Modeling one-mode projection of bipartite networks by tagging vertex information

Jian Qiao[a,*], Ying-Ying Meng[a], Hsinchun Chen[b], Hong-Qiao Huang[a], and Guo-Ying Li[a]

[a]School of Management, Northwestern Polytechnical University, Xi’an 710072, China

[b]Eller College of Management, University of Arizona, Tucson, AZ 85721, USA

Traditional one-mode projection models are less informative than their original bipartite networks. Hence, using such models cannot control the projection’s structure freely. We proposed a new method for modeling the one-mode projection of bipartite networks, which thoroughly breaks through the limitations of the available one-mode projecting methods by tagging the vertex information of bipartite networks in their one-mode projections. We designed a one-mode collaboration network model by using the method presented in this paper. The simulation results show that our model matches three real networks very well and outperforms the available collaboration network models significantly, which reflects the idea that our method is ideal for modeling one-mode projection models of bipartite graphs and that our one-mode collaboration network model captures the crucial mechanisms of the three real systems. Our study reveals that size growth, individual aging, random collaboration, preferential collaboration, transitivity collaboration and multi-round collaboration are the crucial mechanisms of collaboration networks, and the lack of some of the crucial mechanisms is the main reason that the other available models do not perform as well as ours.

1. Introduction

Bipartite networks are an important class of complex networks. A bipartite network is composed of two types of vertices and the edges running only between the vertices of unlike types. Many natural, social, and technical systems can be represented as bipartite networks, such as a coauthors network [1], movie actors network [2], directors network [3], recommendation system [4], and so on [5-8].

In the past several decades, the characteristics of extensive bipartite networks have been analyzed empirically. For example, Newman [1, 9-11] analyzed many statistical properties of scientific collaboration networks in the realms of physics, mathematics, biomedicine and computer science. Lambiotte and Ausloos [12] analyzed the properties of a bipartite network about people sharing their music library. Shang et al. [13] reported the empirical analysis of two large-scale web sites, in which users are connected by music groups and bookmarks, respectively. Zhang et al. [14] presented an empirical study on the Bus Route Networks of Beijing and Yangzhou, Travel Route Network of China, Huai-Yang recipes of Chinese cooked food, and collaboration network of Hollywood actors.

In addition, many new statistical indices for bipartite networks have been proposed in recent years, such as a clustering coefficient based on the fraction of cycles with size four defined by Lind et al. [15], two edge clustering coefficients based on squares and triples, respectively, proposed by Zhang et al. [16], and an index called “collaborative similarity to quantify the diversity of tastes based on the collaborative selection” proposed by Shang et al. [13].

To gain insight into the evolution of bipartite systems, many two-mode network models, such as the sexual contact network model [17], collaboration network model [18], plant-animal mutualistic network model [19], ecological and organizational network model [7], general bipartite network model [20], online bipartite network model [21], and a model for the self-assembly of creative teams [22], have been developed. Two-mode network models are natural in form and can hold information about the complete structure. Another modeling tool, the hyper graph [23], can also hold the complete structure information of bipartite networks [24, 25] because each of its edges, also known as hyper edges, can relate groups of more than two vertices.

Yet, people perhaps are more interested in the one-mode projection of bipartite networks in many scenarios. For instance, for scientific collaboration networks, we usually are more interested in the relationships between scientists rather than the relationships between scientists and their publications. Because of this reason, many unweighted one-mode collaboration models have been proposed. Barabási et al. [26] proposed a model for capturing the temporal evolution of collaboration networks. A one-mode collaboration network model developed by Zhou et al. [27] interpolates between the networks that follow a power-law and an exponential degree distribution. Guimerà et al. [22] designed a model about the self-assembly of creative teams. Zhang et al. [14] suggested a model to understand the evolutionary mechanisms of four non-social systems and a social system. Because many of the informative structures have to be ignored, unweighted one-mode models could not hold the complete structure information of bipartite networks [28, 29]. For instance, from the unweighted projection of a scientist-paper network, we know who are the collaborators of each scientist, but we cannot accurately tell who are the authors of each paper.

Recently, many weighted one-mode models were proposed to contain the structure information of bipartite networks more completely. In the scientist network model designed by Ramasco and Morris [30], the edges of the model are weighted by the times of collaboration. Ke and Ahn [31] proposed a weighted model for reproducing the observed pattern in scientist networks—local clusters consist of dense, weak ties and are interconnected by sparse, strong ties. Zhou et al. [32] proposed a weighting method using asymmetrical weights and self-connection to mimic the information that coauthors might assign a specific paper with different weights. Apparently, weighted one-mode models are more informative than unweighted ones, but such models still have some limitations. First, any weighting method could not exclude subjective factors completely. In some methods, the information of those vertices with one degree is even lost in the projection [32]. Second, the information contained in the models by the static weights of the links is unreliable because the link weight remains constant after being assigned to an edge [33]. Third, the evolving weighted models are time-consuming because a great number of link weights have to be updated instantly at each time step.

In this paper, we propose a new method for modeling a one-mode projection of bipartite networks that overcomes the deficiency of traditional methods. Modeling a one-mode projection of bipartite networks with our method maintains the complete structure information of the original systems, and more importantly, the projection’s topological structure can be controlled very flexibly. As an application example, we design a one-mode collaboration network model with our modeling method to verify the method’s feasibility and explore the crucial evolving mechanisms of some real collaboration networks.

The rest of this paper is arranged as follows. Section two is the introduction of our method for modeling the one-mode projection of bipartite networks. In section three, the statistical indicators used to measure the local and global properties of the one-mode projection of bipartite networks will be introduced. We will infer the possible evolutionary mechanisms of collaboration networks in section four by observing and analyzing the local and global statistical properties of some real data to provide the design of our collaboration network model a strong basis. In section five, we will describe and interpret our one-mode collaboration network model in detail. The feasibility of our modeling method and the performance of our collaboration network model will be examined and discussed through numerical experiments in section six. Finally, we will summarize our research results in the last section.

2. Method for modeling a one-mode projection of bipartite networks

Traditional unweighted and weighted projecting methods cannot completely contain the structure information of a bipartite network; therefore, if such methods are used to model a one-mode projection of the bipartite network, it is impossible to very flexibly control the projection’s topological structure because we cannot precisely choose existing vertices or edges for operation. To solve this problem, we proposed a tag-based one-mode projection modeling method, in which each vertex of the projection has a so-called tag-set, and another type of the vertices in the bipartite network is expressed as tags in the form of incremental natural numbers and stored into the tag-sets of the projection's vertices that there are edges between the two types of vertices in the original bipartite network. Fig. 1 demonstrates the process of producing a growing one-mode collaboration network with our method as well as a map relationship between the one-mode projection and corresponding bipartite network.

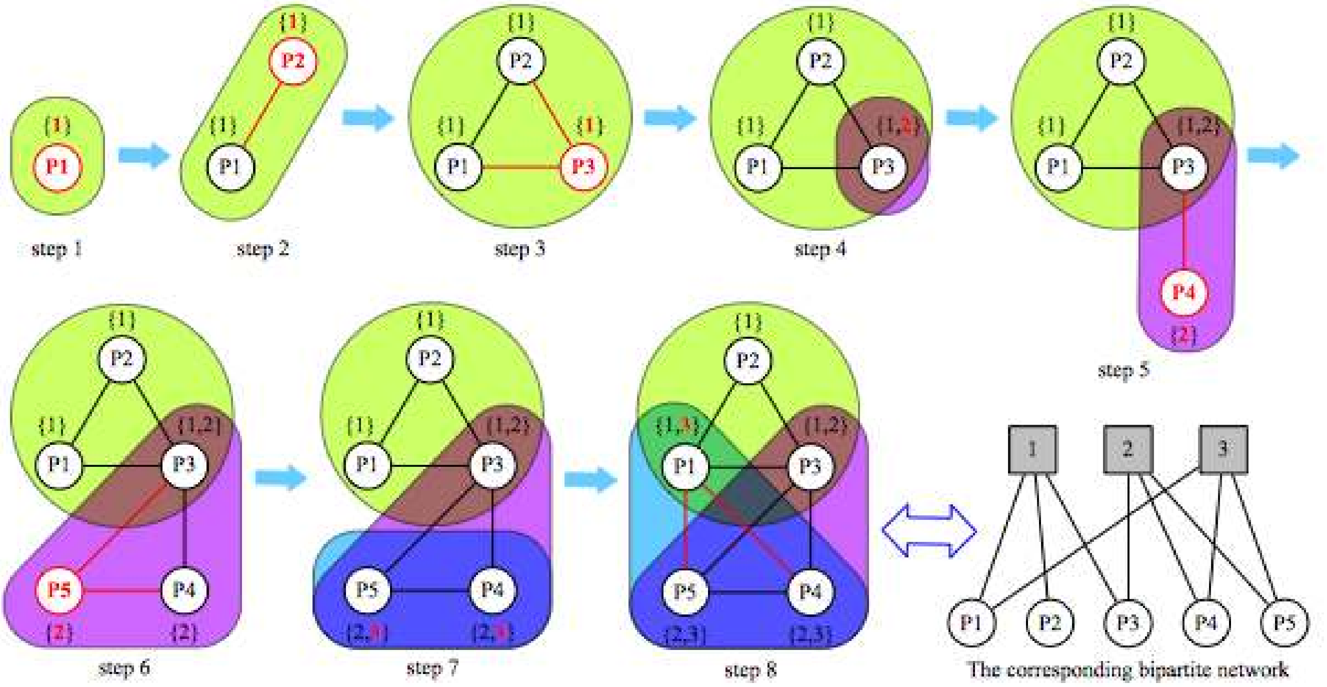

Fig. 1. (Color online) The process of producing a growing one-mode collaboration network with our modeling method as well as a map relationship between the one-mode projection and corresponding bipartite network, where the small circles marked with P1, P2, …, or P5 are participant vertices; the small squares marked with 1, 2, or 3 are the act vertices; the pairs of braces are the tag-sets of the participant vertices; the natural numbers 1, 2, and 3 in the tag-sets are tags, which denote acts; the red vertices, red edges and red tags denote the newly added ones for current step; and each color-shape encompasses the vertices with a same tag. The process starts with an empty network. Then, all of the vertices and tags are added uniformly step-by-step. In step 1, vertex P1 is added in the network and tag 1 is added in the tag-set of P1, which means that P1 is the first participant of act 1; in step 2, P2 is added and tag 1 is selected from the available tags and copied to the tag-set of P2, which means that P2 joins the group of act 1; in step 3, P3 is added and tag 1 is selected and copied to the tag-set of P3, which means that P3 joins the group of act 1; in step 4, P3 is selected from the available vertices and tag 2 is added in the tag-set of P3, which means that P3 is the first participant of act 2; in step 5, P4 is added and tag 2 is selected and copied to the tag-set of P4, which means that P4 also joins act 2; in step 6, P5 is added and tag 2 is selected and copied to the tag-set of P5, which means that P5 is the third participant of act 2; in step 7, P4 and P5 are selected from the group of act 2 and tag 3 is added to the tag-sets of them, which means that P4 and P5 join a new act 3; and in the last step, P1 and tag 3 are selected and the latter is copied to the tag-set of the former, which means that P1 also join act 3.

Fig. 1 shows that our method can control the scope of operation precisely while constructing a one-mode projection of bipartite networks. For example, in step 6, we can make P5 link only with P3 and P4 by utilizing tag 2, and in step 7, we can limit tag 3 to be copied only to portion of those vertices with tag 2. In addition to this most important property, our method also has the following by-product benefits: (i) the one-mode projection contains the complete information of the corresponding bipartite network, (ii) the one-mode projection and corresponding bipartite graph can be converted to each other freely, (iii) the one-mode projection can be simplified to an unweighted form by removing the tag-sets, or to a weighted form by employing a specific weighting method, and (iv) the tags are helpful to analyze the topological properties of the projection. Nevertheless, both unweighted and weighted projections do not have the above benefits because they are less informative compared to the original bipartite graph.

3. Statistics of our one-mode projections

We interpret the statistical indicators introduced in this section with the language of collaboration networks when necessary. We employ the following statistical indicators to measure the local properties of our one-mode projections. $P(k)$ denotes the probability of the degree of any vertex being k; $S(n)$ denotes the number of tags who are owned by n vertices; $Q(q)$ denotes the number of vertices who own q tags; and $C(k)$ denotes the average clustering coefficient of the vertices with degree k. Mathematically,

$$C(k)=\frac{1}{|\{i|k_i=k\}|}\sum\nolimits_{k_i=k} C_i, \tag{1}$$

where the clustering coefficient C_i of vertex i is defined as the quotient of the actual number of edges e_i and the possible number of edges $k_i(k_i-1)/2$ between the nearest neighbors of i, i.e.,

$$C_i=\frac{2e_i}{k_i(k_i-1)}. \tag{2}$$

The C_i reflects the clustering level of the nearest neighbors of i. The average degree $k_{nn}(k)$ of the nearest neighbors of the vertices with degree k reflects the average degree-degree correlation between k-degree vertices and their nearest neighbors and is defined as

$$k_{nn}(k)=\frac{1}{|\{i\,|\,k_i=k\}|}\sum\nolimits_{k_i=k}\langle k_{nn,i}\rangle, \tag{3}$$

where the average degree $\langle k_{nn,i}\rangle$ of the nearest neighbors of i reflects the degree-degree correlation between i and its nearest neighbors and is defined as

$$\langle k_{nn,i}\rangle=\frac{1}{k_i}\sum\nolimits_{j\in V_{nn,i}} k_j, \tag{4}$$

where $V_{nn,i}$ denotes the set of the nearest neighbors of i. A network is referred to as assortative when $k_{nn}(k)$ is an increasing function of k; otherwise, the network is disassortative. The number of the maximal cliques with K vertices is denoted as $N_{clq}(K)$. Here a clique is a graph that any two of its vertices are adjacent, and a maximal clique refers to a graph that is not the subgraph of any other clique.

We use the following statistical indicators to measure the global properties of the one-mode projections of bipartite networks. N_v denotes the number of vertices; N_e is the number of edges; N_a is the number of existing non-copied tags; $\langle k\rangle$ is the average degree of vertices, i.e., the average number of collaborators per participant; $k_{\max}$ is the degree of the vertex with the maximum degree; $\langle q\rangle$ is the average number of tags owned by a participant vertex, i.e., the average number of acts participated in by a participant; $\langle n\rangle$ is the average number of vertices who own the same tag, i.e., the average number of collaborators participating in an act; D is the network density, i.e., the ratio of the number of the actual edges versus the number of possible edges of a network; N_c is the number of connected subgraphs; N_m is the size of the maximum connected subgraph; and C is the global level of the clustering of a network and is defined as:

$$C = \frac{1}{N_v} \sum\nolimits_{i=1}^{N_v} C_i. \tag{5}$$

The symbol L denotes the average length of the shortest paths between all pairs of the vertices of the maximum connected subgraph of a network. Mathematically,

$$L = \frac{1}{N_m(N_m - 1)} \sum\nolimits_{i \neq j} L_{ij}, \tag{6}$$

where the shortest path length L_{ij} from vertex i to vertex j refers to the number of edges on this path between the two vertices. r denotes the degree-degree correlation coefficient, i.e., the Pearson correlation coefficient, and is used to measure the extent of assortativity of a network, which lies in the interval [-1, 1] and is expressed as [34]

$$r = \frac{M^{-1}\sum_i u_i v_i - \left[M^{-1}\sum_i (u_i + v_i)/2\right]^2}{M^{-1}\sum_i \frac{1}{2}(u_i^2 + v_i^2) - \left[M^{-1}\sum_i (u_i + v_i)/2\right]^2}, \tag{7}$$

where u_i and v_i denote the degree of two endpoints of the ith edge, respectively, and M is the number of edges. A network is classified as assortative if $r > 0$; otherwise, it is disassortative.

4. Empirical analysis of real collaboration networks

Before introducing a one-mode collaboration network model designed with our method in the next section, here we perform an in-depth analysis on the evolutionary mechanisms of three real collaboration networks, i.e., the networks composed of actors staring in movies, scientists publishing papers, and directors sitting on boards, to make our model more understandable. The actor network data includes 127,823 movies and was obtained from the Internet Movie Database by Barabási and Albert [35]. Because the number of the movies included in whole data is too large to calculate $N_{clq}(K)$ and L, we only choose the data of the first 15,000 movies. The scientist network data were collected from the condensed matter preprint database at Los Alamos by Newman [1]. The director network data include the "Fortune 1000" US companies from 1982 to 2001 and were collected by Davis et al. [3]. The global statistics of these networks are listed in the "Real data" columns of Table 2.

Both the scientist network and the actor network are growing systems because they all are constructed with accumulated data day by day. The director network is a snapshot of the cross-section data at a specific time. Therefore, the network is an evolving system rather than a growing one, which means that the coming and leaving of participants in the network always coexist. An evolving network also could be regarded roughly as a growing system starting from an empty set and ending with a limited size. In collaboration networks, the participants of an act, such as the authors of a paper, the actors of a movie and the directors of a board, make up a complete subgraph in the one-mode projection.

Statistical indicators of the three real networks shown in Table 2 and Fig. 2 reflect the differences among these networks. From Table 2, we find that N_v, N_a and $\langle q \rangle$ of the scientist network are significantly larger than those of the director network but N_e of the scientist network is definitely smaller than the director network. The cause of these results is: (i) $\langle n \rangle$ of the scientist network is far smaller than the director network, which indicates that the number of authors of a paper is on average much less than number of directors of a board, and (ii) transitivity collaboration, i.e., the collaboration that occurs between the collaborators of a participant, and multi-round collaboration, i.e., a group of participants collaborating with each other repeatedly, are more popular in the scientist network than in the director network. A specific characteristic of multi-round collaboration is that it adds new acts (i.e., new tags) rather than new edges to a collaboration network. In the scientist network, there generally are numerous teams and each of them is composed of young teachers, and/or PhD and master students who are led by a professor. However, common sense tells us there is no similar organization in the director or actor networks. The differences between these networks in $\langle n\rangle$, $\langle q\rangle$ and $Q(q)$ indicate that the collaboration acts in the scientist network often occur locally within the teams with the smallest $\langle n\rangle$, smaller $\langle q\rangle$ and stronger preference, while in the director network the collaboration acts often occur globally with the largest $\langle n\rangle$, smallest $\langle q\rangle$ and weakest preference, and in the actor network, collaborations often occur globally with a very large $\langle n\rangle$, the largest $\langle q\rangle$ and strongest preference. Here, the preference refers to the preference to collaborate, i.e., a participant with more collaboration records has a higher probability of participating in new collaboration acts. Due to the aforementioned factors, the director network holds maximum C, D, and r, and minimum $k_{\max}$; the scientist network holds maximum L, and N_c/N_v, and minimum $\langle k\rangle$, D, and N_m/N_v; and the actor network holds maximum $\langle k\rangle$, $k_{\max}$ and N_m/N_v, and minimum C, r, L, and N_c/N_v. Specially, transitivity collaboration plays a very important role in strengthening C and r of all of the networks.

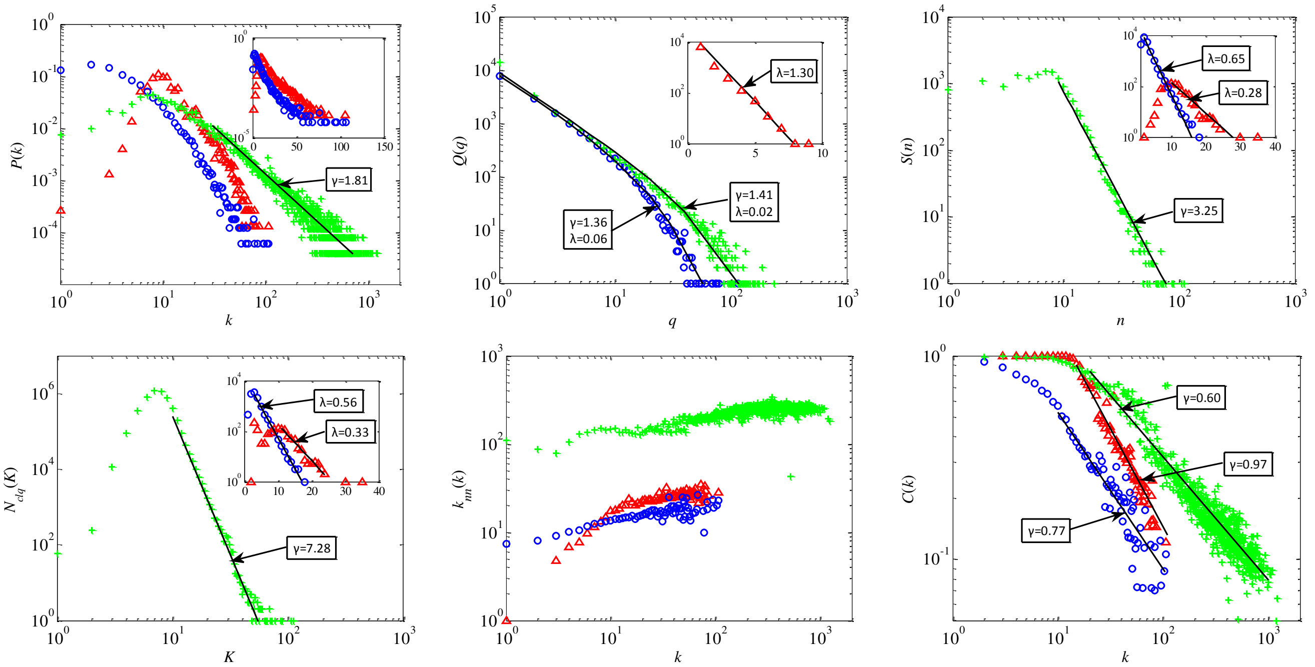

Fig. 2. (Color online) Local statistical properties of three real collaboration networks. The scatter plots with the green crosses represent the actor network, those with the blue circles represent the scientist network, and those with the red triangles represent the director network. Each solid line marked by a single γ value is the fit of a power-law form $k^{-\gamma}$, each solid line marked by single λ value is the fit of an exponential form $e^{-\lambda k}$, and each solid line marked by both γ and λ values is the fit of a power-law form with an exponential cut-off $k^{-\gamma}e^{-\lambda k}$.

From Fig. 2, we find that the mainbody of the $P(k)$ distribution of the actor network follows a power-law, the scientist network is between the exponential and power-law forms, and the director network closely approaches an exponential form. These features indicate that the probability of a participant running into a new collaborator is positively correlated to the number of collaborators this participant has collaborated with, but the correlation level is different from each other. The $Q(q)$ distribution of both the actor network and scientist network follows a power-law with an exponential cut-off, while the director network follows an exponential form. These features reflect that the probability of a participant participating in a new act has a positive correlation with the number of acts this participant has participated in, but the correlation level is different from each other. Multi-round collaboration and the aging of participants have crucial impacts on the $Q(q)$ distribution. The mainbody of the $S(n)$ distribution of the actor network follows a power-law, the scientist network follows an exponential form, and the director network approximately follows a Gaussian distribution. Apparently, in the actor network and scientist network, the number of participants with very few acts is very large and the number participants with very many acts is very small, but the significance level of this feature in the two networks is different from each other. Both $S(n)$ and $N_{clq}(K)$ describe the distribution of the complete subgraphs and their difference is that any complete subgraph described by the $S(n)$ is composed of the participants of an act, while any complete subgraph described by the $N_{clq}(K)$ is composed of the members of a maximal clique, which is why the distributions of $N_{clq}(K)$ are very similar to those of $S(n)$. In addition, as mentioned before, frequently occurred transitivity collaboration and a large $\langle n\rangle$ enhance the assortativity and clustering of the networks. The difference between the histograms of $k_{nn}(k)$ or $C(k)$ is caused by the different $\langle n\rangle$ and $\langle q\rangle$, as well as preferential collaboration.

5. One-mode collaboration network model

In light of the previous analysis of the three real collaboration networks, we designed a growing one-mode collaboration network model using our method. Here the vertices and tags denote participants and acts, respectively. Suppose that N_v, N_a and D of the target one-mode projection are known. Our model starts from an empty network. At each time step, the following four substeps will be performed in order, and the model will end after running N_v time steps:

1) Add a new active vertex i into the network and let $sum_a = sum_a + N_a/N_v$, where sum_a denotes the accumulated number of tags that should be but are not added into the network during the previous and current time steps, hence we set $sum_a = 0$ for the initial time. For simplicity, we add new tags into the network at a constant speed. Thus, the average number of new tags added at each time step is N_a/N_v. If i is the first added vertex, add the first new tag into its tag-set. Otherwise, add a new tag with probability $1-\eta|A|/N_a$ or a copy of a randomly chosen tag (random collaboration) with probability $\eta|A|/N_a$ into the tag-set of i. If the added tag is new, subtract sum_a by 1. Here $|A|$ denotes the number of existing non-copied tags, and η is a scaling factor that is in the interval [0,1]. Apparently, the probability of a newcomer (new added vertex) participating in a new act (new added tag) is inversely proportion to the number of existing acts, which follows common sense. Here, η is used to control the number of connected subgraphs. The process of adding vertices and tags can enhance the right skewness of the distribution of $P(k)$, $Q(q)$ and $S(n)$.
2) If there are at least two active vertices in the network, with the probability of P_a, one of the active vertices is preferentially chosen according to the number of tags every active vertex has. Then, the chosen active vertex inactivated. Obviously, the active vertices with more tags have a bigger probability of being transformed to the inactive (aging) state. New tags or the copies of existing tags cannot be added in the tag-sets of the inactive vertices anymore because of aging.
3) If sum_a is greater than or equal to 1, repeat the following operations to add new tags into the network: first, choose a tag randomly from the existing tags owned by active vertices; second, choose a vertex randomly from the active vertices with the chosen tag; third, add a new tag into the tag-set of the chosen vertex; fourth, with probability P_m, add a copy of the new added tag into the tag-set of each of the chosen vertex's active neighbors with the chosen tag; finally, subtract 1 from sum_a. The first three operations allow the active vertices with more tags have more opportunities of acquiring a new tag (preferential choice); the fourth operation allows some vertices with common tag(s) an opportunity to acquire a new common tag again (multi-round collaboration). The first four operations jointly determine the level of preferential and multi-round collaborations.

4) If $|V|>1$ and $d<|V|^{-\alpha}$, repeat the following operations to increase the density of the network: first, choose an active vertex randomly (random choice) with probability P_r or preferentially (preferential choice) with probability $1-P_r$ according to the number of tags that every active vertex has; second, if the chosen vertex does not have some tags of some of its neighbors, with probability P_t, choose such a neighbor randomly (transitivity choice), and then, choose such a tag of this neighbor randomly and add a copy of this tag into the tag-set of the chosen vertex; otherwise, choose a tag randomly (random choice) and add a copy of the chosen tag into the tag-set of the chosen vertex. Here $|V|$ and d denote current size and density of the network. In real growing networks, the number of actual edges increases far slowly than the potential edges, which results in d decreasing nonlinearly as $|V|$ increases, so we can suppose $d=|V|^{-\alpha}$, where α is a constant and its value is $\alpha=-\log D/\log N_v$. All of the operations in this substep jointly determine the probabilities of random, preferential and transitivity collaborations occurring among active vertices and control both the size of the maximal connected subgraphs and number of connected subgraphs.

6. Experimental results and discussions

We tested the performance of our model by simulating some real collaboration networks with it and comparing the results with other models. The simulated real networks are the scientist network, actor network and director network analyzed in section four. Five candidate collaboration network models for comparison are the RDP model proposed by Ramasco, Dorogovtsev and Pastor-Satorras [18], GUSA model proposed by Guimerà, Uzzi, Spiro and Amaral [22], Zhou model proposed by Zhou et al. [27], Tian model proposed by Tian et al. [20] and ZZL model proposed by Zhang, Zhang and Liu [21].

The RDP model starts from an empty bipartite network. At each time step, a new act with n participants is added to the network, where n is a random number following the $S(n)$ distribution. m of n participants are new without previous experience, where m is a random number following an exponential decay $\bar{m}$. The remaining $n-m$ of the n participants are chosen from existing individuals with a probability proportional to the number of acts attended previously by them. The existing individuals who have attended Q_0 more acts become inactive with a probability given by the complementary of an exponential decay τ. Here, inactive individuals could not participate in new acts anymore. The total number of acts denotes the total number of time steps.

The GUSA model also starts from an empty bipartite network. At each time step, a new act with n participants is added to the network, where n is a random number following the $S(n)$ distribution. If there are some participants of the existing acts who are not the new participants of an act, each of the new participants has a probability, p, of being drawn from such individuals and a probability, $1-p$, of being added as a newcomer. If the participant is drawn from participants of the existing acts and at least one of them has been the participant of the new act, then (i) with probability q, the new participant is randomly selected from among the set of collaborators of a randomly selected participant of the existing acts already participating in the new act; (ii) otherwise, he or she is selected at random from all of the participants of the existing acts. Lastly, participants of the existing acts who remain inactive (i.e., do not participate in new acts) for longer than τ time steps are removed from the network. The total number of acts denotes the total number of time steps.

The Zhou model starts from a one-mode network with m_0 fully connected vertices. Then, at each time step, they add a new vertex into the network to collaborate with some existing vertices. An existing vertex with degree k will be chosen with the probability $\lambda k^{\alpha}/\sum_i k_i^{\alpha}$ to be an actor in the collaboration, where (is a constant and denotes the level of preferential attachment. Using $\langle s\rangle$ to represent the average value of the act-size, such as the mean number of authors per paper, they conclude that $\langle s\rangle=\lambda+1$. Thus, l can be used to control the average act-size of the whole network and it is not free when an idiographic network of known average act-size is simulated. All of the chosen existing vertices will link to the new vertex, and if two chosen vertices have never collaborated so far, there will be a new edge connecting them. The model will stop running when the expected network size is reached.

The Tian model and ZZL model are two bipartite models. A common mechanism of them is that at each time step a new act and a new participant will be added synchronously into the network, which means that the speed of adding new acts is exactly the same as adding new participants. We believe that such a mechanism is far from that of the simulated real networks. For example, it is hard to imagine that new scientists and new papers are born synchronously at the same speed in scientist-paper networks, new actors and new movies are born synchronously at the same speed in actor-movie networks, or new directors and new boards are born synchronously at the same speed in director-board networks. The fact that N_v and N_a of the three real networks listed in Table 2 are distinctly different from each other fully supports our hypothesis. Our experiments also show that the simulation results of the Tian model and ZZL model are far from the real data. To save space, we only show and discuss the simulation comparisons of the RDP, GUSA, Zhou and our models.

Just as we have analyzed in section four, all three real networks that are simulated in this section using the four models can be classified as growing networks without leaving behavior. Therefore, we set $\tau = \infty$ to disable the remove operation when simulating the three real networks with the GUSA model. Table 1 lists the parameter settings of these models.

Table 1. Parameter settings of the four models

Simulated Network	Our model					RDP model			GUSA model			Zhou model		
	η	P_a	P_m	P_r	P_t	$\overline{m}$	Q_0	τ	p	q	τ	α	λ	m_0
Scientist network	0.25	0.25	0.2	0.8	0.8	0.76	15	7	0.75	0.8	∞	0.6	1.66	3
Actor network	0.08	0.2	0.7	0.2	0.65	1.68	50	25	0.8	0.6	∞	1	7.01	8
Director network	0.95	0.6	0.2	1	0.3	8.38	3	1.5	0.3	0.42	∞	2	10.02	11

The global statistics of the three real networks and twelve model networks are listed in Table 2, in which the results in blue color are the best ones. Because N_a is used as the end or control condition of the RDP, GUSA and our models, this indicator of the networks produced by the three models is fully the same as the corresponding real networks. Meanwhile, because N_v and D are also used as the end and control conditions, respectively, of our model, so these indicators as well as N_e and $\langle k \rangle$ of the networks produced by our model are fully same or very close to those of the corresponding real data. Overall, most indicators of our model networks are significantly closer to the real data than the RDP, GUSA and Zhou models.

Table 2. Global statistics of the three real collaboration networks and twelve corresponding model networks

Index	Scientist network & models					Actor network & models					Director network & models				
	Real	Ours	RDP	GUSA	Zhou	Real	Ours	RDP	GUSA	Zhou	Real	Ours	RDP	GUSA	Zhou
N_v	16726	**16726**	16459	19038	**16726**	25231	**25231**	23527	23839	**25231**	7680	**7680**	5282	7151	**7680**
N_a	22016	**22016**	**22016**	**22016**	16724	15000	**15000**	**15000**	**15000**	25224	916	**916**	**916**	**916**	7670
N_e	47594	**47595**	78107	34343	49391	620309	**620314**	577091	575383	647730	55437	56792	54223	47786	**56187**
$\langle k \rangle$	5.6910	**5.6911**	9.4911	3.6078	5.9059	49.170	**49.171**	49.058	48.2724	51.3440	14.437	14.790	20.531	13.365	**14.632**
$k_{\max}$	107	130	**127**	82	147	1223	**1349**	2001	1353	2024	106	121	**92**	85	7605
$\langle q \rangle$	3.5032	2.8639	**3.6477**	1.9123	2.6390	4.7637	**4.6838**	5.1057	4.4148	8.0281	1.3148	**1.2983**	1.8942	1.2960	7.8101
$\langle n \rangle$	2.6615	2.1757	2.7270	1.6536	**2.6393**	8.0129	7.8781	**8.0081**	7.0164	8.0303	11.024	10.885	**10.922**	10.118	7.8203
C	0.3596	**0.3688**	0.1119	0.3988	0.0970	0.1758	**0.1635**	0.1498	0.1981	0.1616	0.5871	**0.5629**	0.3684	0.6647	0.0054
D	0.0003	**0.0003**	0.0005	0.0002	0.0004	0.0019	**0.0019**	0.0021	0.0020	0.0020	0.0019	**0.0019**	0.0039	**0.0019**	**0.0019**
r	0.1846	**0.1682**	0.3108	0.2067	0.1622	0.1510	0.0564	-0.1263	**0.1469**	-0.1408	0.2746	0.2407	0.3932	**0.2616**	-0.5256
L	6.6276	6.5069	4.1689	**6.7358**	4.3894	3.4879	**3.1026**	2.8807	3.0851	2.7769	4.6040	**4.5630**	3.5619	4.9911	1.9990
N_c	1188	**1225**	1543	5470	3212	138	**129**	553	1	30	97	**104**	36	7	1
N_m	13861	**13856**	13964	12288	13515	24842	**25004**	22448	23839	25202	6731	**6977**	4957	7129	7680

Fig. 3 shows the distributions of the local indicators of the three real networks and twelve model networks. Similar to the previous results, our model outperforms the other models in most of the local indicators. Comparatively, the performances of the RDP and GUSA models are generally close to each other, while the Zhou model sometimes performs significantly poorer than the other models, especially for simulating the director network, where we can find all of the indicators of the network produced by the Zhou model are significantly different than the real data.

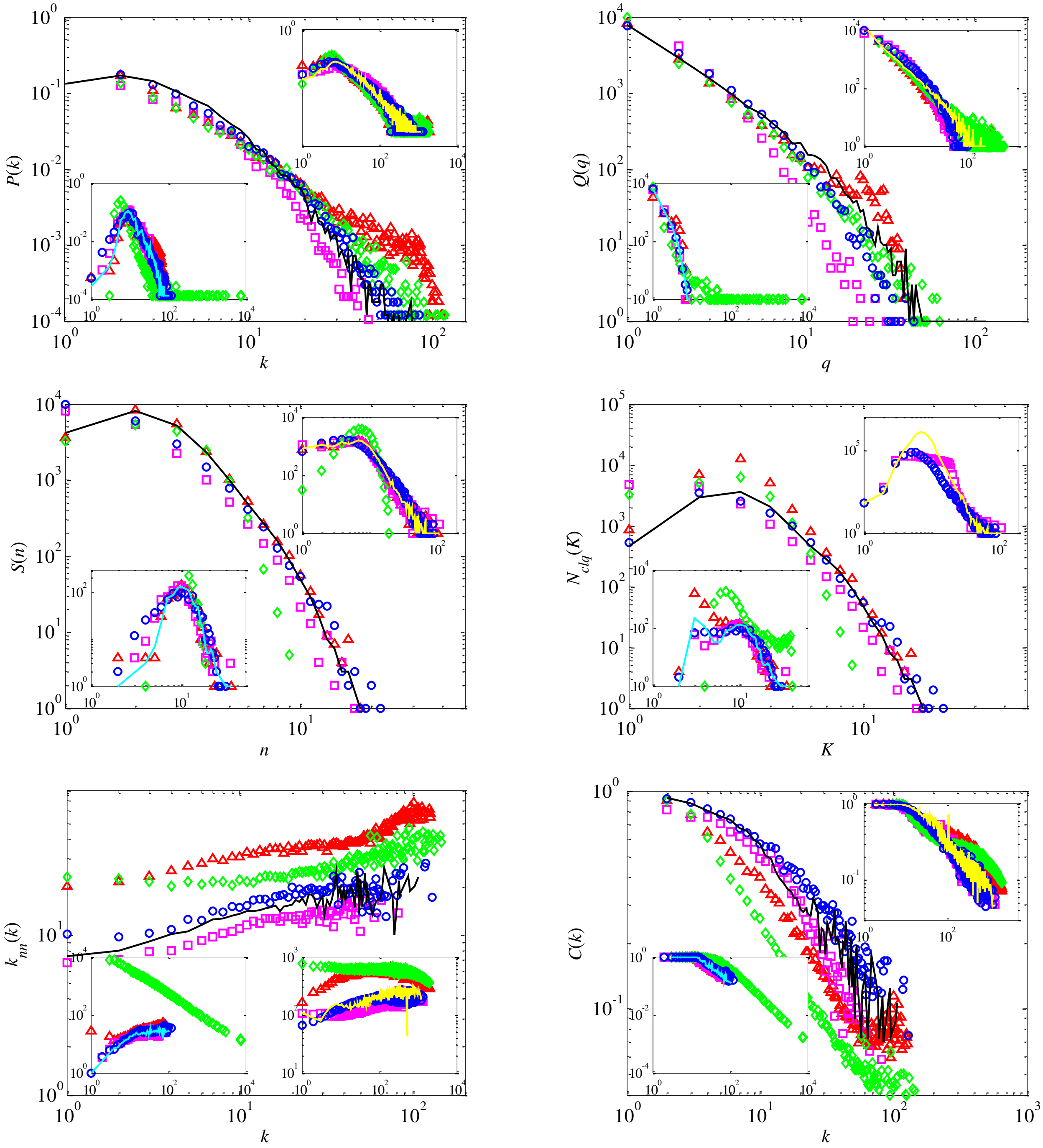

Fig. 3. (Color online) Local statistical properties of the three real collaboration networks and twelve corresponding model networks. The scattered blue circles are the results of our model, the red triangles are the results of the RDP model, the pink squares are the results of the

GUSA model, the green diamonds are the results of the Zhou model, and the black, yellow and cyan solid lines are the results of the real networks. The main plots are the results for the scientist network, the insets with yellow solid lines show the results for the actor network, and the insets with cyan solid lines show the results for the director network. **Note:** In the inset of $N_{clq}(K)$ for the actor network, the results of the networks produced by the RDP and Zhou models cannot be calculated because the corresponding network structure is too complex.

The cause of differing performance among the four models lies in the different mechanisms embedded in the models. In other words, by providing more crucial mechanisms, our model outperforms the RDP, GUSA and Zhou models. In the RDP model, because it lacks multi-round collaboration, the increase of $\langle n\rangle$ and $\langle q\rangle$ of the model networks originates mainly from the creation of a large number of new links between randomly chosen vertices, which leads to the density of the networks increasing quickly and finally being clearly higher than the real data; due to lacking transitivity collaboration, the clustering of the model networks is obviously lower than the real data, even though their density is significantly higher than that the real data. The full preferential choice mechanism of the RDP model also is one of the reasons leading to these results. In the GUSA model, because it lacks preferential collaboration, the clustering coefficients of the model networks always are significantly larger than the real data and other models, and the right skewness of the distributions of $Q(q)$ is significantly weaker than the real data and other models. In the Zhou model, because it lacks more crucial mechanisms, such as random collaboration, transitivity collaboration, multi-round collaboration and individual aging, more statistical indicators of the model networks are significantly different from the real data. Actually, the mechanism "at each time step, add a new vertex to the network to collaborate with some existing vertices" also is an important contributor to the observed results because it is equivalent to adding new acts as fast as adding new vertices, which should be not true in the three real networks.

Additionally, it is well known that any external statistical regularity of a network is essentially the result of its internal mechanisms. Therefore, the prerequisite that the distribution of $S(n)$ must be known beforehand in the RDP and GUSA models makes it possible that some crucial mechanisms of the real systems are put into a black box. In our model, apart from a few probability parameters, any statistical distribution, including $S(n)$, is an output result rather than a known condition, which means that our model favors getting insight into the mechanisms of the collaboration networks.

7. Conclusions

One-mode projection of bipartite networks has a wide range of applications. To overcome the limitation of traditional unweighted and weighted methods for projecting bipartite networks, we developed a tag-based method for modeling a one-mode projection of bipartite networks. The most important advantage of our method is that the existing vertices and edges can be chosen precisely during the producing a one-mode network, which means that the structure of the network is very flexible. Our method has additional merits. For example, the produced one-mode projection can be converted into its original bipartite graph very easily, and it also can be easily converted to an unweighted form by deleting the tag-sets of vertices or to a weighted form by employing a specific weighting method.

Simulation comparisons show that the one-mode collaboration network model designed with our method outperforms the available one-mode and two-mode models significantly and matches three real networks very well. It implies that (i) our method is good at modeling the one-mode projection of bipartite networks in which tagging the vertex information of the original bipartite graphs is crucial to controlling the structure of produced networks freely, (ii) our one-mode collaboration network model matches the mechanisms of real networks more completely and accurately than other models, and (iii) size growth, individual aging, random collaboration, preferential collaboration, transitivity collaboration, and multi-round collaboration are crucial mechanisms of collaboration networks. Certainly, we also note that a few statistical indicators, such as the Pearson correlation coefficient of our model network corresponding to the actor network, do not match those of the real data very well, which indicates that some crucial mechanisms of the collaboration networks might have been partially or completely missed, which we plan to explore in the future.

Acknowledgments

This research is supported by a project sponsored by the Natural Science Basic Research Plan of the Shaanxi Province of China (Program No. 2013JM9001), a project sponsored by the Social Science Foundation of the Shaanxi Province of China (Program No.13Q107), a project sponsored by the Outstanding Young Teacher Training and Funding of School of Management, Northwestern Polytechnical University, China (Program No.W016227), and a project sponsored by the Seed Foundation of Innovation and Creation for Graduate Students of Northwestern Polytechnical University, China.

References

[1] M.E.J. Newman, The structure of scientific collaboration networks, Proceedings of the National Academy of Sciences, 98 (2001) 404-409.
[2] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature, 393 (1998) 440-442.
[3] G.F. Davis, M. Yoo, W.E. Baker, The Small World of the American Corporate Elite, 1982-2001, Strategic Organization, 1 (2003) 301-326.
[4] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Physics Reports, 519 (2012) 1-49.
[5] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, Y. Aberg, The web of human sexual contacts, Nature, 411 (2001) 907-908.
[6] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, The large-scale organization of metabolic networks, Nature, 407 (2000) 651-654.
[7] S. Saavedra, F. Reed-Tsochas, B. Uzzi, A simple model of bipartite cooperation for ecological and organizational networks, Nature, 457 (2009) 463-466.
[8] C.A. Hidalgo, R. Hausmann, The building blocks of economic complexity, Proceedings of the National Academy of Sciences, 106 (2009) 10570-10575.
[9] M.E.J. Newman, Coauthorship networks and patterns of scientific collaboration, Proceedings of the National Academy of Sciences, 101 (2004) 5200-5205.
[10] M.E.J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Physical Review E, 64 (2001) 016132.
[11] M.E.J. Newman, Scientific collaboration networks. I. Network construction and fundamental results, Physical Review E, 64 (2001) 016131.
[12] R. Lambiotte, M. Ausloos, Uncovering collective listening habits and music genres in bipartite networks, Physical Review E, 72 (2005) 066107.
[13] S. Ming-Sheng, L. Linyuan, Z. Yi-Cheng, Z. Tao, Empirical analysis of web-based user-object bipartite networks, EPL (Europhysics Letters), 90 (2010) 48006.
[14] P.-P. Zhang, C. Kan, Y. He, T. Zhou, B.-B. Su, Y. Jin, H. Chang, Y.-P. Zhou, L.-C. Sun, B.-H. Wang, D.-R. He, Model and empirical study on some collaboration networks, Physica A: Statistical Mechanics and its Applications, 360 (2006) 599-616.
[15] P.G. Lind, M.C. González, H.J. Herrmann, Cycles and clustering in bipartite networks, Physical review E, 72 (2005) 056127.
[16] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, Y. Fan, Clustering coefficient and community structure of bipartite networks, Physica A: Statistical Mechanics and its Applications, 387 (2008) 6869-6875.
[17] G. Ergün, Human sexual contact network as a bipartite graph, Physica A: Statistical Mechanics and its Applications, 308 (2002) 483-488.
[18] J.J. Ramasco, S.N. Dorogovtsev, R. Pastor-Satorras, Self-organization of collaboration networks, Physical Review E, 70 (2004) 036106.
[19] K. Takemoto, M. Arita, Nested structure acquired through simple evolutionary process, Journal of Theoretical Biology, 264 (2010) 782-786.
[20] L. Tian, Y. He, H. Liu, R. Du, A general evolving model for growing bipartite networks, Physics Letters A, 376 (2012) 1827-1832.
[21] C.-X. Zhang, Z.-K. Zhang, C. Liu, An evolving model of online bipartite networks, Physica A: Statistical Mechanics and its Applications, 392 (2013) 6100-6106.
[22] R. Guimerà, B. Uzzi, J. Spiro, L.A.N. Amaral, Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance, Science, 308 (2005) 697-702.
[23] C. Berge, Graphs and Hypergraphs, Elsevier, New York, 1973.
[24] J.-W. Wang, L.-L. Rong, Q.-H. Deng, J.-Y. Zhang, Evolving hypernetwork model, The European Physical Journal B, 77 (2010) 493-498.
[25] G.-Y. Yang, J.-G. Liu, A local-world evolving hypernetwork model, Chinese Phys B, 23 (2014) 018901.
[26] A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations,

Physica A: Statistical Mechanics and its Applications, 311 (2002) 590-614.

[27] T. Zhou, B.-H. Wang, Y.-D. Jin, D.-R. He, P.-P. Zhang, Y. He, B.-B. Su, K. Chen, Z.-Z. Zhang, J.-G. Liu, Modelling collaboration networks based on nonlinear preferential attachment, International Journal of Modern Physics C, 18 (2007) 297-314.

[28] M.E.J. Newman, D.J. Watts, S.H. Strogatz, Random graph models of social networks, Proceedings of the National Academy of Sciences of the United States of America, 99 (2002) 2566-2572.

[29] R. Guimerà, M. Sales-Pardo, L.A.N. Amaral, Module identification in bipartite and directed networks, Physical Review E, 76 (2007) 036102.

[30] J.J. Ramasco, S.A. Morris, Social inertia in collaboration networks, Physical Review E, 73 (2006) 016122.

[31] Q. Ke, Y.-Y. Ahn, Tie Strength Distribution in Scientific Collaboration Networks, arXiv preprint arXiv:1401.5027, (2014).

[32] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation, Physical Review E, 76 (2007) 046115.

[33] A. Barrat, M. Barthélemy, A. Vespignani, Weighted Evolving Networks: Coupling Topology and Weight Dynamics, Phys. Rev. Lett., 92 (2004) 228701.

[34] M.E.J. Newman, Assortative Mixing in Networks, Phys. Rev. Lett., 89 (2002) 208701.

[35] A.-L. Barabási, R. Albert, Emergence of Scaling in Random Networks, Science, 286 (1999) 509-512.