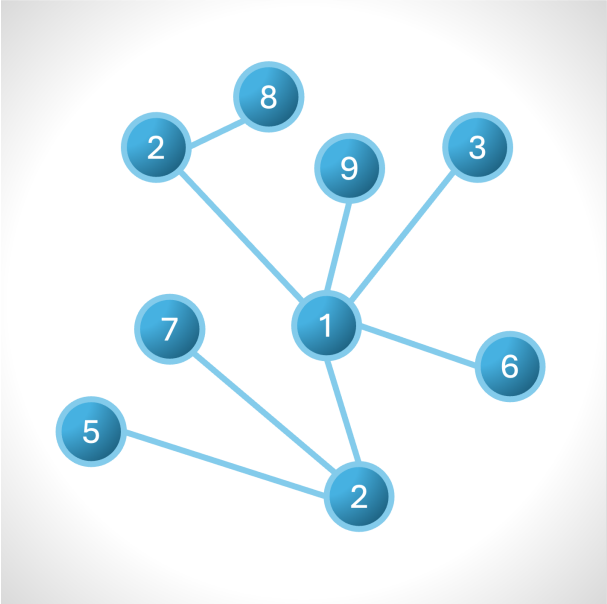
Growing Network Models

Sun Woo P. Kim
CID: 00939015

19th February, 2018
Edited 19th October 2024



Abstract

Three growing network models (GNMs) are studied numerically using Monte Carlo sampling and analytically using mean-field approximation. The first is the Barabási-Albert (BA)/PurePref model, where new nodes are connected preferentially connected to existing nodes with many preexisting links. The second is the PureRand model, where there is no preference. The third is the Mixed model, where node choice probability is chosen between the two models. Apart from rederivation of known results, our original contribution is as follows. We study the node degree evolution directly instead of the steady state distribution, to find an analytic form more accurate with number of edges added per timestep, m. We confirm the results with numerics.

Contents

1 Overview

This is a modified project report for the Masters-level course Complexity and Networks that I took while I was at Imperial College London. Most of the results are known results, but I had some new results using the methods of Sec. 1.4.2 to find a better analytic form for scaling with m, the number of new links added to the new node at each timestep. The resulting expressions are given by Eqs. (23), (35). We expect this to work well for $m \gg 1$, which we can confirm numerically for Figs. 8, 16. Note that the report is very terse, and contain sentences without articles due to the strict word limit that I had.

1.1 Growing Network Models

Growing Network Models (GNMs) are networks where new nodes are added and connected to existing nodes with each timestep. Specifically, the Barabási-Albert/Pure Preferential (BA/PurePref) Model attempt to describe citations in academic journals; it displays fat-tailed degree distributions $p(k)$ observed in data.

A simple graph $\mathcal{G}$ is specified by a growing set of nodes, each node with an index, $\mathcal{V} = \{1, 2, ..., |\mathcal{V}|\}$, set of edges, $\mathcal{E} = \{(i, j) \mid i, j \in \mathcal{V}\}$, a set of unordered pairs.

The dynamics of a GNM is only determined by the degrees; therefore the networks are random graphs (no correlation between nodes). In such circumstance $\mathcal{G}$ can be described by $\mathcal{V}$ and set of degrees, $\mathcal{K} = \{k_1, k_2, ..., k_{|\mathcal{V}|}\}$.

A GNM at time t is specified by $\mathcal{V}$, and $\mathcal{K}$, parameterised by number of edges added per time-step, m, and node choice probability, $\Pi(k)$. It is initialised with initial graph $\mathcal{G}_0$. It is iterated as below:

1. Initialisation. Let initial state be $\mathcal{G}_0$, $t = 0$.

2. Node Addition.

- $t \to t + 1$.
- Add new node: $\{1, 2, ..., |\mathcal{V}(t)|\} \to \{1, 2, ..., |\mathcal{V}(t)|, |\mathcal{V}(t)| + 1\}$
- Add m edges with one end of stub on new node $k_{|\mathcal{V}(t)|+1} = m$,
  and the other end with m existing nodes $k_i \to k_i + 1$, chosen with $\Pi(k)$.
  Cannot choose the same node twice.

3. Iteration. Repeat **Node Addition.** until $N = |\mathcal{V}|$.

For the BA/PurePref model, $\Pi(k) \propto k$; for the PureRand model, $\Pi(k) =$ const., and for the mixed model, Π is chosen between the with ratio q respectively.

1.2 Numerical implementation

The model was implemented in Object-Oriented Python language. Aside from the parameters required to describe the system, an additional list, called `StubList`, contained each stub denoted by the index of the node. This is because choosing a random

Value	Predicted	Measured
$N_{pref,rand,mixd}(t = 1000; N_0 = 3)$	1003	1003
$p(1)/1$	0.018	0.01821
$p(2)/2$	0.018	0.01817
$p(3)/3$	0.018	0.01814
$p(4)/4$	0.018	0.01821
$p(5)/5$	0.018	0.01818

Figure 1: *Test results;* $\#_{repeats} = 1000000$. *All results match up with requirements*

element in `StubList` should choose a node with probability proportional to its degree. This method dramatically reduced code runtime.

1.2.1 Tests for Implementation

Various tests were done to validate the implementation. The first test validated the number of nodes and stubs added per time step. The second validated the method of choosing a random element of `StubList` really produced $\Pi \propto k$ for the BA model. If working as intended, probability of choosing a node from `StubList=[1,2,2,3,3,3,4,4,4,4,5` divided by node index should give a constant value. Lastly, since implementation closely followed derivations, indicates validity of implementation. The results of the tests are summarised in Fig. 1

1.3 Master Equation

Assuming mean-field approximation, GNMs can be described by a master equation, describing state of $t + 1$ as function of state at t:

$$n(k, t+1) = \underset{A}{n(k,t)} + m\underset{B}{\Pi(k-1,t)}n(k-1,t) - m\underset{C}{\Pi(k,t)}n(k,t) + \underset{D}{\delta_{k,m}}, \tag{1}$$

where $n(k, t)$ is the number of nodes with degree k at t, $\Pi(k, t)$ is the probability of choosing a node with degree k. Each term is interpreted as below:
A: Number of nodes of degree k at time t.
B: Average number of nodes promoted to degree k.
C: Average number promoted from degree k to degree $k + 1$ and therefore removed from degree k.
D: The new node is always added with degree m.

Assume that $p(k, t) = \frac{n(k,t)}{N(t)}$. Substituting to (1),

$$p(k,t+1)N(t+1) = p(k,t)N(t) + m\Pi(k-1,t)p(k-1,t)N(t) - m\Pi(k,t)p(k,t)N(t) + \delta_{k,m}.$$

Assume that as $t \to \infty$, probability is steady. Noting that $N(t + 1) = N(t) + 1$,

$$p_\infty(k) = +m\Pi(k-1)p_\infty(k-1)N - m\Pi(k,t)p_\infty(k)N + \delta_{k,m}. \tag{2}$$

Eq. (2) solved with $\Pi(k, t)$ specified, by approximating it as an ODE, $p_{\infty,cont}$, or discretely, $p_{\infty,disc}$.

1.4 Methods to approximate average largest node k_1

$k_1(N; m; \mathcal{G}_0)$ is defined as the largest node for a given N, averaged over the ensemble of graphs. It can be approximated in two methods.

1.4.1 Infinite Probability Sampling

Assume that GNM samples from $p_\infty \; \forall \; N$, and that $p_\infty = \frac{n(k,N)}{N} \; \forall \; N$. For a GNM of size N, on average, expect that the lowest probability represented is $1/N$. Therefore k_1 is estimated by equating cumulative probability of $[k_1, \infty]$ to $1/N$. For $p_{\infty,cont}$,

$$\int_{k_1}^{\infty} p_\infty(k; m) dk = \frac{1}{N} \tag{3}$$

For $p_{\infty,disc}$,

$$\sum_{k_1}^{\infty} p_\infty(k; m) = \frac{1}{N} \tag{4}$$

1.4.2 Node degree evolution

The average number of stubs attached in an iteration is given by

$$1 - (1 - \Pi(k_i(N), N)^m) \approx m\Pi(k_i(N), N) \quad \text{for } \Pi \ll 1. \tag{5}$$

So

$$k_i(N + 1) = k_i(N) + m\Pi(k_i(N), N). \tag{6}$$

So $k_i(N)$ is iterated from its initial degree. Nodes belonging to $\mathcal{G}_0$ will experience the most iterations, so will have largest degree on average. For PurePref/Mixed models, for $m \gg 1$, expect nodes with the highest k's gain ~ 1 degree per iteration; therefore expect chance that nodes added later to overtake initial nodes be small, and tracking the node of the highest degree in $\mathcal{G}_0$ should approximate k_1. For $m \sim 1$, expect significanct chance that later nodes can overtake the initial nodes; expect deviation.

1.5 Choice of Initial Graph $\mathcal{G}_0$

All GNMs add m edges per iteration. To accommodate m edges at the first time step, require

$$m \leq N_0, \tag{7}$$

where N_0 is initial number of nodes. N and E evolve as

$$N(t) = N_0 + t = t\left(1 + \frac{N_0}{t}\right) \Rightarrow N(t \to \infty) = t, \tag{8}$$

$$E(t) = E_0 + mt = mt\left(1 + \frac{E_0}{mt}\right) \Rightarrow E(t \to \infty) = mt. \tag{9}$$

Since long time limit is to be compared with data, want Eqs. (8), (9) to converge to long time limit as fast as possible. Therefore aim to minimise N_0, E_0.

For PurePref, $\Pi(k_i = 0) = 0$. Therefore require $k_i(t = 0) > 0 \; \forall \; i$. It is also assumed that $p_\infty(k < m) = 0$. Since there is a finite probability that nodes with $k_i(t = 0) < m$ stays less than m, choose $k_i(t = 0) \geq m$. k_i is limited by complete graph limit so have

$$m \leq k_i(t = 0) \leq N_0 - 1 \; \forall \; i \in \mathcal{V}_0 \tag{10}$$

Choosing graph with minimum N_0 with requirement $k_i(t = 0) \geq m$, get complete graph of $N_0 = m + 1$. All nodes have degree $k_i(t = 0) = k_0 = m$.

When investigating k_1, choose complete graph of size $N_0 = 2m + 1$, with $k_0 = 2m$. This the minimally sized graph such that $E = mN$ for all times. This is to ensure that form of Π holds at all times for the PurePref and Mixed models.

2 Pure Preferential Attachment

2.1 Justification for Π_{pref}

The naive choice for $\Pi \propto k$ is $\Pi_{pref}(k, t) = \frac{k}{\sum_{i=0}^{N(t)} k_i(t)}$. Since $\sum_{i=0}^{N(t)} k_i(t) = 2E(t)$,

$$\Pi_{pref}(k, t) = \frac{k}{2E(t)}. \tag{11}$$

However, this does not take into account the chance to pick a same node again. Consider the probability to choose the same node twice with the proposed $\Pi_{pref}(k, t)$. Noting that $E(t) = E_0 + mt$,

$$\Pi_{pref,twice}(k, t) = \left(\frac{k}{E_0 + mt} \right)^2.$$

The maximum possible degree, k_{max} has degree $N - 1$. Since $N(t) = N_0 + t$, $k_{max}(t) = N_0 - 1 + t$. So $t \gg 1$, $\Pi_{twice}(k, t)$ scales like

$$\Pi_{pref,twice}(k_{max}, t) \propto \frac{t}{t^2} = \frac{1}{t}$$

which goes to zero as $t \to \infty$. Therefore argue that Eq. (11) is valid as $t \to \infty$.

2.2 Long Term Probability $p_{\infty,pref}(k)$ Derivation

Consider Eq (2). Under long time limit or tuned $\mathcal{G}_0$, $E = mN$, so $\Pi_{pref}(k, t) = \frac{k}{2mN}$. Then

$$p_{pref,\infty}(k) = \frac{(k-1)p_\infty(k-1)}{2} - \frac{kp_\infty(k)}{2} + \delta_{k,m}. \tag{12}$$

2.2.1 Continuous Solution

Rearranging,

$$p_{pref,\infty}(k) = -\frac{1}{2} \left(\frac{kp_{pref,\infty}(k) - (k - \Delta k)p_{pref,\infty}(k - \Delta k)}{\Delta k} \right) + \delta_{k,m}, \to -\frac{1}{2} \frac{\partial \left(kp_{pref,\infty}(k) \right)}{\partial k}.$$

where $\Delta k \coloneqq 1$. For $k \gg 1 \Rightarrow \Delta k \ll k$, since above expression tends to a differential and boundary condition $\delta_{k,m}$ ignored as $k \gg m > 1$.

Substituting trial solution $p_{pref,\infty}(k) = Ak^{-\gamma}$, find that $\gamma = 3$. Since there can be no nodes with degrees $< m$, require normalisation

$$\int_m^\infty p_{pref,\infty}(k)dk = \int_m^\infty Ak^{-3}dk = A\left[\frac{-1}{2}k^{-2}\right]_m^\infty = 1 \Rightarrow A = 2m^2$$

$$\boxed{p_{pref,\infty}(k) = 2m^2k^{-3}\ ;\ k \geq m} \tag{13}$$

2.2.2 Discrete Solution

Return to Eq. (12). Consider when $k > m$. Rearranging,

$$\frac{p_{pref,\infty}(k)}{p_{pref,\infty}(k-1)} = \frac{k-1}{k+2}$$

Use that

$$\frac{f(z)}{f(z+1)} = \frac{z+a}{z+b} \Rightarrow f(z) = A\,\frac{\Gamma(z+1+a)}{\Gamma(z+1+b)} \tag{14}$$

where a, b, A are some constants and $\Gamma(z)$ is the Gamma function. For $p_{\infty,pref}(k)$, $a = -1, b = 2$. Then

$$p_{pref,\infty}(k) = A\,\frac{\Gamma(k)}{\Gamma(k+3)} = \frac{A}{k(k+1)(k+2)}. \tag{15}$$

To normalise, consider Eq. (12) for $k = m$. Any new node must have $k \geq m$, so only initial nodes can have degree $k < m$. At long time limit, the proportion of these go to zero, so $p_{pref,\infty}(k < m) = 0$:

$$p_{pref,\infty}(m) = \frac{m-1}{2}\cancelto{0}{p_{pref,\infty}(m-1)} - \frac{m}{2}p_{pref,\infty}(m) + 1 \Rightarrow p_{pref,\infty}(m) = \frac{2}{2+m} \tag{16}$$

Probability must sum to unity:

$$1 = \sum_{k=m}^{\infty} p_{pref,\infty}(k) = \frac{2}{2+m} + A\sum_{k=m+1}^{\infty}\frac{1}{k(k+1)(k+2)} = \frac{2}{2+m} + A\cdot S_\infty \tag{17}$$

Decomposing S_∞ by partial fraction and considering the (n+2)$^{\text{th}}$ term, can see that the sum tends to

$$S_\infty = \frac{1}{2(m+1)(m+2)}. \tag{18}$$

Combining Eqs. (17), (18),

$$1 = \frac{2}{2+m} + A\,\frac{1}{2(m+1)(m+2)} \Rightarrow A = 2m(m+1).$$

It is easy to show that Eq (16) is equal to Eq (15) for $k = m$. Therefore

$$\boxed{p_{pref,\infty,disc}(k;m) = \frac{2m(m+1)}{k(k+1)(k+2)}\ ;\ k \geq m} \tag{19}$$

2.3 Comparing $\tilde{p}_{pref}(k;m;N)$ with $p_{\infty,pref}(k;m)$

The BA Model was simulated for fixed $N = 10000$ and $m = [2, 4, 8, 16, 32]$.

Since the measured probability, $\tilde{p}_{pref}(k;m;N)$, is to be compared with $p_{\infty,pref}(k;m)$, maximum possible N should be used. $N = 10000$ provided large enough scale-free region while keeping runtime short enough to allow many repeats. m's were chosen to be powers of two to cover log-space evenly. Higher m means more choices and therefore longer runtime, so only went up to 32. This was the case for later sections as well.

$\tilde{p}_{pref}(k;m;N)$ is fat-tailed; it has features in low p. Resolution of the low p's was improve via two methods.

- **Repeats:** Every run, the degree series was frequency-counted, with frequency counted as *zero* if no events. Over multiple runs, the frequency lists were averaged. If degree series were to be simply summed and binned, then the result would bias small probability events, since for a N-sized sample, the minimum measured probability is $1/N$; so probabilities $< 1/N$ would be boosted. $\#_{repeats} = 1000$ was used unless specified otherwise.
- **Log-Binning:** Since many $k \gg 1$ had zero frequency, points were binned where the j^{th} bin covers interval $[a^j, a^{j+1})$, for scale parameter a. For proper normalisation, bins with zero counts were thrown away. $a = 1.05$ was used for all the plots unless specified otherwise. (Code from Complexity Project was edited and used) The effect of log-binning is demonstrated in Fig. 8.

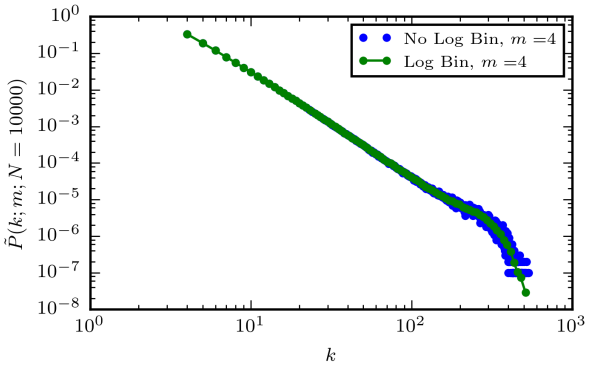


Figure 2: *Plot of $\tilde{p}_{pref}(k;m=4;N=10000)$ with $\#_{repeats} = 1000$, w/o log-binning (blue) and w/ log-binning, $a = 1.05$ (green). Many repeats mean resolution is already good; noise is reduced by log-binning.*

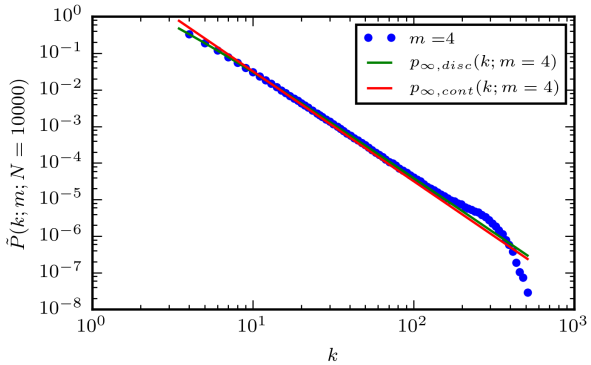


Figure 3: *Plot of $\tilde{p}_{pref}(k;m=4;N=10000)$, compared with $p_{pref,\infty,cont}(k;m)$ and $p_{pref,\infty,disc}(k;m)$. Note that discrete solution agrees better with data especially at low k. This was true for all three models*

$\tilde{p}_{pref}(k;m;N)$ is compared with Eqs. (13), (19) in Fig. 3, which shows that discrete solution more closely resembles the data. This was the case for all GNMs; discrete solution was used for later sections.

$\tilde{p}_{pref}(k; m; N = 10000)$ and $p_{pref,\infty,cont}(k; m)$ for different m's are log-log plotted in Fig 4. The plot reveals two regions: the scale-free region, where $\tilde{p}$ follows p_∞, and the finite-size scaling region, where a characteristic 'bump' is seen, followed by a cut-off. The cutoff is explained by finite size effects. The bumps, which increases with m, can be explained by recalling argument of Sec. 1.4.2: For $m \gg 1$, expect that initial nodes will not be overpassed by added nodes, so initial nodes will have a higher degrees than ones added, resulting in a more pronounced bump.

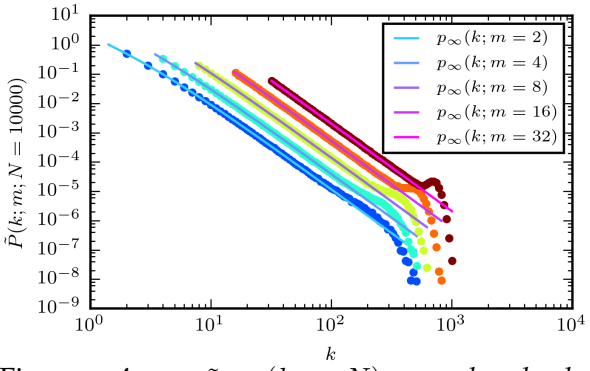


Figure 4: *$\tilde{p}_{pref}(k; m; N)$ overlayed by $p_{\infty,pref}(k; m)$ for $N = 10000, \#_{repeats} = 1000$. Discrete solution fits the data well up to the scaling region. The heights of characteristic 'bumps' increase with m.*

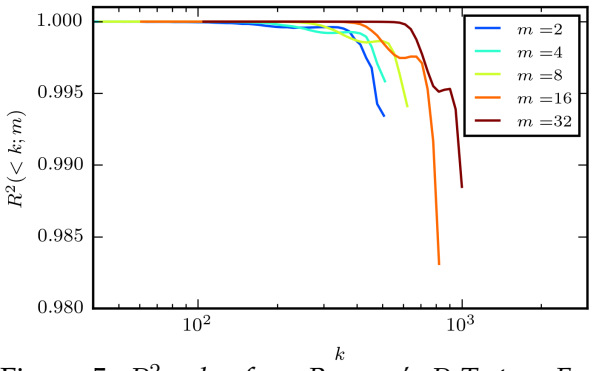


Figure 5: *R^2 value from Pearson's R-Test on Eq. (20), test done on $[m, k]$ of dataset. Note a good R value followed by a drop when including cut-off points.*

many standard statistical tests were considered. Chi-square test requires frequencies of ≥ 5, so data must be binned. However binning loses redolution near cut-off. Visually, a poor fit is seen after the scaling region, so a test that considers regions at a time was required. KS test can only consider entire distributions so was not used.

If two distributions agree,

$$\ln(p_\infty) \text{ vs. } \ln(\tilde{p}_{data}) \tag{20}$$

should result in a straight line. However this test does not provide p-values. A Pearson's R test on was conducted on Eq. (20) for various ranges up to k. The result is plotted in Fig. 5.

From the figure it can be seen that the R^2 value is 1 (3 s.f.) when not including the finite-size scaling region, but drops when included. This indicates that p_∞ is a good fit up to the scaling region.

2.4 k_1 Derivation

Using methods of Sec. 1.4.1, the continuous estimation finds

$$\int_{k_1}^{\infty} p_{pref,\infty}(k; m)dk = m^2 k_1^{-2} = \frac{1}{N} \Rightarrow \boxed{k_{1,pref} = m\sqrt{N}}, \tag{21}$$

discrete estimation finds

$$\sum_{k_1}^{\infty} p_{pref,\infty}(k;m) = \frac{m(m+1)}{k_1(1+k_1)} = \frac{1}{N} \Rightarrow \boxed{k_{1,pref} = \frac{1}{2}\left(\sqrt{4m(m+1)N+1} - 1\right)} \quad (22)$$

$\to m\sqrt{N}$ for $m \gg 1, N \gg 1$. Using methods of Sec. 1.4.2,

$$k_i(N+1) \approx k_i(N) + \frac{\cancel{m}k_i(N)}{2\cancel{m}N} \Rightarrow \frac{k_i(N+1)}{k_i(N)} = \frac{2N+1}{2N} \Rightarrow k_1(N) = k_1(N_0) \prod_{j=N_0}^{N-1} \frac{2j+1}{2j}$$

$$\boxed{k_{1,pref}(N) \approx k_1(N_0)\frac{\Gamma(N+1/2)}{\Gamma(N)}\frac{\Gamma(N_0)}{\Gamma(N_0+1/2)} \approx 2m\sqrt{N}\frac{\Gamma(2m+1)}{\Gamma(2m+1.5)} \to \sqrt{2mN}} \quad (23)$$

(For $m \gg 1, N \gg 1$, since $z \to \infty$, $\Gamma(z+\alpha)/\Gamma(z) \to z^{\alpha}$, and assuming $N_0 = 2m+1, k_1(N_0) = 2m$). It predicts a different scaling for m.

Setting $Y \coloneqq \ln(k_{1,pref})$, $X \coloneqq \ln(N)$, for $N \gg 1$, all three derivations predict that

$$k_{1,pref}(N;m) \propto \sqrt{N} \Rightarrow Y = \frac{1}{2}X + \text{const.}(m) \quad (24)$$

which should result in a gradient of 0.5.

2.5 $\hat{k}_1$ Data Comparison

2.5.1 Measuring $\hat{k}_1$

$\hat{k}_1$ was estimated by running a simulation with the same parameters, recording the highest degree every run, and taking the mean and the standard deviation as the measured value and errors respectively. An example of the measured distribution for m=4,N=10000 is shown in Fig 6. From the figure it can be seen that the distribution is skewed to the right. This is expected due to the preferential attachment nature of the system; if a node gets an advantage, it would snowball since it now has a greater chance of being chosen.

2.5.2 Scaling with N

Shown in Fig. 7, Eq. (24) was plotted and fitted to a linear function for $m = 4$, $N = [500, 1000, ..., 32000]$. The slope was measured to be 0.503 with R of 1 to 2 decimal places. The result agrees with the theoretical predictions.

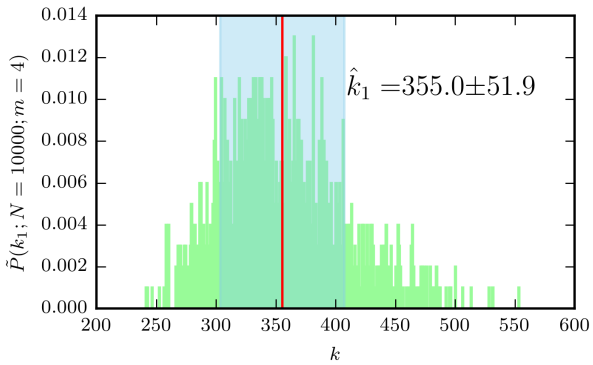


Figure 6: *Example of a distribution of measured k_1's with $\#_{repeats} = 1000$. It is a widely peaked, skewed distribution. The skew is explained by 'win-more' nature of attachment probability.*

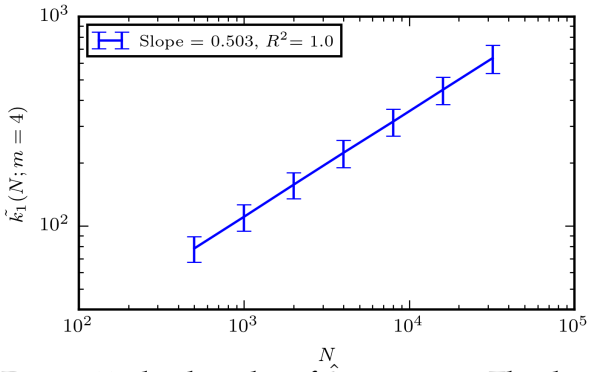


Figure 7: *log-log plot of $\hat{k}_1$ vs. N. The data matches expected slope of 0.5 to two significant figures. high R^2 indicates a good fit.*

2.5.3 Extension: Scaling with m

The three predictions for k_1 were tested against m for $N = 10000$, $\#_{repeats} = 50$. The infinite probability estimations are not good fits but the node degree evaluation method aligns for high m. This agrees with the predictions of Sec. 1.4.2, since derivation was only valid for $m \gg 1$.

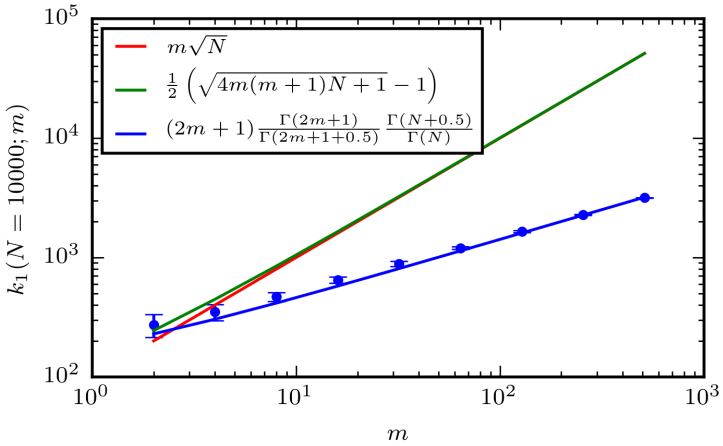


Figure 8: *Measured scaling of k_1 with m for $N = 10000$ compared to the three derivations. Node Degree Evolution method is best, with better accuracy as m increases, as predicted.*

2.6 Data Collapse and Investigation in $\mathcal{G}_0$

2.6.1 Data Collapse Derivation

From the data of Fig. 4, it can be seen that the system displays scaling behaviour with a cutoff. Figure also shows that different m's give different 'bumps'. So, employ Finite-Scaling Ansatz (FSA) used to explain other power-law distributions with cutoffs:

$$p(k; m; N) = p_\infty(k)\ \mathcal{F}\left(\frac{k}{k_1}; m\right) \tag{25}$$

where the scaling function $\mathcal{F}$ behaves as

$$\mathcal{F}(x;m) = \begin{cases} \mathcal{F}_0 + \mathcal{F}_1 x + \ldots & x \ll 1 \\ \text{decaying rapidly} & x \gg 1 \end{cases} \tag{26}$$

and k_1 is the cutoff degree.

In from derivation and data, it was shown that $k_1 \propto \sqrt{N} \Rightarrow k_1 = \text{const.}(m)\sqrt{N}$ for some constant b. Substituting for $p_{\infty,pref}$ and rearranging, and defining $Y = k(k+1)(k+2)p(k;m;N)$ and $X = \frac{k}{\sqrt{N}}$

$$Y = \mathcal{F}\left(\frac{Y}{\text{const.}(m)};m\right) \tag{27}$$

If FSA is valid, then expect a data collapse with Eq. (27).

2.6.2 $\tilde{p}(k;m;N)$ vs. N and Data Collapse

$\tilde{p}(k;m;N)$ was plotted for $m = 3$, $N = [1000, 2000, 4000, 8000, 16000]$ are plotted in Fig. 9, and collapsed in Fig 10. The points line up, indicating that Eq. (27), FSA, is valid.

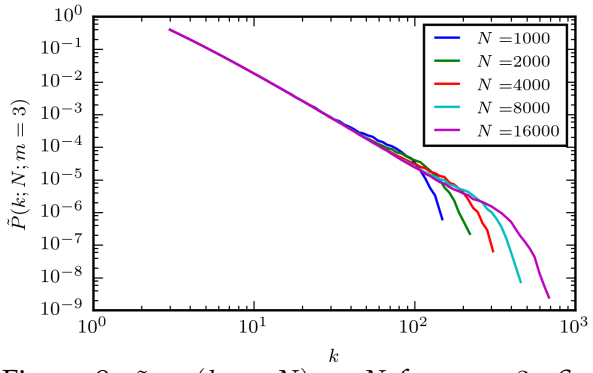


Figure 9: $\tilde{p}_{pref}(k;m;N)$ *vs.* N *for* $m = 3$. *See that cut-off regions shifts as* N *increases.*

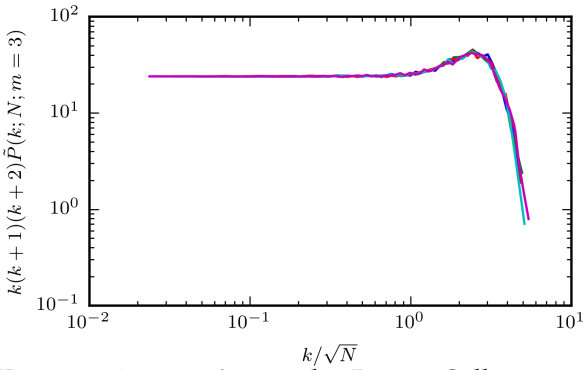


Figure 10: *A good Data Collapse of* $\tilde{p}_{pref}(k;m;N)$ *vs.* N *for* $m = 3$. *Indicates that FSA was valid.*

2.6.3 Extension: Effect of $\mathcal{G}_0$ and Data Collapse

The effect of initial conditions was also studied. For $m = 3$, $N = 1000$, all starting nodes were set to degree $k_0 = m$. N_0 was varied as $N_0 = 4, 8, 16, 32$. The results are plotted in Fig. 11. Figure shows that as N_0 increases, the scale-free region becomes smaller. this confirming the prediction of Sec. 1.5: if N_0 is large, it would take longer for the system to reach the long time limit. Also see that the size of the 'dips' become larger with N_0. Region up to the dip represents contributions only by added nodes. Because the number of head-started initial nodes, N_0 increases, added nodes receive less then they would if there were less initial nodes, resulting in a dip in probability.

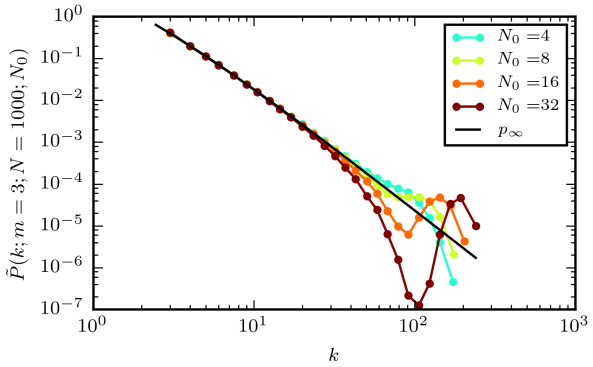


Figure 11: *$\tilde{p}_{pref}$(k;m=3,N=1000,k_0=m,N_0) vs. k for various N_0's. 'Dips' become more pronounced with N_0 as later added nodes receive smaller proportion of nodes, and amplified by 'win-more' node addition.*

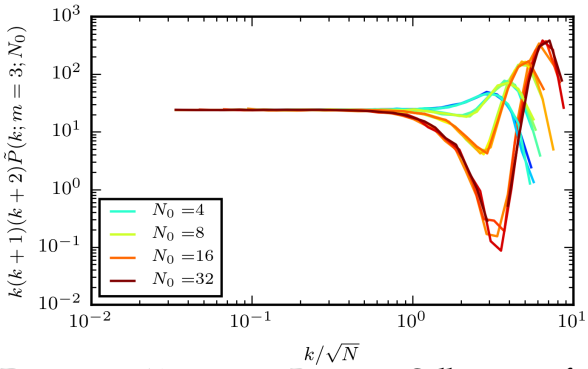


Figure 12: *Data Collapse for $\tilde{p}_{pref}$(k;m=3,N=1000,k_0=m,N_0), $N = [1000, 2000, ..., 8000]$. A separate data collapse for each N_0 indicates that scaling function $\mathcal{F}$ is a function of $\mathcal{G}_0$.*

For each N_0, N was varied as $N = [1000, 2000, ..., 8000]$, and data was collapsed, as shown in Fig 12. For each N_0, there is a good separate data collapse. This indicates that $\mathcal{F}$ is also a function of the initial conditions. Amend FSA, Eq. (25):

$$p(k; m; N) = p_\infty(k)\ \mathcal{F}\left(\frac{k}{k_1}; m; \mathcal{G}_0\right) \tag{28}$$

3 Pure Random Attachment

3.1 Justification for Π_{rand}

A constant, normalised Π_{rand} is

$$\Pi_{rand}(k, t) = \frac{1}{N(t)}. \tag{29}$$

Check probability to pick twice:

$$\Pi_{rand,twice}(k, t) = \frac{1}{N(t)^2} = \frac{1}{(N_0 + t)^2}$$

this scales as $\frac{1}{t^2} \to 0$ as $t \to \infty$. Eq (29) is valid as $t \to \infty$.

3.2 Long Time Probability $p_{\infty,rand}(k)$ Derivation

Using Eqs (2), (29),

$$p_{rand,\infty}(k) = m\left(p_{rand,\infty}(k-1) - p_{rand,\infty}(k)\right) + \delta_{k,m}. \tag{30}$$

3.2.1 Continuous Solution

As Sec. 2.2.1, rearrange to find

$$p_{rand,\infty}(k) = -m\left(\frac{p_{rand,\infty}(k) - p_{rand,\infty}(k - \Delta k)}{\Delta k}\right) + \delta_{k,m}.$$

Where $\Delta k = 1$. For $k \gg m > 1$, approximate as PDE

$$p_{rand,\infty}(k) = -m\frac{\partial p_{rand,\infty}(k)}{\partial k}$$

The trial solution $p_{rand,\infty}(k) = Ae^{-k/m}$ fulfills the requirement of the PDE. Normalising,

$$\int_m^\infty p_{rand,\infty}(k)dk = \int_m^\infty Ae^{-k/m}dk = A\left[-me^{-k/m}\right]_m^\infty = 1 \Rightarrow A = \frac{e}{m}$$

$$\boxed{p_{rand,\infty,cont}(k) = \frac{e^{1-k/m}}{m}\ ;\ k \geq m.} \tag{31}$$

3.2.2 Discrete Solution

Consider Eq. (30) when $k > m$. Rearranging,

$$\frac{p_{rand,\infty}(k)}{p_{rand,\infty}(k-1)} = \frac{m}{m+1}, \Rightarrow p_{rand,\infty}(k) = p_{rand,\infty}(m)\left(\frac{m}{m+1}\right)^{k-m}$$

Now consider $k = m$:

$$p_{rand,\infty}(m) = m(\cancelto{0}{p_{rand,\infty}(m-1)} - p_{rand,\infty}(m)) + 1 \Rightarrow p_{rand,\infty}(m) = \frac{1}{m+1}.$$

$$\boxed{p_{rand,\infty,disc}(k) = \frac{1}{m+1}\left(\frac{m}{m+1}\right)^{k-m}\ ;\ k \geq m} \tag{32}$$

Check normalisation; summing all probabilities,

$$\sum_{k=m}^{\infty} p_{rand,\infty}(k) = \frac{1}{m+1}\sum_{k'=0}^{\infty}\left(\frac{m}{m+1}\right)^{k'} = \frac{1}{m+1}\frac{1}{1-\left(\frac{m}{m+1}\right)} = 1,$$

as required, since it is infinite geometric series.

3.3 Comparing $\tilde{p}_{rand}(k; m; N)$ with $p_{\infty,rand}(k; m)$

$\tilde{p}_{rand}(k; m; N)$ for same parameters as Sec 2.3 are plotted along with the discrete prediction (32) and shown in Fig. 13, with the R^2 value of Eq. (20) fitted to the different parts of the distribution, shown in Fig. 14. As before, there appears to be cutoff due to system size. Secondly, note that with the same parameters, the PureRand model cuts off faster, at $\sim 10^2$ compared the PurePref model $\sim 10^3$. The distribution is smaller tailed.

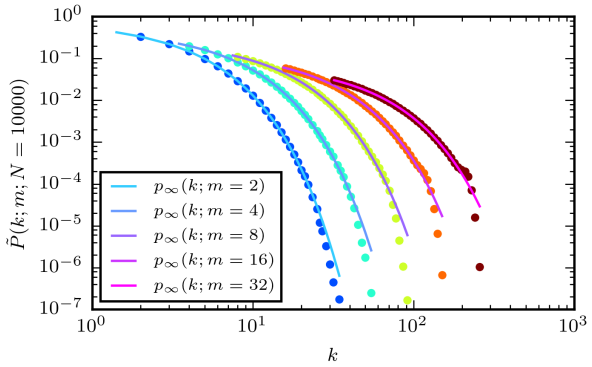


Figure 13: $\tilde{p}_{rand}(k;m;N)$ *overlayed by* $p_{\infty,rand}(k;m)$, *same parameters as before. The 'bumps' are less pronounced compared to PurePref. Cut-off is sharper than PurePref, and finite-size effects are difficult to discern.*

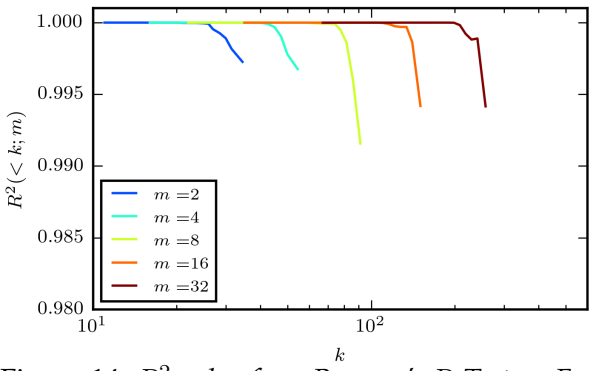


Figure 14: R^2 *value from Pearson's R-Test on Eq. (20), test done on* $[m,k]$ *of dataset. Note a good R value followed by a drop when including cut-off points.*

3.4 k_1 Derivation

Using methods of Sec. 1.4.1, the continuous estimation was found to be

$$\int_{k_1}^{\infty} p_{rand,\infty}(k;m)dk = e^{1-k_1/m} = \frac{1}{N} \Rightarrow \boxed{k_{1,rand} = m(1+\ln(N))} \tag{33}$$

with the discrete case

$$\sum_{k_1}^{\infty} p_{rand,\infty}(k;m) = \left(\frac{m}{m+1}\right)^{k_1-m} = \frac{1}{N} \Rightarrow \boxed{k_{1,rand} = \frac{\ln(N)}{\ln(\frac{m+1}{m})} + m} \tag{34}$$

Using methods of Sec. 1.4.2,

$$k_i(N+1) \approx k_i(N) + \frac{m}{N} \Rightarrow k_i(N) = k_i(N_0) + m\sum_{j=N_0}^{N-1}\frac{1}{j}$$

Using approximation of harmonic numbers, $\sum_{a=1}^{N} 1/a \approx \ln(N)\ ;\ N \gg 1$,

$$\boxed{k_{1,rand}(N) \approx k_i(N_0) + m\ln\left(\frac{N-1}{N_0-1}\right) \to 2m + m\ln\left(\frac{N}{2m}\right)} \tag{35}$$

For $N \gg 1$ and assuming $N_0 = 2m+1, k_1(N_0) = 2m$.

All three derivations indicate for $N \gg m > 1$, $k_{1,rand} \propto \ln(N)$. Letting $Y = k_1$, $X = \ln(N)$, expect

$$Y = \text{const.}(m)X \tag{36}$$

which should give a good fit to linear regression.

3.5 $\hat{k}_1$ Data Comparison

3.5.1 Scaling with N

Eq. (36) was plotted and linearly regressed in Fig. 15, with $R^2 = 1.0$ for 2 significant figures. This validates the derivations.

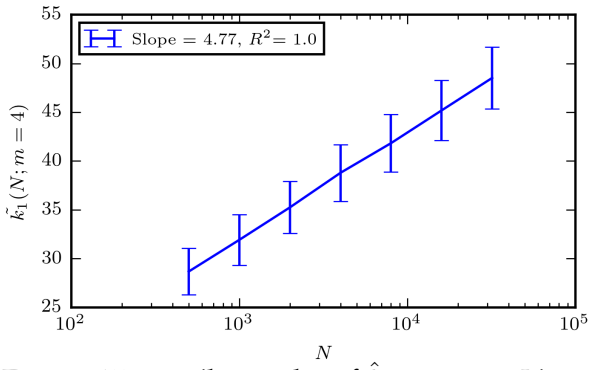


Figure 15: *semilog-x plot of $\hat{k}_1$ vs. N. Linear regression finds a High R^2 value which indicates straight line is a good fit; prediction of scaling is valid.*

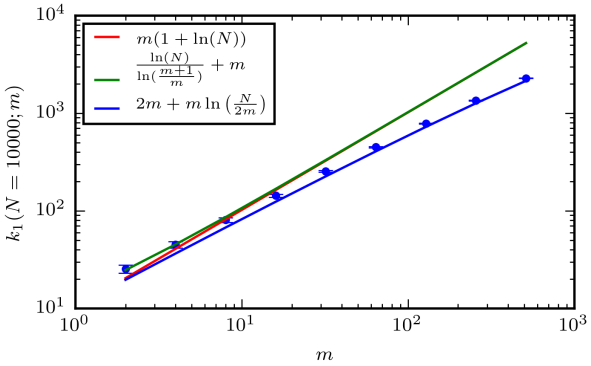


Figure 16: *Measured scaling of k_1 with m for $N = 10000$ compared to the three derivations. Node Degree Evolution method is best, with good accuracy.*

3.5.2 Extension: Scaling with m

Similarly, the scaling with m was investigated, with a similar result showing that degree evolution method gave the best prediction, and getting better with m.

3.6 Data Collapse

3.6.1 Data Collapse Derivation

Using the amended FSA. Eq. (28), with $k_{1,rand}(N) \propto \ln(N) \Rightarrow k_{1,rand}(N) = \text{const.}(m)\ln(N)$ substituting for $p_{\infty,pref}$ and rearranging, and defining $Y = \left(\frac{m}{m+1}\right)^{-k} \tilde{p}_{rand}(k; m; N)$ and $X = \frac{k}{\ln(N)}$, expect

$$Y = \mathcal{F}\left(\frac{X}{\text{const.}(m)}; m; \mathcal{G}_0\right) \tag{37}$$

3.6.2 $\tilde{p}(k; m; N)$ vs. N and Data Collapse

With the same parameters as PurePref model, $\tilde{p}_{rand}(k; m; N)$ vs. N was plotted shown in Fig. 17 and data collapsed in Fig. 18. The spike shown at the edge of Fig. 18 is explained by poor resolution at the tails.

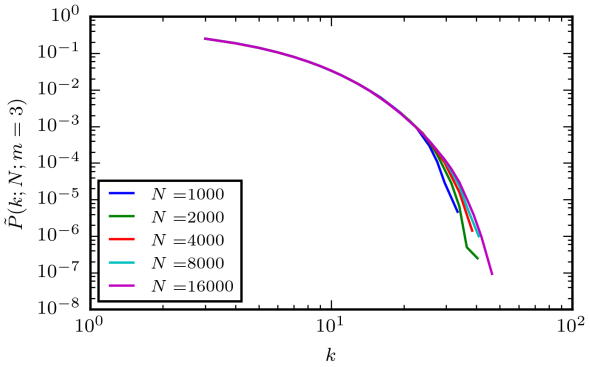


Figure 17: $\tilde{p}_{rand}(k;m;N)$ *vs.* N *for* $m=3$. *Cut-off is difficult to see as even the long time probability has sharp cut-off.*

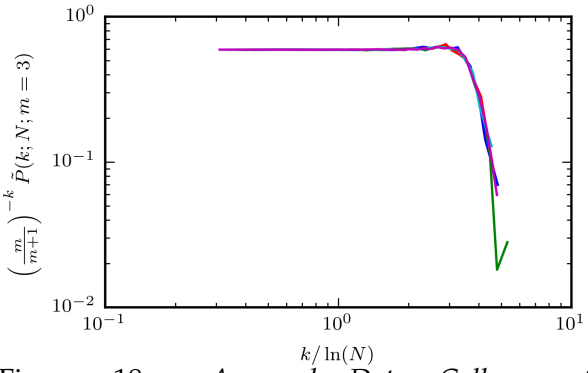


Figure 18: *A good Data Collapse of* $\tilde{p}_{rand}(k;m;N)$ *vs.* N *for* $m=3$. *Indicates that FSA was valid. The form of the bare scaling function is seen, which was hidden before.*

4 Mixed Preferential and Random Attachment

4.1 Justification for Π_{mix}

Π for Mixed Attachment Model is given by

$$\Pi_{rand}(k,t) = q\Pi_{pref}(k,t) + (1-q)\Pi_{rand}(k,t).$$

where q is the ratio of preference. Substituting Eqs. (11), (29), assuming long term limit or tuned $E = mN$,

$$\Pi_{mixd}(k,t) = \frac{q}{2mN(t)} + \frac{1-q}{N(t)} \tag{38}$$

Can show that $\Pi_{mix,twice} \propto 1/t \to 0$ as $t \to \infty$.

4.2 Long Term Probability $p_{\infty,mix}(k)$ Derivation

Using Eqs (2), (29),

$$p_{mix,\infty}(k) = mp_{mix,\infty}(k-1)\left[\frac{q(k-1)}{2m} + (1-q)\right] - mp_{mix,\infty}(k)\left[\frac{qk}{2m} + (1-q)\right] + \delta_{k,m}. \tag{39}$$

4.2.1 Continuous Solution

Rearranging,

$$p_{mix,\infty}(k) = -\frac{q}{2}\left[\frac{kp_{mix,\infty}(k) - (k-\Delta k)p_{mix,\infty}(k-\Delta k)}{\Delta k}\right] + m(1-q)\left[\frac{p_\infty(k) - p_\infty(k-\Delta k)}{\Delta k}\right] + \delta_{k,m}.$$

$\Delta k := 1$. which for $k \gg m > 1$, tends to

$$p_{mix,\infty}(k) = \frac{-q}{2}\frac{\partial(kp_{mix,\infty}(k))}{\partial k} - m(1-q)\frac{\partial p_{mix,\infty}(k)}{\partial k}$$

This is an ODE with the standard result

$$p_{mix,\infty}(k) = A(2m(1-q)+kq)^{-(2/q+1)}$$

Normalise:

$$\int_m^\infty p_{mix,\infty}(k) = A\int_m^\infty (2m(1-q)+kq)^{-(2/q+1)} = 1 \Rightarrow A = 2(m(2-q))^{2/q}$$

$$\boxed{p_{mix,\infty,cont}(k) = 2(m(2-q))^{2/q}(2m(1-q)+kq)^{-(2/q+1)}\ ;\ k \geq m} \tag{40}$$

4.2.2 Discrete Solution

From Eq (39), consider the case $k > m$. Rearranging,

$$\frac{p_{mix,\infty}(k)}{p_{mix,\infty}(k-1)} = \frac{k+(2m(1/q-1)-1)}{k+(2m(1/q-1)+2/q)}$$

Similarly to Sec. 2.2.2, this has the solution of Eq. (14) with $a = 2m(1/q-1)-1$ and $b = k+[2m(1/q-1)+2/q$. So the solution has the form

$$p_{mix,\infty}(k) = A\,\frac{\Gamma(k+2m(1/q-1))}{\Gamma(k+2m(1/q-1)+2/q+1)} \tag{41}$$

for $k > m$. Now consider $k = m$ to find

$$p_{mix,\infty}(m) = \frac{2}{2+m(2-q)} \tag{42}$$

Find A by summing over probabilities:

$$\sum_{k=m}^{\infty} p_{mix,\infty}(k) = \frac{2}{2+m(2-q)} + A\sum_{k=m+1}^{\infty}\frac{\Gamma(k+2m(1/q-1))}{\Gamma(k+2m(1/q-1)+2/q+1)} = 1$$

the infinite series can be found to be

$$S_\infty = \frac{(2+q+m(2-q))\Gamma\left(1+m(2/q-1)\right)}{\Gamma\left(2+2/q+m(2/q-1)\right)}$$

to find

$$\boxed{A(m,q) = \frac{2m(2-q)}{(2+m(2-q))(2+q+m(2-q))}\frac{\Gamma(2+2/q+m(2/q-1))}{\Gamma(1+m(2/q-1))}} \tag{43}$$

Setting $k = m$, (41) is equal to (42). So the discrete solution is

$$\boxed{p_{mix,\infty,disc}(k) = A(m,q)\frac{\Gamma(k+2m(1/q-1))}{\Gamma(k+2m(1/q-1)+2/q+1)}\ ;\ k \geq m} \tag{44}$$

4.3 Comparing $\tilde{p}_{mix}(k; m; N)$ with $p_{\infty,pref}(k; m)$

For $q = 0.5$, with the rest of the parameters identical to Sec. 2.3, the Mixed model was run for different m's, with the resulting distribution plotted alongside Eq. (44) on Fig. 19. Note that the mixed model cuts off in between the two models, around $10^2 \sim 10^3$. Intuitively, this is expected as this model is a 'mix' of the two models.

The R^2 value fitted to different parts of the distribution is shown in Fig. 20, showing a good fit for the scale-free region.

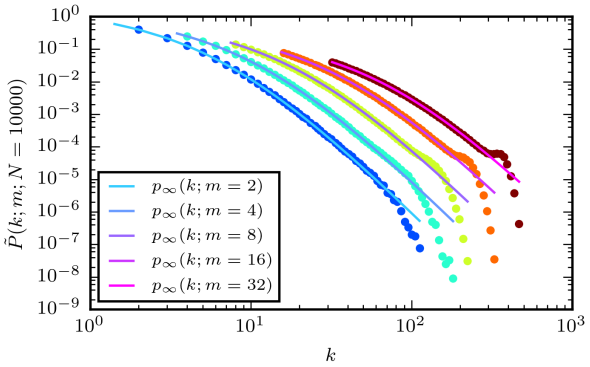


Figure 19: $\tilde{p}_{mix}(k; m; N; q = 0.5)$ *overlayed by* $p_{\infty,rand}(k; m; q = 0.5)$*, same parameters as before.* $p_{\infty,mix,disc}(k; m)$ *is a good fit until scaling region. Cut-off is in between PureRand and PurePref.*

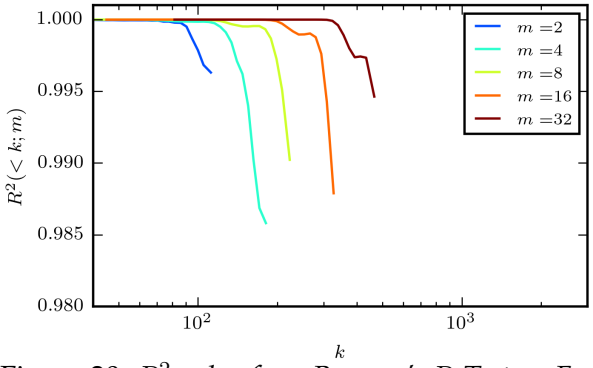


Figure 20: R^2 *value from Pearson's R-Test on Eq.* (20)*, test done on* $[m, k]$ *of dataset. Note a good R value followed by a drop when including cut-off points.*