Estimation of a Scale-Free Network Formation Model*

Anton Kolotilin† and Valentyn Panchenko‡

This version: June, 2018

Abstract

Growing evidence suggests that many social and economic networks are scale free in that their degree distribution has a power-law tail. A common explanation for this phenomenon is a random network formation process with preferential attachment. For a general version of such a process, we develop the pseudo maximum likelihood and generalized method of moments estimators. We prove consistency of these estimators by establishing the law of large numbers for growing networks. Simulations suggest that these estimators are asymptotically normally distributed and outperform the commonly used non-linear least squares and Hill (1975) estimators in finite samples. We apply our estimation methodology to a co-authorship network.

JEL Codes: C15, C45, C51, D85

Keywords: law of large numbers, consistency, degree distribution, scale-free network

*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 various workshops for helpful comments. Kolotilin started working on the paper during his PhD at MIT, whose hospitality is gratefully acknowledged. Kolotilin acknowledges support from the Australian Research Council Discovery Early Career Research Award DE160100964.

†UNSW Business School, School of Economics. Email: akolotilin@gmail.com

‡UNSW Business School, 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 neighbours is approximately proportional to $d^{-\gamma}$ for large d, where γ is a positive constant called the *power-law parameter*. Such networks are called *scale free*. The power-law parameter plays an important role for network topology and network-related phenomena ranging from information dissemination and transmission of viruses to aggregate macro-economic fluctuations (Albert and Barabasi, 2002; Gabaix, 2011; Acemoglu et al., 2012). 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) built the first theoretical model of scale-free network formation. In this model, as the network evolves, new edges are proportionally more likely to connect to higher-degree vertices than lower-degree vertices. Such a process is called *preferential attachment*. Cooper and Frieze (2003) and Cooper (2006) introduced and analyzed a generalized model of scale-free network formation, hereafter referred to as the CF model. Their model nests various scale-free network formation models, including the Barabasi and Albert model and popular hybrid models, such as Jackson and Rogers (2007), in that the CF model is able to generate networks with the same (asymptotic) degree distribution.

In the CF model, there is initially a small fixed network. At each subsequent period, a new vertex and a random number of edges are added. Some of the added edges connect the new vertex with the existing vertices, and others connect the existing vertices between themselves. The endpoints of the added edges are chosen from the existing vertices uniformly at random with some probability and by preferential attachment with the complementary probability. Cooper (2006) shows that the asymptotic degree distribution in the CF model depends only on the subset of parameters, with one parameter being the expected fraction η of edge endpoints added by preferential attachment. Moreover, the asymptotic degree distribution has a power-law tail, where the power-law parameter is $1 + 1/\eta$.

[1] In this paper, we focus on the degree distribution because it is "one of the most fundamental of network properties" (Newman, 2010, p. 243) and it determines many other topological properties (see, e.g., Graham, 2017, p. 1040, and references therein). Other network characteristics, such as clustering, can be parameterized separately from the degree distribution, as suggested, e.g., by Bollobas and Riordan (2003).

Social networks (co-authorship, citation, inventor, movie actor, and sexual relation networks), economic networks (production, interbank market, power-grid networks), biological networks (ecological, cellular, protein, and neural networks) and communication networks (WWW and cell phone networks) often exhibit a power law in the tail of the degree distribution (see, e.g., Newman, 2001; Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2002; Jackson, 2008; Atalay et al., 2011).

The goal of this paper is to develop a rigorous methodology for estimating the parameters of the CF model that determine the asymptotic degree distribution. The challenge is that, in random network formation models, the vertex degrees have non-standard interdependencies and exhibit substantial heterogeneity ("older" vertices have a higher degree than "younger" vertices).

Despite the existence of a variety of theoretical models of scale-free network formation, there is a lack of econometric methods that estimate the structural parameters of these models.[2] Instead, the power-law parameter is often estimated using the log-log rank-degree regression (Gabaix and Ibragimov, 2011, and references therein) or the Hill estimator (Hill, 1975). These popular estimators belong to a large class of tail estimators (Beirlant et al., 2006). Most of these tail estimators, however, are designed for continuous independent random variables that have identical distributions with a specific tail behaviour. Moreover, the performance of tail estimators strongly depends on the appropriate choice of the number of tail observations.

As an alternative to tail estimators, Pennock et al. (2002) and Jackson and Rogers (2007) use non-linear least squares (NLS) to fit the empirical degree distribution to an approximation of the parametrized asymptotic degree distribution. Goldstein et al. (2004) and Clauset et al. (2009) illustrate that such procedures often give highly biased estimates of the model parameters and argue that maximum likelihood estimation is much more robust.[3]

Most related to our paper, Atalay et al. (2011) and Atalay (2013) use pseudo maximum likelihood (PML) to estimate the parameters of random network formation models. Specifically, they calculate the pseudo likelihood assuming that each vertex degree is independent and identically distributed according to the derived asymptotic degree distribution. Since the vertex degrees in their models are interdependent and have different distributions, the asymptotic and finite-sample properties of the PML estimator are not known, but they can be analyzed using our methodology.

To estimate the parameters of the CF model, we develop a class of the generalized method of moments (GMM) and PML estimators. These estimators are computationally simple,

[2]At the same time, there is a growing literature in the econometrics of non-scale-free network formation and network estimation; see Chandrasekhar (2016) and de Paula (2017) for an overview and the recent works of Christakis et al. (2010), Comola and Fafchamps (2014), Goldsmith-Pinkham and Imbens (2013), Chandrasekhar and Jackson (2016), Chandrasekhar and Lewis (2016), Sheng (2016), Graham (2017), Mele (2017), de Paula et al. (2018) among others.

[3]Jackson and Rogers (2007) note that deriving analytically and then computing numerically the true likelihood of the degree sequence appears to be impossible for scale-free network formation models. König (2016) uses likelihood-free Markov-Chain Monte Carlo methods to estimate a similar model.

because they require calculating only a sample average of a moment function, as opposed to calculating the true likelihood of the degree sequence. We show formally that the GMM and PML estimators consistently estimate the parameters of the CF model. We also provide a procedure for conservative variance estimation. The standard consistency results use the uniform laws of large numbers for independent or weakly dependent random variables. But we cannot use these results because the vertex degrees in the CF model have non-standard interdependencies.

To prove consistency of the GMM and PML estimators, we establish the uniform law of large numbers for growing networks from the first principles. Although we rely on certain properties of the degree distribution established for the CF model, our proof is sufficiently general and can be extended to other network formation models. We also establish the weak convergence of the tail empirical measure to formally show consistency of the Hill estimator for the CF model.

Our simulations suggest that the GMM and PML estimators perform well in finite samples; that is, the GMM and PML estimators have a substantially smaller bias and variance than the NLS and Hill estimators. Moreover, the distribution of the GMM and PML estimators is closer to the normal distribution compared to the distribution of the NLS and Hill estimators. We apply our estimation methodology to the network of co-authorship relations among economists, which was investigated in Goyal et al. (2006) and Jackson and Rogers (2007). We build and provide a comprehensive estimation package, which includes the PML and GMM estimators, various implementations of the NLS and Hill estimators, as well as other commonly used tail estimators (see the Supplementary Appendix).

2 Network Formation Model

2.1 Setup

Following Cooper and Frieze (2003) and Cooper (2006), we describe the CF network formation model as a statistical process. We then discuss its relation to other growing network models, economic micro-foundations, and applications in Section 2.3.

Consider a random graph process, $(G(t))_{t\geq 1} = (V(t), E(t))_{t\geq 1}$, where $V(t)$ is a set of vertices and $E(t)$ is a set of edges at the end of each time $t \in \{1, 2, \dots\}$.[4] In economic appli-

[4]Formally, we should refer to this process as a multi-graph process 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

cations, 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 any 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, 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 a random number of edges $M(t)$ 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 $\boldsymbol{p} = (p_0, \ldots, p_m, \ldots, p_P)$ and $\boldsymbol{q} = (q_0, \ldots, q_M, \ldots, q_Q)$, where $p_m = \Pr(m(t) = m)$ and $q_M = \Pr(M(t) = M)$. These distributions characterize 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 $e_i^m(t)$, $i = 1, \ldots, m(t)$, originating from new vertex t. The terminal vertex of each edge $e_i^m(t)$, the vertex with which t connects, is chosen independently from $V(t-1)$ by preferential attachment[5] with probability A_1 and uniformly at random with probability $A_2 = 1 - A_1$:

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

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

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

Define $D_t(d)$ as the number of vertices of the graph $G(t)$ that have degree d. The degree distribution $P_t(d)$ is defined as the fraction of vertices of the random graph $G(t)$ that have

considered process. Furthermore, we treat all edges as undirected, but it is straightforward to extend the analysis to directed graph processes.

[5]Preferential attachment to higher-degree vertices arises naturally in growing networks, more detailed micro-foundations are in Section 2.3. Importantly, the network is scale free if and only if the attachment probability is asymptotically linear in the vertex degree.

degree d; that is, $P_t(d) = D_t(d)/|V(t)|$ is a random variable. Corollary 1 in the next subsection shows that $P_t(d)$ converges in probability to $P(d)$ for all d as $t \to \infty$. The limiting fraction $P(d)$ is called the *asymptotic degree distribution* of the graph process $(G(t))_{t\geq 1}$.

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

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

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

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

As in Cooper and Frieze (2003), we assume that the parameters are such that $0 < \eta < 1$ holds.[6] Note that the structural parameters A_1, B_1, C_1, and $\boldsymbol{q}$ cannot be individually identified from the asymptotic degree distribution and we focus on estimating $\boldsymbol{p}$, $\overline{M}$, and η.

2.2 Asymptotic Degree Distribution

Our derivation of the asymptotic degree distribution relies on the results of Cooper (2006) presented in Appendix A. Proposition 1 restates the concentration results of Cooper (2006) in a form convenient for our analysis.

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

$$\Pr\left(\left|\frac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\right| \geq 2\frac{P(d;\eta,\overline{M},\boldsymbol{p})}{\sqrt{\ln t}}\right) = O\left(\frac{1}{\ln t}\right),$$

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

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

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

Corollary 1 demonstrates the limiting properties of the degree distribution.

[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.

Corollary 1 *We have the following:*

1. *The degree distribution $P_t(d) = D_t(d)/|V(t)|$ of the graph $G(t)$ converges in probability to $P(d;\eta,\overline{M},\boldsymbol{p})$ as $t \to \infty$.*

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

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

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

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.

2.3 Discussion and Examples

The CF model nests many network formation models in that it is able to generate networks with (asymptotic) degree distributions ranging from the exponential degree distribution of the growing Poisson random networks to the power-law degree distribution of preferential attachment networks, including any degree distribution of a hybrid model embedding the elements of both.[7] Importantly, we use the CF model to characterize only the degree distribution, rather than clustering and other characteristics of social and economic networks. Bollobas and Riordan (2003, Theorem 5) suggest that it is possible to introduce any level

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

of clustering in a graph process with preferential attachment without changing the asymptotic degree distribution; Dorogovtsev and Mendes (2002) and Jackson and Rogers (2007) provide examples of such processes. Hence, the CF model may be extended to include any level of clustering, and similar arguments can be made about other network characteristics. Thus, information about clustering and other characteristics is of limited use for estimating parameters $(\eta, \overline{M}, \boldsymbol{p})$, which determine the degree distribution in the CF model.

The statistical network models based on preferential attachment, including the CF model, provide a good fit to real physical, social and economic networks. However, these models lack rigorous micro-foundations for individual strategic behaviour and structural interpretation of game-theoretic link formation models specified in Christakis et al. (2010), König et al. (2014), and Mele (2017) among others. A growing literature is trying to fill this gap.

Jackson and Rogers (2007) built a growing network model related to the CF model using intuitive behavioural principles which motivate preferential attachment. In this model, new vertices connect with existing vertices in two steps: uniformly at random ("meeting strangers"), and through the direct neighbours of the linked vertices ("meeting friends of friends"). This model has the properties of preferential attachment since the probabilities of connecting to friends of friends are proportional to their in-degree.[8] Jackson and Rogers applied the model to web links formation and various social networks including co-authorship, citation, friendship, and romantic relations. This model has been widely used in the literature. Mayer and Puller (2008) used it to describe social networks on university campuses. Chaney (2014) applied a similar model to firms searching for new partners in the context of international trade. Atalay et al. (2011) applied an expanded model to input-output link formation in a production network. Luttmer (2011) proposed a related model of firm growth.

There is a well-established literature with micro-founded strategic network formation models, starting from Jackson and Wolinsky (1996) and Bala and Goyal (2000). In these models, a star network typically emerges as an equilibrium configuration. Introducing noise in the decision-making process of agents in these strategic network formation models leads to the emergence of networks with preferential attachment, similarly to the CF model. König (2016) provides an example of such a model. In particular, he builds a micro-founded network formation model with general marginal payoff functions in the information sharing environment.

[8]The CF model is able to generate the asymptotic degree distribution of the Jackson and Rogers (2007) model by setting $\overline{m} = m_n p_n + m_r p_r,\ q_0 = 1,\ A_1 = p_r m_r/\overline{m},\ A_2 = 1 - A_1$, where m_r and m_n are the numbers of considered edges with strangers and friends of friends, respectively, and p_r and p_n are the corresponding probabilities of creating an edge; and modifying the probability of attachment to $A_1 d_i(v, t-1)/|E(t-1)| + A_2/|V(t-1)|$ to adjust for the undirected graph (compare with (1) in Jackson and Rogers, 2007).

The model is similar to the Jackson and Rogers (2007) growing network, where new agents meet strangers and friends of friends to form edges. The innovation is that the benefits of forming an edge with strangers and friends of friends are modelled explicitly and idiosyncratic noise is added in decision making. With a small amount of noise, centralized star-type networks emerge, but with a large amount of noise and a small pool of potential connections, the networks that exhibit preferential attachment and a power-law tail in the degree distribution emerge. In the latter case, the asymptotic degree distribution (König, 2016, Proposition 2) is analogous to the asymptotic degree distribution in the CF model given by Proposition 1.

3 Methodology

3.1 Preliminaries

We propose the PML and GMM estimators for 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 η, $\overline{M}$, and $\boldsymbol{p} = (p_0, \ldots, p_P)$. Section 3.3 shows that these parameters are identified. Parameter η is of the highest interest in this model as it determines the power-law parameter $1 + 1/\eta$. 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 a P dimensional simplex.[9] 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 where P is unknown, it can be chosen using model selection procedures such as AIC or BIC (see, e.g., Burnham and Anderson, 2002), but we do not explore the asymptotic properties of such procedures.

Let $\boldsymbol{\theta} = (\eta, \overline{M}, \boldsymbol{p})$, i.e., $\boldsymbol{\theta}$ is a multi-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 specify the estimators and establish their consistency as t goes to infinity. This asymptotics is similar to the standard large sample asymptotics, 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 t. Thus, all asymptotic results will continue to hold if we consider alternative asymptotics, 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.

[9] 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 Δ^P.

3.2 Laws of Large Numbers

Before introducing the 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 the standard technical conditions for the distribution function $P(d;\boldsymbol{\theta})$ given by Proposition 1. However, we cannot establish uniform convergence using the standard uniform laws of large numbers for independent or weakly-dependent stationary processes, because the CF model yields substantial heterogeneity in the vertex degree distributions and nonstandard vertex degree interdependencies. The main technical contribution of the paper is the uniform law of large numbers established for the CF model.[10]

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 $\|a(d;\boldsymbol{\theta})\| < Fd$ 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 heuristically 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/η_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

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

and show that each term converges in probability to zero if $d_t^\diamond$ grows to infinity at a rate much slower than $\ln t$ and $d_t^*(\eta_0)$. $\widehat{S}_1 \xrightarrow{P} 0$ 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 $\sum_{d=d_t^\diamond}^{\infty} d D_t(d)/|V(t)|$, which

[10] We closely follow 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).

is equal to the difference between $\widehat{S}_4 = \sum_{d=0}^{\infty} dD_t(d)/|V(t)|$ and $\widehat{S}_5 = \sum_{d=0}^{d_t^{\diamond}-1} dD_t(d)/|V(t)|$. $\widehat{S}_5 \xrightarrow{P} 2(\overline{m}_0 + \overline{M}_0)$ 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.

Using Proposition 1 and Corollary 1, we can also establish the weak convergence of the tail empirical measure, which is a crucial property for consistency of tail estimators.

Proposition 3 *If $k_t/t \to 0$ and $k_t\sqrt{\ln t}/t \to \infty$ as $t \to \infty$, then, for all $x > 0$,*

$$\frac{1}{k_t} \sum_{d=[x\overline{F}^{-1}(k_t/t)]}^{\infty} D_t(d) \xrightarrow{P} x^{-1/\eta_0},$$

where $\overline{F}(z) = \sum_{d=[z]}^{\infty} P(d;\boldsymbol{\theta}_0)$, with $z \geq 0$, and $\overline{F}^{-1}(y) = \inf\{d : \overline{F}(d) \leq y\}$, with $0 < y < 1$.

3.3 Consistency of PML, GMM, and Hill Estimators

The established laws of large numbers allow us to extend the standard consistency results to our network formation model. We define the pseudo log-likelihood based on the asymptotic degree distribution given by Proposition 1 as follows:

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

The true log-likelihood is different from the pseudo log-likelihood, because (i) the vertex degrees are interdependent, and (ii) the finite sample and asymptotic degree distributions differ.

The PML estimator is defined as:

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

The plug-in PML estimator is formally defined as:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{P}}^{\mathrm{PML}} = \arg\max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{L}_t(\boldsymbol{\theta}),$$
$$\text{s.t. } \overline{m} + \overline{M} = \frac{1}{2}\sum_{d=0}^{\infty} d\frac{D_t(d)}{|V(t)|}.$$

That is, $\widehat{\boldsymbol{\theta}}_{\mathrm{P}}^{\mathrm{PML}}$ is obtained by replacing $\overline{m} + \overline{M}$ in $\widehat{L}_t(\boldsymbol{\theta})$ with its estimate[11]

$$\widehat{\overline{m} + \overline{M}} = \frac{1}{2}\sum_{d=0}^{\infty} d\frac{D_t(d)}{|V(t)|}$$

[11]Parameter $\overline{M}$ is first expressed as $(\overline{m} + \overline{M}) - \sum_{m=0}^{P} mp_m$.

and maximizing $\widehat{L}_t(\boldsymbol{\theta})$ over the remaining parameters η and $\boldsymbol{p}$. The estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{P}}^{\mathrm{PML}}$ is faster to compute than $\widehat{\boldsymbol{\theta}}^{\mathrm{PML}}$, because it requires maximization over one less parameter. Notice that $\widehat{\overline{m}+\overline{M}}$ consistently estimates $\overline{m}_0+\overline{M}_0$, by the law of large numbers applied to independent random variables $m(t)$ and $M(t)$.

Consistency of the PML and plug-in PML estimators is established in Proposition 4.

Proposition 4 *Let $\overline{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}$ be compact and $\boldsymbol{\theta}_0 \in \overline{\boldsymbol{\Theta}}$. If $\widehat{\boldsymbol{\theta}}$ satisfies $\widehat{L}_t(\widehat{\boldsymbol{\theta}}) \geq \max_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \widehat{L}_t(\boldsymbol{\theta}) + o_P(1)$, then $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$. In particular, $\widehat{\boldsymbol{\theta}}^{\mathrm{PML}} \xrightarrow{P} \boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{P}}^{\mathrm{PML}} \xrightarrow{P} \boldsymbol{\theta}_0$.*

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

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

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

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

Since Proposition 1 gives the explicit expression for $P(d;\boldsymbol{\theta})$, it is easy to verify whether a given $g(d;\boldsymbol{\theta}_0)$ has zero mean. In particular, from the discussion of the PML estimators, it is evident that this property is satisfied for the moment function vector that consists of the *score function* vector and the *degree function*:

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

Proposition 5 specifies sufficient conditions on moment function $g(d;\boldsymbol{\theta})$ and matrix $\widehat{W}$ for the GMM estimator $\widehat{\boldsymbol{\theta}}$ to be consistent.

Proposition 5 *Let $\widehat{\boldsymbol{\theta}}$ maximize $\widehat{Q}_t(\boldsymbol{\theta})$ where $\widehat{W} \xrightarrow{P} W$, and (i) $W\sum_{d=0}^{\infty} g(d;\boldsymbol{\theta})P(d;\boldsymbol{\theta}_0) = 0$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; (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 $\|g(d;\boldsymbol{\theta})\| < Fd$ for all $d \in \mathbb{N}_+$ and all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, then $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.*

2. *If $g(d;\boldsymbol{\theta}) = \left(\nabla_{\boldsymbol{\theta}} \ln P(d;\boldsymbol{\theta}),\ d - 2(\overline{m}+\overline{M})\right)'$, then $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.*

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

We suggest using the moment function vector that consists of the score function vector and the degree function. With appropriately chosen weights in $\widehat{W}$, the 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 5 shows, to ensure consistency of the GMM estimators based on this moment function vector, 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 4, 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., Newey and McFadden, 1994, p. 2127).[12]

The Hill estimator, which is the most common tail estimator, is defined as

$$\widehat{\eta}^{\text{Hill}} = \frac{1}{k_t} \sum_{d=d_t^{\dagger}+1}^{\infty} D_t(d) \ln \frac{d}{d_t^{\dagger}}$$

for some $d_t^{\dagger}$ and $k_t = \sum_{d=d_t^{\dagger}+1}^{\infty} D_t(d)$.

Consistency of the Hill estimator for the CF model is established in Proposition 6.[13]

Proposition 6 *If $k_t/t \to 0$ and $k_t\sqrt{\ln t}/t \to \infty$ as $t \to \infty$, then $\widehat{\eta}^{\text{Hill}} \xrightarrow{P} \eta_0$.*

[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 $\boldsymbol{\theta}_0$. As Rothenberg (1971) shows, 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 7 we derive G; so for given $\boldsymbol{\theta}_0$ and W, we can numerically verify local identification – in particular, it holds for the GMM estimators used in our simulations.

[13]Using Proposition 6, we can establish consistency of related tail estimators by showing that they have the same probability limit as $\widehat{\eta}^{\text{Hill}}$. For example, for a discrete distribution, Clauset et al. (2009) propose

$$\widehat{\eta}_{\text{C}}^{\text{Hill}} = \frac{1}{k_t} \sum_{d=d_t^{\dagger}+1}^{\infty} D_t(d) \ln \frac{d}{d_t^{\dagger} + 1/2}.$$

This estimator is consistent because $d_t^{\dagger}/[x\overline{F}^{-1}(k_t/t)] \xrightarrow{P} 1$, and thus $(d_t^{\dagger} + 1/2)/d_t^{\dagger} \xrightarrow{P} 1$, by Resnick and Stărică (1995, Proposition 2.1).

3.4 Discussion of Asymptotic Normality and Variance

We now specify sufficient conditions for establishing asymptotic normality of the GMM estimators, and thus of the PML and plug-in PML estimators.

Proposition 7 *Let* $\widehat{\boldsymbol{\theta}}$ *maximize* $\widehat{Q}_t(\boldsymbol{\theta})$*, where* $g(d;\boldsymbol{\theta}) = \left(\nabla_{\boldsymbol{\theta}} \ln P(d;\boldsymbol{\theta}),\ d - 2(\overline{m} + \overline{M})\right)'$*,* $\widehat{W} \xrightarrow{P} W$*,* $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$*, 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 (see Footnote 12). Condition (ii) is an asymptotic normality condition for a sample average of $g(d;\boldsymbol{\theta}_0)$. Asymptotic normality is supported by our simulations but is not proved formally.

To obtain a consistent estimate of the asymptotic variance of $\widehat{\boldsymbol{\theta}}$, we need to find consistent estimates of G and Σ.[14] G can be consistently estimated by $\widehat{G} = \sum_{d=0}^{\infty} g(d;\widehat{\boldsymbol{\theta}}) D_t(d)/|V(t)|$,[15] but obtaining a consistent estimate of Σ is complicated due to vertex degree interdependencies of unknown form. Moreover, some of the nuisance parameters affect Σ but are not identified from the asymptotic degree distribution $P(d;\boldsymbol{\theta})$. To illustrate this point, consider the plug-in PML estimator. The estimate $\widehat{\overline{m} + \overline{M}}$ is asymptotically normally distributed. Indeed,

$$\begin{aligned}\sqrt{|V(t)|}\left(\widehat{\overline{m} + \overline{M}} - \overline{m}_0 - \overline{M}_0\right) &= \sqrt{|V(t)|}\left(\frac{1}{2}\sum_{d=0}^{\infty} d\frac{D_t(d)}{|V(t)|} - \overline{m}_0 - \overline{M}_0\right) \\ &\rightsquigarrow N\left[0, \mathrm{Var}(m(t)) + \mathrm{Var}(M(t))\right],\end{aligned}$$

by the central limit theorem applied to independent random variables $m(t)$ and $M(t)$. But the asymptotic variance $\mathrm{Var}(m(t)) + \mathrm{Var}(M(t))$ depends on the distribution $\boldsymbol{q}$ of $M(t)$ that is not identified from $P(d;\boldsymbol{\theta}_0)$.[16]

[14] By the assumptions of Proposition 7, $\widehat{W} \xrightarrow{P} W$ and $G'WG$ is nonsingular. If, in addition, $\widehat{G} \xrightarrow{P} G$ and $\widehat{\Sigma} \xrightarrow{P} \Sigma$, then by continuous mapping theorem,

$$\left(\widehat{G}'\widehat{W}\widehat{G}\right)^{-1}\widehat{G}'\widehat{W}\widehat{\Sigma}\widehat{W}\widehat{G}\left(\widehat{G}'\widehat{W}\widehat{G}\right)^{-1} \xrightarrow{P} (G'WG)^{-1}G'W\Sigma WG\,(G'WG)^{-1}.$$

[15] Consistency, continuity, and uniform convergence imply:

$$\left\|\widehat{G} - G\right\| \le \left\|\widehat{G} - G(\widehat{\boldsymbol{\theta}})\right\| + \left\|G(\widehat{\boldsymbol{\theta}}) - G\right\| \le \sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\left\|\sum_{d=0}^{\infty} g(d;\boldsymbol{\theta})\frac{D_t(d)}{|V(t)|} - G(\boldsymbol{\theta})\right\| + \left\|G(\widehat{\boldsymbol{\theta}}) - G\right\| \xrightarrow{P} 0.$$

[16] This asymptotic variance could be estimated if the time evolution of the network was observed.

We propose conservative variance estimates of $\widehat{\boldsymbol{\theta}}$ based on a parametric bootstrap (Efron and Tibshirani, 1994, Chapter 6.5) and the principle of maximum entropy (Maasoumi, 1993).[17] The procedure consists of the following steps: (1) compute an estimate $\widehat{\boldsymbol{\theta}}$ from the original network with t vertices; (2) generate N networks with t vertices each by sampling from the CF model with the parameter $\widehat{\boldsymbol{\theta}}$; (3) compute an estimate $\widehat{\boldsymbol{\theta}}^*$ from each network; and (4) calculate the sample variance of the N estimates $\widehat{\boldsymbol{\theta}}^*$. For sufficiently large N, this sample variance approximates the variance of the estimate $\widehat{\boldsymbol{\theta}}$ (see, e.g., Horowitz, 2001).

Since parameters $\boldsymbol{q}$, A_1, B_1, and C_1, which are necessary to sample from the parametric CF model, are not fully identified, we use the principle of maximum entropy in our procedure and choose $\boldsymbol{q}$ to be a geometric distribution, i.e., $q_M = \gamma(1-\gamma)^M$ for $M \in \{0, 1, \dots\}$, where $\gamma = (\widehat{\overline{M}} + 1)^{-1}$; and $A_1 = B_1 = C_1 = 2\widehat{\eta}(\widehat{\overline{m}} + \widehat{\overline{M}})/(\widehat{\overline{m}} + 2\widehat{\overline{M}})$. The principle of maximum entropy aims to specify the least informative distribution subject to (partially) available information. It is well known that our choice of the geometric distribution for $\boldsymbol{q}$ maximizes the entropy of $M(t)$ subject to the constraints that the support of $M(t)$ is $\{0, 1, \dots\}$ and the expectation of $M(t)$ is $\widehat{\overline{M}}$.

We now show that our choice of (A_1, B_1, C_1) asymptotically maximizes the entropy of the fraction of endpoints added by preferential attachment $\widetilde{\eta}$, subject to the constraints that the expected fraction of endpoints added by preferential attachment is $\widehat{\eta}$, and the expected numbers of new-old edges and old-old edges are $\widehat{\overline{m}}$ and $\widehat{\overline{M}}$. By Lyapounov's central limit theorem, $\sqrt{t}\,(\widetilde{\eta} - \widehat{\eta}) \rightsquigarrow N\,[0, \mathrm{Var}\,(\widetilde{\eta})]$, where the asymptotic variance $\mathrm{Var}\,(\widetilde{\eta})$ is given by

$$\mathrm{Var}\,(\widetilde{\eta}) = \frac{\widehat{\overline{m}} A_1(1-A_1) + \widehat{\overline{M}}(B_1(1-B_1) + C_1(1-C_1))}{4(\widehat{\overline{m}} + \widehat{\overline{M}})^2}.$$

It is well known that the entropy of a normally distributed random variable increases with its variance; so maximizing the entropy of $\widetilde{\eta}$ is asymptotically equivalent to maximizing $\mathrm{Var}\,(\widetilde{\eta})$. Finally, it is straightforward to show that our choice of (A_1, B_1, C_1) maximizes $\mathrm{Var}\,(\widetilde{\eta})$ subject to $(\widehat{\overline{m}} A_1 + \widehat{\overline{M}}(B_1 + C_1))/2(\widehat{\overline{m}} + \widehat{\overline{M}}) = \widehat{\eta}$.

If the conditions of Proposition 7 hold, then the GMM estimator with $\widehat{W} \xrightarrow{P} \Sigma^{-1}$ is asymptotically efficient (see Newey and McFadden, 1994, Theorem 5.2), but only in the class of GMM estimators with the moment function vector $g(d; \boldsymbol{\theta}) = \left(\nabla_{\boldsymbol{\theta}} \ln P(d; \boldsymbol{\theta}),\ d - 2(\overline{m} + \overline{M})\right)'$. As a consistent estimator of Σ is not readily available, we also consider the unweighted GMM estimator with $\widehat{W}$ equal to the identity matrix I. Simulations suggest that the GMM estimators with $\widehat{W} = I$ and $\widehat{W} \xrightarrow{P} \Sigma^{-1}$ have a similar bias and variance.

[17] In the next section we demonstrate by simulations that this procedure performs better than a procedure disregarding vertex degree interdependences.

4 Simulations and Application

4.1 Simulation Studies

We investigate the finite sample performance of the PML, plug-in PML_P, unweighted GMM_U, and optimal GMM_O estimators (jointly referred to as the PML-GMM estimators in this section) and compare it to that of the popular NLS and Hill estimators.[18] The weighting matrix for the GMM_O estimator is $\widehat{W} = \widehat{\Sigma}^{-1}$, where $\widehat{\Sigma}$, the sample variance-covariance matrix of the moments, is computed from 100000 replications of the network.

We report the two best performing variants of the NLS estimator: NLS_D, in which observed *distinct* degrees and the corresponding empirical distributions are used as observations, and NLS_R, in which all observed degrees with *repetitions* and the corresponding ordinal ranks are used as observations.

The performance of the Hill estimator heavily depends on the selection of the tail cutoff $d_t^\dagger$. Here, we report the Hill estimator based on a popular selection method in which $d_t^\dagger$ minimizes the asymptotic mean squared error of the estimator (see Beirlant et al., 1996; Matthys and Beirlant, 2000, and the Supplementary Appendix for more details).

The performance of the estimators for the CF model is compared in terms of their bias and standard deviation. In order to support our estimation procedure for the variance, we compute the sample standard deviations of the estimators using (i) the model with the true parameters, SD, (ii) the model where the nuisance parameters are replaced with the values set to maximize the entropy, SD_E,[19] and (iii) the model where degrees are sampled independently directly from the asymptotic degree distribution, SD_I. We also report the Kolmogorov-Smirnov (KS) statistic to assess the closeness to the normal distribution. All simulation results are based on 10000 replications.

As a benchmark, we consider the CF model with the following parameters: $t = 1000$,

[18] The NLS estimator for the CF model is formally derived and various implementations are explained in the Supplementary Appendix. This section reports the results for the Hill estimator, which turns out to be among the best performing tail estimators. The Supplementary Appendix includes an extensive comparison with other popular tail estimators including the discrete maximum likelihood tail estimator (Goldstein et al., 2004), the Hill estimator with continuity correction (Clauset et al., 2009), the log-log rank-degree regression with continuity correction (Gabaix and Ibragimov, 2011), the Pickands (1975) estimator, the Dekkers et al. (1989) moment estimator, and the Smith (1987) maximum likelihood estimator of the Paretian excesses model. We also compare different methods for selecting the tail cutoff. We provide Matlab codes for all these estimators.

[19] In particular, the values of the nuisance parameters, which are not identified from the asymptotic degree distribution $P(d;\boldsymbol{\theta})$, are set to $q_M = \gamma(1-\gamma)^M$ for $M \in \{0,1,\dots\}$, where $\gamma = (\overline{M}+1)^{-1}$; and $A_1 = B_1 = C_1 = 2\eta(\overline{m}+\overline{M})/(2\overline{M}+\overline{m})$.

Table 1: Comparison of various estimators for $t = 1000$

	PML	PML_P	GMM_U	GMM_O	NLS_D	NLS_R	Hill_MS
Bias(η)	0.0009	0.0007	0.0007	0.0007	0.0436	0.0374	0.0816
SD(η)	0.0250	0.0169	0.0167	0.0167	0.0467	0.0176	0.0606
$\text{SD}_\text{E}(\eta)$	0.0312	0.0220	0.0218	0.0218	0.0556	0.0222	0.0639
$\text{SD}_\text{I}(\eta)$	0.0487	0.0526	0.0529	0.0529	0.0805	0.0558	0.1336
KS(η)	0.0083	0.0071	0.0068	0.0070	0.0235	0.0187	0.0216
Bias($\overline{M}$)	0.0019	0.0001	0.0001	0.0001			
SD($\overline{M}$)	0.0418	0.0157	0.0157	0.0157			
$\text{SD}_\text{E}(\overline{M})$	0.0762	0.0605	0.0605	0.0605			
$\text{SD}_\text{I}(\overline{M})$	0.1378	0.1654	0.1653	0.1654			
KS($\overline{M}$)	0.0157	0.0190	0.0138	0.0136			

$p_0 = 1$ ($m(t) = 0$), $q_1 = q_2 = 0.5$ ($\overline{M} = 1.5$), and $A_1 = B_1 = C_1 = 0.5$ ($\eta = 0.5$). As in Bollobas et al. (2001), we assume that the initial graph, $G(1)$, consists of one vertex and a random number $\max\{m(1) + M(1), 1\}$ of loops.[20] In this benchmark, we assume that it is known that $P = 0$.

Table 1 reports the results for the benchmark specification. For parameter η, the PML-GMM estimators show a substantially lower bias and standard deviation relative to the NLS and Hill estimators, with the exception that the NLS_R estimator has a comparable standard deviation to that of the PML-GMM estimators. Among the PML-GMM estimators, the plug-in PML, unweighted GMM, and optimal GMM estimators are very similar in performance. They are closely followed by the PML estimator. Based on our analysis, we recommend using the plug-in PML estimator due to its strong performance and simpler implementation. As anticipated, the maximum entropy approach produces conservative standard deviations, which are higher than those based on the true nuisance parameters. Moreover, the standard deviations are more accurately estimated based on the assumption that the nuisance parameters are set to maximize the entropy rather than the assumption that the degrees are independently identically distributed according to the asymptotic degree distribution. Finally, the KS statistics suggest that the distribution of the PML-GMM estimators is closer to the normal distribution compared to the distribution of the NLS and Hill estimators.

[20] As a robustness check, we have also considered different initial graphs and different parameter values, some of which are reported in the Supplementary Appendix. The qualitative comparison of the estimators remains the same.

Table 2: Absolute bias and (standard deviation) of $\widehat{\eta}$ for various t

t	PML	PML_P	GMM_U	NLS_D	NLS_R	Hill
1000	0.0009	0.0007	0.0007	0.0436	0.0374	0.0816
	(0.0250)	(0.0169)	(0.0167)	(0.0467)	(0.0176)	(0.0606)
10000	0.0001	0.0001	0.0001	0.0405	0.0386	0.0365
	(0.0080)	(0.0054)	(0.0053)	(0.0149)	(0.0052)	(0.0347)
100000	0.0001	0.0000	0.0000	0.0335	0.0394	0.0187
	(0.0026)	(0.0017)	(0.0017)	(0.0055)	(0.0016)	(0.0199)

Table 2 illustrates how the absolute bias and standard deviation of the estimators change with the network size t. The bias of the PML-GMM estimators is small relative to the bias of the NLS and Hill estimators. While the bias of the PML-GMM and Hill estimators tends to vanish as t becomes very large, the bias of the NLS estimator persists. The results agree with formally established consistency of the PML-GMM and Hill estimators. The standard deviation of the PML-GMM and NLS_R estimators appears to decrease at the $\sqrt{t}$-rate, while the standard deviation of the NLS_D and Hill estimators appears to decrease at a slower rate.

Next, we investigate the performance of the plug-in PML estimator in the case of over-specification, when the assumed order P is larger than the actual P, and in the case of misspecification, when the assumed P is smaller than the actual P. Table 3 reports the results for the true model with $P = 1$ and $p_0 = p_1 = 0.5$, and with the other parameters being as in the benchmark. The assumed order of the plug-in PML estimator is indicated by superscript P. The bias and standard deviation of the plug-in PML estimator are higher under overspecification, $P = 2$, than under the correct specification, $P = 1$, but they are still smaller than those of the NLS and Hill estimators. However, under misspecification, $P = 0$, the plug-in PML estimator has a substantial bias which is higher than the bias of the NLS and Hill estimators. In this sense, the plug-in PML estimator is not robust to misspecification. In practice, we may use model selection procedures, such as the BIC, to find an optimal order. The smallest value of the BIC is attained under the correctly specified model. The other PML-GMM estimators perform similarly to the plug-in PML estimator.

4.2 Empirical Application

We illustrate the applicability of the introduced methods by estimating the CF model for the network of co-authorship relations among economists publishing in journals listed by EconLit in the 1990s. This dataset was first considered by Goyal et al. (2006), who constructed a

Table 3: Overspecification and misspecification for $p_0 = p_1 = 0.5$ ($P = 1$)

	$\text{PML}_{\text{P}}^{P=1}$	$\text{PML}_{\text{P}}^{P=2}$	$\text{PML}_{\text{P}}^{P=0}$	NLS_{D}	NLS_{R}	Hill
Bias(η)	0.0006	0.0100	0.1593	-0.0302	-0.0221	0.0721
SD(η)	0.0203	0.0214	0.0163	0.0488	0.0187	0.0543
$\text{SD}_{\text{E}}(\eta)$	0.0244	0.0260	0.0207	0.0594	0.0241	0.0603
Bias($\overline{M}$)	0.0002	0.0404	0.4999			
SD($\overline{M}$)	0.0410	0.0664	0.0223			
$\text{SD}_{\text{E}}(\overline{M})$	0.0722	0.0912	0.0630			
Bias(p_0)	0.0001	0.0151				
SD(p_0)	0.0411	0.0436				
$\text{SD}_{\text{E}}(p_0)$	0.0418	0.0452				
BIC	4714.9	4721.3	4781.5			

network of collaborations in which every publishing author is represented by a vertex, and two authors are connected if they have published at least one paper together in the period of ten years from 1990 and 1999. The network contains $t = 81217$ authors with the average number of co-authors equal to 1.672 (i.e., $\widehat{\overline{m} + \overline{M}} = 0.836$). Jackson and Rogers (2007) estimated the tail parameter of the degree distribution for this network using the NLS estimator.

Since the support of $m(t)$ is not known, we consider various values of P. The models with $P = 1$ and $P = 2$ are close in terms of the BIC. We select the model with $P = 1$, which produces a closer fit to the empirical degree distribution in the tails. The conservative standard errors are computed with 1000 parametric bootstrap replications using the maximum entropy approach introduced in Section 3.4.

Table 4 shows the parameter estimates and their conservative standard errors for the co-authorship network.[21] For the Hill estimator, the tail cutoff that minimizes the asymptotic mean squared error is $d_t^\dagger = 26$. The PML-GMM estimators estimate η to be about 0.21, while the Hill and NLS estimators produce a wide range of estimates from 0.18 to 0.22.[22] The Hill estimator produces a better fit to the empirical distribution in the extreme tails, while the PML-GMM and NLS estimators give a better overall fit.

One of the advantages of the PML-GMM estimators, relative to the Hill and NLS estimators, is that they allow for structural estimation of the CF model; so we can obtain additional

[21] Since the optimal weighting matrix W is difficult to estimate and given the strong performance of the unweighted GMM estimator in the simulations, we do not consider the optimal GMM estimator in the application.

[22] These values are similar to the NLS estimate of $1/\eta = 4.7$ in Jackson and Rogers (2007).

Table 4: Parameter estimates and their standard errors for the co-authorship network

	PML	PML_P	GMM_U	NLS_D	NLS_R	Hill
$\widehat{\eta}$	0.2120	0.2127	0.2126	0.1818	0.2242	0.1814
SE(η)	0.0039	0.0038	0.0039	0.0089	0.0029	0.0422
$\widehat{\overline{M}}$	0.4049	0.4062	0.4062			
SE($\overline{M}$)	0.0047	0.0050	0.0048			
$\widehat{p_0}$	0.5706	0.5703	0.5709			
SE(p_0)	0.0043	0.0044	0.0044			

insights about the network formation process. The structural estimates suggest that the network is formed by an approximately equal number of new-old connections ($\overline{m} = 0.43$) and old-old connections ($\overline{M} = 0.41$). Moreover, a slight majority of new vertices, about 57%, have no initial edges (single-authored papers) and 43% of the new vertices have one initial edge (co-authored papers with one co-author). We also find that about 21% of connections are formed by preferential attachment. The preferential attachment mechanism seems natural in this setting as better known authors are more likely to attract new co-authors and, in this way, grow their network of collaborations.

5 Conclusion

We estimate a general model of scale-free network formation using the PML, GMM, NLS, and Hill estimators. By establishing the laws of large numbers for the CF model, we prove consistency of the PML, GMM, and Hill estimators. We also discuss asymptotic normality and conservative variance estimation of these estimators. Our simulations indicate that the PML and GMM estimators have virtually no bias and a smaller variance than the NLS and Hill estimators. We recommend using the plug-in PML estimator as it has comparable finite sample performance and is simpler in its implementation relative to the GMM estimators.

Our theoretical results are useful for a growing literature on estimation of network formation models. The methodology for establishing the laws of large numbers and consistency of various estimators can be extended to other growing network models. One of the main challenges with growing network models is that the vertex degrees have non-standard interdependencies. We hope that future research will better characterize these interdependencies, which, in turn, will help to prove asymptotic normality and find the asymptotic variance of the introduced PML and GMM estimators.

We focus entirely on the degree distribution in this paper. While this is one of the most important network characteristics, there are other characteristics such as clustering, assortativity, and average distance. Extending the model to include these additional characteristics and developing appropriate estimators are important directions for future research.

Appendix A: Degree Distribution of CF model

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

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

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

To present the main result of Cooper (2006), 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 the 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*
$$\mathbb{E}D_t(d,m) = p_m n_m(d;\eta,\kappa)t\left(1 + O\left(\frac{1}{\ln t}\right)\right),$$
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. Follows from Cooper (2006, Theorem 2.1). ■

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. Denote $P(d;\eta,\overline{M},\boldsymbol{p}) = \sum_{m=0}^{\min\{P,d\}} p_m n_m(d;\eta,\kappa)$.

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

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

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 the expressions from part 1 of Lemma A.1 gives part 1 of Proposition A.1. The following sequence of inequalities establishes part 2:

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

■

Appendix B: Main Proofs

Proof of Proposition 1. The proof follows from Proposition A.1 and the following derivations for large t:

$$\begin{aligned}
&\Pr\left(|D_t(d) - \mathbb{E}D_t(d)| \geq \tfrac{\mathbb{E}D_t(d)}{\sqrt{\ln t}}\right) \\
&= \Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\tfrac{t}{|V(t)|}\left(1 + O\left(\tfrac{1}{\ln t}\right)\right)\right| \geq \tfrac{P(d;\eta,\overline{M},\boldsymbol{p})}{\sqrt{\ln t}}\tfrac{t}{|V(t)|}\left(1 + O\left(\tfrac{1}{\ln t}\right)\right)\right) \\
&= \Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\right| \geq P(d;\eta,\nu,\boldsymbol{p})\left(\tfrac{1}{\sqrt{\ln t}} + O\left(\tfrac{1}{\ln t}\right)\right)\right) \\
&\geq \Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\eta,\overline{M},\boldsymbol{p})\right| \geq 2\tfrac{P(d;\eta,\overline{M},\boldsymbol{p})}{\sqrt{\ln t}}\right).
\end{aligned}$$

■

Proof of Corollary 1.

Part 1. Note that $d_t^*(\eta) \to \infty$ as $t \to \infty$; so $d \leq d_t^*(\eta)$ and Proposition 1 applies.

Part 2. Apply the well-known result (see, e.g., Tricomi and Erdélyi, 1951) that $\Gamma(z+\alpha)/\Gamma(z+\beta) = z^{\alpha-\beta}(1 + O(1/z))$ to $\Gamma(d+\kappa)/\Gamma(d+\kappa+1+1/\eta)$.

Part 3. Apply $\Gamma(z+1) = z\Gamma(z)$ to $n_m(d;\eta,\kappa)$ given by (1):

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

■

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

$$\tfrac{\Gamma(z+\alpha)}{\Gamma(z+1)} > \ (z+1)^{\alpha-1} \quad \text{for } \alpha > 2 \text{ and } z \geq 0.$$

Thus, for $d \geq \max\{1-\kappa_0, P\}$,

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

Thus,

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

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

Part 2. Let $d_t^\diamond = [\sqrt{\ln t}]$. Clearly, $d_t^\diamond \to \infty$, $d_t^\diamond / d_t^*(\eta) \to 0$, and $d_t^\diamond / \ln t \to 0$ as $t \to \infty$. We can write

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

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 there exists $N(\boldsymbol{\theta}_0)$ such that for $0 \leq d \leq d_t^*(\eta_0)$, we have:

$$\Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right| \geq 2\tfrac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right) \leq \tfrac{N(\boldsymbol{\theta}_0)}{\ln t}.$$

Therefore, by definition of $d_t^\diamond$, we have

$$\Pr\left(\exists d \leq d_t^\diamond : \left|\tfrac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right| \geq 2\tfrac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right) \leq \tfrac{N(\boldsymbol{\theta}_0)d_t^\diamond}{\ln t} = O\left(\tfrac{1}{\sqrt{\ln t}}\right).$$

Thus, with probability $1 - O(1/\sqrt{\ln t})$, which approaches one, we have

$$\widehat{S}_2(\boldsymbol{\theta}) \leq \left\|2\textstyle\sum_{d=0}^{d_t^\diamond-1} a(d;\boldsymbol{\theta})\frac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right\| < C_1\tfrac{\|G_0(\boldsymbol{\theta})\|}{\sqrt{\ln t}} \tag{2}$$

for some C_1. The last inequality follows from the uniform convergence of $G_0^n(\boldsymbol{\theta})$ 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 $\|a(d;\boldsymbol{\theta})\| < Fd$, showing $\sum_{d=d_t^\diamond}^{\infty} dD_t(d)/|V(t)| \xrightarrow{P} 0$ 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

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

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

$$\textstyle\sum_{d=0}^{\infty} dP(d;\boldsymbol{\theta}) = \sum_{m=0}^{P} p_m \sum_{d=m}^{\infty} dn_m(d;\eta,\kappa) = 2\left(\overline{m}+\overline{M}\right). \tag{3}$$

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

$$\textstyle\sum_{d=0}^{\infty} d\frac{D_t(d)}{|V(t)|} \xrightarrow{P} 2(\overline{m}_0+\overline{M}_0) = \sum_{d=0}^{\infty} dP(d;\boldsymbol{\theta}_0), \tag{4}$$

where the equality follows from (3). Using (2) and part 1 of this proposition, we get

$$\textstyle\sum_{d=0}^{d_t^\diamond-1} d\frac{D_t(d)}{|V(t)|} = \left(\sum_{d=0}^{d_t^\diamond-1} dP(d;\boldsymbol{\theta}_0)\right)\left(1+O_P\left(\frac{1}{\sqrt{\ln t}}\right)\right) \xrightarrow{P} \sum_{d=0}^{\infty} dP(d;\boldsymbol{\theta}_0). \tag{5}$$

Combining (4) and (5) gives

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

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

Proof of Proposition 3. Let $d_t^\diamond = [(\ln t)^{(1+\eta_0)/2}]$. Clearly, $d_t^\diamond/d_t^*(\eta_0) \to 0$ and $d_t^\diamond/\ln t \to 0$ as $t\to\infty$. Next, we show that $d_t^\diamond/[x\overline{F}^{-1}(k_t/t)] \to \infty$ as $t\to\infty$. To this end, since $P(d;\boldsymbol{\theta}_0)$ is decreasing in d, we can write the following summation-integration inequalities

$$\int_{z+1}^{\infty} P(s;\boldsymbol{\theta}_0)ds \le \overline{F}(z) \le \int_{z}^{\infty} P(s;\boldsymbol{\theta}_0)ds.$$

Combining these inequalities with part 2 of Corollary 1, we obtain

$$\overline{F}(z) = \eta_0 C(\boldsymbol{\theta}_0) z^{-1/\eta_0}\left(1+O\left(\frac{1}{z}\right)\right). \tag{6}$$

Inverting $\overline{F}(z)$ yields

$$\overline{F}^{-1}(y) = \left(\frac{y}{\eta_0 C(\boldsymbol{\theta}_0)}\right)^{-\eta_0}\left(1+O\left(y^{\eta_0}\right)\right). \tag{7}$$

Therefore,

$$\lim_{t\to\infty}\frac{d_t^\diamond}{[x\overline{F}^{-1}(k_t/t)]} = \lim_{t\to\infty}\frac{(\ln t)^{(1+\eta_0)/2}}{x(\eta_0 C(\boldsymbol{\theta}_0)t/k_t)^{\eta_0}} = \frac{1}{x(\eta_0 C(\boldsymbol{\theta}_0))^{\eta_0}}\lim_{t\to\infty}\frac{\sqrt{\ln t}}{(k_t\sqrt{\ln t}/t)^{-\eta_0}} = \infty,$$

where the first equality holds by (7) and definition of $d_t^\diamond$, the second by rearrangement, and the last by requirements $\eta_0\in(0,1)$ and $k_t\sqrt{\ln t}/t\to\infty$.

We can write

$$\textstyle\frac{1}{k_t}\sum\limits_{d=[x\overline{F}^{-1}(k_t/t)]}^{\infty} D_t(d) = \underbrace{\frac{|V(t)|}{k_t}\sum\limits_{d=[x\overline{F}^{-1}(k_t/t)]}^{d_t^\diamond-1}\frac{D_t(d)}{|V(t)|}}_{\widehat{H}_1} + \underbrace{\frac{|V(t)|}{k_t}\sum\limits_{d=d_t^\diamond}^{\infty}\frac{D_t(d)}{|V(t)|}}_{\widehat{H}_2}.$$

To prove the proposition, it suffices to show that $\widehat{H}_1 \xrightarrow{P} x^{-1/\eta_0}$ and $\widehat{H}_2 \xrightarrow{P} 0$.

Proposition 1 implies that there exists $N(\boldsymbol{\theta}_0)$ such that for $0 \leq d \leq d_t^*(\eta_0)$, we have:

$$\Pr\left(\left|\tfrac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right| \geq 2\tfrac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right) \leq \tfrac{N(\boldsymbol{\theta}_0)}{\ln t}.$$

Therefore, by definition of $d_t^\diamond$, we have

$$\Pr\left(\exists d \leq d_t^\diamond : \left|\tfrac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right| \geq 2\tfrac{P(d;\boldsymbol{\theta}_0)}{\sqrt{\ln t}}\right) \leq \tfrac{N(\boldsymbol{\theta}_0)d_t^\diamond}{\ln t} = O\left((\ln t)^{-\frac{1-\eta_0}{2}}\right).$$

Thus, with probability $1 - O\left((\ln t)^{-(1-\eta_0)/2}\right)$, which approaches one, we have

$$\tfrac{|V(t)|}{k_t} \sum_{d=[x\overline{F}^{-1}(k/t)]}^{d_t^\diamond - 1} \left(\tfrac{D_t(d)}{|V(t)|} - P(d;\boldsymbol{\theta}_0)\right) \leq 2\tfrac{|V(t)|}{k_t}\tfrac{\overline{F}([x\overline{F}^{-1}(k_t/t)])}{\sqrt{\ln t}} = 2\tfrac{x^{-1/\eta_0}}{\sqrt{\ln t}}\left(1 + O\left(\left(\tfrac{k_t}{t}\right)^{\eta_0}\right)\right), \quad (8)$$

where the equality follows from (6) and (7). Notice that

$$\tfrac{|V(t)|}{k_t} \sum_{d=[x\overline{F}^{-1}(k_t/t)]}^{d_t^\diamond - 1} P(d;\boldsymbol{\theta}_0) = \tfrac{|V(t)|}{k_t}\overline{F}\left([x\overline{F}^{-1}(k_t/t)]\right) - \tfrac{|V(t)|}{k_t}\overline{F}(d_t^\diamond) \to x^{-1/\eta_0}, \quad (9)$$

where the first term converges to x^{-1/η_0} by (6) and (7) and the second term to 0 by

$$\frac{t}{k_t}\left((\ln t)^{(1+\eta_0)/2}\right)^{-1/\eta_0} = \frac{(\ln t)^{-1/2\eta_0}}{k_t\sqrt{\ln t}/t} \to 0. \quad (10)$$

Combining (8) and (9) gives $\widehat{H}_1 \xrightarrow{P} x^{-1/\eta_0}$.

With probability $1 - O\left((\ln t)^{-(1-\eta_0)/2}\right)$, which approaches one, we have

$$\widehat{H}_2 = \frac{|V(t)|}{k_t}\left(1 - \sum_{d=0}^{d_t^\diamond - 1}\frac{D_t(d)}{|V(t)|}\right) \leq \frac{|V(t)|}{k_t}\overline{F}(d_t^\diamond) + \frac{|V(t)|}{k_t}\frac{2}{\sqrt{\ln t}} \to 0,$$

where the first term converges to 0 by (10) and the second term to 0 by $k_t\sqrt{\ln t}/t \to \infty$. ■

Proof of Proposition 4. 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

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

Thus, for $d \geq P$,

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

where $\underline{C} = \min_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}C(\boldsymbol{\theta}) > 0$. Thus, there is F such that $|\ln P(d;\boldsymbol{\theta})| < Fd$ for all $d \in \mathbb{N}_+$ and all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, so Proposition 2 applies, meaning that $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}}\left|\widehat{L}_t(\boldsymbol{\theta}) - L_0(\boldsymbol{\theta})\right| \xrightarrow{P} 0$, where $L_0(\boldsymbol{\theta}) = \sum_{d=0}^{\infty}\ln P(d;\boldsymbol{\theta})P(d;\boldsymbol{\theta}_0)$ 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:

$$L_0(\boldsymbol{\theta}_0) - L_0(\boldsymbol{\theta}) = -\sum_{d=0}^{\infty} \ln \tfrac{P(d;\boldsymbol{\theta})}{P(d;\boldsymbol{\theta}_0)} P(d;\boldsymbol{\theta}_0) > -\ln\left(\sum_{d=0}^{\infty} \tfrac{P(d;\boldsymbol{\theta})}{P(d;\boldsymbol{\theta}_0)} P(d;\boldsymbol{\theta}_0)\right) = 0. \tag{11}$$

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 Newey and McFadden (1994, Theorem 2.1) 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 $\widehat{\boldsymbol{\theta}}_{\text{P}}^{\text{PML}}$, 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 η 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 $\widehat{L}_t(\widehat{\boldsymbol{\theta}}_{\text{P}}^{\text{PML}}) \geq \widehat{L}_t(\widehat{\eta}^{\text{PML}}, \widehat{\overline{M}}, \widehat{\boldsymbol{p}}^{\text{PML}}) = \widehat{L}_t(\widehat{\boldsymbol{\theta}}^{\text{PML}}) + o_P(1)$, which implies that $\widehat{\boldsymbol{\theta}}_{\text{P}}^{\text{PML}} \xrightarrow{P} \boldsymbol{\theta}_0$. ■

Proof of Proposition 5.

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

Part 2. The moment function has zero mean: $\sum_{d=0}^{\infty} \left(d - 2(\overline{m}_0 + \overline{M}_0)\right) P(d;\boldsymbol{\theta}_0) = 0$ by (3) and $\sum_{d=0}^{\infty} \left(\nabla_{\boldsymbol{\theta}} \ln P(d;\boldsymbol{\theta}_0)\right) P(d;\boldsymbol{\theta}_0) = 0$ by (11) and the interchangeability of summation and differentiation, which follows from Rudin (1976, Theorems 7.10 and 7.17). 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 for $d \geq P$

$$\ln P(d;\boldsymbol{\theta}) = \ln\Gamma(d+\kappa) - \ln\Gamma(d+\kappa+1/\eta+1) + \underbrace{\ln\left(\sum_{m=0}^{P} p_m \frac{\Gamma(m+\kappa+1/\eta)}{\eta\Gamma(m+\kappa)}\right)}_{R(\boldsymbol{\theta})}, \tag{12}$$

where $R(\boldsymbol{\theta})$ collects all terms independent of d. Let $R_x(\boldsymbol{\theta})$ denote a partial derivative of $R(\boldsymbol{\theta})$ with respect to x.

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

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

where $\psi(z) = d\ln\Gamma(z)/dz$ is a polygamma function. Qi et al. (2005) show that for $x > 0$:

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

which implies that there is F such that $\|g(d;\boldsymbol{\theta})\| < Fd$ for all $d \in \mathbb{N}_+$ and all $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$; so condition (iv) of part 1 holds, and thus $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$. ■

Proof of Proposition 6. By Proposition 3, condition (2.1) of Resnick and Stărică (1995) holds and their Proposition 2.4 implies that $\widehat{\eta}^{\text{Hill}} \xrightarrow{P} \eta_0$. ■

Proof of Proposition 7. To prove Proposition 7, we notice that all conditions, except for condition (iv), of Newey and McFadden (1994, Theorem 3.2) are satisfied by assumption. Thus, we only need to check condition (iv) that for compact set $\overline{\boldsymbol{\Theta}}$ such that $\boldsymbol{\theta}_0 \in \overline{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}$, we have $\sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\Theta}}} \|\sum_{d=0}^{\infty} \nabla_{\boldsymbol{\theta}} g(d;\boldsymbol{\theta}_0) D_t(d)/|V(t)| - G(\boldsymbol{\theta})\| \xrightarrow{P} 0$.

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

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

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

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

where $R_{xy}(\boldsymbol{\theta})$ is a second-order partial derivative of $R(\boldsymbol{\theta})$ given in (12) with respect to x and y, and $\psi^{(1)}(z) = d^2 \ln\Gamma(z)/dz^2$ is the Polygamma function of order 1.

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 $\|G(d;\boldsymbol{\theta})\| < Fd$ for all $d \in \mathbb{N}_+$ 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 Rudin (1976, Theorems 7.10 and 7.17). ■

References

Acemoglu, Daron, Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi (2012) "The Network Origins of Aggregate Fluctuations," *Econometrica*, **80** (5), pp. 1977–2016.

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

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.

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, Petra Vynckier, and Jozef L Teugels (1996) "Tail Index Estimation, Pareto Quantile Plots Regression Diagnostics," *Journal of the American Statistical Association*, **91** (436), pp. 1659–1667.

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.

Chandrasekhar, Arun (2016) "Econometrics of Network Formation," in Yann Bramoullé, Andrea Galeotti, and Brian W Rogers eds. *Oxford Handbook of the Economics of Networks*, pp. 303–357.

Chandrasekhar, Arun G and Matthew O Jackson (2016) "A Network Formation Model Based on Subgraphs," working paper, Stanford University.

Chandrasekhar, Arun and Randall Lewis (2016) "Econometrics of Sampled Networks," working paper, Stanford University.

Chaney, Thomas (2014) "The Network Structure of International Trade," *American Economic Review*, **104** (11), pp. 3600–3634.

Christakis, Nicholas A., James H. Fowler, Guido W. Imbens, and Karthik Kalyanaraman (2010) "An Empirical Model for Strategic Network Formation," working paper, Stanford University.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman (2009) "Power-Law Distributions in Empirical Data," *SIAM Review*, **51** (4), pp. 661–703.

Comola, Margherita and Marcel Fafchamps (2014) "Testing Unilateral and Bilateral Link Formation," *Economic Journal*, **124** (579), pp. 954–976.

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.

de Paula, Áureo (2017) "Econometrics of Network Models," in Bo Honoré, Ariel Pakes, Monika Piazzesi, and Larry Samuelson eds. *Advances in Economics and Econometrics: Volume 1: Eleventh World Congress*, pp. 268–323.

de Paula, Áureo, Seth Richards-Shubik, and Elie Tamer (2018) "Identifying Preferences in Networks with Bounded Degree," *Econometrica*, **86** (1), pp. 263–288.

Dekkers, Arnold LM, John HJ Einmahl, and Laurens De Haan (1989) "A Moment Estimator for the Index of an Extreme-Value Distribution," *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 (2011) "The Granular Origins of Aggregate Fluctuations," *Econometrica*, **79** (3), pp. 733–772.

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.

Goldsmith-Pinkham, Paul and Guido W Imbens (2013) "Social Networks and the Identification of Peer Effects," *Journal of Business & Economic Statistics*, **31** (3), pp. 253–264.

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.

Graham, Bryan S (2017) "An Econometric Model of Network Formation with Degree Heterogeneity," *Econometrica*, **85** (4), pp. 1033–1063.

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

Horowitz, Joel L. (2001) "The Bootstrap," in James J. Heckman and Edward Leamer eds. *Handbook of Econometrics*, **5**: Elsevier, pp. 3159–3228.

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.

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 (2016) "The Formation of Networks with Local Spillovers and Limited Observability," *Theoretical Economics*, **11** (3), pp. 813–863.

König, Michael D, Claudio J Tessone, and Yves Zenou (2014) "Nestedness in Networks: A Theoretical Model and Some Applications," *Theoretical Economics*, **9** (3), pp. 695–752.

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.

Luttmer, Erzo GJ (2011) "On the Mechanics of Firm Growth," *Review of Economic Studies*, **78** (3), pp. 1042–1068.

Maasoumi, Esfandiar (1993) "A Compendium to Information Theory in Economics and Econometrics," *Econometric Reviews*, **12** (2), pp. 137–181.

Matthys, Gunther and Jan Beirlant (2000) "Adaptive Threshold Selection in Tail Index Estimation," in Paul Embrechts ed. *Extremes and Integrated Risk Management*, pp. 37–49.

Mayer, Adalbert and Steven L Puller (2008) "The Old Boy (and Girl) Network: Social Network Formation on University Campuses," *Journal of Public Economics*, **92** (1-2), pp. 329–347.

Mele, Angelo (2017) "A Structural Model of Dense Network Formation," *Econometrica*, **85** (3), pp. 825–850.

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.

Newman, Mark (2001) "The Structure of Scientific Collaboration Networks," *Proceedings of the National Academy of Sciences*, **98** (2), pp. 404–409.

——— (2010) *Networks: An Introduction*: Oxford University Press.

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," *Proceedings of the National Academy of Sciences*, **99** (8), pp. 5207–5211.

Pickands, III, James (1975) "Statistical Inference Using Extreme Order Statistics," *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.

Resnick, Sidney and Cătălin Stărică (1995) "Consistency of Hill's Estimator for Dependent Data," *Journal of Applied Probability*, **32** (1), pp. 139–167.

Rothenberg, Thomas J (1971) "Identification in Parametric Models," *Econometrica*, **39** (3), pp. 577–591.

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

Sheng, Shuyang (2016) "A Structural Econometric Analysis of Network Formation Games," working paper, UCLA.

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

Tricomi, F and Arthur Erdélyi (1951) "The Asymptotic Expansion of a Ratio of Gamma Functions," *Pacific Journal of Mathematics*, **1** (1), pp. 133–142.

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

Supplementary Appendix for Estimation of a Scale-Free Network Formation Model

Anton Kolotilin and Valentyn Panchenko

June, 2018

1 NLS Estimator

In Section 2, we derived the asymptotic degree distribution $P(d)$. Below, we approximate the asymptotic degree distribution of the CF model using an alternative method: mean-field approximation. The NLS estimator discussed below is based on the mean-field approximation of the asymptotic degree distribution.

1.1 Mean-Field Approximation of the Degree Distribution

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

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

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

As $t\to\infty$, the differential equation asymptotes to

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

where $\eta\kappa$ is the expected number of edge endpoints added uniformly at random per vertex. The solution to this differential equation is:

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

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, vertices added at an earlier time period ("older" vertices) have a larger expected degree than vertices added at later periods ("younger" vertices). Thus, the cumulative distribution of expected degrees of vertices with the initial degree m can be approximated by (for $d \geq m$):

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

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

$$F^{\mathrm{MF}}(d) = \sum_{m=0}^{\min\{P,d\}} p_m F_t^m(d). \tag{1}$$

For a sufficiently large d and a constant K, the complementary cumulative distribution can be approximated by a power-law distribution:

$$1 - F^{\mathrm{tail}}(d) = C d^{-1/\eta}.$$

This result is analogous to part 2 of Corollary 1 in Section 2.2, which shows that the asymptotic degree distribution $P(d)$ has a power-law tail.

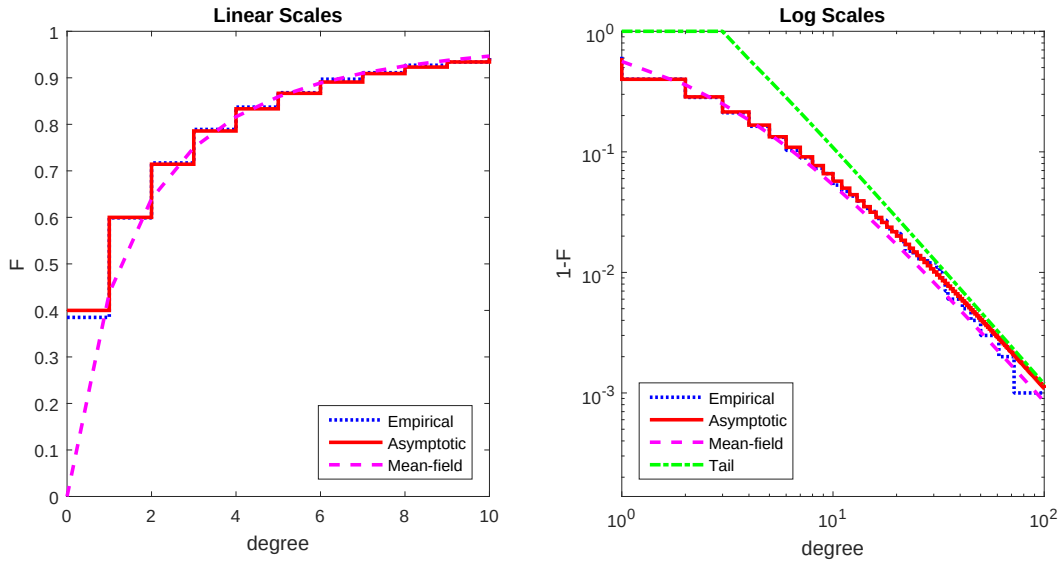
Now we can connect various approximations of the degree distribution to specific estimators: the PML and GMM estimators are derived from the asymptotic degree distribution $P(d)$, the NLS estimator is derived from the mean-field approximation, and the Hill estimator and other tail estimators are based on the power-law approximation in the tail.

Figure 1 compares the cumulative distributions (left panel, linear scale) and the complementary cumulative distributions (right panel, log scale) for the benchmark specification: $t = 1000$, $p_0 = 1$ ($m(t) = 0$), $q_1 = q_2 = 0.5$ ($\overline{M} = 1.5$), and $A_1 = B_1 = C_1 = 0.5$ ($\eta = 0.5$). The empirical cumulative distribution from a simulated network is much closer to the asymptotic approximation than to the mean-field and tail approximations, especially for small degrees d. This may explain inferior performance of the estimators based on the latter two approximations.

1.2 NLS Estimator

We now turn to the NLS estimator commonly used for scale-free network formation models (see Pennock et al., 2002; Jackson and Rogers, 2007; Jackson, 2008).

Figure 1: Degree distributions for a simulation of the CF model



In order to derive the NLS estimator, we need to fix m, that is, assume that $m(t) = m$.[1] Since most of the real networks have vertices with zero degree, we set $m(t) = 0$. Under this assumption, (1) can be expressed as

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

Moreover, $\overline{M}$ can be consistently estimated as[2]

$$\widehat{\overline{M}} = \frac{1}{2} \sum_{d=0}^{\infty} d \frac{D_t(d)}{t}.$$

Parameter η is then estimated by numerically minimizing the quadratic loss:

$$\widehat{\eta}^{\mathrm{NLS}} = \arg\min_{\eta} \sum_{d=0}^{\infty} \left(\ln(1 - \widehat{F}_t(d)) - 1/\eta \left(\ln(2\widehat{\overline{M}}(1/\eta - 1)) - \ln(d + 2\widehat{\overline{M}}(1/\eta - 1)) \right) \right)^2,$$

where $\widehat{F}_t(d)$ is an empirical analogue of the cumulative distribution.

There are several alternatives of what can be used as degree observations for the NLS estimator: (i) *observed distinct* degrees (without repetition), (ii) *consecutive degrees in the range* $[d_{\min}, d_{\max}]$ where $d_{\min} = \min_v d(v,t)$ and $d_{\max} = \max_v d(v,t)$,[3] and (iii) *observed degrees with repetition* (Newman, 2005, Appendix A).

[1] Note that Pennock et al. (2002) assume $m(t) = 0$ and $M(t) = M$ for some constant M, whereas Jackson (2008) and Jackson and Rogers (2007) assume $m(t) = m$ and $M(t) = 0$ for some constant m.

[2] Hereafter, we assume $|V(t)| = t$ for notational simplicity.

[3] Our reconstruction suggests that this method in conjunction with the empirical cumulative distribution for $\widehat{F}_t(d)$ and removed $d_{\max}$ is used in Jackson (2008).

$\widehat{F}_t(d)$ can also be specified in several ways. The empirical cumulative distribution is a common candidate, $\widehat{F}_t(d) = \sum_{i=0}^{d} D_t(i)/t$. But, in this case, $\widehat{F}_t(d_{\max}) = 1$, and hence, $\ln(1-\widehat{F}_t(d_{\max}))$ is not defined. One way to overcome this issue is to remove the observation(s) with $d = d_{\max}$. Alternatively, $\widehat{F}_t(d)$ can be scaled with $1/(t+1)$ instead of $1/t$ (see, e.g., Beirlant et al., 2006, p. 5). Moreover, when the observed degrees with repetition are used, $1-\widehat{F}_t$ can be specified as ordinal ranks (scaled by $1/t$) (see, Newman, 2005, Appendix A). In this case, each observation (vertex) is assigned a distinct ordinal number from 1 to t according to its degree d in descending order. Hence, $\widehat{F}_t$ changes in discrete steps of $1/t$. Gabaix and Ibragimov (2011) advocate using rank $-1/2$ adjustment for improved performance.

The NLS estimators based on the above definitions and adjustments are compared in the Excel spreadsheet of this Supplement. It appears that using (i) observed distinct degrees and (ii) observed degrees with repetition in conjunction with removed $d_{\max}$ yield the best performance. These estimators are reported in the main text of the paper as (i) NLS_{D} and (ii) NLS_{R}.

2 Tail Estimators

There is a well-developed literature on tail estimators, starting from Hill (1975), Pickands (1975), and Smith (1987); see Beirlant et al. (2006) for a detailed analysis and references. These estimators rely on a specific behavior in the tail of the distribution. Since the CF model yields a degree distribution with a power-law tail, tail estimators based on Pareto-type models are appropriate for estimating parameter η, which determines power-law parameter $1+1/\eta$. Most tail estimators are designed for continuous independently identically distributed random variables, but degrees in the CF model are discrete valued, interdependent, and not identically distributed. Moreover, an appropriate choice of the number of observations in the tail, called a tail cutoff, $d_t^\dagger$, after which the tail approximation holds, is crucial for these estimators. We will first assume that $d_t^\dagger$ is known and then, after introducing the estimators, we will discuss various methods for selecting $d_t^\dagger$.

2.1 Pareto-Type Distribution

A simple and popular way to estimate the power-law parameter is to run a rank-degree regression in logs. Specifically, denote increasingly ordered degree observations by $d_1 \leq \cdots \leq d_t$. The regression is $\ln j = c - \frac{1}{\eta} \ln d_{t-j+1}$ for j such that $d_{t-j+1} > d_t^\dagger$. Gabaix and Ibragimov (2011) propose an important simple bias-reducing adjustment. They recommend

using $\ln(j-1/2)$ instead of $\ln j$ in the regression. We implement this estimator and refer to it as the GI estimator.

Hill (1975) estimator is the main tail estimator. It can be derived as a maximum likelihood estimator based on the two assumptions: (i) the tail of the distribution follows continuous Pareto distribution with density $f(d) = \frac{1}{\eta d_t^\dagger}\left(\frac{d}{d_t^\dagger}\right)^{-1-1/\eta}$ conditional on $d > d_t^\dagger$, and (ii) these tail observations are independent,[4]

$$\widehat{\eta}^{\text{Hill}} = \frac{1}{k_t}\sum_{v:d(v,t)>d_t^\dagger} \ln\frac{d(v,t)}{d_t^\dagger} = \frac{1}{k_t}\sum_{d=d_t^\dagger+1}^{\infty} D_t(d)\ln\frac{d}{d_t^\dagger},$$

where $k_t = \sum_{d=d_t^\dagger+1}^{\infty} D_t(d)$ is the number of vertices that have degree greater than $d_t^\dagger$.

For a discrete distribution, Clauset et al. (2009) propose a simple adjustment for the Hill estimator,

$$\widehat{\eta}_{\text{C}}^{\text{Hill}} = \frac{1}{k_t}\sum_{d=d_t^\dagger+1}^{\infty} D_t(d)\ln\frac{d}{d_t^\dagger+1/2}.$$

The discrete counterpart of the Pareto distribution is zeta distribution. Assuming the zeta distribution for the tail, the probability that a vertex has degree d, for $d > d_t^\dagger$, is

$$P(d) = \frac{d^{-1-1/\eta}}{\zeta(1+1/\eta, d_t^\dagger+1)},$$

where $\zeta(1+1/\eta, d_t^\dagger+1) = \sum_{i=0}^{\infty}(i+d_t^\dagger+1)^{-(1+1/\eta)}$ is the Hurwitz zeta function. Goldstein et al. (2004) and Bauke (2007) use a maximum likelihood tail estimator for discrete data,

$$\widehat{\eta}_{\text{G}}^{\text{Hill}} = \underset{\eta}{\text{argmax}} -\frac{1+1/\eta}{k_t}\sum_{d=d_t^\dagger+1}^{\infty} D_t(d)\ln d - \ln\zeta(1+1/\eta, d_t^\dagger+1).$$

2.2 Other Distributions

The Hill estimator, together with its variants discussed above, is applicable for estimating the tail of Pareto-type distributions, and thus of the degree distribution of the CF model. We now introduce other tail estimators applicable for estimating the tails of distributions belonging to the Pareto, Weibull, and Gumbel classes.

Pickands (1975) proposes a tail estimator which is based on sample quantiles in the tails,

$$\widehat{\eta}^{\text{Pic}} = \frac{1}{\ln 2}\ln\frac{d_{t-[k_t/4]} - d_{t-[k_t/2]}}{d_{t-[k_t/2]} - d_{t-k_t}}.$$

[4]Because of degree interdependences in the CF model, this and related estimators, which ignore the interdependences, should formally be referred to as pseudo maximum likelihood estimators. In the main text of the paper, we prove consistency of the Hill estimator.

Dekkers et al. (1989) propose an estimator based on higher moments of the Hill estimator,

$$\widehat{\eta}^{\text{Dek}} = \widehat{\eta}^{\text{Hill}} + 1 - \frac{1}{2}\left(1 - \frac{\left(\widehat{\eta}^{\text{Hill}}\right)^2}{\frac{1}{k_t}\sum_{d=d_t^\dagger+1}^{\infty} D_t(d)\left(\ln\frac{d}{d_t^\dagger}\right)^2}\right)^{-1}.$$

Smith (1987) proposes a maximum likelihood tail estimator assuming generalized Pareto distribution with density $f(d) = \frac{1}{\sigma}\left(1 + \frac{\eta(d-d_t^\dagger)}{\sigma}\right)^{-1/\eta-1}$ conditional on $d > d_t^\dagger$,

$$\widehat{\eta}^{\text{Smith}} = \underset{\eta,\sigma}{\text{argmax}} -\frac{1+1/\eta}{k_t}\sum_{d=d_t^\dagger+1}^{\infty} D_t(d)\ln\left(1 + \frac{\eta(d-d_t^\dagger)}{\sigma}\right) - \ln\sigma.$$

2.3 Selection of Tail Cutoff

Up to this point, we have treated $d_t^\dagger$ as given. Next, we discuss several methods for selecting $d_t^\dagger$ as this is a crucial step for any tail estimator. For the Hill estimator, multiple methods are proposed in the literature (Beirlant et al., 2006, Chapter 4.7). We use a popular analytical method, which we refer to as MS, aiming to balance the asymptotic bias and variance by selecting $d_t^\dagger$ such that it minimizes the asymptotic mean squared error (AMSE) of the estimator (Beirlant et al., 1996; Matthys and Beirlant, 2000), given by

$$\text{AMSE}\left(\widehat{\eta}_{d_t^\dagger}^{\text{Hill}}\right) = \text{ABias}^2\left(\widehat{\eta}_{d_t^\dagger}^{\text{Hill}}\right) + \text{AVar}\left(\widehat{\eta}_{d_t^\dagger}^{\text{Hill}}\right),$$

where $\text{AVar}\left(\widehat{\eta}_{d_t^\dagger}^{\text{Hill}}\right) = \eta^2/k_t$. Lower $d_t^\dagger$ yields higher k_t which, in turn, reduces the variance, but increases the bias. Estimating $\text{ABias}\left(\widehat{\eta}_{d_t^\dagger}^{\text{Hill}}\right)$ relies on the use of scaled log-spacing representation of the Hill estimator as in Beirlant et al. (2002). Define scaled log-spacing as $Z_j = j(\log d_{t-j+1} - \log d_{t-j})$, where $j = 1, \ldots, k_t$. The asymptotic bias can be estimated as[5]

$$\widehat{\text{ABias}}\left(\widehat{\eta}_{d_t^\dagger}^{\text{Hill}}\right) = \frac{6}{k_t}\sum_{j=1}^{k_t}\left(\frac{j}{k_t+1} - \frac{1}{2}\right)Z_j.$$

For the other tail estimators, there are no equivalent methods, but for comparison we apply $d_t^\dagger$ selected by this method to other tail estimators as well.

Clauset et al. (2009) proposes a universal method for any tail estimator: to choose $d_t^\dagger$ so that the distance, D, between the theoretical cumulative distribution of the underlying power-law, $F_{\widehat{\eta}}(d) = 1 - (d/d_t^\dagger)^{-1/\widehat{\eta}}$, with estimated $\widehat{\eta}$ and the empirical cumulative distribution,

[5]For more details on deriving this expression see Chapter 4.5.1 of Beirlant et al. (2006). There $\text{ABias}\left(\widehat{\eta}_{d_t^\dagger}^{\text{Hill}}\right) = \frac{b}{1+\beta}$ and the least-squares estimator for b is given on p. 117. Difficulties of estimating β are discussed on the same page and based on this discussion we set $\beta = 1$.

$\widehat{F}_{k_t}(d) = \sum_{i=d_t^\dagger+1}^{d} D_t(i)/k_t$, is minimized for all $d > d_t^\dagger$. The authors suggest using the Kolmogorov-Smirnov (KS) distance,[6] so that the distance is given by

$$D_{\text{KS}} = \max_{d>d_t^\dagger} \left| F_{\widehat{\eta}}(d) - \widehat{F}_{k_t}(d) \right|.$$

2.4 Simulations

Our simulations (see Supplement, Excel spreadsheet) show that the performance in terms of the mean square error of various tail estimators is substantially better when the tail cutoff is selected using the MS method rather than the KS method.

Comparing the performance of all tail estimators, we find that the Smith estimator outperforms all tail estimators. The Hill estimator is among the best performing tail estimators; it shows a substantial bias for $t = 1000$, which reduces with the number of observations. The continuity correction suggested by Clauset et al. (2009) slightly helps in reducing the bias. The NLS estimators perform better than the tail estimators only in small samples with at most 1000 observations. The introduced PML and GMM estimators outperform both the NLS and tail estimators.

References

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

Bauke, Heiko (2007) "Parameter Estimation for Power-law Distributions by Maximum Likelihood Methods," *The European Physical Journal B*, **58** (2), pp. 167–173.

Beirlant, Jan, Petra Vynckier, and Jozef L Teugels (1996) "Tail Index Estimation, Pareto Quantile Plots Regression Diagnostics," *Journal of the American Statistical Association*, **91** (436), pp. 1659–1667.

Beirlant, J, G Dierckx, A Guillou, and C Staăricaă (2002) "On Exponential Representations of Log-Spacings of Extreme Order statistics," *Extremes*, **5** (2), pp. 157–180.

[6]The first application of the KS distance in this context dates back to Pickands (1975). Clauset et al. (2009) discussed other distances, such as the Cramer-von-Misses (CM) and Anderson-Darling (AD) distance and modified KS distance to penalize tails. We tried these distances, but the results did not change substantially in comparison to the KS distance.

Beirlant, Jan, Yuri Goegebeur, Johan Segers, and Jozef Teugels (2006) *Statistics of Extremes: Theory and Applications*: John Wiley & Sons.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman (2009) "Power-Law Distributions in Empirical Data," *SIAM Review*, **51** (4), pp. 661–703.

Dekkers, Arnold LM, John HJ Einmahl, and Laurens De Haan (1989) "A Moment Estimator for the Index of an Extreme-Value Distribution," *Annals of Statistics*, **17** (4), pp. 1833–1855.

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.

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.

Matthys, Gunther and Jan Beirlant (2000) "Adaptive Threshold Selection in Tail Index Estimation," in Paul Embrechts ed. *Extremes and Integrated Risk Management*, pp. 37–49.

Newman, Mark EJ (2005) "Power Laws, Pareto Distributions and Zipf's Law," *Contemporary Physics*, **46** (5), pp. 323–351.

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," *Proceedings of the National Academy of Sciences*, **99** (8), pp. 5207–5211.

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

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