Expert Systems with Applications 42 (2015) 3420–3432

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An overlapping semantic community detection algorithm base on the ARTs multiple sampling models

Yu Xin [a], Jing Yang [a,*], Zhi-Qiang Xie [b], Jian-Pei Zhang [a]

[a] College of Computer Science and Technology, Harbin Engineering University, Heilongjiang 150001, China
[b] College of Computer Science and Technology, Harbin University of Science and Technology, Heilongjiang 150001, China

ARTICLE INFO

Article history:
Available online 27 December 2014

Keywords:
Semantic Social Network
Community detection
Overlapping communities
Multiple sampling

ABSTRACT

Since the Semantic Social Network (SSN) is a new kind of complex networks, the traditional community detection algorithms require giving the number of the communities and could not detect the overlapping communities. To solve this problem, we propose improving multiple sampling models ARTs, consisting of ART, LART, ARTF and LARTF, sampling the textual information specific to node, link, node field, and link field correspondingly. The proposed ARTs models separate the semantic community detection into context sampling and communities detecting stage. After the context sampling, the quantized semantic coordinate is allocated to each sampling element, by which the cohesion for each sampling field can be established, avoiding the presetting of the number of communities. As the ARTs models are not easy to convergence, we explore the multiple sampling to accelerate the convergence, and the parameters of ARTs are analyzed by experimental analysis. In evaluation aspect, some traditional evaluation models are extended for semantic community measurement. Finally, efficiency of ARTs is verified by experiment.

1. Introduction

In accordance with the development of network communication, the electronic social network, such as Facebook and Twitter, has played an important part in people's daily social communication. Many social networking sites have launched the Community Recommended and Friend Circle Service to enrich people's web life. Thus, the community detection and recommendation algorithms have become the focus on social networks data mining. To date, community detection researching includes the following three aspects: hard community detection, overlapping community detection and semantic community detection.

The hard and overlapping community detection belongs to the topological community detection. The objective of these algorithms is to detect the communities with close internal relationships utilizing the properties of the relationships. The hard community detection is the pioneer work, and the ultimate goal of which is to divide the social networks into several separate networks (Newman, 2006; Newman & Girvan, 2004). The representative algorithms include GN (Girvan & Newman, 2002) and FN (Newman, 2004). In accordance with the development of hard community detection, researchers gradually focus on the case that a node belongs to several communities. Therefore, Palla, Derényi, Farkas, and Vicsek (2005) suggested the CPM algorithm to detect the overlapping structures. After that, overlapping community detection research became the major concern in social networks and many representative algorithms were proposed, such as EAGLE (Shen, Cheng, Cai, & Hu, 2009), LFM (Lancichinetti, Fortunato, & Kertész, 2009), COPRA (Gregory, 2010), UEOC (Jin et al., 2011), et al. The objective of semantic community detection is to cluster the nodes with similar semantic context (microblogging and social labels) into the same community. Since the semantic communities are detected by both context and relationship of the nodes, the result could represent the cohesion of communities more efficiently. For the semantic data mining must be based on the text analysis, many semantic community detection algorithms exploited the latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) model as the core model. According to the applied manner of LDA model, semantic community detection algorithms can be summarized as the following three categories:

(1) The LDA semantic analysis in terms of relationship. Such algorithms treated the topology of the social networks as semantic context, utilizing an improved LDA model to analyze the semantic similarity of nodes. Zhang, Qiu, Giles, Foley, and Yen (2007) proposed the SSN–LDA algorithm, regarding the ID and relationship as semantic context, the similarity of nodes as the training result. Henderson and Eliassi-Rad (2009) proposed the LDA-G algorithm to extend the SSN–LDA model with infinite relational models

* Corresponding author.
E-mail addresses: xinyu@hrbeu.edu.cn (Y. Xin), yangjing@hrbeu.edu.cn (J. Yang), xiezhiqiang@hrbust.edu.cn (Z.-Q. Xie), zhangjianpei@hrbeu.edu.cn (J.-P. Zhang).

http://dx.doi.org/10.1016/j.eswa.2014.11.029
0957-4174/

(IRM) (Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006). The LDA-G combined the LDA model with graph model, allowing it to predict the potential links among the detected communities. Then Henderson, Eliassi-Rad, Papadimitriou, and Faloutsos (2010) proposed the HCDF algorithm, extending the LDA-G with multiple attribute analysis and increasing its stability. The GWN-LDA (Zhang, Giles, Foley, & Yen, 2007) devoting to the directed networks and the HSN-PAM (Zhang et al., 2007) to the hierarchical networks were proposed based on the SSN-LDA. The advantage of such algorithms is the simply structure and the less requirement for input parameters, suitable for handling large-scale data. The disadvantages are that the semantic of such algorithms is not context and the detected community lack of the real semantic relevance.

(2) The LDA semantic analysis in terms of relationship-topic. Such algorithms treat the context of nodes as semantic context, analyzing the similarity of the nodes with semantic context. Most of such algorithms utilize the AT (Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004) model as the basic model. The ART (McCallum, Corrada-Emmanuel, & Wang, 2005) proposed by McCallum is the representative model, which added the recipient sampling into the AT model. The ART promoted the AT research into the field of SSN. After that, McCallum, Wang, and Corrada-Emmanuel (2007) designed the role analysis model (RART) based on the ART, extending the application fields of ART into the Social Computing. Zhou, Manavoglu, Li, Giles, and Zha (2006) applied the user distribution sampling to the AT model, suggesting the CUT model. Cha and Cho (2012) proposed the HLDA model which extract the relational tree model from online social networks on the basis of the relationship of reply context and design a hierarchical LDA to simulate the context relation tree. The advantages of such models are the extension of the context analysis into topological analysis for each node, and the detected community having a higher internal similarity. The disadvantages are that such models merely consider the relationship properties of the social networks, lacking of the consideration on the feature of local field. That would result in the disconnected community.

(3) The LDA semantic analysis in terms of community-topic. Such algorithms add the local field sampling into the relationship-topic model, developing the adjacency sampling to local area sampling. These algorithms avoid the case of disconnection in local field. The GT model (Wang, Mohanty, & McCallum, 2005) suggested by Wang, extending the ART model by replacing the recipient sampling with group recipient, is the representative model. Then, Pathak, DeLong, Banerjee, and Erickson (2008) discussed the necessity of recipient sampling and proposed the CART model, adding the community sampling into the ART model. Recently, community-topic model has become the focus on SSN research. Mei, Cai, Zhang, and Zhai (2008) combining the topic distribution in local field with the modularity, proposed the TMN model and established the topic-community correlation function to optimize the process of community detection. Sachan, Contractor, Faruquie, and Subramaniam (2011, 2012) and Yin, Cao, Gu, and Han (2012) proposed the TURCM and LCTA model, in terms of topic-community and community-topic distribution respectively. The both models above not only increased the difference of the topic distributions in different communities, but also made the result more reasonable. The advantage of such models is the high accuracy of the result. The disadvantages are not only the complex structure and the easy of getting over-fitting result, but the number of communities needs to be preset as the basic LDA model requires the prior parameters. The result tends to be different as the difference of presetting parameter.

Allowing for the advantage of LDA analysis of community-topic on semantic community detection, we adopted the sampling manner of community-topic. To avoid the number of community presetting problem, we separated the community-topic detection into LDA sampling and semantic community detection stage. In the process of LDA sampling we designed the multiple sampling ARTs (consisting of ART, ARTL, ARTF, LARTF) which have a higher weight in the central of sampling field than the marginal. For this manner replaced the community sampling with the field sampling, it has not to preset the number of communities. In the multiple sampling, we analyzed the convergence with various sampling frequency, verifying the optimal sampling frequency for ARTs is 2. In the semantic community detection, we designed the community clustering algorithm. The clustering element is the sampling field which represents the minimal community structure. There exist intersections among different sampling fields. Therefore, the overlapping communities could be obtained. For the clustering process have no requirement for the number of clusters, the semantic community detection could be achieved without presetting the number of communities.

2. ARTs model analysis

2.1. ARTs models

For the typical semantic community analysis algorithms, such as AT, ART and HLDA, sample the context of SSN in the form of point, field, radiation. The difference among them is shown in Fig. 1. Fig. 1(a) is the sampling process of AT model. In the AT the node G_2, G_5 are sampled separately without taking into account the relationships. Therefore, the sampling process of AT model is specific to node. Fig. 1(b) is the sampling process of ART model. In the ART the nodes adjacent to the sampling node are treated as the recipients. One of the recipients (G_1, G_3, G_5) of G_2 is sampled at random when sampling the node G_2. Separately, one of the recipients (G_5, G_9) of G_8 is sampled at random when sampling the node G_8. Essentially, the sampling process of ART is in the form of the field around the node sampled, and the radius of the field is 1. Fig. 1(c) shows the sampling process of HLDA. In the HLDA each nodes is sampled in a hierarchical manner. When sampling the node G_2, the 1-*dis* nodes (G_1, G_3, G_5) are sampled secondly, the 2-*dis* nodes (G_4, G_6, G_7, G_8) are sampled thirdly, and so on. Obviously, the sampling process of HLDA is in the form of radiation.

The ART and HLDA models are the application of AT to the SSN. As the radius of ART is 1, the sampling field is relatively small. The sampling result merely representing the direct relationship could not reflect the community's block feature. The sampling process of HLDA is in the form of radiation without considering the weight, which ignores the impact of distance on sampling. For that, we improve the ART and HLDA model, designing the 3 models Link_ART(LART), ART_Field(ARTF) and Link_ART_Field(LARTF) shown in Fig. 1(d)–(f). Where Fig. 1(d) is the sampling illustration of LART. In the LART, the $link_{2,3}$ (the link between G_2 and G_3) treated as the sampling center, the nodes (G_2, G_3) as the direct sampling nodes, the nodes (G_1, G_4, G_5) as the 1-*dis* sampling nodes, thus the sampling radius of LART is 2. Fig. 1(e) is the sampling illustration of ARTF, In the ARTF, the central node G_2 is treated as the direct sampling node, weighted sampling the 1-*dis* nodes (G_1, G_3, G_5), 2-*dis* nodes (G_4, G_6, G_7, G_8), and so on. For the sampling weight is decreasing with increasing the distance, the sampling field of ARTF forms a convergence region with a high weight in center and a low weight in the edge. Fig. 1(f) shows the LARTF model. In the LARTF, the endpoint (G_2 and G_3) of central link $link_{2,3}$ are treated as the direct sampling node, weighted sampling the 1-*dis* nodes (G_1, G_4, G_5), 2-*dis* nodes (G_6, G_7, G_8), and so on. The LARTF extends the sampling field of ARTF by link centralized sampling.

Eq. (1) is the force formula in the data field (Zhu, Ghahramani, & Lafferty, 2003), representing the force between two elements on topological distance. For the attenuation of context in propagation

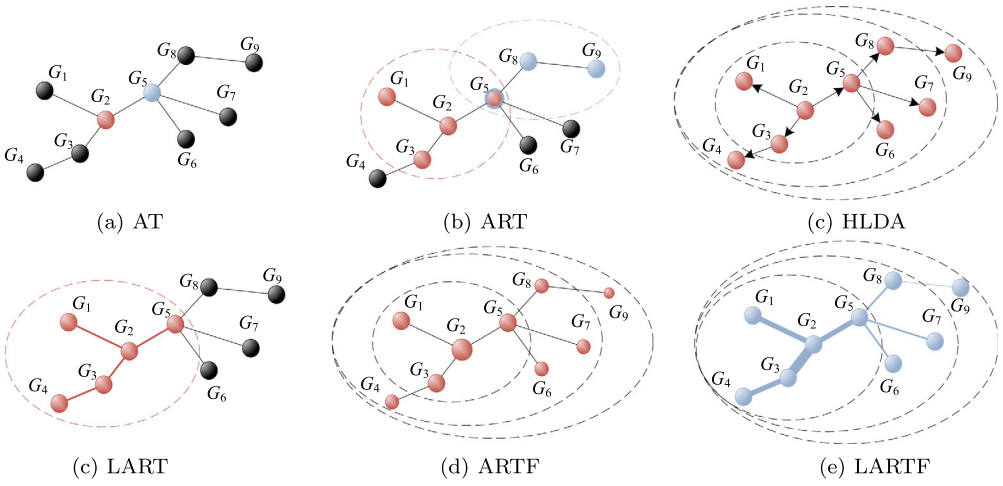

Fig. 1. The sampling process of context.

is increasing with increasing the distance, we choose Eq. (1) to simulate the influence of distance on the relevance of context. Therefore, Eq. (1) is utilized as field sampling weight in ARTF and LARTF.

$$weight_{r,c} = \exp\left(-\left(\frac{dis(G_r, c)}{\sigma}\right)^2\right) \tag{1}$$

where $dis(G_i, c)$ is the distance between G_i and c, σ is the distance controlling factor. According to Eq. (1), the weight of G_i and c is decreasing with increasing the distance of them, forming the field with a high weight in center and a low weight in the edge. As the center of the sampling field is c, the field is denoted as F_c. The probability density function of F_c can be express as follows:

$$fs(x|F_c) = \frac{weight_{x,c}}{\sum_x^{|G|} weight_{x,G_i}} \tag{2}$$

where $|G|$ is the number of nodes in the SSN.

2.2. Gibbs sampling analysis

In this section, the ARTs (ART, LART,ARTF, LARTF) models are described, and the relevant notations are as follows:

G: The global networks, G_i is the ith node.
$|G|$: The number of nodes in G.
L: The links of G, $link_{i,j}$ is the link between G_i and G_j.
$|L|$: The number of links in G.
F_c: The sampling field around c(*node* or *link*).
x: A node chosen from F_c.
N: The number of key words in SSN, N_i is the number of key works in G_i.
D: The number of messages in SSN.
$\boldsymbol{w}$: The vocabulary vector of key words, w_i is the ID of the ith word in vector $\boldsymbol{w}$.
$\boldsymbol{z}$: The topic ID vector corresponding to $\boldsymbol{w}$, z_i is the topic ID the w_i belongs to.
T: The number of topics in SSN.
θ: The topics distribution probability.
φ: The key words distribution probability.
α: The topics priori argument to the topic distribution.
β: The key words priori argument to a special topic.

Fig. 2 shows the plate models of LDA, AT, ART, LART, ARTF and LARTF.

The probabilistic generative process of ARTs can be described as follows:

$x \mid F \sim fs(x \mid F_c)$: Select an node from F_c as the sampling node.
$z \mid \theta \sim Multinomial(\theta)$: Extract a topic from the node in F_c. The topic is obedient to the multinomial distribution with the priori argument θ.
$\theta \mid \alpha \sim Dirichlet(\alpha)$: The argument θ is obedient to the Dirichlet distribution with the priori argument α.
$w \mid \varphi \sim Multinomial(\varphi)$: The key word w in a topic is obedient to the multinomial distribution with the priori argument φ.
$\varphi \mid \beta \sim Dirichlet(\beta)$: The φ is obedient to the Dirichlet distribution with the priori argument β.

$$p(\theta, \varphi, x, z, w|\alpha, \beta, F) = \prod_{i=1}^{|L|}\prod_{j=1}^{|G|} p(\theta_{ij}|\alpha)\prod_{t=1}^{T} p(\varphi_t|\beta) \prod_{d=1}^{D}\prod_{n=1}^{N}(p(x_{dn}|F)p(z_{dn}|\theta_{F,x})p(w_{dn}|\varphi_{dn})) \tag{3}$$

where x_{dn} represents the ID of the node the nth key word belongs to in the dth message in field F; z_{dn} represents the ID of the topic the nth key word belongs to in the dth message in field F; w_{dn} represents the ID of the nth key word in the dth message in field F; $\theta_{F,x}$ and φ_{dn} represent the frequency of the topic z_{dn} and w_{dn} for a specific F when the nth key word of dth message is generating. Integrating over the $\theta_{F,x}$ and φ_{dn} of Eq. (3), the marginal distribution function is obtained in Eq. (4).

$$p(\theta, \varphi, x, z, w|\alpha, \beta, F) = \prod_{i=1}^{|L|}\prod_{j=1}^{|G|} p(\theta_{ij}|\alpha)\prod_{t=1}^{T} p(\varphi_t|\beta) \prod_{d=1}^{D}\prod_{n=1}^{N}(p(x_{dn}|F)p(z_{dn}|\theta_{F,x})p(w_{dn}|\varphi_{dn})) \tag{4}$$

By the derivation of (McCallum et al., 2007), the conditional probability of x and $\boldsymbol{z}$ can be obtained as Eq. (5).

$$P(x_{dn}, z_{dn}|x_{-dn}, z_{-dn}, w, \alpha, \beta, F) \propto \frac{\alpha_{z_{dn}} + n_{F,x_{dn},z_{dn}} - 1}{\sum_{t=1}^{T}(\alpha_t + n_{F,x_{dn},t}) - 1} \times \frac{\beta_{w_{dn}} + m_{z_{dn},x_{dn}} - 1}{\sum_{v=1}^{V}(\beta_v + m_{z_{dn},v}) - 1} \tag{5}$$

The posterior estimates of θ and φ can be calculated by Eq. (6)

$$\hat{\theta}_{Fjz} = \frac{\alpha_z + n_{F,x,z}}{\sum_{t=1}^{T}(\alpha_t + n_{F,x,t})}, \quad \hat{\varphi}_{tw} = \frac{\beta_w + m_{t,w}}{\sum_{v=1}^{V}(\beta_v + m_{t,v})} \tag{6}$$

where V is the number of words in vocabulary, $n_{F,x,t}$ denotes the number of key word belong to the topic t in the node x specific to

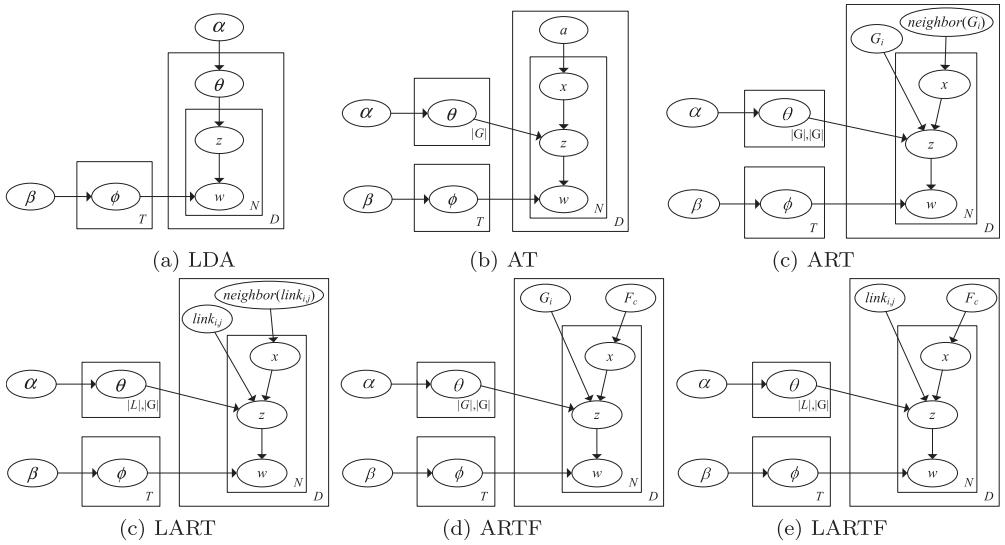

Fig. 2. The plate models.

F, m_{tv} denotes the number of key word v long to the topic t. θ_{Fjz} represents the frequency of the topic z occurrence in the context of G_j specific to F. The Gibbs sampling process of ARTs is described in Algorithm 1.

Algorithm 1. Parameter estimation in Gibbs sampling

1. initialize the node and topic assignments at random
2. repeat
3. **for** $d = 1$ to D
4. **for** $i = 1$ to N
5. draw x_{dn} and z_{dn} form $P(x_{di}, z_{di} | x_{-di}, z_{-di}, w, \alpha, \beta, F)$
6. update $n_{F,x_{dn},z_{dn}}$ and $m_{z_{dn},w_{dn}}$
7. **end for**
8. **end for**
9. until reaching equilibrium
10. calculate the posterior estimates of θ and φ according to Eq. (6)

3. Multiple sampling analysis

3.1. Semantic quantification

According to the analysis of ARTs, 3-dimensional topic distribution θ_{cft} can be obtained by Eq. (6), after the Gibbs sampling process. For the θ_{cft}, the c is the element-dimension, ART (ARTF) containing $|G|$ elements, LART (LARTF) containing $|L|$. The f is the field-dimension (link-dimension) containing $|G|$ elements. The t is the topic-dimension containing T-dimensional vector. The T-dimensional vector represents the topic membership of a node in the field F_c. Given an element in element-dimension, there would be $|G|$ nodes around it in field-dimension. Therefore, for the 3-dimensional topic distribution θ_{cft}, the field-dimension can be seen as the subordinate of element-dimension. Summing over the field-dimension could convert the 3-dimensional θ_{cft} into the 2-dimensional θ_{ct}. The $m_{ct} = \sum_{j=1}^{|G|} \theta_{cjt}$ represents the topic membership of F_c. Thus, the $m_c = (m_{c1}, m_{c2}, \ldots, m_{cT})$ can be treated as the coordinate of F_c in the sematic space.

3.2. Sampling convergence analysis

The General Gibbs sample the sampling field only once for each iteration. There would be large differences of semantic coordinate from each element (node or link), and the convergence is not easy to be obtained. For that, we explore the multiple sampling in each iteration, extending the ARTs into ARTs_Multiple. The relationships of ARTs and ARTs_Multiple are illustrated in Fig. 3.

In order to quantify the effect of multiple sampling, we design the following 3 global similarity measurement $sim1, sim3$ and sim_block.

$$sim1_c = \overline{U(m_{c'}, m_c)}, \quad dis(c', c) = 1 \tag{7}$$

$$sim3_c = \overline{U(m_{c'}, m_c)}, \quad dis(c', c) = 3 \tag{8}$$

$$sim\_block_c = \sum_{dis(c',c) \leqslant 3} weight_{c',c} U(m_{c'}, m_c) \tag{9}$$

where $U(m_{c'}, m_c)$ is the cosine similarity between c' and c, $\overline{U(m_{c'}, m_c)}$ is the average of $U(m_{c'}, m_c)$; $weight_{c',c}$ is the weight coefficient shown in Eq. (1); $sim1_c$ is the average similarity between c and 1-dis element. The $sim1$ reflects the similarity between the sampling center and the neighbors. $sim3_c$ is the average similarity between c and 3-dis element. For the effective radius of the block community (the Minimum community structures) (Girvan & Newman, 2002) is 3, the $sim3$ reflects the similarity the sampling center and the edge elements. sim_block_c is the weighted sum of the similarity of c and the element in 3-dis. For the impact of sampling center on the element around is decreasing with increasing the distance, the sim_block_c reflects the cohesion of the block community. It is more appropriate to measure the sampling result.

We utilize the LFM benchmark (Lancichinetti et al., 2009) to generate the artificial dataset G_500, employing the topology of which to analysis the influence of the sampling frequency fr on $sim1, sim3$ and sim_block. The parameter settings for the generating of G_500 is ($|G| = 500, ad = 3, dmax = 15, cmin = 10, cmax = 35, on = 40, om = 3, mi = 2.5$), where $|G|$ is the number of nodes, ad and $dmax$ are the average and largest degree, $cmin$ and $cmax$ are the numbers of nodes in the smallest and largest community, on is the overlapping nodes, om is the number of the communities an overlapping node belong to. mi is mixing coefficient. The community structure is getting fuzzy with increasing mi. As the simulation of semantic context, we select 30 topics at random, each topic containing 200 key words. We select 30 nodes with the largest degree from the G_500, assigning the 30 topics to the selected 30 nodes, correspondingly. Each of the node which is assigned topics selects 80 key words and assigns them to the neighbor iteratively, until the number of assigned key words in

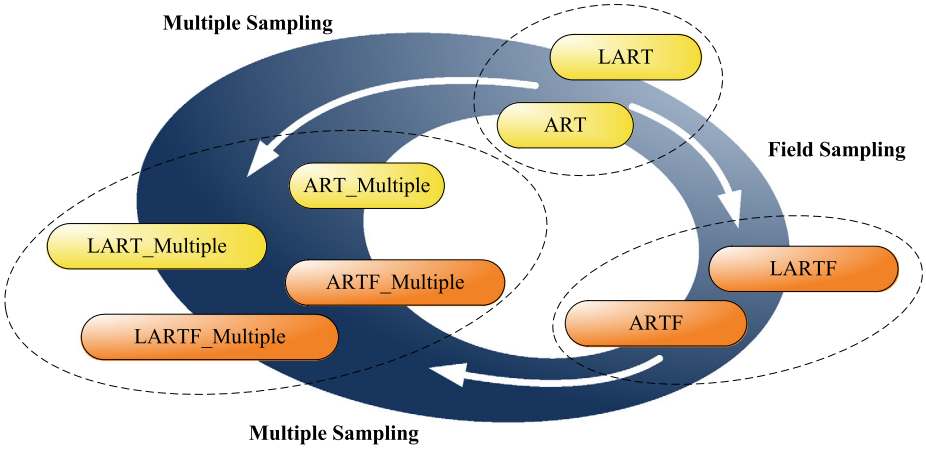

Fig. 3. The relationships of ARTs and ARTs_Multiple.

each node are more than 200. By that, the assigned key words can be seen as the semantic context.

We utilize the ART, ARTF, LART($\sigma = 2$), LARTF($\sigma = 2$) to sample the G_500 for $fr = \{1, 2, 3\}$, obtaining the average of $sim1, sim3, sim\_block$ shown in Figs. 4–6. The demonstration for the comparison of Figs. 4–6 is list as follows:

(1) In Figs. 4–6 the averages of $sim1, sim3$, sim_block are monotonically increasing and tend to convergence with increasing the Gibbs iterations. There exists a critical value of iterations, above which the averages of $sim1, sim3$, sim_block get convergence.
(2) When getting convergence, the averages of $sim1, sim3$, sim_block are increasing with increasing the fr.
(3) The iteration required to convergence is decreasing with increasing the fr.
(4) In Fig. 4 the average $sim1$ of ART and LART are larger than that of ARTF and LARTF, however, the average $sim3$ of ART and LART are smaller than that of ARTF and LARTF. It can be explained that the non-field models (ART and LART) have an impact scope in 1-*dis*, the field model (ARTF and LARTF) in 3-*dis*.
(5) By the comparison of Fig. 6, the averages sim_block of ARTF and LARTF are larger than that of ART and LART. It can be explained that the ARTF and LARTF have a more compact and effective block community.

3.3. Global similarity analysis

As the similarity between two elements is increasing with increasing the sampling frequency fr, thus the global similarity appears to be closed to each other when the fr getting larger. In that case, the semantic coordinate would lose the differentiation from each element, therefore the performance of semantic coordinate is undesirable. To analyze the changing of the global similarity with the increasing fr, we utilize the k-means clustering method to cluster the elements in SSN as follows:

(1) When $U(m_{c'}, m_c)$ the similarity of between two adjacent elements c' and c is larger than a certain *threshold*, the c' and c are combined into a cluster.
(2) The combined cluster is treated as a new element c, the average semantic coordinate of the cluster as the coordinate of c. Repeat the step 1) until any similarity between two adjacent elements is less than the *threshold*.
(3) The ratio of the size of the largest cluster and the size of the global is sim_ratio.

We apply the k-means clustering method in G_500 dataset generated above, plotting the average $sim\_ratio(y)$ against the $threshold(x)$ in Fig. 2. The illustration specific to ART for $fr = 2$ is precisely carried out as follows:

(1) When $threshold < 0.1$, average $sim\_ratio = 1$. In this case, the G is merged into one cluster, implying that any similarity between adjacent elements is larger than 0.1.
(2) When $0.1 < threshold < 0.8, 0.05 <$ average $sim\_ratio < 1$, implying that 95% similarity is in the range (0.1, 0.8).
(3) When $0.8 < threshold < 1$, average sim_ratio close to 0, implying that very little similarity is in the range (0.1, 0.8).

It is can be known from Fig. 7 that the similarity is major in the range (0, 0.5) for $fr = 1$, therefore the similarity is too low. For $fr = 3$, the similarity is major in the range (0.5, 1). In this case, the similarity is too high and the global similarity is close to each other. The desirable case is $fr = 2$, the similarity major in the range (0.2, 0.8). Figs. 4–6 show that the similarity is increasing with increasing the sampling frequency fr, therefore, for $fr > 3$, the similarity is

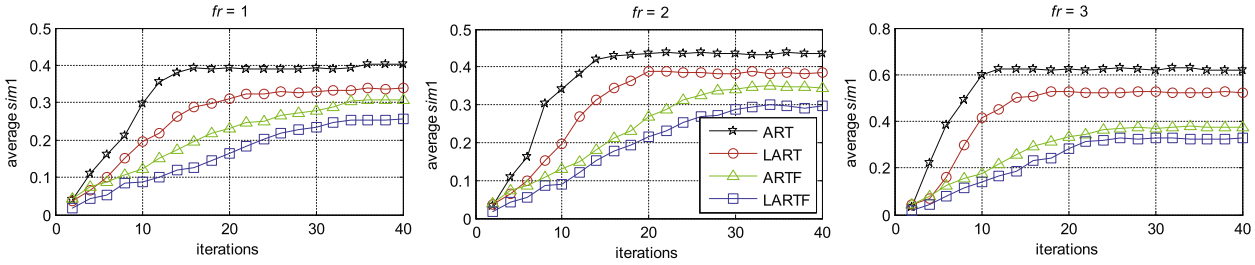

Fig. 4. The comparison of ARTs on average $sim1$ for $fr = \{1, 2, 3\}$.

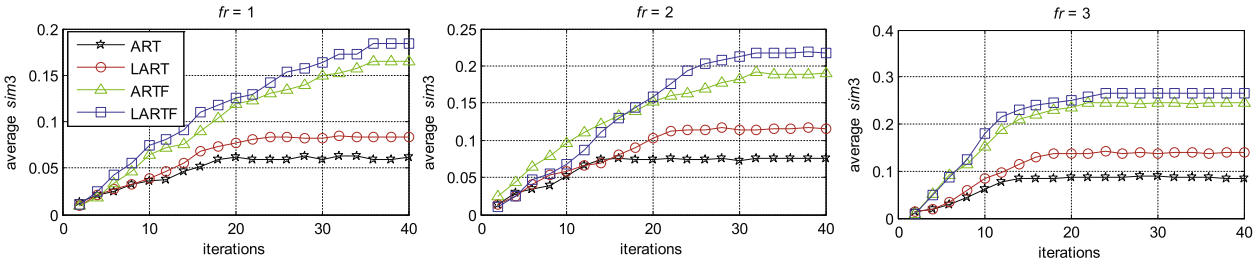

Fig. 5. The comparison of ARTs on average *sim*3 for $fr = \{1, 2, 3\}$.

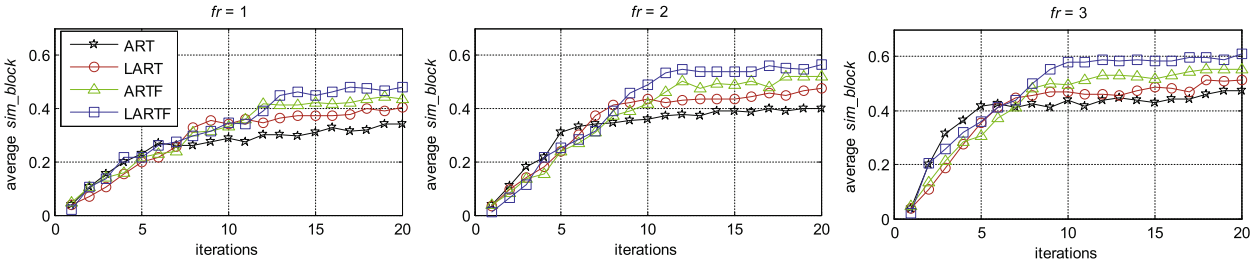

Fig. 6. The comparison of ARTs on average *sim_block* for $fr = \{1, 2, 3\}$.

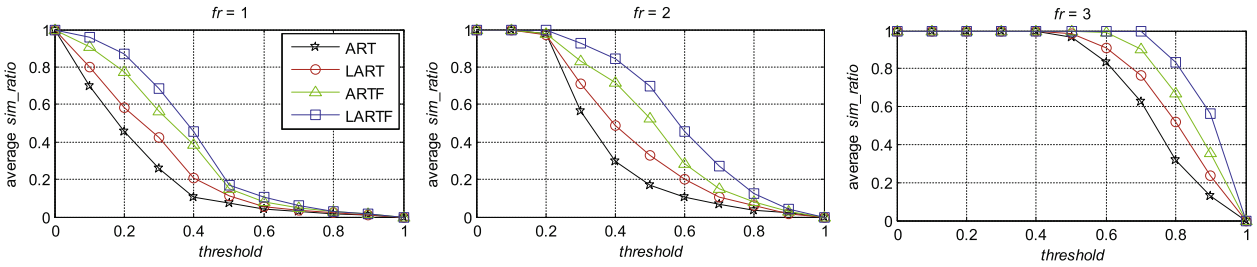

Fig. 7. The average *sim_ratio* against *threshold* for G_500.

larger than 0.5. When ARTs applied to the G_500 dataset got converge, the iterations, $sim1, sim3$, $sim\_block, threshold$ range $(0.1 < sim\_ratio < 0.9)$ for fr=$\{1, 2, \ldots, 6\}$ are listed in Table 1. The comprehensive analysis of Figs. 4–7 and Table 1 shows the optimal fr is 2 specific to G_500.

3.4. The analysis of fr in various scale datasets

To analyze the effective value of fr in various scale datasets, we generate 10 groups of datasets with the size in $\{1000, 2000, \ldots, 10,000\}$, utilizing the generative method of G_500. The iterations, $sim1, sim3$, average $sim\_block, threshold$ range $(10\% < sim\_ratio < 90\%)$ calculated by the ART, LART, ARTF ($\sigma = 2$), LARTF ($\sigma = 2$) are shown in Figs. 8–10. The analysis is as follows:

(1) When the scale gets larger, the iterations, *sim*1, *sim*3, average $sim\_block, threshold$ range $(10\% < sim\_ratio < 90\%)$ are tend to stable. It verified that the correlation between data scale and convergence can be ignored.
(2) When the fr gets larger, the iterations get smaller. For $fr = 3$, the iterations, average *sim_block* and *threshold* range close to the extremum.
(3) The relation of ARTs in the aspects of iterations and average *sim_block* are LARTF > ARTF > LART > ART, implying the LARTF requires the most iteration to get converge and has the best performance at average *sim_block*.
(4) By the comparison of *threshold* in Fig. 10, for $fr = 1$ the maximum of the *threshold* is small, implying the global similarity is too low, while for $fr \geqslant 3$, the global similarity is too high. Therefore, the *threshold* the has an optical solution with $fr = 2$.

From the analysis of Fig. 8–10 above, the optical sampling frequency of ARTs is $fr = 2$ in various scale datasets.

3.5. The analysis of distance controlling factor σ

The σ is the input parameter of ARTF and LARTF. According to Eq. (1) and (2), the σ affects the size of sampling field and the ratio of sampling weight within the sampling field. To demonstrate the influence of σ on the size and sampling weight, we calculate the *dis_ratio* of *n-dis* element with various σ by Eq. (10). The *dis_ratio* for *dis*=$\{0, 1, 2, 3, 4, 5\}$ when σ=$\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5.5\}$ is shown in Fig. 11. It can be known that the *dis_ratio* is close to 0 for $dis > 1$ when $\sigma < 1.5$, implying the scope of sampling field is within 1-*dis*. In this case the size of sampling field is too small. According to Eq. (2), when $\sigma > 2.5$ the valid *n-dis* is larger than 3, moreover the *dis_ratios* of *n-dis* are close to each other. In this case, the size of sampling field is too large, and the differentiation of *dis_ratios* are not obvious. By the demonstration above, the valid value of σ is within (1.5, 2.5) and valid $dis \leqslant 3$.

Table 1
The iterations, *sim*1, *sim*3, *sim_block*, *threshold* range as G_500 getting converge.

ARTs	Measurement	$fr = 1$	$fr = 2$	$fr = 3$	$fr = 4$	$fr = 5$	$fr = 6$
ART	Iterations	17	15	12	12	11	11
	*sim*1	0.34	0.44	0.62	0.63	0.64	0.64
	*sim*3	0.06	0.07	0.08	0.08	0.08	0.08
	sim_block	2.34	3.07	4.5	4.56	4.59	4.61
	threshold	(0.05, 0.4)	(0.27, 0.59)	(0.5, 0.91)	(0.56, 0.93)	(0.62, 0.94)	(0.69, 0.95)
LART	Iterations	22	20	18	17	17	17
	*sim*1	0.3	0.38	0.52	0.52	0.52	0.53
	*sim*3	0.08	0.11	0.13	0.13	0.14	0.14
	sim_block	2.85	3.3	4.62	4.68	4.71	4.72
	threshold	(0.07, 0.51)	(0.28, 0.64)	(0.59, 0.96)	(0.63, 0.97)	(0.71, 0.98)	(0.75, 0.99)
ARTF	Iterations	34	30	22	22	21	21
	*sim*1	0.3	0.35	0.37	0.38	0.38	0.38
	*sim*3	0.16	0.19	0.24	0.24	0.24	0.25
	sim_block	3.35	4.01	5.05	5.21	5.29	5.35
	threshold	(0.1, 0.57)	(0.31, 0.73)	(0.69, 0.97)	(0.76, 0.98)	(0.81, 0.99)	(0.85, 0.99)
LARTF	Iterations	34	30	24	24	24	23
	*sim*1	0.25	0.29	0.32	0.32	0.33	0.33
	*sim*3	0.18	0.21	0.26	0.26	0.27	0.27
	sim_block	3.63	4.49	5.84	5.92	6.11	6.19
	threshold	(0.18, 0.59)	(0.35, 0.81)	(0.76, 0.98)	(0.82, 0.99)	(0.87, 0.99)	(0.90, 0.99)

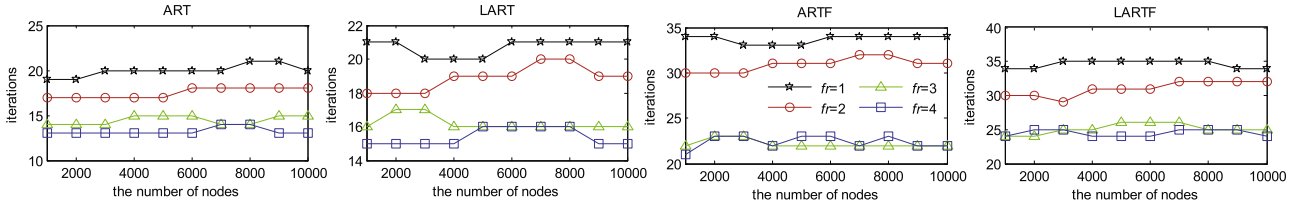

Fig. 8. The comparison of iterations for various scale datasets.

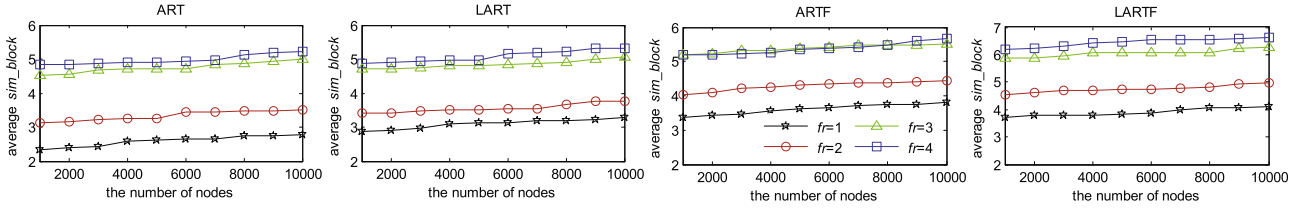

Fig. 9. The comparison of average *sim_block* for various scale datasets.

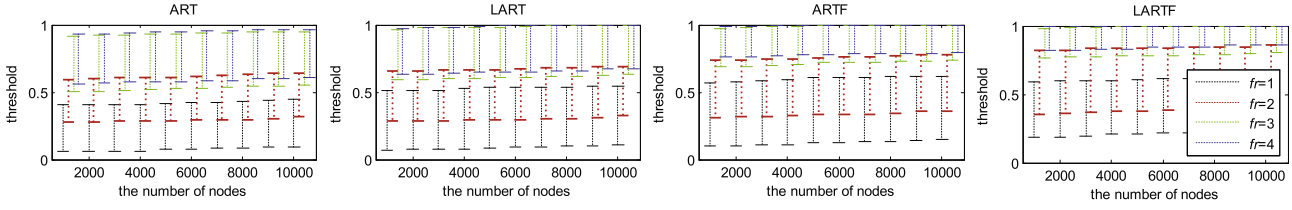

Fig. 10. The comparison of *threshold* ($10\% < sim\_ratio < 90\%$) for various scale datasets.

$$dis\_ratio(dis, \sigma) = \frac{\exp(-(dis/\sigma)^2)}{\sum_{i=1}^{\infty} \exp(-(i/\sigma)^2)} \tag{10}$$

We generate various scale datasets with the size of $\{1000, 2000, \ldots, 10,000\}$, for each size generating 20 times, then separating the 200 dataset into 10 groups by the scale. Fig. 12 is the *threshold* ($10\% < sim\_ratio < 90\%$) distribution box plot of ARTF and LARTF for the 10 groups scale datasets. In Fig. 12 each box represents the distribution of the *threshold* of 20 datasets with the same scale. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme points. In Fig. 12 when $\sigma < 1.5$ the maximum *threshold* of ARTF and LARTF is less than 0.5, implying the 80% global similarities are less than 0.5, and the sampling is insufficient. When $\sigma > 2.5$ the minimum *threshold* of ARTF and LARTF is larger than 0.6, implying the 80% global similarities are

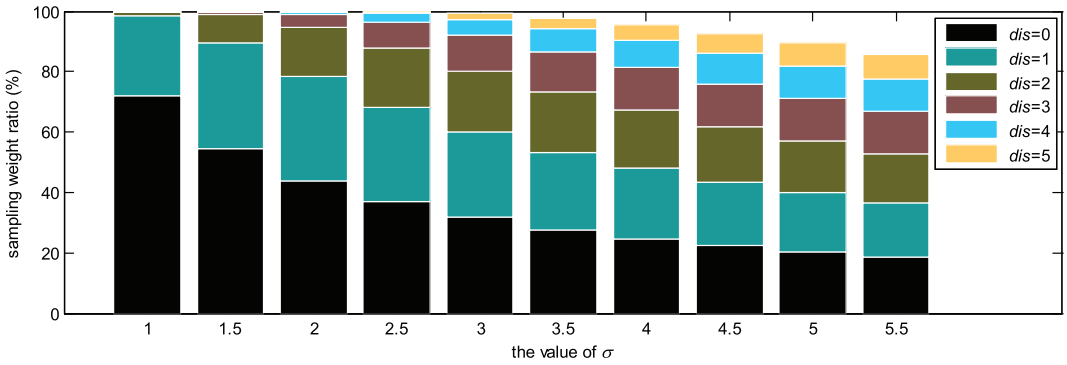

Fig. 11. The histogram of *dis_ratio* against σ for $dis = \{0, 1, 2, 3, 4, 5\}$.

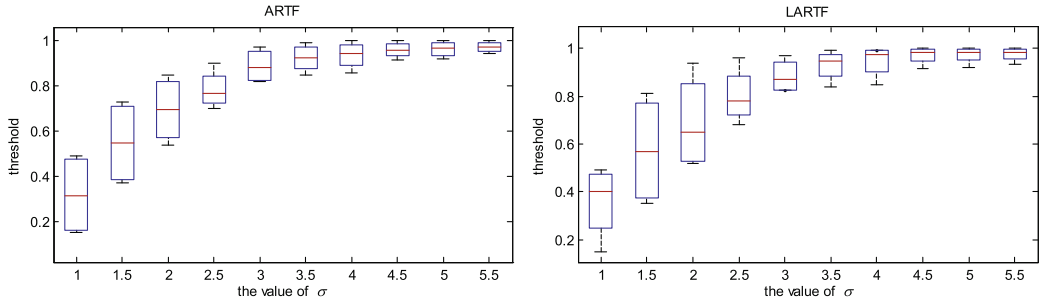

Fig. 12. The distribution box plot of the *threshold* ($10\% < sim\_ratio < 90\%$) for ARTF and LARTF.

larger than 0.6, and the sampling is overfitting. Therefore, the valid value of σ is within (1.5, 2.5), and the valid sampling field radius (*dis*) of ARTF and LARTF is 3.

4. Semantic community detection method and estimation model

4.1. Field clustering method

From the analysis of distance controlling factor σ, the ART and LART can be also seen as the field sampling model with the radius (*dis*) = 1, the radius of ARTF and LARTF is 3. From the semantic quantification, the semantic coordinate m_c of element c represents the multidimensional topic membership of the field F_c. Therefore the cohesion of field F_c can be obtained by m_c. Thus we employ the PCA method to weight the semantic coordinate m_c, obtaining the cohesion W_c of field F_c expressed in Eq. (11).

$$W_c = m_c \cdot \Lambda, \Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_T] \tag{11}$$

where, λ_i is the ith eigenvalue of the correlation matrix of the T-dimensional semantic coordinate.

The field clustering method treats the field as a unit, the clusters as detected community. If two element c and c' have a *dis* equal to 1, the $field_c$ and $field_{c'}$ are more similar to each other. Therefore, the case that the *dis* of the two core elements is 2 is defined as core-related. The clustering process is to cluster the two core-related fields into a new field, the merged two core elements treating as the new core element of the new field. The process can be precisely described as follows:

(1) Sort the semantic cohesion W in descending order, forming the *W-queue*;
(2) Select the top k fields from *W-queue*, guaranteeing the selected k fields cover the entire network justly;
(3) After combining the core-related fields in the selected k fields, the clusters are the detected communities, moreover the intersecting nodes of the clusters are overlapping nodes.

Fig. 13(a) is a representative illustration of ART in field clustering, assuming the G_3, G_4, G_5, G_7 have the larger cohesion, F_3 and F_4, F_4 and F_5 are core-related, G_3–G_5 are the core elements of the cluster formed by merging F_3–F_5. Fig. 13(b) is an illustration of LART, assuming the L_1–L_4 have the larger cohesion, F_2 and F_3 are core-related, L_2–L_3 are the core elements of the cluster formed by merging F_2–F_3. Fig. 13(c) is an illustration of ARTF, assuming the G_1, G_2, G_6, G_7 have the larger cohesion, F_1 and F_2, F_6 and F_7 are core-related, G_1–G_2 are the core elements of the cluster formed by merging F_1–F_2, G_6–G_7 are the core elements of the cluster combined by F_6–F_7. Fig. 13(d) is an illustration of LARTF, assuming the L_2–L_3 have the larger cohesion, F_2 and F_3 are core-related, L_2–L_3 are the core elements of the cluster formed by merging F_2–F_3.

4.2. Evaluation models

According to the quantification of semantic coordinate in ARTs, the semantic coordinate m_c of element c represents the coordinate of node (ART, ARTF) or link (LART, LARTF). For the traditional evaluation model designed specific to node, we convert the link coordinate (LART, LARTF) into node coordinate via Eq. (12). The similarity $U(m_i, m_j)$ between the adjacent node (G_i, G_j) represents the weight of $link_{i,j}$, thus adjacent matrix **S** of G can be obtained by $U(m_i, m_j)$.

$$m_i = \sum_{dis(G_i,G_j)=1} \frac{m_{link_{i,j}}}{degree(G_i)} \tag{12}$$

The traditional evaluation models are listed as follows, where $C_i^{out} = \sum_{p\in C_i, q\notin C_i} A_{pq}$, $C_i^{in} = \sum_{p,q\in C_i} A_{pq}$, $|L| = \sum_{p,q\in G} A_{pq}$, $C_i^{in}(j) = \sum_{p\in C_i} A_{jp}$, $C_i^{out}(j) = \sum_{p\notin C_i} A_{jp}$. In semantic aspects, we take advantage of the formation of traditional evaluation models, improving the variable

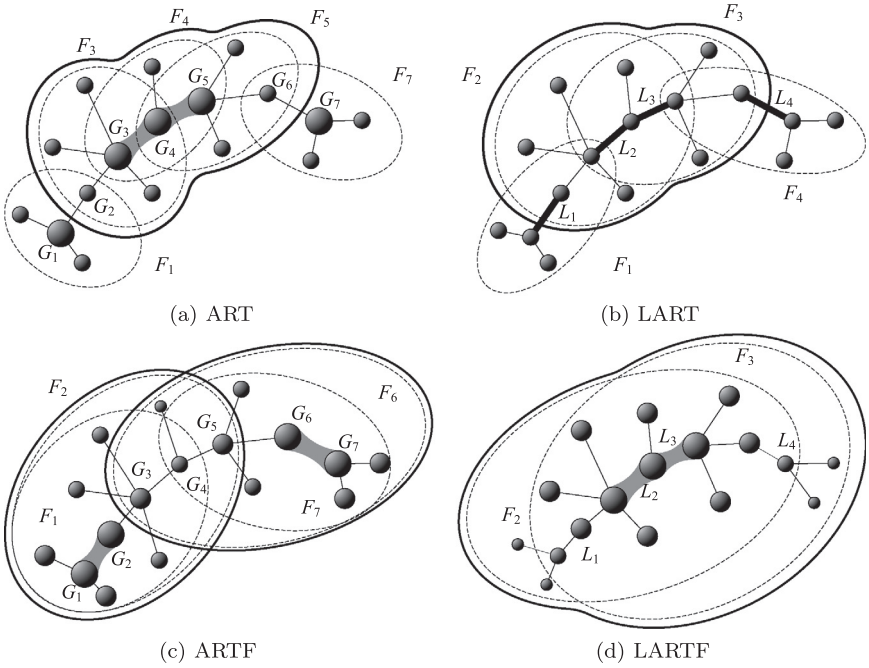

Fig. 13. The sample process of context.

as $C_i^{out} = \sum_{p\in C_i, q\notin C_i} S_{pq}$, $C_i^{in} = \sum_{p,q\in C_i} S_{pq}$, $|L| = \sum_{p,q\in G} S_{pq}$, $C_i^{in}(j) = \sum_{p\in C_i} S_{jp}$, $C_i^{out}(j) = \sum_{p\notin C_i} S_{jp}$, allowing the traditional evaluation to evaluate the semantic communities. The improved models are noted as *s-model*.

$$EQ = \frac{1}{2|L|}\sum_{i=1}^{|C|}\sum_{v,w\in C_i}\frac{1}{O_v O_w}\left(A_{vw} - \frac{degree(v)degree(w)}{2|L|}\right).$$

$$s-EQ = \frac{1}{2|L|}\sum_{i=1}^{|C|}\sum_{v,w\in C_i}\frac{sim(m_i,m_j)}{O_v O_w}\left(A_{vw} - \frac{degree(v)degree(w)}{2|L|}\right).$$

$$AC = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{C_i^{out}}{\min\left(\left(C_i^{out}+\frac{1}{2}C_i^{in}\right),\left(2|L|-C_i^{out}-\frac{1}{2}C_i^{in}\right)\right)}.$$

$$MMC = \sum_{i=1}^{|C|}\frac{2C_i^{out}}{C_i^{in}}.$$

$$Silhouette = \frac{1}{|C|}\sum_{j=1}^{|C|}\left(\frac{1}{|C_j|}\sum_{i\in C_j}\frac{a_i-b_i}{\max(b_i,a_i)}\right),$$

$$a_i = \frac{1}{|C_k|}\sum_{i,j\in C_k}A_{ij},\quad b_i = \max\left(\frac{1}{|C_r|}\sum_{i\notin C_r, j\in C_r}A_{ij}\right).$$

$$Ductance = \sum_{i=1}^{|C|}\frac{C_i^{out}}{C_i^{in}+C_i^{out}}.$$

$$Expansion = \sum_{i=1}^{|C|}\frac{C_i^{out}}{|C_i|}.$$

$$NCut = \sum_{i=1}^{|C|}\frac{C_i^{out}}{C_i^{in}+C_i^{out}} + \frac{C_i^{out}}{2\left(|L|-\frac{1}{2}C_i^{in}\right)+C_i^{out}}.$$

$$AF = \frac{1}{|C|}\sum_{i=1}^{|C|}\sum_{j\in C_i}\frac{C_i^{in}(j)}{\left(C_i^{in}(j)+C_i^{out}(j)\right)^r}.$$

5. Experiment

5.1. The comparison of ARTs on evaluation models

We conduct the experiments on ARTs with non-semantic (traditional) and semantic evaluation models. The experimental process is the following. (1) Generate 100 datasets with 5000 nodes via the method of G_500. (2) Carry out the semantic community detection method and record the detected communities for ART, LART, ARTF ($\sigma = 2$), LARTF ($\sigma = 2$). (3) Employ the non-semantic and semantic evaluation models to evaluate the detected communities. Fig. 14 shows the distribution histogram of ARTs on the 100 datasets. Table 2 is the 25th and 75th percentiles of the ARTs on the 100 datasets, in which *EQ*, *Silhouette*, *Expansion*, *AF* are larger (*AC*, *MMC*, *Ductance*, *Ncut*are smaller) the detected communities are more reasonable. In the same way, *s-EQ*, *s-Silhouette*, *s-Expansion*, *s-AF* are larger (*s-AC*, *s-MMC*, *s-Ductance*, *s-Ncut*) the detected communities are more reasonable.

Taking the case of *EQ* and *s-EQ* in Fig. 14, the ARTs is distributed evenly in terms of *EQ*, implying the ARTs have a similar non-semantic structure. In terms of *s-EQ* the frequency of ARTF and LARTF is less than ART and LART within (0.25, 0.35), and more than ART and LART within (0.35, 0.4), implying the distribution of ARTF and LARTF is larger than that of ART and LART. Therefore, the ARTF and LARTF have a better performance than ART and LART on *s-EQ*. The comprehensive analysis of Fig. 14 and Table 2 shows the ART and ARTF perform better than ARTF and LARTF on non-semantic models (*EQ*, *AC*, *Expansion*, *Ncut* and *AF*), however undesirable on semantic models (*s-EQ*, *s-AC*, *s-MMC*, *s-Silhouette*, *s-Ductance*, *s-Expansion*, *s-Ncut* and *s-AF*). This experimental study has verified the ARTF and LARTF have a better performance than ART and ARTF on semantic community evaluation.

5.2. The comparison on semantic evaluation models

In this section, the performance of semantic evaluation models is compared by increasing the inner similarity of community under the premise the community structure is constant. 20 sets of

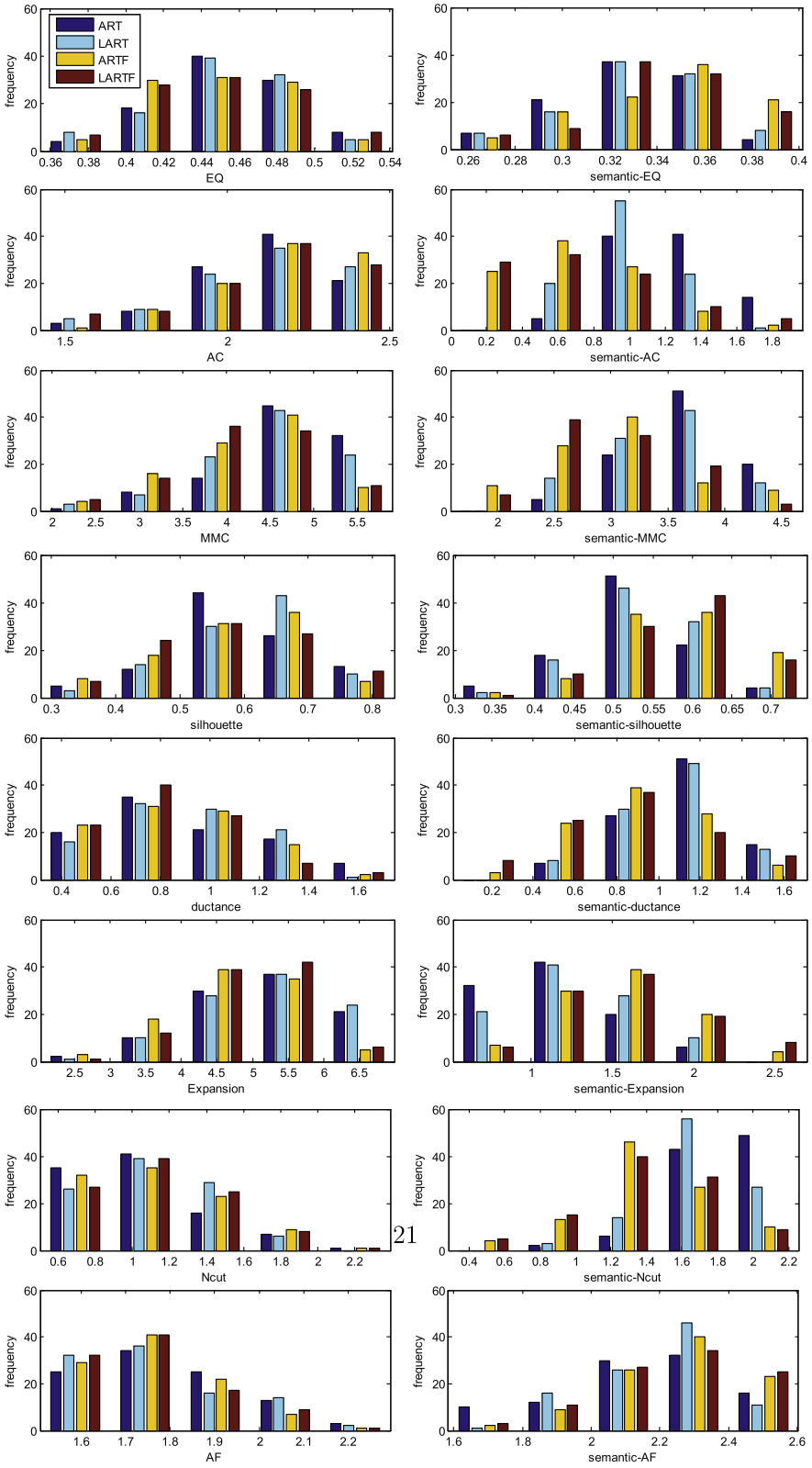

Fig. 14. The histogram or non-semantic and semantic evaluation models.

artificial datasets are generated by LFR (Lancichinetti et al., 2009), with $|G| = 2000$, $ad = 5, dmax = 30, cmin = 10$, $cmax = 100$, $on = 300, om = 4, mi = 2.5$. The semantic coordinate is allocated as follows:

(1) Choose one node in the center of the community as the community label node.
(2) Each community label node propagate its label to its sampling field with the weight in Eq. (2).

Table 2
The 25th and 75th percentiles of the ARTs.

Models	ART	LART	ARTF	LARTF
EQ	(**0.4316**, **0.4763**)	(**0.4327**, **0.4739**)	(0.4159, 0.476)	(0.4149, 0.4773)
s-EQ	(0.3078, 0.3523)	(0.3143, 0.3567)	(**0.3246**, **0.3655**)	(**0.3263**, **0.3612**)
AC	(**1.9791**, **2.2643**)	(**1.9374**, **2.3009**)	(2.0272, 2.3222)	(1.9946, 2.3111)
s-AC	(0.9802, 1.4088)	(0.8393, 1.2011)	(0.3935, 0.9103)	(**0.3256**, **0.9989**)
MMC	(4.3312, 5.2827)	(4.0991, 5.0891)	(**3.6157**, **4.736**)	(**3.5493**, **4.8254**)
s-MMC	(3.3655, 3.961)	(3.0999, 3.7937)	(**2.5991**, **3.3894**)	(**2.4371**, **3.3577**)
Silhouette	(**0.5206**, **0.6611**)	(**0.5401**, **0.6943**)	(0.4944, 0.6559)	(0.4798, 0.6484)
s-Silhouette	(0.4826, 0.5728)	(0.4958, 0.591)	(**0.5104**, **0.6274**)	(**0.5367**, **0.6353**)
Ductance	(0.6617, 1.1225)	(0.6919, 1.1393)	(**0.6331**, **1.0727**)	(**0.6111**, **1.0113**)
s-Ductance	(0.9112, 1.2892)	(0.8878, 1.2474)	(**0.6525**, **1.1373**)	(**0.6453**, **1.0974**)
Expansion	(**4.5877**, **5.9056**)	(**4.5257**, **5.9226**)	(4.1301, 5.4029)	(4.3533, 5.6348)
s-Expansion	(0.849, 1.3999)	(0.9826, 1.5487)	(**1.2756**, **1.7819**)	(**1.2600**, **1.8522**)
Ncut	(0.8696, 1.3252)	(0.8913, 1.2691)	(**0.8585**, **1.4999**)	(**0.8748**, **1.3772**)
s-Ncut	(1.6476, 2.0376)	(1.5481, 1.9766)	(**1.1592**, **1.6848**)	(**1.1637**, **1.6747**)
AF	(**1.6500**, **1.8849**)	(**1.6545**, **1.9025**)	(1.6341, 1.8684)	(1.6452, 1.8558)
s-AF	(1.9865, 2.2200)	(2.0259, 2.3201)	(**2.1479**, **2.3934**)	(**2.1199**, **2.3498**)

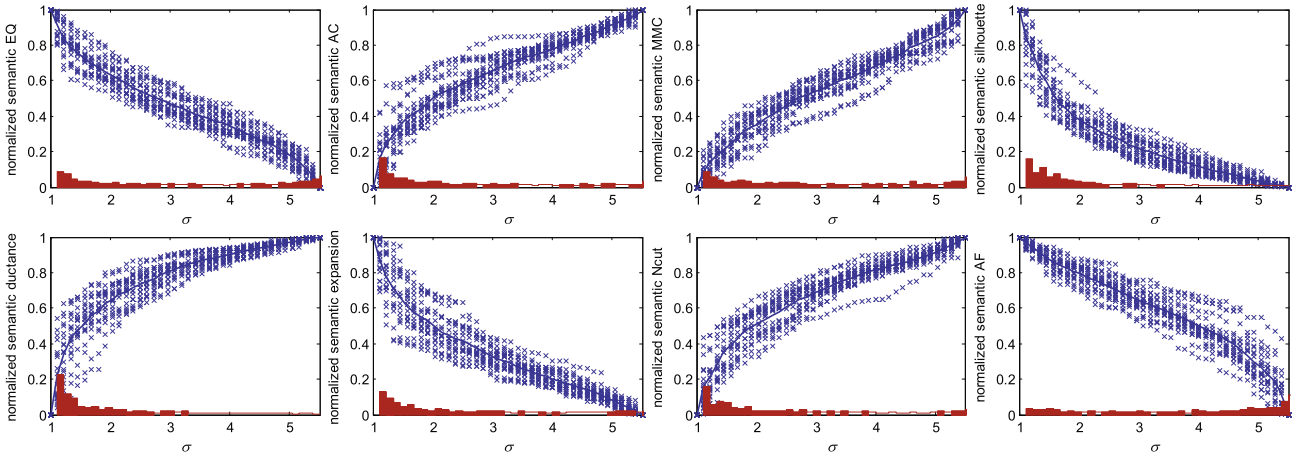

Fig. 15. The histogram or non-semantic and semantic evaluation models.

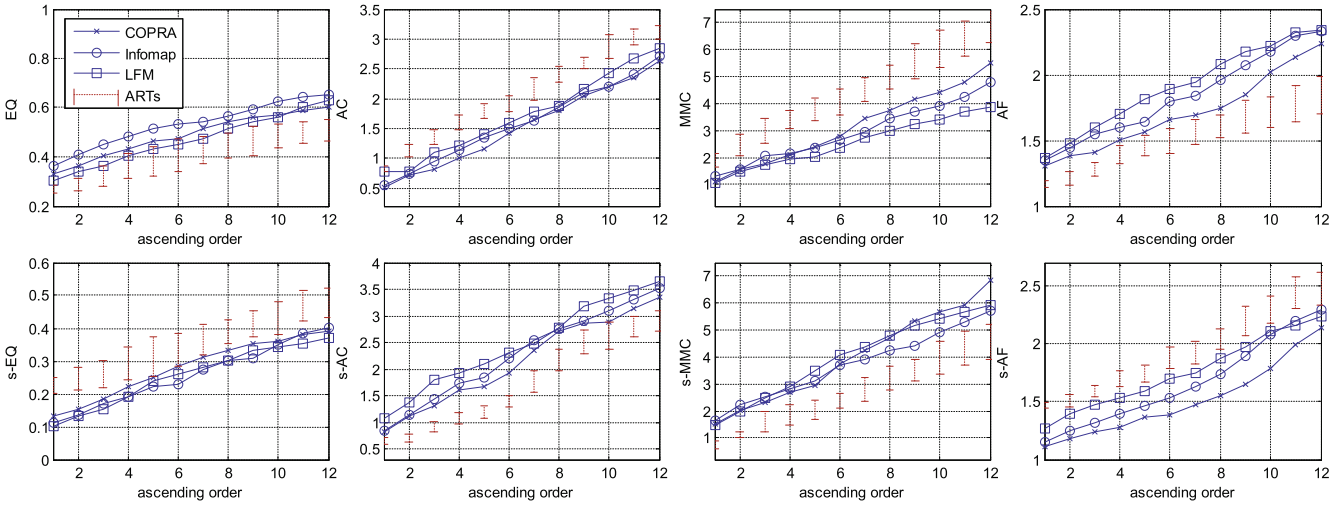

Fig. 16. The comparison of ARTs and non-semantic algorithms.

(3) The received weighted labels are treated as the semantic coordinate for each node.

The original propagation point is center of community, thus the nodes in the same community have the similar semantic coordinate. That approximates the feature of semantic structure. According to the analysis of σ, when σ gets larger the sampling weight tends to evenly. Therefore, when the σ is small, the inner similarity of community is larger than outer. In this case, the semantic community structure is more valid. The experiment is carried out for σ within (1, 5.5), then normalize the result of semantic evaluation models to compare them obviously. The

Table 3
The topics extracted from Enron dataset.

Topic	California power	Gas transportation	Trading	Deals
Word	Power	Gas	Price	Meeting
	Transmission	Energy	Market	Contract
	Energy	Enron	Dollar	Report
	Calpx	Transco	Nymex	Enron
	California	Chris	Trade	Deal

Table 4
The *s-EQ*, *s-AC*, *s-MMC*, *s-AF* of semantic community detection algorithms.

Methods		*CS* = 6	*CS* = 8	*CS* = 10	*CS* = 12	*CS* = 14
TURCM	*s-EQ*	0.198	0.271	0.339	**0.331**	0.283
	s-AC	2.434	2.113	1.812	**1.503**	2.016
	s-MMC	4.231	3.311	2.381	1.864	2.336
	s-AF	1.867	1.922	2.174	2.232	2.031
CART	*s-EQ*	0.152	0.249	0.302	0.294	0.255
	s-AC	2.727	2.468	2.163	1.783	2.364
	s-MMC	4.084	3.293	2.287	**1.665**	2.134
	s-AF	1.934	2.043	2.096	2.241	1.918
CUT	*s-EQ*	0.133	0.231	0.266	0.278	0.227
	s-AC	2.562	2.151	1.837	1.652	2.029
	s-MMC	4.147	3.581	2.599	1.952	2.463
	s-AF	1.882	1.914	1.952	2.169	1.847
LCTA	*s-EQ*	0.164	0.239	0.278	0.311	0.249
	s-AC	2.611	2.282	1.733	1.597	2.257
	s-MMC	3.819	3.083	2.441	1.784	2.515
	s-AF	2.03	2.075	2.133	**2.346**	2.121

The bold figure means the optimal value.

normalized semantic evaluation models are shown in Fig. 15, where the average increment of 20 datasets is plotted using the ladder diagram on the bottom, the filled ladder diagram represents the increment is larger than 0.015.

It can be seen from Fig. 15, for $\sigma > 3.5$, the increment of *s-Silhouette*, *s-Ductance*, *s-Expansion*, *s-Ncut* is less than 0.015, implying the differentiation of them is weak. Therefore the *s-Silhouette*, *s-Ductance*, *s-Expansion*, *s-Ncut* are merely suit to evaluate the communities with big differentiation. The efficiency of *s-EQ*, *s-AC*, *s-MMC*, *s-AF* are better than *s-Silhouette*, *s-Ductance*, *s-Expansion*, *s-Ncut*.

5.3. The comparison on non-semantic community detection algorithms

In this section we choose the representative community detection algorithm COPAR (Gregory, 2010), Infomap (Rosvall & Bergstrom, 2008), LFM, as the non-semantic algorithms, comparing the performance of *EQ*, *s-EQ*, *AC*, *s-AC*, *MMC*, *s-MMC*, *AF*, *s-AF*. We generate 12 sets of datasets utilizing the LFR benchmark (Lancichinetti & Fortunato, 2009) with $|G| = 2000$, $ad = 5$, $dmax = 30$, $cmin = 10$, $cmax = 100$, $on = 300$, $om = 4$, $mi = 2.5$. To intuitively contrast the difference from each algorithm, the averages of 12 datasets are ascending sorted in Fig. 16. The analysis from Fig. 16 is the following: (1) The *EQ*, *AF* of ARTs are lower than that of non-semantic algorithms, however the *AC*, *MMC* are higher than them. (2) The *s-EQ*, *s-AF* of ARTs are higher than that of non-semantic algorithms, however the *AC*, *MMC* are lower than them. That has verified that ARTs has an undesirable topological structure but a more reasonable semantic structure than non-semantic algorithms.

5.4. The comparison on semantic community detection algorithms

In this section, we give a comparison on the representative semantic community detection algorithms which need to preset the number of communities. We choose the Enron dataset (McCallum et al., 2007) which is widely used in semantic community detection as the experimental data. The Enron dataset contains data from about 150 users, mostly senior management of Enron, about 0.5 M items. Table 3 is the four groups of topics extracted from the Enron dataset by LDA analysis. For the *s-EQ*, *s-AC*, *s-MMC*, *s-AF* have been verified to be the appropriate evaluation modes, we choose them to measure the semantic communities. Table 4 is the *s-EQ*, *s-AC*, *s-MMC*, *s-AF* obtained by the semantic algorithms of TURCM (Sachan et al., 2011), CART (Pathak et al., 2008), CUT (Zhou et al., 2006), LCAT (Yin et al., 2012). The *CS* in Table 4 represents the number of communities. By the comparison of Table 4, the optimal number of communities for Enron is 10 for each semantic community detection algorithm. For ARTs, the detected number of communities is 11, the optimal *s-EQ*, *s-AC*, *s-MMC*, *s-AF* is 0.352, 0.134, 0.148, 0.258 respectively. It's verified the result of ARTs approaches to the optimum of all the semantic community detection algorithms. And the advantage of ARTs is need not to preset the number of communities.

6. Conclusion

We have presented the multiple sampling ARTs model, the improving LDA models for overlapping community detection, which avoid presetting the number and the size of communities. The multiple sampling, with sampling frequency equal to 2, can either accelerate the convergence or improve the efficient of sampling. The proposed multiple sampling ARTs models, treating the field as the clustering unit, can achieve overlapping community detection.

For the Semantic Social Network, the ARTs models can be applied to discover the dynamic topic, model the structural transformation, and predict the emotional tendency. The model would form a useful component in systems for routing message recommendation and prioritization, and understanding the interactions in an organization in order to make recommendation about improving organizational efficiency.

The ARTs model explicitly quantizes the semantic information hidden in the social network. Future work will develop the ARTs models to quantize the dynamic community-topic relationship.

Acknowledgements

We acknowledge the support of the National Natural Science Foundation of China under Grant Nos. 61370083, 61073043, 61073041, 61370086; the National Research Foundation for the Doctoral Program of Higher Education of China Nos. 20112304110011, 20122304110012.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Cha, Y., & Cho, J. (2012). Social-network analysis using topic models. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 565–574). ACM.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821–7826.

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics, 12*(10), 103018.

Henderson, K., & Eliassi-Rad, T. (2009). Applying latent Dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on applied computing* (pp. 1456–1461). ACM.

Henderson, K., Eliassi-Rad, T., Papadimitriou, S., & Faloutsos, C. (2010). Hcdf: A hybrid community discovery framework. *SDM* (Vol. 2010, pp. 754–765). SIAM.

Jin, D., Yang, B., Baquero, C., Liu, D., He, D., & Liu, J. (2011). A markov random walk under constraint for discovering overlapping communities in complex networks. *Journal of Statistical Mechanics: Theory and Experiment, 2011*(05), P05031.

Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI* (Vol. 3, p. 5).

Lancichinetti, A., & Fortunato, S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E, 80*(1), 016118.

Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics, 11*(3), 033015.

McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, 3.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artifical Intelligence Research (JAIR), 30*, 249–272.

Mei, Q., Cai, D., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web* (pp. 101–110). ACM.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*(6), 066133.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577–8582.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814–818.

Pathak, N., DeLong, C., Banerjee, A., & Erickson, K. (2008). Social topic models for community extraction. In *The 2nd SNA-KDD workshop* (Vol. 8). Citeseer.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*(4), 1118–1123.

Sachan, M., Contractor, D., Faruquie, T., & Subramaniam, V. (2011). Probabilistic model for discovering topic based communities in social networks. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 2349–2352). ACM.

Sachan, M., Contractor, D., Faruquie, T. A., & Subramaniam, L. V. (2012). Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st international conference on World Wide Web* (pp. 331–340). ACM.

Shen, H., Cheng, X., Cai, K., & Hu, M.-B. (2009). Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications, 388*(8), 1706–1712.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 306–315). ACM.

Wang, X., Mohanty, N., & McCallum, A. (2005). Group and topic discovery from relations and text. In *Proceedings of the 3rd international workshop on link discovery* (pp. 28–35). ACM.

Yin, Z., Cao, L., Gu, Q., & Han, J. (2012). Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(4), 63.

Zhang, H., Giles, C.L., Foley, H.C., & Yen, J. (2007). Probabilistic community discovery using hierarchical latent gaussian mixture model. In *AAAI* (Vol. 7, pp. 663–668).

Zhang, H., Li, W., Wang, X., Giles, C. L., Foley, H. C., & Yen, J. (2007). Hsn-pam: Finding hierarchical probabilistic groups from large-scale networks. In *Data mining workshops, 2007. ICDM workshops 2007. Seventh IEEE international conference on* (pp. 27–32). IEEE.

Zhang, H., Qiu, B., Giles, C. L., Foley, H. C., & Yen, J. (2007). An lda-based community structure discovery approach for large-scale social networks. In *Intelligence and Security Informatics, 2007 IEEE* (pp. 200–207). IEEE.

Zhou, D., Manavoglu, E., Li, J., Giles, C. L., & Zha, H. (2006). Probabilistic models for discovering e-communities. In *Proceedings of the 15th international conference on World Wide Web* (pp. 173–182). ACM.

Zhu, X., Ghahramani, Z., Lafferty, J. et al. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML* (Vol. 3, pp. 912–919).