Networks Project: Numerical analysis of the Barabási-Albert Model

Lingyi Hu
CID: 00919977

26th March, 2017

Abstract: We investigate the behaviour of the Barabási-Albert Model.

Word count: 2591 words in report (excluding front page, figure captions, table captions, acknowledgement and bibliography).

1 Pure preferential attachment

1.1 Degree distribution

1.1.1 Theory

The master equation that describes the evolution of the BA model is given by

$$n(k,t+1) = n(k,t) + m\Pi(k-1,t)n(k-1,t) - m\Pi(k,t)n(k,t) + \delta_{k,m} \tag{1}$$

where k is the total degree of a vertex, $n(k,t)$ is the number of nodes at time t with total degree k, and the probability Π for choosing the existing vertex depends on the model.

In the pure preferential attachment model, we choose an existing edge with $\Pi_{pa} \propto k$, which after normalizing gives $\Pi_{pa} = k/2E(t)$ where $E(t)$ is the number of edges and $2E(t)$ is the normalization constant corresponding to the total degree of the network. Assuming $E(0) = mN(0)$, the number of edges at a given time t is given by $E(t) = mN(t)$, so we get $\Pi = k/2mnN(t)$. Since we are concerned with the degree distribution of the model at large t, we consider the long-time ansatz $n(k,t) \to N(t)p_\infty(k)$.

Substituting these terms into the master equation, we obtain

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

It is clear that $p_\infty(k < m) = 0$, since m edges are added at every stage. So there are 2 cases to consider when solving for the above equation: $k = m$ and $k > m$.

We first consider the case when $k > m$. In this case, $\delta_{k,m} = 0$ and we can rearrange Equation 2 to get

$$\frac{p_\infty(k)}{p_\infty(k+1)} = \frac{k-1}{k+2} \tag{3}$$

To solve this equation, we can substitute in a trial solution of the form

$$f(z) = A\frac{\Gamma(z+1+a)}{\Gamma(z+1+b)} \tag{4}$$

where $\Gamma(z)$ is the Gamma function, which is an extension of the factorial function, with its argument shifted by one, to all real and complex nnumbers except the non-positive integers. Its central property is that

$$\Gamma(z+1) = z\Gamma(z),\ \Gamma(1) = 1. \tag{5}$$

Substituting the trial solution in Equation 4 gives

$$\frac{A\Gamma(z+1+a)}{\Gamma(z+1+b)} \times \frac{\Gamma(z+b)}{A\Gamma(z+a)} \tag{6}$$

which indeed simplifies to give $(z+a)/(z+b)$, using the the property in Equation 5 that $\Gamma(z+a+1)/\Gamma(z+a) = z+a$.

Substituting $a = -1$ and $b = 2$, we get the solution for Equation 3 in terms of A and the Gamma function:

$$p_\infty(k) = A\frac{\Gamma(k)}{\Gamma(k+3)} \tag{7}$$

which simplifies to

$$p_\infty(k) = \frac{A}{k(k+1)(k+2)}. \tag{8}$$

For the second case of $k = m$, Equation 2 becomes

$$p_\infty(m) = \frac{1}{2}[(m+1)p_\infty(m-1) - mp_\infty(m)] + 1. \tag{9}$$

However, we already know that $p_\infty(k < m) = 0$, that is, $p_\infty(m-1) = 0$. Using this, and rearranging Equation 9, we get

$$p_\infty(m) = \frac{2}{m+2}. \tag{10}$$

Substituting $k = m$ and Equation 10 into Equation 8, we get

$$\frac{A}{m(m+1)(m+2)} = \frac{2}{m+2}, \tag{11}$$

giving us the constant A as

$$A = 2m(m+1). \tag{12}$$

For this constant to be physically reasonable, we need to check that the probability satisfies normalization, that is, we need to prove

$$\sum_{k=m}^{\infty} p_\infty(k) = 2m(m+1)\sum_{k=m}^{\infty}\frac{1}{k(k+1)(k+2)} = 1. \tag{13}$$

The term in the summation of Equation 13 can be expanded as a partial fraction:

$$\sum_{k=m}^{\infty}\frac{1}{k(k+1)(k+2)} = \sum_{k=m}^{\infty}\frac{1}{2k} - \sum_{k=m}^{\infty}\frac{1}{k+1} + \sum_{k=m}^{\infty}\frac{1}{2(k+2)} \tag{14}$$

By writing out the first few terms of each summation, we can see that most terms cancel:

$$\begin{aligned}
&\tfrac{1}{2m} - \tfrac{1}{m+1} + \cancel{\tfrac{1}{2(m+2)}} \\
&\quad + \tfrac{1}{2(m+1)} - \cancel{\tfrac{1}{m+2}} + \cancel{\tfrac{1}{m+3}} \\
&\qquad + \cancel{\tfrac{1}{2(m+2)}} - \cancel{\tfrac{1}{m+3}} + \tfrac{1}{2(m+4)} \\
&\qquad\quad + \cancel{\tfrac{1}{2(m+3)}} - \tfrac{1}{m+4} + \tfrac{1}{2(m+5)} \\
&\qquad\qquad + \quad ...
\end{aligned} \tag{15}$$

and from the remaining terms we get the relation in Equation 13

$$\sum_{k=m}^{\infty} p_\infty(k) = 2m(m+1)\left(\frac{1}{2m} - \frac{1}{m} + \frac{1}{2(m+1)}\right) = 2m(m+1)\frac{1}{2m(m+1)} = 1 \tag{16}$$

Hence, we can confirm that the complete exact solution for the probability distribution in the long time limit is

$$p_\infty(k) = \frac{2m(m+1)}{k(k+1)(k+2)}. \tag{17}$$

1.1.2 Numerical analysis

To leverage its speed, `c++` code was used to generate graph data, while `python` was used for data analysis due to its wide range of data analysis tools, such as `numpy`, `scipy`, and `pandas`.

Computational efficiency of our algorithm is important considering that bigger datasets give better and more reliable statistical results. The following algorithm was used In this algorithm, the array M holds the list of edges represented by pairs of vertices, for example, the vertices at $M[0]$ and $M[1]$ are connected, $M[2]$ and $M[3]$ are connected, and so on. In this list, the number of occurences of a vertex is equal to its degree, so it can be used as a sample pool to achieve preferential attachment. To choose m neighbours for each new vertex, we then sample from M. This is also equivalent to choosing an edge at random and then choosing a vertex at random from the edge.

Algorithm 1 Algorithm for preferential attachment

Require: number of vertices N, minimum degree m ▷ $N > m$

Initialize our graph g
Initialize M as an empty array of length $2Nm$
for v in [0, ..., n-1] **do**
    `g.addVertex()` ▷ add new vertex to g
    for i in [0, ..., m-1] **do**
        $M[2(vm+i)] \leftarrow v$
        draw r uniformly at random
        from $[0, ..., 2(vm+i)]$ ▷ Choose random vertex from M
        $M[2(vm+i)+1] \leftarrow M[r]$ ▷ Add edge between vertex $M[r]$ and v

for i in [0, ..., nm-1] **do** ▷ Add all edges stored in M into the graph
    Add edge $(M[2i], M[2i+1])$ to graph g

Clearly this approach produces self loops and multiple edges, especially at the beginning before m vertices have been created. However, since we are concerned with the limit of $N \to \infty$, these effects are insignificant. Also, while unsatisfactory, multiple edges and self-loops do not affect our theoretical result. In the large N limit, the probability of getting multiple edges or self loops is small, and so for simplicity this algorithm was used without modification.

To check that the model was implemented correctly, the degree distribution generated by the model compared checked against the `networkx` implementation of the BA model. By using the same random seed, we could check that degree distribution between the two models were exactly the same. Graphs of fewer than 10 nodes were also generated to ensure that the points followed some basic constraints. This gives confidence that the algorithm is working as expected.

To investigate the degree distribution, the model was run for fixed N but varying m. Several numerical runs were performed for each set of values to improve statistics. It was found that generally above $N = 10^4$, finite size scaling effects are insignificant (see section 1.2.2). In order to reduce finite size effects, $N = 10^7$ was used.

Visually, as can be seen from ??, the numerical results seem to agree with the theoretical model after log-binning, although it understandably falls off near the end due to finite-sized effects. Alternatively, we can look at the complementary cumulative distribution function (ccdf), also known as the tail distribution, to observe the behaviour of the fat tail more clearly. This is shown in ??. We can see that as with the log binned results, the ccdf falls off near the tail, that is, it is lower than the expected theoretical value. This is expected since the probability distribution for our numerical simulations have to be normalized over a range of finite k.

1.2 Largest expected degree

1.2.1 Theory

The finite size of the system imposes a structural cutoff on the largest expected degree. For scale free networks, Aiello et al. (2001) defined the maximum degree to be approximately the value above which there is less than one vertex of that degree in the graph on average, that is, $N \sum_{k=k_1}^{\infty} p_\infty(k) = 1$.

Generally, it is shown (Boguñá, Pastor-Satorras, and Vespignani, 2004) that for a scale free network with $p_\infty(k) \propto k^\gamma$, the largest expected degree will be

$$k_1(N) \sim N^{1/(\gamma-1)}. \tag{18}$$

In our case, this can be easily verified. Starting with the equation

$$N \sum_{k=k_1}^{\infty} p_\infty(k) = 1, \tag{19}$$

we can see that this is almost identical to Equation 13, just factor and lower limit. Hence we have

$$2m(m+1)\frac{1}{2k_1(k_1+1)} = \frac{1}{N}. \tag{20}$$

We can then rearrange this to give us an expression for k_1:

$$k_1 = \frac{-1 + \sqrt{1 + 4Nm(m+1)}}{2} \tag{21}$$

1.2.2 Numerical analysis

2 Pure random attachment

2.1 Degree distribution

2.1.1 Theory

The pure random attachment model can be seen as a limiting case of the BA model. In this model, all existing vertices are chosen with equal probability, i.e. $\Pi = \Pi_{rnd} \propto 1$. This preserves growth but removes preferential attachment.

Similar to the previous section, we start from the master equation in Equation 1, and instead of $\Pi = k/2E(t)$ as in preferential attachment, we use $\Pi_{rnd} = 1/N(t)$. Again, we consider the long-time ansatz $n(k,t) \to N(t)p_\infty(k)$. Substituting these terms into Equation 1, we have

$$p_\infty(k) = mp_\infty(k-1) - mp_\infty(k) + \delta_{k,m}. \tag{22}$$

Considering the case of $k > m$, we obtain the recurrence relation

$$p_\infty(k) = \left(\frac{m}{m+1}\right) p_\infty(k-1) = ... = \left(\frac{m}{m+1}\right)^{k-m} p_\infty(m) \tag{23}$$

Now we consider $k = m$. Substituting $k = m$ into Equation 22 and remembering that $p_\infty(k < m) = 0$, we get

$$p_\infty(m) = -mp_\infty + 1, \tag{24}$$

giving us

$$p_\infty(m) = \frac{1}{m+1}. \tag{25}$$

Combining this result with Equation 23, we get the following formula for $p_\infty(k)$:

$$p_\infty(k) = \frac{1}{m+1}\left(\frac{m}{m+1}\right)^{k-m}. \tag{26}$$

For normalization, we need to check that

$$\sum_{k=m}^{\infty} p_\infty(k) = \frac{1}{m+1}\sum_{k=m}^{\infty}\left(\frac{m}{m+1}\right)^{k-m} = 1. \tag{27}$$

The terms in the summation form a converging geometric series, with the starting term being zero and common ratio being $m/(m+1)$. Hence we have

$$\sum_{k=m}^{\infty}\left(\frac{m}{m+1}\right)^{k-m} = \frac{1}{1-[m/(m+1)]}. \tag{28}$$

By substituting this back into the Equation 27, we can see that normalization is satisfied.

As we can see from Equation 26, the resulting degree distribution in this limit is geometric (Peköz, Röllin, and Ross, 2013), indicating that growth alone is not sufficient to produce a scale free structure.

2.1.2 Numerical analysis

Numerical simulations confirmed that growth alone is not sufficient to produce a scale free structure.

2.2 Largest expected degree

3 Random walks and preferential attachment

3.1 Theoretical degree distribution

One of the weaknesses of the BA model and its generalizations is that this implicitly requires a knowledge of the total degree and a calculation across existing vertices on the graph. This requirements then destroys the potential for this model to exhibit emergent properties based on local behaviour. The attachment by performing a random walk is a solution proposed by Saramäki and Kaski (2004). In this model, a vertex is chosen at random from existing vertices and then executes a random walk of length L from that vertex. The new vertex then attaches to the destination vertex.

This model was thought to be able to reproduce the BA degree distribution even for $L = 1$ (Saramäki and Kaski, 2004; J.P.Saramaki and T.S.Evans, 2004). While this is the case for large L, Cannings and Jordan (2013) later showed that the $L = 1$ degree sequence converges to a degenerate limiting solution in which almost every vertex has degree 1, instead of a power law distribution, and demonstrated that this model is fundamentally different from the BA model, unless we allow an indefinite length for the random walk.

References

Aiello, William et al. (2001). "A Random Graph Model for Power Law Graphs". In: *Experimental Mathematics* 10.1, pp. 53–66. ISSN: 1058-6458. DOI: 10.1080/10586458.2001.10504428. URL: http://www.tandfonline.com/doi/abs/10.1080/10586458.2001.10504428.

Boguñá, M., R. Pastor-Satorras, and A. Vespignani (2004). "Cut-offs and finite size effects in scale-free networks". In: *European Physical Journal B* 38.2, pp. 205–209. ISSN: 14346028. DOI: 10.1140/epjb/e2004-00038-8. arXiv: 0311650 [cond-mat].

Cannings, Chris and Jonathan Jordan (2013). "Random walk attachment graphs". In: *Electronic Communications in Probability* 18, pp. 1–8. ISSN: 1083589X. DOI: 10.1214/ECP.v18-2518. arXiv: 1303.1052.

J.P.Saramaki and T.S.Evans (2004). "Scale Free Networks from Self-Organisation". In: *Physical Review E* 72.2, pp. 1–33. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.72.026138. arXiv: 0411390v2 [cond-mat]. URL: http://link.aps.org/doi/10.1103/PhysRevE.72.026138.

Peköz, Erol A., Adrian Röllin, and Nathan Ross (2013). "Total variation error bounds for geometric approximation". In: *Bernoulli* 19.2, pp. 610–632. ISSN: 1350-7265. DOI: 10.3150/11-BEJ406. URL: http://projecteuclid.org/euclid.bj/1363192040.

Saramäki, Jari and Kimmo Kaski (2004). "Scale-free networks generated by random walkers". In: *Physica A: Statistical Mechanics and its Applications* 341.1-4, pp. 80–86. ISSN: 03784371. DOI: 10.1016/j.physa.2004.04.110. arXiv: 0404088 [cond-mat].