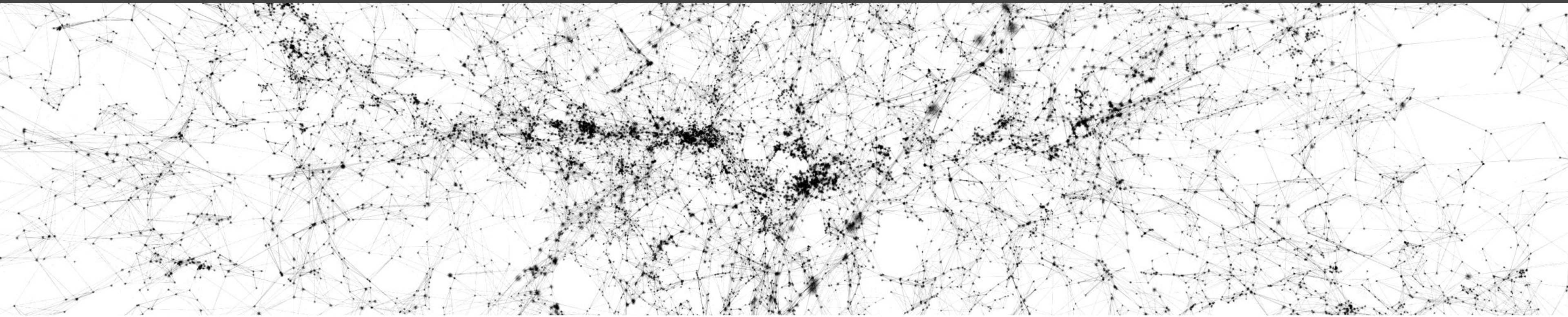
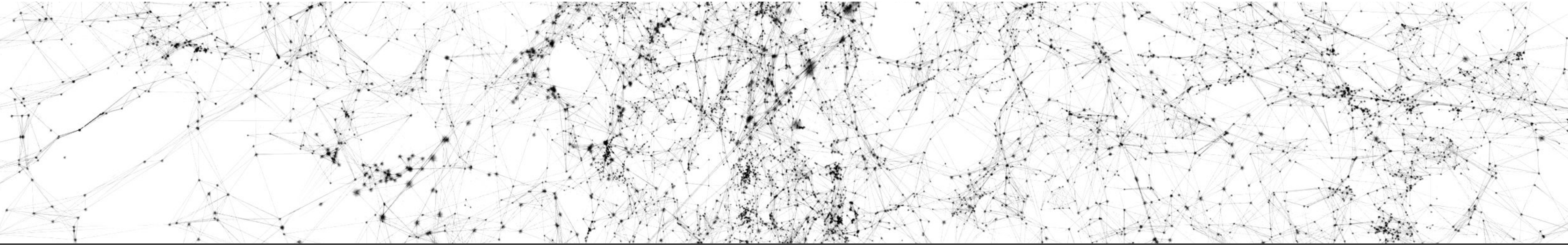


Models

Analysis of complex interconnected data



Quick Notes

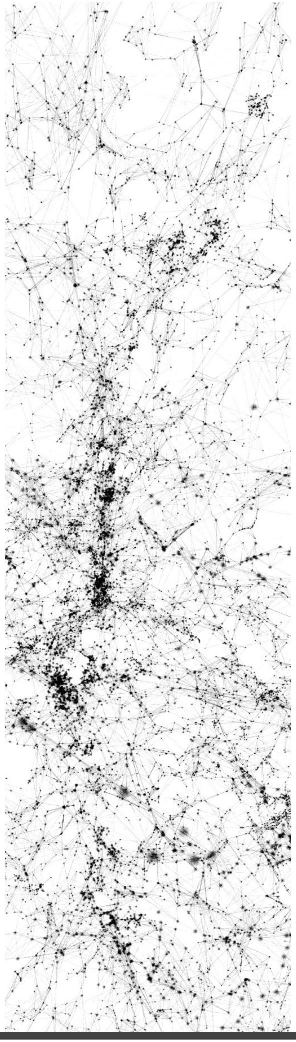
- Reminder, first assignment due in a week
    - http://www.reirab.com/Teaching/NS20/Assignment_1.pdf
    - **Any questions for the assignment?**
    - Submit single entry as a Group in Mycourses
- Use slack for easier communications
    - Let me know if you didn't get an invite
- Anyone new in the class?

Deadlines

- assignment 1 due on Sep. 20th
- assignment 2 due on Oct. 4th
- assignment 3 due on Oct. 18th
- project proposal slides due on Oct. 25th
- project proposal due on Nov. 1th
- Reviews (first round) due on Nov. 8th
- project progress report due on Nov. 22nd
- Reviews (second round) due on Nov. 29th
- project final report slides due on Dec. 1st
- project final report due on Dec. 6th
- Reviews (third round) due on Dec. 13th
- project revised report and rebuttal due on Dec. 20th
- note: dates are tentative, please check them for the updated deadlines

Outline

- **Patterns Quick recap**
- Models
  - ER model
  - BA model
  - SBM
  - Configuration model
  - FF model
  - Kronecker graph model
  - Log likelihood fitting to observed graphs

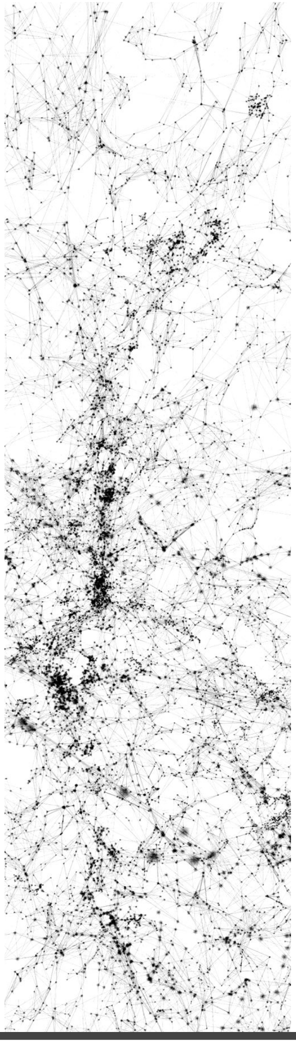


Patterns: quick recap

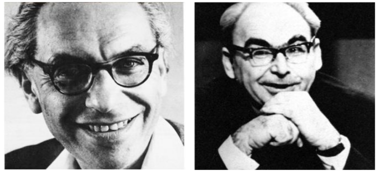
- Sparsity Pattern
  - mean degree << N-1 (or E << Emax)
- Scale Free Pattern
  - heavy tailed degree distribution
- Assortativity Pattern
  - positive or negative correlation between degree of connecting nodes
- Transitivity Pattern
  - high ratio of closed triangles (clustering coefficient)
- Small world Pattern
  - small average shortest path
- Connectivity & eigenvalues of Laplacian matrix
  - number of zero eigenvalues gives the number of connected components

Outline

- Patterns Quick recap
- Models
  - **ER model**
  - SBM
  - Configuration model
  - AB model
  - FF model
  - Kronecker graph model
  - Fitting to observed graphs

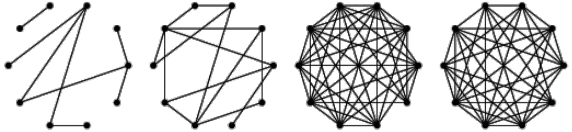


Erdös-Rényi Model (ER)

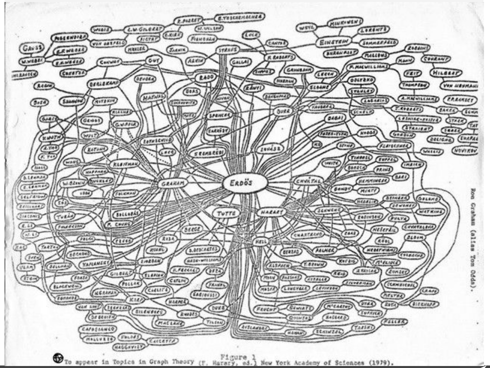


Paul Erdős (1913-1996) **Alfréd Rényi** (1921-1970)

- Introduced in 1960
- Basis of random graph theory
- Simple model that results in **small-world** graphs
- Parameters: ER(n, p) or ER(n, m)
  - n: number of nodes
  - p: probability of an edge between any two nodes
  - m: number of edges
- Generation: all edges are equally likely so toss n(n-1)/2 coins



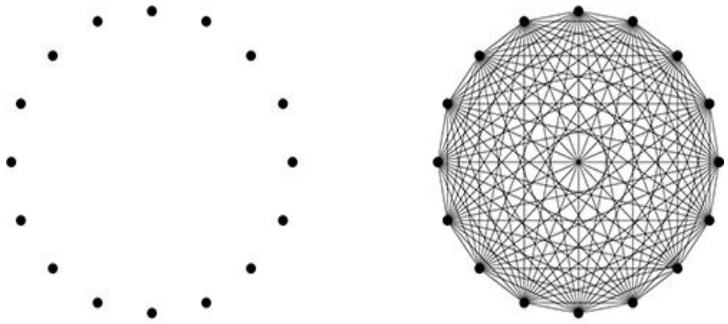
Side note:
What is Erdős number?



Erdös-Rényi Model (ER)

For **p = 0** we have ‹k› = 0, hence all nodes are isolated. Therefore the largest component has size NG = 1 and NG/N→0 for large N.

For **p = 1** we have ‹k›= N-1, hence the network is a complete graph and all nodes belong to a single component. Therefore NG = N and NG/N = 1



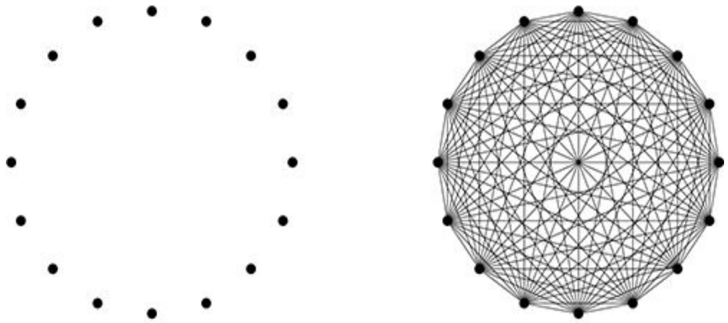
(a) $p = 0$ (b) $p = 1$

At which p we see a giant component?
(NG/N is finite; NG grows in proportion to N)

Erdös-Rényi Model (ER)

For **p = 0** we have ‹k› = 0, hence all nodes are isolated. Therefore the largest component has size NG = 1 and NG/N→0 for large N.

For **p = 1** we have ‹k›= N-1, hence the network is a complete graph and all nodes belong to a single component. Therefore NG = N and NG/N = 1



(a) $p = 0$ (b) $p = 1$

At which p we see a giant component?
(NG/N is finite; NG grows in proportion to N)

$p_c = 1/\ N{-}1 \approx 1/N$

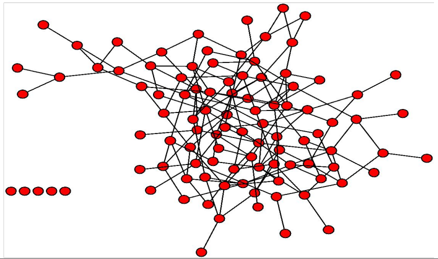
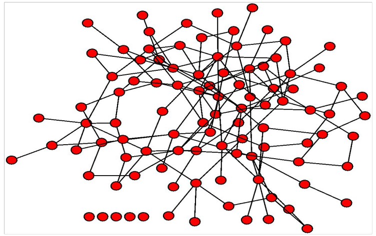
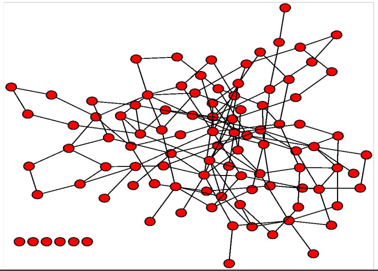
Erdös-Rényi Model (ER): properties

We can derive many properties of ER analytically

derive the expected value of a property as $\langle x \rangle = \sum_G x(G) \times \Pr(G)$

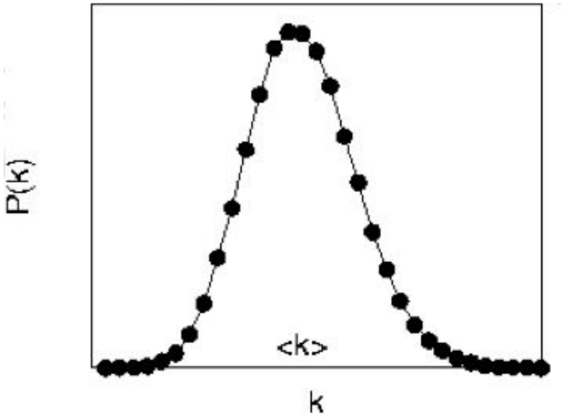
Where probability of observing a given graph is $P(G) = \dfrac{1}{\binom{\binom{n}{2}}{m}}$

or $P(G) = p^m (1-p)^{\binom{n}{2}-m}$

p=0.03, N=100

Erdös-Rényi Model (ER): degree distribution



$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k}$$

Select k nodes from N-1

probability of having *k* edges

probability of missing N-1-k edges

$$<k> = p(N-1) \qquad \sigma_k^2 = p(1-p)(N-1)$$

For large *N* and small *k*, we can use the following approximations:

Poisson degree distribution

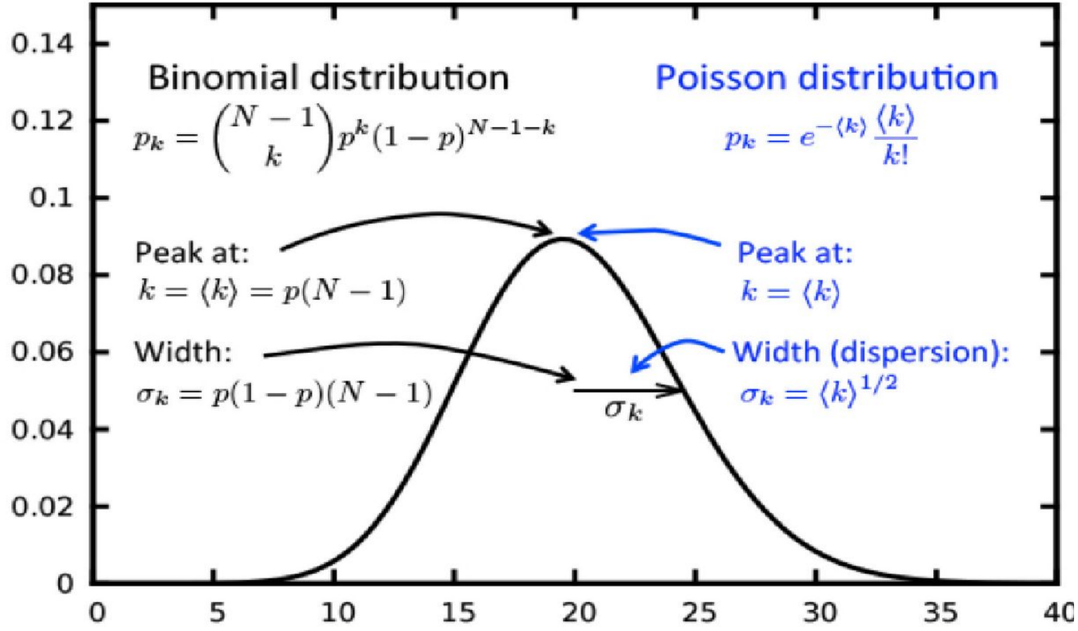
$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$

Erdös-Rényi Model (ER): degree distribution

Exact Result
-binomial distribution-

Large N limit
-Poisson distribution-

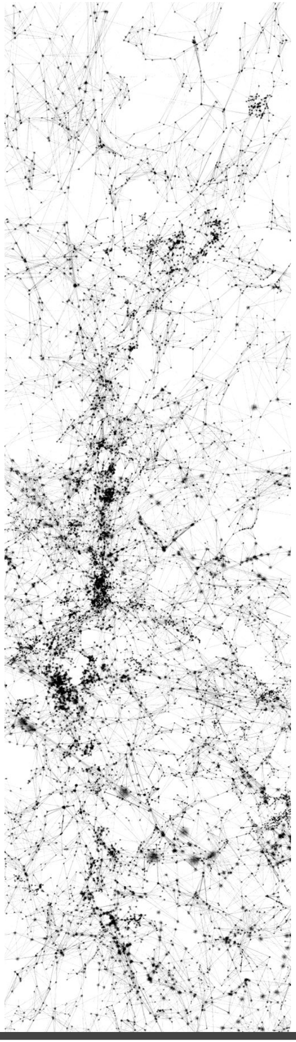
PDF



[From Barbasi's slides]

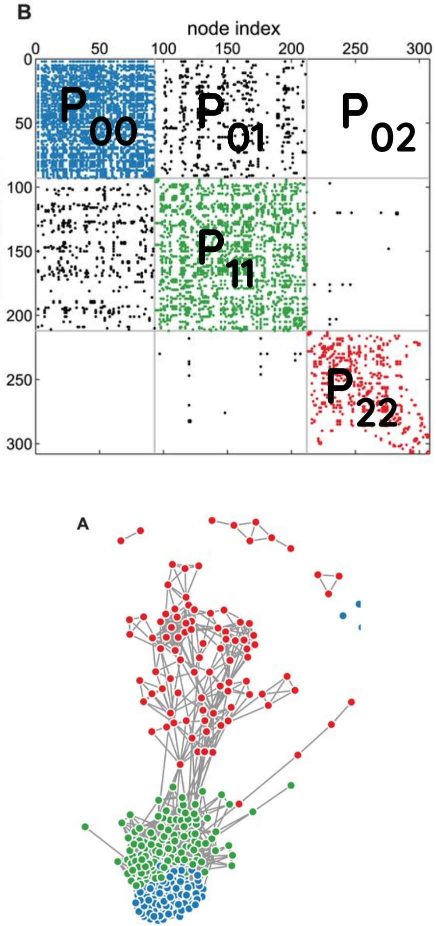
Outline

- Patterns Quick recap
- **Models**
  - ER model
  - **SBM**
  - Configuration model
  - AB model
  - FF model
  - Kronecker graph model
  - Fitting to observed graphs



Stochastic Block Models (SBM)

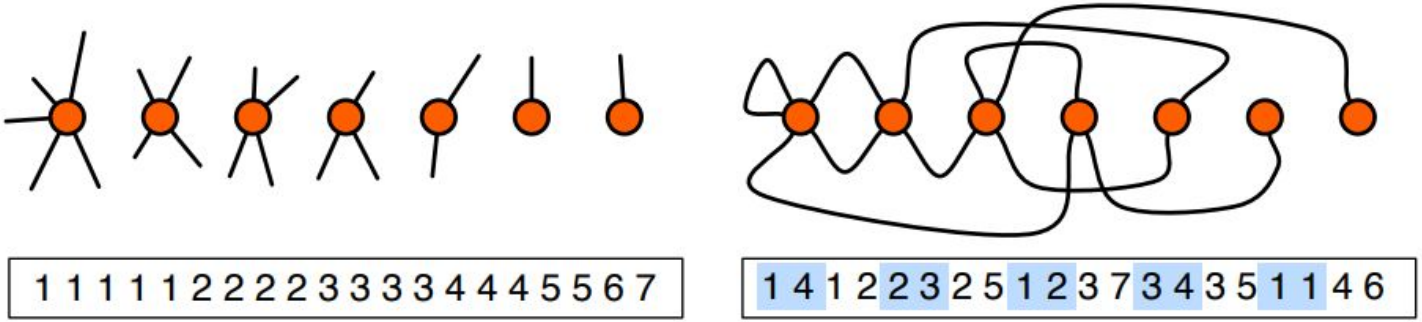
- Generalized ER to created block-structured graphs
- Parameters:
    - n: number of nodes
    - k^2 probabilities: $P_{k \times k}$
    - k disjoint sets that divide the n nodes
- Generation: create (within, between) edges similar to ER for the corresponding subsets of nodes with the corresponding probability



Configuration model

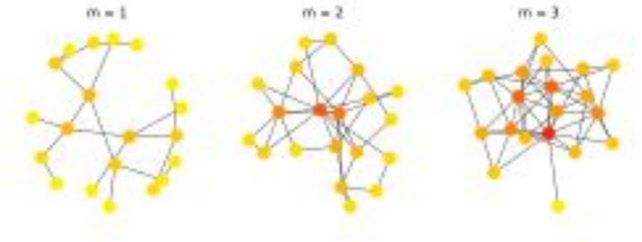
- By Mark Newman, generalizing ER to specific degree distribution
- Parameters: degree sequence (can be easily sampled from any distribution)
- Generation: assign slots, randomly connect them
- Serves as a null model for community detection
  - edges are distributed randomly given the degrees are fixed
  - communities that are not formed randomly should deviate from this
  - more on this later

$$p_{ij} = \frac{k_i k_j}{2m-1} \simeq \frac{k_i k_j}{2m},$$



Albert Barabasi Model (AB)

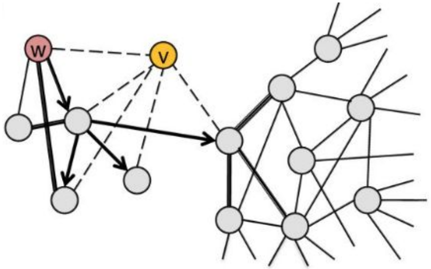
- Introduced in 1999, a.k.a Barabási–Albert (BA) model
- Uses preferential attachment which gives scale-free graphs
- Parameters: BA (n,m)
  - n: number of nodes
  - m: average degree



- Generation:
  - add one node at the time, add **m** connections per new node if possible
  - the probability of forming a connection to an existing node is proportional to its degree, i.e. $\mathbf{p(i) = d_i / \Sigma_j d_j}$

Forest Fire model (FF)

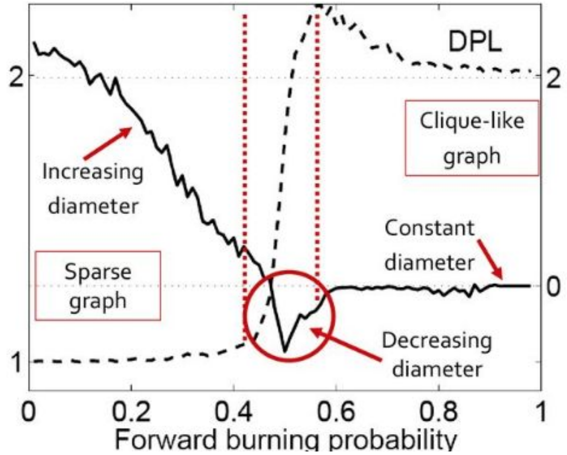
- By Leskovec, 2005
- To follow patterns observed in real-world graphs
  - denser over time, the average degree increasing, and the diameter decreasing
- Parameters: n, p and rp
  - n: number of nodes
  - p: forward burning probability
  - r : backward burning probability



- Generation:
  - add a node at a time, connect the node to an ambassador, chosen uniformly at random
  - then, the new node recursively forms a random number of connections with the neighbours of every node it connects to –outlinks to specific number of inlink and outlink neighbours, drawn from geometric distributions with means of p/(1 –p) and r/(1 –r) respectively

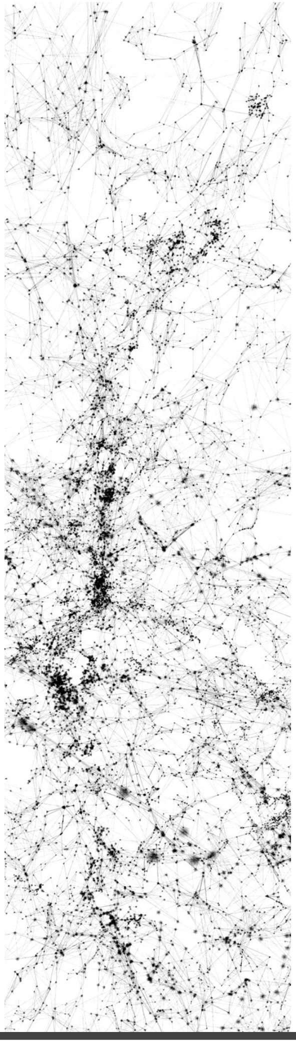
Forest Fire model (FF)

- Heavy-tailed degree distribution
  - rich get richer: older nodes have more chances to become ambassadors
- Densifies
  - newly entered node has more links to neighbours close to its ambassador
- Can result in shrinking diameter
  - Which is observed in real-world networks



Outline

- Patterns Quick recap
- Models
  - ER model
  - SBM
  - Configuration model
  - AB model
  - FF model
  - **Kronecker graph model**
  - Fitting to observed graphs



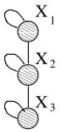
Kronecker graph model

By Leskovec, 2010

Kronecker product of matrices

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \doteq \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}.$$

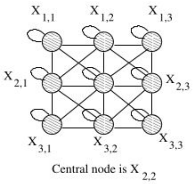
Consider a small initiator matrix, use kronecker products to get the adjacency matrix as $K_k = \underbrace{K_1 \otimes K_1 \otimes \dots K_1}_{k\ times} = K_{k-1} \otimes K_1$



(a) Graph K_1

1	1	0
1	1	1
0	1	1

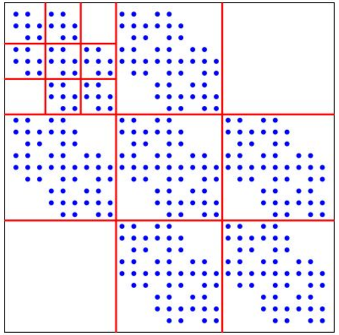
(d) Adjacency matrix of K_1



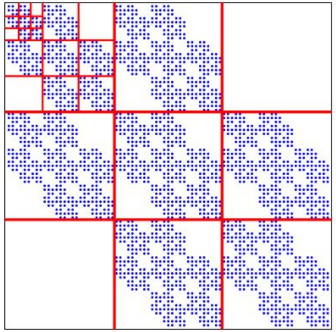
(c) Graph $K_2 = K_1 \otimes K_1$

K_1	K_1	0
K_1	K_1	K_1
0	K_1	K_1

(e) Adjacency matrix of $K_2 = K_1 \otimes K_1$



(a) K_3 adjacency matrix (27×27)



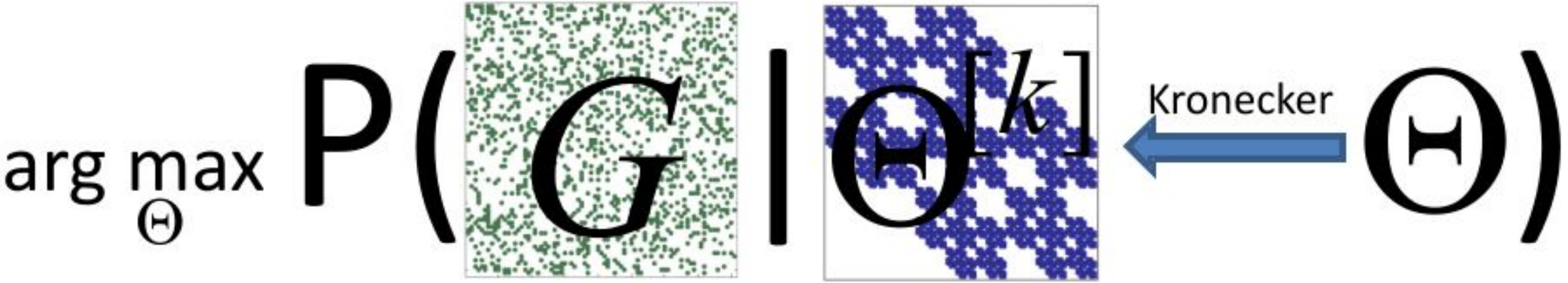
(b) K_4 adjacency matrix (81×81)

Kronecker graph model

Stochastic Kronecker graph, initiator matix is probabilities and edges are drawn for the final graph with the corresponding probabilities

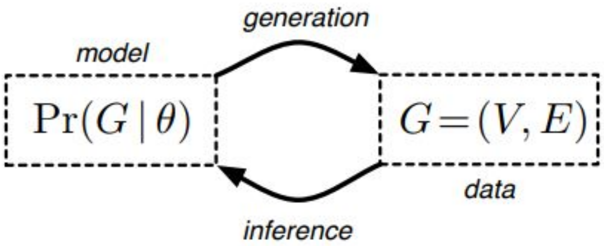
If all probabilities are equal in the initial matrix, this becomes equivalent to ER

the initiator matrix can be set based on real-world data to sample similar graphs, by searching over what matrix is more likely to give the observed



Fitting to observed graphs

- Option 1:
  - Measure and plot different characteristics of the observed graphs
  - Tune the parameters of the model to find a close enough fit to the observed patterns
- Option 2:
  - Define the likelihood of observing a graph, usually assuming edges are independent
  - Use maximum likelihood to find the model parameters



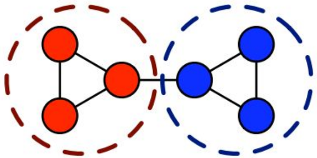
Fitting the SBM to data

Likelihood of G given Probability matrix M and partitioning z

$$\begin{aligned}\mathcal{L}(G \mid M, z) &= \prod_{i,j} \Pr(i \to j \mid M, z) \\ &= \prod_{(i,j)\in E} \Pr(i \to j \mid M, z) \prod_{(i,j)\notin E} 1 - \Pr(i \to j \mid M, z) \\ &= \prod_{(i,j)\in E} M_{z_i,z_j} \prod_{(i,j)\notin E} \left(1 - M_{z_i,z_j}\right) \quad ,\end{aligned}$$

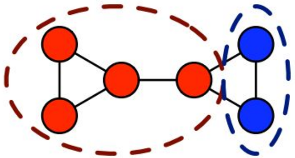
See how to derive the log likelihood here::

http://tuvalu.santafe.edu/~aaronc/courses/5352/csci5352_2017_L6.pdf



$\mathcal{L}_{\text{good}} = 0.043304\ldots$
$\ln \mathcal{L}_{\text{good}} = -3.1395\ldots$

M_{good}	red	blue
red	3/3	1/9
blue	1/9	3/3



$\mathcal{L}_{\text{bad}} = 0.000244\ldots$
$\ln \mathcal{L}_{\text{bad}} = -8.3178\ldots$

M_{bad}	red	blue
red	4/6	2/8
blue	2/8	1/1