Soc. Netw. Anal. Min. (2015) 5:45
DOI 10.1007/s13278-015-0283-z

ORIGINAL ARTICLE

Power-law distributions of attributes in community detection

Tai-Chi Wang[1] · Frederick Kin Hing Phoa[1] · Tun-Chieh Hsu[2]

Received: 8 December 2014 / Revised: 1 July 2015 / Accepted: 6 July 2015 / Published online: 31 July 2015

Abstract Community detection has drawn significant attention as new media generates big data every day. To provide statistical testing procedures for community detection in social networks, a scanning method has been developed based on the likelihood of Poisson random graph. However, the scan statistics did not consider detecting communities of the attributes with power-law distribution. Power-law distribution, generally followed by network attributes, is conspicuous in many scientific situations. This paper aims at extending the scanning method to analyze a social network in which attributes follow power-law distribution. Besides the theoretical construction, simulation studies are performed to verify the feasibility of the proposed method, and an authorship network is used to demonstrate the proposed method.

Keywords Community detection · Scanning method · Power-law distribution · Simulation study · Coauthor relationship network

1 Introduction

Clusters or communities, defined as groups of vertices that share common properties or play similar roles in a network, are one of the most important patterns in social networks. Fortunato (2010) summarized some recent development of the community detection methods. Modularity-based method (Newman and Girvan 2004) is arguably the most popular methods in finding communities of networks. Greedy techniques (Newman 2004) and annealing methods (Guimera et al. 2004) were developed based on this criterion. However, the modularity method lacks statistical significance for deciding if the detected communities are real ones and are criticized for hardly finding communities smaller than a given scale (Fortunato and Barthélemy 2007). Bayesian models (Handcock et al. 2007; Heard et al. 2010) and Latent Dirichlet Allocation (LDA) models (Blei et al. 2003; Liu et al. 2009; Balasubramanyan and Cohen 2011) offer statistical inference of clusters via given model priors, but the selection of priors and computing times are judged like other applications of Bayesian models.

Recent studies of social networks also paid attentions on networks with attributes. For example, a tendency called "homophily" (McPherson et al. 2001) suggested that people usually interact with others who are similar to themselves with some attributes (Kossinets and Watts 2006). Some studies also discussed how homophily affects network integration (Louch 2000). Zhou et al. (2009) developed a distance-based transition probability, based on similarities of both structure and attribute, to construct a clustering algorithm. Yang et al. (2013) modeled the links of network and node attributes to provide a probability regime to detect community memberships.

However, few of them considered attributes with power-law (PL) distributions. The PL distribution is one of the most commonly found distributions that many data sets follow in networks (Clauset et al. 2009), such as the degrees of proteins in the protein interaction network and the degrees of metabolites in the metabolic network. In addition, the properties and problems of PL distributions are also well addressed in Goldstein et al. (2004) and

✉ Tai-Chi Wang
taichi43@stat.sinica.edu.tw

[1] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

[2] Department of Statistics, National Chengchi University, Taipei, Taiwan

Clauset et al. (2009). Thus, we intend to construct a community detection method which can accommodate to PL distributions.

To provide a statistical significance of cluster detection without considering the priors, Wang et al. (2008) provided a scanning method for testing clusters based on the idea of cluster detection in the spatial data analysis (Kulldorff 1997). Under the assumption of Poisson random model, Wang et al. (2008) constructed a scan statistic for detecting structure clusters in social networks. Instead of considering the structure clusters, We intend to generalize scan statistics to consider the attribute of networks with PL distributions. Since the scan statistic is originally applied in temporal (Naus 1966) and spatial domains (Kulldorff 1997), the test statistics are basically constructed by the likelihood of attributes. For this reason, we can extend the scan statistic to accommodate to PL distributions with a similar way.

This study aims to extend the use of the scanning method provided in Wang et al. (2008) to the network whose attributes are PL distributed. In Sects. 2 and 3, we introduce the PL distributions including both "discrete" and "continuous" cases, and the scanning method for community detection in social networks. To verify the proposed method, simulation studies are provided in Sect. 4 and a real data set of authorship is analyzed in Sect. 5. We discuss this method and provide some future works.

2 Power-law distributions

Specifying network clusters with PL-distributed attributes is the main focus in this paper. We only mention the topics related to the scanning methods, including density functions and the maximum likelihood estimates (MLEs). If one is interested in the other properties of PL distribution, such as goodness of fit and additional examples of PL distributions, please refer to Goldstein et al. (2004) and Clauset et al. (2009). Note that, we use different notations for density functions and parameters to distinguish between the "discrete" and "continuous" PL distributions.

2.1 Discrete power-law distribution

The probability density function of discrete PL distribution is expressed as

$$p(x) = Mx^{-\alpha},$$

where M is a normalized constant and is usually expressed as

$$M = 1/\zeta(\alpha, x_{\min}) = 1/\sum_{w=0}^{\infty}(w + x_{\min})^{-\alpha}$$

when a lower bound $x_{\min} > 0$ is considered. In a special case of $x_{\min} = 1$, $\zeta(\alpha, 1)$ is equivalent to the Riemann zeta function and is abbreviated as $\zeta(\alpha)$. Observed from its density function, the probability is confined by the value of α and has a heavy tail.

A likelihood ratio of PL distribution is required when applying a PL distribution. Suppose $\{x_1, x_2, \ldots, x_n\}$ is a random vector following PL distribution and only the parameter α is considered. The log-likelihood is

$$\ln L(\alpha) = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln(x_i). \quad (1)$$

By differentiating Eq. (1) at α,

$$\frac{\zeta'(\alpha, x_{\min})}{\zeta(\alpha, x_{\min})} = -\frac{1}{n}\sum_{i=1}^{n} \ln(x_i), \quad (2)$$

where $\zeta'(\alpha)$ is the derivative of the zeta function. Since Eq. (2) contains an infinity summation, there is no closed form of the MLE for α, and the MLE of this distribution cannot be directly found. Numerical algorithm thus becomes a possible way to obtain its solution. When $x_{\min} = 1$, we can quickly solve it by computing the Riemann zeta function whose derivation is provided in most mathematical softwares like Matlab and Maple. On the other hand, an approximated estimate is considered.

$$\hat{\alpha} \simeq 1 + n\left[\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min} - 1/2}\right)\right]^{-1}. \quad (3)$$

when $x_{\min} \geq 6$ (Clauset et al. 2009). In this paper, the data are restricted to $x_{\min} = 1$, so Eq. (2) is applied to construct the likelihood ratio test of the scanning method. If $x_{\min} \geq 6$, we would suggest to directly use the approximate estimation Eq. (3). If $x_{\min}$ is not included in above cases ($1 < x_{\min} < 6$), iterative computation is required to solve the equation (Gillespie 2013). We did not consider the parameter $x_{\min}$ in this study, because it is easier to assume $x_{\min}$ is a known constant and this assumption is usually true in our experience. If the $x_{\min}$ is not certain and has to be estimated, it will be more complicated than the discussion in this study. Clauset et al. (2009) also gave some details of the estimations for parameters α and $x_{\min}$.

2.2 Continuous power-law distribution

Similar to what we have done in the case of discrete PL case, the continuous case is introduced as follows. Since the formulation of the continuous PL distribution can be obtained by integration, the estimation procedures are easier. We also assume that the $x_{\min}$ is a fixed and known parameter. The continuous PL density function is expressed as

$$f(x) = Kx^{-\beta},$$

where K is a normalized constant and is expressed as

$$K = (\beta - 1)x_{\min}^{\beta-1}.$$

Suppose a random sample is collected as $\{x_1, \ldots, x_n\}$. By directly differentiating the parameter α, the MLE of α is obtained as

$$\hat{\beta} = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}. \quad (4)$$

3 The scanning method for PL distributions

The scan statistic is one of the most popular cluster detection methods applied in spatial domain (Kulldorff 1997). However, there are few studies and methods applying this approach to detect clusters in social networks. Wang et al. (2008) first used a scan statistic to detect clusters in a social network. We briefly introduce the basic idea of the scanning method and extend it to networks with PL-distributed attributes.

3.1 Scanning window and test statistic for structure pattern in networks

A scanning window is used to separate the studied region/network into two parts, where a likelihood ratio test is used to evaluate the difference between the selected observations and their complementary observations. Usually, a scanning window is circularly expanded and constructed by a center and a radius. In the network structure, the center is one node in a given network, and the radius is the shortest path from the center node to the other nodes. We demonstrate the construction and move of a scanning window via a grid network in Fig. 1.

Based on this 5 by 5 grid network (Fig. 1a), we clearly observe the distances (the length of shortest paths) among nodes. Thus, it is easy to transform and demonstrate the network into a radius expansion. In this study, we use circular windows to be elective subgraphs. That is, a scanning window is generated based on a center with a corresponding radius, and the set of nodes within the window is the elective subgraph. Take the node 13 as the center for example. In Fig. 1b, the scanning windows with gray boundaries are generated from the center 13, and each window is expanded based on the length of the shortest paths. Suppose we generate a scanning window by the center node 13 and a radius 1. The elective subgraph is demonstrated in Fig. 1c, and we can compare the difference between the selected part and its unselected counterpart. One may check that all the vertices are contained in the windows according to Fig. 1a. In fact, the number of testing regions for a scan statistic is decided based on the number of vertices and the number of radiuses.

Wang et al. (2008) provided explicit descriptions on how to test the structure based on the scan statistics. We recall some notations that appear in the rest of our paper. We only focus on undirected graphs in this study. Let $G = (V_G, E_G)$ be an undirected graph with vertex set $V_G = \{v_1, \ldots, v_{|V_G|}\}$ and edge set E_G, and the degrees of vertices are $\mathbf{k} = \{k_1, \ldots, k_{|V_G|}\}$. In addition, we define the sum of total degrees as $k_G = \sum_{i=1}^{|V_G|} k_i$ and the total number of edges as $|E_G| = k_G/2$. Then, by considering the Poisson random graph model (Erdős and Rényi 1959) with degree vector $\mathbf{k}$, the number of expected edges connecting the pair nodes (v_i, v_j) is expressed as $e_{ij} = (k_i k_j)/(2|E_G|)$ for $i \neq j$, and $e_{ii} = k_i^2/(4|E_G|)$.

Suppose a subgraph $Z = (V_Z, E_Z)$ is selected. Similar notations is used to describe the quantities of Z: $k_Z = \sum_{i \in V_Z} k_i$ and $|E_Z| = k_Z/2$. To test if a network is composed by two different subgraphs, the number of edges in G is equal to Poi $(\lambda = \gamma\mu(Z \bigcap G) + \eta\mu(Z^C \bigcap G))$, where γ and η represent the strengths for subgraph Z and its complementary subset Z^C, and $\mu(Z \bigcap G)$ and $\mu(Z^C \bigcap G)$ represent the expected numbers of edges under the null hypothesis. Under the Poisson random graph assumption, $\mu(G)$, $\mu(Z)$, and $\mu(Z^C)$ are, respectively, defined as

$$\mu(G) = \frac{k_G^2}{4|E_G|},$$
$$\mu(Z) = \frac{k_Z^2}{4|E_G|}, \quad \text{and}$$
$$\mu(Z^C) = \mu(G) - \mu(Z).$$

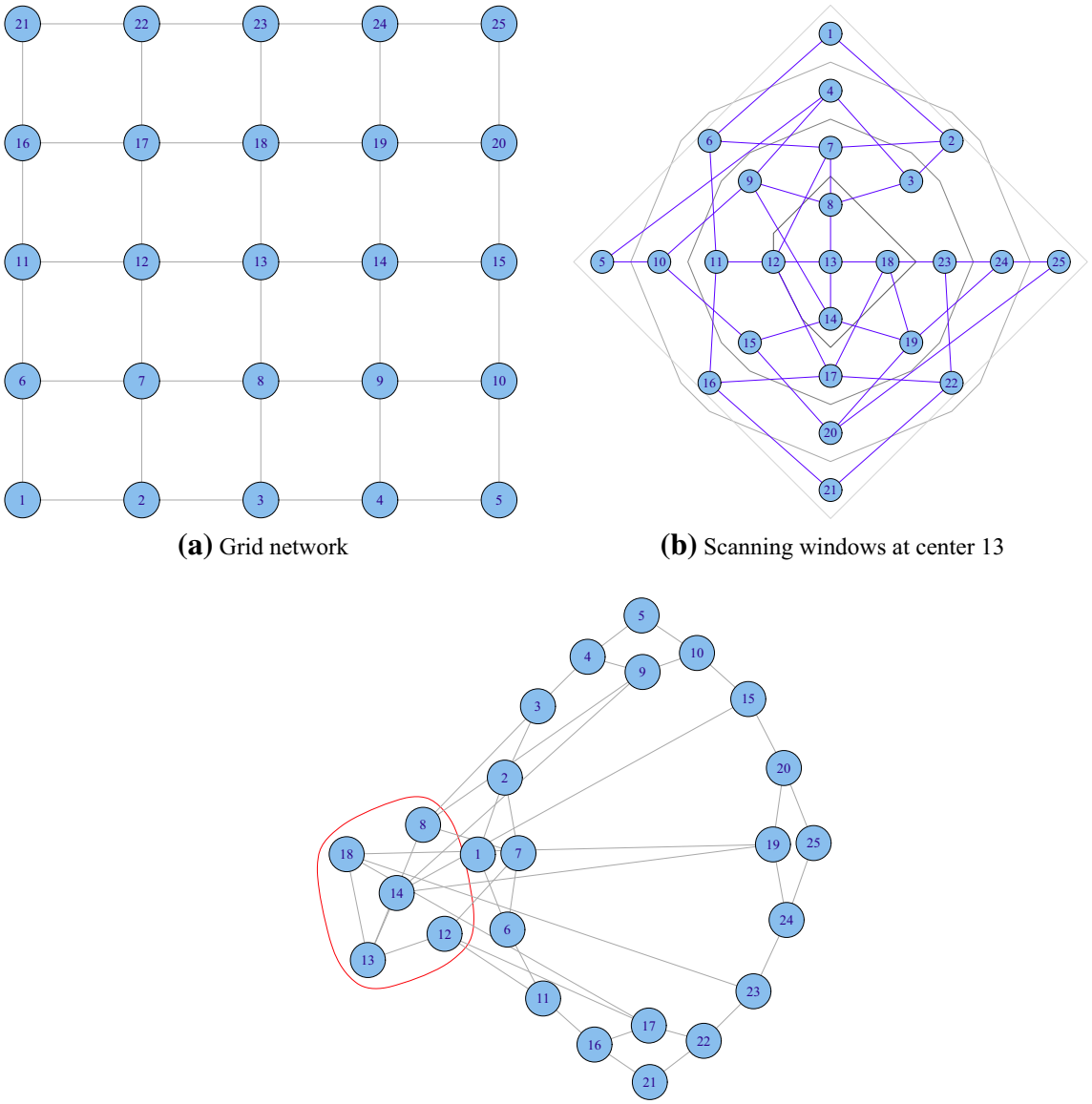
Thus, the likelihood ratio statistic of a selected subgraph Z is

$$LR(Z) = \frac{L_Z}{L_0} = \begin{cases} \left(\frac{|E_Z|}{\mu(Z)}\right)^{|E_Z|} \left(\frac{|E_G| - |E_Z|}{\mu(G) - \mu(Z)}\right)^{|E_G|-|E_Z|} & \text{if} \quad \hat{\gamma} > \hat{\eta} \\ 1 & \text{otherwise,} \end{cases} \quad (5)$$

where $\hat{\gamma} = \frac{|E_Z|}{\mu(Z)}$ and $\hat{\eta} = \frac{|E_G|-|E_Z|}{\mu(G)-\mu(Z)}$. By scanning the whole region, the test statistic is the one with the maximum logarithmic likelihood ratio

$$\lambda_S(Z) = \max_Z \ln LR(Z).$$

The subgraph Z with maximum $LR(\cdot)$ is identified as a cluster if the null hypothesis is rejected.



(a) Grid network

(b) Scanning windows at center 13

(c) Selected subgraph and the unselected counterpart

Fig. 1 Example of scanning window in a grid network

3.2 The hypothesis and likelihood ratio test for discrete PL distribution

A testing statistic for attribute is constructed in a similar manner. Since the data are divided into two parts via a scanning window, the likelihood ratio evaluates the likelihoods between the values within the selected window and the unselected counterpart. In general, the test suggests $H_0 : F(Z) = F(Z^c)$ vs. $H_a : F(Z) < F(Z^c)$, where F is the distribution function. Suppose a subgraph is selected as Z and the parameter of interest is θ. The likelihood ratio statistic is expressed as

$$\lambda(Z, Z^c) = \frac{\sup_{\Theta} L(\theta|x)}{\sup_{\Theta_0} L(\theta|x)},$$

where Θ_0 is the parameter space under the null hypothesis and Θ is the entire parameter.

In this study, the distribution of interest is PL, and we consider the cluster with higher value of PL distribution, or equivalently, the cluster with a smaller parameter α in the PL distribution. Thus, it is equivalent to consider the test $H_0 : \alpha_z \geq \alpha_c$ vs. $H_a : \alpha_z < \alpha_c$, where α_z and α_c are the parameters of PL distribution for the selected observations and their complementary observations, respectively.

Set the minimum value of the PL distribution as 1. Based on the null hypothesis that $H_0 : \alpha_Z = \alpha_{Z^c}$, the joint likelihood of the distribution is

$$L_0(\alpha|\mathbf{x}) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \zeta(\alpha, x_{\min}) x_i^{-\alpha}.$$

By taking logarithm on L_0 and differentiating it with respect to α, the MLE of α is the solution of

$$\frac{\zeta'(\hat{\alpha}_0)}{\zeta(\hat{\alpha}_0)} = -\frac{1}{n}\sum_{i=1}^{n} \ln(x_i).$$

Thus, the denominator of the testing statistic λ is $\sup_{\Theta_0} L(\theta|x) = L_0(\hat{\alpha}_0|\mathbf{x})$.

Two cases on the numerator of the test statistic are considered; $\alpha_z \geq \alpha_c$ and $\alpha_z < \alpha_c$. When $\alpha_z \geq \alpha_c$, the numerator of the test statistic reduces to the null hypothesis and obtains the same estimate of the denominator. When $\alpha_z < \alpha_c$, the joint likelihood is viewed as two parts; one belongs to the selected subgraph Z and the other one is the complementary set of Z, Z^c. Then, the joint likelihood of the distribution for this case is

$$L_\Theta(\alpha|\mathbf{x}) = \prod_{i\in Z} p(x_i) \prod_{j\in Z^c} p(x_j).$$

Since x_i and x_j belong to different αs and are independent, we separately discuss the estimates of α_z and α_c. The MLEs of them are the solutions of

$$\begin{aligned} \frac{\zeta'(\hat{\alpha}_z)}{\zeta(\hat{\alpha}_z)} &= -\frac{1}{n_z}\sum_{i\in Z} \ln(x_i) \quad \text{and} \\ \frac{\zeta'(\hat{\alpha}_c)}{\zeta(\hat{\alpha}_c)} &= -\frac{1}{n_c}\sum_{j\in Z^c} \ln(x_j), \end{aligned} \tag{6}$$

where n_z and n_c are the numbers of nodes in Z and Z^c, respectively. Denote the estimates as $\hat{\alpha}_z$ and $\hat{\alpha}_c$. The numerator of the testing statistic is $\sup_\Theta L(\theta|x) = L(\hat{\alpha}_z, \hat{\alpha}_c|\mathbf{x})$ when $\hat{\alpha}_z < \hat{\alpha}_c$, and $\sup_\Theta L(\theta|x) = L_0(\hat{\alpha}_0|\mathbf{x})$ otherwise.

According to above description, the test statistic is $\lambda(Z) = \frac{L(\hat{\alpha}_z, \hat{\alpha}_c|\mathbf{x})}{L(\hat{\alpha}_0|\mathbf{x})}$ when $\hat{\alpha}_z < \hat{\alpha}_c$, and $\lambda(Z) = 1$ otherwise. By considering the real form of the PL distribution, the likelihoods for the null hypothesis and the alternative hypothesis are

$$L(\hat{\alpha}_0) = \prod_{i=1}^{n}\left[\frac{x_i^{-\hat{\alpha}_0}}{\zeta(\hat{\alpha}_0, x_{\min})}\right], \quad \text{and}$$

$$L(\hat{\alpha}_z, \hat{\alpha}_c) = \prod_{i\in Z}\left[\frac{x_i^{-\hat{\alpha}_z}}{\zeta(\hat{\alpha}_z, x_{\min})}\right] \prod_{j\in Z^c}\left[\frac{x_j^{-\hat{\alpha}_c}}{\zeta(\hat{\alpha}_c, x_{\min})}\right].$$

By scanning the whole region via some predetermined radii, the test statistic for detecting clusters is

$$\lambda_A(Z) = \max_{Z\in\Omega} \lambda(Z).$$

The logarithm form of the test statistic is expressed as

$$\begin{aligned} \Lambda_A(Z) &= \ln \lambda_A(Z) \\ &= -\hat{\alpha}_z \sum_{i\in Z} \ln(x_i) - n_z \ln(\zeta(\hat{\alpha}_z, x_{\min})) - \hat{\alpha}_c \sum_{j\in Z^c} \ln(x_j) \\ &\quad - n_c \ln(\zeta(\hat{\alpha}_c, x_{\min})) + \hat{\alpha}_0 \sum_{i\in\Omega} \ln(x_i) + n \ln(\zeta(\hat{\alpha}_0, x_{\min})), \end{aligned} \tag{7}$$

where n, n_z, and n_c are the numbers of nodes in the whole graph, the selected subgraph Z, and the complementary subgraph Z^c, respectively.

3.3 The hypothesis and likelihood ratio test for continuous PL distribution

Similar to the discrete PL distribution, we construct the testing statistic for the continuous PL distribution. Since we have the MLE of the continuous case from Eq. (4), we can directly apply the same construction of the test statistic for the continuous PL distribution. For the null hypothesis, $H_0 : \beta_Z \geq \beta_{Z^c}$, the joint likelihood of the distribution is

$$L_0(\beta|\mathbf{x}) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} (\beta - 1) x_{\min}^{\beta-1} x_i^{-\beta}.$$

By taking logarithm on L_0 and differentiating it with respect to β, the MLE of β is

$$\hat{\beta}_0 = 1 + n\left[\sum_{i=1}^{n} \ln\frac{x_i}{x_{\min}}\right]^{-1}.$$

Thus, the denominator of the testing statistic λ is $\sup_{\Theta_0} L(\theta|x) = L_0(\hat{\beta}_0|\mathbf{x})$.

Furthermore, for the case $\beta_z < \beta_c$, the MLEs of them are

$$\begin{aligned} \hat{\beta}_Z &= 1 + n\left[\sum_{i\in Z} \ln\frac{x_i}{x_{\min}}\right]^{-1} \quad \text{and} \\ \hat{\beta}_Z^c &= 1 + n\left[\sum_{j\in Z^c} \ln\frac{x_j}{x_{\min}}\right]^{-1}. \end{aligned} \tag{8}$$

where n_z and n_c are the numbers of nodes in Z and Z^c, respectively.

According to above description, the test statistic is $\lambda(Z) = \frac{L(\hat{\beta}_z, \hat{\beta}_c|\mathbf{x})}{L(\hat{\beta}_0|\mathbf{x})}$ when $\hat{\beta}_z < \hat{\beta}_c$, and $\lambda(Z) = 1$ otherwise. By

considering the real form of the PL distribution, the likelihoods for the null hypothesis and the alternative hypothesis are

$$L(\hat{\beta}_0) = \prod_{i=1}^{n}\left[\left(\hat{\beta}_0 - 1\right)x_{\min}^{\hat{\beta}_0-1}x_i^{-\hat{\beta}_0}\right],$$

and

$$L(\hat{\beta}_z, \hat{\beta}_c) = \prod_{i\in Z}\left[\left(\hat{\beta}_z - 1\right)x_{\min}^{\hat{\beta}_z-1}x_i^{-\hat{\beta}_z}\right]\prod_{j\in Z^c}\left[\left(\hat{\beta}_c - 1\right)x_{\min}^{\hat{\beta}_c-1}x_i^{-\hat{\beta}_c}\right].$$

Thus, the test statistic for detecting clusters with the continuous PL distribution is expressed as

$$\begin{aligned}\Lambda_A(Z) = \ln\lambda(Z) = {} & n_z\ln\left(\hat{\beta}_z - 1\right) - \hat{\beta}_z\sum_{i\in Z}\ln\frac{x_i}{x_{\min}} \\ & + n_c\ln\left(\hat{\beta}_c - 1\right) - \hat{\beta}_c\sum_{i\in Z^c}\ln\frac{x_i}{x_{\min}} \\ & - n\ln\left(\hat{\beta}_0 - 1\right) + \hat{\beta}_0\sum_{i\in\Omega}\ln\frac{x_i}{x_{\min}}, \end{aligned} \quad (9)$$

where n, n_z, and n_c are the numbers of nodes in the whole graph, the selected subgraph Z, and the complementary subgraph Z^c, respectively.

3.4 Testing procedure

Due to a large set of selected subgraphs, scanning method often suffers the multiple testing problem. The Monte Carlo testing is a suggested solution to this problem Kulldorff (1997). For the case of attribute pattern, we directly apply the method provided in Kulldorff (1997) in which a randomized permutation of observation is suggested. For example, when a PL distribution is applied, the Monte Carlo procedure randomly permutes the observations and assigns the values for each node.

When a new graph is generated, we evaluate the new data by the same test statistics provided in Sects. 3.2 and 3.3. Suppose a simulation with a large number of iteration, such as 99 or 999, is executed. A Monte Carlo p value with R runs is computed as

$$p = \frac{\#\{\Lambda_r \geq \Lambda_{\text{obs}}\} + 1}{R + 1}, \quad (10)$$

i.e., the probability of finding more extreme values than the observed value. If the p value is smaller than a prespecified criterion (e.g., 0.05), it is statistically significant to declare that there is a cluster. That is, the observed value is less likely to happen under the null hypothesis. The details about the Monte Carlo simulations in other distributions are referred to Kulldorff (1997).

4 Simulation study

4.1 Simulation settings

In this section, we generate a series of random data to testify the type I error and the testing power for detecting a single cluster with power-law distributions based on our proposed method. For the network structure, we restrict the study region with 100 nodes and the edges among them are set to follow a Bernoulli distribution with connection probability $p_0 = 1/20$, i.e., expected degree of each node is 5. To verify the power of our proposed method, we set a cluster in the network with a variate size S (10, 15, and 20) with higher connection probability of edges ($p_c = 1/4$, 1/2, 3/4, and 1). For the consideration of attribute, we set a power-law distribution for usual nodes with the parameters $\alpha = 2.5$ and $x_{\min} = 1$, and set that for cluster nodes with the lower α (from 1.5 to 2.0 in steps by 0.1).

We illustrated a simulated network in Fig. 2. In this example, we set the cluster size as 20, connection probability as 1/2 for the cluster nodes and 1/20 for the usual nodes, and the parameter αs of discrete PL attribute is 1.8 and 2.5 for the cluster nodes and usual nodes, respectively. The node labels are the values of discrete PL attributes. From Fig. 2, the cluster nodes clearly have higher connections than other nodes. However, it is not trivial to observe the difference by eye for the cluster of attribute, even if we set a large difference between usual nodes and cluster nodes. Thus, it is necessary to have an automatic tool to specify the clusters.

4.2 Type I error

For testing the feasibility of our proposed method, we conduct a simulation for type I error. In this subsection, different connection probabilities ($p_0 =$ 1/20, 1/15, 1/10, and 1/5) are considered to check the performance of our proposed method. In each simulation case, we executed 1,000 runs to assess sampling fluctuations. If at least a subgraph among the constructed network is statistically significant, the type I error occurs. Table 1 shows that the type I errors are very consistent and well performed around 0.05 for both the discrete and continuous power-law distributions.

We will use $p_0 = 1/20$ in the following simulations since $p_0 = 1/20$ is the usual case in real data.

4.3 Testing the attribute clusters

In our past studies, we realize that the cluster size and connection strength greatly affect the detection power, so

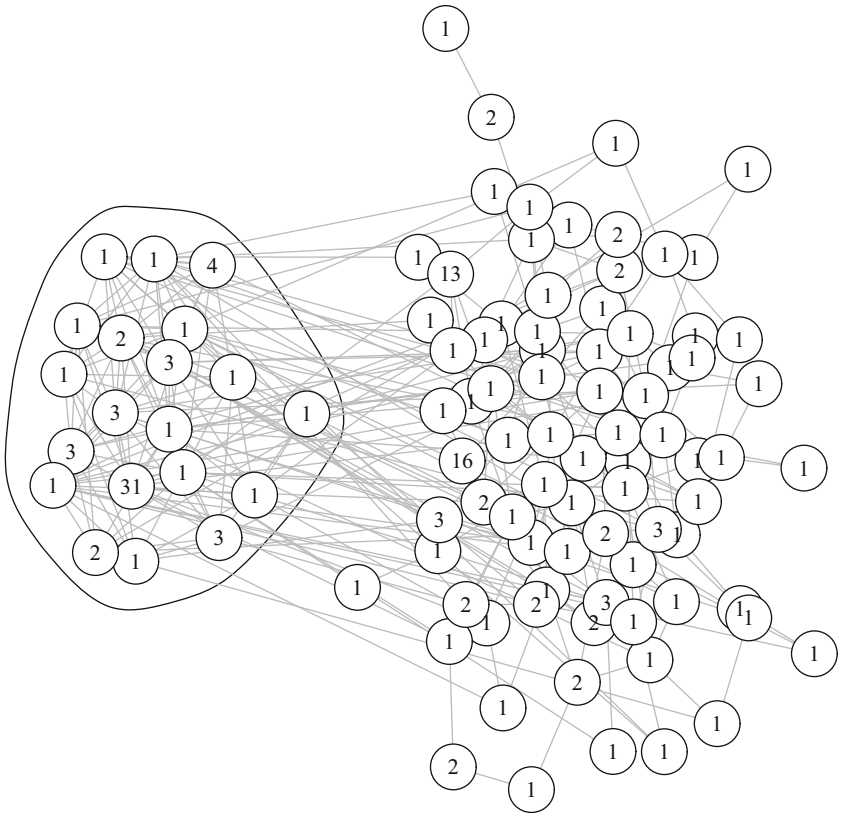


Fig. 2 A simulated cluster network with size 20, connection probability 1/2, and discrete power-law distribution $\alpha = 1.8$

Table 1 Type I error

Pattern	Connection probability p_0			
	1/5	1/10	1/15	1/20
Continuous	0.040	0.048	0.051	0.050
Discrete	0.040	0.054	0.053	0.039

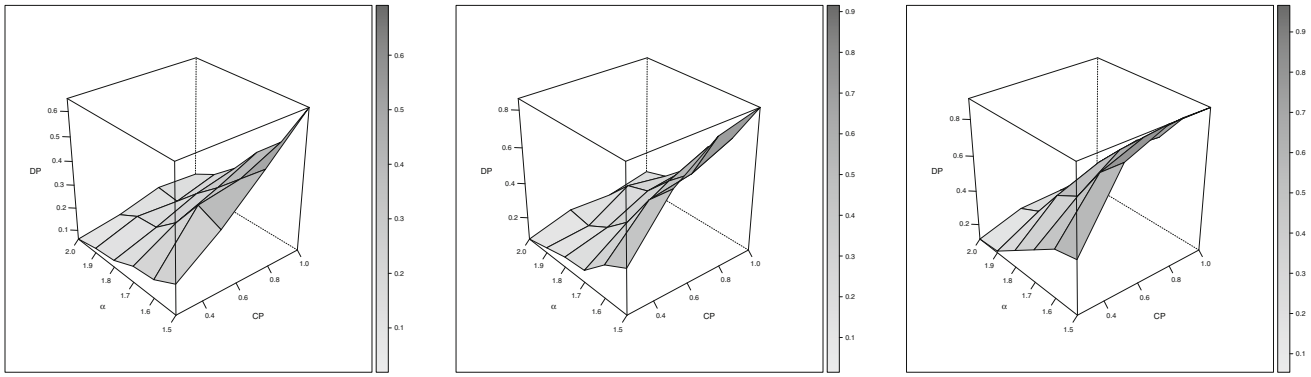
one should pay attention to these changes. The simulation for this purpose is set up as follows. p_c are selected from 1/4, 1/2, 3/4, and 1; cluster sizes are selected from 10, 15, and 20; the parameters α and β of discrete and continuous power-law attributes for cluster nodes are selected from 1.5 to 2.0 in steps by 0.1 and that of usual nodes is fixed as 2.5. We also try to set the α and β for usual nodes as 3. However, the behavior of each node is too similar to one another (most of them are 1), and the cluster may not be easily noticeable.

Figures 3 and 4 illustrate the detection powers for continuous and discrete PL attributes, respectively. The results show few differences between continuous and discrete cases, and cluster size does not appear to affect the testing power. In contrast, connection probability is the most important factor for detecting clusters. If the clusters are not highly connected, it is hard to see the similarity of attributes even we set a large value for cluster nodes. On the other hand, parameter value has impact when it has a lower value (smaller than 1.7). In Fig. 2, most data are 1 and few nodes have extreme values. It is hard to verify the significant difference. To see the influence of the parameter values, we list the estimation results for some selected cases in Table 2.

Table 2 shows the average values and standard deviations of estimations in 100 simulations for each combination case. The estimated averages are acceptable, but they are underestimated for large true values and are overestimated for small true values. In addition, the standard deviations (the bracket values) show interesting changes. When connection probability gets higher, the standard deviation gets lower. The estimations under the null hypothesis, values of $\hat{\alpha}_0$ and $\hat{\beta}_0$, are equally important but not included in Table 2. The average values of the parameters under null hypothesis range between 2.2 and 2.4 and the standard deviations range between 0.12 and 0.16. These results suggest that a good power can be achieved when the cluster parameter is not low.

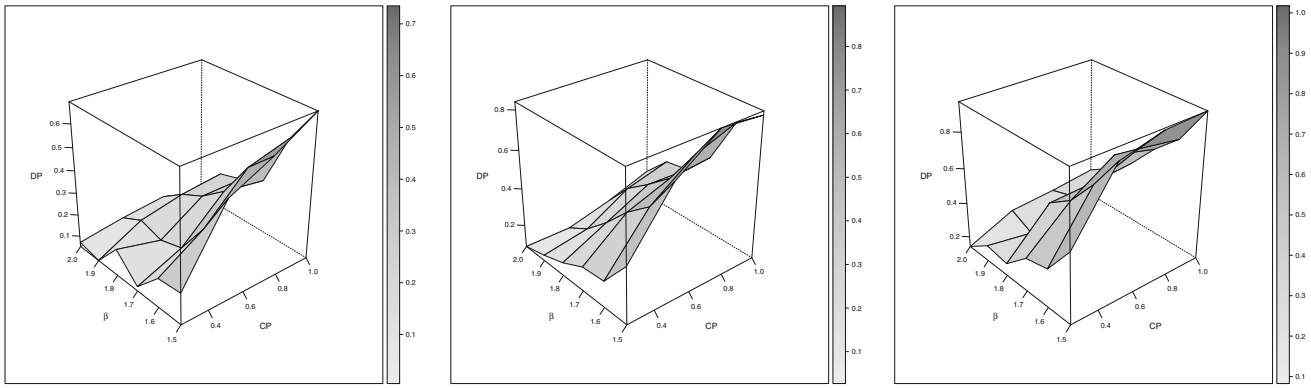
5 Empirical study

We apply our proposed method to the authorship data in this section. The authors collaboration network was produced from the BibTeX bibliography (Beebe 2002)



(a) Size = 10 **(b)** Size = 15 **(c)** Size = 20

Fig. 3 The detection powers of discrete PL-attributed clusters with different structure and parameter α



(a) Size = 10 **(b)** Size = 15 **(c)** Size = 20

Fig. 4 The detection powers of continuous PL-attributed clusters with different structure and parameter β

Table 2 Estimation results of parameter α for cluster nodes

Size	CP	Discrete true value of α			Continuous true value of β		
		2.0	1.7	1.5	2.0	1.7	1.5
10	0.25	1.88 (0.25)	1.77 (0.21)	1.65 (0.19)	1.81 (0.25)	1.75 (0.23)	1.68 (0.21)
	0.5	1.83 (0.24)	1.75 (0.22)	1.67 (0.2)	1.83 (0.23)	1.78 (0.21)	1.65 (0.2)
	0.75	1.81 (0.25)	1.75 (0.2)	1.62 (0.19)	1.86 (0.23)	1.76 (0.21)	1.63 (0.19)
	1	1.86 (0.21)	1.76 (0.19)	1.6 (0.16)	1.85 (0.21)	1.74 (0.17)	1.62 (0.16)
20	0.25	1.85 (0.22)	1.72 (0.19)	1.61 (0.15)	1.82 (0.2)	1.72 (0.2)	1.59 (0.17)
	0.5	1.84 (0.21)	1.71 (0.18)	1.56 (0.14)	1.83 (0.22)	1.73 (0.17)	1.58 (0.14)
	0.75	1.86 (0.2)	1.71 (0.14)	1.6 (0.12)	1.87 (0.18)	1.7 (0.14)	1.6 (0.13)
	1	1.87 (0.17)	1.75 (0.14)	1.62 (0.13)	1.89 (0.18)	1.72 (0.14)	1.62 (0.12)

The brackets are the standard deviations from 100 estimation results of each combination

obtained from the Computational Geometry Database and was well organized in Pajek datasets (http://vlado.fmf.uni-lj.si/pub/networks/data/collab/geom.htm). We only consider the vertices with at least one paper in this study. Thus, the network of interest consists of 6158 vertices and 11,898 edges. The maximum number of papers for an author is 697 in this data. There are strong connections between coauthors. Most researches only consider the network structure, but few of them mentioned that the number of papers itself may form another patterns. We apply our method to the data with PL distribution as a demonstration.

First, we inspect the data properties and check if the data follows a PL distribution. The log–log plot and Q–Q plot with estimated parameter α of PL distribution are shown in

Fig. 5. Clearly from Fig. 5, the coauthor data possess the features of the PL distribution such as the heavy tail and logarithm linear form, together with some biased values on both ends.

Following the testing procedures in Sect. 3, the results of structure and attribute clusters are listed in Table 3. We show the results of the most significant cluster of these two types of clusters. In addition, the Newman's modularity method (Newman and Girvan 2004) is also applied to detect the structure clusters. The modularity measure is defined as

$$Q = \frac{1}{2|E_G|}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2|E_G|}\right)s_i s_j = \frac{1}{4|E_G|}s^T B s,$$

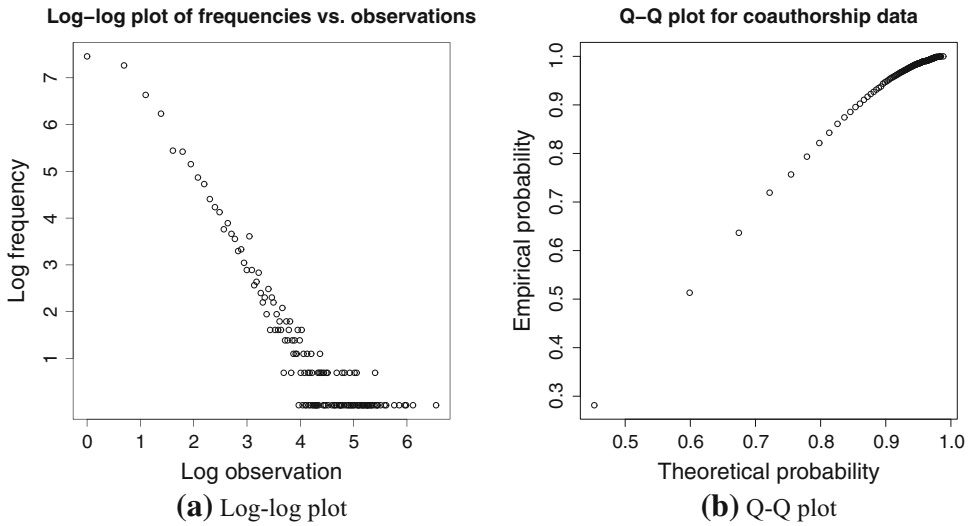
where the notations $|E_G|$, k_i, and k_j are equivalent to those described in Sect. 3.1, A is the adjacent matrix based on the given network, and s is a ± 1 vector in which 1 represents an element belongs to the target group and -1 represents an element does not belong to the target group if only two groups are considered.

Table 3 lists all the results of our proposed method and the Newman's modularity method. The bracket of the method column indicates the pattern types of the detected cluster. When considering the structure cluster (S), the scan method detect a larger cluster but with lower connection strength within the cluster S_Z. The modularity values Q of the Scan(S) and Newman(S) are interesting. It is common that Newman's method usually finds the best modularity when considering the existence of two clusters. However, the scan method is able to a higher modularity (although there is more than one cluster detected), because the scan method is more flexible to find the clusters in this case. Compare the attribute cluster with the structure clusters, the attribute cluster (A) detected by the proposed method is apparently larger than the structure clusters detected by the other two method. That means the expansion of the attribute seems quicker than the network connection.

6 Discussion

In this paper, we extend the scan statistic to consider PL distributions in both discrete and continuous cases for testing the attribute clusters in networks. We first review the properties and the estimations of PL distributions, and then we generate the likelihood ratio test statistic of PL distributions for the scanning method. We further construct simulations to verify the feasibility of our detecting approach. Finally, we use the method to analyze the authorship data and compare the results with the modularity method from (Newman and Girvan 2004).

In practice, the scanning method can be applied to other distributions such as binomial, Poisson, normal, and multinomial distributions which were constructed in spatial data analysis. The power-law distribution, which is rarely mentioned in many statistical applications, is however one



(a) Log-log plot **(b)** Q-Q plot

Fig. 5 Exploratory data analysis of the coauthor data

Table 3 Information of detected clusters

Method	Λ_S	S_0	S_z	Q	Λ_A	A_0	A_z	Size
Scan(S)	1487.67	0.80	1.75	0.32				556
Newman(S)	2061.38	0.88	3.35	0.23				333
Scan(A)					191.43	1.63	1.47	2904

of the most important distributions in social networks. We believe the potential of extending this distribution to other statistical applications. Besides, a truncated power-law distribution draws much attention in recent years (Burroughs and Tebbens 2001). Some studies believe that it is more realistic for real data, but the estimation is more difficult.

There are few problems of the scanning method. One of the most obvious problem is the computing burden. Because the scan statistics are decided by the scanning windows, the number of scanning windows is tremendous in a large network. Furthermore, we applied the Monte Carlo method to obtain a statistical significance. This also increases the computing burden. The other problem is the connection probability and clustered size can influence the detection power. We are not sure if this problem exists in a large but spare network like the authorship data in this study, since it is not feasible in terms of the computing loading for our proposed method to test all generated windows and verify clusters by Monte Carlo procedure in such a large network. To resolve the computing burden listed above, We are looking forward to parallel computing method which can evaluate the Monte Carlo simulations at the same time to facilitate the computing efficiency. In addition, we also try to create a new algorithm to reduce the number of scanning windows rather than searching the whole region.

We only generate circularly scanning windows in this study, but there are many different ways to construct these windows, such as elliptical shape windows (Kulldorff et al. 2006) and flexible scanning windows (Tango and Takahashi 2005) proposed in the spatial statistics. It is a vital future work that we will try to utilize these diverse methods to see if the detection accuracy can be improved.

Acknowledgments This work was supported by (a) Career Development Award of Academia Sinica (Taiwan) grant number 103-CDA-M04 and National Science Council (Taiwan) grant number 102-2628-M-001-002-MY3, for Phoa, (b) Thematic Research Program of Academia Sinica (Taiwan) grant number AS-103-TP-C03 for Phoa and Wang.

References

Balasubramanyan R, Cohen WW (2011) Block-lda: Jointly modeling entity-annotated text and entity-entity links. SDM, SIAM 11:450–461

Beebe NH (2002) Nelson hf beebe bibliographies page. http://www.math.utah.edu/∼beebe/bibliographies.html

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Burroughs SM, Tebbens SF (2001) Upper-truncated power laws in natural systems. Pure Appl Geophys 158(4):741–757

Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703

Erdős P, Rényi A (1959) On random graphs. Publ Math Debr 6:290–297

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174

Fortunato S, Barthélemy M (2007) Resolution limit in community detection. In: Proceedings of the National Academy of Sciences 104:36–41

Gillespie CS (2013) Fitting heavy tailed distributions: the poweRlaw package. R package version (20):2

Goldstein ML, Morris SA, Yen GG (2004) Problems with fitting to the power-law distribution. Eur Phys J B Condens Syst 41(2):255–258

Guimera R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70(2):025–101

Handcock MS, Raftery AE, Tantrum JM (2007) Model-based clustering for social networks. J Roy Stat Soc A Stat 170(2):301–354

Heard NA, Weston DJ, Platanioti K, Hand DJ (2010) Bayesian anomaly detection methods for social networks. Ann Appl Stat 4(2):645–662

Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. Science 311(5757):88–90

Kulldorff M (1997) A spatial scan statistic. Commun Stat Theor M 26(6):1481–1496

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. Stat Med 25(22):3929–3943

Liu Y, Niculescu-Mizil A, Gryc W (2009) Topic-link lda: joint models of topic and author community. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp 665–672

Louch H (2000) Personal network integration: transitivity and homophily in strong-tie relations. Soc Netw 22(1):45–64

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annu Rev Sociol pp 415–444

Naus JI (1966) Some probabilities, expectations and variances for the size of largest clusters and smallest intervals. J Am Stat Assoc 61(316):1191–1199

Newman ME (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69(6):066–133

Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026–113

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 4(1):11

Wang B, Phillips JM, Schreiber R, Wilkinson DM, Mishra N, Tarjan R (2008) Spatial scan statistics for graph clustering. In: SDM, pp 727–738

Yang J, McAuley J, Leskovec J (2013) Community detection in networks with node attributes. In: Data Mining (ICDM), 2013 IEEE 13th International Conference on, pp 1151–1156

Zhou Y, Cheng H, Yu JX (2009) Graph clustering based on structural/attribute similarities. In: Proceedings of the VLDB Endowment 2:718–729