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

Jimmy Ting-Yuan Kuo[*]

Abstract

Using the estimation method developed in Wan et al. (2017), I fitted a 5-parameter linear preferential attachment model, where power law for both the in- and out-degree distribution is generated in the limit, for directed graphs on a bitcoin trust-network dataset using only one snapshot observation of the network process. By comparing the real network with the simulated network, I performed several modeling checking and predictive checking procedures to assess how well the model describes real data. I compared the degree distribution, in-out degree correlation, age versus degree relationship, parameter dynamics, clustering coefficient, and assortativity coefficient of the real and simulated network, and found that 1) the bitcoin trust-network is not scale-free; 2) nodes in the real network have a significantly higher correlation in its in- and out- degree; 3) contrary to preferential attachment, the positive age-degree relationship is much weaker in the real network; 4) the out-degree is especially poorly modeled; and 5) there is a greater degree of assortative behavior in the simulated network, whereas the real network is dis-assortative. Further analysis suggests that preferential attachment is the incorrect mechanism generating the empirical data. Kondor et al. (2014) studied properties of the bitcoin transaction level network, this study complements Kondor et al. (2014) with a detailed analysis on the trust-network level data with a model driven approach.

1 Introduction

A well-discussed property thought to be common among the degree distribution of networks is that it obeys a power law distribution. Namely, let $f(d)$ be the fraction of nodes with degree d, then

$$f(d) \propto d^{-\alpha}, \quad \text{for } d \geq d_{\min}$$

$\alpha > 1$ is said to be the power law exponent. The distribution has infinite first moment when $1 < \alpha < 2$, infinite second-moment when $2 < \alpha < 3$. It is dubbed interchangeably between networks which have a power law degree distribution and networks which are "scale-free" since power law is the only functional form which satifies scale-invariance,

$$f(cd) \propto (cd)^{-\alpha} = c^{-\alpha} d^{-\alpha} \propto f(d)$$

for some constant c. In order to give some theoretical underpinning to the seemingly wide-spread phenomena of scale-free networks, Barabási and Albert (1999) proposed the following simple canonical linear preferential attachment (PA) model for undirected, unweighted networks. Let $\{G(t)\}_{t=t_0}^{T}$ be a sequence of graphs indexed by time t. Let $G(t_0)$ be the initial given graph with m_0 nodes. Then at each $t > t_0$, generate $G(t)$ from $G(t-1)$ by adding a new node, v, and form an edge between v and each of the $m(\leq m_0)$ existing nodes, with the probability proportional to the node degree,

$$\mathbb{P}[\text{ choose } w \in G(t-1)] = \frac{{k_w}^{\alpha}}{\sum_j {k_j}^{\alpha}}, \quad \alpha = 1$$

where k_w is the degree of node w. It is linear because $\alpha = 1$. Such linear PA mechanism would generate the "rich-get-richer" effect, and as rigorously proved in Bollobás et al. (2001), in the limit, $f(d) \sim Cd^{-\alpha}$ with $\alpha = 3$. However, Krapivsky, Redner, and Leyvraz (2000) has shown that under this model, the network is only scale-free if $\alpha = 1$: if $\alpha < 1$ then the limiting degree distribution is streched exponential; if $\alpha > 1$ we would get winner take all - one node that connects to nearly all other nodes. Many growth models have been subsequently proposed which generate scale-freeness in the limit, some use a preferential attachment

[*] Department of Mathematics and Statistics, Boston University. Email: jimmykuo@bu.edu

mechanism and some not, and often require model parameters to fall on some range or exact point, see Table 3 of R. Albert and Barabasi (2001) for a summary of analytical results.

Most work on testing for the scale-freeness of networks have been exclusively on determining whether the resulting degree distribution obeys power law, not fitting the hypothesized growth model. However, as Broido and Clauset (2018) have pointed out, many such works rely on less rigorous statistical methods as fitting fat-tailed distributions is difficult, and often used small sample datasets. Moreover, likelihood ratio tests are seldem done to compare power law to other fat-tailed distributions. This is crucial as illustrated by the Barabasi-Albert-type model, streched exponential and power law are all possible limiting distributions and the parameter requirement for scale-freeness is exact.

Besides the statistical concerns, stricly analyzing the resulting degree distribution is fundementally not a network problem. In order to understand the network processes which could generate scale-freeness, it would be of great benefit to directly fit the underlying process to provide another tool-kit to analyze the formation of degree distribution. However, most work on fitting growth models rely on having the entire history of the graph process, relatively few (Wiuf et al. 2006, Bezáková, Kalai, and Santhanam (2006), Bloem-Reddy and Orbanz (2016), Guetz and Holmes (2011), Leskovec et al. (2010), Wan et al. (2017)) developed methods to fit growth models based on a single snapshot, which significatly limited the number of datasets that researchers can analyze. Therefore, the goal of this paper is to fit a network growth model based on a single snapshot of the graph and conduct simulation and predictive checking studies to see how well the model explains the empirical degree distribution of the network, as well as other network characteristics.

The model I will be fitting is a 5-parameter linear preferential attachment model for directed, unweighted network, described below.

2 The Model

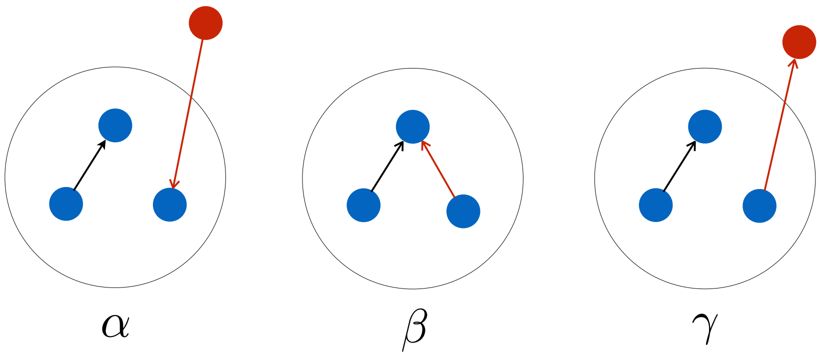

Figure 1: Edge/Node-Addition Schemes

The model is a random linear PA model for directed graphs as described in Bollobás et al. (2003), with the parameter vector with all positive elements $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta_{in}, \delta_{out})$. As usual, $G(n) = (V(n), E(n))$, where $V(n)$ denotes the set of nodes at time n and $E(n)$ denotes the set of edges at time n.

For this linear PA model, at $n+1$ an *edge* is added to $G(n)$ to form $G(n+1)$. Following the notation in Wan et al. (2017), let $n \equiv |E(n)|$, where $|\cdot|$ denotes the cardinality, and let $N(n) \equiv |V(n)|$. For every $u \in V(n)$, $D_{in}^{(n)}(u)$ and $D_{out}^{(n)}(u)$ denotes the in- and out-degree of u in $G(n)$, respectively. Let (v, w) denote a directed edge from v to w. We assume that there is a given initial finite directed graph $G(n_0)$ with at least one node and n_0 edges. For all $n > n_0$ and given $G(n-1)$, generate $G(n)$ as follows,

1. Toss a three-sided coin J_n with $\Omega = \{1, 2, 3\}$ and the following mass function, $\mathbb{P}(J_n = 1) = \alpha$, $\mathbb{P}(J_n = 2) = \beta$, $\mathbb{P}(J_n = 3) = \gamma$. Assume $0 < \alpha, \ \beta, \ \gamma < 1$ to avoid degeneracy. $\{J_n\}$ then forms a multinomial process.

- 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 $w \in V(n-1)$. The choice of w is based on the following probability,

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

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

- 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,

$$\mathbb{P}[\text{choose } (v, w)] = \Big(\frac{D_{in}^{(n-1)}(v) + \delta_{in}}{n - 1 + \delta_{in} N(n-1)} \Big) \Big(\frac{D_{out}^{(n-1)}(w) + \delta_{out}}{n - 1 + \delta_{out} N(n-1)} \Big) \tag{2}$$

  In other words, v and w are being chosen independently and the probability of w being chosen is proportional to its in-degree and the probability of v being chosen is proportional to its out-degree.

- 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,

$$\mathbb{P}[\text{ choose } v \in V(n-1)] = \frac{D_{out}^{(n-1)}(v) + \delta_{out}}{n - 1 + \delta_{out} N(n-1)} \tag{3}$$

Figure 1 shows the edge-addition schemes based on the coin toss. In summary: in the α-scheme, add a new node and directs it to an existing node where the existing node is being chosen proportional to its in-degree; in the β-scheme, no new node is added, but add a new edge between two existing node with the probability as being proportional to the product of their in and out-degrees; in the γ-scheme, add a new node, and direct an existing node to the new node added, where the existing node is being chosen proportional to its out-degree.

2.1 Power law as the limiting degree distribution

Bollobás et al. (2003) studied the limiting in- and out-degree distribution of this model and showed that the marginal in and out-degree distribution of the graph has the power law property in the following way. Let $x_i(n)$ denote the number of nodes in $G(n)$ with in-degree i, so $x_i(n)/N(n)$ is the fraction of nodes with in-degree i at time step n. Similarly, let $y_i(n)$ denote the number of nodes in $G(n)$ with out-degree i, and write $y_i(n)/N(n)$. Then by Theorem 3.1 of Bollobás et al. (2003), there exists constants p_i, q_i for fixed $i \geq 1$ such that as $n \to \infty$, almost surely

$$\frac{x_i(n)}{N(n)} \to p_i, \quad \frac{y_j(n)}{N(n)} \to q_i \tag{4}$$

See equation 3.10 of Bollobás et al. (2003) for the closed form solution of p_i and q_i. Then taking the limit as $i \to \infty$,

$$p_i \sim C_1 i^{\kappa_{in}} \quad \text{if } \alpha\delta_{in} + \gamma > 0, \quad \kappa_{in} = 1 + \frac{1 + \delta_{in}(\alpha + \gamma)}{\alpha + \beta} \tag{5}$$

$$q_i \sim C_2 i^{\kappa_{out}} \quad \text{if } \gamma\delta_{out} + \alpha > 0, \quad \kappa_{out} = 1 + \frac{1 + \delta_{out}(\alpha + \gamma)}{\beta + \gamma} \tag{6}$$

where C_1, C_2 are positive constants and $f_i \sim g_i$ denotes $f_i/g_i \to 1$ as $i \to \infty$.

2.2 Parameter Estimation, Inference, and Simulation

Wan et al. (2017) has proposed a MLE approximation procedure to estimate $\boldsymbol{\theta}$ just based on a single snapshot of the graph. Let $X_{>i}(n) = \sum_{j>i} x_j(n)$ and $Y_{>i}(n) = \sum_{j>i} y_j(n)$ denote the number of nodes with in-degree greater than i in $G(n)$ and the number of nodes with out-degree greater than j in $G(n)$, respectively. Let $\tilde{\boldsymbol{\theta}}$ be the estimator for $\boldsymbol{\theta}$. Then Wan et al. (2017) p.13-14 proposed a simple algorithm to compute $\tilde{\boldsymbol{\theta}}$ where $\tilde{\boldsymbol{\theta}}$ is strongly consistent, i.e. $\tilde{\boldsymbol{\theta}} \overset{a.s.}{\to} \boldsymbol{\theta}$ element-wise. I won't outline the complete derivation here, but the basic idea is that, as they have shown, $\{X_{>i}(n)\}_i$, $\{Y_{>i}(n)\}_i$, and $\{J_n\}_n$ (the multinomial coin process) are sufficient statistics for $\boldsymbol{\theta}$, but $\{J_n\}_n$ is unknown from a single snapshot. However, $\boldsymbol{\theta}$ can in fact be well-approximated by functions of $\{X_{>i}(n)\}_i$ and $\{Y_{>i}(n)\}_i$ only, if n is large. The result is that $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ can be computed in closed-form, with $\tilde{\beta} = 1 - N(n)/n$ which intuitively makes sense, and obtaining $\tilde{\delta}_{in}$, $\tilde{\delta}_{out}$ each requires numerically solving an implicit function. The authors suggested a procedure to re-normalize $\tilde{\boldsymbol{\theta}}$ so that $\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma} = 1$. It should be noted that from my experience, without the re-normalization procedure the parameters are not guaranteed to be strictly positive, as assumed by the model.

As for inference, there are no formal inference procedure propsed, instead the authors suggested the following bootstrapping procedure to estimate the variance of $\tilde{\boldsymbol{\theta}}$. Use $\tilde{\boldsymbol{\theta}}$ to simulate 10^4 independent bootstrap replicates of network with $n = 10^5$ edges. For each simulated network, compute the snapshot esimate $\tilde{\boldsymbol{\theta}}^*_{\boldsymbol{n}} \equiv (\tilde{\alpha}^*, \tilde{\beta}^*, \tilde{\delta}^*_{in}, \tilde{\delta}^*_{out})$ and take the sample variance, $\hat{\text{Var}}(\tilde{\boldsymbol{\theta}}^*_{\boldsymbol{n}})$, of $\tilde{\boldsymbol{\theta}}^*_{\boldsymbol{n}}$ over the 10^4 snapshot estimates to approximate $\text{Var}(\tilde{\boldsymbol{\theta}})$. Then by assuming asymtotic normality, construct the two-sided confidence interval of $(1 - \epsilon)$ as usual,

$$\{\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}\}_i \pm z_{\epsilon/2}\sqrt{\hat{\text{Var}}(\{\tilde{\boldsymbol{\theta}}^*_{\boldsymbol{n}}\}_i)}, \quad i = 1, 2, 3, 4$$

Where $z_{\epsilon/2}$ is the upper $\epsilon/2$ quantile of the standard normal.

Wan et al. (2017) also proposed an efficient simulation algorithm for such linear AP network on p.5 where the cost of simulation is $O(n)$. When choosing which node to connect to in the existing graph, naively sampling with the multi-nomial distribution would require $O(N(n))$ evaluations, and $N(n)$ increases linearly with n by construction, so the total cost of sampling would be $O(n^2)$. Their algorithm utilizes a trick with sampling once from the uniform distribution, instead of the multi-nomial distribution, therefore significantly decreases the computational cost.

In this model, since we get scale-freeness so long as all paramters are positive, which is not a very stringent test, we can only test for scale-freeness based on predictive checking methods and see how the properties of the simulated graph line up with the empirical one.

3 Illustration: Bitcoin Network

3.1 Data

The data is bitcoin data `bitcoin-otc` download from the Stanford SNAP website[1], compiled by Kumar et al. (2017). It is a who-trusts-whom network of people who trade using Bitcoin on a platform called Bitcoin

[1] http://snap.stanford.edu/data/soc-sign-bitcoinotc.html

OTC. Since there isn't a centralized credit rating system in the Bitcoin market, there is a need for a P2P trust-rating system in detecting fraudulent and risky users. Each each edge (v, w) represents a rating from v to w on w's perceived trust-worthiness. A rating can range from -10 (total distrut) to 10 (total trust). Since a rating likely results in a trade, this who-trusts-whom network can also be viewed as a proxy for transaction network data. I won't be using the ratings in the analysis and instead will focus on the network structure.

The data is originally temporal, with a timestamp on edge appearance. It has 5,881 nodes and 35,592 edges. I will pool all data to form a single snapshot. It should be noted that the linear PA model here considered allows for self-loops and multiple-edges (from equation 2), but this is not allowed/non-existent for bitcoin data. This will be further discussed in section (3.2.4).

Kondor et al. (2014) also studied properties of the bitcoin network, but directly on transaction level data. This study seeks to complement Kondor et al. (2014) with a detailed analysis on the trust-network level data of bitcoin with a model driven approach.

3.2 Result

Fitting the model on bitcoin data, we get

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}_{in}, \tilde{\delta}_{out}) = (0.0029,\ 0.8348,\ 0.1623,\ 1.4834,\ 3.9572) \\ &\implies \tilde{\kappa}^{in} = 2.4862 \qquad \tilde{\kappa}^{out} = 2.6588 \end{aligned} \tag{7}$$

I then simulate a network, G_{sim}, with 35,592 edges to match the data using $\tilde{\boldsymbol{\theta}}$, and obtained 5,873 nodes, which is very close to the real data. Due to computational constraints, I won't be doing the boostraping inference procedure.

3.2.1 Sanity Checks

To make sure that the estimation and simulation procedure works well, I will perform some sanity checks on G_{sim}. Estimating $\boldsymbol{\theta}_{sim}$, we get

$$\tilde{\boldsymbol{\theta}}_{sim} = (0.0033, 0.8350, 0.1617, 1.6574, 4.1046)$$

Which lines up with true parameter values. Moreover, we also know for sure that $\kappa_{sim}^{in} = 2.4862$ and $\kappa_{sim}^{out} = 2.6588$. We can then directly fit the resulting degree distribution of G_{sim} to power law and see if the estimates line up with the true $\kappa_{in},\ \kappa_{out}$ values. Using state-of-the-art MLE discrete power law method developed by Clauset, Shalizi, and Newman (2009) implemented via the package `poweRlaw`, which estimates α by selecting the $\hat{\alpha}^{MLE}$ (a function of x_{min}) that corresponds to the $\hat{x}_{min}$ value that minimizes the following Kolmogorov-Smirnov distance,

$$D = \max_{x \geq x_{min}} |S(x) - P(x)|$$

where $S(x)$ is the empirical CDF when the observations are at least x_{min}, and $P(x)$ is the power law CDF that best fits the data in $x \geq x_{min}$, we get that

$$\hat{\kappa}_{sim}^{in} = 2.4140 \qquad \hat{\kappa}_{sim}^{out} = 2.3887$$

with 95% CI, $\big(2.2131,\ 2.6149\big)$ and $\big(2.1758,\ 2.6016\big)$, generated by boostrapping. Hence, the MLE power law estimates of κ^{out} is off, this points to some of the difficulty in analyzing power law data. In terms of goodness of fit, generated via boostraping as suggested by Clauset, Shalizi, and Newman (2009), we can accept the null hypothesis that both the in and out-degree distribution of the simulated network follows power law, with $p = 0.14$ and $p = 0.15$, respectively. This means that the graph size is large enough so that the graph is approximately scale-free.

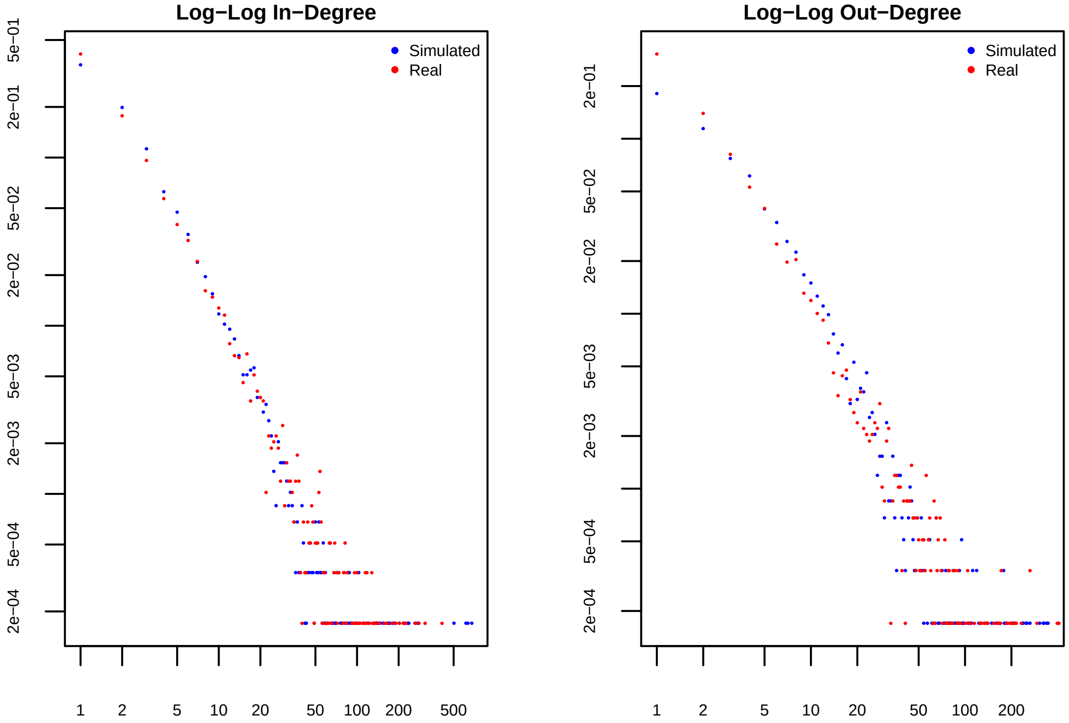

Figure 2: Degree Distribution of the Real and Simulated Network

3.2.2 Degree Distribution

I now turn to comparing in the difference between the degree distributions of the simulated and real network. Figure 2 compares the in and out-degree distribution of the real and simulated network. To quantify the distance between the real and simulated distribution, we use the Kolmogorov-Smirnov test,

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

Now $E(x)$ is the empirical CDF of the degree distribution and $S(x)$ is the CDF of the degree distribution of the simulated network. For hypothesis testing, we have

$$\mathcal{H}_0 = \text{ Data comes from the simulated degree distribution}$$

$$\mathcal{H}_1 = \text{ Data does not come from the simulated degree distribution}$$

KS test shows that, $D^{in} \approx 0.051$ and $D^{out} \approx 0.107$ for the in and out-degree respectively, with p-values, $p^{in} < 10^{-6}$ and $p^{out} < 10^{-15}$ so we reject the null hypothesis for both degree distributions. If the min and/or max value of the simulated and/or empirical degree distribution is removed, KS test statistic does not change by much. This indicates that the model is not a great fit.

In terms of directly fitting the real degree distribution to power law using the Clauset, Shalizi, and Newman (2009) method, we get that

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

with 95% CI, $\big(2.1141,\ 2.4275\big)$ and $\big(1.9847,\ 2.1341\big)$ for in and out-degree respectively. Recall that our parametric model says,

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

hence there is disagreement between the power law MLE method and network model fitting, with κ_{out} being especially off. Goodness of fit indicates that the null (in and out-degree comes from power law) is rejected,

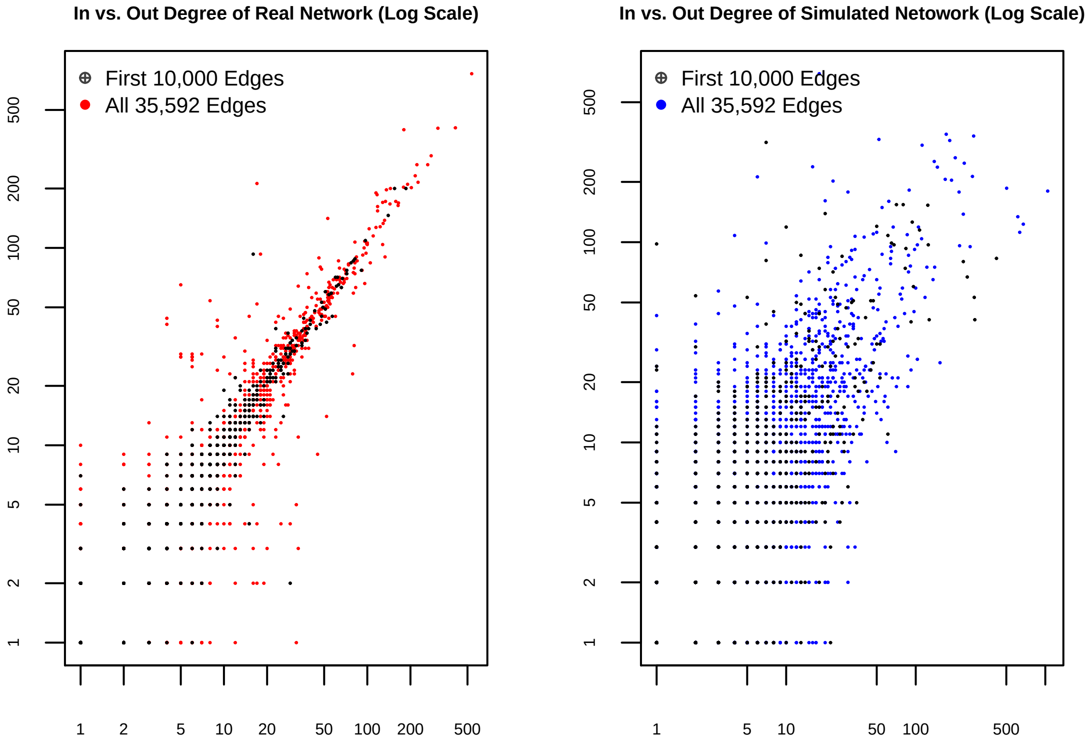

Figure 3: In-Degree (y-axis) vs. Out-Degree (x-axis) of the Real and Simualted Network

with p-values, $p = 0.02$ and $p < 10^{-6}$. Judging from the power law exponent MLE estimation, the smaller power law exponents in the real network indicate that the real network has fatter tails.

As for likelihood ratio tests, using the method developed in Vuong (1989), a comparison with the log-normal distribution for the out-degree yields Likelihood $\text{Ratio}_{out} = -3.028$, with $p = 0.002$. Here the null is that both distributions are equally far off the true distribution and a negative-signed LR favors the alternative distribution. Hence testing against log-normal, data once again does not support power law. As for the in-degree, Likelihood $\text{Ratio}_{in} = -2.203$, $p = 0.028$.

We can draw two conclusions from this analysis. 1) Both predictive checking from the linear PA model and direct statistcal tests of the power law fit indicate that both the in and out-degree of the Bitcoin OTC network are not scale-free, therefore since direct MLE of the simulated suggest that the graph size is large enough so that power law generated, the KS test directly rejects the model since scale-freeness is a guaranteed property. 2) The fact that $D_{in} < D_{out}$ in the KS test and that $\hat{\kappa}^{out} < \hat{\kappa}^{in}$ in the power law MLE fit suggest that the out-degree deviates more from power law and has fatter tails, compared to the in-degree.

3.2.3 In-Out Degree Correlation and Degree Growth

Figure 3 shows in- versus out-degree of the real and simulated network. Black dots denote the joint degree only for the first 10,000 edges. It is clear that the real network has a much higher correlation between in-degree and out-degree for each node: for the real network, the correlation is about 0.95; for the simulated network, the correlation is about 0.54. Moreover, the region where the in and out-degree seems uncorrelated is smaller in the real network. In the linear PA Model, since the probabilities of being chosen as a in-coming or out-going node only depends on its in-degree or out-degree (and the δ's) respectively, the degree correlation in the simulated network is entirely from time (i.e. older nodes have a higher probability of being connected), with the δ's controlling how much preferential attachment depends on its past. The overt in-out-degree correlation hence cannot be explained by the model. The black dots show that the in-out-degree correlation is already present in the real network at the first third of the network process; whereas for the simulated

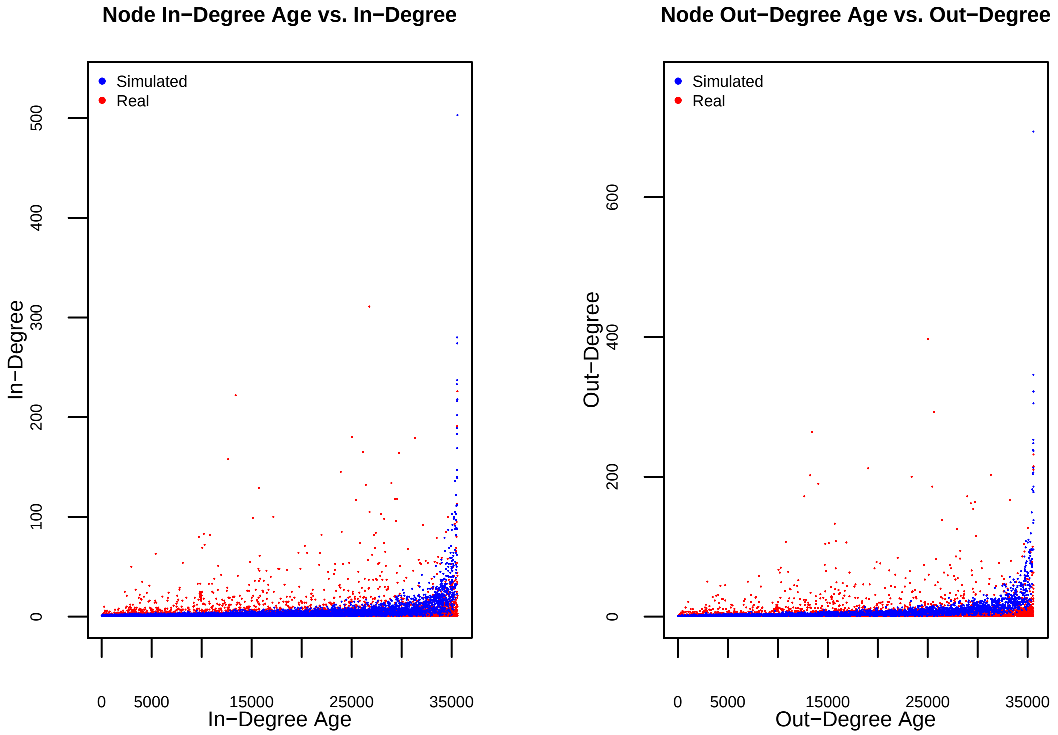

Figure 4: Age of Node vs. Degree

network, we can see that blue dots dominate the tail end where there is in-out-degree correlation coming from older age nodes.

Figure 4 supports the observation that in the real network, as supposed to preferential attachment models where in and out-degree correlation comes from time, the age of the node and its degree has less clear of a relationship therefore it is unlikely that the excess in-out-degree correlation in the real network is due to time/preferential attachment. The left panel shows the relationship between a node's in-degree age, which is defined as the amount of time steps since the node has been introduced as the "receiver" of an edge (i.e. the w in (v, w)), and its in-degree. While the right panel shows the relationship between out-degree age and out-degree. We can see that the positive relationship is much clearer in the simulated network for both the in- and out-degree. The top panel of Figure 6 in the appendix shows the same graphs but on the log-log scale, where the positive relationship is clearer for both the simulated and real network, however the contrast is still big. The figures in the regular scale outlines the many nodes in the real data which have high degrees but are not the oldest nodes, whereas in the simulated network, this is almost completely non-existent. To quantify the relationship between a node's age and degree, I estimate for both the real and simulated networks the following OLS regressions,

$$\log(\text{In-Degree}_{i,j}) = \alpha_j^{in} + \beta_j^{in} \log(\text{In-Degree Age}_{i,j}) + \epsilon_i \tag{8}$$

$$\log(\text{Out-Degree}_{i,j}) = \alpha_j^{out} + \beta_j^{out} \log(\text{Out-Degree Age}_{i,j}) + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{9}$$

where $j \in \{\text{Real Netowrk, Simulated Network}\}$, and $i \in V(n)$. Nodes which have degree zero are excluded (have neither been an in-coming/out-going node). The result is that,

$$\hat{\beta}_{real}^{in} = 0.225, \quad \hat{\beta}_{real}^{out} = 0.169, \quad \hat{\beta}_{sim}^{in} = 0.600, \quad \hat{\beta}_{sim}^{out} = 0.7362$$

all significant at the 95% level, which strongly supports the visual illustrations in Figure 4. To see whether the in-degree or out-degree growth rate is more well captured by the linear PA model, I used $\hat{\beta}_{sim}^{in}$ and $\hat{\beta}_{sim}^{out}$

to predict (8) and (9) respectively for the real network and calculate the corresponding mean square error, rescaling to the linear scale, obtaining

$$MSE_{in} = 318.5 \qquad MSE_{out} = 535.4$$

showing that the out-degree is more signifcantly off from model prediction, consistent with previous analysis on the degree distribution and power law exponent estimates. These mean square error values are quite sizable, since the average degree in the real network is around 6 for both the in- and out-degree, while the maximum for the out degree distribution is 535 and for the in-degree distribution is 763.

Overall, the in and out-degree correlation and degree growth rates are not well captured by the model, with the out-degree being especially unrelated with age and off from model prediction. One might suspect that perhaps either, in the real network probabilites equations (1) - (3) also depend on some function of total degree, explaining the high in-out-degree correlation and low correlaton bewteen the in/out-degree and its respective degree age, or more simply that preferential attachment is the wrong model here and there are external factors affecting the choice probabilities. The bottom panel of Figure 6 in the appendix shows that the first scenario is unlikely. The right graph of the bottom panel is now the logrithm of in degree versus the logrithm of the total age, defined as the time steps passed since the node's *first* appearance (regardless of whether as an in-coming or out-going node). If the PA mechanism also depends on the total degree, then the total age should be more predictive of the node's in/out degree. However, as we can see in Figure 6, the total age-degree relationship is still noisey for the real data, and far off from the simulated network. The out-degree, however, seems to be more correlated with the total age than the out-degree age, as fitting the log out-degree as a function of the total age now yields,

$$\log(\text{Out-Degree}_{i,real}) = \alpha_{real} + \beta^{total}_{real} \log(\text{Total Age}_{real}) + \epsilon_i$$

with $\hat{\beta}^{total}_{real} = 0.176$ significant at the 95% level, which is larger than previously using out-degree age as the predictor ($\hat{\beta}^{out}_{real} = 0.169$). This might point to some insight about why the out-degree distribution is especially off model prediction, but the relationship is still weak nevertheless.

3.2.4 Dynamics, Assortativity, and the Clustering Coefficient

I will now turn to examining parameter dynamics and other network characteristic measures, which might shed light on the appropriateness of modeling with preferential attachment for bitcoin data and explain how weak relationship between node age and degree in the data. The first three panels of Figure 5 shows how the parameters and network characteristic coefficients update through time steps. For each parameter/coefficient, I estimated how its value update every 1,186 more time steps, from the beginning of the network through the end. The x-axis represent the time steps. Dotted lines represent the values for the simualted network. We can see that for all values of $\boldsymbol{\theta}$, the values of the simulated and real network do indeed converge. α, β, and γ are particularly well-behaved. Recall that δ_{in} and δ_{out} represent the proportion which the choice probabilites does not depend on preferential attachment, so we can view the difference $|\tilde{\delta}^{real}_{in,out} - \tilde{\delta}^{sim}_{in,out}|$ as the extent to which our linear PA model is not capturing the non-preferential attachment factors affecting the choice probabilities at each time step. We can see that for δ_{in} and δ_{out}, deviations from the simulated estimates are larger than α, β, and γ and more unstable, suggesting that there are a lot of non-preferential attachment factors, particularly for the out-degree and early on in the network. A good model should show the $\tilde{\delta}$'s being relatively constant. This confirms that there is a significant amount of external factors driving the excess in-out-degree correlation and weakening the positive age-degree relationship in the data, especially for the out-degree. Moreover, the fact that $\tilde{\delta}^{real}_{in}$ and $\tilde{\delta}^{real}_{out}$ are generally decreasing through time suggests that preferential attachment matters more as n increases.

The third panel shows the evolution of assortativity and clustering coefficients. Clustering coefficient, C, for the whole graph here is defined as,

$$C = \frac{1}{|V'|} \sum_{v \in V'} \frac{\tau_{\triangle}(v)}{\tau_3(v)}$$

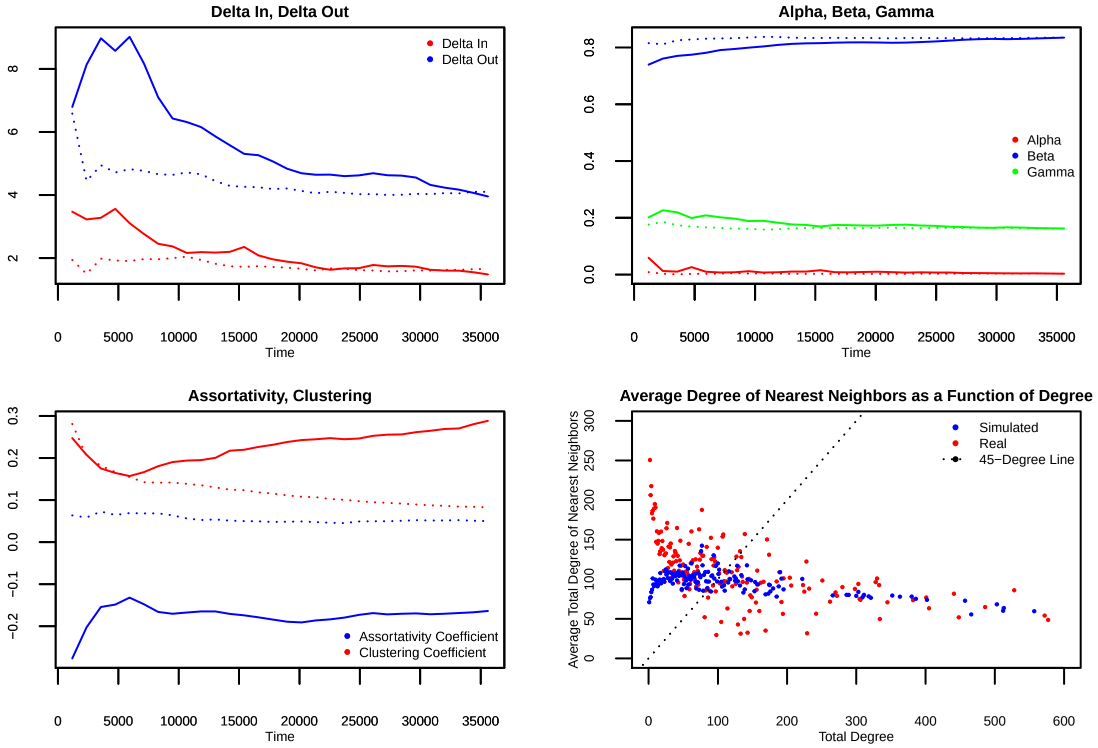

Figure 5: Parameter/Coefficient Evolution and Assortativity. Dotted colored lines represent the simulated network.

where $V' \equiv \{v \in V(G): \text{ total degree of } v \geq 2\}$, $\tau_{\triangle}(v) \equiv$ number of traingles in G which involve v, and $\tau_3(v) \equiv$ connected triples where v is the center. While there is no analytical solution for the clustering coefficient of this linear PA model (that I know of), using a mean-field approach, A. Fronczak, Fronczak, and Hołyst (2003) shows that the clustering coefficient for the cannonical BA model described in the introduction is, for large t,

$$C_{BA} = \frac{m-1}{8}\frac{(\ln t)^2}{t}$$

that is it should be decreasing in time. This coincides with what we see above for the simulated model. While for the real network, the cluster coefficient matches with the simulated networks' early on, but diverges after around 6,000 edges and starts increasing. For comparison, the clustering coefficient of a Erdos-Renyi graph with the same amount of edges and nodes is about 0. Overall, however, it is hard to interpret what part of the model is or is not capturing the behavior of the clustering coefficient of the real network, firstly because the coefficient is just one of many ways to measure the community structures and how subsets of the graph cluster (moreover, the coefficient used here is a coarse aggregate/average measure), and secondly, just based on the choice probabilities in equations (1) - (3), it is not directly obvious and intuitive about how clusters should form.

As for the assortativity coefficient, ρ, it is defined as in M. Newman (2003) as the Pearson Correlation Coefficient,

$$\rho = \frac{\sum_{j,k} jk\big(r_{j,k} - \sum_i r_{j,i} \sum_i r_{i,k}\big)}{\sigma(\sum_i r_{j,i})\sigma(\sum_i r_{i,k})}$$

where $r_{j,k}$ is the fraction of edges which direct a node with out-degree k to a node with in-degree j, and $\sigma(\cdot)$ denotes the standard deviation. Hence ρ attempts to measure how likely is a node to direct to another node with an in-degree similar to the out-degree of itself. If $\rho = 1$, then the network is perfectly assortative; if $\rho = -1$, then the network is perfectly dis-assortative. We can see that clearly, the linear PA model is slightly assortative but close to being non-assortative, this perhaps makes sense as since $\tilde{\beta}$ is large, therefore by equation (2), high in-degree nodes and high out-degree nodes will more likely be simultaneously chosen (but completely independently) to connect. For the real network, however, there is a tendency for a high out-degree node to direct an edge to a low in-degree node, which suggests that preferential attachment does not work in an average sense (as ρ is an average), given that $\tilde{\beta}$ is large.

The last panel of Figure 5 plots the average nearest neighbor total degree as a function of total degree (i.e. each point represents the average total degree of a set of nodes' nearest neighbors, where the set of nodes share the same total degree), as introduced by Barrat et al. (2004). The dotted line represent the 45 degree line. This function is an alternative way to examine assortativity but now as a function of degree and provides a different perspective than the assortativity coefficient. If all points of a graph lie on the 45 degree line, then the graph is perfectly assortative. This figure has excluded out extreme points. As shown in the figure, for the simulated network, the lower total degree nodes are assortative, but nodes with total degree rougly larger than 50 are dis-assortative with the slightly negative relationship. On the other hand, for the real network the function appreas to be strictly negative, matching the result in Kondor et al. (2014) for transaction level data. This suggests that the model seems to have only captured the dis-assortative nature of large total degree nodes. However, since this is all based on a single simulation of the network process, it is unclear if the dis-assortative behavior for the large degree nodes in the simulated network is a just a random artifact.

It should be pointed out that however, the precise reason for the assortativity behavior of the linear PA model considered here is not exactly clear, since 1) not a lot of analytical work has been done on the behavior of assortativity for directed scale-free graphs (most are on directed graphs), and particularly, graphs which also allow self-loops, multi-edges, and internal edge addition possible under the β scheme; and 2) these two methods of measuring the assortativity coefficients are not the only ways of measurement. For example, Litvak and Van Der Hofstad (2013) shows that for scale-free networks, ρ has the statistical property of approaching zero when the graph size is large, explaining what is found in M. E. Newman (2002) for undirected linear preferential attachment models like that of BA, and instead suggested using Spearman's rho. Williams and Genio (2014) studied the degree-correlation structure for scale-free directed networks using large scale simulations, and found that out of the four possible correlation combinations (in-in, out-in, in-out, out-out), only out-in correlation, which is the one used above, has the tendency to be negative while the rest all suggest no correlation.

Though these studies line up with the results discovered here, none of them studied models which allow self-loops and multi-edges. This scenario is particularly relevant here, precisely because if β is large, there is a high probability for large degree nodes to be simultaneously chosen thus displaying assortativity. Whereas for the cannonical BA model, dis-assortativity is intuitive because old nodes which have a high degree will more likely be chosen by new nodes which have low degrees, moreover, in the absence of multi-edges, high-degree nodes are not allowed to repeatedly connect with each other. This physical reasoning is also applied in Park and Newman (2003) and Maslov, Sneppen, and Zaliznyak (2004) to explain the dis-assortativity in the World Wide Web network, where there is restriction to single-edges. Therefore, given that $\tilde{\beta}$ is large, it is hard to image how preferential attachment in this model would generate dis-assortativity present in the real data.

By examining parameter dynamics and the assortative coefficient, we have found that the large difference $|\tilde{\delta}^{real}_{in,out} - \tilde{\delta}^{sim}_{in,out}|$ early on in the network process and particularly for the out-degree suggest that there are non-preferential attachment factors affecting the choice probabilities, explaining the weak age-degree relationship and excess in-out-degree correlation inconsistent with the preferential attachment mechanism and the poor fit for the out-degree. Further analysis on assortativity suggests that preferential attachment wich allows multiple edges might be the reason why there is discrepancy in the degree of assortativity between the real and simulated network.

4 Discussion and Conclusion

In this paper, I fitted a bitcoin trust-network dataset to a 5-parameter scale-free linear PA model and performed several predictive model checking procedures by comparing the real network to the simulated network. The main findings are that: 1) bitcoin data is not scale-free, whether generated from the model (KS test) or by direct statistical testing of the degree distribution; 2) nodes in the real network have a significantly higher correlation in its in- and out- degree; 3) contrary to preferential attachment, the positive age-degree relationship is much weaker in the real network; 4) the out-degree is especially poorly modeled; and 5) there is a greater degree of assortative behavior in the simulated network, whereas the real network is dis-assortative.

None of these phenomena are tested jointly however, so it is hard to systematically pinpoint exactly which aspect the model is failing, but we can still provide a few conclusions. (1) directly rejects the model via the KS test since scale-freeness is a guaranteed property of the model, moreover since direct maximum liklihood fit on the degree distribution of the simulated indeed cannot reject power law, we know that the graph size is large enough so that the distribution being test against in the KS test is approximately power law. (3) suggests that linear preferential attachment is not a great fit, and (5) suggests that linear preferential attachment which allows multiple edges is the incorrect mechanism that causes the mis-match in assortativity between the real data and the simulation. In particular, the evolution and instability of $\tilde{\delta}_{in}$ and $\tilde{\delta}_{out}$ suggest that there are a lot of external factors besides preferential attachment affecting how a node is to be chosen for connection, especially early on in the network and for the out-degree. This could explain why the age-degree relationship of a node is weaker in the real network and the excess in-out degree correlation found in section 3.2.2 as there are common factors driving the in- and out-degree simultaneously. Further analysis on using the total age to predict out-degree shows that the choice probabilities do not depend much more on total age, ruling out the dependence of total degree, thus further suggests that preferential attachment is the incorrect mechanism.

As for directions for future research, we then could suggest that the addition of modeling individual node-fitness may provide a better model. For example, let the choice probability in (1) be instead,

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

and similarly for other choice probabilities. The fitness parameter, η_w, seeks to capture the intrinsic individual attractiveness of the node and is drawn from a probability distribution. While there are no analytical results that I know of under this model with fitness, for the BA model, the network is no longer scale-free but the degree distribution insteads depends on the distribution $p(\eta_w)$[2]. Under this framework, the age-node relationship is weaker, since we may have a node with high fitness that enters at a later stage which becomes highly connected; moreover, modeling fitness would also give rise to the presence of high in-out-degree correlation, if η_w is common across the in and out degree. We can also further restrict the model to allowing only single-edges, which might capture the dis-assortative nature of the data.

Overall, goodness-of-fit of the model is likely to be dataset-dependent. This paper provides a model-based approach to assess the fit based on one snapshot observation of the graph, providing another tool-kit to analyze random scale-free graphs.

5 Reference

Albert, Reka, and Albert-Laszlo Barabasi. 2001. “Statistical mechanics of complex networks.” *Reviews of Modern Physics*. doi:10.1103/RevModPhys.74.47.

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.

Barrat, A., M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. 2004. “The architecture of complex weighted networks.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (11): 3747–52. doi:10.1073/pnas.0400087101.

Bezáková, Ivona, Adam Kalai, and Rahul Santhanam. 2006. “Graph model selection using maximum

[2] See R. Albert and Barabasi (2001) for a summary of results. Addtionally, if $p(\eta_w)$ follows the log-normal distribution, Simkin et al. (2012) shows that the resulting degree distribution is also log-normal, which is favored over power law by the likelihood ratio test in section (3.2.2)

likelihood." *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 105–12. doi:10.1145/1143844.1143858.

Bloem-Reddy, Benjamin, and Peter Orbanz. 2016. "Random Walk Models of Network Formation and Sequential Monte Carlo Methods for Graphs," 1–35. http://arxiv.org/abs/1612.06404.

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. 2009. "Power-law distributions in empirical data." *SIAM Review* 51 (4): 661–703. https://epubs.siam.org/doi/pdf/10.1137/070710111.

Fronczak, Agata, Piotr Fronczak, and Janusz Hołyst. 2003. "Mean-field theory for clustering coefficients in Barabási-Albert networks." *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 68 (4): 1–4. doi:10.1103/PhysRevE.68.046126.

Guetz, Adam N., and Susan P. Holmes. 2011. "Adaptive importance sampling for network growth models." *Annals of Operations Research* 189 (1): 187–203. doi:10.1007/s10479-010-0685-2.

Kondor, Dániel, Márton Pósfai, István Csabai, and Gábor Vattay. 2014. "Do the rich get richer? An empirical analysis of the Bitcoin transaction network." *PLoS ONE* 9 (2): 1–9. doi:10.1371/journal.pone.0086197.

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.

Kumar, Srijan, Francesca Spezzano, V. S. Subrahmanian, and Christos Faloutsos. 2017. "Edge weight prediction in weighted signed networks." *Proceedings - IEEE International Conference on Data Mining, ICDM*, 221–30. doi:10.1109/ICDM.2016.175.

Leskovec, Jure, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. 2010. "Kronecker graphs: An Approach to Modeling Networks." *Journal of Machine Learning Research*. doi:10.1090/S0002-9939-1962-0133816-6.

Litvak, Nelly, and Remco Van Der Hofstad. 2013. "Uncovering disassortativity in large scale-free networks." *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 87 (2): 1–11. doi:10.1103/PhysRevE.87.022801.

Maslov, Sergei, Kim Sneppen, and Alexei Zaliznyak. 2004. "Detection of topological patterns in complex networks : correlation profile of the internet" 333: 529–40. doi:10.1016/j.physa.2003.06.002.

Newman, M. E.J. 2002. "Assortative Mixing in Networks." *Physical Review Letters* 89 (20): 1–5. doi:10.1103/PhysRevLett.89.208701.

Newman, Mark. 2003. "Mixing patterns in networks." *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 67 (2): 13. doi:10.1103/PhysRevE.67.026126.

Park, Juyong, and M. E.J. Newman. 2003. "Origin of degree correlations in the Internet and other networks." *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 68 (2): 7. doi:10.1103/PhysRevE.68.026112.

Simkin, M V, V P Roychowdhury, My T Thai, and Panos M Pardalos. 2012. *Fitness-Based Generative Models for Power-Law Networks*. Vol. 57. doi:10.1007/978-1-4614-0754-6.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57 (2): 307–33.

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.

Williams, Oliver, and Charo I.Del Genio. 2014. "Degree correlations in directed scale-free networks." *PLoS ONE* 9 (10): 1–6. doi:10.1371/journal.pone.0110121.

Wiuf, Carsten, Markus Brameier, Oskar Hagberg, and Michael P H Stumpf. 2006. "A likelihood approach to analysis of network data." *Proceedings of the National Academy of Sciences of the United States of America* 103 (20): 7566–70. doi:10.1073/pnas.0600061103.

5.1 Appendix

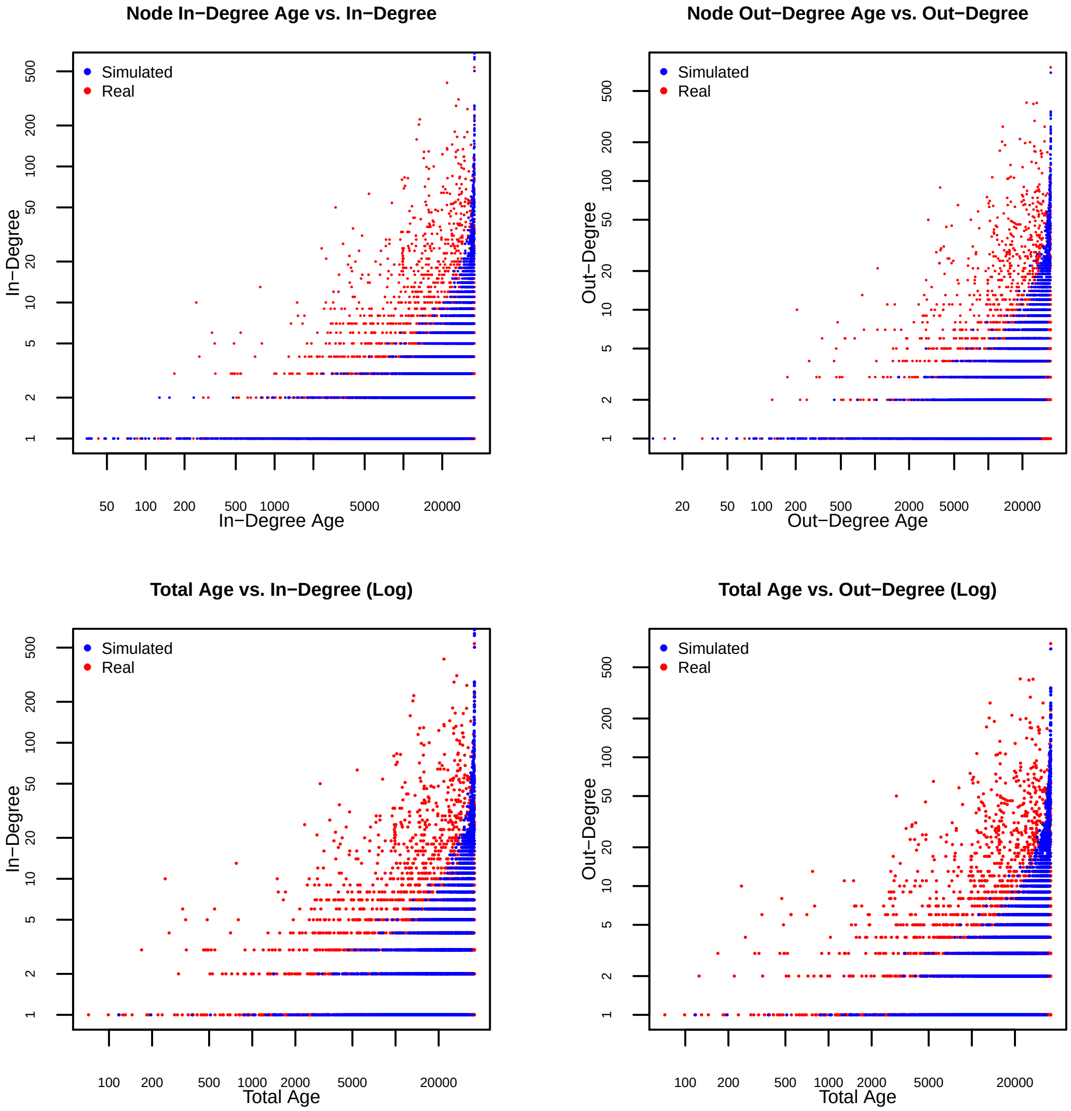

Figure 6: Age of Node vs. Degree (Log-Log Scale)