Asymptotic degree distributions in large (homogeneous) networks: A little theory and a counterexample[a][b]

Armand M. Makowski

University of Maryland at College Park
(and National Science Foundation (CISE/CCF/CIF))

[a]Supported by NSF Grant CCF-1217997 and a Lady Davis Trust Fellowship.
[b]Joint work with Siddarth Pal (ECE, now with Raytheon BBN/ARL)

References

- S. Pal and A.M. Makowski, "On the asymptotics of degree distributions," in Proceedings of the 54th IEEE Conference on Decision and Control (CDC 2015), Osaka (Japan), December 2015.
- S. Pal and A.M. Makowski, "Asymptotic degree distributions in large (homogeneous) random networks – A little theory and a counterexample," *IEEE Transactions on Network Science and Engineering* (2015). Accepted for publication (2019).
- A.M. Makowski and S. Pal, "Asymptotic degree distributions in random threshold graphs," Submitted to *Applied Probability* (2019).

Good generative random graph models

In many surveys we are told: At the beginning there were Erdős-Rényi graphs . . .

In Erdős-Rényi graphs, the degree distribution is “Poisson”-like

---

Story line?

Erdős-Rényi graphs $\mathbb{G}(n;p)$:

$$D_{n,k}(p) =_{st} \text{Bin}(n-1;p), \qquad \begin{array}{c} k = 1, \ldots, n \\ n = 2, 3, \ldots \\ 0 \leq p \leq 1 \end{array}$$

Homogeneity – There is **a** (**common**) degree distribution:

$$D_n(p) =_{st} D_{n,k}(p), \quad k = 1, 2, \ldots, n$$

Theoretical justification via **Poisson convergence** – For some $\lambda > 0$,

$$D_n(p_n) \Longrightarrow_n \boldsymbol{p}_\lambda \quad \text{if } p_n \sim \frac{\lambda}{n}$$

where $\boldsymbol{p}_\lambda$ is the Poisson pmf on $\mathbb{N}$ with parameter λ.

We are also told

Erdős-Rényi graph models are deemed **inappropriate** for many "complex networks" because observations/measurements point to a degree distribution which is **not** Poisson-like.

(**Vague empirical statement**) For large networks (large n),

$$\frac{N_n(d)}{n} \simeq C d^{-\alpha} \tag{1}$$

for some α in the range $[2, 3]$ and $C > 0$ with

$$N_n(d) = \text{Number of nodes with degree } d \text{ in } \mathbb{G}_n$$

Power/Pareto-like degree distribution. **Scale-free** networks.

Preferential attachment models (Barabási and Albert (1999))

Generative model based on **growth** and **preferential attachment** – Markov sequence of random graphs

$$\mathbb{G}_n = (V_n, \mathbb{E}_n), \ n = 1, 2, \ldots$$

No more a **single** node degree distribution but instead an (empirical) **networkwide** degree distribution:

$$P_n(d) = \frac{N_n(d)}{n}, \qquad \begin{array}{l} n = 1, 2, \ldots \\ d = 0, 1, \ldots \end{array}$$

with

$$N_n(d) = \text{Number of nodes with degree } d \text{ in } \mathbb{G}_n$$

It can be shown that

$$P_n(d) \xrightarrow{P}_n p_{\mathrm{SF}}(d) \quad d = 0, 1, \ldots$$

with pmf $\boldsymbol{p}_{\mathrm{SF}}(d) = (p_{\mathrm{SF}}(d),\ d = 0, 1, \ldots)$ on $\mathbb{N}$ such that

$$p_{\mathrm{SF}}(d) \sim C d^{-3} \quad (d \to \infty)$$

Original model

Barabási and Albert (1999), Bollobás and Riordan (2000) – Many variations on this theme

Back to Erdős-Rényi graphs

Apples vs. oranges?

With scaling $p : \mathbb{N}_0 \to [0,1]$, **identically** distributed rvs

$$D_{n,k}(p_n), \quad k = 1, 2, \ldots, n$$

with a **common** distribution $\text{Bin}(n-1, p_n)$ vs. an **empirical** measure

$$P_n(d) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\left[D_{n,k}(p_n) = d\right], \quad d = 0, 1, \ldots$$

Is there a relationship (in the large n limit) between the pmf of the generic degree and the empirical degree pmf?

Resolving the difference

Lemma 1 *If* $p_n \sim \frac{\lambda}{n}$ *for some* $\lambda > 0$, *then for each* $d = 0, 1, \ldots$,

$$\frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\left[D_{n,k}(p_n) = d\right] \xrightarrow{P}_n p_\lambda(d)$$

where $\boldsymbol{p}_\lambda = (p_\lambda(d),\ d = 0, 1, \ldots)$ *is the Poisson pmf on* $\mathbb{N}$ *with parameter* λ *given by*

$$p_\lambda(d) = \frac{\lambda^d}{d!} e^{-\lambda}, \quad d = 0, 1, \ldots$$

Why should one care? To be able to establish a **meaningful** comparison between Erdős-Rényi graph models and preferential attachment models

---

A natural question: Can we generalize Lemma 1 to a larger class of random graphs models?

**A general framework
(and an easy result)**

Consider a sequence of random graphs $\{\mathbb{G}_n, \ n = 2, 3, \ldots\}$, in the **homogeneous** case, with

$$D_{n,1} =_{st} D_{n,2} =_{st} \ldots =_{st} D_{n,n} =_{st} D_n.$$

Q: If $D_n \Longrightarrow_n D$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left[D_n = d\right] = \mathbb{P}\left[D = d\right], \quad d = 0, 1, \ldots$$

then is it true that

$$\frac{N_n(d)}{n} \xrightarrow{P}_n \mathbb{P}\left[D = d\right], \quad d = 0, 1, \ldots$$

where

$$N_n(d) = \sum_{k=1}^{n} \mathbf{1}\left[D_{n,k} = d\right]?$$

A: Yes **provided**

- Homogeneity
- Existence of an asymptotic degree distribution
- Asymptotic uncorrelatedness

These conditions hold for many classes of **generative** random graph models

Assumption 1 *(Homogeneity) For each $n = 2, 3, \ldots$, the degree rvs in $\mathbb{G}_n$ are* **equidistributed** *in the sense that*

$$D_{n,k} =_{st} D_{n,1}, \quad k = 1, 2, \ldots, n$$

and

$$(D_{n,k}, D_{n,\ell}) =_{st} (D_{n,1}, D_{n,2}) \quad \begin{array}{c} k \neq \ell \\ k, \ell = 1, \ldots, n \end{array}$$

Assumption 2 *(Existence of an asymptotic degree distribution) Under Assumption 1, there exists an $\mathbb{N}$-valued rv D such that*

$$D_{n,1} \Longrightarrow_n D.$$

Let $\boldsymbol{p} = (p(d),\ d = 0, 1, \ldots)$ denote the pmf of the limiting rv D.

Assumption 3 *(Asymptotic uncorrelatedness) Under Assumption 1, for each* $d = 0, 1, \ldots$, *the rvs* $\mathbf{1}[D_{n,1} = d]$ *and* $\mathbf{1}[D_{n,2} = d]$ *are* **asymptotically uncorrelated** *in the sense that*

$$\lim_{n\to\infty} \mathrm{Cov}\left[\mathbf{1}[D_{n,1} = d], \mathbf{1}[D_{n,2} = d]\right] = 0.$$

Assumptions 1 and 2 are the baseline assumptions

Assumption 3 amounts to

$$\lim_{n\to\infty} \left(\mathbb{P}[D_{n,1} = d, D_{n,2} = d] - \mathbb{P}[D_{n,1} = d]\,\mathbb{P}[D_{n,2} = d]\right) = 0.$$

Main (but easy) result

Proposition 1 *Under Assumptions 1-3, we have*

$$\frac{N_n(d)}{n} \xrightarrow{P}_n p(d), \quad d = 0, 1, \ldots$$

where the pmf $\boldsymbol{p} = (p(d),\ d = 0, 1, \ldots)$ *is postulated in Assumption 2.*

Here, there is **equivalence** between L^2-convergence and convergence in probability:

- Convergence in probability implies L^2-convergence by bounded convergence
- L^2 convergence implies convergence in probability by Tchebychev's inequality: With $\varepsilon > 0$, Tchebychev's inequality gives

$$\mathbb{P}\left[\left|\frac{N_n(d)}{n} - p(d)\right| > \varepsilon\right] \leq \frac{1}{\varepsilon^2} \cdot \mathbb{E}\left[\left|\frac{N_n(d)}{n} - p(d)\right|^2\right].$$

Thus, as we **mimic** the proof of WLLNs, we need only prove that

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\frac{N_n(d)}{n} - p(d)\right|^2\right] = 0, \quad d = 0, 1, \ldots$$

Lemma 2 *If Assumptions 1-2 hold, then*

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\frac{N_n(d)}{n} - p(d)\right|^2\right] = \lim_{n\to\infty} \mathrm{Cov}\left[\mathbf{1}\left[D_{n,1} = d\right], \mathbf{1}\left[D_{n,2} = d\right]\right]$$

for each $d = 0, 1, \ldots$ with the understanding that if one of the limits exists, so does the other and the limiting values coincide.

Fix $n = 2, 3, \ldots$ and $d = 0, 1, \ldots$: Centering about the mean, we have

$$\mathbb{E}\left[\left|\frac{N_n(d)}{n} - p(d)\right|^2\right] = \mathrm{Var}\left[\frac{N_n(d)}{n}\right] + \left|\mathbb{E}\left[\frac{N_n(d)}{n}\right] - p(d)\right|^2.$$

$$\begin{aligned} \mathbb{E}\left[N_n(d)\right] &= \mathbb{E}\left[\sum_{k=1}^{n} \mathbf{1}\left[D_{n,k} = d\right]\right] \\ &= \sum_{k=1}^{n} \mathbb{P}\left[D_{n,k} = d\right] = n \cdot \mathbb{P}\left[D_{n,1} = d\right] \end{aligned}$$

Assumption 1

so that

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{N_n(d)}{n}\right] = \lim_{n\to\infty} \mathbb{P}\left[D_{n,1} = d\right] = p(d)$$

Assumption 2

$$\lim_{n\to\infty} \left|\mathbb{E}\left[\frac{N_n(d)}{n}\right] - p(d)\right|^2 = 0$$

$$\begin{aligned}\mathrm{Var}\left[N_n(d)\right] &= n \cdot \mathrm{Var}\left[\mathbf{1}\left[D_{n,1}=d\right]\right] \\ &\quad + n(n-1)\cdot \mathrm{Cov}\left[\mathbf{1}\left[D_{n,1}=d\right], \mathbf{1}\left[D_{n,2}=d\right]\right]\end{aligned}$$

Assumption 1

so that

$$\begin{aligned}\mathrm{Var}\left[\frac{N_n(d)}{n}\right] &= \frac{1}{n}\cdot \mathrm{Var}\left[\mathbf{1}\left[D_{n,1}=d\right]\right] \\ &\quad + \frac{n-1}{n}\cdot \mathrm{Cov}\left[\mathbf{1}\left[D_{n,1}=d\right], \mathbf{1}\left[D_{n,2}=d\right]\right]\end{aligned}$$

$$\lim_{n\to\infty} n^{-1}\mathrm{Var}\left[\mathbf{1}\left[D_{n,1}=d\right]\right] = 0$$

Always

Combining, under Assumption 1 and Assumption 2, we conclude that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[\left|\frac{N_n(d)}{n} - p(d)\right|^2\right] \\
&= \lim_{n\to\infty}\left(\operatorname{Var}\left[\frac{N_n(d)}{n}\right] + \left|\mathbb{E}\left[\frac{N_n(d)}{n}\right] - p(d)\right|^2\right) \\
&= \lim_{n\to\infty}\operatorname{Var}\left[\frac{N_n(d)}{n}\right] \\
&= \lim_{n\to\infty}\operatorname{Cov}\left[\mathbf{1}\left[D_{n,1} = d\right], \mathbf{1}\left[D_{n,2} = d\right]\right]
\end{aligned}$$

The desired result

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\frac{N_n(d)}{n} - p(d)\right|^2\right] = 0$$

is **equivalent** to Assumption 3

Corollary 1 *Under Assumptions 1-2, we have*

$$\frac{N_n(d)}{n} \xrightarrow{P}_n p(d), \quad d = 0, 1, \ldots$$

where the pmf $\boldsymbol{p} = (p(d),\ d = 0, 1, \ldots)$ *is postulated in Assumption 2* **if and only if** *Assumptions 3 holds, namely*

$$\lim_{n\to\infty} \mathrm{Cov}\left[\mathbf{1}\left[D_{n,1} = d\right], \mathbf{1}\left[D_{n,2} = d\right]\right] = 0$$

Convergence of

$$\left\{ \frac{N_n(d)}{n}, \; n = 1, 2, \ldots \right\}$$

in either mode **fails** if

$$\lim_{n \to \infty} \mathrm{Cov} \left[\mathbf{1} \left[D_{n,1} = d \right], \mathbf{1} \left[D_{n,2} = d \right] \right] > 0$$

A commonly occurring framework

Given is an underlying **parametric** family of random graphs

$$\{\mathbb{G}(n;\alpha),\ n=2,3,\ldots\},\quad \alpha\in A\subset\mathbb{R}^p$$

Typically, for each $\alpha\in A$, the degree rvs $D_{n,1}(\alpha),\ldots,D_{n,n}(\alpha)$ are **exchangeable**.

There **exists** a scaling $\alpha^\star:\mathbb{N}_0\to A$ such that

$$D_{n,1}(\alpha^\star_n)\Longrightarrow D$$

for some **non**-degenerate $\mathbb{N}$-valued rv – Existence of a **maximal** component

Set

$$\mathbb{G}_n = \mathbb{G}(n; \alpha_n^\star), \quad n = 1, 2, \ldots$$

Assumptions 1 and 2 are **automatically** satisfied, so that only Assumption 3 need to be verified

---

The result applies to a long list of models:

- Erdős-Rényi graphs $\mathbb{G}(n; p)$ $\qquad p_n^\star \sim \frac{\lambda}{n}$
- Geometric random graphs $\mathbb{G}(n; \rho)$ $\qquad \pi \left(\rho_n^\star\right)^2 \sim \frac{\lambda}{n}$
- Random key graphs $\mathbb{K}(n; \theta)$ $\qquad \frac{(K_n^\star)^2}{P_n^n} \sim \frac{\lambda}{n}$
- $\ldots$

Always?

Random threshold graphs
(A counterexample)

Motivation

Preferential attachment model is based on **growth** and **preferential attachment**

- Cumulative advantage – “The rich get richer”
- Predicated on **information** about the degree of each vertex being available to newly added nodes, either explicitly or implicitly – Questionable assumption in some cases

---

An answer to this modeling issue: **Hidden variable** models (Caldarelli et al. (2002))

- Creation of a link between two nodes expresses a **mutual** benefit based on **intrinsic** attributes

Random threshold graphs

Fitness – **I.i.d.** rvs $\{\xi, \xi_i,\ i = 1, 2, \ldots, n\}$

"The rich know the richer" – With **threshold** θ in $\mathbb{R}$,

$$\mathbb{T}(n;\theta): \quad i \sim j \quad \text{if and only if} \quad \xi_i + \xi_j > \theta$$

Non-negative fitness – The fitness rvs $\xi_1, \ldots, \xi_n$ are assumed to be **non-negative** i.i.d. rvs — We need consider only $\theta > 0$.

Already a large body of work concerning this class of random graph models, e.g., degree distribution, clustering, degree correlations, etc.

In particular, power law for the degree distribution can emerge in the many node limit ($n \to \infty$) when $\xi \sim \mathrm{Exp}(\lambda)$ under scaling $\theta^\star : \mathbb{N}_0 \to \mathbb{R}_+$ such that

$$\theta_n^\star = \lambda^{-1} \log n, \quad n = 1, 2, \ldots$$

Fujihara et al. (2003) show that $D_{n,1}(\theta_n^\star) \Longrightarrow_n D$, namely

$$\lim_{n\to\infty} \mathbb{P}\left[D_{n,1}(\theta_n^\star) = d\right] = p_{\mathrm{Fuj}}(d), \quad d = 0, 1, \ldots$$

with

$$p_{\mathrm{Fuj}}(d) = \mathbb{E}\left[\frac{\left(e^{\lambda\xi}\right)^d}{d!} e^{-e^{\lambda\xi}}\right], \quad d = 0, 1, \ldots$$

Conditional Poisson pmf

(with rate $e^{\lambda\xi}$)

Claim (Caldarelli et al.): Random threshold graphs with exponential fitness provide an **alternative** to the preferential attachment model due to the fact that

$$p_{\mathrm{Fuj}}(d) \sim d^{-2} \quad (d \to \infty)$$

Beware!

Applying the theory ($\xi \sim \text{Exp}(\lambda)$)

Here we focus on

$$\mathbb{G}_n = \mathbb{T}(n; \theta_n^\star), \quad n = 1, 2, \ldots$$

Assumption 1 holds: Trivial by exchangeability

Assumption 2 holds with $\boldsymbol{p} = \boldsymbol{p}_{\text{Fuj}}$ as shown by Fujihara et al. (2003)

Assumption 3 **fails**!

Formally,

Proposition 2 *With* $\xi \sim \mathrm{Exp}(\lambda)$ *for some* $\lambda > 0$*, for each* $d = 0, 1, \ldots$*, the limit*

$$C(d) \equiv \lim_{n\to\infty} \mathrm{Cov}[\mathbf{1}\left[D_{n,1}(\theta_n^\star) = d\right], \mathbf{1}\left[D_{n,2}(\theta_n^\star) = d\right]]$$

exists with $C(d) > 0$.

For instance,

$$C(0) = \mathbb{E}\left[e^{-\max(e^{\lambda\xi_1}, e^{\lambda\xi_2})}\right] - \mathbb{E}\left[e^{-(e^{\lambda\xi_1} + e^{\lambda\xi_2})}\right] > 0$$

since $\max(e^{\lambda\xi_1}, e^{\lambda\xi_2}) < e^{\lambda\xi_1} + e^{\lambda\xi_2}$

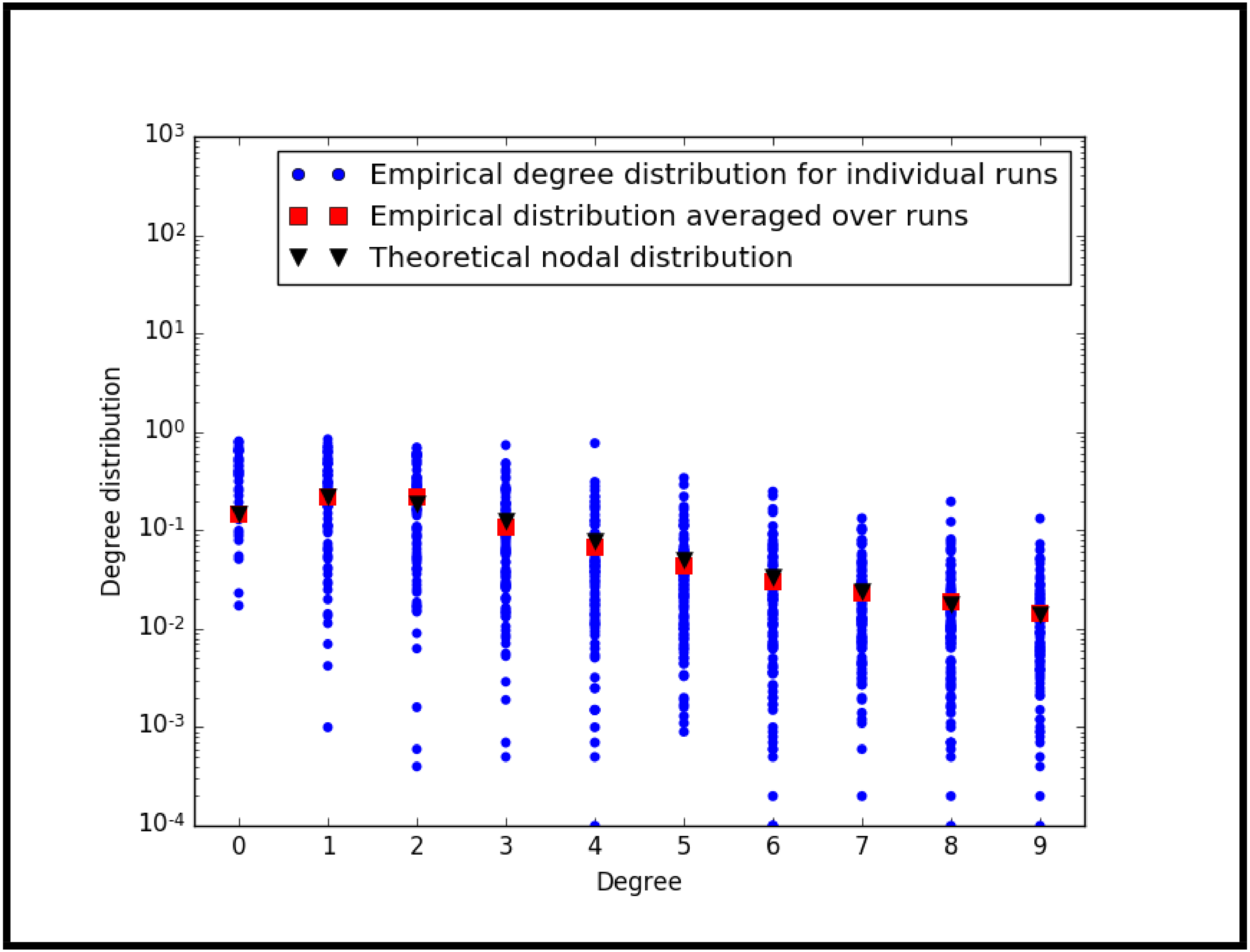
Empirical degree distribution for individual runs
Empirical distribution averaged over runs
Theoretical nodal distribution
Degree distribution
Degree
10^3
10^2
10^1
10^0
10^-1
10^-2
10^-3
10^-4
0 1 2 3 4 5 6 7 8 9

Figure 1:

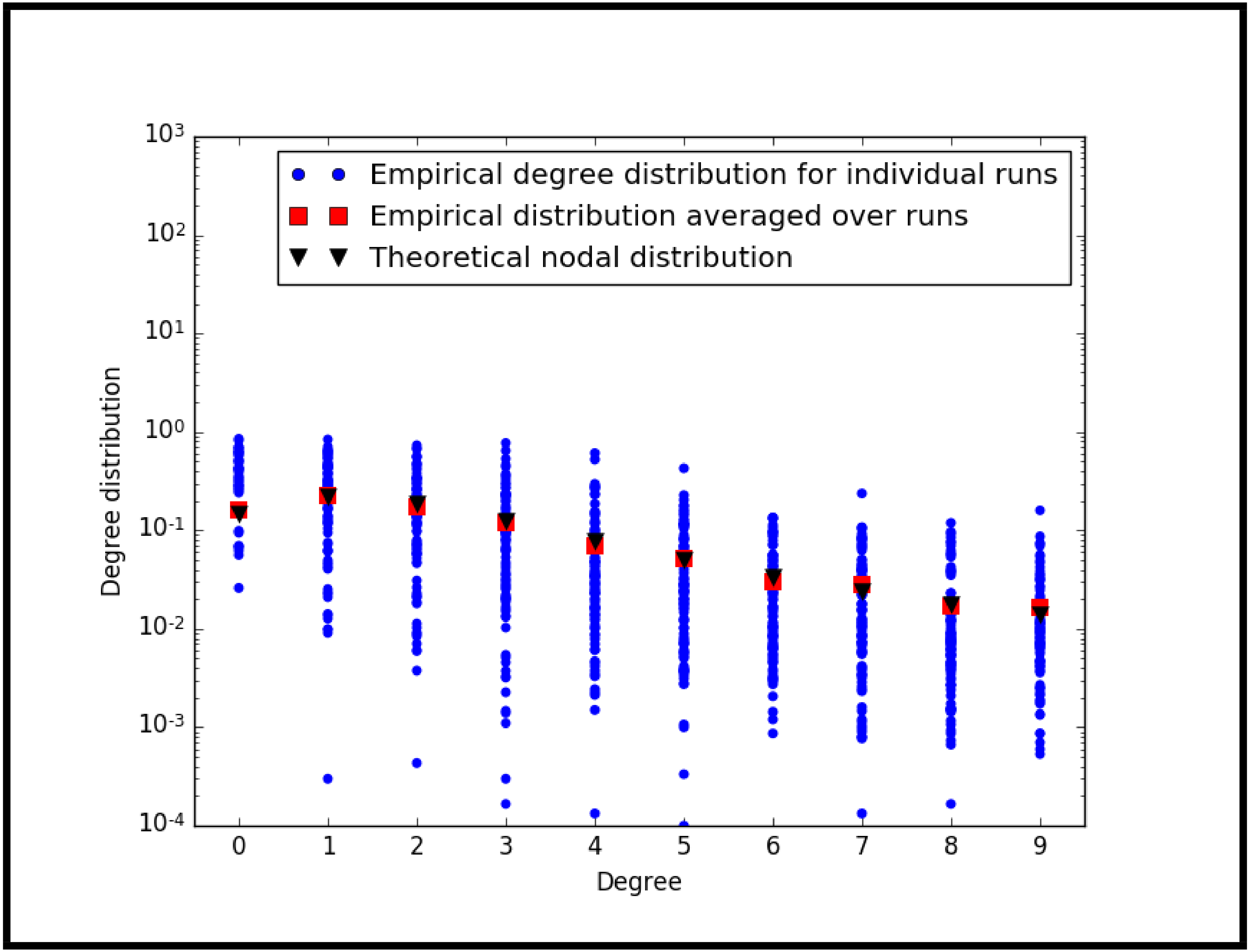
Empirical degree distribution for individual runs
Empirical distribution averaged over runs
Theoretical nodal distribution
Degree distribution
Degree
10^3
10^2
10^1
10^0
10^-1
10^-2
10^-3
10^-4
0
1
2
3
4
5
6
7
8
9

Figure 2:

Limiting empirical distribution?

Theorem 1 *For each* $d = 0, 1, \ldots,$

$$\frac{N_n(d)}{n} \Longrightarrow_n \Pi(d)$$

where $\Pi(d)$ *is a* **non-degenerate** $[0,1]$*-valued rv with*

$$\mathbb{E}\left[\Pi(d)\right] = p_{\mathrm{Fuj}}(d) \quad \textit{and} \quad \mathrm{Var}\left[\Pi(d)\right] > 0.$$

Basic ideas of the proof

- Multi-dimensional version of Fujihara et al. – For each $r = 1, 2, \ldots,$

$$\lim_{n\to\infty} \mathbb{P}\left[D_{n,1} = d_1, \ldots, D_{n,r} = d_r\right]$$

- Method of moments:

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{N_n(d)}{n}\right)^r\right] = \lim_{n\to\infty} \mathbb{P}\left[D_{n,1} = d, \ldots, D_{n,r}(d)\right]$$

- Characteristic function

$$\lim_{n\to\infty} \mathbb{E}\left[e^{it\frac{N_n(d)}{n}}\right], \quad t \in \mathbb{R}$$

Non-homogeneous random networks

. . .

Consider the sequence of random graphs $\{\mathbb{G}_n, \ n = 1, 2, \ldots\}$ **without** any additional assumptions.

For each $n = 2, 3, \ldots$, write $V_n = \{1, \ldots, n\}$ and set

$$\Sigma_n \equiv \{(k, \ell) \in V_n \times V_n : \ k \neq \ell\}$$

Consider the rv $(\nu_n, \mu_n) : \Omega \to \Sigma_n$ which is **uniformly** distributed over Σ_n, i.e.,

$$\mathbb{P}\left[\nu_n = k, \mu_n = \ell\right] = \frac{1}{n(n-1)}, \qquad \begin{array}{c} k \neq \ell \\ k, \ell \in V_n. \end{array}$$

The Σ_n-valued rv (ν_n, μ_n) models the randomly uniform selection of two nodes in V_n (without repetition). Each of the rvs ν_n and μ_n is **uniformly** distributed over V_n. The selection rv (ν_n, μ_n) is assumed to be **independent** of the random graph $\mathbb{G}_n$.

Proposition 3 *Under the foregoing assumptions, with* $d = 0, 1, \ldots$, *the convergence*

$$\frac{N_n(d)}{n} \xrightarrow{P}_n L(d)$$

holds for some scalar $L(d)$ *in* $\mathbb{R}$ *if and only if*

$$\lim_{n\to\infty} \mathbb{P}\left[D_{n,\nu_n} = d\right] = L(d) \tag{2}$$

and

$$\lim_{n\to\infty} \mathrm{Cov}\left[\mathbf{1}\left[D_{n,\nu_n} = d\right], \mathbf{1}\left[D_{n,\mu_n} = d\right]\right] = 0. \tag{3}$$

$$\mathbb{P}\left[D_{n,\nu_n} = d\right] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\left[D_{n,k} = d\right]\right] = \mathbb{E}\left[\frac{N_n(d)}{n}\right]$$

Remarks

- **True generalization:** Under Assumption 1, for each $n = 1, 2, \ldots$ we have the distributional equalities $D_{n,\nu_n} =_{st} D_{n,1}$ and $D_{n,\mu_n} =_{st} D_{n,1}$, and

$$(D_{n,\nu_n}, D_{n,\mu_n}) =_{st} (D_{n,1}, D_{n,2})$$

- **No** assumption on the sequence $\{\mathbb{G}_n,\ n = 2, 3, \ldots\}$, but also no operational ability of **equating** asymptotically the two different degree distributions available in the homogeneous case. Yet, . . .
- Applies when the graphs $\{\mathbb{G}_n,\ n = 2, 3, \ldots\}$ are (i) deterministic (ii) **non**-homogeneous Barabási-Albert model (and other growth models) and (iii) many other models

Conclusions

- The degree distribution of a **node** vs. the degree distribution of the **network**
  - A word of caution even in **homogeneous** parametric networks!
- Equality in the limit under the appropriate scaling: Often true but still needs to be established
  - Theory given earlier provides a framework for doing so
- **Counterexample:** Preferential attachment vs. hidden variable models (random threshold graphs with exponential fitness)
  - **Not** comparable!

  - Convergence **in distribution** of empirical frequencies to a non-degenerate rv
  - **Do not always believe what you read!**

- Non-homogeneous models and random sampling
  - A new approach to convergence in growth models?