$S_{t+1} = S_t \setminus I_{t+1}$ (ξ_{xy} Bernoulli in E-R)

$v \in V_{n,p}$: $C(v)$ = connected component in $G_{n,p}$ containing v.

(people that get infected by an infection starting at v)

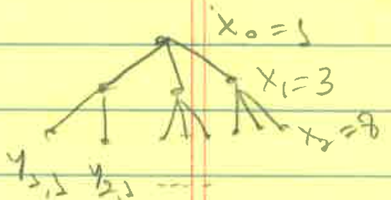
Interested in: asymptotic size of $C(v)$ as $n \to \infty$?

Setting above: $C(1)$ of interest. Note $C(1) = \bigcup_{t \geq 0} I_t$.

$\boxed{\lambda = np \text{ constant}}$ One approach: Construct a branching process (BP) approximation to I_t.

Key: Identify a BP $\{Z_t\}$ s.t. $|I_t| \leq Z_t$ and $\sum_{t=1}^{\infty} E(Z_t - |I_t|) \leq \frac{c}{n}$ if $\lambda < 1$.

What is a BP? a seq. of R.V.'s $\{X_k\}_{k \geq 0}$ s.t. $X_{n+1} = \sum_{i=1}^{X_n} Y_{n,i}$ (1)



$Y_{n,i}$ indep & dist'd according to $\{p_k\}_{k \geq 0}$.

X_n = # of nodes in gen' on n.

$Y_{n,i}$ = # of offsprings of node i in generation n

$\boxed{BP}$ – constructing Z_t is a bit technical, but it depends on the ξ_{xy}'s & is chosen s.t. (1) is satisfied.

Note: $\lambda < 1$: $E(Z_t - |I_t|) \leq \frac{c}{n}$

$\lambda > 1$: $E(Z_t - |I_t|) \leq \frac{c}{n} \cdot \lambda^{2t+2}$ → good approx. for initial times.

This approx'n using the BP Z_t is important because it is one way in which important connectivity properties of the E-R graph are proven.

→ Important because giant / largest component comes up in applications of any real networks, not just E-R r. graphs.

Thm Case 1 Subcritical regime $\boxed{\lambda < 1}$.

$\exists\ a = a(\lambda) = a(np) > 0$ s.t.

$$\lim_{n \to \infty} P(|C_1| \le a \log n) = 1.$$

i.e. Largest connected component is at most size $O(\log n)$, smaller than the whole population.

Case 2 Supercritical regime $\boxed{\lambda > 1}$.

Pext: extinction prob of BP with offspring distribution Poisson(λ)

$0 < P_{ext} < 1$!

Then $\exists\ a = a(\lambda) > 0$; $\forall\ \delta > 0$:

$$\lim_{n \to \infty} P\left(\left| \frac{|C_1|}{n} - (1 - P_{ext}) \right| < \delta,\ |C_2| \le a \log n \right) = 1\ \forall \delta.$$

i.e. Largest component has size a fixed fraction of n, all others are pockets of size $O(\log n)$.

Case 3 Critical regime $\boxed{\lambda = 1.}$

$$P(|C_1| = O(N^{2/3})) = 1\ a.s.$$

Note: v. similar to theory of phase transitions in material science

Proof

For case 1. Introduce an important way of exploring nodes via r. walk.

Pick arbitrary node $v \in \{1, 2, \dots, n\}$. $C(v)$ its connected component

A_k = set of "active" nodes in $C(v)$

B_k = set of "explored" nodes in $C(v)$.

Initially $(t=0)$: $\begin{cases} A_0 = \{v\} \\ B_0 = \phi. \end{cases}$

Iteration: (1) At step k, choose arbitrary $v_{k-1} \in A_{k-1}$.

(2) D_k = neighbors of v_{k-1}

(3) $A_k = A_{k-1} \cup D_k \setminus \{v_{k-1}\}$

(4) $B_k = B_{k-1} \cup \{v_{k-1}\}$

$|A_k| = |A_{k-1}| + \xi_k - 1.$

$T = \min\{k > 0 \mid |A_k| = 0\}.$ (i.e. done exploring)

$|A_T| = 1 + \sum_{i=1}^{T} \xi_i - T \Rightarrow 0 = 1 + \sum_{i=1}^{T} \xi_i - T$

$\Rightarrow T = 1 + \sum_{i=1}^{T} \xi_i.$

$T = |B_T| = C(v)$ all nodes that have been explored.

$$P(|C(v)| > k) = P(T > k) = P(|A_0| > 0, |A_1| > 0, \ldots, |A_k| > 0)$$

$$\leq P(|A_k| > 0) = \quad \text{(Bin. dist'n)}$$

$$= P(\mathrm{Bin}(n-1, 1-(1-p)^k) \geq k) \leq$$

$$\leq P(\mathrm{Bin}(n, kp) \geq k) \quad \left(1-(1-p)^k \leq kp\right)$$

$$= P(e^{\theta \cdot \mathrm{Bin}(n,kp)} \geq e^{k\theta}) \leq$$

$$\leq E[e^{\theta \cdot \mathrm{Bin}(n,kp)}] \cdot e^{-k\theta} \quad \text{(Markov inequality)}$$

$$P(|C(v)| > k) \leq (1 + kp(e^\theta - 1))^n e^{-k\theta} \leq$$

$$\leq e^{nkp(e^\theta - 1)} \cdot e^{-k\theta} = \quad (1 + x \leq e^x)$$

$$= e^{-k(\theta - \lambda(e^\theta - 1))} \quad (np = \lambda)$$

For $\lambda < 1$, small θ; $\theta - \lambda(e^\theta - 1) > 0$

$\rightarrow$ some choice of θ: $P(|C(v)| > k) \leq e^{-bk}$; $b > 0$.

$$P(|C_1| > b^{-1} \cdot \delta \cdot \log n) \leq e^{-b \cdot b^{-1} \delta \log n} =$$

$$= n^{-\delta}.$$

Choose $\delta > 0$; $n \to \infty \Rightarrow p \to 0$ ($\lambda = np = \text{cst.}$)

Let $a = \text{cst} = b^{-1}\delta > 0$.

$\Rightarrow P(|C_1| \leq a \log n) = 1$ as $n \to \infty$.

• <u>Connectivity in E-R r. graph</u> $G(n, p)$ (Continued)

$u \in V$; $deg(u) = \sum_{v \in V} \xi_{uv}$; $\xi_{uv} = \begin{cases} 1 \text{ if } (u,v) \in E \\ 0 \text{ else.} \end{cases}$

u isol'd if $deg(u) = 0$. <u>Isol'd iso</u>: # of isolated nodes.

$v \in V$; $\mathfrak{Z}_v = \begin{cases} 1 \text{ if } deg(v) = 0. \\ 0 \quad \text{else.} \end{cases}$

$$\mathfrak{Z}_v = \prod_{v \neq u} \mathbb{1}\{\xi_{uv} = 0\} = \prod_{v \neq u} (1 - \xi_{vu})$$

<u>Note</u> If one node is not isolated, the other node is automatically not isolated either, so r.v's not iid.

X = # of isolated nodes = $\sum_v \mathfrak{Z}_v \rightarrow$ not quite Poisson, but v. close

$\boxed{\text{Thm}}$ <u>Scaling</u> : $np = \log n + c$ (constant)

Then $d_{TV}(X, Poi(e^{-c})) \rightarrow 0$ as $n \rightarrow \infty$.

<u>Proof</u> : Uses Stein-Chen method.

So, P(no isol'd nodes) $= P(X=0) \doteq e^{-e^{-c}}$ as $n \rightarrow \infty$.

In fact, can show that P ($\exists$ conn'd comp of size $2, \ldots, k$) $\xrightarrow[n\to\infty]{} 0$.

So P ($G_{n,p}$ connected) $=$ P (no isol'd nodes) $= e^{-e^{-c}}$.

<u>Note</u> : for $\lambda = np =$ cst scaling, P($G_{n,p}$ connected) was 0 <u>(Thm p.7)</u>.

<u>Diameter of E-R r. graph</u> : it can be proven (technical) that the diameter has values in a small range of values around Diam $= \frac{\ln N}{\ln(pN)} = \frac{\ln N}{\ln \langle k \rangle}$ ☆

$\rightarrow$ Same for the average shortest path $L \sim O\left(\frac{\ln N}{\ln \langle k \rangle}\right)$.

<u>Why?</u> Average # of neighbors a distance l away is λ^l where $\lambda = (n-1) \cdot p \simeq np$ for $n \rightarrow \infty$.

To get to the whole network, $\lambda^L = N \Rightarrow L = \frac{\ln N}{\ln \lambda} = \frac{\ln N}{\ln \langle k \rangle}$.

Note : Since $L \sim O\left(\frac{\ln N}{\ln \langle k \rangle}\right)$, slower than $\log(N) \Longrightarrow$ the E-R r-graph reproduces the small-world scenario.

- Clustering coefficient

$$C = p = \frac{\langle k \rangle}{n} \; ; \; \text{Why? } \quad c_i = \frac{p\,k(k-1)/2}{k(k-1)/2} = p$$

edges among neighbors of a node

$$\& \quad C = \frac{\sum c_i}{n} = \frac{p \cdot n}{n} = p. \checkmark$$

$$\Rightarrow \quad C = p = \frac{\lambda}{n} \to 0 \text{ as } n \to \infty \quad (\text{Not realistic!})$$

- Degree distribution : Poisson → unrealistic, no correlation between degs of adjacent vertices, no community structure → inadequate to describe most observed dist'n (~ power laws).

There exist extensions of the E-R r.graph.

$\boxed{\text{But}}$: important main model b/c ideas of giant component, phase transitions are present in all the more sophisticated models.

Summary E-R r. graph

- Poisson degree distribution; not realistic
- clustering : $O\left(\frac{1}{n}\right) \xrightarrow[n \to \infty]{} 0$: not realistic
- no community structure
- characteristic path length : $O(\log n)$ → reproduces small-world phenomen.

III Generalized random graphs

- make E-R more realistic
- easiest property to change: non-Poisson degree distribution.

1) The Configuration Model

Def'd in the following way: specify degree dist'n p_k; p_k = fraction of vertices w/ degree k.

- Choose a degree sequence: n vals of the degrees k_i of vertices $i=\overline{1,n}$, from this dist'n.
- Give each vertex i in our graph k_i "stubs" or "spokes" sticking out of it → i.e. ends of edges-to-be.
- Then choose pairs of stubs at random from the network & connect them together → gives every top. of a graph w/ the given seq. w/ equal prob'ty

Main results – results on size of giant component can be proven here via powerful formalism of a generating function.

Probability generating fcn of $\bar{X}$ (takes values k w/ probability $p(k)$) is:

$$G(z) = E(z^{\bar{X}}) = \sum_{k=0}^{\infty} p(k) z^k .$$

Note: $G'(z) = \sum_{k=1}^{\infty} k\, p(k) z^{k-1}$

$$G'(1) = \sum_{k=1}^{\infty} k\, p(k) = E(\bar{X})$$

Back to config model: Degree of a vertex that we reach by following a randomly chosen edge is not p_k.

$\exists$ k edges that arrive at a vertex of deg k ⇒ k times as likely to arrive at that vertex than at one of degree 1.

⇒ Deg. dist'n of the vertex @ the end of a randomly chosen edge is $k \cdot p_k$. Many times int'd in the # of edges that leave a vertex (excess degree)

→ dist'n $q_k = \dfrac{(k+1)\cdot p_{k+1}}{\sum_k k\, p_k} = \dfrac{(k+1)\, p_{k+1}}{\langle k \rangle} = \dfrac{(k+1)\, p_{k+1}}{z}$

Recall: $q_k = \dfrac{(k+1)\,p_{k+1}}{\sum k p_k}$

Define 2 gen'ing fcns for dist's p_k & q_k:

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k, \quad G_1(x) = \sum_{k=0}^{\infty} q_k x^k.$$

Note: $G_1(x) = \dfrac{G_0'(x)}{z}$. , $z = \langle k \rangle = \sum k p_k$.

- Gen'ing fcn $H_1(x)$ for the total # of vertices reachable by following an edge:

$$\boxed{H_1(x) = x\, G_1(H_1(x))}$$

Won't prove, but here's the intuition:

– when following an edge, we find at least a vertex at the other end (x); + some other clusters of vertices (repr'd by H_1) reachable by following other edges attached to that one vertex.

excess degree $\sim q_k \Rightarrow G_1(x)$

- Gen'ing fcn $H_0(x)$ = total # of vertices reachable from a randomly chosen vertex: $\boxed{H_0(x) = x\, G_0(H_1(x))}$

$\hookrightarrow$ idea of giant comp

- Mean component size in the region of no giant component is:

$$\langle s \rangle = \underbrace{H_0'(1)}_{\text{exp value}} = 1 + \frac{G_0'(1)}{1 - G_1'(1)} = 1 + \frac{z_1^2}{z_1 - z_2} \quad ★$$

where $z_1 = z = \langle k \rangle = G_0'(1)$.

$z_2 = \langle k^2 \rangle - \langle k \rangle = G_0'(1) \cdot G_1'(1)$.

- Divergence in ★ when $z_1 = z_2$, i.e. when $G_1'(1) = 1$, i.e when $\boxed{\sum_k k(k-2)\,p_k = 0.}$ ← critical cond'n

← phase transition at which a giant comp. appears..

i.e. $\sum > 0 \Rightarrow$ giant comp. a.s. (occupies a fraction of the graph).

$\sum < 0 \Rightarrow$ largest comp. is $O(\log N)$

more extensions: – directed graphs, bipartite graphs w/ 2 types of nodes.

– still use generating fcn framework.

References used :

1. M. E. J. Newman, "The Structure and Function of Complex Networks", SIAM REVIEW, 45, 2003 (167-256)

2. S. Boccaletti et al, "Complex networks: Structure and dynamics", Physic Reports, 424, 2006 (175-308)

3. M. Draief and L. Massoulie, "Epidemics and Rumours in Complex Networks", Cambridge University Press, 2010.

4. D. J. Watts, "Small Worlds: The Dynamics of Networks between order and randomness", Princeton University Press, 2003.