# Estimation of a Scale-Free Network Formation Model*

Anton Kolotilin† and Valentyn Panchenko‡

This version: February, 2016

**Abstract**

Growing evidence suggests that many social and economic networks are scale free in that their degree distribution has a power-law tail. The most widespread explanation for this phenomenon is a random network formation process with preferential attachment. For a general version of such a process, we develop PMLE and GMM estimators. By establishing the uniform law of large numbers for growing networks, we prove consistency of these estimators. Simulations suggest asymptotic normality of these estimators. In contrast to these estimators, the commonly used NLLS and local tail-index estimators perform poorly in finite samples. We apply our estimation methodology to a co-authorship network.

*JEL Codes: C15, C45, C51, D85*

*Keywords: consistency, degree distribution, network formation, scale-free network*

*Anton Kolotilin started working on the paper during his PhD at MIT, whose hospitality is gratefully acknowledged. We thank Denis Chetverikov and Victor Chernozhukov for numerous helpful comments that significantly improved the paper. We also thank Isaiah Andrews, Arun Chandrasekhar, Jerry A. Hausman, Guido Imbens, Anna Mikusheva, Whitney Newey, Chad Syverson, and the participants at MIT Econometrics Lunch and at the 2nd Sydney Econometrics Research Group Workshop for helpful comments. All errors are our.

†UNSW, School of Economics. Email: a.kolotilin@unsw.edu.au

‡UNSW, School of Economics. Email: v.panchenko@unsw.edu.au

# 1 Introduction

Many real networks have a *degree distribution with a power-law tail*.[1] That is, the fraction $P(d)$ of vertices that have $d$ neighbors is approximately proportional to $d^{-\gamma}$ for large $d$, where $\gamma$ is a positive constant called the *power-law parameter*. Such networks are called *scale-free*. The power-law parameter plays a important role for the topology of a network and for the network's statistical properties, such as learning, the spread of viruses, the size of the largest component, the connectivity, the searchability, and the robustness to errors and attacks (Albert and Barabasi, 2002). In this paper, we estimate the power-law parameter and other parameters for a general model of random scale-free network formation.

Barabasi and Albert (1999) build the first theoretical model of scale-free network formation (hereafter the BA model):

> "...starting with a small number ($m_0$) of vertices, at every time step we add a new vertex with $m(\leq m_0)$ edges that link the new vertex to $m$ different vertices already present in the system. To incorporate preferential attachment, we assume that the probability $\Pi$ that a new vertex will be connected to vertex $i$ depends on the connectivity $k_i$ of that vertex, so that $\Pi(k_i) = k_i / \sum_j k_j$. After $t$ time steps, the model leads to a random network with $t + m_0$ vertices and $mt$ edges."

The idea of the model is that the rich get richer: more "popular" vertices get more links than less popular vertices as a network evolves. Such a process is called *preferential attachment*. The BA model initiated further wide-range investigation and modelling of scale-free networks.[2]

Cooper and Frieze (2003) and Cooper (2006) introduce and analyze a general model of scale-free network formation (hereafter the CF model). This model nests many models of scale-free network formation, including the BA model and popular hybrid models, such as Jackson and Rogers (2007).[3]

[1] Such networks include social networks (coauthorship, citation, movie actor, and sexual relation networks), biological networks (ecological and food webs, cellular, protein and neural networks) and other networks (WWW, Internet, Fedwire interbank market, and power grid networks). See Albert and Barabasi (2002) and Dorogovtsev and Mendes (2002).

[2] For a general overview of scale-free network literature, see Albert and Barabasi (2002), Dorogovtsev and Mendes (2002) and Chapter 5 of Jackson (2008). For a rigorous mathematical treatment of scale-free random graph processes, see Bollobas and Riordan (2003).

[3] Specifically, the CF model is able to reproduce the same (asymptotic) degree distribution as in these models. However, clustering obtained in Jackson and Rogers (2007) cannot be reproduced by the CF model.

In the CF model, initially, there is a small fixed network. At each subsequent period, a new vertex with a random number of edges is added. Some of added edges connect a new vertex with the existing network, and others connect old vertices between themselves. The sampling method for choosing link endpoints is decided uniformly at random with some probability and by preferential attachment with the complementary probability.

Cooper (2006) shows that the asymptotic degree distribution depends only on the subset of parameters including the initial degree distribution of added vertexes, $\boldsymbol{p}$, the share of edges added by preferential attachment, $\eta$. Moreover, the asymptotic degree distribution is a linear combination (with coefficients $\boldsymbol{p}$) of modified Yule-Simon distributions (see Simon, 1955). Most interestingly, the asymptotic degree distribution has a power-law tail, where the power-law parameter is determined as $1 + 1/\eta$. The goal of this paper is to develop rigorous methodology for estimating the parameters determining the asymptotic degree distribution.

Despite a variety of theoretical models of network formation, there is a lack of rigorous econometric methods that estimate structural parameters of network formation models. Many papers just plot the transformed degree distribution for real networks on the log-log scale and calculate the slope of the tail of the distribution to estimate the power-law parameter. The local tail exponent estimators, starting from Pickands (1975), Hill (1975), Smith (1987), their generalisations and similar alternative estimators are popular formal procedures for estimating a tail exponent. However, many of these estimators assume continuous distributions, rely on a specific tail behaviour and all of them strongly depend on the appropriate choice of the number of tail observations. From the perspective of structural model estimation Pennock et al. (2002) and Jackson and Rogers (2007), apply nonlinear least squares procedures (hereafter NLLS) to fit the empirical degree distribution to an approximation of the parametrized asymptotic degree distribution. Goldstein et al. (2004) illustrate that although such procedures give good graphical fits of empirical and estimated distributions on the log-log scale, they may give biased and inaccurate estimates of model parameters. Goldstein et al. (2004) also argue that the maximum likelihood estimation is much more robust. Unfortunately, as Jackson and Rogers (2007) note, deriving analytically and then computing numerically the true likelihood of all possible degree distributions appears to be intractable for scale-free network formation models. König (2015) provides behavioural foundations for the model of Jackson and Rogers (2007) and estimates the model using Bayesian methods.

Atalay et al. (2011) and Atalay (2013) estimate parameters of a network formation model using a pseudo maximum likelihood estimator (hereafter PML). Specifically, they calculate the

The methods presented in this paper will rely on certain properties derived for the CF model, but not yet available (and not trivially derivable) for more general models allowing for clustering.

likelihood assuming that each node degree is independent and identically distributed according to a derived asymptotic degree distribution. However, in their network formation models, node degrees are interdependent and have different distributions even in asymptotics ("old" nodes have a much higher degree than "young" nodes). Under these conditions the asymptotic properties the PML estimator are not well understood. Our results can be extended to derive the properties of this estimator.

To estimate parameters of the CF model, we develop a class of generalized method of moments (hereafter GMM) estimators, which includes the PML estimators as a special case. This GMM estimation is computationally simple, because it requires calculating only a sample average of a moment function, as opposed to the true likelihood of all possible degree distributions. We show formally that the GMM estimators give consistent estimates of the CF model parameters. Standard consistency results rely on the uniform law of large numbers for independent or stationary data. We cannot rely on these results because node degrees are interdependent in a non-standard way in the CF model. Our main technical contribution is the prove of the uniform law of large numbers using first principles. While the proof relies on certain properties of the degree distribution sequence established for the CF model, the proof is sufficiently general and can be extended to other network-formation models. Relying on the introduced uniform law of large number we establish the asymptotic properties the GMM and PMLE estimators. Simulations suggest that the GMM and PMLE estimators perform well in finite sample, i.e., the distribution of estimates is close to normal, there practically no bias even for small networks and the variance is small in comparison with other estimators. distributed. The NLLS estimator and local tail exponent estimators exhibit larger biases and higher variances. We illustrate the usefulness of our estimation methodology with an application to the network of co-authorship relationship among economists which was earlier investigated in Goyal et al. (2006) and Jackson and Rogers (2007).

The rest of the paper is organized as follows. Section 2 introduces and discusses the model. Section 3 describes the estimation methodology by establishing new uniform law of large numbers for growing networks, introducing the PMLE and GMM estimators and deriving the asymptotic properties of the these estimators. Section 4 compares finite sample statistical properties of the PMLE and GMM estimators with currently applied NLLS and local tail exponent estimators using Monte Carlo simulations. Section 5 illustrates an application of the estimation methodology. Section 6 concludes. Appendix A restates the key results of Cooper (2006) used in this paper. All proofs are relegated to Appendix B. Appendix C derives the NLSS and local tail exponent estimators for the CF model.

# 2 Network Formation Model and Discussion

## 2.1 Setup

We study the CF model introduced by Cooper and Frieze (2003) and further analyzed by Cooper (2006). Following them, we describe network formation as a statistical process, but this process can be economically microfounded using benefits and costs of initiating and/or removing an edge (see Section 2.3).

Consider a random graph process, $(G(t))_{t\geq 1} = (V(t), E(t))_{t\geq 1}$, where $V$ is a set of vertices, $E$ is a set of edges, and $t \in \{1, 2, \dots\}$ is time.[4] In economic applications, the vertices typically represent economic agents and the edges represent their connections. Let $G(1)$ be an initial graph that contains $|V(1)| \geq 1$ vertices and $|E(1)| \geq 1$ edges (the number of elements of an arbitrary finite set $X$ is denoted by $|X|$ hereafter). For $t \geq 2$, the random graph $G(t)$ is obtained from $G(t-1)$ as follows. A *new* vertex born at time $t$ and indexed by its birth-time $t$ is added to the graph. The new vertex forms a random number of edges $m(t)$ connecting it with some existing (*old*) vertices in $V(t-1)$. At the same time, old vertices in $V(t-1)$ form $M(t)$ edges between themselves. Both $m(t)$ and $M(t)$ are bounded from above by integers $P$ and $Q$, and are independently distributed (among themselves and across time) according to finite support distributions $\boldsymbol{p} = (p_0, \dots, p_m, \dots, p_P)$ and $\boldsymbol{q} = (q_0, \dots, q_M, \dots, q_Q)$, where $p_m = \Pr(m(t) = m)$ and $q_M = \Pr(M(t) = M)$. These distributions characterise agents behaviour in forming new connections over time. Denote an average number of new-old edges added at $t$ by $\overline{m} = \mathbb{E}(m(t))$ and an average number of old-old edges added at $t$ by $\overline{M} = \mathbb{E}(M(t))$. We assume that there is a positive probability that at least one edge is added, i.e. $\overline{m} + \overline{M} > 0$. Denote the degree (i.e., the number of immediate neighbours) of a vertex $v$ of the graph $G(t)$ by $d(v, t)$.

Next we define with whom the agents form connections. First, consider edges $e_i^m(t)$, $i = 1, \dots, m(t)$, originating from new vertex $t$. Their terminal endpoints, vertices with whom $t$ connects, are chosen independently with probability $A_1$ by preferential attachment from $V(t-1)$ (i.e. the probability that an old vertex, $v$, is the terminal endpoint of $e_i^m(t)$ is proportional to the degree of this vertex $d(v, t-1)$),[5] and with probability $A_2 = 1 - A_1$

[4] Formally we should refer to this process as a multi-graph as we allow for *loops* (i.e., edges joining a vertex to itself) and *multiple edges* (i.e., several edges joining the same two vertices). However, relying on Bollobas et al. (2001), we expect that the fraction of multiple edges and loops goes to 0 as $t \to \infty$ for the considered process. Also, we treat all edges as undirected, but it is straightforward to extend the analysis to directed graph processes.

[5] Preferential attachment arises naturally in information sharing networks, more detailed microfounda-

uniformly at random from $V(t-1)$. To summarize,

$$p_A(v,t) \equiv \Pr(v \text{ is a terminal endpoint of } e_i^m(t)) = A_1 \frac{d(v,t-1)}{2|E(t-1)|} + A_2 \frac{1}{|V(t-1)|}.$$

Second, consider edges $e_i^M(t)$, $i = 1, \ldots, M(t)$, connecting old vertices in $V(t-1)$. The initial endpoint vertex and the terminal endpoint vertex of each edge $e_i^M(t)$ are chosen independently with probability $B_1$ and $C_1$ by preferential attachment from $V(t-1)$ and with probability $B_2 = 1 - B_1$ and $C_2 = 1 - C_1$ uniformly at random from $V(t-1)$. Thus, we obtain

$$\begin{aligned} p_B(v,t) &\equiv \Pr(v \text{ is an initial endpoint of } e_i^M(t)) = B_1 \frac{d(v,t-1)}{2|E(t-1)|} + B_2 \frac{1}{|V(t-1)|}, \\ p_C(v,t) &\equiv \Pr(v \text{ is a terminal endpoint of } e_i^M(t)) = C_1 \frac{d(v,t-1)}{2|E(t-1)|} + C_2 \frac{1}{|V(t-1)|}. \end{aligned}$$

The degree distribution, $P_t(d)$, of a random graph is itself a random object. Define $D_t(d)$ as the number of vertices of the graph $G(t)$ that have degree $d$. Then the degree distribution defined as the fraction of vertices of the random graph $G(t)$ that have degree $d$ is $P_t(d) \equiv D_t(d)/|V(t)|$, which is a random variable. Corollary 1 below shows that for all $d$ the fraction $P_t(d)$ converges in probability to $P(d)$ as $t$ goes to infinity. The limiting fractions $P(d)$ are called the *asymptotic degree distribution* of the graph process $(G(t))_{t\geq 1}$.

Corollary 1 also shows that the asymptotic degree distribution of the graph process $(G(t))_{t\geq 1}$ is fully characterized by the initial degree probability distribution of the newly added vertexes, $\boldsymbol{p}$, the average number of old-old edges, $\overline{M}$, and the limiting fraction of the endpoints inserted by the preferential attachment, $\eta$, defined as

$$\eta \equiv \frac{\overline{m}A_1 + \overline{M}(B_1 + C_1)}{2(\overline{m} + \overline{M})}.$$

Parameter $\boldsymbol{p}$ uniquely defines $\overline{m}$ which we will often use to simplify notation. In this vein, we will also use parameter $\kappa \geq 0$ defined as

$$\kappa \equiv \frac{(\overline{m} + 2\overline{M})}{\eta} - 2(\overline{m} + \overline{M}).$$

As in Cooper and Frieze (2003), we assume that parameters are such that $0 < \eta < 1$ holds.[6] Note that the structural parameters $A_1, B_1, C_1$ and $\boldsymbol{q}$ can only be partially identified from the asymptotic degree distribution, so we will focus on estimating $\boldsymbol{p}$, $\overline{M}$, and $\eta$.

tions are in Section 2.3. Importantly, the network is scale-free if and only if the attachment probability is asymptotically linear.

[6]In contrast to Cooper and Frieze (2003) who assume that $p_0 = 0$ and $q_0 = 0$, (i.e., each vertex has at least one neighbour), our model allows vertices to have zero degree to conform with real network data. Nevertheless, if $0 < \eta < 1$, all results and corresponding proofs from Cooper (2006) remain valid.

## 2.2 Asymptotic Degree Distribution

Our derivations of the asymptotic degree distribution rely on the results of Cooper (2006) presented in Appendix A. Proposition 1 restates expected degree sequence and concentration results of Cooper (2006) in a form convenient for our analysis.

**Proposition 1** *For* $0 \leq d \leq d^*(t;\eta)$, *where* $d^*(t;\eta) = \min\{t^{\eta/3}, t^{1/6}/\ln^2 t\}$, *and for any* $K > 1$, *we have the following*

$$\Pr\left(\left|\frac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\right| \geq K\frac{P(d;\eta,\overline{M},\boldsymbol{p})}{\sqrt{\ln t}}\right) = O\left(\frac{1}{\ln t}\right),$$

*where* $P(d;\eta,\overline{M},\boldsymbol{p})$ *is the asymptotic degree distribution given by*

$$P(d;\eta,\overline{M},\boldsymbol{p}) = \left(\sum_{m=0}^{\min\{P,d\}} p_m \frac{\Gamma(m+\kappa+1/\eta)}{\eta\Gamma(m+\kappa)}\frac{\Gamma(d+\kappa)}{\Gamma(d+\kappa+1+1/\eta)}\right) \tag{1}$$

*and* $\Gamma(\cdot)$ *is the gamma function.*

Corollary 1 demonstrates the limit properties of the degree distribution.

**Corollary 1** *We have the following:*

1. *The fraction* $P_t(d)$ *of vertices of the graph* $G(t)$ *that have degree* $d$ *converges in probability to* $P(d;\eta,\overline{M},\boldsymbol{p})$ *as* $t \to \infty$.

2. *The asymptotic degree distribution* $P(d;\eta,\overline{M},\boldsymbol{p})$ *has a power-law tail with the power-law parameter* $1 + 1/\eta$ :

   $$P(d;\eta,\overline{M},\boldsymbol{p}) = C(\eta,\overline{M},\boldsymbol{p})d^{-1-1/\eta}\left(1 + O\left(\frac{1}{d}\right)\right),$$

   *where* $C(\eta,\overline{M},\boldsymbol{p}) = \sum_{m=0}^{P} p_m\Gamma(m+\kappa+1/\eta)/(\eta\Gamma(m+\kappa))$.

3. *When the probability of preferential attachment tends to zero, the asymptotic degree distribution approaches a distribution proportional to the geometric distribution:*

   $$\lim_{\eta\to 0} P(d;\eta,\overline{M},\boldsymbol{p}) = \left(\sum_{m=0}^{\min\{P,d\}} p_m(1-\lambda)^{-m}\right)\lambda(1-\lambda)^d,$$

   *where* $\lambda = (2\overline{M} + \overline{m} + 1)^{-1}$ *is the parameter of the geometric distribution.*[7]

[7]The geometric distribution is the discrete analogue of the exponential distribution, as $(1-\lambda)^d = e^{\ln(1-\lambda)\,d}$.

## 2.3 Discussion and Examples

The CF model nests many network formation models in a sense that it is able to generate networks with degree distribution patterns ranging from the exponential degree distribution generated by the growing random graphs to the power law degree distribution of the preferential attachment networks and any hybrid models embedding the elements of both.[8] Importantly, we use the CF model to model only the degree distribution, rather than clustering and other characteristics of the social and economics networks. In particular, Theorem 5 of Bollobas and Riordan (2003) suggests that it is possible to introduce any level of clustering in a graph process with preferential attachment without changing the asymptotic degree distribution. Dorogovtsev and Mendes (2002) provide an example of such a process in Section IX C. Hence, the CF model may be extended to include any level of clustering. Similar arguments can be made about other network characteristics. Therefore, information about clustering and other characteristics is of limited use for estimating parameters $(\eta, \overline{M}, \boldsymbol{p})$ which solemnly determine the graph degree distribution.

The network models based on preferential attachment, including the CF model, provide good fit to observed physical, social and economic networks, but may lack rigorous microfoundations for individual strategic behaviour and structural interpretation of game-theoretic link formation models, as in Christakis et al. (2010) and Mele (2013). A growing literature attempts to fill this gap.

Jackson and Rogers (2007) build a growing network model using intuitive behavioural principles which may explain preferential attachment. Similarly to the CF model, in their model, a new vertex is added at every period. The new vertex considers edges with $m_r$ old vertices uniformly at random by "meeting strangers". In addition, the new vertex considers edges with $m_n$ direct neighbours of the previously "befriended" vertices by "meeting friends of friends". The new vertex chooses to create a considered edge when net marginal benefits of creating this edge are positive. The net marginal benefits of forming an edge are not

[8] Specifically, to obtain the preferential attachment graph of Barabasi and Albert (1999), set $p_m = 1,\ q_0 = 1$, and $A_1 = 1$, hence, $\eta = 1/2$ and $\overline{M} = 0$; for the hybrid graph in Chapter 5 of Jackson (2008) set $p_m = 1,\ q_0 = 1$, and $A_1 = 1 - \alpha$; for the hybrid graph in Pennock et al. (2002) set $p_0 = 1,\ q_m = 1$, and $B_1 = C_1 = \alpha$. The Dorogovtsev et al. (2000) and Buckley and Osthus (2004) setting, where the probability to be connected to a new vertex is proportional to the sum of initial attractiveness $A$ and degree $d(v, t-1)$, can be reflected in the CF model by setting $p_m = 1$, $q_0 = 1$, and $A_1 = 1/(1 + A/2m)$. The copying model of Kleinberg et al. (1999) and Kumar et al. (2000), in a simple version of which a new vertex either forms a random edge (with probability $\alpha$) or copies one edge from existing vertex (with probability $1 - \alpha$), is also covered by the CF model with $p_m = 1$, $q_0 = 1$, and $A_1 = 1 - \alpha$.

modelled explicitly, but are assumed to be positive with probability $p_r$ for strangers and with probability $p_n$ for friends of friends. The network is directed in a sense that edges go from the new vertex to the old vertices. The model has the properties of preferential attachment since the probability to be connected to the friend of the friend is proportional to its in-degree, $d_i(v, t-1)$. Therefore, the directed version of the CF model can replicate the asymptotic degree distribution of the Jackson and Rogers (2007) model by setting $\overline{m} = m = m_n p_n + m_r p_r$, $q_0 = 1, A_1 = p_r m_r/m, A_2 = 1 - A_1$ and modifying $p_A(v,t) = A_1 d_i(v, t-1)/|E(t-1)| + A_2/|V(t-1)|$ to reflect a directed graph.[9] Jackson and Rogers (2007) apply the model to web site links and various social networks including a co-authorship network, a citation network, friendship and romantic relationship networks. Atalay et al. (2011) apply a similar model incorporating growth and decay features to buyer-supplier relationships.

There is a well established literature with rigorous micro-foundations on strategic network formation starting from Jackson and Wolinsky (1996) and Bala and Goyal (2000). In these types of models a star network typically emerges as an equilibrium configuration. However, introducing some noise in the decision making process of agents in these strategic network formation models leads to emergence of networks with the preferential attachment, a special case of the CF model. Babus and Ule (2008) suggest a stylized example, in which agents gain access to information from others as long as they are at most two edges away in the network. They specify a simple marginal payoff an agent might gain from connecting to agent $v$ at time $t$, but also add idiosyncratic random noise in decision making. Using discrete choice logit framework, they show that the probability of forming an edge with an existing agent is proportional to the degree of this agent as in the preferential attachment network.

König (2015) considers more general marginal payoff functions with some idiosyncratic noise in the information sharing environment. Similarly to Jackson and Rogers (2007) setting, at every period a newly born agent can sample a given number (called observation radius) of existing agents (strangers) selected uniformly at random and can also observe neigbours of previously met agents (friends of friends). The key innovation of the paper is that benefits of forming an edge with sampled strangers and friends of friends are modelled explicitly. König (2015) finds that for small noise centralized star-type networks emerge irrespectively of the observation radius, but for larger noise and smaller observation radius the networks exhibiting preferential attachment and power law in degree distribution emerge. In this case the derived asymptotic degree distributions (König, 2015, Proposition 2) is similar to the asymptotic degree distribution in the CF model given by (1).

[9]The derivation follows directly from equation (1) in Jackson and Rogers (2007).

# 3 Methodology

## 3.1 Preliminaries

We propose the PMLE and GMM type estimators to estimate the parameters of the asymptotic degree distribution of the CF model. As shown in Corollary 1, the asymptotic degree distribution depends only on the subset of parameters of the model; specifically on $\eta$, $\overline{M}$, and $\boldsymbol{p} = (p_0, \ldots, p_P)$. Section 3.3 shows that these parameters are identified. From the setup of the model it is clear that $\eta \in (0,1)$, $\overline{M} \in [0,\infty)$ and $\boldsymbol{p} \in \Delta^P$, where $\Delta^P = \{\boldsymbol{p} \in \mathbb{R}_+^{P+1} : \sum_{i=0}^P p_i = 1\}$ is $P$ dimensional simplex.[10] We assume that the dimensionality $P$ of $\boldsymbol{p}$ is known; i.e., it is known how many parameters we need to estimate. In applications if $P$ is unknown, it can be chosen using information criteria such as AIC or BIC (see, e.g., Burnham and Anderson, 2002), but we do not explore asymptotic properties of such procedures.

Parameter $\eta$ is of the highest interest in this model. First, it determines the power-law parameter $1+1/\eta$, which is important for the statistical properties of the network. Second, it is equal to the limiting fraction of the edge endpoints inserted by the preferential attachment.

Let $\boldsymbol{\theta} = (\eta, \overline{M}, \boldsymbol{p})$, i.e. $\boldsymbol{\theta}$ is $P+3$ dimensional parameter with the domain $\boldsymbol{\Theta} = (0,1) \times [0,\infty) \times \Delta^P$. To represent the true value, a generic value, and an estimate, we write $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}$, and $\widehat{\boldsymbol{\theta}}$ respectively.

In Section 3.3 we derive asymptotic properties of the introduced estimators, $\widehat{\boldsymbol{\theta}}$, as $t$ goes to infinity. This asymptotic is similar to the standard large sample asymptotic, in which the number of observations goes to infinity. In the random graph process that we consider, one vertex and at most $P+Q$ edges are added at each time step $t$. Thus all asymptotic results will continue to hold if we consider an alternative asymptotic, in which the number of vertices $|V(t)|$ or the number of edges $|E(t)|$ of the graph $G(t)$ goes to infinity, since $|V(t)| \to \infty$, $|E(t)| \to \infty$ and $t \to \infty$ are equivalent.

## 3.2 Uniform Law of Large Numbers

Before introducing estimators and establishing their consistency we establish the uniform law of large numbers under non-standard conditions prevalent in growing network models. The standard regularity conditions for establishing consistency of estimators are continuity and uniform convergence. We can establish continuity by checking standard technical conditions for the distribution function $P(d;\boldsymbol{\theta})$ given by (1). However, we cannot establish uniform

[10] Formally, because of the assumption $\overline{m} + \overline{M} > 0$, whenever $\overline{M} = 0$ we should eliminate point $\boldsymbol{p} = (1, 0, ..., 0)$, which corresponds to $p_0 = 1$, from simplex $\Delta^P$.

convergence by using the standard uniform laws of large numbers for independent or weakly-dependent stationary data processes, because the CF model yields substantial heterogeneity in the vertex degrees and nonstandard vertex degree interdependencies. The main technical contribution of the paper is the uniform law of large numbers established for the CF model.[11]

**Proposition 2** *If $a(d;\boldsymbol{\theta})$ is a matrix of functions continuous in $\boldsymbol{\theta}$ on a compact set $\overline{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}$, and there is $F$ such that $\|a(d;\boldsymbol{\theta})\| < F \cdot (d+1)$ for all $d \in \mathbb{N}$ and all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, where $\|a(d;\boldsymbol{\theta})\| = \left(\sum_{j,k} a_{jk}^2\right)^{1/2}$ is the Euclidean norm, then we have the following.*

1. *$G_0(\boldsymbol{\theta}) = \sum_{d=0}^{\infty} a(d;\boldsymbol{\theta}) P(d;\boldsymbol{\theta}_0)$ is continuous in $\boldsymbol{\theta}$.*
2. *$\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \left\| \widehat{G}_t(\boldsymbol{\theta}) - G_0(\boldsymbol{\theta}) \right\| \xrightarrow{P} 0$, where $\widehat{G}_t(\boldsymbol{\theta}) = \sum_{d=0}^{\infty} a(d;\boldsymbol{\theta}) D_t(d)/|V(t)|$.*

To prove Proposition 2, we do not impose any specific dependence structure on the vertex degrees, but instead use the concentration result of Proposition 1. To illustrate the key steps of the proof, suppose that $a(d;\boldsymbol{\theta})$ is a function $a(d)$ that does not depend on $\boldsymbol{\theta}$ and satisfies $0 < a(d) < d$. Part 1 of Proposition 2 holds because $\sum_{d=0}^{n} a(d) P(d;\boldsymbol{\theta_0})$ is a converging series as follows from $\eta_0 < 1$, $a(d) < d$, and $P(d;\boldsymbol{\theta_0})$ being approximately proportional to $d^{-1-1/\eta_0}$ by part 2 of Corollary 1. To prove part 2, we bound $\left|\widehat{G}_t(\boldsymbol{\theta}) - G_0(\boldsymbol{\theta})\right|$ by the sum of the three terms as follows

$$\left|\widehat{G}_t(\boldsymbol{\theta}) - G_0(\boldsymbol{\theta})\right| \leq \underbrace{\sum_{d=\widetilde{d}(t)}^{\infty} a(d) P(d;\boldsymbol{\theta}_0)}_{\widehat{S}_1} + \underbrace{\sum_{d=0}^{\widetilde{d}(t)-1} a(d) \left| \frac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0) \right|}_{\widehat{S}_2} + \underbrace{\sum_{d=\widetilde{d}(t)}^{\infty} a(d) \frac{D_t(d)}{|V(t)|}}_{\widehat{S}_3},$$

and show that each term converges in probability to zero if $\widetilde{d}(t)$ grows to infinity but much slower than $d^*(t;\eta)$ and $\ln t$. $\widehat{S}_1 \to 0$ again by part 2 of Corollary 1. $\widehat{S}_2 \xrightarrow{P} 0$ by the concentration result of Proposition 1. Finally, $\widehat{S}_3$ can be bounded above by $\sum_{d=\widetilde{d}(t)}^{\infty} d D_t(d)/|V(t)|$, which is equal to the difference between $\widehat{S}_4 = \sum_{d=0}^{\infty} d D_t(d)/|V(t)|$ and $\widehat{S}_5 = \sum_{d=0}^{\widetilde{d}(t)-1} d D_t(d)/|V(t)|$. $\widehat{S}_5 \xrightarrow{P} 2(\overline{m}_0 + \overline{M}_0)$ by Proposition 1 and part 1 of Corollary 1. Finally, $\widehat{S}_4 \xrightarrow{P} 2(\overline{m}_0 + \overline{M}_0)$ by the law of large numbers applied to independent draws of $m(t) + M(t)$. Therefore, $\widehat{S}_3 \xrightarrow{P} 0$, and part 2 of Proposition 2 follows.

[11] This section closely follows Newey and McFadden (1994) notation. Symbols $\rightsquigarrow$ and $\xrightarrow{P}$ stand for convergence in distribution and probability respectively. $O_P(1)$ and $o_P(1)$ are stochastic order symbols, formally defined in van der Vaart (2000).

### 3.3 Consistency of PML and GMM Estimators

The established uniform law of large numbers allows us to extend the standard consistency results to our network formation model. We define the pseudo log-likelihood based on the asymptotic degree distribution in (1) as follows:

$$\widehat{L}_t(\boldsymbol{\theta}) \;=\; \sum_{d=0}^{\infty} \frac{D_t(d)}{|V(t)|} \ln P(d;\boldsymbol{\theta}). \tag{2}$$

The true log-likelihood is different from the pseudo log-likelihood, because (i) degrees of vertices are interdependent, and (ii) a finite sample (not asymptotic) degree distribution should be used.

The PML estimator is defined as:

$$\widehat{\boldsymbol{\theta}}^{\text{PML}} = \arg\max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{L}_t(\boldsymbol{\theta}). \tag{3}$$

The plug-in PML estimator is formally defined as:

$$\widehat{\boldsymbol{\theta}}^{\text{plPML}} = \arg\max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{L}_t(\boldsymbol{\theta}), \tag{4}$$

$$\text{s.t. } \overline{m} + \overline{M} = \frac{1}{2}\sum_{d=0}^{\infty} d\frac{D_t(d)}{|V(t)|}. \tag{5}$$

That is, $\widehat{\boldsymbol{\theta}}^{\text{plPML}}$ is obtained by replacing $\overline{m}+\overline{M}$ in (2) with its estimate $\widehat{\overline{m}+\overline{M}} = \frac{1}{2}\sum_{d=0}^{\infty} d\frac{D_t(d)}{|V(t)|}$ and maximizing (2) over the remaining parameters $\eta$ and $\boldsymbol{p}$; so $\widehat{\boldsymbol{\theta}}^{\text{plPML}}$ is faster to compute than $\widehat{\boldsymbol{\theta}}^{\text{PML}}$, because it requires maximization over one less parameter, $\overline{M}$. A rationale for this estimator is that $\widehat{\overline{m}+\overline{M}}$ is a consistent estimate of $\overline{m}_0 + \overline{M}_0$ with the standard asymptotic:

$$\sqrt{|V(t)|}\left(\widehat{\overline{m}+\overline{M}} - \overline{m}_0 - \overline{M}_0\right) \rightsquigarrow \mathcal{N}(0, \text{Var}(m(t)) + \text{Var}(M(t))),$$

because $m(t)$ and $M(t)$ are independent of each other and across time $t$.

Consistency of the PML and the plug-in PML estimators is established in Propositions 3.

**Proposition 3** *Let $\overline{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}$ be compact and $\boldsymbol{\theta}_0 \in \overline{\boldsymbol{\Theta}}$. If $\widehat{\boldsymbol{\theta}}$ satisfies $\widehat{L}_t(\widehat{\boldsymbol{\theta}}) \geq \max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{L}_t(\boldsymbol{\theta}) + o_P(1)$ where $\widehat{L}_t(\boldsymbol{\theta})$ is given by (2), then $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$. In particular, $\widehat{\boldsymbol{\theta}}^{\text{PML}} \xrightarrow{P} \boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}^{\text{plPML}} \xrightarrow{P} \boldsymbol{\theta}_0$.*

We now consider a more general class of GMM estimators. A GMM estimator $\widehat{\boldsymbol{\theta}}$ is defined as $\boldsymbol{\theta}$ that maximizes

$$\widehat{Q}_t(\boldsymbol{\theta}) = -\left[\sum_{d=0}^{\infty} g(d;\boldsymbol{\theta})\frac{D_t(d)}{|V(t)|}\right]' \widehat{W} \left[\sum_{d=0}^{\infty} g(d;\boldsymbol{\theta})\frac{D_t(d)}{|V(t)|}\right], \tag{6}$$

where $\widehat{W}$ is a positive semi-definite matrix and the *moment function* vector $g(d;\boldsymbol{\theta})$ satisfies

$$\sum_{d=0}^{\infty} g(d;\boldsymbol{\theta}_0)P(d;\boldsymbol{\theta}_0) = 0. \tag{7}$$

Since (1) gives the explicit expression for $P(d;\boldsymbol{\theta})$, it is easy to verify whether a given $g(d;\boldsymbol{\theta})$ satisfies (7). In particular, from the discussion of the PML estimators, it is evident that (7) is satisfied for the moment function vector that consists of the *score function* vector and the *degree function*:

$$g(d;\boldsymbol{\theta}) = \left(\nabla_{\boldsymbol{\theta}} \ln P(d;\boldsymbol{\theta}),\ d - 2(\overline{m} + \overline{M})\right)'. \tag{8}$$

Proposition 4 specifies sufficient conditions on moment function $g(d;\boldsymbol{\theta})$ and matrix $\widehat{W}$ for the GMM estimate $\widehat{\boldsymbol{\theta}}$ to be consistent.

**Proposition 4** *Let $\widehat{\boldsymbol{\theta}}$ maximize (6) where $\widehat{W} \xrightarrow{P} W$, (i) $W \sum_{d=0}^{\infty} g(d;\boldsymbol{\theta})P(d;\boldsymbol{\theta}_0) = 0$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; and (ii) $\boldsymbol{\theta}_0 \in \overline{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}$ where $\overline{\boldsymbol{\Theta}}$ is compact.*

1. *If (iii) $g(d;\boldsymbol{\theta})$ is continuous on $\overline{\boldsymbol{\Theta}}$; and (iv) there is $F$ such that $\|g(d;\boldsymbol{\theta})\| < F \cdot (d+1)$ for all $d \geq 0$ and all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, then $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.*

2. *If $g(d;\boldsymbol{\theta})$ is given by (8), then $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.*

Conditions (i), (ii), and (iii) are standard identification, compactness, and continuity assumptions (see Theorem 2.6 of Newey and McFadden, 1994). Condition (iv) is required for the uniform law of large numbers established in Proposition 2.

We suggest using the moment function vector given by (8). With appropriately chosen weights in $\widehat{W}$, GMM estimators based on this moment function vector nest the PML and plug-in PML estimators when they are viewed as solutions to their first-order conditions. In particular, $\widehat{\boldsymbol{\theta}}^{\text{PML}}$ is a solution to

$$\sum_{d=0}^{\infty} \nabla_{\boldsymbol{\theta}} \ln P(d;\boldsymbol{\theta}) \frac{D_t(d)}{|V(t)|} = 0;$$

so it can be viewed as a GMM estimator with $\widehat{W}$ that puts the full weight on the score function (and no weight on the degree function).

As part 2 of Proposition 4 shows, for the consistency of GMM estimators based on the moment function vector (8), we only need to check the identification condition (i). In contrast to the identification of the PML and plug-in PML estimators established in Proposition 3, it is difficult to specify primitive conditions on $g(d;\boldsymbol{\theta})$ and $W$ such that the identification

condition holds. A common practice in the GMM literature, therefore, is to simply assume identification (see, e.g., p. 2127 of Newey and McFadden, 1994).[12]

## 3.4 Discussion of Asymptotic Normality and Variance

We now specify sufficient conditions for establishing asymptotic normality of GMM estimators based on the moment function vector (8), and, thus, of the PML and plug-in PML estimators.

**Proposition 5** *Let* $\widehat{\boldsymbol{\theta}}$ *maximize (6) where* $g(d;\boldsymbol{\theta})$ *is given by (8),* $\widehat{W} \xrightarrow{P} W$, $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$, *and* $\boldsymbol{\theta}_0 \in$ *interior*$(\boldsymbol{\Theta})$. *If (i)* $G'WG$ *is nonsingular where* $G = \sum_{d=0}^{\infty} \nabla_{\boldsymbol{\theta}} g(d;\boldsymbol{\theta}_0) P(d;\boldsymbol{\theta}_0)$ *and (ii)* $\sqrt{|V(t)|}\sum_{d=0}^{\infty} g(d;\boldsymbol{\theta}_0) D_t(d)/|V(t)| \rightsquigarrow N(0,\Sigma)$, *then*

$$\sqrt{|V(t)|}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \rightsquigarrow N\left[0, (G'WG)^{-1} G'W\Sigma WG\,(G'WG)^{-1}\right].$$

Condition (i) holds under local identification. Condition (ii) is an asymptotic normality condition for a sample average of $g(d;\boldsymbol{\theta}_0)$. If we could assume independence this condition would follow from a central limit theorem (CLT). [13] Simulations suggest that this condition holds. Informally asymptotic normality is suggested by the results of Cooper (2006, Theorem 2.2), who shows that most vertex degrees are asymptotically independently distributed according to a negative binomial distribution. The asymptotic normality of the estimators is supported by our simulations. However, the simulations show that the independence does not hold in finite samples which has implications for the asymptotic variance.

To get a consistent estimate of the asymptotic variance of $\widehat{\boldsymbol{\theta}}$, we need to find consistent estimates of $G$ and $\Sigma$.[14] A consistent estimate of $G$ can be obtained by $\widehat{G} = \sum_{d=0}^{\infty} g(d;\widehat{\boldsymbol{\theta}}) \frac{D_t(d)}{t}$,[15]

[12] It is easier to verify a local identification condition, which requires that there is a unique solution to $W\sum_{d=0}^{\infty} g(d;\boldsymbol{\theta}) P(d;\boldsymbol{\theta}_0) = 0$ only in some neighbourhood of $\boldsymbol{\theta}_0$. By Rothenberg (1971), a sufficient condition for local identification is that $WG$ has full column rank, where $G = \sum_{d=0}^{\infty} \nabla_{\boldsymbol{\theta}} g(d;\boldsymbol{\theta}_0) P(d;\boldsymbol{\theta}_0)$. At the end of the proof of Proposition 5 we derive $G$ for the moment function vector (8); so for given $\boldsymbol{\theta}_0$ and $W$, we can verify local identification – in particular, it holds for GMM estimators used in our simulations.

[13] Under independence and certain additional assumptions this condition follows from a central limit theorem (CLT). Formally, we cannot assume independence because of interdependencies in vertex degrees. Jenish and Prucha (2009, 2012) establish CLT for weakly dependent spatial processes, which requires certain mixing or near-epoch dependence. It is non-trivial to verify whether these conditions hold in our context.

[14] By assumption (ii) of Proposition 5, $\widehat{W} \xrightarrow{P} W$ and $G'WG$ is nonsingular. If in addition $\widehat{G} \xrightarrow{P} G$ and $\widehat{\Sigma} \xrightarrow{P} \Sigma$, then by continuous mapping theorem, $\left(\widehat{G}'\widehat{W}\widehat{G}\right)^{-1}\widehat{G}'\widehat{W}\widehat{\Sigma}\widehat{W}\widehat{G}\left(\widehat{G}'\widehat{W}\widehat{G}\right)^{-1} \rightarrow (G'WG)^{-1}G'W\Sigma WG\,(G'WG)^{-1}$.

[15] Consistency, continuity, and uniform convergence imply: $\left\|\widehat{G} - G\right\| \leq \left\|\widehat{G} - G(\widehat{\boldsymbol{\theta}})\right\| + \left\|G(\widehat{\boldsymbol{\theta}}) - G\right\| \leq \sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \left\|\sum_{d=0}^{\infty} g(d;\widehat{\boldsymbol{\theta}})\frac{D_t(d)}{t} - G(\boldsymbol{\theta})\right\| + \left\|G(\widehat{\boldsymbol{\theta}}) - G\right\| \xrightarrow{P} 0$.

but it is difficult to get a consistent estimate of $\Sigma$, because vertex degrees are interdependent and are not identically distributed.

To overcome this issue, we propose using parametric bootstrap (see, e.g., Efron and Tibshirani, 1994) to compute standard errors of the parameter estimates. The procedure amounts to resampling from the parametric CF model given the estimated values of the parameters. The complication with this procedure is that some of the structural parameters of the network are only partially identified. Namely, parameters $A_1$, $B_1$ and $C_1$ are between 0 and 1 and are bound by the relationship in (1). When $\overline{M} = 0$ ($q_0 = 1$), we have complete identification, $A_1 = 2\eta$, and we can easily implement the parametric bootstrap. In other cases parameter $\boldsymbol{q}$ is not identified, but if we assume a lower bound and an upper bound for $M$, say 0 and $Q$, by specifying $q_0 = 1 - \overline{M}/Q$ and $q_Q = \overline{M}/Q$ for given $\overline{M}$ the variance of the estimator for $\overline{M}$ will be maximized given all other parameters. The minimum variance for the estimator of $\overline{M}$ can be achieved by setting $q_{\lfloor \overline{M} \rfloor} = \overline{M} - \lfloor \overline{M} \rfloor$ and $q_{\lceil \overline{M} \rceil} = \lceil \overline{M} \rceil - \overline{M}$, where $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the largest integer not greater than $x$ and the smallest integer not less than $x$, respectively. The variance of the estimator for $\eta$ will be maximized by using $\boldsymbol{q}$ that maximizes the variance of $\overline{M}$ and after assuring that the maximized variance of $\overline{M}$ is greater than the variance of $\overline{m}$, setting $A_1 = 0$ and $B_1$ as close as possible to $C_1$, e.g. for $\eta = 0.5, p_0 = 1$ set $B_1 = C_1 = 0.5$. The variance for $\eta$ will be minimised for $\boldsymbol{q}$ minimizing the variance of $\overline{M}$ and after making sure that is smaller than the variance of $\overline{m}$ setting $A_1 = 0$ and $B_1$ as apart as possible to $C_1$, e.g., for $\eta = 0.5, p_0 = 1$ set $B_1 = 1, C_1 = 0$. This way we can find the lower bound and the upper bound for standard errors using the parametric bootstrap. In application we use 1000 replications.

**Asymptotic efficiency** Newey and McFadden (1994, Theorem 5.2) implies that the GMM estimator with $\widehat{W} \xrightarrow{P} \Sigma^{-1}$ is asymptotically efficient in the class of GMM estimators. We applied the above bootstrap procedure to estimate variance of the moments and use its inverse as an estimate of the optimal weighting matrix. However, our simulations indicate that in finite samples using the identity weighting matrix produces similar and sometimes better results in terms of the MSE in comparison to using the estimate of the optimal weighting matrix. Similar finite sample results are often found in the GMM literature (see, e.g., Altonji and Segal, 1996).

# 4 Simulations

We investigate finite sample performance of the GMM (including PML), NLLS, and several local tail exponent estimators including the discrete local tail exponent (DLTE) estimator, the log-log rank-degree regression with Gabaix and Ibragimov (2011) correction (hereafter GI), and Hill (1975) estimator with Clauset et al. (2009) correction (hereafter Hill). The NLLS and the local tails exponent estimators are formally discussed and derived for the CF model in Appendix C.

We compare estimators using sample statistics of parameter estimates, namely, sample mean, sample standard deviation, bias, a difference between the sample mean of an estimate and its true value, and the mean squared error between an estimate and the true value. We also perform the Anderson-Darling (AD) test for normality of the estimate and reports its $p$-value. For the local tail exponent estimators we report the median value of trashhold $d_{\text{tr}}$ selected using the AD distance (see Appendix C for details). All simulation results are based on 10000 replications.

As a benchmark, we consider the CF model with the following parameters: $t = 1000$, $\eta = 0.5$, $p_0 = 1$ ($m(t) = 0$), and $q_1 = q_2 = 1/2$ ($\overline{M} = 1.5$). Similar to Bollobas et al. (2001), we assume that initial graph $G(1)$ consists of one vertex and a random number $\max\{m(1)+M(1), 1\}$ of loops.[16] We compare the GMM estimator using the moment function vector with the score function and the degree function as in (8) and the $\text{GMM}_{-\overline{M}}$ estimator excluding $\overline{M}$-component of the score function from the moment vector. In this treatment, the PML and GMM estimators assume the correct model ($P = 0$).

Table 1 compares the results for all considered estimators. In terms of $\eta$, the PMLE and GMM class estimators are effectively unbiased, while the NLLS and the local tail exponent estimators show some bias. Similarly in terms of the standard error the PMLE and GMM outperform the other estimators. The AD test suggests normality for that $\eta$ estimates from the PMLE and GMM estimators and rejects normality for NLLS and the local tail exponent estimators. Overall, the GMM estimator based on the score function and the degree function attains the smallest MSE, though, its performance is closely comparable to the plug-in PLME and $\text{GMM}_{-\overline{M}}$ estimators.

**<<Place Table 1 about here>>**

Table 2 investigates how bias of the considered estimators changes with sample size or

[16] As a robustness check we also considered different initial graphs, but the results were not effected, which supports ergodicity for the CF model.

the number of vertices, $t$, (assuming $|V(1)| = 1$). We consider the benchmark parameters and vary $\eta$. The bias of the GMM class of estimators (including PMLE) is virtually 0 and reduces with $t$. The bias of the NLLS estimator is relatively high and reduction with $t$ is rather slow. The bias of DLTE and Hill estimators is initially large, but there is some reduction with $t$, while the bias of the GI estimator is not very large initially, but it does not reduce (but sometimes increase) with $t$. These results support consistency for the GMM class of estimators and suggest that convergence of NLLS estimator is rather slow, while the GI estimator does not converge to the true value even in large samples. Also the bias depends on the value of $\eta$ in a non-trivial way.

**<<Place Table 2 about here>>**

Figure 1 explores the behaviour of the standard deviations of the estimates for $\eta$ for benchmark parameters and different $\eta$-s as $t$ increases. For brevity for the class of the GMM estimators we report only the PMLE and the GMM using the score and the degree moment function. The standard errors of the plug-in PMLE and $\text{GMM}_{-\overline{M}}$ estimators are very close to this GMM estimator. The GMM-based estimates attain the smallest standard deviations for all $t$ and different values of parameter $\eta$. Moreover, the standard deviation of the PMLE and GMM estimates of $\eta$ appears to decrease with the rate $\sqrt{t}$, which suggests that these estimators are $\sqrt{t}$-consistent. While normality of the GMM-class based estimates sometimes does not hold for $t = 1000$, the simulations (not shown here for brevity) suggest that for larger samples the normality seem to hold. For the other considered estimators the convergence results are noisy and depend on the value to the parameters $\eta$. Out of these estimators the NLLS seem to perform better than the local tail exponent estimators for large $t$, while the GI-based estimates show smaller standard deviation for small $t$. Similarly, for these estimators normality is rejected even in large samples.

**<<Place Figure 1 about here>>**

Next, we investigate the behaviour of the GMM estimator under overspecification, when the assumed $P$ is larger than the actual $P$, and misspecification, when the order is reversed. Table 3 report the results for the true model with $P = 1$ and $p_0 = p_1$ and the rest of the parameters as in the benchmark. The assumed order is the GMM of estimators is indicated by superscript $P$, e.g., $\text{GMM}^0$ denoted the misspecified model with $P = 0$. In case of overspecification when the assumed order is $P = 2$, the GMM estimator is less efficient for all parameters, but the bias remains rather small. For $\eta$ it is still smaller than for the

NLLS and the local tail exponent estimators. However, in case of misspecification, when the assumed order is $P = 0$, the GMM estimator shows a substantial bias, in both $\eta$ and $\overline{M}$ which is much higher than the bias of NLLS and the local tail exponent estimators. In the this sense the GMM estimator is nor robust to the misspecification. We can use BIC criterion the perform model selection. The last raw of Table 3 reports an average values of the BIC criterion, which suggest that on the correct specified model with $P = 1$ would be often selected. The behaviour of the other types of the GMM estimators (including the PMLE) is qualitatively similar.

**<<Place Table 3 about here>>**

Lastly, Table 4 explores how the properties of the GMM class of estimators change with other parameters. Columns $1 - 4$ present the results for values of $q_0 = q_3 = 1/2$ keeping all other parameters fixed at the benchmark level. The GMM estimator outperforms the PMLE estimator, but generally the bias is small. Columns $5 - 8$ present results for another modification, when $B_1 = 0$ and $C_1 = 1$ instead of $B_1 = C_1 = 1/2$; i.e., the initial vertex of each edge is chosen uniformly at random and the terminal vertex of each edge is chosen by preferential attachment. Again, the GMM estimator performs a bit better than the PMLE estimator.

**<<Place Table 4 about here>>**

# 5 Application to Co-authorship Network

We illustrate the usefulness of the introduced methods by estimating the CF model for the network of coauthorship relations among economists publishing in journals listed by EconLit in the 1990s. This dataset was first considered by Goyal et al. (2006). They produced a network of collaboration in which every publishing author is a vertex in the network, and two vertices are linked with an edge if they have published a paper or more together within in the period of ten years between 1990 and 1999. The network contains $t = 81217$ authors with average number of coauthors 1.672, i.e. $\widehat{\overline{m} + \overline{M}} = 0.836$. Jackson and Rogers (2007) used the NLLS to estimate the parameters for this network.

As we do not know the support of $m(t)$, we use the BIC criterion, which selects order $P = 2$, that is, $m(t) \in \{0, 1, 2\}$. We compute the standard errors of the parameters using the parametric bootstrap. As the estimate of $\overline{M}$ is nearly 0, all structural parameters of the

model are identified and it is straightforward to perform the parametric bootstrap. We use 1000 replications.

Table 5 presents the parameter estimates and their standard errors for the coauthorship network. For local exponent estimators we also report thrashhold $d_{\text{tr}}$. There is a substantial difference between the parameter estimates found by the GMM estimators, the NLLS and the local tail exponent estimators. Our estimate of $\eta$ using the NLLS is 0.19 which is close to the estimate of 0.21 in Jackson and Rogers (2007).[17] One of the peculiarities of this data is that about 99% of the observations have degree between 0 and 10, while the highest degree in the sample is 54. This explain the substantial variation in the parameter estimates. Under these conditions the tail exponent estimators are very sensitive to $d_{\text{tr}}$ and we observe rather different results for all three local tail exponent estimators.

Structurally, the CF model suggests that the number of old-old egdes is negligibly small and the network is mostly formed by new-old edges with majority of new vertexes about 43% having one edge (co-author). In 37% of case the new vertices have no edges (single-authored papers) and only in about 20% of cases new vertex have two co-authors. Also, we conjecture that about 76% of new-old connection are done by preferential attachment. There is a large number of collaborations between graduate students and their (former) supervisors, but it is not common to coauthor with more than one supervisor. Single-authors papers are fairly standard for graduating economists. Preferential attachment mechanism is natural in this setting as successful professors attract more new graduate students and grow their network of collaborations.

**<<Place Table 5 about here>>**

# 6 Conclusion

We estimate a general model of scale-free network formation using the PMLE, GMM, NLLS, and local tail exponent estimators. We prove consistency of the GMM class of estimators and derive conditions for asymptotic normality of these estimators. Our simulations indicate that GMM class of estimators (including PMLE) produce considerably better estimates with virtually no bias and smaller standard errors than the other considered estimators if the model is correctly specified. However, the GMM estimators are less robust to misspecification compared to the NLLS and the local tail exponent estimator. Our simulations suggest that the misspecification can be avoided using model selection based on BIC.

[17] Their parametrization is different and they actually estimate $1/\eta$ which they find to be equal to 4.7

The results of this paper are useful for a new and growing literature on estimation of network formation models. The methodology for establishing the UNLL can be extended to other growing network models. For example, Atalay et al. (2011) and Atalay (2013) use the PML estimator to estimate similar network formation models. Using the methodology developed in this paper, one can prove consistency of the PML estimator for these models. Moreover, our simulations suggest that some GMM estimators yield 33% smaller standard errors of estimates than the PML estimator does; so one can estimate model parameters more precisely using our general class of GMM estimators, which includes the PML estimator as a special case.

One of the challenges with the growing networks models is non-trivial dependencies in the degree distribution. We hope that further research will lead to better characterisation of these dependencies, which, in turn, will help formalising the procedures for establishing the normality and estimating the asymptotic variance of the estimators.

## Appendix A: Results for Degree Distribution

Following Cooper (2006), define $d^*(t;\eta)$ as $d^*(t;\eta) = \min\{t^{\eta/3}, t^{1/6}/\ln^2 t\}$ and $n_m(d;\eta,\kappa)$ for $d \in \{m, m+1, \dots\}$ as

$$n_m(d;\eta,\kappa) = \frac{B(d+\kappa, 1+1/\eta)}{B(m+\kappa, 1/\eta)} = \frac{\Gamma(m+\kappa+1/\eta)}{\eta\Gamma(m+\kappa)} \frac{\Gamma(d+\kappa)}{\Gamma(d+\kappa+1+1/\eta)}, \tag{9}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function and $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ is the Beta function. The second equality in (9) follows from $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$. Notice that $n_m(d;\eta,\kappa)$ is a probability distribution because $n_m(d;\eta,\kappa) \geq 0$ for $d \geq m$ and $\sum_{d=m}^\infty n_m(d;\eta,\kappa) = 1$, where the latter follows from $B(x+1,y) = B(x,y)x/(x+y)$.

To present Cooper (2006)'s main result, we define $D_t(d,m)$ as the number of vertices of the graph $G(t)$ with initial degree $d(v,v) = m$ and current degree $d(v,t) = d$. Following Cooper (2006), the equations with terms like $O(1/\ln t)$ should be treated as inequalities giving upper and lower bounds (no explicit functional form is implied). Constants in error terms like $O(1/\ln t)$ may depend on parameters of the model but not on $d$.

**Lemma A.1** *For $m \leq d \leq d^*(t;\eta)$, we have the following:*

1. *expected degree sequence*

$$\mathbb{E}D_t(d,m) = p_m n_m(d;\eta,\kappa)t\left(1 + O\left(\frac{1}{\ln t}\right)\right),$$

*2. concentration*

$$\Pr\left(|D_t(d,m) - \mathbb{E}D_t(d,m)| \geq \frac{\mathbb{E}D_t(d,m)}{\sqrt{\ln t}}\right) = O\left(\frac{1}{\ln t}\right).$$

**Proof of Lemma A.1.** See the proof of Theorem 2.1 in Cooper (2006). ■

Since the initial degrees of vertices are not observed in real networks, we need to extend Lemma A.1 in the following way for our analysis.

**Proposition A.1** *For $0 \leq d \leq d^*(t;\eta)$, we have the following:*

*1. expected degree sequence*

$$\mathbb{E}D_t(d) = P(d;\eta,\overline{M},\boldsymbol{p})t\left(1 + O\left(\frac{1}{\ln t}\right)\right),$$

*2. concentration*

$$\Pr\left(|D_t(d) - \mathbb{E}D_t(d)| \geq \frac{\mathbb{E}D_t(d)}{\sqrt{\ln t}}\right) = O\left(\frac{1}{\ln t}\right).$$

**Proof of Proposition A.1.** Summing up expressions from part 1 of Lemma A.1 gives part 1 of Proposition A.1. The following sequence of inequalities establishes part 2:

$$\begin{aligned}
\Pr\left(|D_t(d) - \mathbb{E}D_t(d)| \geq \tfrac{\mathbb{E}D_t(d)}{\sqrt{\ln t}}\right) &= \Pr\left(\left|\textstyle\sum_{m=0}^{\min\{P,d\}}\left(D_t(d,m) - \mathbb{E}D_t(d,m)\right)\right| \geq \tfrac{\sum_{m=0}^{\min\{P,d\}}\mathbb{E}D_t(d,m)}{\sqrt{\ln t}}\right) \\
&\leq \Pr\left(\textstyle\sum_{m=0}^{\min\{P,d\}}|D_t(d,m) - \mathbb{E}D_t(d,m)| \geq \tfrac{\sum_{m=0}^{\min\{P,d\}}\mathbb{E}D_t(d,m)}{\sqrt{\ln t}}\right) \\
&\leq \Pr\left(\exists m : |D_t(d,m) - \mathbb{E}D_t(d,m)| \geq \tfrac{\mathbb{E}D_t(d,m)}{\sqrt{\ln t}}\right) \\
&\leq \textstyle\sum_{m=0}^{\min\{P,d\}}\Pr\left(|D_t(d,m) - \mathbb{E}D_t(d,m)| \geq \tfrac{\mathbb{E}D_t(d,m)}{\sqrt{\ln t}}\right) = O\left(\tfrac{1}{\ln t}\right).
\end{aligned}$$

■

# Appendix B: Main Proofs

**Proof of Proposition 1.** The proof follows from Proposition A.1 and the following sequence of equalities and inequalities:

$$\begin{aligned}
&\Pr\left(|D_t(d) - \mathbb{E}D_t(d)| \geq \tfrac{\mathbb{E}D_t(d)}{\sqrt{\ln t}}\right) \\
&= \Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\tfrac{t}{t+|V(1)|-1}\left(1 + O\left(\tfrac{1}{\ln t}\right)\right)\right| \geq \tfrac{P(d;\eta,\overline{M},\boldsymbol{p})}{\sqrt{\ln t}}\tfrac{t}{|V(t)|}\left(1 + O\left(\tfrac{1}{\ln t}\right)\right)\right) \\
&= \Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\right| \geq \tfrac{P(d;\eta,\nu,\boldsymbol{p})}{\sqrt{\ln t}}\left(1 + O\left(\tfrac{1}{\ln t}\right)\right)\right) \\
&\geq \Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\right| \geq K\tfrac{P(d;\eta,\overline{M},\boldsymbol{p})}{\sqrt{\ln t}}\right),
\end{aligned}$$

for any $K > 1$ and sufficiently large $t$. ■

**Proof of Corollary 1.** To prove part 1, we note that $d^*(t;\eta) \to \infty$ as $t \to \infty$; so $d \leq d^*(t;\eta)$ and Corollary 1 applies. Part 2 follows from the well-known result (see, e.g., Palumbo, 1998) $\Gamma(z+\alpha)/\Gamma(z+\beta) = z^{\alpha-\beta}(1+O(1/z))$ applied to $\Gamma(d+\kappa)/\Gamma(d+\kappa+1+1/\eta)$. Part 3 follows from $\Gamma(z+1) = z\Gamma(z)$ applied to (9):

$$n_m(d;\eta,\overline{M},\overline{m}) = \frac{\left(d-1-2(\overline{m}+\overline{M})+(\overline{m}+2\overline{M})/\eta\right)\ldots\left(m-2(\overline{m}+\overline{M})+(\overline{m}+2\overline{M})/\eta\right)}{\eta\left(d-2(\overline{m}+\overline{M})+(\overline{m}+2\overline{M}+1)/\eta\right)\ldots\left(m-2(\overline{m}+\overline{M})+(\overline{m}+2\overline{M}+1)/\eta\right)} \underset{\eta\to 0}{\to} \frac{1}{\overline{m}+2\overline{M}+1}\left(\frac{\overline{m}+2\overline{M}}{\overline{m}+2\overline{M}+1}\right)^{d-m}.$$

■

**Proof of Proposition 2.**

**Part 1.** We first prove that $G_0^n(\boldsymbol{\theta}) = \sum_{d=0}^{n} a(d;\boldsymbol{\theta})P(d;\boldsymbol{\theta}_0)$ converges uniformly on $\overline{\boldsymbol{\Theta}}$ to $G_0(\boldsymbol{\theta})$. Palumbo (1998) shows that

$$\frac{\Gamma(k+\lambda)}{\Gamma(k+1)} > \ (k+1)^{\lambda-1} \quad \text{for } \lambda > 2 \text{ and } k \geq 0.$$

Thus for $d \geq 1-\kappa_0$

$$P(d;\boldsymbol{\theta}_0) = \sum_{m=0}^{\min\{P,d\}} p_{m0}\frac{\Gamma(m+\kappa_0+1/\eta_0)}{\eta_0\Gamma(m+\kappa_0)}\frac{\Gamma(d+\kappa_0)}{\Gamma(d+\kappa_0+1+1/\eta_0)} < C(\boldsymbol{\theta}_0)(d+\kappa_0)^{-1-1/\eta_0}, \tag{10}$$

where $C(\boldsymbol{\theta}_0) = \sum_{m=0}^{P} p_{m0}\Gamma(m+\kappa_0+1/\eta_0)/(\eta_0\Gamma(m+\kappa_0))$. Thus,

$$\|a(d;\boldsymbol{\theta})P(d;\boldsymbol{\theta}_0)\| < C(\boldsymbol{\theta}_0)(d+\kappa_0)^{-1-1/\eta_0}F(d+1) = J_d.$$

Evidently, $\sum_{d=0}^{\infty} J_d < \infty$. Thus $G_0^n(\boldsymbol{\theta})$ converges uniformly on $\overline{\boldsymbol{\Theta}}$ to $G_0(\boldsymbol{\theta})$ (Theorem 7.10 of Rudin, 1976). Moreover, since $G_0^n(\boldsymbol{\theta})$ is continuous on $\overline{\boldsymbol{\Theta}}$, $G_0(\boldsymbol{\theta})$ is also continuous on $\overline{\boldsymbol{\Theta}}$ (Theorem 7.12 of Rudin, 1976).

**Part 2.** Clearly, there exists $\widetilde{d}(t)$ such that $\widetilde{d}(t) \leq d^*(t;\eta)$, $\widetilde{d}(t) \to \infty$, and $\widetilde{d}(t)/\ln t \to 0$ as $t \to \infty$. Then,

$$\begin{aligned}\left\|\widehat{G}_t(\boldsymbol{\theta}) - G_0(\boldsymbol{\theta})\right\| &= \left\|\sum_{d=0}^{\widetilde{d}(t)-1} a(d;\boldsymbol{\theta})\left(\frac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right) + \sum_{d=\widetilde{d}(t)}^{\infty} a(d;\boldsymbol{\theta})\left(\frac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right)\right\| \\ &\leq \underbrace{\left\|\sum_{d=\widetilde{d}(t)}^{\infty} a(d;\boldsymbol{\theta})P(d;\boldsymbol{\theta}_0)\right\|}_{\widehat{S}_1(\boldsymbol{\theta})} + \underbrace{\left\|\sum_{d=0}^{\widetilde{d}(t)-1} a(d;\boldsymbol{\theta})\left|\frac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right|\right\|}_{\widehat{S}_2(\boldsymbol{\theta})} + \underbrace{\left\|\sum_{d=\widetilde{d}(t)}^{\infty} a(d;\boldsymbol{\theta})\frac{D_t(d)}{|V(t)|}\right\|}_{\widehat{S}_3(\boldsymbol{\theta})}.\end{aligned}$$

To prove $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\left\|\widehat{G}_t(\boldsymbol{\theta}) - G_0(\boldsymbol{\theta})\right\| \xrightarrow{P} 0$, it suffices to show that $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\widehat{S}_1(\boldsymbol{\theta}) \xrightarrow{P} 0$, $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\widehat{S}_2(\boldsymbol{\theta}) \xrightarrow{P} 0$, and $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\widehat{S}_3(\boldsymbol{\theta}) \xrightarrow{P} 0$.

Because $G_0^n(\boldsymbol{\theta})$ uniformly converges to $G_0(\boldsymbol{\theta})$ on $\overline{\boldsymbol{\Theta}}$, we have $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\widehat{S}_1(\boldsymbol{\theta}) \xrightarrow{P} 0$.

Proposition 1 implies that for any $K > 1$ there exists $N(\boldsymbol{\theta}_0)$ such that for $0 \leq d \leq d^*(t;\eta)$, we have:

$$\Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right| \geq K\tfrac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right) \leq \tfrac{N(\boldsymbol{\theta}_0)}{\ln t}.$$

Therefore, by definition of $\widetilde{d}(t)$, we have

$$\Pr\left(\exists d \leq \widetilde{d}(t) : \left|\tfrac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right| \geq K\tfrac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right) \leq \tfrac{N(\boldsymbol{\theta}_0)\widetilde{d}(t)}{\ln t} \to 0. \tag{11}$$

Thus, with probability approaching one:

$$\widehat{S}_2(\boldsymbol{\theta}) \leq \left\|\textstyle\sum_{d=0}^{\widetilde{d}(t)-1} a(d;\boldsymbol{\theta}) K\tfrac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right\| < C_1\tfrac{K\|G_0(\boldsymbol{\theta})\|}{\sqrt{\ln t}},$$

for some $C_1$. The last inequality follows from the uniform convergence of $G_0^{\eta}(\boldsymbol{\theta})$ on $\overline{\boldsymbol{\Theta}}$. Since $G_0(\boldsymbol{\theta})$ is continuous on a compact set $\overline{\boldsymbol{\Theta}}$, $\|G_0(\boldsymbol{\theta})\|$ is bounded on $\overline{\boldsymbol{\Theta}}$; so $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{S}_2(\boldsymbol{\theta}) \xrightarrow{P} 0$.

Since $\|a(d;\boldsymbol{\theta})\| < F(d+1)$, showing $\sum_{d=\widetilde{d}(t)}^{\infty} d\tfrac{D_t(d)}{|V(t)|} \xrightarrow{P} 0$ is sufficient for $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{S}_3(\boldsymbol{\theta}) \xrightarrow{P} 0$. Using the definition of $n_m(d;\eta,\kappa)$ and the property $B(x+1,y) = B(x,y)x/(x+y)$, we can represent $\sum_{d=m}^{\infty} dn_m(d;\eta,\kappa)$ as the composition of an infinite geometric series and its derivative, which simplifies to

$$\textstyle\sum_{d=m}^{\infty} dn_m(d;\eta,\kappa) = \tfrac{\kappa\eta+m}{1-\eta}.$$

Next, using the definition of $P(d;\boldsymbol{\theta})$ and $\kappa$, we obtain

$$\textstyle\sum_{d=0}^{\infty} dP(d;\boldsymbol{\theta})) = \sum_{m=0}^{P} p_m \sum_{d=m}^{\infty} dn_m(d;\eta,\kappa) = 2\left(\overline{m} + \overline{M}\right). \tag{12}$$

Since $m(t) + M(t)$ are i.i.d. with finite variance, the law of large numbers implies

$$\textstyle\sum_{d=0}^{\infty} d\tfrac{D_t(d)}{|V(t)|} \xrightarrow{P} 2(\overline{m}_0 + \overline{M}_0) = \sum_{d=0}^{\infty} dP(d;\boldsymbol{\theta}_0)), \tag{13}$$

where the equality follows from (12). Using (11) and part 1 of this proposition, we get

$$\textstyle\sum_{d=0}^{\tilde{d}(t)-1} d\tfrac{D_t(d)}{|V(t)|} = \left(\sum_{d=0}^{\tilde{d}(t)-1} dP(d;\boldsymbol{\theta}_0)\right)\left(1 + O_P\left(\tfrac{1}{\sqrt{\ln t}}\right)\right) \xrightarrow{P} \sum_{d=0}^{\infty} dP(d;\boldsymbol{\theta}_0). \tag{14}$$

Combining (13) and (14) gives

$$\textstyle\sum_{d=\widetilde{d}(t)}^{\infty} d\tfrac{D_t(d)}{|V(t)|} = \sum_{d=0}^{\infty} d\tfrac{D_t(d)}{|V(t)|} - \sum_{d=0}^{\widetilde{d}(t)-1} d\tfrac{D_t(d)}{|V(t)|} \xrightarrow{P} 0,$$

which completes the proof of $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{S}_3(\boldsymbol{\theta}) \xrightarrow{P} 0$. ■

**Proof of Proposition 3.** Denote $\underline{\eta} = \min_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \eta$ and $\overline{\kappa} = \max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \kappa$. Palumbo (1998) shows that

$$\tfrac{\Gamma(k+\lambda)}{\Gamma(k+1)} < \left(k + \tfrac{\lambda}{2}\right)^{\lambda-1} \quad \text{for } \lambda > 2 \text{ and } k \geq 0.$$

Thus,

$$|\ln P(d;\boldsymbol{\theta})| = -\ln P(d;\boldsymbol{\theta}) \leq \ln\left(\tfrac{\Gamma(d+\kappa+1/\eta+1)}{\underline{C}\Gamma(d+\kappa)}\right) < -\ln\underline{C} + \left(1+1/\underline{\eta}\right)\ln\left(d+\overline{\kappa}+1/(2\underline{\eta})\right),$$

where $\underline{C} = \min_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} p_{m_o}\frac{\Gamma(m_o+\kappa+1/\eta)}{\eta\Gamma(m_o+\kappa)} > 0$ and $m_o$ is $\min(m)$ such that $p_m > 0$. Thus, there is $C$ such that $|\ln P(d;\boldsymbol{\theta})| < C\ln(d+1)$ and Proposition 2 applies; i.e., $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\left|\widehat{L}_t(\boldsymbol{\theta}) - L_0(\boldsymbol{\theta})\right| \xrightarrow{P} 0$, where $L_0(\boldsymbol{\theta}) \equiv \sum_{d=0}^{\infty}\ln P(d;\boldsymbol{\theta})P(d;\boldsymbol{\theta}_0)$ is a continuous function.

$L_0(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$ by information inequality. Indeed, it is clear that $\sum_{d=0}^{\infty}|\ln P(d;\boldsymbol{\theta})|\,P(d;\boldsymbol{\theta}_0) = -L_0(\boldsymbol{\theta}) < \infty$ for all $\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}$. Moreover, if $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$, then there exists $d$ such that $P(d;\boldsymbol{\theta}) \neq P(d;\boldsymbol{\theta}_0)$ and thus by the strict version of Jensen's inequality:

$$L_0(\boldsymbol{\theta}_0) - L_0(\boldsymbol{\theta}) = -\sum_{d=0}^{\infty}\ln\tfrac{P(d;\boldsymbol{\theta})}{P(d;\boldsymbol{\theta}_0)}P(d;\boldsymbol{\theta}_0) < \ln\left(\sum_{d=0}^{\infty}\tfrac{P(d;\boldsymbol{\theta})}{P(d;\boldsymbol{\theta}_0)}P(d;\boldsymbol{\theta}_0)\right) = \ln\left(\sum_{d=0}^{\infty}P(d;\boldsymbol{\theta})\right) = 0. \tag{15}$$

Thus, if $\widehat{L}_t(\widehat{\boldsymbol{\theta}}) \geq \max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\widehat{L}_t(\boldsymbol{\theta}) + o_P(1)$, then all conditions of Theorem 2.1 of Newey and McFadden (1994) are satisfied and thus $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$. By definition $\widehat{L}_t(\widehat{\boldsymbol{\theta}}^{\text{PML}}) = \max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\widehat{L}_t(\boldsymbol{\theta})$; so $\widehat{\boldsymbol{\theta}}^{\text{PML}} \xrightarrow{P} \boldsymbol{\theta}_0$. To solve for an estimate $\widehat{\boldsymbol{\theta}}^{\text{plPML}}$, we substitute $(\overline{m}+\overline{M}) = \widehat{(\overline{m}+\overline{M})}$ in $\widehat{L}_t(.)$ and maximize $\widehat{L}_t(\eta,\widehat{\overline{M}},\boldsymbol{p})$ over $\eta$ and $\boldsymbol{p}$, where $\widehat{\overline{M}} = \widehat{(\overline{m}+\overline{M})} - \overline{m}$. Since $\widehat{\overline{M}}$ is continuous, $\widehat{(\overline{m}+\overline{M})} \xrightarrow{P} (\overline{m}_0+\overline{M}_0)$, and $\widehat{L}_t(\boldsymbol{\theta})$ uniformly converges to a continuous function $L_0(\boldsymbol{\theta})$, it follows that $\widehat{L}_t(\widehat{\boldsymbol{\theta}}^{\text{plPML}}) \geq \widehat{L}_t(\widehat{\eta}^{\text{PML}},\widehat{\overline{M}},\widehat{\boldsymbol{p}}^{\text{PML}}) = \widehat{L}_t(\widehat{\boldsymbol{\theta}}^{\text{PML}}) + o_P(1)$, which implies that $\widehat{\boldsymbol{\theta}}^{\text{plPML}} \xrightarrow{P} \boldsymbol{\theta}_0$. ■

**Proof of Proposition 4.**

**Part 1.** See the proof of Theorem 2.6 in Newey and McFadden (1994) and replace Lemma 2.4 with our Proposition 2 in the argument.

**Part 2.** Equation (7) holds for $= d - 2(\overline{M}+\overline{m})$ by (12) and for $\nabla_{\boldsymbol{\theta}}\ln P(d;\boldsymbol{\theta})$ by (15) and interchangeability of summation and differentiation (see Theorems 7.10 and 7.17 of Rudin, 1976). We now verify conditions of part 1 to make sure that $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$. Condition (iii), clearly, holds. To verify condition (iv), it is convenient to use the following representation

$$\ln P(d;\boldsymbol{\theta}) = \ln\Gamma(d+\kappa) - \ln\Gamma(d+\kappa+1/\eta+1) + \underbrace{\ln\left(\sum_{m=0}^{\min(P,d)} p_m\frac{\Gamma(m+\kappa+1/\eta)}{\eta\Gamma(m+\kappa)}\right)}_{R(\boldsymbol{\theta})},$$

where $R(\boldsymbol{\theta})$ collects all terms independent of $d$, for any $d \geq P$. Let $r_x(\boldsymbol{\theta})$ denote a partial derivative of $R(\boldsymbol{\theta})$ with respect to $x$ and $r_{xy}(\boldsymbol{\theta})$ denote a second-order partial derivative with respect to $x$ and $y$.

The score function $s(d;\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln P(d;\boldsymbol{\theta})$ can be written as

$$\begin{aligned}
s_{\eta}(d;\boldsymbol{\theta}) &= -\frac{\overline{m}+2\overline{M}}{\eta^2}\psi\left(d+\kappa\right) + \frac{\overline{m}+2\overline{M}+1}{\eta^2}\psi\left(d+\kappa+1+1/\eta\right) + r_{\eta}(\boldsymbol{\theta}) \\
s_{\overline{M}}(d;\boldsymbol{\theta}) &= 2\left(\frac{1}{\eta}-1\right)\Big(\psi\left(d+\kappa\right) - \psi\left(d+\kappa+1+1/\eta\right)\Big) + r_{\overline{M}}(\boldsymbol{\theta}), \\
s_{p_m}(d;\boldsymbol{\theta}) &= m\left(\frac{1}{\eta}-2\right)\Big(\psi\left(d+\kappa\right) - \psi\left(d+\kappa+1+1/\eta\right)\Big) + r_{p_m}(\boldsymbol{\theta}),
\end{aligned}$$

where $\psi(.)$ is a polygamma function. Polygamma function of order $n$ is defined as $\psi^{(n)}(z) = \left(\frac{d}{dz}\right)^{n+1} \ln\Gamma(z)$ with $\psi(z) = \psi^{(0)}(z)$. Qi et al. (2005) show that for $x > 0$:

$$\frac{1}{2x} - \frac{1}{12x^2} < \psi\left(x+1\right) - \ln x < \frac{1}{2x},$$

which implies that there is $F$ such that

$$\|g(d;\boldsymbol{\theta})\| < F\left(d+1\right), \tag{16}$$

for all $d \geq 0$ and all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$; so condition (iv) of part 1 holds, and, therefore, $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$. ■

**Proof of Proposition 5.** To prove Proposition 5, we notice that all conditions, except for condition (iv), of Theorem 3.2 in Newey and McFadden (1994) are satisfied by assumption. Thus, we only need to check condition (iv) that $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \|\sum_{d=0}^{\infty} \nabla_{\boldsymbol{\theta}} g(d;\boldsymbol{\theta}_0) D_t(d)/|V(t)| - G(\boldsymbol{\theta})\| \xrightarrow{P} 0$ for some compact set $\overline{\boldsymbol{\Theta}}$ such that $\boldsymbol{\theta}_0 \in \overline{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}$.

Denote $G(d;\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} g(d;\boldsymbol{\theta})$ and recall that $\boldsymbol{\theta} = (\eta, \overline{M}, \boldsymbol{p})$. Then the last row of $G(d;\boldsymbol{\theta})$ is given by

$$\nabla_{\boldsymbol{\theta}}\left(d - 2(\overline{M}+\overline{m})\right) = \begin{pmatrix} 0 & -2 & 0 & \dots & -2m & \dots \end{pmatrix}.$$

Next, we calculate $h(d;\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ln P(d;\boldsymbol{\theta})$

$$
\begin{aligned}
h_{\eta\eta}(d;\boldsymbol{\theta}) &= \frac{2(\overline{m}+2\overline{M})}{\eta^3}\psi\left(d+\kappa\right) - \frac{2(\overline{m}+2\overline{M}+1)}{\eta^3}\psi\left(d+\kappa+1+1/\eta\right) \\
&\quad + \frac{(\overline{m}+2\overline{M})^2}{\eta^4}\psi^{(1)}\left(d+\kappa\right) - \frac{(\overline{m}+2\overline{M}+1)^2}{\eta^4}\psi^{(1)}\left(d+\kappa+1+1/\eta\right) + r_{\eta\eta}(\boldsymbol{\theta}) \\
h_{\eta\overline{M}}(d;\boldsymbol{\theta}) &= -\frac{2}{\eta^2}\big(\psi\left(d+\kappa\right) - \psi\left(d+\kappa+1+1/\eta\right)\big) + 2\left(\frac{1}{\eta}-1\right)\cdot \\
&\quad \left(-\frac{\overline{m}+2\overline{M}}{\eta^2}\psi^{(1)}\left(d+\kappa\right) + \frac{\overline{m}+2\overline{M}+1}{\eta^2}\psi^{(1)}\left(d+\kappa+1+1/\eta\right)\right) + r_{\eta\overline{M}}(\boldsymbol{\theta}) \\
h_{\eta p_m}(d;\boldsymbol{\theta}) &= -\frac{m}{\eta^2}\big(\psi\left(d+\kappa\right) - \psi\left(d+\kappa+1+1/\eta\right)\big) + m\left(\frac{1}{\eta}-2\right)\cdot \\
&\quad \left(-\frac{\overline{m}+2\overline{M}}{\eta^2}\psi^{(1)}\left(d+\kappa\right) + \frac{\overline{m}+2\overline{M}+1}{\eta^2}\psi^{(1)}\left(d+\kappa+1+1/\eta\right)\right) + r_{\eta p_m}(\boldsymbol{\theta}) \\
h_{\overline{M}\overline{M}}(d;\boldsymbol{\theta}) &= 4\left(\frac{1}{\eta}-1\right)^2\left(\psi^{(1)}\left(d+\kappa\right) - \psi^{(1)}\left(d+\kappa+1+1/\eta\right)\right) + r_{\overline{M}\overline{M}}(\boldsymbol{\theta}) \\
h_{\overline{M}p_m}(d;\boldsymbol{\theta}) &= 2m\left(\frac{1}{\eta}-1\right)\left(\frac{1}{\eta}-2\right)\left(\psi^{(1)}\left(d+\kappa\right) - \psi^{(1)}\left(d+\kappa+1+1/\eta\right)\right) + r_{\overline{M}p_m}(\boldsymbol{\theta}) \\
h_{p_mp_m}(d;\boldsymbol{\theta}) &= m^2\left(\frac{1}{\eta}-2\right)^2\left(\psi^{(1)}\left(d+\kappa\right) - \psi^{(1)}\left(d+\kappa+1+1/\eta\right)\right) + r_{p_mp_m}(\boldsymbol{\theta})
\end{aligned}
$$

Qi et al. (2005) shows that for $x > 0$

$$
\begin{aligned}
\frac{1}{2x} - \frac{1}{12x^2} &< \psi\left(x+1\right) - \ln x < \frac{1}{2x}, \\
\frac{1}{2x^2} - \frac{1}{6x^3} &< \frac{1}{x} - \psi^{(1)}\left(x+1\right) < \frac{1}{2x^2} - \frac{1}{6x^3} + \frac{1}{30x^5},
\end{aligned}
$$

which implies that there is $F$ such that

$$
\|G(d;\boldsymbol{\theta})\| < F(d+1), \tag{17}
$$

for all $d \geq 0$ and all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$.

In addition $G(d;\boldsymbol{\theta})$ is continuous; so Proposition 2 applies. Therefore, condition (iv) of Theorem 3.2 in Newey and McFadden (1994) holds and

$$
G(\boldsymbol{\theta}) = \sum_{d=0}^{\infty}\begin{pmatrix}
h_{\eta\eta}(d;\boldsymbol{\theta}) & h_{\eta\overline{M}}(d;\boldsymbol{\theta} & h_{\eta p_0}(d;\boldsymbol{\theta}) & ... \\
h_{\eta\overline{M}}(d;\boldsymbol{\theta}) & h_{\overline{M}\overline{M}}(d;\boldsymbol{\theta}) & h_{\overline{M}p_0}(d;\boldsymbol{\theta}) & ... \\
h_{\eta p_0}(d;\boldsymbol{\theta}) & h_{\overline{M}p_0}(d;\boldsymbol{\theta}) & h_{p_0p_0}(d;\boldsymbol{\theta}) & ... \\
... & ... & ... & ... \\
0 & -2 & 0 & ...
\end{pmatrix} P(d;\boldsymbol{\theta}_0).
$$

Notice that we interchange the order of summation and differentiation using Theorems 7.10 and 7.17 of Rudin (1976). ■

# Appendix C: NLSS and Local Tail Exponent Estimators

We start with approximating the asymptotic degree distribution of the CF model by a mean-field method. First, this method provides the intuition as to why the asymptotic degree distribution has a power-law tail. Second, this method also suggests which parameters of the random graph process are crucial for the asymptotic degree distribution. Third, the NLLS method discussed further is based on the mean-field approximation of the asymptotic degree distribution.

## Mean-Field Approximation of the degree distribution

Using the mean-field methods of Barabasi and Albert (1999) we approximate the CF network formation process by a continuous time process such that

$$\begin{aligned}\frac{d\mathbb{E}\left(d(v,t)\right)}{dt} &= \frac{\left(\overline{m}A_1+\overline{M}(B_1+C_1)\right)\mathbb{E}\left(d(v,t)\right)}{2\mathbb{E}|E(t-1)|}+\frac{\overline{m}A_2+\overline{M}(B_2+C_2)}{\mathbb{E}|V(t-1)|}\\ &= \frac{\left(\overline{m}A_1+\overline{M}(B_1+C_1)\right)\mathbb{E}\left(d(v,t)\right)}{2(\overline{m}+\overline{M})(t-2)+2|E(1)|}+\frac{\overline{m}A_2+\overline{M}(B_2+C_2)}{t-2+|V(1)|},\end{aligned}$$

where $\overline{m}A_1+\overline{M}(B_1+C_1)$ and $\overline{m}A_2+\overline{M}(B_2+C_2))$ is the expected number of edge endpoints added at time $t$ by preferential attachment and uniformly at random, respectively.

Assuming $t \gg \max\{|V(1)|,|E(1)|\}$ the differential equation becomes

$$\frac{d\mathbb{E}\left(d(v,t)\right)}{dt}=\frac{\eta\mathbb{E}\left(d(v,t)\right)}{t}+\frac{\eta\kappa}{t},$$

where $(\eta\kappa)$ is the number of edges added uniformly at random. The solution to this differential equation is:

$$\phi_t^m(v)\equiv\mathbb{E}\left(d(v,t)\right)=\left(m(v)+\frac{\nu}{\eta}\right)\left(\frac{t}{v}\right)^{\eta}-\kappa,$$

where $m(v)$ is the degree of a newly added vertex at time $v$. The function $\phi_t^m(v)$ is decreasing in $v$, which means that given an initial degree, "older" vertices have a larger expected degree than "younger" vertices. Thus the distribution of expected degrees of vertices with initial degree $m$ can be approximated by (for $d \geq m$):

$$F_t^m(d)=\frac{p_m|\{i:\phi_t^m(i)\leq d\}|}{p_m t}=1-\frac{\phi_t^{m(-1)}(d)}{t}=1-(m+\kappa)^{\frac{1}{\eta}}(d+\kappa)^{-\frac{1}{\eta}}. \tag{18}$$

Thus the distribution of expected degrees of graph $G(t)$ can be approximated by:

$$F^{\mathrm{MF}}(d)=\sum_{m=0}^{\min\{P,d\}}p_m F_t^m(d), \tag{19}$$

where $F_t^m(d)$ are given by (18).

Claim 1 part 2 shows that for sufficiently large $d$ asymptotic degree distribution can be approximated by power law distribution, which we denote $P^{\text{tail}}(d;\eta)$,

$$P^{\text{tail}}(d;\eta) \equiv Cd^{-1-1/\eta}. \tag{20}$$

This and related distributions will appear in the local tail exponent methods. We compare the approximations with asymptotic and empirical distribution of the the degree for the benchmark specification $t = 1000$, $\eta = 0.5$, $p_0 = 1$ ($m(t) = 0$), and $q_1 = q_2 = 1/2$ ($\overline{M} = 1.5$). Figure 1 shows the simulation-based empirical cumulative distribution function (ECDF) of the degree distribution, $\widehat{F}_t(d) \equiv \sum_{v=1}^{t} \mathbf{1}(d(v,t) \leq d)/|V(t)|$, and the cumulative distribution functions (CDF) based on the asymptotic degree distribution $F(d) = \sum_{\widetilde{d}=0}^{d} P(\widetilde{d})$, the mean-field approximation, $F^{\text{MF}}(d)$, and the power law approximation, $F^{\text{tail}}(d) = 1 - \sum_{\widetilde{d}>d} P^{\text{tail}}(d)$, on linear and log-log scales. As we can see, the CDF of the asymptotic distribution, $F(d)$, is close to the ECDF, $\widehat{F}_t(d)$, but the CDF based on the mean-field approximation, $F^{\text{MF}}(d)$, and the CDF based on the power law approximation, $F^{\text{tail}}(d)$, deviate from the ECDF for small $d$. This may explain subpar performance of the estimators based on the latter two approximation.

**<<Place Figure A.1 about here>>**

## NLLS Estimator

We now turn to the NLLS method commonly used to estimate scale-free network formation models (see Pennock et al., 2002; Jackson and Rogers, 2007; Jackson, 2008). Pennock et al. (2002) assume $m(t) = 0$ and $M(t) = M$ for some $M$, whereas Jackson (2008) and Jackson and Rogers (2007) assume $m(t) = m$ and $M(t) = 0$ for some $m$. More generally, the NLLS method can be used only when $m(t) = m$ for some $m$ but $M(t)$ can have an arbitrary distribution, $\boldsymbol{q}$. Since most real networks have vertices with zero degree, we use the NLLS method for the case of $m(t) = 0$.

Under this assumption (19) can be rewritten as

$$\ln(1 - F^{\text{MD}}(d)) = 1/\eta \left( \ln\left(2\overline{M}(1/\eta - 1)\right) - \ln\left(d + 2\overline{M}(1/\eta - 1)\right)\right).$$

Moreover, $\overline{M}$ cab consistently estimated by $\sum_{v=1}^{t} d(v,t)/2|V(t)|$ as shown in (**??**); denote the estimate by $\widehat{\overline{M}}$. $F(d)$ can be estimated by ECDF $\widehat{F}_t(d)$.[18] Thus, the only parameter to be

[18]Since $\widehat{F}_t(d_{\max}) = 1$, $\ln(1 - \widehat{F}_t(d))$ is not defined. As a remedy, we drop observations with $d = d_{\max}$ as in Jackson (2008).

estimated is $\eta$, which is done by numerically minimizing the quadratic loss:

$$\widehat{\eta^{\text{NLLS}}} = \text{argmin} \sum_d \left(\ln(1-\widehat{F}_t(d)) - 1/\eta \left(\ln(2\widehat{\overline{M}}(1/\eta-1)) - \ln(d+2\widehat{\overline{M}}(1/\eta-1))\right)\right)^2,$$

where the sum is taken over all observed distinct degrees as in Barabasi and Albert (1999).[19]

## Local Tail Exponent Estimators

There is a well developed literature on local tail exponent estimators starting from Pickands (1975), Hill (1975), Smith (1987), see Beirlant et al. (2006) for detailed treatment and references. These estimators rely on behaviour in the tail of the distribution and, therefore, are robust against misspecification in the rest of the distribution. However, an appropriate choice of the number of observations in the tail, or a threshold, $d_{\text{tr}}$ after which the tail approximation holds is crucial for these estimators. For the moment we will assume that $d_{\text{tr}}$ is known and will discuss various methods for finding $d_{\text{tr}}$ after introducing the estimators.

A simple and popular way to estimate the tail exponent is for to run a log-log rank-degree regression, $\log(r) = c - \frac{1}{\eta}\log(d)$, for $d \geq d_{\text{tr}}$, where $r$ is rank of $d$. Ordinal ranking assigns each observation a distinct ordinal number according to their degree $d$ in ascending order. We use fractional ranking to determine $r$, which, in case of the ties, when different observations have equal $d$-s, assigns these observations their average ordinal rank. The log-log rank-degree regression is based on the fact that approximating the degree by continuous distribution from (20) we can find the probability that a degree is higher than $d$, $1 - F^{\text{tail}}(d) \propto d^{-1/\eta}$, which, in turn, can be estimated by rank $r$ up to a normalizing constant. Gabaix and Ibragimov (2011) proposed a simple, yet important bias-reducing adjustment, to use $r - 1/2$ instead of $r$, which we implement for this estimator and refer to it as GI estimator.

Hill (1975) estimator is another well-known local tail exponent estimator. It can be derived as an MLE estimator based on the assumption that the tail of the distribution follows continuous Pareto distribution with pdf $f(d) = \frac{d}{\eta}\left(\frac{d}{d_{\text{tr}}}\right)^{-1/\eta}$ for $d \geq d_{\text{tr}}$ and these tail observations are independent. For discrete distribution Clauset et al. (2009) propose an adjustment for Hill estimator, which we adopt here. Let $n_{\text{tail}} \equiv |\{v : d(v,t) \geq d_{\text{tr}}\}|$ be the number of vertices that have degree at least $d_{\text{tr}}$, then the adjusted Hill estimator is given by

$$\widehat{\eta^{\text{Hill}}} = \frac{1}{n_{\text{tail}}} \sum_{v:d(v,t)\geq d_{\text{tr}}} \ln \frac{d(v,t)}{d_{\text{tr}} - 1/2}. \tag{21}$$

[19] An alternative would be to use all degrees in the range $[d_{\min}, d_{\max}]$ where $d_{\min} = \min_v d(v,t)$ and $d_{\max} = \max_v d(v,t)$ as in Jackson (2008). Asymptotically these methods are equivalent, but our simulations show that taking actually observed degrees works better in finite samples.

There are various generalisations of the Hill estimator proposed in the literature; e.g. Dekkers et al. (1989) uses higher moments, which we also implement for comparison. We also consider Pickands (1975) estimator which is based on sample quantiles in the tails.

The discrete counterpart of the Pareto distribution is zeta distribution, which can be used to derive local tail exponent estimator using (20) without the continuous approximation (see, e.g., Clauset et al., 2009). Conditional on having a degree not lower than $d_{\text{tr}}$, a vertex chosen uniformly at random has degree $d$ with probability

$$P(d) = \frac{d^{-1-1/\eta}}{\zeta(1+1/\eta, d_{\text{tr}})}, \tag{22}$$

where $\zeta(1+1/\eta, d_{\text{tr}}) = \sum_{i=0}^{\infty}(i+d_{\text{tr}})^{-(1+1/\eta)}$ is the Hurwitz zeta function. The discrete local tail exponent (DLTE) estimator is found by maximizing the pseudo log-likelihood conditional on having degree not lower than $d_{\text{tr}}$ given by

$$\widehat{L}_t(\eta|d_{\text{tr}}) = -\sum_{v:d(v,t)\geq d_{\text{tr}}} (1+1/\eta)\ln d(v,t) + \ln\zeta(1+1/\eta, d_{\text{tr}})\,.$$

Until now, we treated $d_{\text{tr}}$ as given. Next, we discuss several methods for selecting $d_{\text{tr}}$ as this is a crucial step for any local tail exponent estimator. For discrete distribution Handcock and Jones (2004) propose to select $d_{\text{tr}}$ using AIC or BIC criteria on the basis of the complete log-likelihood. For each $d < d_{\text{tr}}$ they assume separate probability $\Pr(d=k) = \pi_k$ for $k < d_{\text{tr}}$, where $\pi_k$ are treated as parameters to be estimated. Then the log-likelihood can be written as

$$\widetilde{L}_t(\pi,\eta) = \sum_{d=0}^{d_{\text{tr}}-1} D_t(d)\ln\pi_d + n_{\text{tail}}\ln\left(1-\sum_{d=0}^{d_{\text{tr}}-1}\pi_d\right) + \widehat{L}_t(\eta|d_{\text{tr}}). \tag{23}$$

Handcock and Jones (2004) show that empirical frequencies $\widehat{\pi}_d = \frac{D_t(d)}{|V(t)|}$ and the DLTE estimator, $\widehat{\eta}$, maximize (23). Then $d_{\text{tr}}$ is selected by minimizing AIC or BIC:

$$\begin{aligned} AIC &= -2\widetilde{L}_t(\widehat{\pi},\widehat{\eta}) + 2\,(d_{\text{tr}}+1)\,, \\ BIC &= -2\widetilde{L}_t(\widehat{\pi},\widehat{\eta}) + (d_{\text{tr}}+1)\ln|V(t)|. \end{aligned}$$

For comparison we use (23) and plug-in $\widehat{\eta}$ for the other local tail exponent estimators described above to use AIC and BIC criteria.

Clauset et al. (2009) suggests an alternative approach which is based on choosing $d_{\text{tr}}$ in such a way that the distance between the CDF of the corresponding power-law distribution with estimated $\widehat{\eta}$ and the ECDF is minimised for all $d \geq d_{\text{tr}}$. The authors suggest using Kolmogorov-Smirnov (KS), or supremum, distance, and demonstrate superior performance

of this measure in comparison to the above AIC and BIC criteria. They argue that there are many other possible distance measures. Hence, in addition to the KS distance, we implement Cramer-von-Misses (CM) and Anderson-Darling (AD) distance measures. Our simulations showed that the AD distance-based criterion resulted in the lowest MSE of parameter $\eta$ for most considered local tail exponent estimators and parametrisation. Out of all considered local tail exponent estimators, GI, adjusted Hill and DLTE performed best. For brevity in Section 4, we report the results for these three local tail exponent estimators with $d \geq d_{\text{tr}}$ selection based on the AD distance. The complete set of results is available on request.

# References


Albert, Reka and Albert-Laszlo Barabasi (2002) "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics*, **74** (1), pp. 47–97.

Altonji, Joseph G and Lewis M Segal (1996) "Small-sample bias in GMM estimation of covariance structures," *Journal of Business & Economic Statistics*, **14** (3), pp. 353–366.

Atalay, Enghin (2013) "Sources of Variation in Social Networks," *Games and Economic Behavior*, **79**, pp. 106–131.

Atalay, Enghin, Ali Hortacsu, James Roberts, and Chad Syverson (2011) "Network Structure of Production," *Proceedings of the National Academy of Sciences*, **108** (13), pp. 5199–5202.

Babus, Ana and Aljaz Ule (2008) "Limited Connections: Economic Foundations for a Preferential Attachment Model," mimeo.

Bala, Venkatesh and Sanjeev Goyal (2000) "A Noncooperative Model of Network Formation," *Econometrica*, **68** (5), pp. 1181–1229.

Barabasi, Albert-Laszlo and Reka Albert (1999) "Emergence of Scaling in Random Networks," *Science*, **286** (5439), pp. 509–512.

Beirlant, Jan, Yuri Goegebeur, Johan Segers, and Jozef Teugels (2006) *Statistics of extremes: theory and applications*: John Wiley & Sons.

Bollobas, Bela, Oliver Riordan, Joel Spencer, and Gabor Tusnady (2001) "The Degree Sequence of a Scale-Free Random Graph Process," *Random Structures and Algorithms*, **18** (3), pp. 279–290.

Bollobas, Bela and Oliver Riordan (2003) "Mathematical Results on Scale-Free Random Graphs," in Stefan Bornholdt and Heinz Georg Schuster eds. *Handbook of Graphs and Networks*, Berlin: Wiley, pp. 1–34.

Buckley, Pierce G. and Deryk Osthus (2004) "Popularity Based Random Graph Models Leading to a Scale-Free Degree Sequence," *Discrete Mathematics*, **282** (1-3), pp. 53–68.

Burnham, Kenneth P. and David R. Anderson (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, New York: Springer Verlag.

Christakis, Nicholas A., James H. Fowler, Guido W. Imbens, and Karthik Kalyanaraman (2010) "An Empirical Model for Strategic Network Formation," Working Paper 16039, National Bureau of Economic Research.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman (2009) "Power-law distributions in empirical data," *SIAM review*, **51** (4), pp. 661–703.

Cooper, Colin (2006) "Distribution of Vertex Degree in Web-Graphs," *Combinatorics, Probability and Computing*, **15** (5), pp. 637–661.

Cooper, Colin and Alan Frieze (2003) "A General Model of Web Graphs," *Random Structures and Algorithms*, **22** (3), pp. 311–335.

Dekkers, Arnold LM, John HJ Einmahl, and Laurens De Haan (1989) "A moment estimator for the index of an extreme-value distribution," *The Annals of Statistics*, **17** (4), pp. 1833–1855.

Dorogovtsev, Sergey N., Jose F.F. Mendes, and Alexandr N. Samukhin (2000) "Structure of Growing Networks with Preferential Linking," *Physical Review Letters*, **85** (21), pp. 4633–4636.

Dorogovtsev, Sergey N. and Jose F.F. Mendes (2002) "Evolution of Networks," *Advances in Physics*, **51** (4), pp. 1079–1187.

Efron, Bradley and Robert J Tibshirani (1994) *An introduction to the bootstrap*: CRC press.

Gabaix, Xavier and Rustam Ibragimov (2011) "Rank – 1/2: A Simple Way to improve the OLS Estimation of Tail Exponents," *Journal of Business & Economic Statistics*, **29** (1), pp. 24–39.

Goldstein, Michel L., Steven A. Morris, and Gary G. Yen (2004) "Problems with Fitting to the Power-Law Distribution," *European Physical Journal B*, **41** (2), pp. 255–258.

Goyal, Sanjeev, Marco J. Van der Leij, and Jose Luis Moraga-Gonzalez (2006) "Economics: An Emerging Small World," *Journal of Political Economy*, **114** (2), pp. 403–412.

Handcock, Mark S. and James Holland Jones (2004) "Likelihood-Based Inference for Stochastic Models of Sexual Network Formation," *Theoretical Population Biology*, **65** (4), pp. 413–422.

Hill, Bruce M. (1975) "A Simple General Approach to Inference about the Tail of a Distribution," *Annals of Statistics*, **3** (5), pp. 1163–1174.

Jackson, Matthew O. (2008) *Social and Economic Networks*, Princeton: Princeton University Press.

Jackson, Matthew O. and Brian W. Rogers (2007) "Meeting Strangers and Friends of Friends: How Random Are Social Networks?" *American Economic Review*, **97** (3), pp. 890–915.

Jackson, Matthew O and Asher Wolinsky (1996) "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, **71** (1), pp. 44–74.

Jenish, Nazgul and Ingmar R Prucha (2009) "Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random fields," *Journal of Econometrics*, **1** (150), pp. 86–98.

——— (2012) "On Spatial Processes and Asymptotic Inference Under Near-epoch Dependence," *Journal of Econometrics*, **170** (1), pp. 178–190.

Kleinberg, Jon M, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S Tomkins (1999) "The web as a graph: measurements, models, and methods," in *Computing and combinatorics*: Springer, pp. 1–17.

König, Michael (2015) "The Formation of Networks with Local Spillovers and Limited Observability," *Theoretical Economics*. Forthcoming.

Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal (2000) "Stochastic Models for the Web Graph," in *41st Annual Symposium on Foundations of Computer Science*, pp. 57–65.

Mele, Angelo (2013) "A Structural Model of Segregation in Social Networks," mimeo, John Hopkins University.

Newey, Whitney K. and Daniel McFadden (1994) "Large Sample Estimation and Hypothesis Testing," in Robert F. Engle and Daniel L. McFadden eds. *Handbook of Econometrics*, **4**, Amsterdam: Elsevier, pp. 2111–2245.

Palumbo, Biagio (1998) "A Generalization of Some Inequalities for the Gamma Function," *Journal of Computational and Applied Mathematics*, **88** (2), pp. 255–268.

Pennock, David M., Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles (2002) "Winners Don't Take All: Characterizing the Competition for Links on the Web," *PNAS*, **99** (8), pp. 5207–5211.

Pickands, III, James (1975) "Statistical Inference Using Extreme Order Statistics," *the Annals of Statistics*, **3** (1), pp. 119–131.

Qi, Feng, Run-Qing Cui, Chao-Ping Chen, and Bai-Ni Guo (2005) "Some Completely Monotonic Functions Involving Polygamma Functions and an Application," *Journal of Mathematical Analysis and Applications*, **310** (1), pp. 303–308.

Rothenberg, Thomas J (1971) "Identification in Parametric Models," *Econometrica*, pp. 577–591.

Rudin, Walter (1976) *Principles of Mathematical Analysis*, New York: McGraw-Hill.

Simon, Herbert A. (1955) "On a Class of Skew Distribution Functions," *Biometrika*, **42** (3-4), pp. 425–440.

Smith, Richard L. (1987) "Estimating Tails of Probability Distributions," *Annals of Statistics*, **15** (3), pp. 1174–1207.

van der Vaart, A.W. (2000) *Asymptotic Statistics*, New York: Cambridge University Press.

Table 1: Comparison of various estimations for benchmark parameters

| | PMLE | plPMLE | GMM | $\text{GMM}_{-\overline{M}}$ | NLLS | DLTE | Hill | GI |
|---|---|---|---|---|---|---|---|---|
| $mean(\eta)$ | 0.5009 | 0.5007 | 0.5007 | 0.5007 | 0.5441 | 0.6267 | 0.5287 | 0.4854 |
| $std(\eta)$ | 0.0250 | 0.0169 | 0.0167 | 0.0169 | 0.0472 | 0.0600 | 0.0522 | 0.0213 |
| $bias(\eta)$ | 0.0009 | 0.0007 | 0.0007 | 0.0007 | 0.0441 | 0.1267 | 0.0287 | -0.0146 |
| $MSE(\eta)$ | 0.0006 | 0.0003 | 0.0003 | 0.0003 | 0.0042 | 0.0196 | 0.0035 | 0.0007 |
| $AD\ p\text{-}val(\eta)$ | 0.1584 | 0.1656 | 0.2438 | 0.1656 | <0.0005 | <0.0005 | <0.0005 | <0.0005 |
| $median(d_{\text{tr}})$ | | | | | | 12 | 11 | 12 |
| $mean(\overline{M})$ | 1.5019 | 1.5001 | 1.5001 | 1.5001 | | | | |
| $std(\overline{M})$ | 0.0418 | 0.0157 | 0.0157 | 0.0157 | | | | |
| $bias(\overline{M})$ | 0.0013 | 0.0001 | 0.0001 | 0.0001 | | | | |
| $MSE(\overline{M})$ | 0.0018 | 0.0002 | 0.0002 | 0.0002 | | | | |
| $AD\ p\text{-}val(\overline{M})$ | <0.0005 | <0.0005 | 0.0035 | <0.0005 | | | | |

Table 2: Bias of considered estimations for increasing $t$ and different $\eta$.

| $t$ | PMLE | plPMLE | GMM | $\text{GMM}_{-\overline{M}}$ | NLLS | DLTE | Hill | GI |
|---|---|---|---|---|---|---|---|---|
| | | | | $\eta = 0.5$ | | | | |
| 1000 | 0.0009 | 0.0007 | 0.0007 | 0.0007 | 0.0441 | 0.1267 | 0.0287 | -0.0146 |
| 10000 | -0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0407 | 0.0615 | 0.0096 | -0.0344 |
| 100000 | -0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0333 | 0.0306 | 0.0038 | -0.0464 |
| | | | | $\eta = 0.2$ | | | | |
| 1000 | 0.0013 | 0.0012 | 0.0012 | 0.0012 | 0.0463 | 0.1512 | 0.0914 | 0.1271 |
| 10000 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0379 | 0.0692 | 0.0520 | 0.0981 |
| 100000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0306 | 0.0266 | 0.0319 | 0.0778 |
| | | | | $\eta = 0.8$ | | | | |
| 1000 | 0.0023 | 0.0002 | 0.0002 | 0.0002 | 0.0111 | 0.1120 | -0.0209 | -0.1892 |
| 10000 | -0.0003 | 0.0000 | 0.0000 | 0.0000 | -0.0099 | 0.0452 | -0.0239 | -0.2180 |
| 100000 | -0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0097 | 0.0190 | -0.0141 | -0.2340 |

Table 3: GMM behaviour under overspecification and misspecification in comparison with other estimators. True $P = 1 : \ p_0 = p_1 = 0.5$.

| | GMM$^1$ | GMM$^2$ | GMM$^0$ | NLLS | DLTE | Hill | GI |
|---|---|---|---|---|---|---|---|
| $mean(\eta)$ | 0.5006 | 0.5125 | 0.3424 | 0.4698 | 0.6147 | 0.5236 | 0.4833 |
| $std(\eta)$ | 0.0207 | 0.0224 | 0.0162 | 0.0488 | 0.0526 | 0.0464 | 0.0179 |
| $bias(\eta)$ | 0.0006 | 0.0125 | -0.1576 | -0.0302 | 0.1147 | 0.0236 | -0.0167 |
| $MSE(\eta)$ | 0.0004 | 0.0007 | 0.0251 | 0.0033 | 0.0159 | 0.0027 | 0.0006 |
| $AD\ p\text{-}val(\eta)$ | 0.0016 | 0.1709 | 0.7329 | 0.0017 | <0.0005 | <0.0005 | <0.0005 |
| $median(d_{\mathrm{tr}})$ | | | | | 13 | 12 | 13 |
| $mean(\overline{M})$ | 1.5001 | 1.4559 | 1.9993 | | | | |
| $std(\overline{M})$ | 0.0417 | 0.0704 | 0.0223 | | | | |
| $bias(\overline{M})$ | 0.0001 | -0.0441 | 0.4993 | | | | |
| $MSE(\overline{M})$ | 0.0017 | 0.0069 | 0.2498 | | | | |
| $AD\ p\text{-}val(\overline{M})$ | <0.0005 | <0.0005 | 0.1254 | | | | |
| $mean(p_0)$ | 0.5002 | 0.4826 | | | | | |
| $std(p_0)$ | 0.0419 | 0.0449 | | | | | |
| $bias(p_0)$ | 0.0002 | -0.0174 | | | | | |
| $MSE(p_0)$ | 0.0018 | 0.0023 | | | | | |
| $AD\ p\text{-}val(p_0)$ | <0.0005 | 0.2713 | <0.0005 | | | | |
| *average BIC* | 4715 | 4721 | 4781 | | | | |

Table 4: Different distributions of $M(t)$ and values of $B_1$ and $C_1$

| | $q_0 = q_3 = 1/2$ | | | | $B_1 = 0, C_1 = 1$ | | | |
|---|---|---|---|---|---|---|---|---|
| | PMLE | plPMLE | GMM | GMM$_{-\overline{M}}$ | PMLE | plPMLE | GMM | GMM$_{-\overline{M}}$ |
| $mean(\eta)$ | 0.4990 | 0.5005 | 0.5006 | 0.5005 | 0.5010 | 0.5006 | 0.5007 | 0.5006 |
| $std(\eta)$ | 0.0287 | 0.0200 | 0.0198 | 0.0200 | 0.0210 | 0.0137 | 0.0135 | 0.0137 |
| $bias(\eta)$ | -0.0010 | 0.0005 | 0.0006 | 0.0005 | 0.0010 | 0.0006 | 0.0007 | 0.0006 |
| $MSE(\eta)$ | 0.0008 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0002 | 0.0002 | 0.0002 |
| $AD\ p\text{-}val(\eta)$ | 0.3070 | 0.0029 | 0.0059 | 0.0029 | 0.0082 | 0.1224 | 0.1700 | 0.1224 |
| $mean(\overline{M})$ | 1.4962 | 1.5001 | 1.5000 | 1.5001 | 1.5020 | 1.5001 | 1.5001 | 1.5001 |
| $std(\overline{M})$ | 0.0638 | 0.0469 | 0.0469 | 0.0469 | 0.0368 | 0.0157 | 0.0157 | 0.0157 |
| $bias(\overline{M})$ | -0.0038 | 0.0001 | 0.0000 | 0.0001 | 0.0020 | 0.0001 | 0.0001 | 0.0001 |
| $MSE(\overline{M})$ | 0.0041 | 0.0022 | 0.0022 | 0.0022 | 0.0014 | 0.0002 | 0.0002 | 0.0002 |
| $AD\ p\text{-}val(\overline{M})$ | 0.0065 | 0.0018 | 0.0066 | 0.0018 | <0.0005 | <0.0005 | <0.0005 | <0.0005 |

Table 5: Parameter estimates and their standard error for the co-authorship network

| | PMLE | plPMLE | GMM | GMM$_{-\overline{M}}$ | NLLS | DLTE | Hill | GI |
|---|---|---|---|---|---|---|---|---|
| $\eta$ | 0.3994 | 0.4024 | 0.3734 | 0.3732 | 0.1912 | 0.1753 | 0.0959 | 0.1228 |
| $se(\eta)$ | 0.0039 | 0.0049 | 0.0033 | 0.0033 | 0.0087 | 0.0352 | 0.0267 | 0.0228 |
| $d_{\text{tr}}$ | | | | | | 30 | 29 | 27 |
| $\overline{M}$ | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | | |
| $se(\overline{M})$ | 0.0054 | 0.0093 | 0.0054 | 0.0054 | | | | |
| $p_0$ | 0.3551 | 0.3401 | 0.3675 | 0.3676 | | | | |
| $se(p_0)$ | 0.0037 | 0.0054 | 0.0035 | 0.0035 | | | | |
| $p_1$ | 0.4333 | 0.4351 | 0.4253 | 0.4253 | | | | |
| $se(p_1)$ | 0.0030 | 0.0030 | 0.0029 | 0.0029 | | | | |
| $p_2$ | 0.2117 | 0.2248 | 0.2072 | 0.2071 | | | | |
| $se(p_2)$ | 0.0032 | 0.0046 | 0.0031 | 0.0031 | | | | |

Figure 1: Sample standard deviation of $\eta$ estimates as a function of $t$.

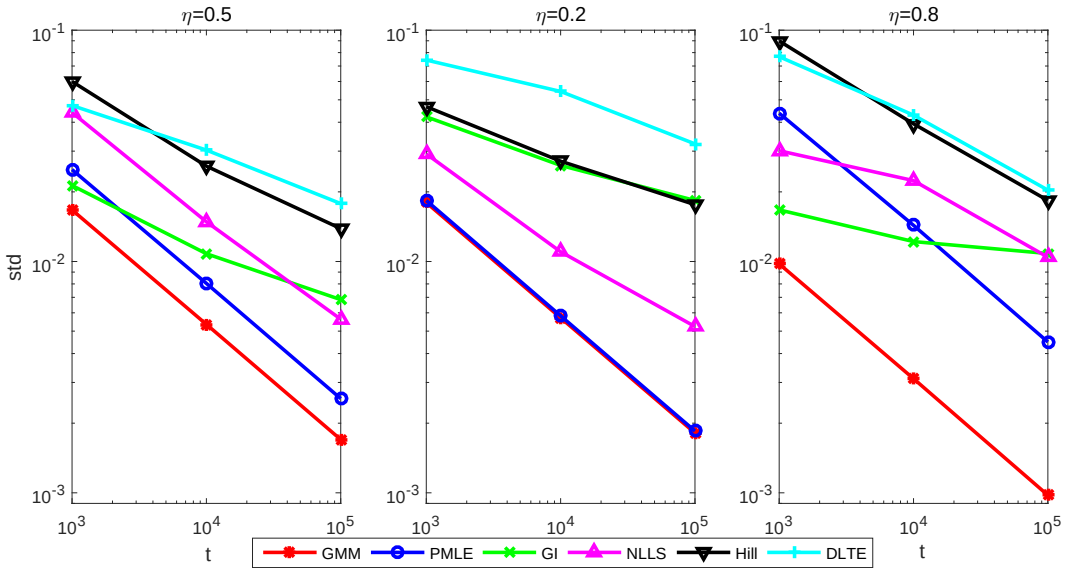


Figure A.1: Degree distributions for a simulation of the CF model with benchmark parameters.