Fitting and Model Checking a Linear Preferential Attachment Model for Directed Graphs

Ting-Yuan Kuo

26th April, 2018

Linear Preferential Attachment (Linear PA)

- Motivated to provide an explaination to the phenomenom of scale-free networks ($\iff$ power law), first proposed by Barabási and Albert (1999)
- Belongs to the class Network Growth Models
- Will generate "rich-get-richer" effect
- Gets a power law asympototically for its degree distribution.

Canonical Linear PA Model: Barabasi-Albert (1999)

- A random growth model for undirected, unweighted graphs
- Let there be an initial graph, $G(t_0)$, with node size m_0
- Add a node at each time-step t to $G(t)$ and connect the newly added node to $m(\leq m_0)$ nodes present in $G(t-1)$ by

$$\mathbb{P}[\text{choose } v \in G(t-1)] = \frac{k_v}{\sum_j k_j}$$

  where k_v is the degree of node $v \in G(t-1)$
- Linear in the sense that

$$\frac{{k_v}^{\alpha}}{\sum_j {k_j}^{\alpha}}, \ \alpha = 1$$

Power Law

- Bollobás et al. (2001) shows that as $t \to \infty$

$$f_d \propto d^{-\gamma}, \ \gamma = 3$$

  where f_d is the number of nodes with degree d, iff $\alpha = 1$
- In the limit f_d has a power law exponent of 3
  - If $\alpha < 1$, we get stretched exponential; if $\alpha > 1$, a single node connects to nearly all other nodes (Krapivsky, Redner, and Leyvraz 2000).
- Likely too simple for empirical data, so I will fit another linear PA model for directed graphs

Goal

1) To fit a linear PA model for directed graphs based only on a snap-shot of the graph and 2) check whether the model is a good fit in terms of the degree distribution

- Important since we often cannot observe the full history of the graph
- Most empirical work on testing for power law of the distribution rely on directly testing the resulting degree distribution
  - Hard to fit fat-tailed distributions (Broido and Clauset 2018)
  - Hard to distinguish between fat-tailed distributions, more rigorous method is by likelihood ratio tests, but seldom done
- By directly fitting the model, we can do predictive checking and provide another tool kit to test the power law hypothesis

The Linear PA Model for Directed Graphs

Notation: Let $D_{in}^{(n)}(u)$ and $D_{out}^{(n)}(u)$ denote the in- and out-degree of node u in $G(n)$, respectively.

1. At each time-step n, toss an unfair three-sided coin J_n with $\Omega = \{1, 2, 3\}$ and the mass function, $\mathbb{P}(J_n = 1) = \alpha,\ \mathbb{P}(J_n = 2) = \beta,\ \mathbb{P}(J_n = 3) = \gamma.$ Assume $0 < \alpha,\ \beta,\ \gamma < 1.$
2. If $J_n = 1$ (α-scheme): Add a new node, v, to $G(n-1)$ and an edge (v, w) leading from v to a previously existing $w \in V(n-1)$. Choose w by,

$$\mathbb{P}[\text{ choose } w \in V(n-1)] = \frac{D_{in}^{(n-1)}(w) + \delta_{in}}{n - 1 + \delta_{in} N(n-1)}$$

   That is choose w with the probability proportional to its in-degree and corrected by a bias parameter δ_{in}

Cont.

3. If $J_n = 2$ (β-scheme): Add a directed edge (v, w) to $E(n-1)$ where $v, w \in V(n-1)$ (no new node is added). Choose (v, w) as such,

$$\mathbb{P}[\text{choose } (v,w)] = \Big(\frac{D_{in}^{(n-1)}(v) + \delta_{in}}{n - 1 + \delta_{in} N(n-1)}\Big)\Big(\frac{D_{out}^{(n-1)}(w) + \delta_{out}}{n - 1 + \delta_{out} N(n-1)}\Big)$$

4. If $J_n = 3$ (γ-scheme): Add a new node w to $G(n-1)$ and an edge (v, w) leading from an existing node v to w with the probability,

$$\mathbb{P}[\text{ choose } v \in V(n-1)] = \frac{D_{out}^{(n-1)}(v) + \delta_{out}}{n - 1 + \delta_{out} N(n-1)}$$

- Note: this is a 5 parameter model with $\theta = (\alpha, \beta, \gamma, \delta_{in}, \delta_{out})$

Power Law for this model

- Bollobás et al. (2003) showed that for this model the in- and out-degree distribution also has power law

Roughly, let p_i and q_i be the in and out degree distribution respectively (i denotes the degree count), then in the limit,

$$p_i \propto i^{\kappa_{in}}, \quad \text{if } \alpha\delta_{in} + \gamma > 0$$

$$q_i \propto i^{\kappa_{out}}, \quad \text{if } \gamma\delta_{out} + \alpha > 0$$

Where,

$$\kappa_{in} = 1 + \frac{1 + \delta_{in}(\alpha + \gamma)}{\alpha + \beta}$$

$$\kappa_{out} = 1 + \frac{1 + \delta_{out}(\alpha + \gamma)}{\beta + \gamma}$$

Estimation, Inference, and Simulation

- Most work on estimating network growth models rely on having the full history the graph, $\{G(t)\}_{t=t_0}^{m}$. But in most practical circumstances we can only observe some snap-shot $G(t^*)$
- Wan et al. (2017) proposes an approximate MLE estimator $\tilde{\theta}$ for θ that is strongly consistent (i.e. $\tilde{\theta} \overset{a.s.}{\to} \theta$ as $n \to \infty$). See p.13-14 of their paper for the algorithm. They also came up with a fast simulation algorithm.
- No formal inference procedure, instead suggested an ad-hoc boostrapping procedure to compute the sample variance of $\tilde{\theta}$ based on repeated indepedent simulations. Will not be doing this due to computational constraints.

Picture

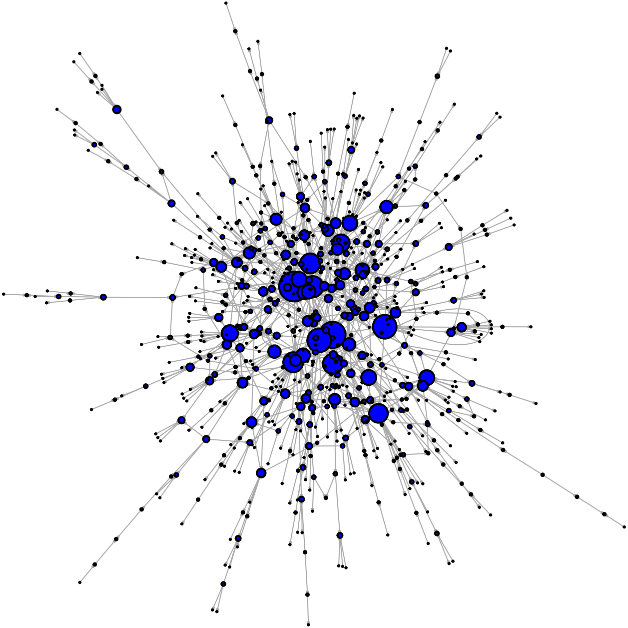

Illustration: Bitcoin Network Data

- Downloaded from the Stanford SNAP website.
- Directed and temporal. Has 35,592 edges, 5881 nodes.
- Transaction data: edge (v, w) means v sold to w. (w, v) for vice versa
- I will pool together all data to pretend that we only have one snap-shot (i.e. only has the final adj. matrix and do not know the true edge/node permutation)

Estimation and Simulation

- Fitting the network, we get

$$\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}_{in}, \tilde{\delta}_{out}) = (0.0029,\ 0.8348,\ 0.1623,\ 1.4834,\ 3.9572)$$

$$\implies \tilde{\kappa}^{in} = 2.4862, \qquad \tilde{\kappa}^{out} = 2.6588$$

- Then use $\tilde{\theta}$ to simulate a network, $G(t)_{sim}$, with 35,592 edges (matching data) - we get 5873 nodes! (recall real network has 5881 nodes)

Sanity Check for the Estimates

- $\tilde{\theta}_{sim}$ of $G(t)_{sim}$ is (0.0033, 0.8350, 0.1617, 1.6574, 4.1046)
- Fitting the simulated degree distribution using a power law MLE method proposed by Clauset, Shalizi, and Newman (2007),
$$\hat{\kappa}^{in}_{sim} = 2.4140 \quad \hat{\kappa}^{out}_{sim} = 2.3887$$
with 95% CI, $(2.2131,\ 2.6149)$ and $(2.1758,\ 2.6016)$
  - We know for sure, $\kappa^{in}_{sim} = 2.4862$ and $\kappa^{out}_{sim} = 2.6588$
- In terms of goodness of fit, we can accept the null hypothesis that both the in and out-degree distribution of the simulated network follows power law, $p = 0.14$ and $p = 0.15$, respectively

Back to Bitcoin Data

- Recall for the Bitcoin network, our parametric model says that

$$\tilde{\kappa}^{in} = 2.4862 \qquad \tilde{\kappa}^{out} = 2.6588$$

- On the other hand, by the Clauset, Shalizi, and Newman (2007) power law MLE method on the empirical distribution

$$\hat{\kappa}^{in} = 2.2708 \qquad \hat{\kappa}^{out} = 2.0594$$

  with 95% CI, $(2.1141,\ 2.4275)$ and $(1.9847,\ 2.1341)$. Goodness of fit says we reject the null hypothesis that the in and out-degree comes from power law, $p = 0.02$ and $p < 10^{-6}$.

$\implies$ Some disagreement. By direct MLE estimation of the degree distributions, real data has much fatter tails

Predictive Checking: Degree Distribution

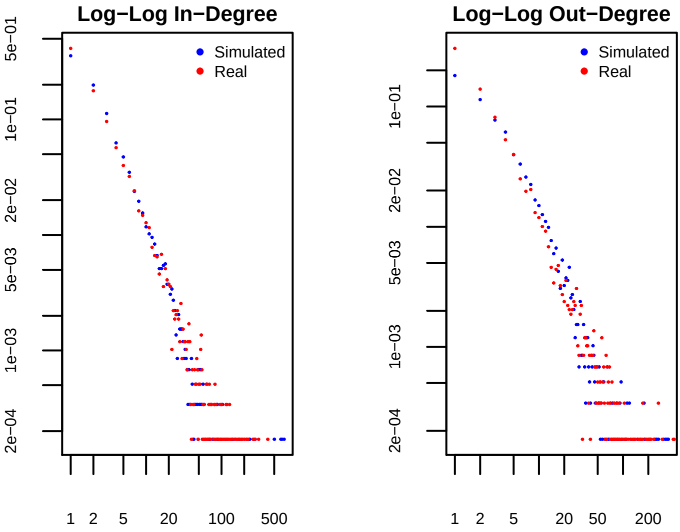

Figure 2: Degree Distribution of the Real and Simulated Network

Predictive Checking: Kolmogorov-Sminrov (KS) Test

- To quantify the distance between the empirical and simualted degree distribution, we can use the KS statistic, defined as

$$D = \sup_x |E(x) - S(x)|$$

  Here, $E(x)$ is the empirical CDF and $S(x)$ is the CDF for the simulated distribution. We test,

$$\mathcal{H}_0 : E(x) = S(x), \forall x \qquad \mathcal{H}_1 : E(x) \neq S(x)$$

  To reject $\mathcal{H}_0$, we need about $D > 0.025$.
- KS test shows that, $D^{out} \approx 0.107$ and $D^{in} \approx 0.051$ for the out and in-degree respectively

$\implies$ **Not a good fit**

Joint In-Out Degree Distribution

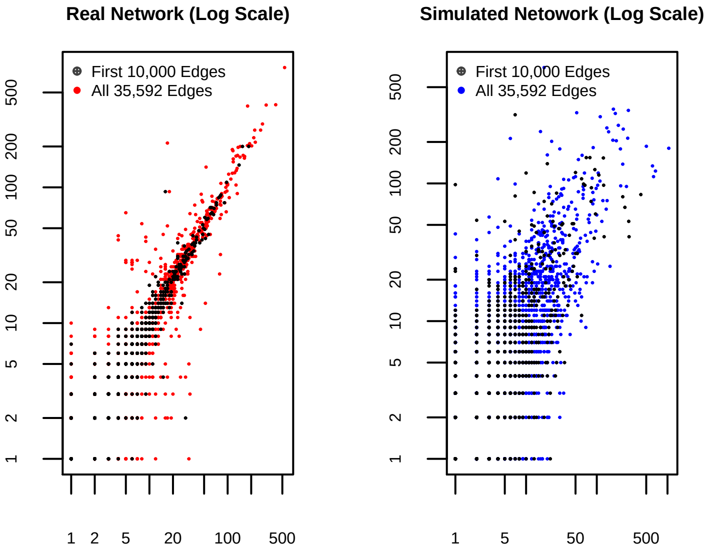

Figure 3: In vs. Out Degree of the Real and Simualted Network

Joint In-Out Degree Distribution

- Much higher correlation between in and out degree in the real network
  - $\implies$ Using some notion of total degree for the PA mechanism?
- High in and out degree correlation early on in the real network
  - $\implies$ Accelerated dynamics? Latent factors (covariates) that dominate the effects of staying longer in the network?
- Higher concentration of in and out degree in high degree regions for the real data
  - $\implies$ Non-linear attachment?
    - For example, in the Barabasi-Albert model, if the attachment exponent $\alpha > 1$, then we get a winner takes all situation (Krapivsky, Redner, and Leyvraz 2000)
  - Confirms the observation of fatter tails for real data

Heatmap of the Adjacency Matrix

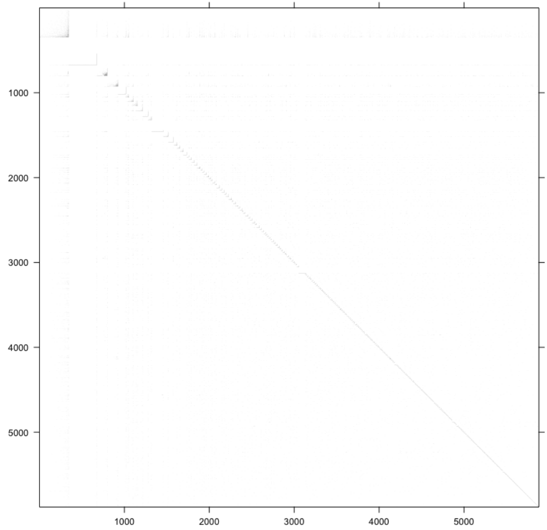

Figure 4: Heatmap for the real network

Heatmap of the Adjacency Matrix

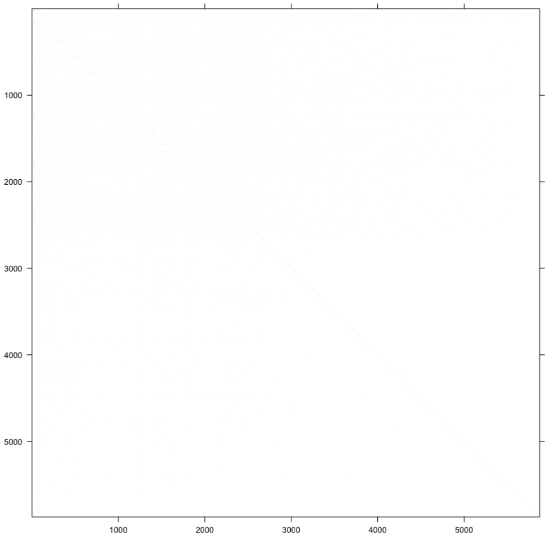

Figure 5: Heatmap for the simulated network

Next Steps

- Clustering Coefficients
- Connectivity
- Dynamics
- More sample of simulated network (if time allows)

Reference I

Barabási, Albert László, and Réka Albert. 1999. "Emergence of scaling in random networks." *Science* 286 (5439): 509–12. doi:10.1126/science.286.5439.509.

Bollobás, B, C Borgs, J Chayes, and O Riordan. 2003. "Directed scale-free graphs." *Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms*, 132–39.

Bollobás, B, Oliver Riordan, Joel Spencer, and Gabor Tusnady. 2001. "Random Graph Process." *Random Structures & Algorithms* 18 (3): 279–90.

Broido, Anna D., and Aaron Clauset. 2018. "Scale-free networks are rare," 26–28. `http://arxiv.org/abs/1801.03400`.

Clauset, Aaron, C Rohilla Shalizi, and M E J Newman. 2007. "Power-law distributions in empirical data." *ArXiv Preprint ArXiv:0706.1062* 64 (4): 661–703.

Reference II

https://epubs.siam.org/doi/pdf/10.1137/070710111.

Krapivsky, P. L., S. Redner, and F. Leyvraz. 2000. "Connectivity of growing random networks." *Physical Review Letters* 85 (21): 4629–32. doi:10.1103/PhysRevLett.85.4629.

Wan, Phyllis, Tiandong Wang, Richard A. Davis, and Sidney I. Resnick. 2017. "Fitting the linear preferential attachment model." *Electronic Journal of Statistics* 11 (2): 3738–80. doi:10.1214/17-EJS1327.