On the Scale-Free Property of Citation Networks: An Empirical Study

Xiaoshi Zhong*
School of CS & Tech, Beijing Institute of Technology
CAT Institute, China North Industries Group Corporation
Beijing, China
xszhong@bit.edu.cn

Huizhi Liang*
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
hzliang@bit.edu.cn

ABSTRACT

Citation networks have been thought to exhibit scale-free property for many years; however, this assertion has been doubted recently. In this paper, we conduct extensive experiments to resolve this controversial issue. We firstly demonstrate the scale-free property in scale-free networks sampled from the popular Barabasi-Albert (BA) model. To this end, we employ a merged rank distribution, which is divided into outliers, power-law segment, and non-power-law data, to characterize network degrees, and propose a random sample consensus (RANSAC)-based method to identify power-law segments from merged rank distributions, and use the Kolmogorov-Smirnov (KS) test to examine the scale-free property in power-law segments. Subsequently, we apply the same methods to examine the scale-free property in real-world citation networks. Experimental results confirm the scale-free property in citation networks and attribute previous skepticism to the presence of outliers.

CCS CONCEPTS

• **Networks** → *Network Properties*; Citation Networks.

KEYWORDS

Citation networks, scale-free property, merged rank distribution, power-law segment, random sample consensus (RANSAC)

ACM Reference Format:
Xiaoshi Zhong and Huizhi Liang. 2024. On the Scale-Free Property of Citation Networks: An Empirical Study. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651541

1 INTRODUCTION

Citation networks has served as invaluable tools for analyzing the impact of scientific discoveries, researchers, and publishing venues [11]. Their significance underscores the importance of understanding their properties, among which a controversial one is the scale-free property. A network is said to exhibit scale-free property if the distribution of either its degrees or its nodes can be characterized by a power-law model. Citation networks have been thought to exhibit scale-free property for many years [1, 7–9]; however, a recent research casted doubt on this longstanding assertion [4].

In this paper, we conduct extensive experiments to resolve this controversial issue. Specifically, we firstly demonstrate the scale-free property in networks sampled from scale-free models through a three-step process: (1) employing a merged rank distribution to characterize the degrees of these sampled scale-free networks and partitioning the merged rank distribution into three categories:*outliers*; representing the highest-degree nodes deviating from the fitted power-law line; *power-law segment*, indicating a well-fitted segment with a power-law; and *non-power-law data*, comprising nodes with degrees smaller than a threshold and deviating from the fitted power-law line; (2) introducing a method based on random sample consensus (RANSAC) [2] to identify power-law segments from these merged rank distributions; and (3) utilizing the Kolmogorov-Smirnov (KS) test to examine the scale-free property in these identified power-law segments. After demonstrating the scale-free property in sampled scale-free networks, we apply the same methods to examine the scale-free property in real-world citation networks. Our experimental results on nine versions of citation networks, spanning over a decade, clearly demonstrate the presence of scale-free property in citation networks. Furthermore, our experiments indicate that it is the outliers that lead Golosovsky (2017) [4] to argue against the scale-free nature of citation networks.

Compared to previous studies, our research presents several distinct contributions. Firstly, while previous studies predominantly focus on examining the scale-free property through degree distributions of citation networks, our research advocates for analyzing the scale-free property in a merged rank distribution, which we argue to be a more effective strategy. Secondly, while prior studies tend to characterize the entire distribution or its long tails in networks using a power-law model, our research suggests that the scale-free property occurs primarily within a specific segment of citation network distributions. Thirdly, whereas previous studies typically examine the scale-free property solely in real-world citation networks, our research advocates a two-step examination: firstly confirming the scale-free property in sampled scale-free networks and subsequently examining it in real-world citation networks.

2 METHODOLOGY

2.1 Merged Rank Distribution

In a network with n nodes, we sort the n nodes by their degrees in descending order and let k_r denote the degree of the r-th node or let

*Both authors contributed equally. X. Zhong is the corresponding author.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651541

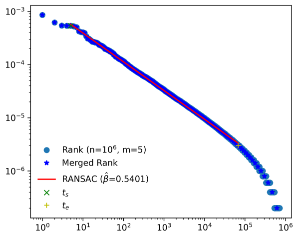


Figure 1: Rank distribution ("Rank") and merged rank distribution ("Merged Rank") of a sampled scale-free network

$p(r)$ denote the proportion of degrees of the r-th node. The degrees of the network are deemed to follow a rank power-law distribution if the data points $\{(r, k_r)\}$ or $\{(r, p(r))\}$ (where $1 \le r \le n$) can be characterized by Eq. (1) or Eq. (2), respectively:

$$k_r = C \cdot r^{-\beta} \quad (1)$$

$$p(r) = C' \cdot r^{-\beta} \quad (2)$$

where $\beta > 0$ represents the scaling exponent while $C > 0$ and $C' > 0$ are normalized constants.

In empirical networks, it is common to encounter data points in which different r values have equal k_r or $p(r)$ values; namely $\exists r_i \neq r_j \Rightarrow k_{r_i} = k_{r_j}$ or $p(r_i) = p(r_j)$. This violates the strictly decreasing property of power-law functions like Eq. (1) or Eq. (2): $\forall r_i < r_j \Rightarrow k_{r_i} > k_{r_j}$ or $p(r_i) > p(r_j)$. However, these data points with equal k_r values appear with consecutive r values. Suppose the r values that correspond to all equal k_r values range from r_s to r_e, we can merge all of them into a single data point (r_c, k_r), where $r_c = \sqrt{r_s \cdot r_e}$ represents the center of r_s and r_e on double-logarithmic axes or the geometric average on original axes.

$\{(r_c, k_r)\}$ or $\{(r_c, p(r_c))\}$ is termed the "merged rank distribution" (referred to as "mRankDist").[1] Figure 1 illustrates the RankDist and the mRankDist of a network sampled from the Barabasi-Albert (BA) model. Unlike RankDist, which generally lacks the strictly decreasing property in empirical networks, mRankDist consistently exhibits this property. In line with [12], we observe that not all high-degree nodes in citation networks conform to a power-law distribution. We also observe that the mRankDist can be divided into three distinct segments: *outliers*, *power-law segment*, *non-power-law data*. The power-law segment represents a portion of the mRankDist that aligns well with a power-law model. Outliers denote the highest-degree nodes deviating from the fitted power-law line, while non-power-law data comprises nodes with degrees smaller than a designated threshold k_{min} and deviating from the fitted power-law line. As depicted in Figure 1, the delineation between the three segments within mRankDist is marked by data points t_s and t_e.

It is worth noting that when each distinct r corresponds to a unique k_r or $p(r)$, the mRankDist and RankDist are the same.

2.2 RANSAC: Identifying Power-Law Segments from Merged Rank Distributions

We propose a method based on RANSAC [2] to identify power-law segments from merged rank distributions. Suppose the mRankDist is represented by $p(r_c) = A \cdot r_c^{-\beta}$, then it is equivalent to Eq. (3):

$$s = \beta \cdot t + a \quad (3)$$

[1] We designate the normal rank distribution(s) as "RankDist" or "nRankDist" while using "mRankDist" specifically for the merged rank distribution(s).

Algorithm 1: Identifying Power-Law Segment

Input: $\{(t_i, s_i)\}_1^N$, $\hat{\beta}$, $\hat{a}$, $\hat{s}_i \leftarrow t_i \times \hat{\beta} + \hat{a}$, δ_{err}, δ_{count}
Output: t_s, t_e
Initiate: $t_s = 1$, $t_e = N$, $\delta_{count} = 10$, $\delta_{err} = 0.1$
for $i \leftarrow 1$ **to** N **do**
  $flag \leftarrow$ FALSE
  for $j \leftarrow i$ **to** $i + \delta_{count}$ **do**
    if $|\hat{s}_j - s_j|/s_j > \delta_{err}$ **then**
      $flag \leftarrow$ TRUE and break
  if $flag$ *is TRUE* **then**
    $t_s \leftarrow t_i$ and break
for $i \leftarrow N$ **to** *1* **do**
  $flag \leftarrow$ FALSE
  for $j \leftarrow i$ **to** $i - \delta_{count}$ **do**
    if $|\hat{s}_j - s_j|/s_j > \delta_{err}$ **then**
      $flag \leftarrow$ TRUE and break
  if $flag$ *is TRUE* **then**
    $t_e \leftarrow t_i$ and break

where $s = -\log p(r_c)$, $t = \log r_c$, and $a = -\log A$.

Given a set of data points $\{(t_i, s_i)\}$, where $1 \le i \le N$, characterized by a line as described in Eq. (3), the RANSAC algorithm aims to estimate the parameters (i.e., $\hat{\beta}$ and $\hat{a}$) of the power-law model and identify the power-law segment (which is separated by t_s and t_e). Initially, the classic RANSAC algorithm is employed to estimate $\hat{\beta}$ and $\hat{a}$. Subsequently, an algorithm outlined in Algorithm 1 is designed to identify the power-law segment. This algorithm begins by setting error and count thresholds, denoted as δ_{err} and δ_{count} respectively, then iteratively scans the mRankDist from start to finish. At each data point, it checks if the subsequent points, say δ_{count}, confom to the linear model within the error threshold. The first point meeting this criterion is marked as the starting point t_s of the segment. A similar reverse iteration is performed to determine the segment's end point t_e, effectively isolating the segment that best fits the linear model while excluding outliers.

2.3 Goodness-of-Fit Testing

In line with prior studies [3, 12, 13], we utilize the KS test [5, 6] to examine the goodness-of-fit. The KS test is a nonparametric method that evaluates the null hypothesis H_0: a sample with N observations is drawn from a hypothesized population. The KS statistic is defined by Eq. (4):

$$D_n = \sup_x |F_n(x) - F(x)| \quad (4)$$

where $F_n(x)$ represents the cumulative density function (CDF) of the sample, $F(x)$ denotes the CDF of hypothesized population, and sup_x denotes the supremum of the set of distances.

We employ the critical KS statistic (D_n^α) with a significance level of $\alpha = 0.05$ for hypothesis testing, as per the table provided by Massey [6] (refer to its Table 1). Specifically, for sample sizes $N > 35$, the critical KS statistic is defined as $D_n^{0.05} = \frac{1.36}{\sqrt{N}}$. For a calculated value of D_n, we compare it to $D_n^{0.05}$: if $D_n > D_n^{0.05}$, then reject H_0; otherwise, accept H_0. It is worth noting that while non-rejection does not equate to acceptance, it is customary in this area to consider non-rejection as indicative of acceptance.

Table 1: Statistics of the nine versions of citation networks. Nodes represent papers while edges represent citations.

Dataset	Year	# Nodes	# Edges
DBLP-V3	2010	333,977	2,265,340
DBLP-V5	2011	982,768	24,288,367
DBLP-V6	2013	1,375,399	38,339,598
DBLP-V7	2014	517,577	4,001,021
DBLP-V8	2016	650,218	6,340,666
DBLP-V9	2017	33,726	1,618,844
DBLP-V10	2017	2,356,111	1,08,445,406
DBLP-V11	2019	2,973,396	67,170,765
DBLP-V12	2020	3,538,030	82,226,114

3 EXPERIMENTS

We conduct experiments to firstly demonstrate the scale-free property in sampled scale-free networks and subsequently examine the scale-free property in real-world citation networks.

3.1 Experimental Setup

3.1.1 Datasets. In our experiments, we utilize two types of datasets: (1) sampled scale-free networks and (2) real-world citation networks. For sampled scale-free datasets, we sample networks from the BA model with different settings: $n = 10^4, 10^5, 10^6$ and $m = 5, 10$; here n denotes the number of total nodes while m indicates the number of edges added with each new node to the network. We sample 500 networks for each (n, m) setting and focus on in-degree networks, due to the directed nature of citations. For real-world citation networks, we collect various versions of the DBLP citation networks[2] spanning over a decade [10]. Table 1 summarizes the statistics of these citation networks.

3.1.2 State-of-the-art Baselines. We compare the fitting results of our method to two state-of-the-art models: GA2019 [3] and LS_{avg} [12]. GA2019 employs maximum-likelihood estimation while LS_{avg} utilizes least-squares fitting to fit power-law distributions.

3.1.3 Implementation and Evaluation Metrics. GA2019 is applied on RankDist, while both LS_{avg} and our RANSAC-based method are applied on mRankDist. For GA2019 and LS_{avg}, the KS test is performed on the entire distribution; while for RANSAC, the KS test is conducted on both the identified power-law segment and the entire mRankDist. In terms of evaluation metrics, we utilize the KS test, as described in Section 2.3, to assess the goodness-of-fit.

3.2 Experimental Results

3.2.1 Results on Sampled Directed Scale-Free Networks. Table 2 reports the fitting and testing results of RANSAC, GA2019, and LS_{avg} on rank distributions of sampled directed scale-free networks. The table shows that RANSAC achieves the best D_n^{seg} with all 500 accepts; by contrast, both GA2019 and LS_{avg} reject all sampled networks as scale-free networks. This highlights RANSAC's superior performance over GA2019 and LS_{avg}. Conversely, when considering outliers and non-power-law data, RANSAC identifies some sampled networks as non-scale-free, suggesting the validity of segmenting mRankDist for scale-free networks. Furthermore, Figure 2 visually depicts examples of fitting to sampled scale-free networks, displaying that scale-free property occurs in partial segments rather than across the entire distributions.

3.2.2 Results on Real-World Citation Networks. Table 3 presents the fitting and testing results of RANSAC, GA2019, and LS_{avg} on rank distributions of real-world citation networks. Notably, RANSAC successfully identifies almost all power-law segments from mRankDist, passing the KS test, with the exception of the DBLP-V9 network due to its significantly smaller size compared to other networks. This performance surpasses that of GA2019 and LS_{avg}. Due to limited space, Figure 3 illustrates some of these fitting and testing results. These results align with those observed in sampled scale-free networks, as reported in Table 2 and depicted in Figure 2. This consistency suggests the scale-free property inherent in citation networks, with outliers and non-power-law data causing citation networks to be non-scale-free.

4 CONCLUSION

In this paper, we resolve the controversial issue surrounding the scale-free nature of citation networks. We achieve this by characterizing network degrees through a merged rank distribution with three distinct portions and identifying power-law segments from merged rank distributions using a RANSAC-based method. Experimental results on both sampled scale-free networks and real-world citation networks conclusively demonstrate the presence of scale-free property in citation networks.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable feedback and constructive suggestions.

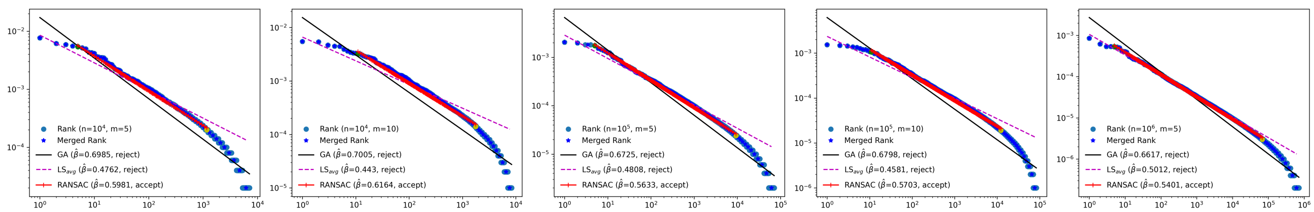
REFERENCES

[1] Derek de Solla Price. 1976. A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American society for Information science* 27, 5 (1976), 292–306.
[2] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.
[3] Martin Gerlach and Eduardo G. Altmann. 2019. Testing Statistical Laws in Complex Systems. *Physical Review Letters* 122, 16 (2019), 168301.
[4] Michael Golosovsky. 2017. Power-law citation distributions are not scale-free. *Physical Reviewe E* 96, 3 (2017), 032306.
[5] Andrey Kolmogorov. 1941. Confidence Limits for an Unknown Distribution Function. *The Annals of Mathematical Statistics* 12, 4 (1941), 461–463.
[6] Jr. Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (1951), 68–78.
[7] Derek J. De Solla Price. 1965. Networks of Scientific Papers. *Science* 149, 3683 (1965), 510–515.
[8] Sidney Redner. 1998. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B* 4, 2 (1998), 131–134.
[9] Sidney Redner. 2005. Citation Statistics from more than a Century of Physical Review. *Bulletin of the American Physical Society* (2005), 1–17.
[10] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD'08*. 990–998.
[11] Ludo Waltman. 2016. A review of the literature on citation impact indicators. *Journal of informetrics* 10, 2 (2016), 365–391.
[12] Xiaoshi Zhong, Muyin Wang, and Hongkun Zhang. 2022. Is Least-Squares Inaccurate in Fitting Power-Law Distributions? The Criticism is Complete Nonsense. In *Proceedings of The ACM Web Conference 2022*. 2748–2758.
[13] Xiaoshi Zhong, Xiang Yu, Erik Cambria, and Jagath C. Rajapakse. 2023. Marshall-Olkin Power-Law Distributions in Length-Frequency of Entities. *Knowledge-Based Systems* 279 (2023), 110942.

[2] https://www.aminer.org/citation

Table 2: Fitting and testing result of RANSAC, GA2019, and LS_{avg} on rank distributions of sampled directed scale-free networks. The reported results are based on the average of 500 sampled networks. KS denotes the number of "accept/reject" results among the 500 sampled networks. For RANSAC, after computing $\hat{\beta}$, the KS test is conducted on two sets of data points from mRankDist: (1) the power-law segment (which is indicated by "*seg*") and (2) all data points ("*all*").

(n, m)	RANSAC					GA2019			LS_{avg}		
	$\hat{\beta}$	D_n^{seg}	KS_{seg}	D_n^{all}	KS_{all}	$\hat{\beta}$	D_n	KS	$\hat{\beta}$	D_n	KS
$(10^4, 5)$	0.6121	0.0062	500/0	0.0173	500/0	0.6954	0.2605	0/500	0.4721	0.0264	0/500
$(10^4, 10)$	0.6237	0.0106	500/0	0.0306	118/382	0.6975	0.3250	0/500	0.4272	0.0398	0/500
$(10^5, 5)$	0.5665	0.0015	500/0	0.0051	493/7	0.6736	0.2740	0/500	0.4885	0.0083	0/500
$(10^5, 10)$	0.5734	0.0025	500/0	0.0092	377/123	0.6803	0.3335	0/500	0.4570	0.0138	0/500
$(10^6, 5)$	0.5402	0.0004	500/0	0.0014	500/0	0.6616	0.2839	0/500	0.4986	0.0022	0/500
$(10^6, 10)$	0.5448	0.0005	500/0	0.0026	452/48	0.6695	0.3415	0/500	0.4797	0.0040	0/500

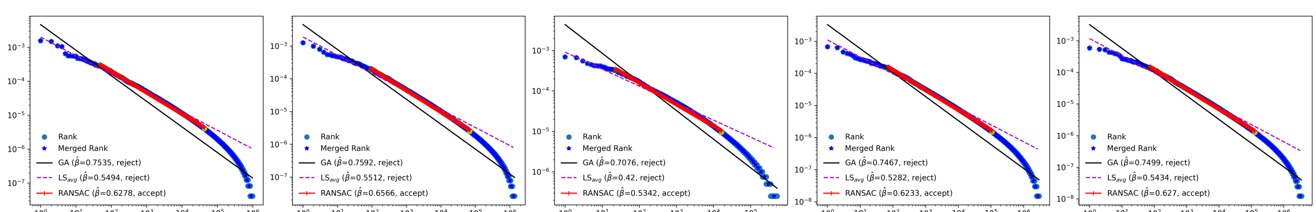


(a) n=10^4, m=5 (b) n=10^4, m=10 (c) n=10^5, m=5 (d) n=10^5, m=10 (e) n=10^6, m=5

Figure 2: Examples of fitting and testing results of RANSAC, GA2019, and LS_{avg} on rank distributions of sampled directed scale-free networks (best viewed in color). The horizontal axis indicates ranks r while the vertical one indicate $p(r)$. For a fitting by RANSAC, *outliers*, *power-law segment*, and *non-power-law data* are separated by the green and yellow data points.

Table 3: Fitting and testing result of RANSAC, GA2019, and LS_{avg} on rank distributions of real-world citation networks. The notations are consistent with those used in Table 2, with the exception that KS here indicates the result of the KS test conducted on individual networks.

(n, m)	RANSAC					GA2019			LS_{avg}		
	$\hat{\beta}$	D_n^{seg}	KS_{seg}	D_n^{all}	KS_{all}	$\hat{\beta}$	D_n	KS	$\hat{\beta}$	D_n	KS
DBLP-V3	0.5236	0.0012	**accept**	0.0028	reject	0.7012	0.3284	reject	0.4747	0.0043	reject
DBLP-V5	0.6278	0.0015	**accept**	0.0121	reject	0.7535	0.4832	reject	0.5494	0.0096	reject
DBLP-V6	0.6566	0.0014	**accept**	0.0186	reject	0.7592	0.4900	reject	0.5512	0.0121	reject
DBLP-V7	0.5342	0.0013	**accept**	0.0055	reject	0.7076	0.3541	reject	0.4200	0.0078	reject
DBLP-V8	0.5846	0.0017	**accept**	0.0070	reject	0.7375	0.3814	reject	0.4949	0.0122	reject
DBLP-V9	0.6336	0.0113	reject	0.0464	reject	0.6412	0.2527	reject	0.3816	0.0407	reject
DBLP-V10	0.6280	0.0008	**accept**	0.0085	reject	0.6631	0.4513	reject	0.5441	0.0066	reject
DBLP-V11	0.6233	0.0010	**accept**	0.0111	reject	0.7467	0.4788	reject	0.5282	0.0085	reject
DBLP-V12	0.6270	0.0007	**accept**	0.0100	reject	0.7499	0.4842	reject	0.5434	0.0077	reject



(a) DBLP-V5 (b) DBLP-V6 (c) DBLP-V7 (d) DBLP-V11 (e) DBLP-V12

Figure 3: Fitting and testing results of RANSAC, GA2019, and LS_{avg} on rank distributions of real-world citation networks