

An Empirical Study on the Role of Macro-Meso-Micro Measures in Citation Networks

Rishabh Narang[1], Sanjay Misra[2], and Rinkaj Goyal[1(✉)]

[1] University School of Information, Communication and Technology, GGS Indraprastha University, Dwarka, New Delhi, India
rnradon17@gmail.com, rinkajgoyal@gmail.com
[2] Covenant University, Ota, Nigeria
sanjay.misra@covenantuniversity.edu.ng

Abstract. Due to the growing number of articles published every year as the research output, it is imperative to analyze their impact on future and current research. In this cito-analytical study, we use two publicly available citation datasets, i.e., arXiv's High-Energy Physics Citation Theory Network and Cora Citation Network. This study employs different macro-meso-micro level indicators such as K-cores, centrality measures, and clustering coefficient in identifying relevant network characteristics and establishing their inter-relationships to determine impactful research. While the meso-level feature identifies the type of citation network, the micro-level indicators (centrality measures) help in recognizing the individual node (research paper) strength and macro-level statistics comments upon the global network characteristics. The current exposition empirically demonstrates the relevance of using macro-meso-micro level statistics together as the unit in determining influential and significant research output. While previous researchers have independently used these metrics in other academic networks, we, however, showed their importance and inter-relationship using an integrationist approach in citation networks.

Keywords: Scientometrics · Social media analytics · Citation networks · Impact analysis

1 Introduction

In recent years, pattern exploration in the scholarly and scientific literature community to examine the knowledge flows, knowledge networks, and research impact has gathered a considerable interest and curiosity from the researcher's community [44]. Due to the rapid growth of the number of articles being published globally, information scientists have started analyzing such academic networks by focusing on collaborations, citations, and co - citation networks using Social Network Analysis (SNA) techniques [7,18,33,34].

This work was supported by the FRGS [Grant number: FRGS-2018-60-1115], Guru Gobind Singh Indraprastha University, New Delhi.

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11622, pp. 340–356, 2019.
https://doi.org/10.1007/978-3-030-24305-0_26

These techniques take the complex network of academic activities (Citation Network, Co-authorship Network, Repository Analysis, etc.) into consideration to conceptualize relevant theories and develop models using relational processes or structural outcomes of some key concepts.

In academia, the citations are an integral part of the process of referencing and acknowledging the information gathered from books, papers, or web sites. Researchers' use citations as one of the metrics to determine the popularity and importance of an article. Citation network is a complex directed graph in which the articles are defined as nodes of the graph and the new paper citing the studies of old papers form a directed edge between them. In an attempt to evaluate how specific areas have evolved and matured, different efforts have been made to analyze the emergence of principal authors and the types of papers that are currently influencing the field [8].

Studies that examine the trends in scientific citations using graph-theoretic and SNA techniques are called 'cito-analytical' studies or 'citations analysis'. Previous studies have investigated academic citation networks to describe their diverse patterns, clustering of nodes, structural cohesiveness, and macro, meso, micro-level network properties to determine the structure of such networks as a function of the detail level [16,27,36].

Prior studies in this area either focus on macro-level statistics to investigate the network's overall structure or on micro-level indicators to obtain the vertex importance for co-citation or co authorship networks (Table 1). We, however, in this study empirically identify the importance of macro-meso-micro indicators (a Social Network Analysis characteristic) in the field of cito-analytics.

Table 1. Previous studies dealing with macro-meso-micro features of social network analysis.

S.No	Study	Measures
1	Luukkonen et al. [25]	Macro (Frequency Distribution)
2	Glänzel et al. [15]	Macro (Frequency Distribution and Empirical Density)
3	Guardio et al. [17]	Macro and Meso (Degree Distribution, Clustering)
4	Yan and Ding [42]	Meso (Clustering)
5	Bibi et al. [4]	Micro (Centrality)
6	Zare et al. [43]	Micro (Centrality)
7	Bollen et al. [5]	Micro (Centrality)

2 Method and Materials

2.1 Networks Following Power-Law

In the field of statistical analysis, power-law provides a functional relationship between the two metrics in consideration. A relative change in one metric resonates as a proportional relative change in the other metric. This change is independent of the initial size of those metrics. Thus, in other words, one metric varies

as a power of the other. During the network evolvement, the dis-proportionality of the vertices increases because of the formation of new links over the nodes which already have a large number of edges. This thereby results in a heavy-tailed degree distribution graph of the network. Any network following such characteristics is termed as a scale-free network [2]. The term scale-free network refers that magnifying on any part of the distribution of the network would not change its shape because of the presence of a few highly connected vertices and a large number of weakly connected nodes.

2.2 Influencing Factors for Cito-Analytical Studies

Freeman [11] developed centrality measures based upon the betweenness properties of the nodes in the network. He further introduced four types of centralities: closeness, betweenness, degree, and eigenvector to be used in the domain of social networks. In citation networks, we use the same centrality measure to determine the structural importance and impact of the articles in the network.

Closeness Centrality. It determines the direct or indirect closeness of a node with respect to all other nodes in the network [24]. The value of closeness centrality for a node v is defined as the sum of the reciprocal of distances to the node v to all other nodes in the network.

$$C_c\,(v) \;=\; \Sigma\,\frac{1}{d_{v,t}} \tag{1}$$

where $d_{v,t}$ denotes the shortest distance between node v and t.

Betweenness Centrality. Betweenness centrality quantifies how much a node controls the communication between all pairs of vertices in the network by acting as a bridge between the two randomly chosen vertice's shortest path [19,24]. It is calculated as the ratio of the number of shortest paths passing through the given vertex to the total number of shortest paths.

$$C_b\,(v) \;=\; \Sigma_{i\,\neq v\neq j}\,\frac{\sigma_{ij}\,(v)}{\sigma_{ij}}, \tag{2}$$

where $\sigma_{ij}(v)$ is the number of those paths that intersect node v, and σ_{ij} is total number of shortest paths from node i to node j in graph G.

Degree Centrality. Degree Centrality defines the number of ties a node has with its immediate neighboring nodes in the network, i.e., the number of edges that are adjacent to the given vertex [19,24]. Degree centrality is computed as

$$C_d\,(v) \;=\; \frac{k_v}{n-1} = \Sigma_{j\in G}\,\frac{a_{vj}}{n-1} \tag{3}$$

where k_v is the degree of the node, n is the total number of the nodes in the network, and a_{vj} is the number of nodes j that are adjacent to node v.

PageRank. PageRank algorithm is the part of Google search engine to rank web sites based upon the direction of the links. Brin and Page [6] proposed it as an alternative of Eigenvector Centrality, which works only with undirected networks.

PageRank has constantly been used for cito-analytical studies to determine the importance of cited nodes by measuring the impact of citing vertices like authors, articles, journals, etc. [39]. PageRank algorithm measures the popularity and importance of a scholar from a citation perspective by computing the ranking of the authors depending upon the network of the incoming links.

2.3 Dataset and Description

In this paper, we consider two publicly available cito-analytic data sets: arXiv's High-Energy Physics Citation Theory Network [12,22] and Cora Citation Network [1,21,37]. We used datasets of sufficient sizes for conclusion stability. These datasets have more nodes and edges in comparison to DBLP dataset used in other studies. These are directed networks with each node representing a scientific paper ID. In the given set U consisting of node ID, the citing relation is given as $R \subseteq U \times U$, $uRv \equiv v\ cites\ u$ which provides a citation network $N = (U, R)$ [3]. Alternatively, the directed edge $(A \rightarrow B)$ denotes that publication A cites publication B. The basic description of the two data sets is as follows:

1. **arXiv's High-Energy Physics Citation Theory Network:**–The data set contains citations from e-print arXiv and covers the period of papers from January 1993 to April 2003 (124 months). arXiv was launched on August 14, 1991, and thus represents essentially the complete history of its HEP-TH section. The dataset consists of 27,770 articles (nodes) with a total of 352,807 citations (edges).
2. **Cora Citation Network:**– This consists of computer science papers and references gathered through crawling and automatic parsing of the bibliographies of the papers. This data set includes 23,166 nodes and 91,500 edges.

Henceforth, these datasets will be referred as Network-1 and Network-2 respectively.

3 Results and Discussions

Table 2 illustrates the work flow behind the cito-analytical study that is followed.

3.1 Macro-Level Statistics

We compute the macro-level statistics to determine the global network characteristics of each network (data set). Table 3 describes the network metrics such as the number of nodes (scientific papers), edges (citations), average degree connectivity (average of the nearest neighbor of nodes with degree k), and network density (proportion of the direct links in the network with respect to the total

ties possible) of each network. We further use Degree Distribution, Gini Coefficient, Lorenz Curve, Giant Component Size, and K-cores to identify the global properties of the considered networks.

Table 2. Research methodology for cito-analytical study.

Property	Method	Indicator type
Verify real-world network	(a) Clustering Coefficient	(a) Meso-level
Verify scale-free network	(a) Degree Distribution	(a) Macro-level
	(b) Frequency Distribution on Centrality Measures	(b) Micro-level
Evaluate network distribution and reachability of nodes	(a) Lorenz Curve and Gini Coefficient	(a) Macro-level
Identify key paper in the network	(a) K-Cores	(a) Macro-level
	(b) Centrality Measures	(b) Micro-level
Measure the redundancy of metrics used for ranking articles	(a) Kendall's Concordance Based Correlation Coefficient for Centrality Metrics	(a) Micro-level

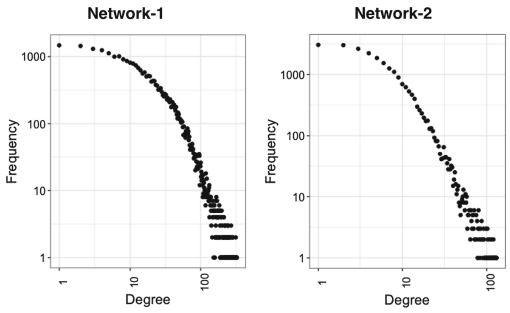

Fig. 1. Degree distribution

Degree Distribution. Since degree distribution tells us about how often an event of a given size occurs, we focus on the relationship between the activity of two users citing each others article. Figure 1 represents that the networks used in this study are scale-free networks because degree distribution exhibits a power-law characteristic (see Sect. 2.1). The fraction of $P(k)$ of nodes having k connections with other nodes of the network is defined as $P(k) \sim k^{-\gamma_\kappa}$, where $-\gamma_\kappa$ is the scale-free degree exponent of the network [29].

Lorenz Curve and Gini Coefficient. Corrado Gini [14] proposed a coefficient (Gini Coefficient) to measure statistical depression, which is commonly used to calculate inequality of wealth and income distribution in economics sector.

However, it can also be used to measure other distributions as well, including quantification of social, or scientific collaboration in researchers network [38]. The coefficient (G) is calculated as the area between the Lorenz Curve of the given distribution and a straight line drawn at a 45° angle from the origin ($G = A/(A+B)$, where A is the area between the straight line drawn at a 45° angle from the origin and Lorenz Curve, and B is the area under the Lorenz Curve). Gini coefficient lies in between 0 to 1 and with a low value tending towards 0 indicates a more equal distribution while a high value tending towards 1 illustrates a more unequal distribution among the parts.

We calculate the Gini Coefficient by plotting a Lorenz Curve (Fig. 2) between the fraction of actors (nodes) and their edges. The coefficient value obtained for Network-1 is 0.570 or 57% and that for Network-2 is 0.521 that is 52.1%. Following this value, we observe that the latter network is slightly more equally distributed amongst the two since the value of Gini Coefficient for Network-2 is less than Network-1 (0.570 and 0.521 for Network-1 and Network-2 respectively).

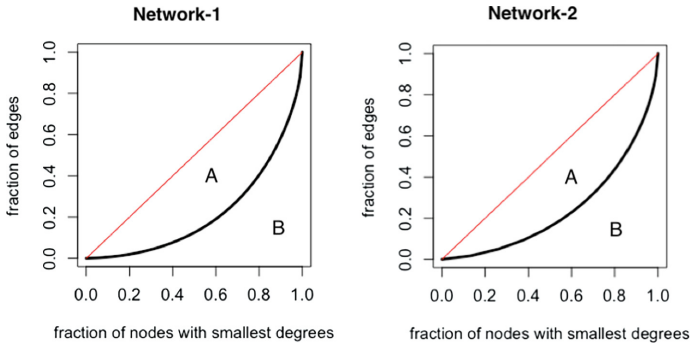

Fig. 2. Lorenz curve

Table 3. Macro-level statistics.

Dataset	Total nodes	Total edges	Average degree connectivity	Network density	Giant component size
Network-1	27770	352807	0.000174	0.000457	7464
Network-2	23166	91500	0.00116	0.000170	3991

Giant Component Size. The largest strongly connected component of a directed graph is the largest association of nodes in the network in which all the nodes are reachable by following the direction of the edges [32]. It can be used to represent the main core of research activities [10]. Therefore, we measure the portion of the network that subjects to be the part of the main core. The GCC of the Network-1 contains 7464 nodes (refer Table 3) that compute to be 26.89% of the total nodes of the network, while for Network-2 is 17.92% of the total actors of the network; thus depicting that a small portion of the network is strongly connected.

K-Cores. For the directed citation networks used in this study, we use another measure, K- Core. Seidman first proposed it for unweighted graphs and since then it has found widespread in the analysis of real networks in social sciences [35,41]. K-core measure identifies the number of times an author is cited. This aids in generating the cores of key articles leading to the impact analysis in the network. Each node present in the k-core has a degree no less than k. To calculate the number of authors in each core, we consider the values of k ranging from 1 to the maximum degree of the node that is 37 in case of Network-1 and 15 for Network-2. For the directed cito-analytic dataset, the K-Core measures the innermost core of the highly cited authors in the network [40]. Table 4 depicts the number of authors in each core k for the Network-1 and Network-2 respectively.

Table 4. Total nodes present in each kth core of Network-1 and Network-2

Network-1				Network-2	
K-Core	Nodes	K-Core	Network	K-Core	Network
1	27769	20	7749	1	23166
2	26099	21	7289	2	19815
3	24395	22	6825	3	16138
4	22764	23	6410	4	12744
5	21188	24	6020	5	9673
6	19705	25	5661	6	7026
7	18429	26	5227	7	4787
8	16936	27	4832	8	3086
9	15488	28	4394	9	1684
10	14406	29	4011	10	1089
11	13303	30	3515	11	535
12	12215	31	2938	12	96
13	11366	32	2406	13	32
14	10668	33	1728	14	30
15	10003	34	1364	15	27
16	9500	35	112		
17	9002	36	108		
18	8601	37	52		
19	8175				

3.2 Meso-Level Features

Meso-level features use clustering as a central mechanism to identify the structural characteristics of the network. Table 5 shows the meso-level features such as the average clustering coefficient and transitivity. The value of the average global

clustering coefficient is close to 0 that is 0.312 and 0.265 for both Network-1 and Network-2 respectively. The global clustering coefficient lies between 0 to 1 where value closer to 1 depicts that the network is fully connected and not a real network, while a real-world network has the value significantly less than 1 [30]. It is measured as:

$$C = \frac{3 \times Number\ of\ triangles}{Number\ of\ connected\ triples\ of\ vertices} \tag{4}$$

This measure can be applied to both directed and undirected networks, and refers to the concept of "fraction of transitive triples" used in sociology [41].

Table 5. Meso-level feature

Dataset	Average clustering coefficient	Transitivity
Network-1	0.312	0.132
Network-2	0.265	0.127

3.3 Micro-level Indicators

After employing the macro-level statistics and meso-level measures, we further investigate the micro-level indicators of both the networks to analyze the individual node's power for their network impact analysis. The four centrality measures described in Sect. 2.2 are computed for determining the influencing nodes of the network. Figures 3 and 4 depict the frequency distributions of Closeness, Degree, Betweenness and PageRank Centralities for Network-1 and Network-2 respectively. The vertical axis depicts the frequency corresponding to the metric values on the horizontal axis. The frequency distribution of Degree, Betweenness, and PageRank Centralities for both the datasets again follow a power law where the majority of the articles have a low value of centrality while a very small amount of articles possess a higher value (refer Sect. 2.1). While the Closeness Centrality frequency distribution follows a normal curve.

Thereafter, using the same micro-level parameter, i.e. centrality measures, we examine the key articles of influence. Table 6 and 7 describes top 30 article IDs based on different centrality measures and citation counts. While centrality measures give the important nodes responsible for information flow as described in Sect. 2.2, citation count allows us to determine the impact of an article according to the number of times it has been cited by other researchers. To determine the key article for impact analysis, we can not focus solely on a single metric. Table 6 and 7 both show some inconsistency with respect to the rankings of centrality measures and citation count. The most cited article in Network-1 with article ID 9711200 does not even come under the top 30 ranking articles based upon Closeness and Betweenness Centrality. It has a relatively lower rank (rank 5) under the PageRank metric for computing the influencing node. Whereas, few articles such

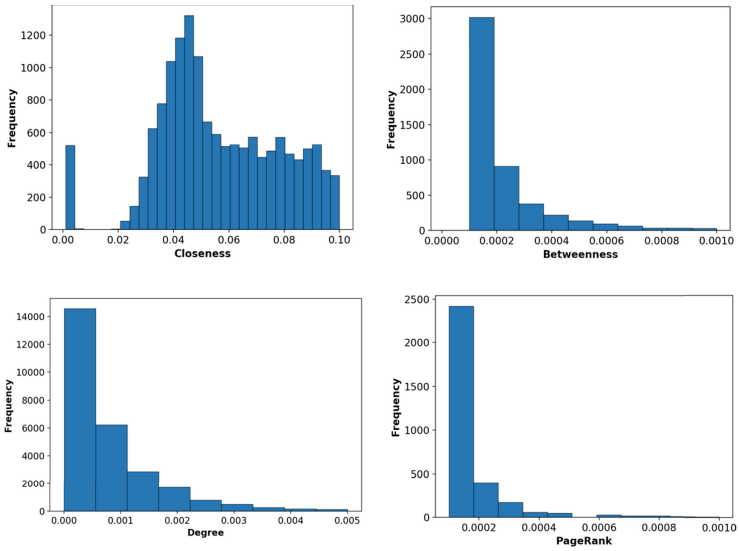

Fig. 3. Frequency distribution based on closeness centrality, betweenness centrality, degree centrality and PageRank for Network-1

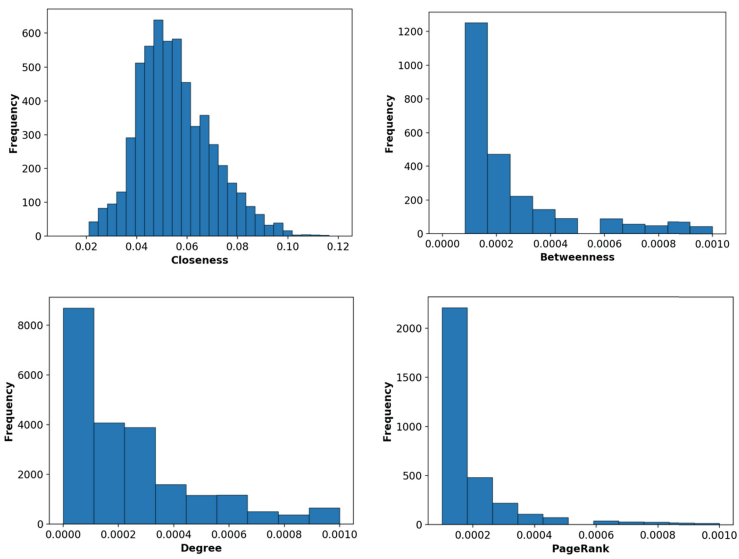

Fig. 4. Frequency distribution based on closeness centrality, betweenness centrality, degree centrality and PageRank for Network-2

Table 6. Network-1's top 30 article ids based on different centrality measures and citation count

Rank	CC based[a] Article ID	DC based[b] Article ID	BWC based[c] Article ID	PR based[d] Article ID	Most cited Article ID
1	9407087	9711200	9905111	9407087	9711200
2	9503124	9802150	9810008	9503124	9802150
3	9510017	9802109	206223	9510017	9802109
4	9401139	9905111	9509140	9402044	9407087
5	9402002	9407087	9803001	9711200	9610043
6	9410167	9908142	9912210	9410167	9510017
7	9408099	9610043	9902121	9408099	9908142
8	9408074	9510017	9607239	9207016	9503124
9	9510135	9503124	206182	9402002	9906064
10	9402032	9906064	9907085	9610043	9408099
11	9207016	9408099	9904179	9205068	9905111
12	9301042	9510209	9904036	9510135	9510209
13	9305185	9711162	9903033	9201015	9711162
14	9304154	9510135	111093	9205027	9510135
15	9501068	9410167	3136	9304154	9410167
16	9504090	9611050	9905046	9802150	9611050
17	9412184	9601029	204158	9401139	9601029
18	9510209	9602022	9906064	9504090	9602022
19	9205027	9803131	206196	9802109	9603142
20	9610043	9603142	9605009	208020	9803131
21	9602022	9603167	9908105	9305185	9603167
22	9512059	9711165	9908142	9207053	9711165
23	9207053	9412184	302221	9307049	9611230
24	9512077	9401139	110123	9204102	9409089
25	9302103	9611230	108172	9510209	9412184
26	9411149	9602052	9611016	9501068	9703166
27	9510169	9703166	204004	9906064	9401139
28	9402044	9608024	110234	9402032	9411149
29	9602052	9710046	9711200	9301042	9602052
30	9511030	9409089	9906151	9403198	9510169

[a] Closeness Centrality
[b] Degree Centrality
[c] Betweenness Centrality
[d] PageRank

Table 7. Network-2's top 30 article ids based on different centrality measures and citation count

Rank	CC[a] based Article ID	DC[b] based Article ID	BWC[c] based Article ID	PR[d] based Article ID	Most Cited Article ID	Rank	CC[a] based Article ID	DC[b] based Article ID	BWC[c] based Article ID	PR[d] based Article ID	Most cited Article ID
1	1435	660	1435	1219	660	16	2955	1475	1044	1589	1475
2	1219	1435	1561	660	1435	17	1257	190	513	4301	300
3	1248	322	5399	1435	322	18	2689	1452	647	1044	1546
4	290	681	923	689	681	19	654	113	2442	1452	113
5	482	3005	1546	1248	3005	20	1076	491	278	7691	1219
6	11320	124	3427	262	124	21	1218	1219	152	407	491
7	742	352	224	1218	352	22	2442	3717	4074	124	3717
8	352	647	1012	1690	647	23	1710	1640	3342	6108	554
9	4187	1546	1465	3941	1248	24	681	399	1886	3333	399
10	3005	1248	367	681	262	25	7212	554	6613	399	331
11	5399	262	1437	5585	1044	26	689	532	159	647	1481
12	1589	1044	9053	322	1646	27	8953	1924	4004	168	144
13	1452	1646	2637	278	1407	28	576	331	8966	2660	1182
14	2552	1407	2582	2988	190	29	1407	1481	17151	4392	532
15	4392	300	1640	2827	1452	30	6725	497	7922	3778	278

as article ID 1435 in Network-2 (Table 7) are consecutively high ranked attaining a position of 1, 2, 1, 3 and 2 in Closeness, Degree, Betweenness, PageRank and Most Cited Article ID metrics respectively. Paper with high degree centrality is important as it suggests that there is an exchange of information with a large number of different papers. Similarly, articles with high closeness centrality rank imply that there is an information flow in the nearby neighbors of the connected component of the node under consideration. Moreover, nodes with high betweenness centrality value have more control over the flow of information as the majority of the information passes through that bridge node. Consecutively, since PageRank centrality takes link's direction into account, it suggests that the information can flow only in one direction that is from the cited paper to the article citing it. Hence, the node with the high PageRank centrality means that there are many articles pointing to it or a paper with a high PageRank is pointing to it thereby the node has an influence extending into the wider network, delimiting the direct connections [9].

Various metrics such as different centrality measures have been studied rigorously for the identification of the important nodes within a complex network [11,31]. Amongst these metrics, K-Cores decomposition (refer Sect. 3.1) is considered as one of the well-established measures to identify the node's influencing power in a network. Kitsak et al. [20] argued that K-Core is a better metric to

Table 8. Centrality metric rankings of all the Article ID's present in the Kth core of Network-1

Article ID in Kth core	Rank in CC[a]	Rank in DC[b]	Rank in BWC[c]	Rank in PR[d]	Article ID in Kth Core	Rank in CC[a]	Rank in DC[b]	Rank in BWC[c]	Rank in PR[d]
110242	10205	147	12950	746	202157	10421	649	4544	2465
201081	10233	132	5332	839	202186	10399	662	757	2745
112044	10255	93	298	649	202190	10485	2448	9923	4786
202021	10235	7213	314	559	203018	10478	4443	245	4123
202109	10368	130	458	978	203080	10464	1807	7929	4156
202153	10426	787	5709	2595	203082	10448	918	382	3943
202179	10431	429	2483	2549	203231	14948	829	5228	3801
203028	12015	1037	1680	3788	204002	15041	2300	6078	9372
203093	10496	1376	4573	4481	204026	14220	1867	4045	6156
203124	10450	385	42	3935	204033	14998	1249	171	5433
203134	10498	1113	869	4554	204146	15019	611	1210	2886
203140	14942	737	2388	4994	204234	15079	989	1209	12093
203164	10810	1248	10820	4347	205008	15623	2301	12402	15328
203186	12083	1225	3434	5022	205109	15154	1922	10550	16613
203229	10503	955	2362	4293	205134	15149	1058	5588	10466
203249	14925	1281	4640	4645	205200	14364	550	92	9671
203255	14938	1806	5370	3792	205266	16513	1547	8873	15045
203257	14954	1153	2962	5411	205279	16510	830	3654	5038
204001	14981	2668	4353	9135	205303	16536	1589	8639	17366
204004	9228	597	27	4012	205311	16509	638	2332	4736
204054	14951	333	681	3930	206010	16559	738	3905	11415
204174	15048	1757	4828	9339	206029	16557	990	7267	19814
204196	15035	1645	2232	8314	206062	16573	875	4595	16536
205067	15066	769	3634	12121	207037	16586	1282	4268	14128
205089	15022	391	725	2747	207175	18090	1719	13329	22706
202111	9170	373	564	1971	211178	17717	245	6382	19583

identify the influential vertices in the network as compared to a node having the highest degree in the network. The nodes that are the part of the last core of the network are considered as the most influential vertices. A node even if it has a lower degree centrality value, but is placed strategically in the core of the network can disseminate information at a higher extent. Furthermore, the authors also stated that K-Core is more efficient than some of the other centrality measures which require high computations like closeness and betweenness centrality.

Table 9. Centrality metric rankings of all the Article ID's present in the Kth core of Network-2

Article ID in Kth Core	Rank in CC[a]	Rank in DC[b]	Rank in BWC[c]	Rank in PR[d]	Article ID in Kth Core	Rank in CC[a]	Rank in DC[b]	Rank in BWC[c]	Rank in PR[d]
155	648	86	639	443	2880	1065	107	636	42
223	278	539	1279	7110	3174	1031	67	642	35
273	1221	130	666	231	3529	6750	1038	1095	821
375	11048	665	13952	13952	4112	133	1649	774	1898
399	112	24	583	25	4113	6334	2005	775	2467
917	858	577	776	2939	4897	7434	436	721	497
1125	1166	125	680	65	5555	1601	116	674	71
1472	765	254	736	301	5680	12738	15188	2760	15188
1473	331	48	638	124	7800	7397	1698	753	861
1546	5	9	346	34	7804	6812	2273	759	1230
1770	587	176	754	814	8056	13717	917	15942	15942
1865	920	104	921	222	9202	1708	190	720	237
2376	3065	780	1294	7589	10119	4294	1721	11499	6263
2620	258	489	745	984					

We thus represent the article IDs found in the kth core that is 37 for Network-1 and 15 for Network-2 (refer Tables 8 and 9) and compare the rankings of these articles with the rankings obtained from different centrality metrics.

From Table 4 we observe that the majority of the article ID's present in the last core of both the networks are not the top ranking nodes obtained from the centrality metrics [26]. Hence, to understand the relationship between micro-level statistics using centrality measures and macro-level indicators using K-Cores, we calculate the Kendall's Concordance Based Correlation amongst all the centrality metrics and K-Core Index Value (normalised by the maximum k value [13]). Tables 10 and 11 enlist the values of Kendall's Concordance Based Correlation Coefficient. While previous researchers have used this metric in co-authorship networks to determine the relationship between centrality measures [4], we use it to establish the relationship between Micro-level Indicators (Centrality Measures) and Macro-level Statistics (K-Core Index) in cito-analyical studies. Though besides Kendall's Concordance Based Correlation Coefficient, other similar measures also exist such as Pearson's Product-Moment Correlation Coefficient, and Spearman's Rank-based Correlation Measure [23], however, we use Kendall's Rank Order Correlation Coefficient to discern the importance of micro-macro-level indicators because it has a more intuitive interpretation. Further, the metric gives higher accuracy as compared to Spearman's Rank-based Correlation on smaller samples [28]. The resultant values range between -1 to 1, where values close to 1 indicate strong agreement between the two rankings, while -1 indicates strong disagreement.

Table 10. Kendall's rank-based correlation coefficient values for Network-1

	CC[a]	DC[b]	BWC[c]	PR[d]	KCI[e]
CC[a]		0.365	0.417	0.606	0.332
DC[b]			0.543	0.319	0.863
BWC[c]				0.366	0.503
PR[d]					0.235
KCI[e]					

[e]K-Core Index

Table 11. Kendall's rank-based correlation coefficient values for Network-2

	CC[a]	DC[b]	BWC[c]	PR[d]	KCI[e]
CC[a]		0.472	0.645	0.798	0.415
DC[b]			0.530	0.439	0.852
BWC[c]				0.604	0.465
PR[d]					0.332
KCI[e]					

From Tables 10 and 11 the highest correlation amongst the centrality metrics can be seen in Closeness-PageRank Centrality, while the least correlation exists for Degree-PageRank Centrality for both the datasets. Whereas, comparing the Kendall's Concordance Based Correlation Coefficient amongst the macro-micro metrics, we see that Degree Centrality-K-Core Index has the highest value (0.863 for Network-1 and 0.852 for Network-2). This suggests that Degree Centrality and K-Core Index Value have a high relationship tendency. Higher Kendall's Correlation Coefficient measure in terms of ranking of nodes provides a greater similarity in the order of ranking as can be observed in Closeness-PageRank Centrality Metrics. When two metrics are highly correlated, the information gain from their perspective becomes redundant while uncorrelated metrics can provide unique insights regarding the network. Thus in order to determine the most impactful articles of the network, it is better to use both macro (K-Core) and micro (Centrality Measures) indicators together.

4 Conclusions

The analysis of academic networks such as co-authorship, co-citation, bibliographic, citation, repository networks using social network analysis (SNA) techniques has recently attracted considerable interests of researchers community. Prior studies in this area show the effectiveness of SNA based approaches in revealing meaningful patterns in these networks. In this study, we empirically show the significance of using macro-meso-micro level statistics together while analysing citation networks. Initially, the type of citation network can be identified by applying meso-level features (clustering) to ascertain that whether the network is a small-world or a real-world network. The type of network can also be verified using macro-level statistics (degree distribution) and the frequency distribution of micro-indicators (centrality measures). Through the empirical analysis of two citation datasets: arXiv's High-Energy Physics Citation Theory Network and Cora Citation Network, we observed that it is essential to consider both macro-level statistics (K-Cores) and micro-level indicators (Centrality Metrics) to find the most impactful nodes of the network. Since the majority of the nodes present in the last k^{th}core of the network which is considered to be the

most impactful core of the network are not the top influential vertices according to the different centrality measures, it is imperative to consider both the metrics while finding the influential nodes of the network.

Furthermore, Kendall's Correlation measure reveals that amongst the centrality measures, Closeness-PageRank Centrality Metric has the highest correlation value, while Degree Centrality and K-Core has the maximum correlation when considering the relationship amongst the centrality metrics, or between the centrality measures and K-Cores respectively of the two given networks. Thus, instead of Closeness Centrality which is a computationally high centrality metric, a computationally low centrality measure, i.e., PageRank can be used to calculate the influential nodes in the two citation network.

References

1. Cora citation network dataset - KONECT, September 2016. http://konect.uni-koblenz.de/networks/subelj_cora
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
3. Batagelj, V.: Efficient algorithms for citation network analysis. arXiv preprint cs/0309023 (2003)
4. Bibi, F., Khan, H., Iqbal, T., Farooq, M., Mehmood, I., Nam, Y.: Ranking authors in an academic network using social network measures. Appl. Sci. **8**(10), 1824 (2018)
5. Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R.: A principal component analysis of 39 scientific impact measures. PloS ONE **4**(6), e6022 (2009)
6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
7. Crane, D.: Invisible Colleges: Diffusion of Knowledge in Scientific Communities (1972)
8. Dawson, S., Gašević, D., Siemens, G., Joksimovic, S.: Current state and future trends: a citation network analysis of the learning analytics field. In: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, pp. 231–240. ACM (2014)
9. Ding, Y., Yan, E., Frazho, A., Caverlee, J.: Pagerank for ranking authors in co-citation networks. J. Am. Soc. Inf. Sci. Technol. **60**(11), 2229–2243 (2009)
10. Fatt, C., Ujum, E., Ratnavelu, K.: The structure of collaboration in the journal of finance. Scientometrics **85**(3), 849–860 (2010)
11. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977)
12. Gehrke, J., Ginsparg, P., Kleinberg, J.: Overview of the 2003 KDD cup. ACM SIGKDD Explor. Newsl. **5**(2), 149–151 (2003)
13. Giatsidis, C., et al.: A k-core decomposition framework for graph clustering. arXiv preprint arXiv:1607.02096 (2016)
14. Gini, C.: Variabilità e mutabilità. In: Pizetti, E., Salvemini, T. (ed.) Reprinted in Memorie di metodologica statistica. Libreria Eredi Virgilio Veschi, Rome (1912)
15. Glänzel, W., Debackere, K., Thijs, B., Schubert, A.: A concise review on the role of author self-citations in information science, bibliometrics and science policy. Scientometrics **67**(2), 263–277 (2006)

16. Gondal, N.: The local and global structure of knowledge production in an emergent research field: an exponential random graph analysis. Soc. Netw. **33**(1), 20–30 (2011)
17. Guardiola, X., Guimera, R., Arenas, A., Diaz-Guilera, A., Streib, D., Amaral, L.: Macro-and micro-structure of trust networks. arXiv preprint cond-mat/0206240 (2002)
18. Jones, B.F., Wuchty, S., Uzzi, B.: Multi-university research teams: shifting impact, geography, and stratification in science. Science **322**(5905), 1259–1262 (2008)
19. Kas, M., Carley, L.R., Carley, K.M.: Monitoring social centrality for peer-to-peer network protection. IEEE Commun. Mag. **51**(12), 155–161 (2013)
20. Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identification of influential spreaders in complex networks. Nat. Phys. **6**(11), 888 (2010)
21. Kunegis, J.: KONECT - the Koblenz network collection. In: Proceedings of International Conference on World Wide Web Companion, pp. 1343–1350 (2013). http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-koblenz-network-collection.pdf
22. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 177–187. ACM (2005)
23. Li, C., Li, Q., Van Mieghem, P., Stanley, H.E., Wang, H.: Correlation between centrality metrics and their application to the opinion model. Eur. Phys. J. B **88**(3), 65 (2015)
24. Liu, B.: Social network analysis. In: Web Data Mining. Data-Centric Systems and Applications. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19460-3_7
25. Luukkonen, T., Persson, O., Sivertsen, G.: Understanding patterns of international scientific collaboration. Sci. Technol. Hum. Values **17**(1), 101–126 (1992)
26. Malliaros, F.D., Rossi, M.E.G., Vazirgiannis, M.: Locating influential nodes in complex networks. Sci. Rep. **6**, 19307 (2016)
27. Marchiori, M., Possamai, L.: Micro-macro analysis of complex networks. PloS ONE **10**(1), e0116670 (2015)
28. Meghanathan, N.: Evaluation of correlation measures for computationally-light vs. computationally-heavy centrality metrics on real-world graphs. J. Comput. Inf. Technol. **25**(2), 103–132 (2017)
29. Muchnik, L., et al.: Corrigendum: origins of power-law degree distribution in the heterogeneity of human activity in social networks. Sci. Rep. **5**, 15932 (2015)
30. Newman, M.E.: Models of the small world. J. Stat. Phys. **101**(3–4), 819–841 (2000)
31. Newman, M.E.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
32. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proc. Natl. Acad. Sci. **101**(suppl 1), 5200–5205 (2004)
33. Parreira, M.R., Machado, K.B., Logares, R., Diniz-Filho, J.A.F., Nabout, J.C.: The roles of geographic distance and socioeconomic factors on international collaboration among ecologists. Scientometrics **113**(3), 1539–1550 (2017)
34. Popp, J., Balogh, P., Oláh, J., Kot, S., Harangi Rákos, M., Lengyel, P.: Social network analysis of scientific articles published by food policy. Sustainability **10**(3), 577 (2018)
35. Seidman, S.B.: Network structure and minimum degree. Soc. Netw. **5**(3), 269–287 (1983)

36. Small, H., Griffith, B.C.: The structure of scientific literatures I: identifying and graphing specialties. Sci. Stud. **4**(1), 17–40 (1974)
37. Šubelj, L., Bajec, M.: Model of complex networks based on citation dynamics. In: Proceedings of the WWW Workshop on Large Scale Network Analysis, pp. 527–530 (2013)
38. Subramanian, S.V., Kawachi, I.: Income inequality and health: what have we learned so far? Epidemiol. Rev. **26**(1), 78–91 (2004)
39. Waltman, L., Yan, E.: PageRank-related methods for analyzing citation networks. In: Ding, Y., Rousseau, R., Wolfram, D. (eds.) Measuring Scholarly Impact, pp. 83–100. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10377-8_4
40. Wang, S., Rohe, K., et al.: Discussion of "coauthorship and citation networks for statisticians". Ann. Appl. Stat. **10**(4), 1820–1826 (2016)
41. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)
42. Yan, E., Ding, Y.: Scholarly network similarities: how bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. J. Am. Soc. Inf. Sci. Technol. **63**(7), 1313–1326 (2012)
43. Zare-Farashbandi, F., Geraei, E., Siamaki, S.: Study of co-authorship network of papers in the journal of research in medical sciences using social network analysis. J. Res. Med. Sci. Off. J. Isfahan Univ. Med. Sci. **19**(1), 41 (2014)
44. Zhao, D., Strotmann, A.: Analysis and visualization of citation networks. Synth. Lect. Inf. Concepts Retrieval Serv. **7**(1), 1–207 (2015)