Aalto University
School of Science
and Technology

CS-E5745
Mathematical Methods for Network Science

Mikko Kivelä

Department of Computer Science
Aalto University, School of Science
mikko.kivela@aalto.fi

January 25, 2024

Generating functions and their use in networks

- Learning goals this week:
  - Recap of probability generating functions from last week
  - Learn how to use PGFs to solve component size distributions in networks
- We will be following the Section 13 in Newman: *Networks, An Introduction*

PGFs from last week

- Definition:

$$g(z) = p(0) + p(1)z + p(2)z^2 \cdots = \sum_{k=0}^{\infty} p(k)z^k \qquad (1)$$

- $p(k)$ can be extracted through derivation:

$$p(k) = \left[\frac{1}{k!}\frac{d^k}{dz^k}g(z)\right]_{z=0} \qquad (2)$$

- Moments can also be calculated through derivation:

$$\langle X^m \rangle = \left[z\frac{d}{dz}\dots z\frac{d}{dz}g(z)\right]_{z=1} = \left[(z\frac{d}{dz})^m g(z)\right]_{z=1} \qquad (3)$$

PGFs from last week

- Sums of independent RVs

$$g_{X_1+X_2}(z) = g_{X_1}(z) * g_{X_2}(z) \tag{4}$$

$$g_{\sum_{i=1}^{N} X_i}(z) = [g_{X_i}(z)]^N \tag{5}$$

$$g_{X_1+c}(z) = g_{X_1}(z) * z^c \tag{6}$$

- If N is also a RV in $S = \sum_{i=1}^{N} X_i$:

$$g_S(z) = g_N(g_{X_i}(z)) \tag{7}$$

Notation for networks (from Newman)

- For the degree distribution $p(k)$:

$$g_0(z) = \sum_{k=0}^{\infty} p(k) z^k$$

- For the excess degree distribution $q(k)$:

$$g_1(z) = \sum_{k=0}^{\infty} q(k) z^k$$

- These two are related:

$$g_1(z) = \frac{1}{\langle k \rangle} \frac{d}{dz} g_0(z)$$

Solving the Galton-Watson process for networks

- Last week we derived the equations for Galton-Watson processes

$$
\begin{aligned}
g_{K_1}(z) &= g_0(z) \\
g_{K_d}(z) &= g_{K_{d-1}}(g_1(z))
\end{aligned}
$$

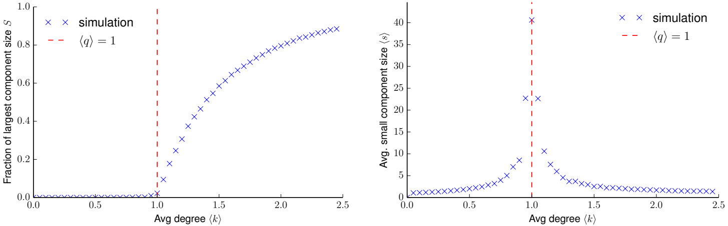
- Solution for the expected value:

$$
\langle K_d \rangle = \langle q \rangle^{d-1} \langle k \rangle = \left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right)^{d-1} \langle k \rangle
$$

- A criterion for the percolation threshold!

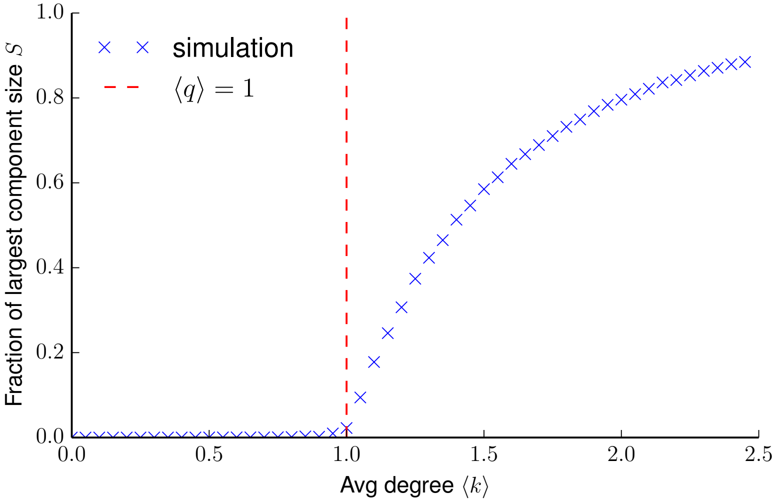
Solving the Galton-Watson process for networks

- We have the percolation threshold, but we are still missing
  - Shape of the relative giant size curve ($\sum_d K_d$ in the giant)
  - Size distributions of small components ($\sum_d K_d$ when not in the giant)
- Note: GW process approximation for networks only works when d is small



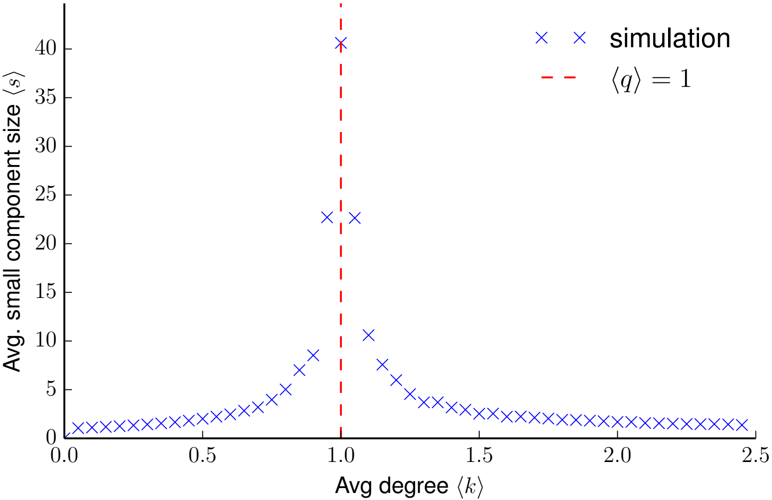
Giant component size in ER networks

- Largest component size in ER networks with $N = 10^5$



Expected small component size in ER networks

- Expected size of components other than the largest component in ER networks with $N = 10^5$



Uniqueness of the giant component

- Assume that there are two large components in ER networks with $S_1, S_2 > 0$ proportion of all nodes in them
- Number of nodes: $S_1 N$ and $S_2 N$
- Possible edges between the components: $S_1 N S_2 N = S_1 S_2 N^2$
- Probability that no two pairs are connected: $q = (1-p)^{S_1 S_2 N^2} = (1 - \frac{\langle k \rangle}{N-1})^{S_1 S_2 N^2}$
- In "thermodynamic limit" ($N \to \infty$, s.t. $\langle k \rangle$ constant):

$$q = q_0 e^{-\langle k \rangle S_1 S_2 N} \to 0$$

Solving for giant component size

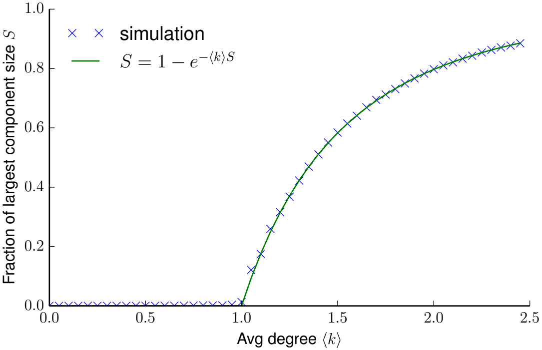
- Idea: Instead of the GW process, we write down "self-consistency equations"
  - Write down the probability random node not being in giant u as a function of u
  - The fixed point where $u = f(u)$ gives us u
  - The probability of random node in giant $S = 1 - u$

Solving for giant component size for ER networks

- Definition: u is the probability that (uniformly) randomly selected node doesn't belong to the giant component
  - Note that for Poisson degree distributions: $p(k) = q(k)$
- If node i doesn't belong to the giant then for every other node j:
  - There is no connection between i and j (probability $1 - p$), or
  - There is connection, but j doesn't belong to the giant (probability pu)
- In total the following must hold $u = [(1 - p) + pu]^{N-1}$
  - $u = e^{-\langle k \rangle (1-u)}$, when $N \to \infty$

Solving for giant component size for ER networks

- Shape of the relative giant size curve for ER network (with $N = 10^5$ nodes)



Solving for giant component size for configuration model (1/3)

- Definition: u is the probability that following an edge (and removing it) doesn't lead to the giant component
- If node i doesn't belong to the giant then none of the neighbors j belong to the giant
  - The number of new neighbors the node i has k is distributed as the excess degree $q(k)$
  - Probability that none of the k neighbors is in the giant is u^k
- In total the following must hold $u = \sum_{k=0}^{\infty} q(k) u^k$

Solving for giant component size for configuration model (2/3)

- In total the following must hold $u = \sum_{k=0}^{\infty} q(k) u^k$
  - Using the definition of the PGF [Eq. (1)]: $u = g_1(u)$
  - Probability of uniformly randomly selected node not being in the giant is $\sum_{k=0}^{\infty} p(k) u^k = g_0(u)$
- Returns the previous result for ER networks: $u = e^{-\langle k \rangle (1-u)}$

Solving for giant component size for configuration model (3/3)

- We can also solve the component size distributions of non-giant components
- Write equations for component size distributions using component size distributions
  - Leads to nice equations when done using PGF's

Solving for component size distributions

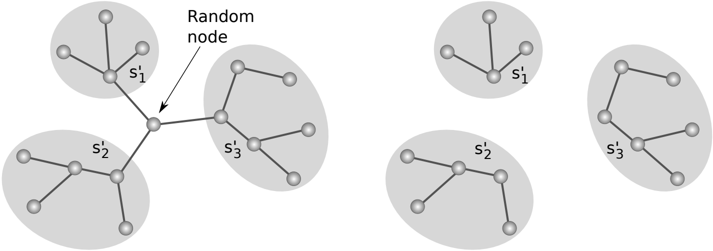
- Separate giant component size and small component size distribution
- S: Probability that uniformly randomly selected node belongs to the giant component
- RV s: size of non-giant component where uniformly randomly selected node belongs to
- Notation from Newman:
  - $P(s = k) = \pi_k$
  - $h_0(z) = g_s(z)$
- Sum of the above probabilities is one: $h_0(1) + S = 1$

Solving for component size distributions

- RV s': Follow a link and remove it. s' is the size of the non-giant component after that link is removed ("excess small component size")
- Notation from Newman:
  - $P(s' = k) = \rho_k$
  - $h_1(z) = g_{s'}(z)$

Solving for component size distributions

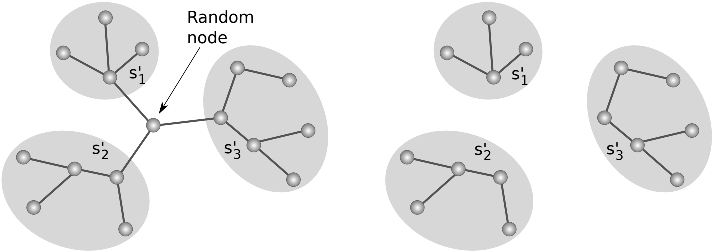
- $s = 1 + s'_1 + s'_2 + s'_3$



Solving for component size distributions

- RVs s, s', and the first neighborhood size k_1 are related

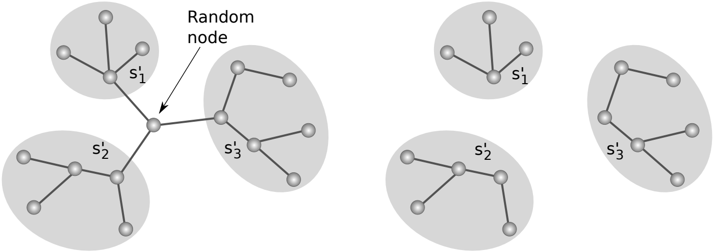
$$s = 1 + \sum_{k=0}^{K_1} s'$$



Solving for component size distributions

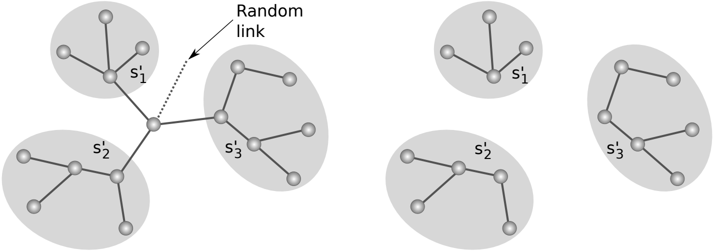
- Notation from Newman: $h_0(z) = g_s(z)$ and $h_1(z) = g_{s'}(z)$
- Using GF properties

$$
\begin{aligned}
s &= 1 + \sum_{k=0}^{K_1} s' \iff \\
h_0(z) &= z g_0(h_1(z))
\end{aligned}
$$



Solving for component size distributions

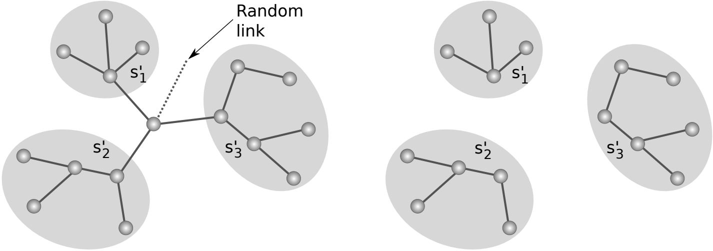
- $s' = 1 + s'_1 + s'_2 + s'_3$



Solving for component size distributions

- s' (h_1) and the excess degree $K_{1,i}$ (g_1) are related

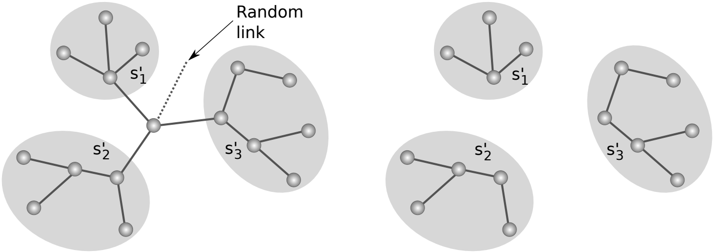
$$s' = 1 + \sum_{k=0}^{K_{1,i}} s'$$



Solving for component size distributions

- Using GF properties

$$
\begin{aligned}
s' &= 1 + \sum_{k=0}^{K_{1,i}} s' \iff \\
h_1(z) &= z g_1(h_1(z))
\end{aligned}
$$



Solving for component size distributions

- In total we have

$$h_0(z) = zg_0(h_1(z)) \tag{8}$$
$$h_1(z) = zg_1(h_1(z)) \tag{9}$$

- Solving $h_0(z)$ possible but not easy
- Solving $h_0(1)$ easier
  - Remember: $h_0(1) + S = 1$

Solving for component size distributions

- Using $h_0(1) + S = 1$ and Eq. (8):

$$S = 1 - h_0(1) = 1 - g_0(h_1(1))$$

- The value $h_0(1)$ can be solved from Eq. (9):

$$h_1(1) = g_1(h_1(1))$$

- Using Newmans notation: $u = h_1(1)$ (=probability that following a link does not lead to the giant component):

$$S = 1 - g_0(u) \quad (10)$$

$$u = g_1(u) \quad (11)$$

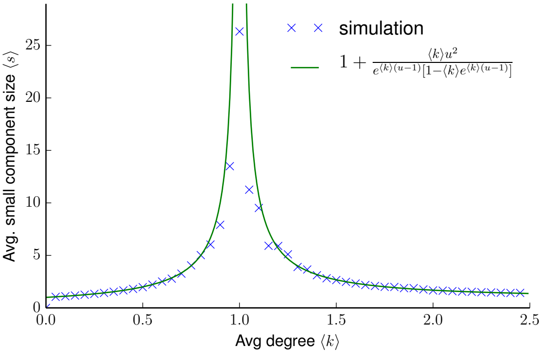
Solving for component size distributions

- With some (home)work one can use previous equations to solve for the expected small component size

$$\langle s \rangle = 1 + \frac{g_0'(1)u^2}{g_0(u)[1 - g_1'(u)]} \tag{12}$$

Solving for giant component size for ER networks

- Shape of the expected size of the small components in ER network (with $N = 10^5$ nodes)



Solving for component size distributions

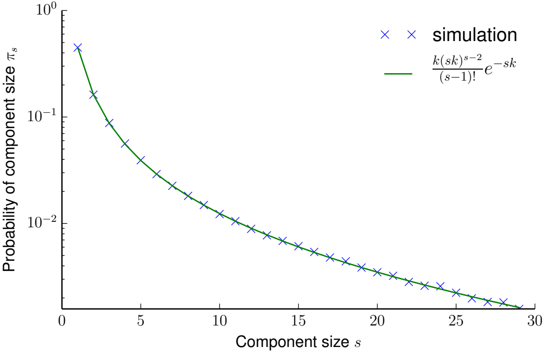
- With even more work one can find a formula for the whole small component size distribution!
  - See Newman pp. 468–469

$$\pi_s = \frac{\langle k \rangle}{(s-1)!} \left[\frac{d^{s-2}}{dz^{s-2}} \left[g_1(z) \right]^s \right]_{z=0} \tag{13}$$

- Note that the formula works only for $s > 1$ and $\pi_1 = p_0$

Solving for component size distribution for ER networks

- Component sizes in ER network (with $N = 10^7$ nodes and $\langle k \rangle = 0.8$)



Summary on component sizes

- We can now solve:
  - Threshold for connectivity (percolation threshold)
  - The (expected) number of nodes d steps away in BFS process
  - The size of the giant component
  - Component size distribution
- ... with some limitations:
  - Equations are derived for large configuration models (but might give good approximations for real networks)
  - We can write the equations but they might not have closed form solutions
  - Some equations do not work for the critical point (one can also solve for this separately)

Extensions

- Similar ideas (PGF's, self-consistency equations ...) can be used for many extensions
  - Networks with degree correlations
  - Networks with triangles
  - Mutual connectivity
  - ...
- Some of these extensions are done in the possible projects

Percolation problems

- In percolation problems one removes nodes or edges can calculates the connected components
- Equivalent to finding component size distributions with altered degree distributions! (if done for configuration model or ER networks)
  - If K and K' are RV's of the original and the modified degree distribution, then percolation methods are given by $P(K' = k|K = m)$
  - $P(K' = k) = \sum_{m=0}^{\infty} P(K' = k|K = m) * P(K = m)$
  - $g_{K'}(z) = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} P(K' = k|K = m) * P(K = m) z^k$
  - Some projects might include percolation done in this way