Int. j. inf. tecnol. (March 2022) 14(2):1061–1073
https://doi.org/10.1007/s41870-019-00344-4

ORIGINAL RESEARCH

Empirical analysis of synthetic and real networks

Pranav Nerurkar[1] · Madhav Chandane[1] · Sunil Bhirud[1]

Received: 13 April 2018 / Accepted: 31 July 2019 / Published online: 6 August 2019

Abstract With increasing digitization a wide variety of systems from diverse domains such as computer science, transportation, social science have become available in the form of networks. It is argued that to understand complex systems a deep understanding of the networks behind them is needed. A network theoretic perspective provides valuable insights into the structure and trends of systems. Datasets belonging to different domains have their own unique features and behavioural trends and the current inquiry aims to highlight this. In this inquiry, a comprehensive analysis of synthetic and real-world published benchmark data-sets, evaluation methods, and open source projects is performed. The aim is to provide novice and expert users with tools for algorithmic designs and methodologies. Empirical studies are used to compare the performance of network theoretic tools on common data-sets. Finally, limitations of the network perspective on systems are listed and research directions to facilitate future study are elaborated.

Keywords Statistical analysis · Social networks · Network structure

1 Introduction

A network in graph theory is a tuple $G = (V, E)$ where V is a (finite) set of vertices and E is a (finite) set of edges [1]. Each edge is either a one or two element subset of V. When a system is represented in the form of a network, the entities of the system are denoted as the nodes of the network. If a pair of entities have an interaction or a relationship with each other then this is denoted as an edge between the entities [2]. For instance, if the transportation network of country is represented as a network, the vertices (nodes) of this network would be the different cities of a country. The edges of the network would denote the presence of a direct transport link between one or more cities. Depending on the type of interaction this transportation network aims to capture, the edges could be directed or un-directed and weighted or un-weighted. Undirected edges would represent presence or absence of a direct route between the cities. Weighted edges would represent the volume of traffic between cities in the transportation network. Thus, graph representations offer the flexibility to capture different aspects of systems [3].

Network representation of a system also allows the application of Network science for its analysis. Network science concepts have their roots in graph theory, a fertile field of mathematics. Graph theory is concerned with proving theorems and developing algorithms that can be applied on arbitrary graphs (irrespective of what the graph models in the real world). What distinguishes network science from graph theory is its empiric [4]. It is this empirical aspect that makes Network analysis interesting [2]. Network science researchers do not study graphs from an abstract point of view but instead study graph representations of real world systems to understand their properties.

✉ Pranav Nerurkar
panerurkar_p16@ce.vjti.ac.in

Madhav Chandane
mmchandane@it.vjti.ac.in

Sunil Bhirud
sgbhirud@ce.vjti.ac.in

[1] Department of CE and IT, VJTI, Mumbai, India

1.1 Background

Usefulness of network representations in the examination of complex systems is illustrated in the origin story of Google [5]. In their seminal paper, L Page et al. [6] argued that identification of authoritative web-pages on the internet could be done by representing the internet as a network (Fig. 1). The web-pages (entities) could be the nodes of the network and web-pages connected by a hyperlink could be shown in the network as nodes linked by a directed edge. In the web-graph that is thus formed, authoritative web-pages would be those with high eigen-vector centrality ranking [5]. The PageRank ranking technique proposed by the authors for information retrieval was based on this intuition [5]. Thus, network science provided a novel approach to analyze the internet. This proved useful in increasing the efficacy of information retrieval on the internet. This analogy shows how a concept of graph theory was used on a network representation model to develop an efficient search engine.

In addition to this, there are several other useful statistical concepts in graph theory which can be applied on network models to draw inferences about the nature of the systems they represent [3, 7]. Measurements of path length, diameter, connectivity, transitivity and density allow inferences on the structural characteristics of systems [1, 2, 4, 8, 9]. Locations of nodes in a network model can be used to draw parallels about the importance of those entities in the overall scheme of the system [2]. For instance, nodes located at the boundaries in a network are important for information exchange with nodes outside the network and nodes located at the center of a network are hubs which keep the network intact and play a key role in information interchange within the network.

In the model given in Fig. 2 edges connect scientists that have coauthored at least one paper. Symbols indicate the research areas of the scientists. As seen in Fig. 2, density of edges between researchers in a particular domain are high compared to density of edges between researchers of different domains. The position of scientists (nodes) in the Santa Fe Institute network provides insights regarding their importance to the research community of the institution. The key takeaway from this example is that graph theory has several concepts that can be applied to networks to understand the behavior or trends in the real-world system they represent [10]. Similarly, data or systems across diverse domains such as computer science, transportation, social science, economics and biology too have been investigated using a network perspective [11].

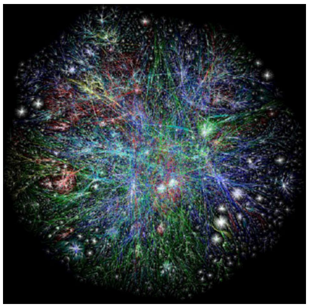


Fig. 1 Internet network. Reprinted figure with permission from www.opte.org

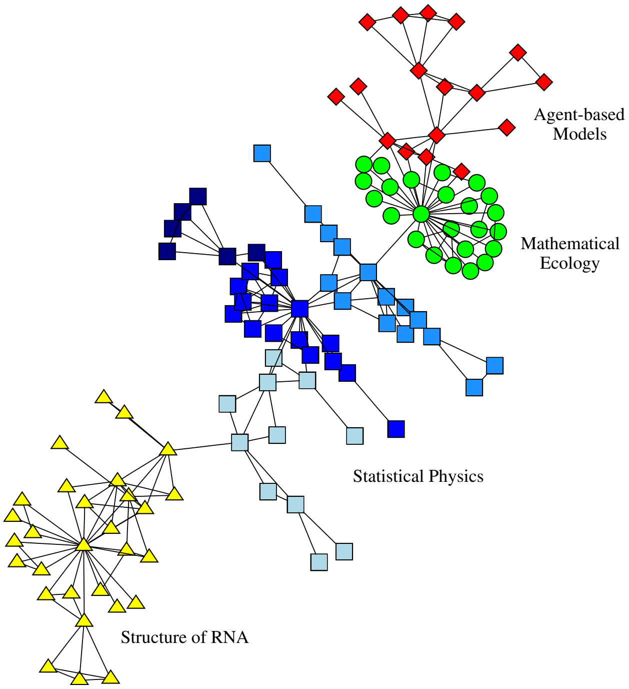


Fig. 2 A network model representing the collaborations between scientists working at the Santa Fe Institute (SFI). Reprinted figure from [2]

1.2 Objective

This inquiry provides statistical analysis of large social network data-sets available on Stanford Network Analysis Project as well as standard synthetic benchmark data-sets such as GN-benchmark, LFR benchmark [12], dynamic LFR benchmark, small world model [13], Erdos Renyi Random Graph [14], Barabasi Albert Preferential Attachment graph and Forest Fire Graph [11]. The focus is on detailed analysis of these data-sets to uncover behavioural and structural characteristics. Domain specific phenomenon that occur in these data-sets are explained using intuition and empirical evidence is offered to support them.

The rest of the paper is organized into three main sections. Section 2 presents a review of literature which includes network theoretic concepts [15, 16]. Section 3 presents the detailed descriptions of the data-sets and their specifications. In case of synthetic benchmarks the specifications are used as provided in research papers where the generative models have been proposed. Section 3 also hosts a critical discussion of the results obtained and an explanation for the same. The concluding remarks of this inquiry are presented in Sect. 4.

2 Review of literature

2.1 Empirical analysis of social networks

Statistical features of social networks such as number of nodes n, average degree d, diameter D and average path length L of graphs can be used to infer the rate of diffusion processes amongst nodes. Shorter diameters and path lengths would indicate a faster diffusion of information Eq. 1. Modularity is a measure of the community structure in the graph and ranges from [-1, 1] for pure random graphs to perfect community structure. Reciprocity index of a graph is the proportion of mutual connections in a graph and ranges from [0, 1]. Assortativity measures homophily in a graph. Transitivity is defined as the probability of $e_{j,k}$ in a graph where for nodes i, j, k the edges $e_{i,j}$ and $e_{i,k}$ exist.

$$D \propto \frac{ln(n)}{ln(d)}; L \propto \frac{ln(n)}{ln(d)}. \tag{1}$$

Adhesion or edge connectivity E for connected graph G is the minimum number of edges $\lambda(G)$ whose deletion from a graph G disconnects G.

Diameter is the length $max_{(u,v)} d(u, v)$ of the "longest shortest path" (i.e., the longest graph geodesic) between any two graph vertices (u, v) of a graph, where $d(u, v)$ is a graph distance.

Average path length $L = \sum_1^E (G) \frac{d(u,v)}{E(G)}$

Degree distribution of graph $P(k) = \frac{n_k}{n}$ is fraction of nodes in the network with degree k i.e. n_k where n is the Graph order.

Modularity Q is a measure of quality of separation of different vertex types from each other.

$$Q = \frac{1}{2m} * \sum A_{vw} - \frac{k_v * k_w}{2m} * \delta(c_v, c_w) \tag{2}$$

where, m is the number of edges in Eq. 2; A_{vw} is the element of the A adjacency matrix in row v and column w; k_v, k_w is the degree of v and w; c_v, c_w is the type (or component) of v and w; $\delta(c_v, c_w)$ is 1 if $c_v = c_w$ otherwise 0.

Verification of power laws $f(k) \propto k^{-\alpha}$ in networks related to eigen-vectors distribution $x_1, x_2, \ldots x_{20}$, component distribution $C_1, C_2, \ldots, C_k$

Assortativity measures the level of homophily of the graph.

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2} \tag{3}$$

where, q_k is the number of edges leaving the node, other than the one that connects the pair j, k; σ_q is the standard deviation of q in Eq. 3; e_{jk} is the refers to the joint probability distribution of the remaining degrees of the two vertices.

Graph density (G_D) is the number of edges present graph G amongst all possible edges in G. G_D for undirected and directed graphs is given by below Eqs. 4 and 5 respectively.

$$\frac{2|E|}{|V|(|V| - 1)} \tag{4}$$

$$\frac{|E|}{|V|(|V| - 1)}. \tag{5}$$

Reciprocity ρ as given in Eq. 6 is the measure of the likelihood of vertices in a directed network to be mutually linked.

$$\rho = \frac{\sum_{i \neq j(a_{ij} - \overline{a})(i \neq j(a_{ji} - \overline{a})}}{sum_{i \neq j(a_{ij} - \overline{a})^2}}. \tag{6}$$

The betweenness centrality of a node $g(v)$ is given by the Eq. 7:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st} v}{\sigma_{st}} \tag{7}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

McGlohon et al. [7] observed that In and Out degree distributions of social networks to follow a power law, the weights of edges $w_{i,j}$ between two nodes with weights w_i, w_j follow a relation given by Eqn 8, the distribution of component sizes of the social networks follow a power law, number of edges $E(t)$ and total weight of a graph $W(t)$, at time t follow a power law and the first twenty eigenvalues of a social network follow a power law. In this inquiry the validity of the phenomena is tested on diverse datasets both synthetic and real.

$$w_{i,j} \propto (\sqrt{(w_i - w_{i,j}) * (w_j - w_{i,j})})^{\gamma}. \tag{8}$$

Tan et al. [17] observed that indegrees of nodes matched their outdegrees and presence of densely connected hubs of high degree nodes in social networks. The authors proposed other measures for social network analytics such as change in group clustering coefficient with group size, relation of out degree of a node with the number of groups to which it belongs, relation of clustering coefficient with degree of a node and degree of assortativity and reciprocity in social networks. Freidman et al. [18] observed higher graph density in online social networks such as Facebook, Twitter etc compared to social networks of other domains. Gewerc et al. [19] observed that node centrality and graph density, provide insight into how friendship would be formed in an online social network, how peers interact and how all this affects the network evolution over time. Hoff et al. [8] have argued that probability of relation between actors depends on the position of actors in an unobserved social space. Dwyer et al. [20] have applied measures of concern for privacy and trust to members of different sites, and looked for variances in behaviour. Gomez et al. [3] have observed that social networks of various domains present common features of traditional social networks such as a presence of a large connected component, small average path length and high clustering, but differs from them in showing moderate reciprocity and neutral assortativity by degree. Using K–S Goodness of fit test, the authors show that the degree distributions are better explained by log-normal instead of power-law distributions. Another interesting observation is a high reciprocity in links in the online social network Slashdot.

The literature review highlights works that have uncovered new trends or phenomena in social networks. The focus of the current inquiry is to further this line of research i.e. use of network science to uncover patterns or trends seen in systems belonging to various domains. A second aim is to provide data driven analysis of various interesting data-sets obtained from Stanford Network Analysis Project. This shall provide novice and expert users with tools for algorithmic designs and methodologies. However, there are certain issues of the network representation models. These issues are elaborated along with possible solutions to them.

2.2 Synthetic data-sets

2.2.1 *Barabasi Albert (BA): preferential attachment*

Intuition behind this model is that when new nodes enter a network they prefer to attach to popular nodes (high indegree) over others. The generative process of the network initializes with a single node. Then at each time step a node is created that initiates outgoing edges to nodes existing in the system. The probability that an existing node i is chosen by an outgoing edge is given by Eq. 9:

$$P[i] \propto k[i]^{\alpha} \tag{9}$$

α is the exponent of preferential attachment; $k[i]$ is the indegree of vertex i in the current time step.

Thus the probability $P(k)$ that a newly created nodes links with k existing nodes decays as a power law. The graph generated using this model has power law distribution of degrees. However, this stochastic process assumes only linear relation for preferential attachment [9].

2.2.2 *Erdős–Rényi random graph: random attachment*

These graphs are of two types $G(n, p)$ and $G(n, m)$. $G(n, p)$ has n vertices and probability of an edge between them is constant p. $G(n, m)$ has n vertices and m edges such that m edges are chosen uniformly at random from a set of all possible edges [9]. In both $G(n, p)$ and $G(n, m)$ ER generative models, it is assumed that the nodes decide to form edges with other nodes based on a constant probability ($G(n, p)$) or uniformly at random $G(n, m)$. Hence, no preferential attachment pattern is observed.

2.2.3 *Preferential attachment and aging*

This is a discrete time step model of a growing random graph. At each time step a single node is added and it initiates links to node already existing in the network. The probability of a node k getting an initiated edge is given by $P[k]$ in Eq. 10 [9]. This model thus enriches the Barabasi Albert model.

$$P[k] = (c * k[i]^{\alpha} + a) * (d * l[i]^{\beta} + b) \tag{10}$$

c, d is the coefficient of degree and age; $k[i]$, $l[i]$ is the indegree and age of node i at current time step; a, b is the attractiveness of node with no adjacent edge and zero age; α, β is the preferential attachment exponent, aging exponent.

2.2.4 *Watts–Strogatz model*

A generative model which creates a lattice structured graph. Each node is connected to all nodes within its neighbourhood. The lattice structure that is formed is rewired i.e. edges are selected at random with a probability p and connected to nodes outside their immediate neighbourhood. This is done without altering the number of nodes or edges. The rewiring procedure, creates a “Small World” Effect i.e. reduction in the average path length of the graph [9].

2.2.5 *J–R model*

Jackson et al. proposed a network generative model where nodes of the social network are allowed to form links to other nodes using a hybrid strategy that encapsulates elements of preferential attachment model and the Erdős-Rényi model. Thus if there are pre-existing m nodes in a network then a newborn node links to $a * m$ of them chosen uniformly at random and $(1 - a) * m$ using a neighborhood search strategy (choice based links) and attaches to them. The hyper-parameter a is ratio of chance based interactions to choice based interactions.

2.2.6 Girvan Newman benchmark, LFR benchmark

GN benchmarks are developed using the stochastic block models. These graphs have 4 communities with 32 nodes each. There are a total of 128 nodes with each having a degree of 16. The mixing parameter μ as given in Eq. 11 decides each community association and each node has disjoint membership to one community.

$$\mu = \frac{k_o}{k_i} + k_o \tag{11}$$

k_o is the number of edges connecting vertices in different communities; k_i is the number of edges connected to a vertex.

Girvan Newman benchmarks [2] produce networks with poisson distribution. This is a drawback as real world graphs have power law distributed network sizes. To overcome this drawback, LFR benchmarks [2, 21] were proposed that had vertex degrees and community sizes power law distributed. LFR benchmark graphs are basically configuration models with built in communities which may be overlapping or disjoint. Another benchmark designed to model dynamic communities was proposed by Granell et al. [12] based on the planted l-partition model. Communities are allowed to grow, shrink, merge and split. However, at each time step the sub-graphs are proper communities in the probabilistic sense [2].

2.2.7 Forest fire network model

The Forest Fire network is a generative model where one vertex is added at a time. This vertex a connects to *ambs* vertices already present in the network, chosen uniformly at random. For each chosen vertex v the following procedure is performed:

– Generate two random numbers that are geometrically distributed with means $\frac{p}{1-p}$ and $\frac{rp}{1-rp}$ such that p is forward probability, r is backward probability.

Based on these probabilities outgoing and incoming neighbours of v are connected to a. If v has neighbours below a threshold value then all of them are connected to a.

2.3 Real networks

Tables 1 and 2 are a collection of network data-sets publicly available on platforms such as SNAP [22] and UCIML [23]. The description of these data-sets are given in Tables 3, 4 and 5.

Table 1 Description of the Social Relationship in networks (Part- I)

Sr No	Dataset	Description	Social relationship $e_{i,j}$
1	Amazon-Net	Items frequently purchased with one another on Amazon.com	item i frequently purchased with j
2	Arxiv-Net [11]	Citation graph of papers from Arxiv High Energy physics category	Paper i cites paper j
3	CondMat-Net	Collaboration network of scientist working on condensed matter research	Scientists i and j have collaborated
4	Epi-Net	Trust network between users on Epinion.com	User i trusts j
5	Fb-Net	Friendship network from Facebook	User i and j are friends
6	Gnut-Net	Peer2Peer file sharing network of users from Gnutella.com	User i shared file with j
7	Gow-Net [12]	Friendship network of users from Gowalla.com	User i and j are friends

Table 2 Description of the social relationship (part-II)

S. no.	Dataset	Description	Social relationship $e_{i,j}$
8	Slash-Net [19]	Friendship network from users of Slashdot.com	User i and j are friends
9	Twt-Net	Followers network from Twitter.com	User i follows j
10	Wiki-Net	Voters network from Wikipedia	User i has voted for user j
11	Bitcoin-OTC [24]	Bitcoin transaction between users	Edge $e_{i,j}$ means user i trusts user j
12	EU-Net [11]	Email communication network	Edge $e_{i,j}$ means i sent at least one message to j
12	Google-Net [25]	Web-graph	Edge $e_{i,j}$ denoted hyperlink from website i to j
13	CAIDA-Net [11]	Internet Topology graph	Edge $e_{i,j}$ denoted router i connected to j
14	Road-PA [25]	Road network of Pennsylvania, USA	Edge $e_{i,j}$ denoted intersection i connected to j

Table 3 Description of networks—part 1

Description	Fb-Net	Twt-Net	Epi-Net	Slash-Net	Gow-Net	Bitcoin-OTC
Nodes	4039	81306	75879	77360	196591	5881
Edges	88234	1768149	508873	905468	950327	35592
Ratio of nodes in largest WCC	1	1	1	1	1	–
Ratio of edges in largest WCC	1	1	1	1	1	–
Ratio of nodes in largest SCC	1	0.84	0.42	0.91	1	–
Ratio of edges in largest SCC	1	0.95	0.87	0.98	1	–
Avg. clustering coeff.	0.61	0.57	0.14	0.06	0.24	–
Fraction of closed triangles	0.26	0.06	0.02	0.01	0.007	–
Diameter	8	7	14	10	14	–
90-percentile effective diameter	4.7	4.5	5	4.7	5.7	–

Table 4 Description of networks—part 2

Description	EU-Net	ArXiv-Net	Google-Net	Amazon-Net	CondMat-Net
Nodes	265214	34546	875713	262111	23133
Edges	420045	421578	5105039	1234877	93497
Ratio of nodes in largest WCC	0.85	0.97	0.98	1	0.92
Ratio of edges in largest WCC	0.94	1	0.99	1	0.98
Ratio of nodes in largest SCC	0.13	0.37	0.5	0.92	0.92
Ratio of edges in largest SCC	0.36	0.33	0.67	0.92	0.97
Avg. clustering coeff.	0.07	0.28	0.51	0.42	0.63
Fraction of closed triangles	0.001	0.05	0.02	0.09	0.107
Diameter	14	12	21	32	14
90-percentile effective diameter	4.5	5	8	11	6.5

Table 5 Description of networks—part 3

Description	Wiki-Net	CAIDA-Net	Gnut-Net	Road-PA
Nodes	7115	1696415	62586	1088092
Edges	103689	11095298	147892	1541898
Ratio of nodes in largest WCC	0.99	0.99	1	1
Ratio of edges in largest WCC	1	1	1	1
Ratio of nodes in largest SCC	0.18	0.99	0.22	1
Ratio of edges in largest SCC	0.38	1	0.34	1
Avg. clustering coeff.	0.14	0.25	0.01	0.05
Fraction of closed triangles	0.05	0.001	0.001	0.02
Diameter	7	25	11	786
90-percentile effective diameter	3.8	6	7	530

3 Experimental study

Measures based on literature review in Sect. 2.1 are highlighted in Table 6 and are used to evaluate synthetic and real data-sets enlisted in Sect. 2.1, 2.3. The aim is to understand and explain underlying social phenomena and derive interesting insights about the behavior of these networks.

3.1 Results

3.1.1 Barabasi Albert: preferential attachment [BA-game]

A connected, directed graph is generated by through this model with parameters $N = 10000, \alpha = 1, a = 1$. Adhesion and edge connectivity is 0 and hence the graph is disconnected. A small diameter of 14 (given in Table 7) indicates diffusion of information can take place fast. However, 0.1% nodes have most of the incoming edges. Thuss “Rich

Table 6 Measures of network analysis

S. no.	Description
1	Adhesion (edge connectivity)
2	Diameter
3	Average path length
4	Mathematical model for degree distribution
5	Modularity
6	First 20 eigenvalues follow power law or not
7	Component distribution follows power law or not
8	Does the total weights and total edges of the graph are power law distributed
9	Does the weights and degree of the graph follow a power law
10	Weights of an edge has a relation with the weights of the nodes it connects
11	Assortativity
12	Graph density
13	Reciprocity
14	Vertex betweenness
15	Local tansitivity
–	–

getting richer” phenomenon is observed in this model. Modularity value indicates no inherent community structure. Preferential attachment model creates hubs which capture most of the incoming edges.

3.1.2 Erdos Renyi random graph [ER-game]

An undirected graph is generated through this model with parameters $N = 10000, E = 10000$. Adhesion is 0 and hence the graph is disconnected. A diameter of 29 indicates slower diffusion rate compared to Preferential attachment model. A longer diameter is due to lower average degree. The generated graph is not connected. Modularity value indicates no community structure. Thus, it would be ideal as a null benchmark for testing community detection algorithms. The strongly connected components distribution follows a power law with $\gamma = 3.09$ and intercept of 1. The first 20 eigenvalues of this graph also followed a power law distribution with $\gamma = 32$ and intercept of 3.32. Both the previous results were verified on K–S test.

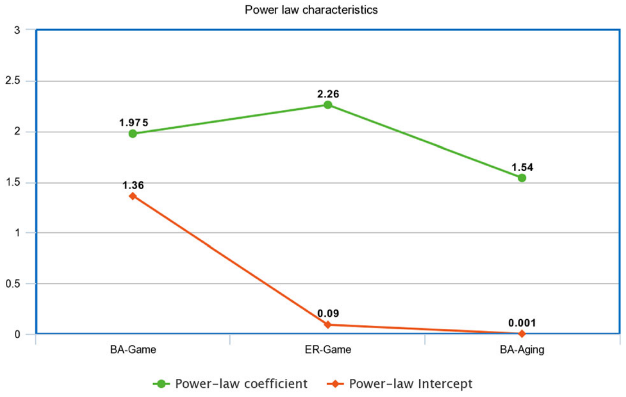


Fig. 3 Power-law co-efficient for BA-game, ER-game and BA-aging. All results indicate a good fit on K–S test

3.1.3 Evolving random graph with preferential attachment and ageing [BA-aging]

An directed graph is generated through this model with parameters set at $N = 10000, c = 1, d = 1, a = 1, b = 0, \alpha = 1, \beta = 1$. Adhesion is 0 and hence the graph is easily disconnected. A diameter of 11 indicates the fast rates for diffusion of information. The generated graph is connected. Modularity value indicates no community structure. Thus, it would be ideal as a null benchmark for testing community detection algorithms. The co-efficients of power law fi to the degree distribution is given in Fig. 3.

3.1.4 Watts–Strogatz model [WS-model]

The Small World network generated by this model with parameters $N = 100, L = 1$ has strong adhesion due to high clustering between nodes. The graph shows a small diameter $D = 10$. Small diameter indicates possibility of fast diffusion of information through nodes. As the entire network is connected, the component distribution is not available. Also as most of the nodes have same degree, the degree distribution is also not available. However, it was found that first 20 eigenvalues followed power law distribution with $\gamma = 3.34, \text{intercept} = 5.313$ and a good fit as per the K–S test. This graph captures the “Small World phenomenon” observed in real networks but fails to capture other effects such as a power law distributed degree distribution and component size distribution.

Table 7 Summary of results

S. no.	Description	BA-game	ER-game	BA-aging	WS-model	GN-model
1	Adhesion	0	0	0	10	32
2	Modularity	0	0	0	0	0
3	Diameter	14	29	11	10	4
4	Is connected	True	False	True	True	True
5	Components	1	1616	1	1	1

3.1.5 GN benchmark graph [GN-model]

The GN Benchmark model with parameters set at $N = 128, Deg_{Avg} = 16$ has strong adhesion due to high clustering between nodes. The graph shows a low degree of separation with diameter $D = 4$. The graph also has inbuilt communities but community sizes don't follow a power law distribution. As the entire network is connected, the component distribution is not available. As all the nodes have same degree and hence a degree distribution is not available. It was found that first 20 eigenvalues followed power law distribution with $\gamma = 3.15$, intercept $= 7$ and a good fit as per the K–S test. This model fails to have a power law distributed degree distribution and component size distribution.

3.1.6 LFR benchmark graph

LFR benchmark models given in Table 8 provide a graph with power law distributed community sizes and degree distributions. At low values of μ community structure is absent and hence the poor score for modularity. The benchmark is a good substitute for real data as it follows properties seen in real networks such as having a large connected component, a small diameter and power law distributed first 20 eigenvalues. The power law fit was verified using p-value of K–S Test as given in Fig. 4.

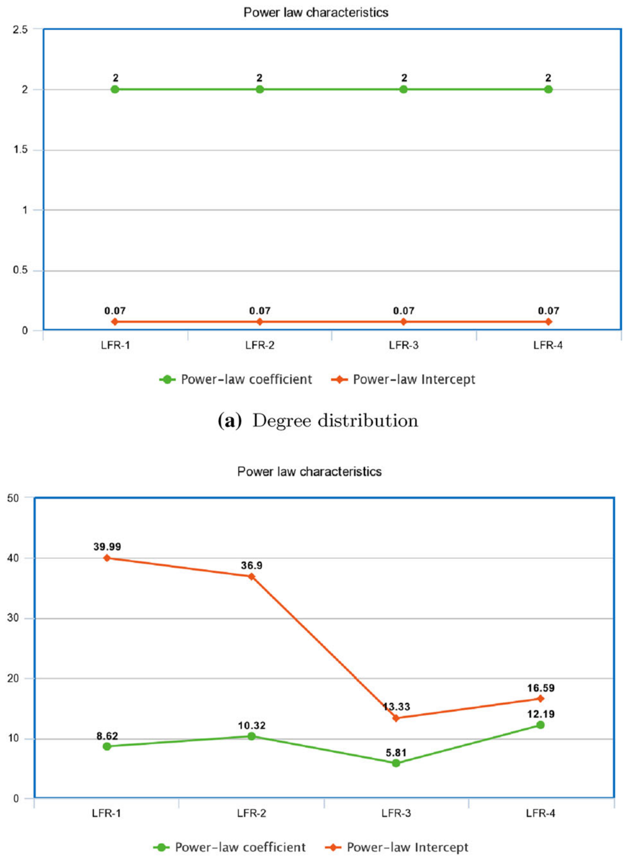
Static LFR Weighted benchmarks are used to verify additional properties as listed in Table 9.

3.1.7 Dynamic benchmark graph

The dynamic benchmarks given in Table 10 do not show the effect seen in real graphs such as "Gelling Point" [7]. The evolution of the edges with time do not show a power law relation with the nodes unlike that seen in real graphs [7].

Table 8 LFR benchmark graph

Description	LFR-1	LFR-2	LFR-3	LFR-4
Nodes	1000	1000	1000	1000
Avg. degree	15	15	15	15
Edge type	Undirected	Undirected	Directed	Directed
Weighted	Unweighted	Weighted	Unweighted	Weighted
μ	0.1	0.1	0.1	0.1
Edges	7767	7692	15,381	15,381
Adhesion	14	14	0	0
Modularity	0	0	0	0
Diameter	6	6	8	5
Is connected	True	True	True	True



(a) Degree distribution

(b) Distribution of first 20 eigenvalues

Fig. 4 Power law co-efficient of LFR benchmark graphs

3.1.8 Forest fire network model

The graph generated by this model has no adhesion or edge connectivity as given in Table 11. The diameter and average path length are small which may promote fast diffusion of data. The graph has poor assortativity by degree and reciprocity. A giant connected component exists in the graph consisting of all nodes. No central hubs are present. Modularity is poor indicating no inherent communities. The clustering coefficient has negative correlation with out degree, suggesting that there is significant clustering among low-degree nodes. Majority nodes have incoming edges equal to their outgoing edges. The in-degree and out-degree distributions of nodes fit the power law well as indicated by R^2 values. Phenomena seen in real networks such as homophily, reciprocity are absent.

Table 9 LFR benchmark graph

Description	LFR-weighted benchmark graph
Does the total weights and total edges of the graph follow a power law	$\gamma = -1$
Does the out weights and out degree of the graph follow a power law	$\gamma = -1.335$
Does the in weights and in degree of the graph follow a power law	$\gamma = -1.459$
Weights of an edge has a relation with the weights of the nodes it connects	$\gamma = -0.408$

Table 10 Dynamic benchmark graph

Description	Std-grow	Std-merge	Std-mixed
Nodes	128	128	128
Communities	4	4	4
Nodes per vommunity	32	32	32
Probability that a vertex can link to nodes in its community	0.5	0.5	0.5
Probability that a vertex can link to nodes of other community	0.01	0.05	0.05
Time period (iterations)	100	100	100

Table 11 Forest fire network model

S. no.	Description	Value
1	Nodes	10,000
2	Ambassador nodes	1
3	fw.prob, bw.factor	0.38, 0.35
4	Edges	20,638
5	Adhesion	0
6	Diameter	13
7	Avg Path length	3.4
8	Assortativity	0.2
9	Reciprocity	0
10	Is connected	True
11	Vertex betweenness	0
12	Modularity	-0.004
13	Triangles	969
14	In-degree power law distributed	$\gamma = 1.7$, Int=1.7, $R^2 = 0.86$
15	Out-degree power law distributed	$\gamma = 2.97$, Int=-0.7, $R^2 = 0.94$

3.1.9 *Online social networks*

The commonality identified during the analysis of the social networks of these websites was the power law distribution of their in-degree with in-weights and out-degree with out-weights as given in Fig. 5. The degree distributions were also power law distributed with $1.18 \leq \gamma \leq 1.74$. Adhesion for Facebook, Gowalla were 1 which indicated that users of these websites were reachable through atleast 1 route. Reciprocity of directed networks such as Epinions and Slashdot indicated higher level of mutual agreements between members. Vertex betweenness for the social networks did not indicate presence of central hubs. For social networks of such large scale lack of a central hub is quite intuitive. Poor score for modularity and graph density is

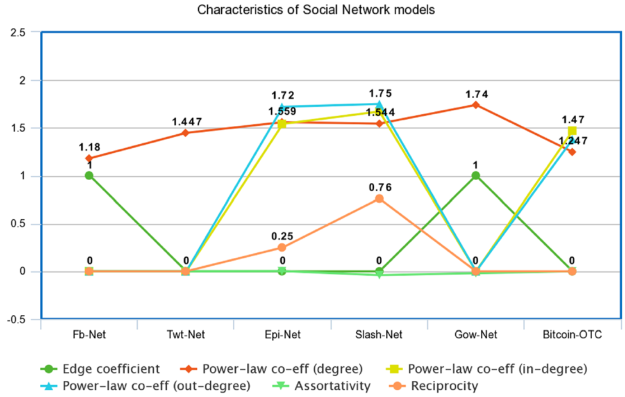


Fig. 5 Characteristics of online social network data-sets

seen in all these social networks, this is also common for such large scale graphs. Epinions and Slashdot see high reciprocity for links as the social relationship is that of trust.

3.1.10 Communication networks

Average number of frequent contacts a person has is three in the Organization under review. However an email account exists that has 7636 outgoing links. This might indicate the presence of a facility for Employee-Management communication. Assortativity by degree has a negative value which means highly active emailers send most of their emails to less active emailers. Inherent community structure is present in the graph which means that the organization could possibly have departments which communicate more internally than with other departments.

3.1.11 Citation networks

Nodes with zero citations exist which is common for citation networks. Also nodes with zero outdegree also exist (possibly because the graph is incomplete). Average number of citations a paper has is twelve. Max number of citations a paper receives is 846, max number of citations a paper makes is 411. Adhesion is zero, as many papers exist with few citations or nill citations. Assortativity in citation graphs is low as most of the links are from papers with no citations to papers with large number of citations. Reciprocity doesn't exist in citation graphs as the social relationship is "Academic status" and is not reciprocated frequently unlike friendships.

3.1.12 Web graphs

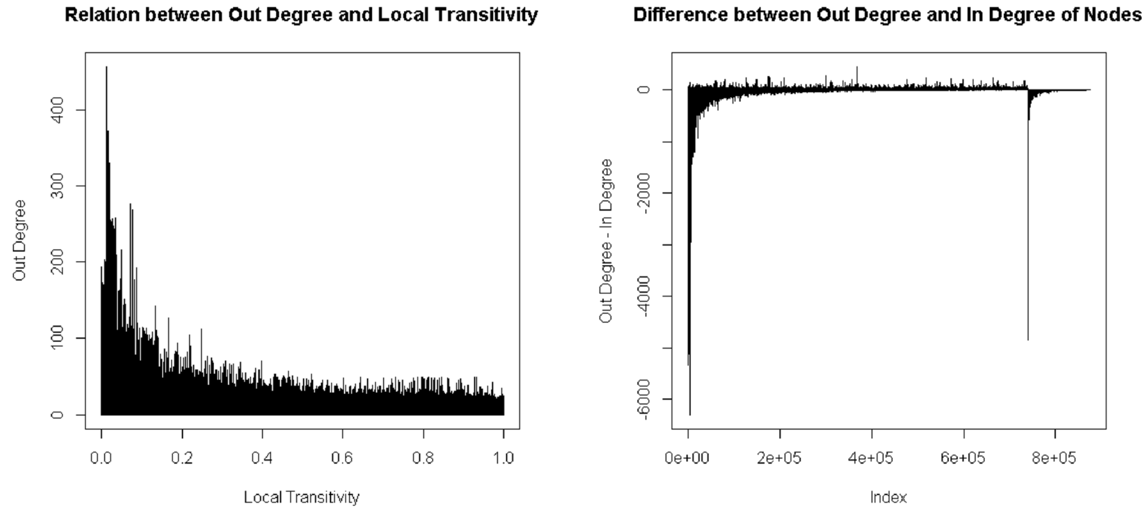
Google Web graph has adhesion and edge connectivity 0 as the graph is not connected. It is also characterised by lack of a central hub and lack of inherent community structure. The high reciprocity value for hyperlinks between webpages is peculiar and might not be held true for a larger sample size of the internet graph. Figure 6a shows analysis of local transitivity indicated a higher level of clustering between neighbours of nodes of low degree than that of nodes of high degree. Figure 6b shows that indegree of nodes matches their outdegree except for popular webpages.

3.1.13 Product co-purchasing networks

Analysis of the product co-purchasing network of Amazon reveals that Incoming edges for products are higher than outgoing edges. A Hub is present in the graph which has 420 co-purchased products. Analysis of the modularity value indicates that there exist communities of products that are co-purchased together. ≈ 40% links are reciprocal indicating consumers opt for *Customers Who Bought This Item Also Bought* feature of the Amazon.

3.1.14 Collaboration networks

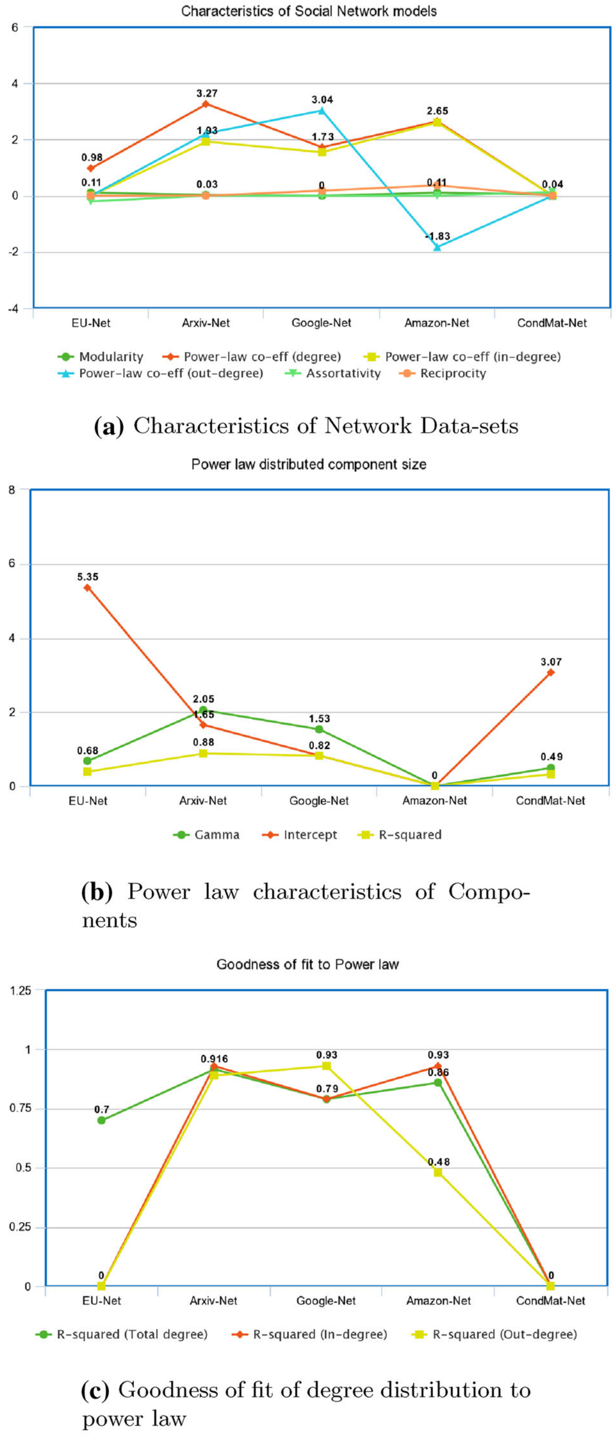
The analysis of the degree of the nodes in ArXiv Collaboration networks reveals that an average scientist has collaborated with 16 others and in rare cases upto 550 others for his research. The scientists that collaborate with few people for research show a higher degree of transitivity between their collaborators. Average path length of this



(a) Local Transitivity

(b) OutDegree - InDegree Relation

Fig. 6 **a** Local transitivity, **b** outdegree–indegree relation



(a) Characteristics of Network Data-sets

(b) Power law characteristics of Components

(c) Goodness of fit of degree distribution to power law

Fig. 7 Analysis of network data-sets

network is ≈ 5 indicating fast diffusion of information in the network as given in Fig. 7.

3.1.15 Other datasets

The analysis of the social networks from Voting trends (Wiki-Vote), Internet Topology (CAIDA-Net), P2P networks (GnutellaP2P) and Road networks (Road-PA) reveal a adhesion and edge connectivity of 0. Assortativity by degree on Road networks show that busy (high degree) intersections are connected to other busy intersection as given in Fig. 8. Such a design requirement is very intuitive and could be common to other road networks too. All the social networks exhibit poor community structure and also do not contain central hubs. Figure 9a shows that users that have voted for many participants have low transitivity. Transitivity patterns of all the networks under analysis reveal the same trend. Figure 9b indicates users that have supported large number of other users do not necessarily receive support from others. Component and degree distributions are analysed in Figs. 10 and 11 respectively.

4 Conclusion

The analysis of social networks belonging to domains such as online social networks to citation graphs was performed in this inquiry. Network theoretic concepts were used for the analysis. Observations made in the previous inquiries were validated on social network data-sets of different domains. It was observed that degree distributions on real and synthetic data-sets were power law distributed. However, component size distribution has a poor fit for the power law except in GnutellaP2P file sharing websites. In the case of the road networks of Pennsylvania state of USA, it was found that busy intersections would be connected to other busy intersection. Probably the roads were designed in this way to ensure smooth vehicle movement. Commonality between data-sets of all domains was the absence of central hubs, presence of community structure and low graph density.

Social networks are a particular class of networks that represent the sum of all personal or professional ties between the members of the system. A network perspective revealed that social networks shared properties such as negative assortativity, domination of a few members, high edge density, power law distributed degree and component sizes, high transitivity, high reciprocity and small average path length and diameter.

Representing the systems in the form of a social network facilitated easier but conceptually sound analysis. However, there are also several limitations of network representation models. The calculation of certain characteristics such as diameter and average path length require $n * n$ computations for a graph of n nodes. This made the calculation expensive for large graphs. Machine learning applications on graphs such as node classification, link prediction, expert recognition etc. require hand-engineering features to be calculated initially. This makes the results of the exercise dependent on the ability of the

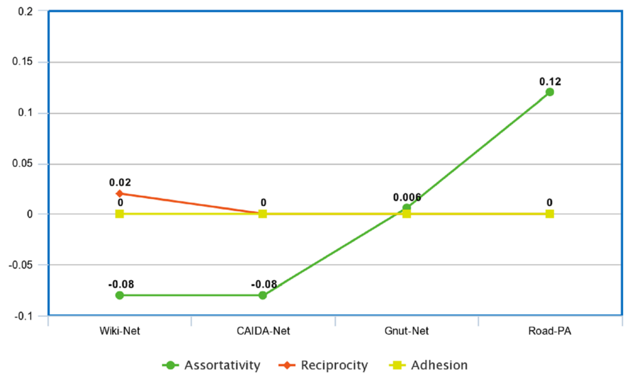
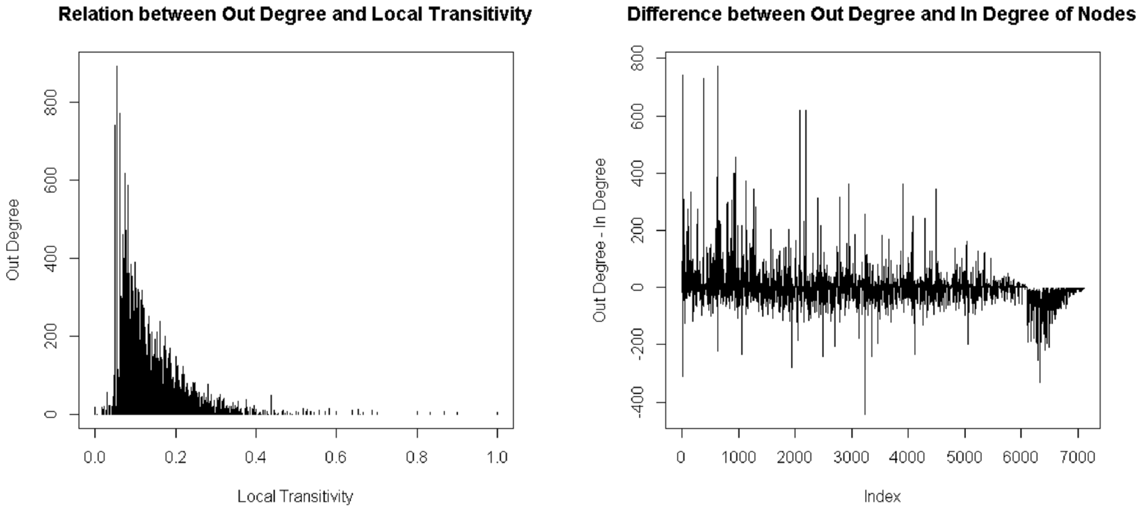


Fig. 8 Degree of assortativity, reciprocity and adhesion in the networks

researcher. Machine learning cannot be directly applied on the network as the data-points on the network are not *i.i.d*. The data-points are connected to each other by links or edges and thus the *i.i.d* assumption for machine learning is not satisfied.

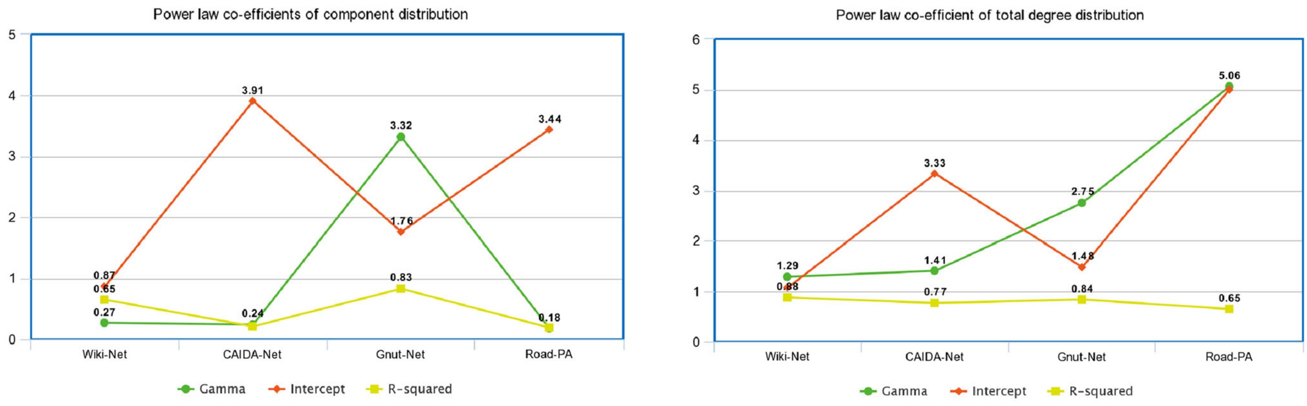
A solution for these drawbacks is network representation learning (NRL). The adjacency matrix of network $G(V, E)$ is denoted by $A \in R^{|V|*|V|}$ and $V^a \in R^{|V|*p}$ is used to denote the vertex attribute matrix if present otherwise, $V^a = \phi$. The problem is defined as follows: given a network $G = (V, E)$ and associated attributes, the aim is to represent each node u in a low-dimensional vector space y_u by learning a mapping $f : V, V^a \rightarrow R^d$, namely $y_v = f(v, V^a) \forall v \in V$. It is required that $d \ll |V|$ and the function f preserve a proximity measure defined on the



(a) Local Transitivity

(b) OutDegree - InDegree Relation

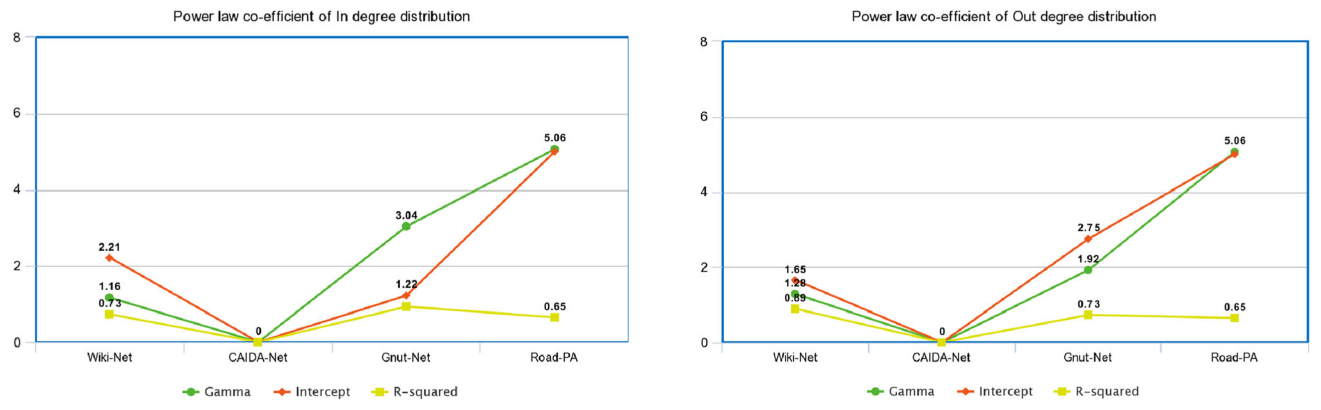
Fig. 9 **a** Local transitivity, **b** outdegree–indegree relation



(a) Power law co-efficient of component distribution of networks

(b) Power law co-efficient of total degree distribution of networks

Fig. 10 Component distributions of network data-sets



(a) Power law co-efficient of In degree distribution of networks

(b) Power law co-efficient of Out degree distribution of networks

Fig. 11 Degree distribution of networks

graph G. Intuitively, if two nodes u and v are “similar” in graph G, their embedding y_u and y_v should be close to each other in the embedding space i.e. ${y_u}^T y_v \sim 1$. The notation $f(G) \in R^{|V|*d}$ is used for the embedding matrix of all nodes in the graph G. As the nodes are converted to vector embeddings, the drawbacks of network representation models are overcome. Designing efficient NRL techniques is a promising research direction if machine learning applications have to be developed for graphs.

References

1. Handcock MS, Gile KJ (2010) Modeling social networks from sampled data. Ann Appl Stat 4(1):5
2. Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44
3. Gomez V, Kaltenbrunner A, Lopez V (2008) Statistical analysis of the social network and discussion threads in slashdot. In: Proceedings of the 17th international conference on World Wide Web. ACM, New York, pp 645–654
4. Barabási A-L et al (2016) Network science. Cambridge University Press, Cambridge
5. Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: bringing order to the web. In: Proceedings of the 7th international world wide web conference, Brisbane, Australia, pp 161–172. https://www.bibsonomy.org/bibtex/2ac49c33e114ca171db40cece6a0ae4d6/sac
6. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117
7. McGlohon M, Akoglu L, Faloutsos C (2011) Statistical properties of social networks. In: Social network data analytics. Springer, New York, pp 17–42
8. Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. J Lashdot Stat Assoc 97(460):1090–1098
9. Jackson MO (2010) Social and economic networks. Princeton University Press, Princeton
10. Shi Y, Gui H, Zhu Q, Kaplan L, Han J (2018) Aspem: embedding learning by aspects in heterogeneous information networks. In: Proceedings of the 2018 SIAM International Conference on Data Mining. SIAM, pp 144–152
11. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. ACM Trans Knowl Discov Data 1(1):2–49
12. Granell C, Darst RK, Arenas A, Fortunato S, Gómez S (2015) Benchmark model to assess community structure in evolving networks. Phys Rev E 92(1):12–19
13. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393(6684):440–442
14. Erdos P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5(1):17–60
15. Ricci V (2005) Fitting distributions with R. https://www.bibsonomy.org/bibtex/289d38b1135b797469c33d514a7677ff2/folke
16. Johnson, RA, Wichern, DW et al. (2002) Applied multivariate statistical analysis, vol 5. Prentice Hall, Upper Saddle River, NJ
17. Tan W, Brian Blake M, Saleh I, Dustdar S (2013) Social-network-sourced big data analytics. IEEE Int Comput 17(5):62–69
18. Friedman BD, Burns MJ, Cao J (2014) Enterprise social networking data analytics within Alcatel-Lucent. Bell Labs Tech J 18(4):89–109
19. Gewerc A, Marteiro E (2016) Academic social networks and learning analytics to explore self-regulated learning: a case study. IEEE Rev Iberoam de Tecnol del Aprendiz 11(3):159–166
20. Dwyer C, Hiltz S, Passerini K (2007) Trust and privacy concern within social networking sites: A comparison of facebook and myspace. AMCIS 2007 proceedings, pp 324–339
21. Lancichinetti A, Fortunato S (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys Rev E 80(1):118–129
22. Leskovec J, Krevl A (2014) SNAP datasets: stanford large network dataset collection. http://snap.stanford.edu/data
23. Dua D, Taniskidou K (2017) UCI: machine learning repository. http://archive.ics.uci.edu/ml
24. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, New York, pp 29–42
25. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. Internet Math 6(1):29–123