Quality sorting model of network formation

Jong Hyun Chung

1 Introduction

The preferential attachment model predicts power law degree distribution often observed in the real world networks. The model, however, has a strong normative implication about these networks. The degree difference across nodes at a given point in time depends solely on the initial degrees of the nodes (i.e. their ages) and is unaffected by the fundamental characteristics of the nodes. Applied to the sales distribution of products or citation distribution of academic papers, the model implies the superstars are just as good as the countless no-names. The idea that highly skewed distribution may arise in the absence of quality differentiation dates back to at least Adler (1985).

Literature, however, suggests this prediction of preferential attachment model does not hold in the real world. The initial degree alone does not fully predict the node's future links. The past studies, however, fall short of quantifying the relative importance of luck (or the node's initial condition in the network, such as the size of the network at its birth) and quality (or the node's inherent ability to attract links independent of its position in the network).

This assessment is difficult in part because the quality of node is difficult to observe directly. And yet, the model with node-specific quality proposed by Bianconi and Barabasi (2001) is sensitive to the quality distribution across nodes. It is therefore difficult to apply the model to quantitatively assess the importance of node quality.

In this paper, I construct a model where each agent forms a link following a two-step procedure. First, it searches for potential nodes to form a link. Second, it observes the quality of the potential nodes and form a link with the highest quality nodes. The key difference with the fitness model is that quality of a node is ordinal so the distribution of the quality is no longer relevant. Still, the model would allow, for example, measuring the correlation between the citation count and the quality ranking to quantitatively assess how well the network can discover high-quality nodes.

The model results provide a formal justification of an empirical approach taken by Newman (2009) and Newman (2014). These papers use the citation count conditional on the publication date as a proxy measure of the paper's quality. Compared with other papers with the same initial citation count, these papers gain earn more citations. Consistent with this observation, the model predicts high quality nodes are more likely to receive future links conditional on its current degree.

In the second half of the paper, I introduce high energy physics citation data as a ground for testing and fitting the model. I estimate the model parameters to fit the model-predicted distribution to the observed citation count distribution. The parameters that allow quality-based selection provide better fit to the data than the parameter that predicts no selection. Consequently, the calibrated model suggests quality premium so that the high-quality papers receive on around 6 more citations relative to the mean citation of 12.

2 Model

A network is formed over discrete time $t = \{1,2,\ldots\}$. At each period t a new node is added with a permanent quality $\eta \in \mathbb{R}$ drawn from a common distribution F. Denote the node added at time t by t and let $N_t = \{1,\ldots,t\}$ be the set of all nodes in the network by the end of period t. Links are formed in a two-stage procedure. In the first stage, node t surveys s number of nodes among N_{t-1}. In the second stage, the node forms links with $m \leq s$ highest quality nodes among the surveyed ones.

The survey process can be specified in a flexible manner. I assume that the survey process does not depend on the node quality – node quality is considered unobserved until the node has been surveyed. With this assumption, the probability that node $i < t$ with in-degree $d_i(t)$ and quality η_i receives a new link in period $t+1$ can be expressed as

$$\Pr[t+1 \text{ links } i|\eta_i] = \Pr[t+1 \text{ surveys } i] + \Pr[t+1 \text{ links } i \mid t+1 \text{ surveys } i, \eta_i].$$

The survey process can be specified flexibly, but as a baseline case, I consider the random survey process so that each node selects s nodes randomly from N_{t-1}. In this case, $\Pr[t+1 \text{ surveys } i] = \frac{s}{t}$. The probability node $t+1$ links to node i conditional on having surveyed it is the probability that η_i is greater than at least $s-m$ other surveyed nodes, whose quality distribution is the same as the population distribution F due to the random survey process. Denoting the quality quantile as $q \equiv F(\eta)$,

$$\begin{aligned}\Pr[t+1 \text{ links } i \mid t+1 \text{ surveys } i, \eta_i] &= \sum_{k=0}^{m-1} \binom{s-1}{k} (1-q_i)^k q_i^{s-k-1} \\ &= I_q(s-m, m),\end{aligned}$$

where $I_x(\alpha,\beta)$ is the regularized incomplete beta function.

Following Barabasi, Albert and Jeong (1999), I use mean-field method to estimate the degree distribution. Denoting d_q as the in-degree of a node with quality quantile q, the continuous approximation of the degree evolution is

$$\frac{\partial d_q}{\partial t} = \frac{sI_q(s-m,m)}{t} \equiv \frac{C(s,m,q)}{t}.$$

The solution to the differential equation yields the degree of node i at time t

$$d_q(t) = C(s,m,q)\log\left(\frac{t}{i}\right)$$

and the degree distribution conditional on quality quantile

$$F_t(d \mid q) = 1 - \exp\left(-\frac{d}{C(s,m,q)}\right).$$

Thus, the conditional distribution is exponential with mean $C(s,m,q)$. As one would expect, the mean degree of a node is increasing in its age $t-i$ and its quality η since $C(s,m,q)$ is increasing in q. In particular, this implies that conditional on age, higher quality nodes are more likely to have higher degree.

The unconditional distribution can be found by taking the mean with respect to q. Since q is uniformly distributed regardless of F, the unconditional degree distribution does not depend on the quality distribution. While deriving the general solution is non-trivial, it is instructive to consider the case of $m = 1$ so that the new node surveys s nodes and links with the highest quality node among the surveyed ones. If $s = 1$, then quality is irrelevant and the model collapses to a random growth network so that the degree distribution is $F_t(d) = 1 - e^{-d}$. When $s > 1$, the distribution is given by

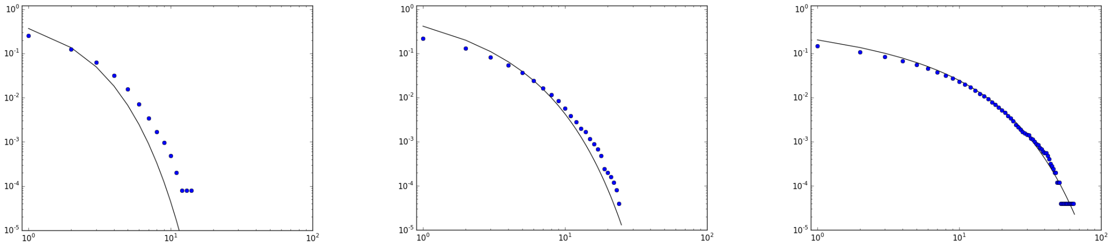
$$F_t(d) = 1 - \frac{1}{s-1}E_{\frac{s}{s-1}}\left(\frac{d}{s}\right),$$

where $E_n(x) \equiv \int_1^{\infty} \frac{e^{-xt}}{t^n}\,dt$ is the generalized exponential integral function. Given the limit $E_n(x) \to x^{-1}e^{-x}$ as $x \to \infty$, the distribution can be approximated as $1 - \frac{s}{s-1}\frac{e^{-d/s}}{d}$ for large d.

Interestingly, model predicts a different distribution from preferential attachment model (power law) as well as from uniform attachment model (exponential), providing a testable implication.

To test the goodness of mean-field approximation, I plot the complementary cumulative distribution function (CCDF) predicted by the model and given by a simulation with $T = 100000$ periods. The approximation provides a reasonably good fit, especially for larger s.

An interesting aspect of the model is how different values of s and m affect the rich-get-richer phenomenon. Table 1 reports the Gini coefficient of the in-degree distribution for different values of s and m, where the degree distributions are found through simulations. The result suggests that the inequality increasing s and decreasing in m. Intuitively, increased search intensity through larger s (broader survey) and lower m (pickier link) concentrate links to high-quality nodes, generating greater degree inequality in the resulting network.



(a) $s = 1$

(b) $s = 3$

(c) $s = 10$

Figure 1: Simulated and predicted distributions, $m = 1$

m	s			
	1	3	6	10
1	0.667	0.767	0.850	0.897
2	-	0.644	0.767	0.842
5	-	-	0.572	0.699

Table 1: In-degree Gini coefficients

The model presented in this milestone is particularly parsimonious, characterized by only two parameters s and m. While I derive analytic results for $m = 1$ case, simulating the model is similarly trivial for higher m. Fitting the model parameters given an observed degree distribution, therefore, would be straightforward.

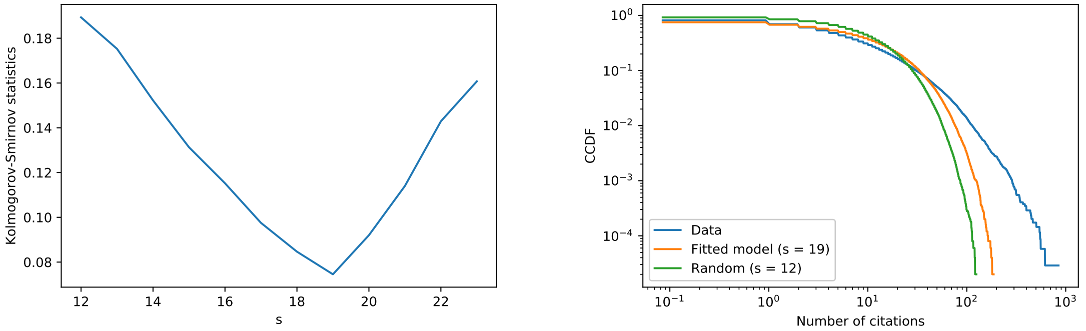
In the final version of the project, I hope to explore other variants of the model in order to better understand how the survey process can affect the network dynamics. In one direction, I may consider a preferential survey processes: Instead of randomly selecting randomly, a newly arrived node may choose the nodes to survey based on their degrees. In another direction, allowing node quality to contain pair-specific component would be useful. In this case, broader survey may offset the rich-get-richer process by allowing nodes to find idiosyncratic matches and connect with low-degree but well-matched nodes.

3 Citation network

I examine the high-energy physics citation network data covering 34,546 HEP-PH (high energy physics phenomonology) papers from January 1993 to April 2003. The total number of citations is 421,578, and the average number of citations for each paper is 12.2.

I fit the model to the data by fixing $m = 12$ and choosing $s \geq m$ such that the Kolmogorov-Smirnov statistics between the empirical distribution and the model distribution based on the simulation is minimized. Figure 2(a) reports the result of this exercise by plotting the Kolmogorov-

Smirnov statistics against the varying values of s. The model matches the observed distribution the best at $s = 19$.



(a) Kolmogorov-Smirnov statistics

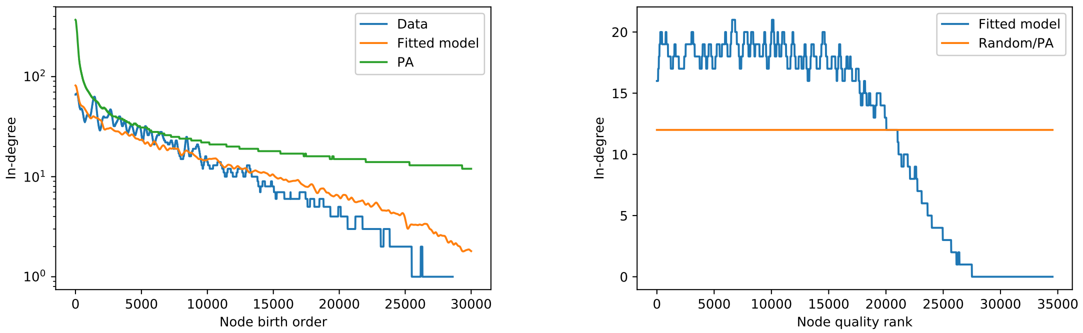
(b) Complementary CDF, data and model

Figure 2: Random survey model goodness-of-fit

For visual check of how well the model fits the data, Figure 2(b) plots the CCDF of citations received by papers in the data and predicted by the model, with $s = 19$ and $s = 12$. The latter case reduces to a random attachment model where a new node randomly chooses set of existing nodes to link. Compared to the random attachment model, the model with $s = 19$ clearly provides fits the empirical distribution better. However, even with the selection from $s > m$, the model fails to fully account for the tail distribution and under-predicts the frequency of "superstar" papers. This suggests that the survey process is at least somewhat preferential to high-degree nodes. Later in the paper, I explore the preferential survey process.

On the other hand, simple preferential attachment model over-predicts the importance of node age. Figure 3(a) plots the in-degree of node ranked by age, smoothed by Gaussian kernel. For the data plot, the node age is based on the paper publication date, so the paper published earlier has higher age rank. On a log-plot, the in-degree has linear relationship with the node age rank in the data. The preferential attachment model, however, predicts exponential relationship on the log-plot, predicting much higher importance on the birth order in the degree dispersion. In contrast, the sorting model the degree dispersion is in part due to the quality differences, so the node age plays more modest role. As the figure shows, its prediction on the age-degree relationship fits the data much more closely than the preferential attachment model.

As a way to quantify the effect of quality, Figure 3(b) plots the in-degree on the quality ranking. Both the preferential attachment and random attachment model predict no role of node quality, so the relationship is flat at the mean degree of 12. The fitted model predicts modest role of quality. For sufficiently high-quality nodes, the in-degree is about 18-19. Compared to the average citation of 12, quality premium is about 6-7, which is modest given the standard



(a) In-degree on node age

(b) In-degree on node quality

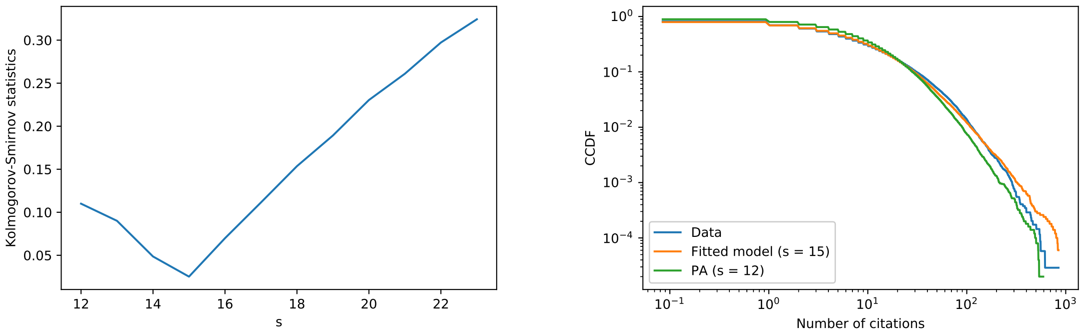
Figure 3: Conditional in-degree

deviation of 25. The marginal effect of quality diminishes in this model due to the fact that top $\frac{m}{s}$ quality nodes will be always linked conditional on being surveyed. Hence the marginal role of quality is diminished above the threshold, as reflected in the initial plateau in the graph. For mid-low range, however, quality plays an important role in the number of links a node gathers.

The analysis so far insisted on random survey process, where each node existing node has equal probability of being surveyed by a new node. While this assumption provides a parsimonious model with simple intuition, its restrictiveness limits the ability of the model to fit the data, as illustrated in Figure 2(b). An alternative assumption is to model the survey process as following a variant of preferential attachment. Formally, I consider the process formalized by Mitzenmacher (2004). With probability p, a new node picks an existing node randomly to survey. With probability $1-p$, it picks an existing node with probability proportional to in-degree. In the context of citation network, this survey process captures the idea that more cited papers become more well-known and thus become more likely to be read by other researchers.

To fit the model, I first estimate the parameter p by estimating the power-law coefficient α following Clauset, Shalizi, and Newman (2009). More precisely, given a cutoff value x^*, the power law degree exponent $\hat{\alpha}$ is estimated through maximum likelihood using the data points x such that $x \geq x^*$. The choice of the cutoff is then determined to minimize the Kolmogorov-Smirnov statistics between the estimated Pareto distribution with shape $\hat{\alpha}$ and minimum x^* against the empirical distribution above x^*. From the corresponding $\hat{\alpha}$, the probability parameter p is estimated from the relation $\alpha = 1 + \frac{1}{1+p}$. From the citation data, I find $\hat{\alpha} = 4.12$ and $p = 0.68$.

I still fix $m = 12$ and choose s with the same criterion as before. Figure 4(a) shows that $s = 15$ provides the best fit based on Kolmogorov-Smirnov statistics. Figure 4(b) shows the complementary CDF of the fitted model. Given the additional degree of freedom in the model, the model is able to fit the degree distribution better than the random survey model. Relative to the no-sorting model ($s = 12$), the sorted model is better able to capture the CCDF curvature in the actual distribution. Nonetheless, it still has strong linearity in the tail distribution to fit the data completely.

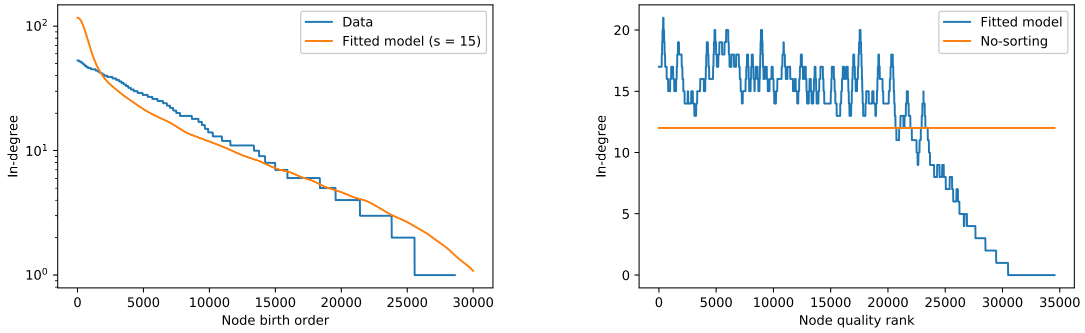


(a) Kolmogorov-Smirnov statistics

(b) Complementary CDF, data and model

Figure 4: Preferential survey model goodness-of-fit

Figure 5 provides the results of the same set of exercises as with the random survey model. The age-conditional degree follows the data reasonably well, although like the no-sorting preferential attachment model, the age-premium is over-estimated. The quality-premium shown in 5(b) is similar to the one found with random survey model in Figure 3(b), although the quality premium seems to be slightly lower with the preferential survey.



(a) In-degree on node age

(b) In-degree on node quality

Figure 5: Preferential survey model conditional in-degree

4 Conclusion

Despite the evidence that the node-specific qualities contribute to its network properties, the difficulty of observing and measuring such qualities has challenged quantitatively analyzing how node qualities interact with the network formation and resulting degree distribution. In this paper, I propose a model with node-specific qualities that are ordinal in nature. This allows structural estimation of the model without specifying the quality distribution across nodes and consequently study, for example, the role of quality differences in node degree distribution. Consistent with the past empirical results, the model predicts that older and higher quality nodes on average have higher degrees.

Using citation data, I estimate the model and indeed find an evidence that paper quality plays a non-trivial role in the number of citations it gathers. While the model is unable to estimate the quality of each individual node, it can predict the aggregate relationship between node quality and degree. I find that high-quality papers receive about 18 citations compared to the average of 12, suggesting a modest role of quality. This result appears to be robust against survey process specification.

While the proposed model provides an improvement over simple preferential attachment or random attachment models, it still falls short in matching the degree distribution as well as the age-conditional distribution observed in the data, suggesting the model could improve from additional considerations. An important extension for future study may allow both node-specific and pair-specific components to the node quality. For example, in the context of the citation network, pair-specific component would capture the relevance of the papers. Such feature could allow studying how the degree concentration due to node-specific quality affects the ability of new nodes to discover relevant (but possibly low-quality) nodes. On the empirical side, a future study may explore how to specify the survey process so that the resulting distribution provides a good fit for the given data.

References

[1] Adler, M. (1985). “Stardom and Talent.” *American Economic Review* 75(1), p.208–212.

[2] Barabasi, A.L., Albert, R., and H. Jeong (1999). “Mean-field theory for scale-free random networks.” *Physica A* 272, p.173-187.

[3] Bianconi, G. and A.L. Barabási (2001). “Competition and Multiscaling in Evolving Networks.” *Europhysics Letters* 54(4), p.436-42.

[4] Clauset, A., Shalizi, C.R., and M.E.J. Newman (2009). “Power-law distributions in empirical data.” *SIAM Review* 51(4), p.661-703.

[5] Gehrke, J., Ginsparg, P., and J. M. Kleinberg (2003). “Overview of the 2003 KDD Cup.” *SIGKDD Explorations* 5(2), p.149-151.

[6] Jackson, M.O. and B.W. Rogers (2007). "Meeting Strangers and Friends of Friends: How Random Are Social Networks?" *The American Economic Review* 97(3), p.890-915.

[7] Kong, J.S., Sarsahr N., and V.P. Roychowdhury (2008). "Experience Versus Talent Shapes the Structure of the Web." *Proceedings of the National Academy of Sciences* 105(37), p.13724-9.

[8] Leskovec J., Kleinberg J., and C. Faloutsos (2005). "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations." ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

[9] Newman, M.E.J. (2009). "The First-mover Advantage in Scientific Publication." *Europhysics Letters* 86(6), 68001.

[10] Newman, M.E.J. (2014). "Prediction of Highly Cited Papers." *Europhysics Letters* 105(2), 28002.