Theoretical Computer Science 551 (2014) 22–38

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Fitting truncated geometric distributions in large scale real world networks

Swarup Chattopadhyay *, C.A. Murthy, Sankar K. Pal

Center for Soft Computing Research, Indian Statistical Institute, Kolkata-700108, India

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form 13 April 2014
Accepted 5 May 2014
Available online 15 May 2014
Communicated by A. Skowron

Keywords:
Social networks
Natural computing
Power-law distributions
Heavy-tailed distributions
Maximum likelihood
Truncated geometric distribution

ABSTRACT

Degree distribution of nodes, especially a power-law degree distribution, has been regarded as one of the most significant structural characteristics of social and information networks. However it is observed here that for many large scale real world networks, the power-law does not fit properly because of the presence of large fluctuations and sparsity in upper and lower tails of the distribution. Here we have proposed to fit the truncated geometric distribution on three distinct and non-overlapping parts of the degree frequency table. Extensive experiments on twenty three (23) real world networks revealed that the proposed model fitted better than the power-law and other distributions.

1. Introduction and related works

A network is a set of items (named as vertices or nodes) with connections between them, called edges. An edge represents the relation between two vertices or social entities, and a network represents a collection of such relations. A large scale real world network is one in which the number of nodes is large (of the order of thousands or millions). We can consider such examples as, Twitter, Facebook, LinkedIn network, etc. to be large scale real world networks. Apart from these, there are some classic examples of real world networks formed by connecting small number of nodes [1] such as the Zachary's karate club network, Dolphin social network, American College football network, etc. Usually, in a real world network, many pair of individuals (or nodes or vertices) is joined by a path to form a connected component. Thus a large scale real world network can be considered to be a collection of connected components.

Understanding and modeling real world network structures has been a focus of attention in a number of diverse fields, including physics, biology, computer science, statistics, and social sciences. Recent research has focused on the analysis of structural characteristics like degree distribution and clustering coefficient. The node degree distribution has been viewed as an important structural characteristic of social and information network. The degree distribution of many man-made and/or naturally occurring phenomena, including city sizes, incomes, word frequencies, earthquake magnitude and number of phone call per customer [2] generally follows power-law [3–6]. But this distribution produces significant fitting error [7] and then it becomes core research problem in the newly emerging network science discipline.

In 1999, Barabasi and Albert [8–10] modeled the degree distribution of the World wide Web (WWW) using a power-law. Since then, this structural behavior has been broadly investigated in many other types of real-world networks, including

* Corresponding author. Tel.: +91 33 2575 3104/3100; fax: +91 33 2578 3357.
E-mail addresses: swarupchatt@gmail.com (S. Chattopadhyay), murthy@isical.ac.in (C.A. Murthy), sankar@isical.ac.in (S.K. Pal).

http://dx.doi.org/10.1016/j.tcs.2014.05.003
0304-3975/

metabolic networks, gene regulatory networks, collaboration networks, communication networks and social networks. A real world network model incorporates knowledge about the individuals and their interactions. Such networks are dynamical in nature, by virtue of their growing vertices and/or edges. This dynamical behavior can be captured using natural computing techniques. Natural computing refers to computational processes which investigate models inspired by natural system such as neural computation, evolutionary computation, granular computation, cellular automata, swarm intelligence, artificial immune systems and artificial life systems in order to understand the real world problems in terms of information processing.

There are several natural processes related to signal processing, data visualization, data mining, etc. which, like real world networks, can be viewed as information processing, where natural computing models have been applied successfully [11,12]. Description of some natural computing models, applied for the identification of real world networks, social networks, networks formed by society of agents or objects, molecules can be found at [11,13]. One such natural computing technique, Adaptive Double Self-Organizing Map (ADSOM) [14] involves a clustering strategy for the modeling of such real world gene regulatory networks. Situated Cellular Agents (SCA) [15], another such technique, has been used to simulate crowd dynamics in a society of multi-agent systems based on the agent interaction. The degree distribution of real world networks involving collaboration networks, communication networks, social networks, and biological networks, etc. were reported to follow power-law [16–18]. People have also observed that many human activities, ranging from communication to entertainment [19] and the dynamics of human behavior based on blogosphere follows heavy tailed distribution [20]. Recently, it has been well studied and empirically shown that the power-law distribution does not fit well in all cases [17,21]. Several models have also been proposed for studying the structural characteristics of real-world networks [22]. There has been some recent studies involving social networks gathered from the records of a large mobile phone operator [2] and Facebook [7], which suggested the existence of log-normal, Pareto log-normal (PLN), double Pareto–lognormal (DPLN), drift power-law, etc. distributions instead of power-law.

In this article, we shall be dealing with models mainly related to power-law, the estimation of parameters with respect to it, and the corresponding consequences. We will also be discussing some of the other models *viz.* Poisson, log-normal, drift power-law, power-law with exponential cutoff and Pareto [7,17] for comparison purpose of the degree distribution of real world networks.

Mathematically, a quantity x follows a power-law if it is drawn from a probability distribution

$$P(x) \propto x^{-\alpha}, \tag{1}$$

where, the parameter α is a positive constant, and is known as exponent or scaling parameter. α is to be estimated from the given data. Usually, the power-law is applied only for values greater than some minimum (say, x_{min}).

Often, simple graphical methods are used for fitting the empirical data to a power-law distribution. Such graphical analysis, based on linear fitting of log–log transformed data, can be grossly erroneous [18,23]. Therefore, without a quantitative measure of goodness-of-fit, it is difficult to assess how well a data approximates a power-law distribution. Due to some reasons sometimes the underlying process may not actually generate power-law distributed data, such as biased estimation of parameters or biased data collection technique. The current broadly used methods for fitting the power-law distribution over real world networks tend to provide significant error because of large fluctuations and sparsity in upper and lower tail of the distribution.

This paper demonstrates that multiple distributions are appropriate to fit the data and they provide better fit than the fit provided by power-law and other distributions alone. Here, the observed fit of the distribution is evaluated by using a chi-square test. As the real world networks grow to sizes far beyond the possibility of manual processing, it is very difficult to collect and analyze the data over a certain period. Better fit distributions help in modeling the phenomenon better. Techniques such as natural computing [11], statistical modeling [24], graph theoretic approaches [25], etc. can aid in the study of the evolvability of real world networks and help in better modeling.

In this article, we will be providing a more accurate fit using truncated geometric distribution to model the node degree distribution of a network compared to power-law, log-normal, Pareto, drift power-law and power-law with exponential cutoff distributions.

The remainder of the paper is organized as follows. Section 2 provides the details of power-law distribution, how to fit, how to estimate the scaling parameter and the quality of fit. We describe some common statistical distributions and derive their truncated versions in Section 3. We propose and analyze a new distribution for fitting the data in Section 4. Goodness-of-fit test and the testing over the real world networks are provided in Section 5 and Section 6. Section 7 concludes the paper.

2. Power-law: definitions, properties and fit

Let us provide here some basic definitions required for fitting the power-law distribution to data sets.

2.1. Continuous and discrete power-law behavior

Conventionally, a power-law distributed quantity can be either continuous or discrete. A continuous power-law distribution is one described by a probability density function $p(x)$ such that

$$p(x)dx = Pr(x < X \leq x + dx) = Cx^{-\alpha}dx \tag{2}$$

where x is the observed value and C is a normalization constant. Clearly, this density diverges as x tends to 0, and hence Eq. (2) cannot hold for all $x \geq 0$; there must be some lower bound to the power-law behavior. We will denote this bound by x_{min}. Then, provided $\alpha > 1$, it is easy to calculate the normalizing constant and it was found [17] that

$$p(x)dx = \frac{(\alpha - 1)}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha} \tag{3}$$

In the discrete case, x can take only a discrete set of values. In this paper we considered only the case of integer values with a probability mass function of the form

$$p(x)dx = Pr(X = x) = Cx^{-\alpha} \tag{4}$$

2.2. Procedure for fitting power-law to empirical data

We turn now to the first of the main goals of this paper, the fitting of power-law distribution to discrete data. Studies of empirical distributions that follow power-law usually give some estimate of the scaling parameter α [17,26]. The tool most often used for this task is the simple histogram. Taking the logarithm of both sides of Eq. (4), we see that the power-law distribution obeys $\ln p(x) = -\alpha \ln x + const$, implying that it follows a straight line on a log–log plot.

Typically the slope α is extracted by performing a least-squares linear regression on the logarithm of the histogram [16]. There are several ways of estimating α from the observed data. We have used here maximum likelihood estimate (MLE) for estimating the scaling parameter alpha since other methods provided worse result than MLE.

2.3. Estimating the scaling parameter

First, let us consider the estimation of the scaling parameter α. Estimating α requires, as we will see, a value for the lower bound x_{min} of power-law behavior in the data. For the moment, let us assume that this value is known. The method of choice for fitting parameterized models such as power-law distributions to observed data is the method of maximum likelihood, which gives an accurate estimate of the parameter when the sample size is large [27,28]. Assuming that our data is drawn from a distribution that follows a power-law exactly for $x \geq x_{min}$, we can derive maximum likelihood estimators (MLEs) of the scaling parameter for both the discrete and continuous cases [16,17].

The MLE for the continuous case is

$$\hat{\alpha} = 1 + n\left[\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{min}}\right)\right]^{-1}$$

where $x_i, i = 1, 2...n$ are the observed values of x such that $x_i \geq x_{min}$. Here and elsewhere we use "hatted" symbols such as $\hat{\alpha}$ to denote estimates derived from data.

The MLE for the case where x is a discrete integer variable is less straightforward. Seal [29] and more recently Goldstein et al. [30] treated the special case $x_{min} = 1$, showing that the appropriate estimator for α is given by the solution to the transcendental equation [31,32]:

$$\frac{\zeta'(\alpha)}{\zeta(\alpha)} = (-1/n)\sum_{i=1}^{n} \ln x_i$$

For estimating x_{min} for a discrete data set, usually, a few probable values of x_{min} are considered. The better one among them is taken as the estimate for x_{min}.

2.4. Quality of fit of power-law to large scale networks

Though power-law is popular for fitting in large scale real world networks, some problems were found in the power-law fit in many cases. For the whole range of variable x_{min}, a single power-law distribution is found to be inadequate in representing the characteristics of the data well. The following table provides the quantitative error in favor of rejecting the power-law hypothesis.

In Table 1, the chi-square statistic is computed while fitting power-law distribution over the whole data by a proper choice of x_{min}. That is, for every data set, the reported value of x_{min} is the best observed value. In other words, the other x_{min} values provided worse results than the reported x_{min}. The observed chi-square value is larger than the actual chi-square value in every cases and hence power-law distribution did not seem to fit properly. Therefore this inadequacy of the single distribution (power-law) motivates us to use truncated distribution. Hence our aim is to split the whole data into some segments and try to fit appropriate distribution over each segment.

Table 1
Error in fitting power-law to some data sets.

Data set [38]		No. of nodes	No. of edges	X_{min}	Estimated α	No. of intervals	Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
Social network	Who Trust Whom	75 879	508 837	2	2.75	695	25 392	762.661	Rejected
	Live Journal	4 847 571	68 993 773	3	2.26	770	27 247	866.991	Rejected
Citation network	HEPP	34 546	421 578	3	2.18	390	23 451	447.632	Rejected
	HEPTh	27 770	352 807	2	2.01	405	5078	447.632	Rejected
Collaboration network	Condensed Matter	23 133	186 936	1	1.64	70	1234	90.531	Rejected
	General Relativity	5242	28 980	2	2.01	65	583	84.821	Rejected

3. Truncated distributions

In this section we discuss some of the decay distributions, *viz.* power-law, log-normal, geometric, Pareto, drift power-law, power-law with exponential cutoff, exponential, Poisson, and their corresponding truncated versions to model the degree distribution of a real world network.

Given a discrete or continuous distribution, if we truncate the distribution either at one end point or at both end points, the probability of points falling outside the range (outside the truncated portion) will be zero and the probability inside the range will be adjusted accordingly.

Let, X be a discrete, unbounded, non-negative random variable having probability mass function $P(x); x = 0, 1, 2, \ldots\ldots$ with cumulative distribution function $F(x)$.

We shall define a new random variable Y as follows.

$$P(Y = i) = \begin{cases} \frac{P(X=i)}{F(b)-F(a)}, & i = a+1, a+2, \ldots, b-1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $F(x) = \sum_{i=0}^{x} P(X = i);\ x = 0, 1, 2, 3, \ldots\ldots$

Then Y is a random variable, obtained from truncating X on the left and right at a and b respectively.

For continuous case, let X be a random variable having probability density function $f(x)$, with cumulative distribution function $F(x)$ both of which have infinite support. Let Y be a random variable obtained from X after truncating X on the left and the right are a and b respectively. Then the probability density function $g(y)$ of Y can be defined as follows.

$$g(y) = \begin{cases} \frac{f(x)}{F(b)-F(a)}, & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $F(x) = \int_{-\infty}^{x} f(x)dx$.

We have given here some common statistical distributions and their corresponding truncated versions while truncating at the point k in the following table.

Name	Distribution				
	pmf/pdf	cdf		Truncated pmf/pdf	
Geometric	$(1-p)^{x-1}p$	$1-(1-p)^x$	$x = 1, 2, 3, \ldots\ldots$	$\frac{(1-p)^{x-1}p}{1-(1-p)^{k-1}}$	$x = 1, 2, 3, \ldots, k$
Power-law	$Cx^{-\alpha}$	$\sum_{x=1}^{k} Cx^{-\alpha}$	$x = 1, 2, 3, \ldots\ldots$	$\frac{Cx^{-\alpha}}{\sum_{x=1}^{k} Cx^{-\alpha}}$	$x = 1, 2, 3, \ldots, k$
Pareto	$\alpha \frac{x_m^\alpha}{x^{\alpha+1}}$	$\sum_{x=1}^{k} \alpha \frac{x_m^\alpha}{x^{\alpha+1}}$	$x = 1, 2, 3, \ldots\ldots$	$\frac{\alpha \frac{x_m^\alpha}{x^{\alpha+1}}}{\sum_{x=1}^{k} \alpha \frac{x_m^\alpha}{x^{\alpha+1}}}$	$x = 1, 2, 3, \ldots, k$
Poisson	$\frac{e^{-\mu}\mu^x}{x!}$	$\sum_{x=0}^{k} \frac{e^{-\mu}\mu^x}{x!}$	$x = 0, 1, 2, \ldots\ldots$	$\frac{\frac{e^{-\mu}\mu^x}{x!}}{\sum_{x=0}^{k} \frac{e^{-\mu}\mu^x}{x!}}$	$x = 0, 1, 2, \ldots, k$
Log-normal	$\frac{1}{x\sqrt{2\pi\sigma^2}} exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]$	$\frac{1}{2} + \frac{1}{2} erfc(\frac{(\ln x-\mu)}{\sqrt{2}\sigma})$	$x \in (0, \infty)$	$\frac{\frac{1}{x\sqrt{2\pi\sigma^2}} exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]}{\int_{x=0}^{k}(\frac{1}{2} + \frac{1}{2} erfc(\frac{(\ln x-\mu)}{\sqrt{2}\sigma}))dx}$	$x \in (0, k]$
Exponential	$\lambda(e^{-\lambda x})$	$1 - e^{\lambda x}$	$x \in [0, \infty)$	$\frac{\lambda(e^{-\lambda x})}{1-e^{\lambda x}}$	$x \in [0, k]$
Power-law exp. cutoff	$Cx^{-\alpha}e^{-\lambda x}$	$\sum_{x=1}^{k} Cx^{-\alpha}e^{-\lambda x}$	$x = 1, 2, 3, \ldots\ldots$	$\frac{x^{-\alpha}e^{-\lambda x}}{\sum_{x=1}^{k} Cx^{-\alpha}e^{-\lambda x}}$	$x = 1, 2, 3, \ldots, k$
Drift power-law	$C(x+\theta)^{-\alpha}$	$\sum_{x=1}^{k} (x+\theta)^{-\alpha}$	$x = 1, 2, 3, \ldots\ldots$	$\frac{(x+\theta)^{-\alpha}}{\sum_{x=1}^{k} (x+\theta)^{-\alpha}}$	$x = 1, 2, 3, \ldots, k$

In this paper, the data sets used are all finite data sets. For a finite sample space it is expected that the truncated distributions provide better approximation compared to non-truncated distributions. Better approximation means the expected frequency becomes more close to the actual frequency.

Table 2
Error in fitting geometric distribution (truncated and non-truncated).

No. of points	Observed p	Truncated			Non-truncated	
		Truncation point	Estimated p	Error	Estimated p	Error
10 000	0.35	21	0.35273	5204	0.35301	5479
20 000	0.25	31	0.24962	13 156	0.25079	13 416
50 000	0.2	48	0.20012	28 987	0.20044	29 343
100 000	0.15	71	0.14951	92 286	0.14953	92 386

Table 3
Error in fitting Poisson distribution (truncated and non-truncated).

No. of points	Observed λ	Truncated			Non-truncated	
		Truncation point	Estimated λ	Error	Estimated λ	Error
10 000	5	14	4.9908	10 675	4.9885	10 738
20 000	7	21	7.0143	19 509	7.0142	19 621
50 000	4	15	4.0055	21 213	4.0056	21 303
100 000	6	20	6.0016	32 841	6.0015	32 901

We have here experimented it over two discrete distributions *viz.* Geometric and Poisson by generating random points from these two distributions and it has been found that the truncated versions provide us better approximation compared to non-truncated versions. The error in approximation is being calculated by sum of squares of the difference between actual and expected frequencies.

Tables 2–3 provide results for both truncated and non-truncated cases by generating random points from geometric and Poisson distribution. In each case, it has been found that the truncated distributions provide us the better approximation compared to non-truncated distributions.

4. Proposed method of fitting multiple truncated geometric distributions

Usually, one may try to fit a single distribution to a data set, or fit different distributions on different segments of the data. Fitting a single well known distribution is always better, provided the distribution fits the data well. Fitting distributions on different segments of the data is not always the first choice because of the following drawbacks. (a) One needs to determine the number of segments appropriately, and provide valid reasons for each of the segments. Determining the number of segments may be done heuristically (which may not be proper in some situations), or mathematically (by assuming a model that may or may not work always well). (b) After finding the number of segments, the data is to be divided into those many segments. The process of division of the data in those many segments may not be unique. (c) Then a distribution is to be fit on each segment.

It is clear from the graph (Fig. 1) of the degree distribution that a decay distribution needs to be used for modeling. It has been further observed that a single standard decay distribution is unable to fit the data properly. In such a situation where no standard existing distribution can be used for fitting properly, one may resort to fitting distributions on different data segments. For the 23 data sets considered by us (see Section 4.2), we tried to the fit the truncated versions of several well known distributions such as the Poisson, Pareto, Geometric, Log-normal, Power-law, Drift power-law, and power-law with exponential cutoff. No distribution was found to fit any one of these data sets properly. The measure of fit used here is the Chi Square Test.

By looking at the graph (Fig. 1) of the degree distribution of a real world network, *viz.* collaboration network, communication network, citation network, social network, etc. considered by us, we can intuitively divide the degree distribution of the network into three parts based on their slopes. We have observed that the first part has a sharp slope with the change of frequency being abrupt, the slope in the second part is gradually decreasing with a slow change in the frequency, the change in the slope of the third part is very-very slow which is also reflected by the slight change in the frequency. Thus, we tried to segment the data sets, and fit distributions on each segment. That is, there is a possibility of fitting several decay distributions over each part with certain parameters. It was empirically found that the geometric distribution provides a better fit than the others, based on their statistical significance. These observations were made on the basis of where and when the standard distributions are failing in fitting the data exactly. The value for the number of segments (*i.e.*, three) is arrived at subjectively.

There exist two ways in which the data may be divided into three parts. One way is to make the division independent of the data set. The second is to make the division data dependent, *i.e.*, for different data sets, the segment lengths, the starting point and end point of the intervals are different. The divisions are made in data dependent fashion here.

Generally, in a society people can be classified to belong to one of the three categories, *viz.* low income family, medium income family and high income family, based on their income. Similarly, in an educational system a student can also be categorized into low, medium or high merit, based on his/her performance. Occasionally, further categorizations like too low, too high or above average, are also used in the system/society. The three broad categories, low, medium and high seems

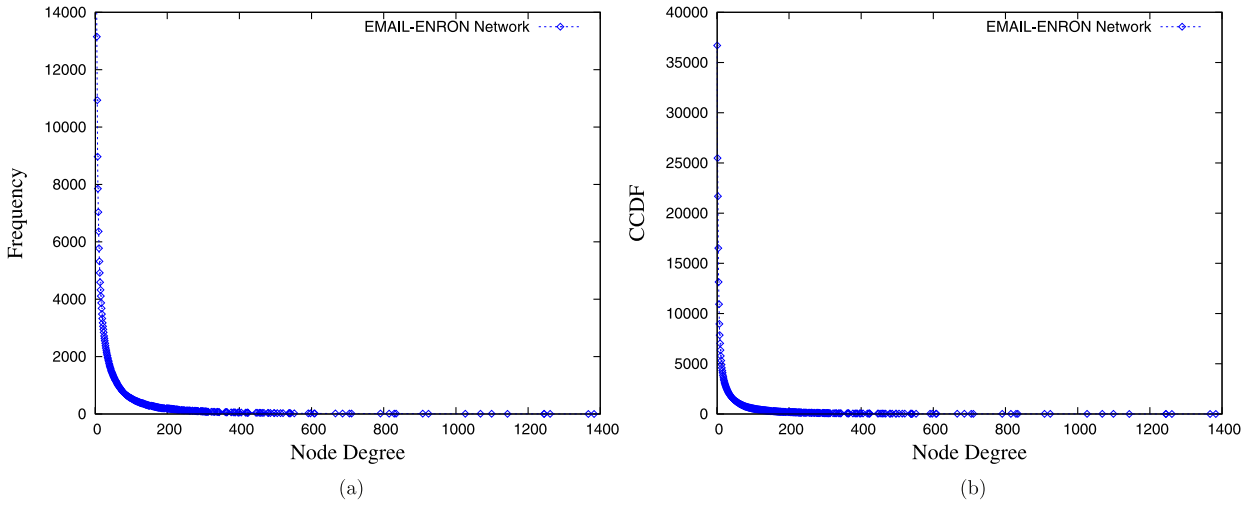

Fig. 1. Degree distribution of EMAIL ENRON network considering: (a) Degree-Frequency and (b) Degree-Complementary Cumulative Frequency.

to be more rational than further categorization and found to be justified. So, even though the degree distribution can be divided into more than three parts or less, *viz.* four parts or two (say), but the division into three parts has a crucial role for extracting significant contextual information from the data. This is exemplified below.

In the field of collaboration networks, an author having higher collaborative degree means that the author is well familiar in the respective field. In the context of this article, the first part corresponds to the set of authors having less number of collaborations (say, ~ 10). Authors belonging to the second part have relatively higher number of collaborations (say, ~ 100), and the third part consists of authors having a very high number of collaborative (say, > 200). These authors are extremely familiar figures in the community. In the case of citation networks, the first part corresponds to the set of papers which have lower citation, the second part consists of papers having moderate number of citations, while the third part contains the highly cited papers. This division is quite apparent from the data considered by us.

Similarly, in the case of social networks, various well-known websites like those of the Facebook, Twitter, LinkedIn, etc. have become successful due to their ability in connecting people, exchanging their ideas, and through this influencing a large population in a short period of time. Hence it is natural to target one's messages to highly connected/influential people who will propagate them further in the social network. Now finding a group of influential people efficiently in large-scale social networks has become a challenging problem in current situation. In the context of this article, the group of less influential people belong to the first part, the second part consists of people with medium influence, while high influential people belong to the third part.

Before fitting a distribution over each segment of every data set, sometimes, pre-processing of the data is needed. Simple rules are applied to partition the data.

4.1. Data pre-processing and dividing rules

Given a degree distribution data of a real world network, our first task is to check if any noise cleaning, or smoothing operation is necessary on the data. If the frequency distribution table of degree of nodes exhibits many local variations in its histogram, then those local variations need to be smoothened. Smoothing algorithms remove noise or local fluctuation from data sets while preserving the underlying patterns and is a standard practice in Time Series analysis [33]. One of the ways of smoothing a histogram is to use moving average technique with the help of a window of length 3 (usually the window length is taken as 3), and the window is moved over the entire histogram [34,35]. This is a standard procedure for removing local variations in the data. However, we note here that, as we move towards the tail of the distribution, we have many nodes with degree frequency zero, and moving average method may not be applied at this portion of the data. We have here moved the window over degree frequency table until we get two consecutive zeros as degree frequency.

While doing moving average over the entire histogram, the total frequency does not change abruptly. Let $f_1, f_2, ..., f_N$ be the original frequencies given in the table and $g_1, g_2, ..., g_N$ be the modified frequencies after moving average of window size 3 over the entire frequencies. Now g_i can be written as

$$g_i = \frac{f_{i-1} + f_i + f_{i+1}}{3}, \quad i = 1, 2, ..., M$$

If we pad the null frequencies f_{-1}, f_0 and f_{M+1}, f_{M+2} with the original frequencies over head and tail, then it can be easily proved that

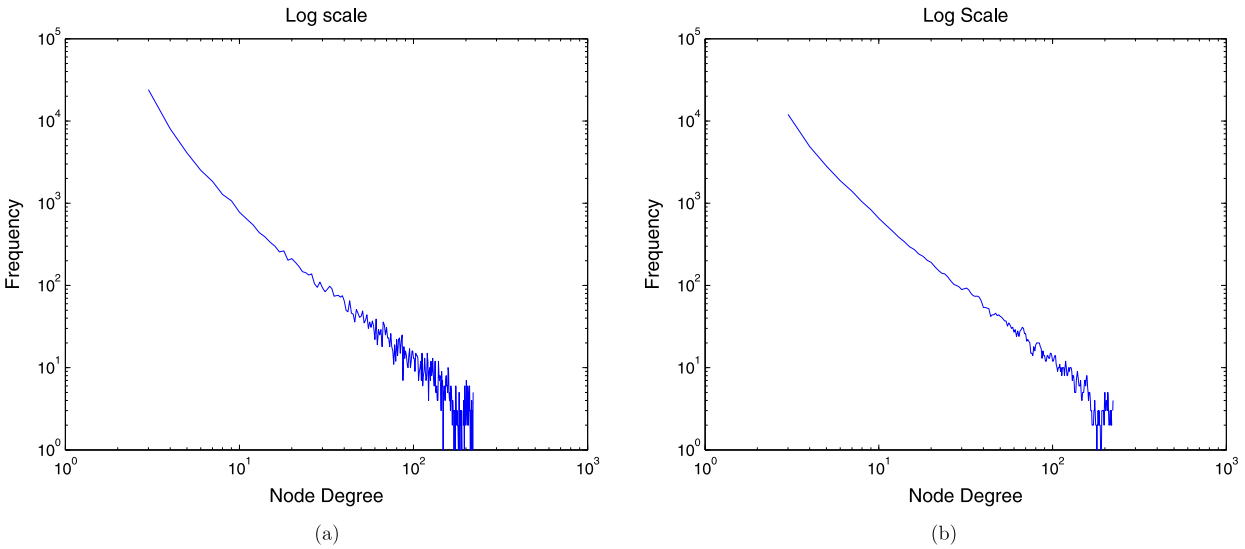

Fig. 2. Degree distribution of Who Trust Whom network of Epinions (a) before moving average and (b) after moving average.

$$\sum_{i=0}^{M+1} g_i = \sum_{i=1}^{M} f_i.$$

So in the rest of the paper for experimental purpose, we will be dealing with the modified frequencies $g_1, g_2, ..., g_N$ without changing the total frequency a lot. Fig. 2 depicts the changes by applying the moving average method of window size 3 over the degree frequency table of Who Trust Whom network of Epinion in loglog scale. From the above figure it is very clear that by applying the moving average method, we are only removing the local variance from the data by keeping the original distribution as it is.

While fitting power-law and other distributions over the whole data, it has been observed that the distributions does not fit properly in the upper and lower/tail part of the data due to the presence of huge nonlinearity and sparsity. Here we are trying to remove the nonlinearity and sparsity from the upper and lower/tail part of the data by taking log and grouping it into some intervals for better fitting using proposed method.

Now let us divide the modified degree frequency table into three distinct and non-overlapping groups based on the frequency value of certain degree. We create the first group by considering all the degrees and their corresponding frequencies such that the frequency for every cell is at least 32. This group has huge nonlinearity in the frequencies, is also the reason for the popularity of power-law. Another way of analyzing the powers is to consider log (which makes it linear), and work on the log values. Considering logarithms for analyzing nonlinear systems is a popular ploy in many applications. The reason for considering 32 is that, it becomes 5 once logarithm with base 2 is applied on the degree distribution, and for a chi-square test, generally, the frequency of every cell is taken to be minimum 5 [36]. We tried to fit a distribution to the values of the original variable and its corresponding modified frequency value. Where the modified frequency value is at least 5. It may be noted that the chi square goodness of fit test works well for those data sets for which every observed frequency value is at least 5. Thus we considered here all those bins, where for each such bin, the modified frequency value is not less than 5.

Next we create the second group by considering all the degrees and their corresponding frequencies such that the frequency of every cell is at least 5 and not more than 31. Thus, in this case, no log-plot is considered for the frequency distribution table. Original frequencies are considered in this case. That is, unlike the previous case, there are no transformations from one space to another.

Lastly the third group is formed by considering remaining degrees and their corresponding frequencies. Here in this group no distribution will fit properly because of sparseness of the data, so we need to accumulate the data into some intervals of fixed size and estimate the frequencies using this distribution. Here we have considered the interval length as 25 for the third segment.

4.2. Fitting truncated distributions

After the data is divided into three segments, we tried to fit several distributions (truncated and non-truncated) on the data sets. They are Poisson, power-law, exponential, geometric, Pareto, log-normal, drift power-law and power-law with exponential cutoff. On every segment of every data set, every one of the five distributions is fit in two ways, truncated and non-truncated. While fitting a truncated distribution, the range of the variable is taken to be the length of the line segment under consideration.

Table 4
Error in fitting truncated log-normal distribution to some data sets.

Name of the data sets	No. of nodes	No. of edges	No. of intervals		Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
CIT-HEP-PH	34 546	421 578	segment 1	81	31.24	100.75	Accepted
(in degree)			segment 2	61	234.54	77.93	Rejected
			segment 3	10	46.31	16.91	Rejected
SOC-EPINIONS	75 879	508 837	segment 1	68	92.31	85.99	Rejected
(in degree)			segment 2	95	110	116.51	Accepted
			segment 3	20	49.32	31.24	Rejected
CA-ASTRO-PH	18 772	396 160	segment 1	76	32.54	95.08	Accepted
(in degree)			segment 2	75	212.43	93.94	Rejected
			segment 3	5	29.45	11.07	Rejected
WEB-GOOGLE	875 713	5 105 039	segment 1	141	192.86	167.51	Rejected
(in degree)			segment 2	152	643.56	179.58	Rejected
			segment 3	21	67.43	31.41	Rejected

Table 5
Error in fitting truncated power-law distribution to some data sets.

Name of the data sets	No. of nodes	No. of edges	No. of intervals		Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
CIT-HEP-PH	34 546	421 578	segment 1	81	145.98	100.75	Rejected
(in degree)			segment 2	61	193.75	77.93	Rejected
			segment 3	10	23.12	16.91	Rejected
SOC-EPINIONS	75 879	508 837	segment 1	68	34.34	85.99	Accepted
(in degree)			segment 2	95	88.41	116.51	Accepted
			segment 3	20	33.37	31.24	Rejected
CA-ASTRO-PH	18 772	396 160	segment 1	76	132.56	95.08	Rejected
(in degree)			segment 2	75	443.32	93.94	Rejected
			segment 3	5	21.63	11.07	Rejected
WEB-GOOGLE	875 713	5 105 039	segment 1	141	201.69	167.51	Rejected
(in degree)			segment 2	152	656.78	179.58	Rejected
			segment 3	21	75.57	31.41	Rejected

Table 6
Error in fitting truncated Pareto distribution to some data sets.

Name of the data sets	No. of nodes	No. of edges	No. of intervals		Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
CIT-HEP-PH	34 546	421 578	segment 1	81	151.98	100.75	Rejected
(in degree)			segment 2	61	195.2	77.93	Rejected
			segment 3	10	26.12	16.91	Rejected
SOC-EPINIONS	75 879	508 837	segment 1	68	76.34	85.99	Accepted
(in degree)			segment 2	95	91.64	116.51	Accepted
			segment 3	20	34.66	31.24	Rejected
CA-ASTRO-PH	18 772	396 160	segment 1	76	134.49	95.08	Rejected
(in degree)			segment 2	75	448.92	93.94	Rejected
			segment 3	5	29.3	11.07	Rejected
WEB-GOOGLE	875 713	5 105 039	segment 1	141	199.99	167.51	Rejected
(in degree)			segment 2	152	612.89	179.58	Rejected
			segment 3	21	69.13	31.41	Rejected

Truncated distributions are expected to fit better than the non-truncated distributions since the probability of points outside the range is taken to be zero. Even under the assumption of truncated variable, the standard distributions are not fitting properly for the data sets under consideration. Some of these results are shown in the table given below.

Note that Tables 4–10 provide results on some data sets, and using some distributions (truncated log-normal, truncated power-law, truncated Pareto, truncated Poisson, truncated drift power-law, truncated power-law with exponential cutoff, truncated exponential). The parameter estimation for every one of these distributions (truncated or non-truncated) is done

Table 7
Error in fitting truncated Poisson distribution to some data sets.

Name of the data sets	No. of nodes	No. of edges	No. of intervals		Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
CIT-HEP-PH	34 546	421 578	segment 1	81	187.32	100.75	Rejected
(in degree)			segment 2	61	384.32	77.93	Rejected
			segment 3	10	56.23	16.91	Rejected
SOC-EPINIONS	75 879	508 837	segment 1	68	111.87	85.99	Rejected
(in degree)			segment 2	95	532.67	116.51	Rejected
			segment 3	20	41.83	31.24	Rejected
CA-ASTRO-PH	18 772	396 160	segment 1	76	121.39	95.08	Rejected
(in degree)			segment 2	75	419.49	93.94	Rejected
			segment 3	5	28.92	11.07	Rejected
WEB-GOOGLE	875 713	5 105 039	segment 1	141	232.78	167.51	Rejected
(in degree)			segment 2	152	789.48	179.58	Rejected
			segment 3	21	67.32	31.41	Rejected

Table 8
Error in fitting truncated drift power-law distribution to some data sets.

Name of the data sets	No. of nodes	No. of edges	No. of intervals		Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
CIT-HEP-PH	34 546	421 578	segment 1	81	64.42	100.75	Accepted
(in degree)			segment 2	61	102.33	77.93	Rejected
			segment 3	10	35.2	16.91	Rejected
SOC-EPINIONS	75 879	508 837	segment 1	68	49.21	85.99	Accepted
(in degree)			segment 2	95	125.1	116.51	Rejected
			segment 3	20	42.32	31.24	Rejected
CA-ASTRO-PH	18 772	396 160	segment 1	76	62.12	95.08	Accepted
(in degree)			segment 2	75	114.35	93.94	Rejected
			segment 3	5	28.91	11.07	Rejected
WEB-GOOGLE	875 713	5 105 039	segment 1	141	181.55	167.51	Rejected
(in degree)			segment 2	152	620.73	179.58	Rejected
			segment 3	21	79.42	31.41	Rejected

Table 9
Error in fitting truncated power-law cutoff distribution to some data sets.

Name of the data sets	No. of nodes	No. of edges	No. of intervals		Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
CIT-HEP-PH	34 546	421 578	segment 1	81	69.22	100.75	Accepted
(in degree)			segment 2	61	116.2	77.93	Rejected
			segment 3	10	39.32	16.91	Rejected
SOC-EPINIONS	75 879	508 837	segment 1	68	59.73	85.99	Accepted
(in degree)			segment 2	95	103.2	116.51	Accepted
			segment 3	20	45.81	31.24	Rejected
CA-ASTRO-PH	18 772	396 160	segment 1	76	70.11	95.08	Accepted
(in degree)			segment 2	75	101.23	93.94	Rejected
			segment 3	5	25.28	11.07	Rejected
WEB-GOOGLE	875 713	5 105 039	segment 1	141	172.55	167.51	Rejected
(in degree)			segment 2	152	723.51	179.58	Rejected
			segment 3	21	82.3	31.41	Rejected

using maximum likelihood principle. Even with the non-truncated distributions the results are similar. The results are also similar for other data sets too. There are totally 23 data sets. 22 data sets are divided into three segments. One data set is divided into two segments. In total there are 68 segments.

No distribution is fitting well for any segment of any data sets except for few where truncated Pareto, power-law, log-normal, drift power-law, power-law with exponential cutoff and exponential fitting well only for first and second part of the data. This resulted in fitting a truncated geometric distribution on every segment of every data set.

Table 10
Error in fitting truncated exponential distribution to some data sets.

Name of the data sets	No. of nodes	No. of edges	No. of intervals		Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$
CIT-HEP-PH	34 546	421 578	segment 1	81	192.35	100.75	Rejected
(in degree)			segment 2	61	76.2	77.93	Accepted
			segment 3	10	32.1	16.91	Rejected
SOC-EPINIONS	75 879	508 837	segment 1	68	151.23	85.99	Rejected
(in degree)			segment 2	95	111.21	116.51	Accepted
			segment 3	20	43.25	31.24	Rejected
CA-ASTRO-PH	18 772	396 160	segment 1	76	143.22	95.08	Rejected
(in degree)			segment 2	75	89.11	93.94	Accepted
			segment 3	5	18.23	11.07	Rejected
WEB-GOOGLE	875 713	5 105 039	segment 1	141	252.1	167.51	Rejected
(in degree)			segment 2	152	192.5	179.58	Rejected
			segment 3	21	60.38	31.41	Rejected

Now we are proposing to fit truncated geometric distribution, a special case of negative binomial distribution and it has been found that in all the cases it fits properly (*i.e.* observed chi-square value is less than the actual chi-square value) over each segment of data and produces less significant error compared to power-law and other distributions. Here we elaborately describe the procedure for fitting truncated geometric distribution over the data sets.

The geometric distribution is a discrete distribution for $x = 1, 2, \ldots$ having probability mass function

$$f(x) = p(1-p)^{x-1} = pq^{x-1} \tag{7}$$

where $0 < p < 1$, $q = 1 - p$, and the cumulative distribution function is

$$F(x) = \sum_{k=1}^{x} f(k) = 1 - q^x, \quad x = 1, 2, 3, \ldots$$

Let $X_1, X_2, \ldots, X_n$ be a random sample of independent observations from a geometric distribution with parameter p, then the maximum-likelihood estimate of p is (or p can be estimated as)

$$\hat{p} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1} = n/\sum_{i=1}^{n} X_i$$

Let X be a random variable having probability mass function $f(x)$ (Eq. (7)) and if the truncation points on the left and the right are at a and b respectively then the truncated probability mass function [37] of X from Eq. (7) is

$$f(x) = \begin{cases} \frac{pq^{x-1}}{q^a - q^{b-1}}, & x = a+1, a+2, \ldots, b-1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The transformations $y = (x - a)$, and $d = (b - a)$ reduce 8 to

$$g(y) = \begin{cases} \frac{pq^{y-1}}{1-q^{d-1}}, & y = 1, 2, \ldots, d-1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Eq. (9) is a geometric distribution singly truncated on the right at the point $y = d$. In the derivation of the maximum likelihood estimator of the transformed variable Y, the mass function $g(y)$ (Eq. (9)) will be used.

Maximum Likelihood estimate (MLE) of parameters of truncated geometric distribution For a random sample of size n, the likelihood function is

$$L = \prod_{i=1}^{n} g(y_i) = \left(1 - q^{d-1}\right) p^n q^{\left(\sum_{i=1}^{n} y_i - n\right)}$$

For convenience we typically work with the logarithm of L, which is equal to the log of the likelihood, denoted by ℓ and given by,

$$\ell = \log(L) = -n\log\left(1 - q^{d-1}\right) + n\log(p) + \left(\sum_{i=1}^{n} y_i - n\right)\log(q)$$

Now we calculate the most likely value of q by maximizing the likelihood with respect to q, which is the same as maximizing the log likelihood, since the logarithm is a monotonic increasing function. Setting $\frac{\partial \ell}{\partial q} = 0$, we find

$$(nd-n)\left(\frac{q^{d-2}}{1-q^{d-1}}\right) - \frac{n}{1-q} + \left(\sum_{i=1}^{n} y_i - n\right)\frac{1}{q} = 0$$

$$(nd-n)q^{d-1}(1-q) - nq\left(1-q^{d-1}\right) + \left(\sum_{i=1}^{n} y_i - n\right)(1-q)\left(1-q^{d-1}\right) = 0$$

$$\hat{q}\left(\sum_{i=1}^{n} y_i - nd + n\right) + \hat{q}^{d-1}\left(nd - \sum_{i=1}^{n} y_i\right) - \hat{q}\left(\sum_{i=1}^{n} y_i\right) + \left(\sum_{i=1}^{n} y_i - n\right) = 0 \tag{10}$$

The above equation is a d^{th} degree polynomial in $\hat{q}$, and it gives the maximum likelihood estimate of $q(\hat{p} = 1-\hat{q})$. Given value of d, n, and $\sum_{i=1}^{n} y_i = n\bar{y}$, one can compute the value of $\hat{q}$ using an iterative technique such as Newton–Raphson method to solve Eq. (10). By substituting $\sum_{i=1}^{n} y_i = n\bar{y}$, Eq. (10) becomes

$$\bar{y} = \frac{(\hat{q}^d(d-1) - d\hat{q}^{d-1} + 1)}{(\hat{q}^d - \hat{q}^{d-1} - \hat{q} + 1)} \tag{11}$$

While fitting truncated geometric distribution over the degree frequency table, we will be using Eq. (11) for estimating the parameters.

5. Goodness-of-fit test

Given an observed data set (degree frequency table) and a truncated geometric distribution from which it is hypothesized that the data is drawn, we want to know whether that hypothesis is valid except for the last part. That is, could the data we see have plausibly been drawn from the specified truncated geometric distributions? Questions of this type can be answered using goodness-of-fit tests that compare the observed data to the hypothesized distributions. One of the simplest and widely used procedures is to do a chi-square test on the expected frequencies with respect to the observed frequencies.

As we have seen, one can measure how closely a hypothetical distribution resembles the actual distribution of an observed set of samples by calculating chi-square statistic. The calculation returns a single number that is smaller for hypothesized distributions that are a better fit to the data.

Our approach in this section is to calculate this number in three distinct non-overlapping parts of the observed data set and the best fit truncated geometric distribution computed as described above. Then, if this value is suitably small we can say that the truncated geometric distribution is a plausible fit to the data.

6. Testing for real world networks

In this section we examine a large number of real world complex networks representing measurements of quantities whose distribution, it has been conjectured, follow power-laws. Over each distinct part of every data set, we fit and test geometric distribution using the methods described in the previous section. For strengthening the argument that the truncated geometric distribution provides better fitting compared to power-law, log-normal, Pareto, drift power-law and power-law with exponential cutoff distribution over each segment of the data, we have tested it over 23 large real world networks data.

The data sets we study here come from a broad variety of different branches. We present results of fitting truncated geometric distribution to 23 large real-world networks which are available online [38]: Large online social networks (Who Trust Whom Network of Epinions.com (SOC-EPINIONS), Live Journal online social network (SOC-LIVE JOURNAL), Salshdot social network from November 2008 (SOC-SALSHADOT 2008), Salshdot social network from February 2009 (SOC-SALSHADOT 2009), Wikipedia Who Votes On Whom network (WIKI VOTE)), communication networks (Email network from EU research institution (EMAIL-EU-ALL), Email Communication network from Enron (EMAIL-ENRON), Wikipedia talk network (WIKI-TALK)), citation networks (Arxiv High Energy Physics paper citation network (CIT-HEP-PH), Arxiv High Energy Physics Theory citation network (CIT-HEP-TH), Citation network among US Patents (CIT-PATENTS)), collaboration networks of co-authorships from DBLP and various areas of physics (Collaboration network of Arxiv Astro Physics (CA-ASTRO-PH), Collaboration network of Arxiv Condensed Matter (CA-COND-MAT), Collaboration network of Arxiv General Relativity (CA-GR-QC), Collaboration network of Arxiv High Energy Physics (CA-HEP-PH), Collaboration network of Arxiv High Energy Physics Theory (CA-HEP-TH)), product co-purchasing networks (Amazon product co-purchasing network from March 2 2003 (AMAZON-03-02), Amazon product co-purchasing network from March 12 2003 (AMAZON-03-12), Amazon product co-purchasing network from May 5 2003 (AMAZON-05-05), Amazon product co-purchasing network from June 1 2003 (AMAZON-06-01)), web

Table 11
Social networks.

Name of the data sets	No. of nodes	No. of intervals		Estimated q	Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$ and $\gamma = 0.01$
SOC-EPINIONS	75 879	segment 1	68	0.99	11.28	85.99	Accepted(α)
(in degree)		segment 2	95	0.95	43.86	116.51	Accepted(α)
		segment 3	20	0.84	31.24	31.42	Accepted(α)
SOC-LIVE JOURNAL	4 847 571	segment 1	450	0.997	14.313	500.14	Accepted(α)
(in degree)		segment 2	320	0.94	151.76	341.39	Accepted(α)
		segment 3	54	0.913	102.53	72.13	Rejected
SOC-SALSHADOT 2008	77 360	segment 1	94	0.998	16.54	115.39	Accepted(α)
(in degree)		segment 2	119	0.96	76.97	143.25	Accepted(α)
		segment 3	17	0.825	18.62	26.29	Accepted(α)
SOC-SALSHADOT 2009	82 168	segment 1	96	0.999	23.83	117.62	Accepted(α)
(in degree)		segment 2	120	0.969	81.38	144.35	Accepted(α)
		segment 3	16	0.839	20.62	24.99	Accepted(α)
WIKI VOTE	7115	segment 1	39	0.999	1.533	52.19	Accepted(α)
(out degree)		segment 2	66	0.972	31.357	83.76	Accepted(α)
		segment 3	11	0.614	8.79	18.39	Accepted(α)

Table 12
Communication networks.

Name of the data sets	No. of nodes	No. of intervals		Estimated q	Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$ and $\gamma = 0.01$
EMAIL-EU-ALL	265 214	segment 1	21	0.98	11.94	30.14	Accepted(α)
(in degree)		segment 2	65	0.96	32.57	82.53	Accepted(α)
		segment 3	28	0.89	45.35	46.96	Accepted(γ)
EMAIL-ENRON	36 692	segment 1	51	0.99	11.89	66.34	Accepted(α)
(in degree)		segment 2	75	0.96	39.98	93.94	Accepted(α)
		segment 3	18	0.83	26.31	27.58	Accepted(α)
WIKI-TALK	2 394 385	segment 1	103	0.99	21.32	125.45	Accepted(α)
(in degree)		segment 2	183	0.95	124.54	213.39	Accepted(α)
		segment 3	26	0.89	38.61	38.85	Accepted(α)

and blog graphs (Web Graph from Google (WEB-GOOGLE)) and Biological Networks (Protein–protein interaction network in budding yeast (YEASTPPI) [39] and a network of disorders and disease genes (DISEASOME) [40]).[1]

The fit was found to be good for the first two segments of 23 different real world data sets. For the third segment of the data set, we were having intervals of degree of nodes instead of degree of nodes. Due to the unknown nature of selection of interval length and also due to very sparse data, firm conclusions cannot be made on this part of data set. The above table tells us about the limitations of the size of the data (no. of vertices). If the size of the data is small, then segment 3 may not exist. On the other hand if the size is too large, then segment 3 exists but geometric distribution may not always fit properly because of sparseness of the data. If we consider the interval length 25, then in most of the cases segment 3 exists and it has been found that geometric distribution is fitting well here.

Tables 11–15 provide the quantitative measures of the chi-square statistics for goodness of fit test of the proposed method of fitting truncated geometric distributions over different segments of the data. From the above tables it has been clear that over every segment of each data set, the observed chi-square value is less than the actual chi-square value and hence truncated geometric distribution fits well over each segment of the data.

It is to be noted here that the earlier authors, who proposed the power-law, used p-values for validation of their claims [17]. Other authors who consolidated their claim [21] also used p-values in their works. Accordingly, we have also found p-values for some data sets in this regard, both for power-law, log-normal, Pareto, drift power-law, power-law cutoff and for the proposed set up of multiple distributions. However the obtained p-values can't be directly compared since the number of parameters estimated is not same in both the cases. Additionally, the proposed method divides the dataset into a few segments whereas the power-law is fit for the entire dataset, not for segments. Therefore a comparison between the p-values here is unfair.

[1] http://wiki.gephi.org/index.php/Datasets.

Table 13
Citation networks.

Name of the data sets	No. of nodes	No. of intervals		Estimated q	Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$ and $\gamma = 0.01$
CIT-HEP-PH	34 546	segment 1	81	0.986	2.05	100.75	Accepted(α)
(in degree)		segment 2	61	0.97	17.97	77.93	Accepted(α)
		segment 3	10	0.72	10.02	16.91	Accepted(α)
CIT-HEP-TH	27 770	segment 1	57	0.983	0.97	73.31	Accepted(α)
(in degree)		segment 2	69	0.97	14.244	87.11	Accepted(α)
		segment 3	12	0.77	17.54	19.67	Accepted(α)
CIT-PATENTS	3 774 768	segment 1	111	0.986	1.81	134.36	Accepted(α)
(in degree)		segment 2	64	0.96	36.39	81.38	Accepted(α)
		segment 3	4	0.61	7.94	9.48	Accepted(α)

Table 14
Collaboration networks and Web Graph.

Name of the data sets	No. of nodes	No. of intervals		Estimated q	Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$ and $\gamma = 0.01$
CA-ASTRO-PH	18 772	segment 1	76	0.988	2.19	95.08	Accepted(α)
		segment 2	75	0.974	32.764	93.94	Accepted(α)
		segment 3	5	0.61	10.5	11.07	Accepted(α)
CA-COND-MAT	23 133	segment 1	38	0.976	0.06	50.99	Accepted(α)
		segment 2	27	0.94	4.08	37.65	Accepted(α)
		segment 3	5	0.45	10.29	11.07	Accepted(α)
CA-GR-QC	5242	segment 1	18	0.95	0.95	26.29	Accepted(α)
		segment 2	51	0.93	69.57	74.91	Accepted(γ)
		segment 3	*NF*[a]	...	...	...	...
CA-HEP-PH	12 008	segment 1	46	0.99	10.14	60.48	Accepted(α)
		segment 2	64	0.975	74.05	81.38	Accepted(α)
		segment 3	12	0.83	11.25	19.67	Accepted(α)
CA-HEP-TH	9877	segment 1	24	0.99	17.6	33.92	Accepted(α)
		segment 2	17	0.98	24.11	24.99	Accepted(α)
		segment 3	4	0.56	2.53	9.48	Accepted(α)
WEB-GOOGLE	875 713	segment 1	141	0.991	6.69	167.51	Accepted(α)
(in degree)		segment 2	152	0.988	104.29	179.58	Accepted(α)
		segment 3	21	0.86	30.97	31.41	Accepted(α)

[a] Not found.

We have provided statistical evidences of the improved fitting obtained with the truncated geometric distribution model. We leverage one of the popular statistical methods *viz.* Chi-square test to evaluate the Goodness-of-Fit tests of the above mentioned distributions.

Another method of verification is to plot the results obtained, and compare them visually. For this purpose, the log–log plots of the complementary cumulative distribution function (CCDF) of the original frequency, estimated frequency by truncated geometric distribution and the frequency estimated by power-law, log-normal, Pareto, drift power-law and power-law cutoff distribution are made for all the data sets. We have plotted the distribution curve in the original scale after doing the inverse log for segment 1 keeping the rest part as it is. Eight examples are provided in Figs. 3–10. They are the Citation Network (CIT-HEP-PH), Communication Network (Email Communication network from Enron (EMAIL-ENRON)), Social Network (SOC-EPINIONS), Collaboration network of Arxiv Astro Physics (CA-ASTRO-PH), Collaboration network of Arxiv Condensed Matter (CA-COND-MAT), Web Graph from Google (WEB-GOOGLE) and Biological Network (Protein–protein interaction network in budding yeast (YEAST-PPI) and Diseasome Network (DISEASOME)). It is visually clear that the proposed geometric distribution based fit is better than the fit of the power-law, log-normal, Pareto, drift power-law and power-law cutoff distribution in most of the data sets. In some cases log-normal, drift power-law and power-law cutoff provides better fitting than proposed method in the upper part (segment 1) of the distribution.

In all the above cases for the 3rd part of the data, *i.e.*, for segment 3, one needs to do the interval estimation and we found out that, here also geometric distribution fits properly according to the chi-square goodness of fit test and proper choice of interval length. In all the experimental results showed in Tables 4–15, we have considered the interval length as 25 for the third segment.

The following Table 16 provides information about the difference between the actual and mapped frequencies in each segment (Seg. diff.) of HEP Citation network while fitting several distributions including the proposed one. Total difference

Table 15
Product co-purchasing networks and biological networks.

Name of the data sets	No. of nodes	No. of intervals		Estimated q	Observed chi-square value	Actual chi-square value	Null hypothesis decision at $\alpha = 0.05$ and $\gamma = 0.01$
AMAZON-03-02 (in degree)	262 111	segment 1	50	0.97	1.103	65.17	Accepted(α)
		segment 2	35	0.954	6.18	47.41	Accepted(α)
		segment 3	13	0.61	26.52	27.68	Accepted(γ)
AMAZON-03-12 (in degree)	400 727	segment 1	91	0.986	1.54	112.02	Accepted(α)
		segment 2	86	0.978	29.98	106.39	Accepted(α)
		segment 3	15	0.784	27.13	29.14	Accepted(γ)
AMAZON-05-05 (in degree)	410 236	segment 1	92	0.986	1.73	113.14	Accepted(α)
		segment 2	84	0.979	27.41	104.14	Accepted(α)
		segment 3	15	0.79	20.5	23.68	Accepted(α)
AMAZON-06-01 (in degree)	403 394	segment 1	94	0.987	1.39	115.39	Accepted(α)
		segment 2	87	0.98	27.28	107.52	Accepted(α)
		segment 3	12	0.78	17.87	19.67	Accepted(α)
YEAST-PPI	2361	segment 1	16	0.9582	1.32	23.68	Accepted(α)
		segment 2	18	0.885	7.13	28.87	Accepted(α)
		segment 3	2	0.8611	0.287	5.99	Accepted(α)
DISEASOME	3926	segment 1	10	0.95	1.1	18.31	Accepted(α)
		segment 2	20	0.89	5.95	31.41	Accepted(α)
		segment 3	3	0.68	0.5	7.82	Accepted(α)

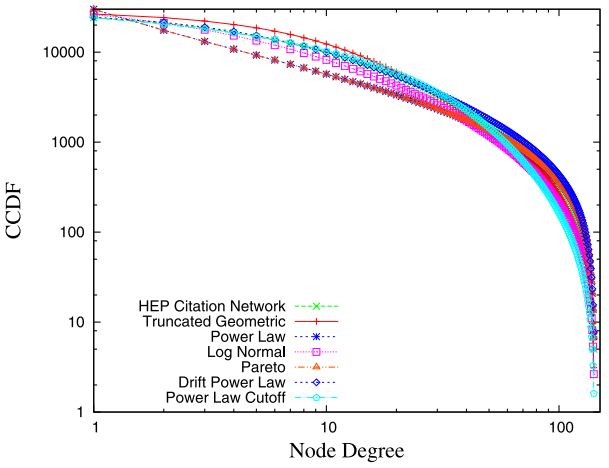

Fig. 3. Fitting the degree distribution of High Energy Physics (HEP) paper citation network in log scale.

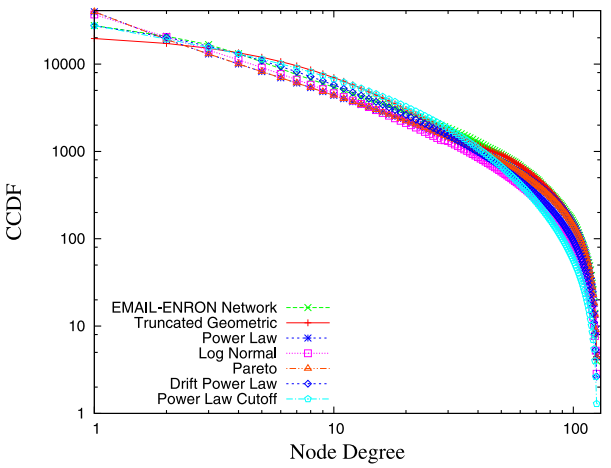

Fig. 4. Fitting the degree distribution of Email Communication network from Enron (EMAIL-ENRON) in log scale.

(Tot. diff.) is the sum of the differences of all the segments. It is clear from the table that the proposed method of fitting truncated geometric distribution provides small differences compared to other distributions, *viz.* power-law, log-normal, Pareto, drift power-law, power-law cutoff in the last two segments.

Table 17 produces the total difference between the actual and mapped frequencies of the whole degree distribution during fitting several distributions (power-law, log-normal, Pareto, drift power-law and power-law cutoff) including the proposed one over some data sets considered by us for visual plotting.

Our proposed methodology provides a way of grouping the individuals in a real world network by dividing it into three distinct and non-overlapping parts based on the degree of the individuals. It assures that different groups come from a specified distribution. Hence, in particular, in an influential social network the present work makes sure the presence of low influential, medium influential and high influential set of individuals whose word-of-mouth affects in the real sense in companies, industries for promoting their products, services, etc. By simulating the parameters of a proposed fit distributions, one can easily capture the spatial structure and dynamics of each group.

7. Conclusions

We have proposed a new scheme of fitting degree distribution instead of power-law over real world networks. The common way of identifying and estimating power-law distribution by the approximately straight line behavior of a histogram

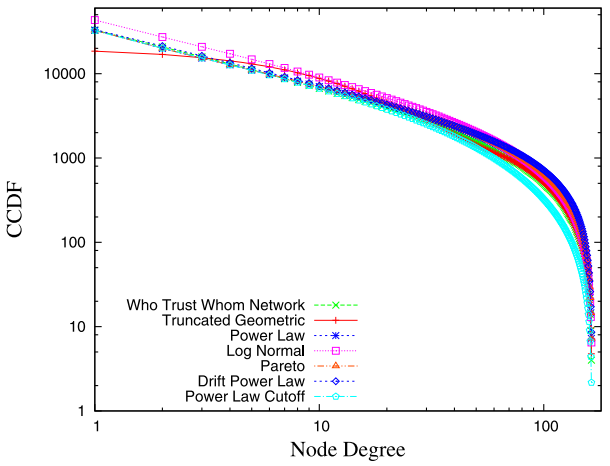

Fig. 5. Fitting the degree distribution of Who Trust Whom network of Epinion in log scale.

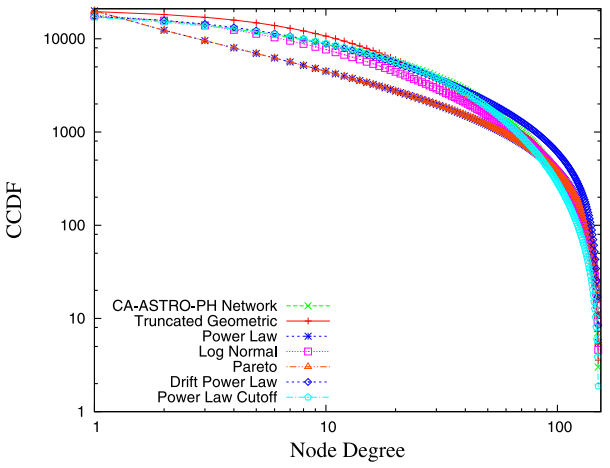

Fig. 6. Fitting the degree distribution of Collaboration network of Arxiv Astro Physics (CA-ASTRO-PH) in log scale.

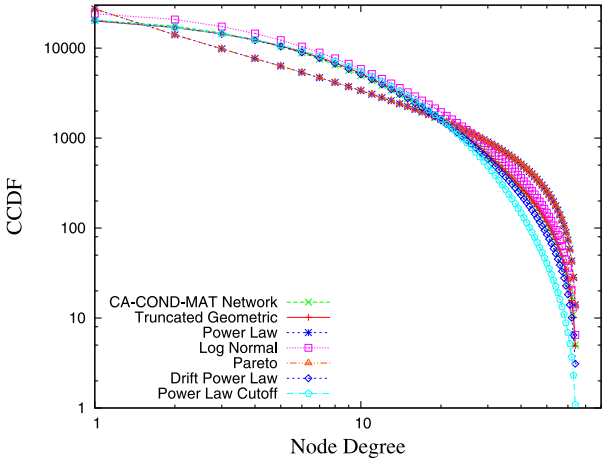

Fig. 7. Fitting the degree distribution of Collaboration network of Arxiv Condensed Matter (CA-COND-MAT) in log scale.

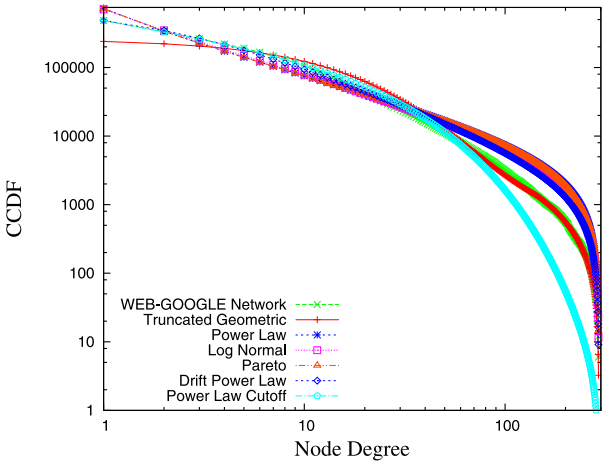

Fig. 8. Fitting the degree distribution of Web-Google network in log scale.

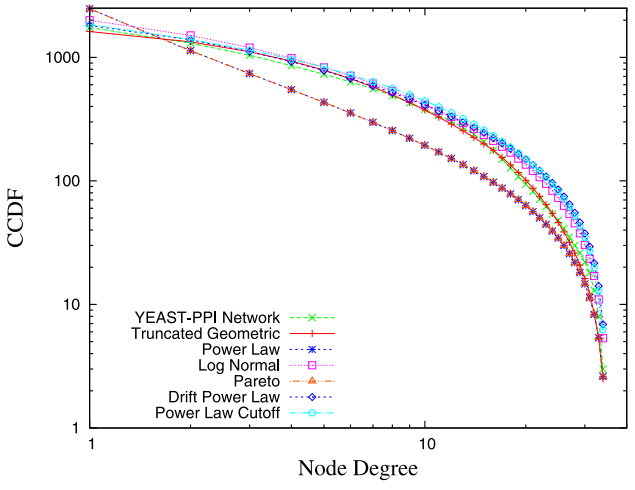

Fig. 9. Fitting the degree distribution of Yeast-PPI network in log scale.

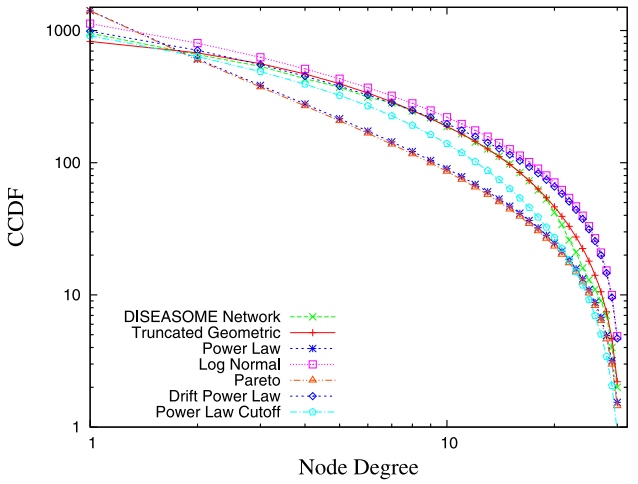

Fig. 10. Fitting the degree distribution of DISEASOME network in log scale.

Table 16
Difference in the actual and mapped plot of HEP Citation network.

No. of intervals		Power-law		Log-normal		Pareto		Drift power-law		Power-law cutoff		Proposed method	
Seg. No.	Int. len.	Seg. diff.	Tot. diff.	Seg. diff.	Tot. diff.	Seg. diff.	Tot. diff.	Seg. diff.	Tot. diff.	Seg. diff.	Tot. diff.	Seg. diff.	Tot. diff.
Seg. 1	81	7408	7624	3234	3374	7412	7630	1124	1415	2424	2662	2962	3060
Seg. 2	61	150		92		151		220		160		62	
Seg. 3	10	66		48		67		71		78		36	

Table 17
Total difference between the actual and mapped plots.

Name of data sets	Total difference					
	Power-law	Log-normal	Pareto	Drift power-law	Power-law cutoff	Proposed method
HEP Citation network	7624	3374	7630	1415	2662	3060
EMAIL-ENRON network	3445	2316	3453	2514	2643	3411
Who Trust Whom network	2592	3031	2601	890	1740	4051
CA-ASTRO-PH network	5224	2245	5226	1785	1984	4104
CA-COND-MAT network	4891	3689	4892	1711	2907	1645
WEB-GOOGLE network	45213	48124	45219	24056	28350	46238
YEAST-PPI network	318	168	320	178	171	153
DISEASOME network	268	141	270	122	123	85

on a log–log plot is known to give a biased result and should not be used in most cases. Here in this paper we have described a simple and alternative way of fitting the truncated geometric distributions over the whole degree frequency table by dividing it into three distinct and non-overlapping parts. We have applied the methods to large number of data sets from a broad range of different fields, it has been found that our method produces less fitting error compared to that of power-law using the chi-square test. In association with the maximum likelihood estimate methods, the chi-square test table given here has been used to provide a quantitative analysis of goodness-of-fit for modeling empirical real world data by multiple truncated geometric distributions.

The aforesaid investigation, resulting in a new concept of modeling the data with multiple truncated geometric distributions, will enable one to capture the structural characteristics of networks. It was empirically found for real world networks involving social, collaboration, communication, citation, biological, etc. that there existed three parts/segments in each of these data sets. Our methodology provides a mode of grouping the individuals, which has an effect on the real world networks. Natural computing models can be used for reliable data analysis in real world networks by dividing it into three distinct and non-overlapping segments. They can be used to transform available heterogeneous data from real world networks into knowledge and provide a proper insight, as in the case of the present work which involves the fitting of multiple truncated geometric distribution. The proposed investigation for identifying the better fit distribution can also be well applicable to capture the dynamics of networks formed by society of agents or objects, molecules. This technique as developed by us provides a plausible and reasonable way of fitting distribution to a wide variety of data sets, but in no way is the only way of fitting a distribution.

Acknowledgements

The authors gratefully acknowledge the financial assistance received in the form of a grant, INT/BRAZIL/IT-P/05/2010 from the Department of Science and Technology, Ministry of Science and Technology, Government of India. S.K. Pal acknowledges the J.C. Bose fellowship of the Govt. of India. The authors thank the anonymous referees for their valuable comments and suggestions. S. Chattopadhyay gratefully acknowledges Mr. R. Das for his helpful discussion.

References

[1] M. Newmann, Network data, available at http://www-personal.umich.edu/~mejn/netdata/.
[2] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, J. Leskovec, Mobile call graphs: beyond power-law and lognormal distributions, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'08, New York, USA, 2008, pp. 596–604.
[3] V. Sessions, Power-law distributions, mitiq.mit.edu x (5) (2005) 1–9.
[4] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, Internet Math. 2 (1) (2004) 226–251.
[5] T. Fenner, M. Levene, G. Loizou, A model for collaboration networks giving rise to a power-law distribution with an exponential cutoff, Soc. Netw. 29 (1) (2007) 70–80.
[6] L.A. Adamic, B.A. Huberman, Zipf's law and the Internet, Glottometrics 3 (2002) 143–150.
[7] A. Sala, H. Zheng, B. Zhao, S. Gaito, Brief announcement: revisiting the power-law degree distribution for social graph analysis, in: Proceeding of the 29th ACM SIGACT–SIGOPS Symposium on Principles of Distributed Computing, New York, USA, 2010, pp. 400–401.
[8] A.L. Barabasi, R. Albert, H. Jeong, G. Bianconi, Power-law distribution of the world wide web, Science 287 (2000) 2115.
[9] A. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[10] R. Albert, H. Jeong, A. Barabasi, Diameter of the world wide web, Nature 401 (1999) 130–131.
[11] L.N. de Castro, Fundamentals of natural computing: an overview, Phys. Life Rev. 4 (2007) 1–36.
[12] L. Kari, G. Rozenberg, The many facets of natural computing, Commun. ACM 51 (10) (2008) 72–83.
[13] S. Mitra, R. Das, Y. Hayashi, Genetic networks and soft computing, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (2011) 94–107.
[14] H. Ressom, D. Wang, P. Natarajan, Clustering gene expression data using adaptive double self-organizing map, Physiol. Genomics 14 (2003) 35–46.
[15] S. Bandini, S. Manzoni, G. Vizzari, Situated cellular agents: a model to simulate crowding dynamics, IEICE Trans. Inf. Syst. E87-D (3) (2004) 669–676.
[16] M.E.J. Newman, Power-law distributions in empirical data, Physics (2000).
[17] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661.
[18] M.E.J. Newman, Power law distribution, SIAM Rev. 45 (2003) 157–256.
[19] A.-L. Barabasi, The origin of bursts and heavy tails in humans dynamics, Nature 435 (2005) 207.
[20] Y. Song, C. Zhang, M. Wu, The study of human behavior dynamics based on blogosphere, in: 2010 International Conference on Web Information Systems and Mining, Sanya, China, 2010.
[21] M.L. Goldstein, S.A. Morris, G.G. Yen, Problems with fitting to the power-law distribution, Eur. Phys. J. B 41 (2) (September 2004) 255–258.
[22] R. Durrett, Random Graph Dynamics, Cambridge University Press, 2007.
[23] R. Albert, A. Barabasi, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (2002) 47–97.
[24] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, Z. Ghahramani, Kronecker graphs: an approach to modeling networks, J. Mach. Learn. Res. 11 (2010) 985–1042.
[25] D. Chakrabarti, Y. Zhan, C. Faloutsos, R-mat: a recursive model for graph mining, in: SDM, 2004.
[26] A. Clauset, M. Young, K.S. Gleditsch, On the frequency of severe terrorist events, J. Confl. Resolut. 51 (1) (February 2007) 58–87.
[27] L. Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, Verlag, Berlin, 2002.
[28] O.E. Barndorff-Nielsen, D.R. Cox, Inference and Asymptotic, Chapman and Hall, London, 1995.
[29] H.L. Seal, The maximum likelihood fitting of the discrete Pareto law, J. Inst. Actuar. 78 (1952) 115–121.
[30] M.L. Goldstein, S.A. Morris, G.G. Yen, Problem with power law distribution, Eur. Phys. J. B 41 (2004) 255.
[31] H. Bauke, Parameter estimation for power-law distributions by maximum likelihood methods, Eur. Phys. J. B 58 (2) (2007) 167–173.
[32] D.J. Choi, On a generalization of the Hurwitz zeta function, Indian J. Pure Appl. Math. 23 (2) (February 1992) 83–91.
[33] J.D. Hamilton, Time Series Analysis, Princeton University Press, Princeton, 1994.
[34] C.A. Murthy, S.K. Pal, Fuzzy thresholding: mathematical framework, bound functions and weighted moving average technique, Pattern Recogn. Lett. 11 (1990) 197–206.
[35] C.A. Murth, S.K. Pal, Histogram thresholding by minimizing gray level fuzziness, Inform. Sci. 60 (1/2) (1992) 107–135.
[36] A.M. Goon, M.K. Gupta, D. Gupta, Fundamentals of Statistics, vol. I, World Press, Calcutta, 1962.
[37] R.L. Thomasson, C.H. Kapadia, On estimation the parameter of a truncated geometric distribution, Themis Signal Analysis Statistics Research Program, Tech. Rep. 5, Sept. 6 1968.
[38] SNAP, Stanford network analysis project (snap) data sets, available at http://snap.stanford.edu/.
[39] S. Sun, L. Ling, N. Zhang, G. Li, R. Chen, Topological structure analysis of the protein–protein interaction network in budding yeast, Nucleic Acids Res. 31 (9) (2003) 2443–2450.
[40] K.I. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabasi, The human disease network (the human diseasome), Proc. Natl. Acad. Sci. USA 104 (2007) 8685–8690.