Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 567 – 576

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Characterizing the Structure of Affiliation Networks

Dajie Liu[a,*], Norbert Blenn[a], Piet Van Mieghem[a]

[a] *Faculty of Electrical Engineering, Mathematics and Computer Science*
Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands

Abstract

Our society contains all types of organizations, such as companies, research groups and hobby clubs. Affiliation networks, as a large and important portion of social networks, consist of individuals and their affiliation relations: Two individuals are connected by a link if they belong to the same organization(s). Affiliation networks naturally contain many fully connected cliques, since the nodes of the same organization are all connected with each other by definition. In this paper, we present methods which facilitate the computation for characterizing the real-world affiliation networks of ArXiv coauthorship, IMDB actors collaboration and SourceForge collaboration. We propose a growing hypergraph model with preferential attachment for affiliation networks which reproduces the clique structure of affiliation networks. By comparing computational results of our model with measurements of the real-world affiliation networks of ArXiv coauthorship, IMDB actors collaboration and SourceForge collaboration, we show that our model captures the fundamental properties including the power-law distributions of group size, group degree, overlapping depth, individual degree and interest-sharing number of real-world affiliation networks, and reproduces the properties of high clustering, assortative mixing and short average path length of real-world affiliation networks.

Keywords: affiliation network, hypergraph, line graph, eigvalue, power law

1. Introduction

Social networks, as one type of real-world complex networks, are widely studied [1, 2, 3, 4]. Social networks are defined as networks where nodes are individuals and links are relations between individuals, reflecting acquaintances, friendships, sexual relations, collaboration, common affiliation, etc. Apart from many common properties of the real-world networks, such as a high clustering coefficient, a short characteristic path length and a power-law degree distribution, social networks also exhibit assortative mixing and community structure [4, 5, 6, 7].

Affiliation networks, an important type and a large portion of social networks, have not yet been well understood and modeled. The nodes of affiliation networks represent individuals. If two nodes in an affiliation network have the same membership, for instance, they belong to the same institute or they work for the same project, then they are connected by a link. Examples of affiliation networks include movie actor networks (nodes represent the actors and two actors have a link if they have been casted together in one or more movies), science coauthorship networks (nodes represent the scientists and two nodes are connected if they have coauthored one or more articles), journal editor

*Corresponding author
Email address: d.liu@tudelft.nl (Dajie Liu)

doi:10.1016/j.procs.2012.04.061

networks (nodes as the editors and two editors are adjacent if they serve on the same editorial boards of journals) and sports player networks (nodes as players and two players who played in the same games are connected)[1].

Affiliation networks naturally contain many fully connected subnetworks which are called cliques or complete subgraphs in the language of graph theory, since the nodes of the same group, such as a movie cast, are all connected with each other by definition. The clique structure of affiliation networks increases largely the percentage of triangles among the three hops walks, consequently resulting in high clustering coefficient. The affiliation networks also show high assortativity. Besides the statistics of individuals such as clustering coefficient, characteristic path length and nodal degree, we are also interested to answer the following questions: How many groups are there? How many individuals does a group have? How many groups does an individual belong to? How many individuals do two groups have in common? how many groups do two individuals belong to together (It is be useful for recommandation systems)? And to how many groups is a group adjacent (Two groups are adjacent if they have individuals in common)? In Section 4, we attempt to gain more information on the structure properties of affiliation networks by answering these questions in the cases of the ArXiv coauthorship networks of subjects of "General Relativity and Quantum Cosmology" and "High Energy Physics - Theory", the IMDB movie actors collaboration network and the SourceForge software collaboration network. In Section 3, we introduce the analytical properties on the topology and spectra of the affiliation networks. In Section 5, we propose a preferential attachment based growing hypergraph model for affiliation networks. The nodes of the hypergraph model represent the groups of affiliation networks, and the hyperedges, connecting multiple nodes, represent the individuals. Numerical analyses show that our hypergraph model reproduces all the properties of affiliation networks.

2. The representation of affiliation networks

2.1. Preliminaries

Suppose the affiliation network under consideration has N individuals and M groups, where an individual may belong to multiple groups. The membership number m_j of an individual j is defined by the number of groups of which j is a member. The degree d_j of an individual j equals the number of individuals who have the same membership in one or more groups. The interest-sharing number $\alpha_{i,j}$ of individuals i and j is defined by the number of groups to which they both belong, which indicates how many common interests they share. The group size s_k of group k is the number of individuals that belong to group k. The group degree u_k of group k equals the number of groups sharing individual(s) with group k. The overlapping depth $\beta_{k,l}$ of two groups k and l equals the number of individuals that they share. An affiliation network is linear if $\beta_{k,l} \leq 1$ for all $k, l \in [1, M]$, where M is the number of groups. If the membership number $m_j = m$ for $j \in [1, N]$, the affiliation network is called a m-uniform affiliation network.

We use the graphs in Figure 1a to exemplify the definitions of d_j, m_j, $\alpha_{i,j}$, s_k, u_k, and $\beta_{k,l}$. The graph in Figure 1a (1) has labeled five nodes which are members of at least two groups. Obviously, $d_1 = 24$, $d_2 = 12$, $d_3 = d_4 = 8$ and $d_5 = 9$. Nodes $1 - 5$ belong to 5, 3, 2, 2 and 2 groups respectively, thus $m_1 = 5$, $m_2 = 3$ and $m_3 = m_4 = m_5 = 2$. Individual 1 and 2 belong to only one common group, hence $\alpha_{i,j} = 1$. As shown in Figure 1a (2), the groups $I - IV$ have 6, 5, 5 and 6 nodes respectively, hence, $s_I = s_{IV} = 6$ and $s_{II} = s_{III} = 5$. Evidently, the overlapping widths: $\beta_{I,II} = 2, \beta_{I,III} = 1, \beta_{I,IV} = 3, \beta_{II,III} = 2, \beta_{II,IV} = 0$ and $\beta_{III,IV} = 1$. The group degree: $u_I = u_{III} = 3$, $u_{II} = u_{IV} = 2$.

An affiliation network is usually described by a graph where the nodes represent the individuals and two nodes are connected by a link if they both belong to a group or several groups. If a set C_I of individuals belong to group I, the set C_I of individuals comprise a fully connected clique. If a set C_{II} ($C_{II} \subseteq C_I$) of individuals also belong to another group II, we cannot represent the group II by this graph description, because the set C_{II} of individuals are already fully connected inside the group I. Newman et al. [8] suggested a bipartite graph model with all information preserved by representing a group with one type of nodes and individuals with the other type of nodes, where links only connect nodes of different types, as shown in Figure 1b. Lattanzi et al. [9] proposed a bipartite-graph-based generative model for affiliation networks. However, the bipartite-graph-based model does not reproduce all the affiliation networks' topological properties shown in Section 3. Hence, we introduce the hypergraph representation of affiliation networks.

[1]Some biological networks can also be classified as affiliation networks including protein interaction networks with proteins as nodes, of which two are connected by a link if they involve in the same functional category or more. In this paper, we focus on affiliation networks of social networks.

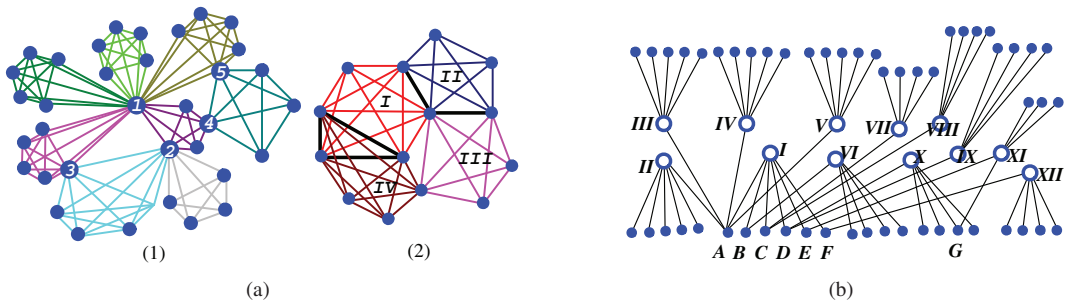


Figure 1: (a) The example affiliation networks for illustration of the definitions of d_j, m_j, $\alpha_{i,j}$, s_k, u_k, and $\beta_{k,l}$. The nodes denote individuals. The groups consist of links of the same color and the shared thick black link(s), and the nodes incident to the links of both colors. (b) The bipartite graph representation of the affiliation network of the NAS group. There are two types of nodes: the blue circles denoting the groups and the solid blue disks denoting the individuals. If an individual belongs to a group, the corresponding two nodes are connected by a link. The corresponding hypergraph is shown in Figure 2(a).

Table 1: The groups and the group members of the exemplary affiliation network of NAS.

Index	Names of groups	Individuals	Index	Names of groups	Individuals
I	NAS-TU Delft	A, B, C, D, E, F	VII	A research group-KPN	$C, C_1, \cdots, C_4$
II	A research group-MIT	$A, A_1, \cdots, A_5$	$VIII$	A piano club	$C, C_5, \cdots, C_8$
III	A research group-Cornell Univ.	$A, A_6, \cdots, A_{10}$	IX	A research group-TNO	$D, D_1, \cdots, D_4$
IV	IEEE/ACM ToN editorial board	$A, A_{11}, \cdots, A_{15}$	X	A rock band	D, D_5, D_6, D_7
V	A research group-KSU	$A, A_{16}, \cdots, A_{20}$	XI	A soccer team	$E, E_1, \cdots, E_4$
VI	A research group-Ericsson	$B, B_1, \cdots, B_4$	XII	Bioinformatics-TU Delft	$F, F_1, \cdots, F_4$

2.2. Hypergraph representation

A hypergraph is the generalization of a simple graph[2]. A hypergraph $H(M, N)$ has M nodes and N hyperedges[3]. Its nodes are of the same type as those of a simple graph, as shown in Figure 2 (a). The hyperedges of hypergraphs can connect multiple nodes, like hyperedge A in Figure 2 (a) connecting nodes $I, II, \cdots, V$. A hypergraph is linear if each pair of hyperedges intersects in at most one node. Hypergraphs where all hyperedges connect the same number m of nodes are defined as m-uniform hypergraphs with the special case that 2-uniform hypergraphs are simple graphs. If an affiliation network is linear, the representing hypergraph is linear; if an affiliation network is m-uniform, the representing hypergraph is also m-uniform.

We propose to describe an affiliation network with M groups and N individuals by a hypergraph $H(M, N)$: M nodes represent the M groups; N hyperedges represent N individuals; and an hyperedge is incident to a node if the corresponding individual is a member of the corresponding group.

The line graph of a hypergraph $H(M, N)$ is defined as the graph $l(H)$, of which the node set is the set of the hyperedges of $H(M, N)$ and two nodes are connected by a link of weight t, when the corresponding hyperedges share t node(s). The degree d_j of an individual j, defined in subsection 2.1, equals the number of individuals that connect to j in the line graph $l(H)$. The line graph $l(H)$ is an unweighted graph when the corresponding hypergraph is linear; otherwise is weighted, and the weight of link $i \sim j$ equals the interest-sharing number $\alpha_{i,j}$.

2.3. An illustrative example

In this subsection, we give an exemplary affiliation network and then represent it by a hypergraph. Table 1 describes an affiliation network based on the affiliations of members of the NAS research group (Network Architectures and Services Group at Delft University of Technology). Individuals A, B, C, D, E, F are members of NAS and the

[2] A simple graph is an unweighted, undirected graph containing no self-loops nor multiple links between the same pair of nodes

[3] We use the term "hyperedge" instead of "hyperlinks" in order not to make confusion with hyperlinks of WWW webs.

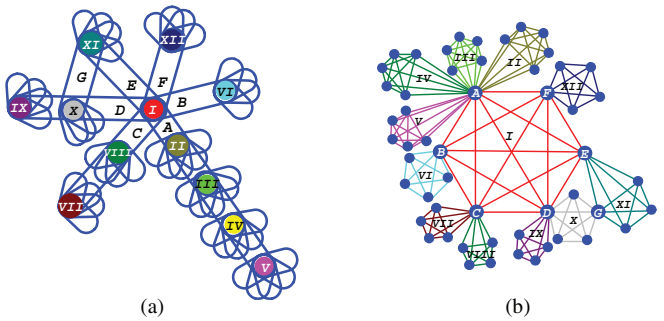


Figure 2: (a) The hypergraph representation of the NAS affiliation network described in Table 1. The hyperedges are the blue ellipse-like closed curves. The nodes are the disks with different colors marked with Roman numerals. A node and a hyperedge are incident if the node is surrounded by the hyperedge. The hyperedges and nodes represent the individuals and the groups respectively. Individuals participate in multiple groups, implying that the groups overlap with each other. (b) The line graph of the hypergraph in (a), which is a simple graph. The nodes here denote the individuals while the communities consist of links of the same color and the nodes which are incident to them. Note that this graph is also the line graph of the hypergraph.

other individuals are the members of groups which overlap with the NAS group. Figure 1b depicts the bipartite graph representation of the NAS affiliation network with the blue circles representing the groups and the solid blue disks representing the individuals. Two nodes are linked when the corresponding individual belongs to the corresponding group.

We represent this network by the hypergraph $H(12, 53)$ shown in Figure 2 (a). The nodes of the hypergraph denote the groups and the individuals are denoted by the hyperedges. There are 12 groups as described in Table 1, corresponding to the 12 nodes of the hypergraph in Figure 2 (a), and there are 53 individuals among whom 6 NAS members with the membership number $m_A = 5$, $m_C = m_D = 3$, $m_B = m_E = m_F = 2$. If an individual belongs to multiple groups, the corresponding nodes are connected by the hyperedge specifying that individual.

Figure 2 (b) depicts the line graph $l(H)$ of the hypergraph $H(12, 53)$ in Figure 2 (a), which represents the exemplary NAS affiliation network. In the line graph $l(H)$, the individuals are denoted by nodes and the groups are denoted by links of the same color and the nodes which are incident to those links. The line graph $l(H)$ is unweighted since the NAS affiliation network is linear.

3. Properties of affiliation networks

3.1. Topological properties

The line graph $l(H)$ has N nodes and L links. The topology of $l(H)$ can be described by its adjacency matrix A, a $N \times N$ matrix, where the element a_{ij} equals the linkweight of link $i \sim j$ if there is a link between node i and node j, else $a_{ij} = 0$. Since $l(H)$ is undirected, the adjacency matrix A is symmetric.

The following equalities are valid for all affiliation networks,

$$N = \sum_{k=1}^{M} s_k - \sum_{k,l=1}^{M} \beta_{k,l}$$

$$L = \frac{1}{2}\sum_{j=1}^{N} d_j = \sum_{k=1}^{M} \frac{s_k\,(s_k - 1)}{2} - \sum_{k,l=1}^{M} \frac{\beta_{k,l}\,(\beta_{k,l} - 1)}{2}$$

$$\sum_{j=1}^{N} \left(m_j - 1\right) = \sum_{k,l=1}^{M} \beta_{k,l}$$

If $\beta_{k,l} \leq 1$ for all $k, l \in [1, M]$, where M is the number of groups, which implies that the affiliation networks are linear, we have,

$$d_j = \sum_{\substack{\text{All the groups to} \\ \text{which individual } j \text{ belongs}}} (s - 1)$$

where s is the group size; And

$$u_k = \sum_{\substack{\text{All the individuals} \\ \text{that group } k \text{ contains}}} (m-1)$$

where m is the membership number of an individual. When the affiliation network is linear, we also have $\alpha_{i,j} \leq 1$.

The adjacency matrix $A^{l(H)}_{N\times N}$ of the line graph $l(H)$ of a hypergraph $H(M,N)$ which represents an affiliation network with M groups and N individuals, can be expressed by the unsigned incidence matrices $R_{M\times N}$ of $H(M,N)$

$$A^{l(H)}_{N\times N} = \left(R^T R\right)_{N\times N} - \mathrm{diag}(R^T R) \tag{1}$$

where the entry r_{ij} of R is 1 if node i and hyperedge j are incident, otherwise $r_{ij} = 0$. Basically, the adjacency matrix $A^{l(H)}$ equals the matrix $R^T R$ setting all the diagonal entries to zero. The interest-sharing number $\alpha_{i,j}$ of individual i and j equals the entry $a^{l(H)}_{ij}$of $A^{l(H)}$

$$\alpha_{i,j} = a^{l(H)}_{ij} \tag{2}$$

The membership number m_j of an individual j equals,

$$m_j = \sum_{i=1}^{M} r_{ij} = (R^T R)_{jj} \tag{3}$$

The group size s_k of group k is

$$s_k = \sum_{l=1}^{N} r_{kl} = (RR^T)_{kk} \tag{4}$$

Let $W_{M\times M} = \left(RR^T\right)_{M\times M} - \mathrm{diag}(RR^T)$, then the overlapping depth $\beta_{k,l}$ of two groups k and l equals,

$$\beta_{k,l} = w_{kl} \tag{5}$$

where w_{kl} is an entry of $W_{M\times M}$.

The individual degree d_j equals the number of nonzero entries in the jth row/column of $A^{l(H)}_{N\times N}$, with the special case $d_j = \sum_{i=1}^{N} a^{l(H)}_{ij}$ when the affiliation network is linear. Similarly, the group degree u_k equals the number of nonzero entries in the kth row/column of $W_{M\times M}$.

3.2. Spectral properties

3.2.1. The adjacency spectra of $l(H)$ of m-uniform affiliation networks

A m-uniform affiliation network can be represented by m-uniform hypergraphs $H_m(M,N)$, of which the unsigned incidence matrix R has exactly *m one-entries* and *M − m zero-entries* in each column. Thus, all the diagonal entries of $R^T R$ are m. The adjacency matrix of the line graph of $H_m(M,N)$ can be written as,

$$A^{l(H_m)}_{N\times N} = R^T R - mI \tag{6}$$

where $R^T R$ is a Gram matrix [10][11].

Lemma 1. *For all matrices $A_{N\times M}$ and $B_{M\times N}$ with $N \geq M$, it holds that $\lambda(AB) = \lambda(BA)$ and $\lambda(AB)$ has $N - M$ extra zero eigenvalues*

$$\lambda^{N-M} \det(BA - \lambda I) = \det(AB - \lambda I)$$

Lemma 1 and (6) yields,

$$\det\left(A^{l(H_m)}_{N\times N} - (\lambda - m)I\right) = \lambda^{N-M} \det\left(\left(RR^T\right)_{M\times M} - \lambda I\right)$$

The adjacency matrix $A^{l(H_m)}_{N\times N}$ has at least $N - M$ eigenvalues of $-m$. We have

$$x^T \left(R^T R\right) x = (Rx)^T Rx = \|Rx\|_2^2 \geq 0$$

and

$$x^T \left(RR^T\right) x = \left(R^T x\right)^T R^T x = \left\|R^T x\right\|_2^2 \geq 0$$

where $x_{L\times 1}$ is an arbitrary vector. Hence, both $\left(R^T R\right)_{N\times N}$ and $\left(RR^T\right)_{M\times M}$ are positive semidefinite, hence all eigenvalues of $\left(R^T R\right)_{N\times N}$ are non-negative. Due to (6), the adjacency eigenvalues of $A^{l(H_m)}_{N\times N}$ are not smaller than $-m$.

3.2.2. The adjacency spectra of $l(H)$ of non-uniform affiliation networks

A non-uniform affiliation network with maximum membership number $m_{\max}$ can be represented by a non-uniform hypergraph $H(M, N)$. The unsigned incidence matrix R of $H(M, N)$ has at most $m_{\max}$ *one-entries* in each column. Therefore, the largest diagonal entry of $R^T R$ is $m_{\max}$. The adjacency matrix of the line graph of non-uniform hypergraph $H(M, N)$ is,

$$A^{l(H)}_{N\times N} = R^T R + C - m_{\max} I \tag{7}$$

where $C = \text{diag}\left(\begin{array}{cccc} c_{11} & c_{22} & \cdots & c_{LL} \end{array}\right)$ and $c_{jj} = m_{\max} - (R^T R)_{jj} \geq 0$ for $j \in [1, N]$.

Since

$$\begin{aligned} x^T \left(R^T R + C\right) x &= x^T \left(R^T R\right) x + x^T \left(\sqrt{C}^T \sqrt{C}\right) x \\ &= \|Rx\|_2^2 + \left\|\sqrt{C}x\right\|_2^2 \geq 0 \end{aligned}$$

where $x_{L\times 1}$ is an arbitrary vector and $\sqrt{C} = \text{diag}\left(\begin{array}{cccc} \sqrt{c_{11}} & \sqrt{c_{22}} & \cdots & \sqrt{c_{LL}} \end{array}\right)$, $R^T R + C$ is also positive semidefinite, thus, the adjacency eigenvalues of $A^{l(H_m)}_{N\times N}$ are not smaller than $-m_{\max}$.

4. Real-world affiliation networks

4.1. ArXiv coauthorship network

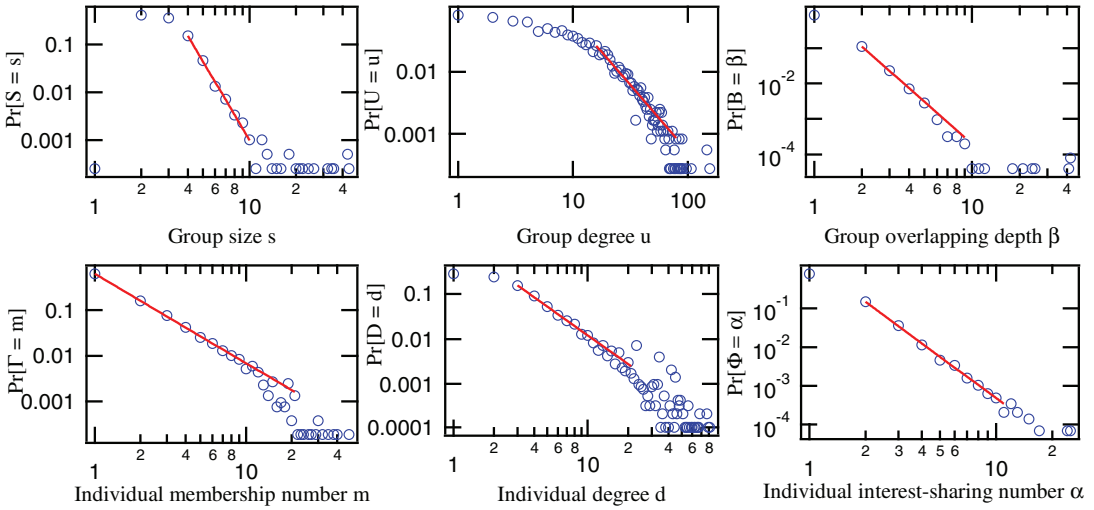
We analyze the arXiv data of subjects of "General Relativity and Quantum Cosmology" (GR-QC) and "High Energy Physics - Theory" (HEP-TH) in the period from January 1993 to April 2003, which were collected by J. Leskovec et al. [12]. We construct the hypergraph with the papers as nodes and the authors as hyperedges. A hyperedge is incident to a node if the corresponding author authors/coauthors the corresponding paper. In this manner we construct the hypergraph of the arxiv GR-QC coauthorship network with 5855 authors and 13454 papers, and the hypergraph of the arXiv HEP-TH coauthorship network with 9877 authors and 21568 papers. We fit the data of s, β, m, d and α with the power function $f(x) = x^{-\gamma}$. The values of γ are shown in Table 2. The group size s follows a power-law distribution. In this case of coauthorship network, the group size s means the number of authors a paper has. As shown in Figure 3a and 3b, We see that, in the coauthorship networks of both subjects, the papers with only one author and with more than ten authors are very rare. Most of papers have two or three authors. The group degree u in both Figure 3a and 3b has a power-law tail. The group overlapping depth β follows a power-law distribution[4]. The membership membership m of an individual here means the number of papers he or she authors and coauthors, and also follows a power-law distribution. The interesting sharing number α, denoting the number of papers in which two individuals participate together, follows a power-law distribution[5]. The ArXiv coauthorship networks of both subjects possess high clustering coefficient, large assortativity coefficient and short average path length as shown in Table 2.

4.1.1. IMDB movie actors collaboration network

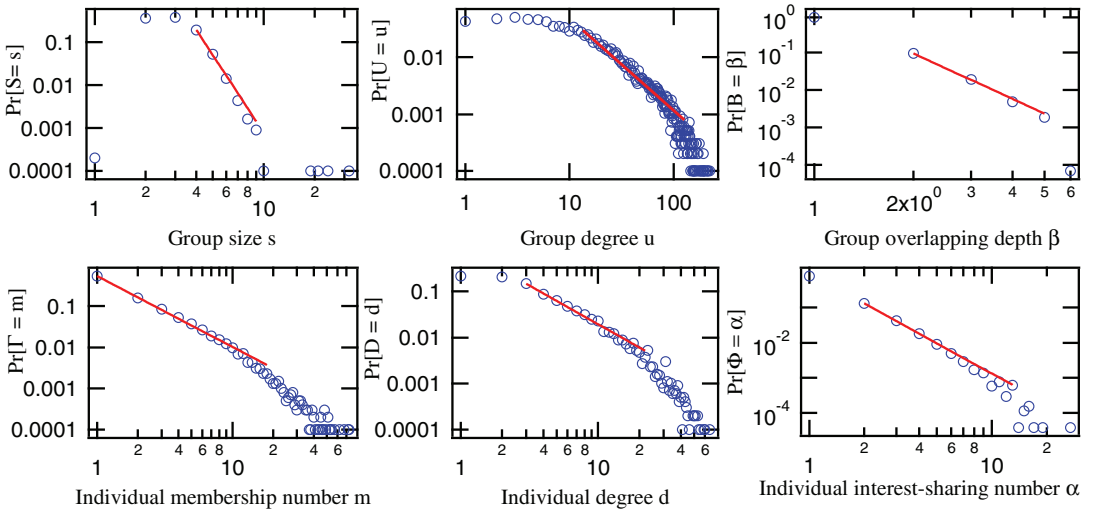
The data of IMDB movie actors collaboration network with 127823 movies and 392340 actors, were collected by Hawoong Heong from Internet Movie Database (based on www.imdb.com). We construct the hypergraph of IMDB

[4]Most of the pairs of groups have no overlap. We only consider the group pairs which overlaps with each other.

[5]We only consider the individual pairs who have nonzero interest-sharing number.



(a)



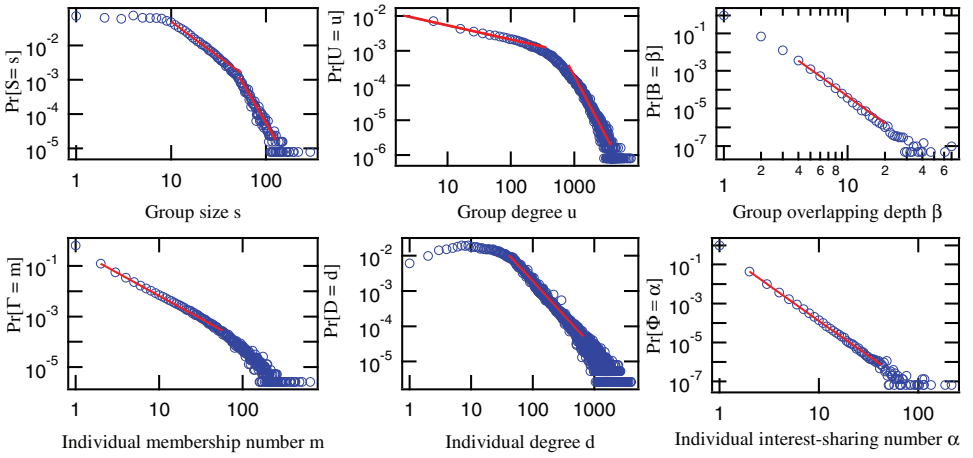
(b)

Figure 3: The probability density distribution of group size s, group degree u, group overlapping depth β (the first row from left to right), individual membership number m, individual degree d, individual interest-sharing number α (the second row from left to right) of the ArXiv coauthorship networks of (a) "General Relativity and Quantum Cosmology" category and (b) "High Energy Physics - Theory" category. They all follow power-law distributions.

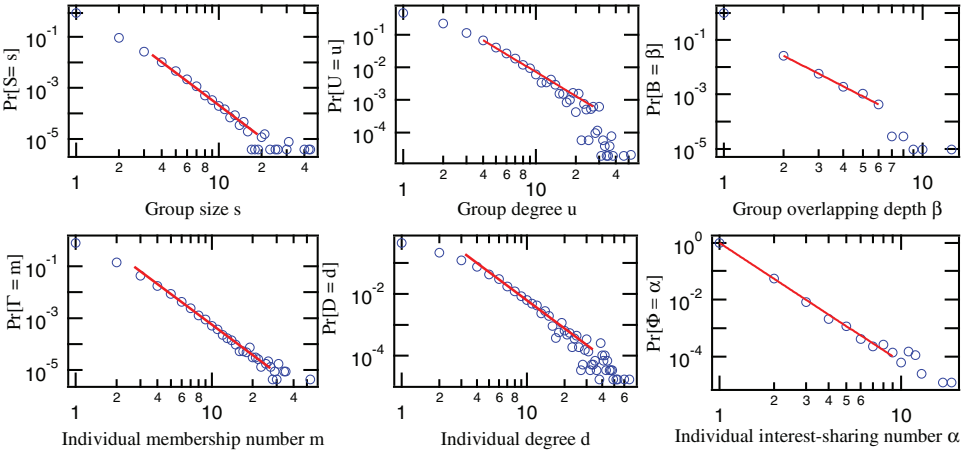
Table 2: The clustering coefficients C, the assortativity coefficients ρ_D, the average path lengths l, and the exponents γ of power-law fittings of s, u, β, m, d and α of the arXiv GR-QC and HEP-TH coauthorship networks, the IMDB actor collaboration network, the SourceForge software collaboration network and the proposed growing hypergraph model.

Network	$\gamma(s)$	$\gamma(u)$	$\gamma(\beta)$	$\gamma(m)$	$\gamma(d)$	$\gamma(\alpha)$	C	ρ_D	l
ArXiv GRQC coauthorship	5.50	2.14	3.93	1.95	1.84	3.56	0.637	0.584	6.50
ArXiv HEP-TH coauthorship	6.24	1.63	3.56	1.72	1.68	2.86	0.289	0.382	4.89
IMDB actors collaboration	2.04/5.35	0.407/3.40	4.80	1.81	1.91	3.62	0.762	0.682	4.29
SourceForge collaboration	3.91	2.45	3.76	3.48	2.61	4.60	0.636	0.401	7.06
Growing hypergraph model	2.55	1.17	6.02	2.02	1.27	3.15	0.54	0.71	4.98

moive actors collaboration network with the movies as nodes and the actors as hyperedges. A hyperedge is incident to a node if the corresponding actor appears in the corresponding movie. We fit the data of s, u, β, m, d and α with the power function $f(x) = x^{-\gamma}$, as shown in Figure 4a and Table 2. The data of s are fitted with two power functions in different regions. The group degree u appears also to follow two power-law distribution in two regions. All the values of γ are shown in Table 2. The IMDB movie actors collaboration network exhibits high clustering, assortative mixing and short average path length.



(a)



(b)

Figure 4: The probability density distribution of group size s, group degree u, group overlapping depth β (the first row from left to right), individual membership number m, individual degree d, individual interest-sharing number α (the second row from left to right) of (a) the IMDB movie actors collaboration network and (b) the SourceForge software collaboration network.

4.1.2. The SourceForge software collaboration network

SourceForge is a web-based project repository assisting developers to develop and distribute open source software projects. SourceForge facilitates developers by providing a centralized storage and tools to manage the projects. Each project has multiple developers. We construct the hypergraph of the SourceForge software collaboration network by taking software projects as nodes and the developers as hyperedges. A hyperedge is incident to a node if the corresponding developer participates in the corresponding software project. The SourceForge software collaboration network has 259252 software projects and 161653 developers. We fit the data of s, u, β, m, d and α with the power

function $f(x) = x^{-\gamma}$. As shown in Figure 4b, the pdfs of all the six metrics d_j, m_j, $\alpha_{i,j}$, s_k, u_k, and $\beta_{k,l}$ are well fitted by power law functions with exponents γ shown in Table 2. The SourceForge network also has a high clustering coefficient, a high assortativity coefficient and an small average path length, which are shown in Table 2.

5. Modeling of affiliation networks

5.1. Model description

As stated before, we use the nodes of hypergraph to represent the groups and the hyperedges to represent the individuals. In the description of our model, the nodes and groups, the hyperedges and individuals are used interchangeably. Our model is a growing hypergraph model, starting with a small hypergraph which represent the initial groups and individuals. Later on, new individuals and new groups are added to the network in the growing process.

We notice that the number of group M is larger the number of individuals N in ArXiv networks and Sourceforge network, and M is smaller than N in IMDB network. Making a movie needs more efforts and labor force than writing a paper or developing an open-source software. In our model, we take $\frac{M}{N} = 1$, assuming that each coming individual start a new group. Note that the group size of real-world affiliation network follow a power-law distribution. We employ preferential attachment of individual to the existing groups to achieve the power-law distributed group size. The tricky issue is to determine the membership number of each new coming individuals, namely to decide how many nodes that a new hyperedge should connect to. The analysis of real-world affiliation networks tells a power-law distribution of the membership number, hence we preproduce a power-law distributed sequence of numbers, taking them as the membership numbers of new coming individuals.

Our hypergraph model is described by the following procedure:

1. Start with a seed hypergraph $H_0(M_0, N_0)$ with M_0 groups and N_0 hyperedges.
2. Suppose that the desired number of individuals (hyperedges) of the network to be generated is $N+N_0$. Determine the membership numbers for the N new hyperedges: $\Gamma = \begin{bmatrix} m_1 & m_2 & \cdots & m_L \end{bmatrix}$. Note that the membership number vector Γ is the input parameter of our hypergraph model.
3. At growing step j, $j = 1, 2, \cdots, L$, add a new hyperedge j and a new group to the hypergraph. Make the new hyperedge j and the new group incident, and the membership number of j becomes 1.
   (a) Connect the new hyperedge j to the existing group i with probability $p_i = s_i / \sum s_i$, where s_i is the group size of i and $\sum s_i$ is the sum of group sizes of all the existing groups.
   (b) Repeat 3a) $m_j - 1$ times so that the membership number of the hyperedge j increases to m_j.
4. Repeat 3) until the number of hyperedges increases to $N + N_0$.

Compute the metrics d_j, m_j, $\alpha_{i,j}$, s_j,u_j and $\beta_{i,j}$ using the methods given in Section 3.1 including the formulas (1) to (5).

5.2. Simulation results

We use a hypergraph $H(20, 20)$ with the membership number $m_j = 1$, $j = 1, 2, \cdots 20$, as the starting seed. We add 5000 new hyperedges (individuals) and 5000 new nodes (groups) to the starting seed through 5000 growing steps. Hence, all the hypergraphs we generate have 5020 nodes and 5020 hyperedges. We generate a sequence of natural numbers following a power-law distribuion with the pdf $\Pr[\Gamma = m] = m^{-2.02}$. In the growing process, we apply this sequence of natural numbers as the membership numbers. We denote the group size and group degree of a random group by S and U, the group overlapping depth of a random pair of groups by B, the individual degree of a random individual by D, and the interest-sharing number of a random pair of hyperedges by Φ.

Due to the principle of preferential attachment, we expect that the group size of all the generated hypergraphs follow power law distributions, which are confirmed by Figure 5. Futhermore, group degree u, individual degree d, group overlapping depth β and individual interest-sharing number α also follow power-law distributions. Our model reproduces the power-law distributions of s, u, d, β and α observed in real-world affiliation networks with similar exponents as shown in Table 2, and also reproduces the properties of high clustering, assortative mixing and short average path length exhibited by real-world affiliation networks.

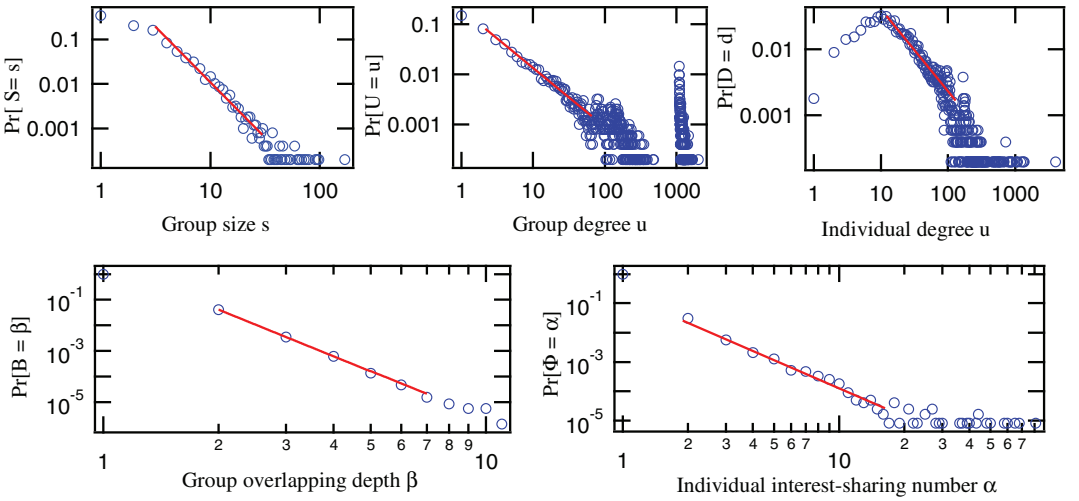


Figure 5: The probability density distribution of group size s, group degree u, individual degree d, group overlapping depth β and individual interest-sharing number α of our growing preferential-attaching hypergraph model. The membership numbers are input parameters, follows a power-law distribution. The hypergraph model reproduces the power-law distributions of s, u, d, β and α observed in real-world affiliation networks.

6. Conclusion

Affiliation networks are an important type of social networks. We propose a hypergraph representation which reproduces the clique structure of affiliation networks. We give analytically the topological and spectral properties of affiliation networks. We also present formulas which facilitate the computation for characterizing the real-world affiliation networks of ArXiv coauthorship, IMDB actors collaboration and SourceForge collaboration. We propose a preferential attachment based growing hypergraph model for affiliation networks. Numerical analyses show that our hypergraph model reproduces the power-law distributions of group size, group degree, overlapping depth, individual degree and interest-sharing number of real-world affiliation networks, and reproduces the properties of high clustering, assortative mixing and short average path length of real-world affiliation networks.

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of modern physics 74 (2002) 47–96.
[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Physics Reports 424 (2006) 175–308.
[3] M. E. J. Newman, D. J. Watts, S. H. Strogatz, Random graph models of social networks, Proc. Natl. Acad. Sci. USA 99 (2002) 2566–2572.
[4] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America 99 (12) (2002) 7821–7826.
[5] Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466 (7307) (2010) 761–764.
[6] M. E. J. Newman, Mixing patterns in networks, Phys. Rev. E 67 (2) (2003) 026126.
[7] P. Van Mieghem, H. Wang, X. Ge, S. Tang, F. A. Kuipers, Influence of assortativity and degree-preserving rewiring on the spectra of networks, The European Physical Journal B - Condensed Matter and Complex Systems (2010) 643–652.
[8] M. E. J. Newman, S. H. Strogatz, D. J. Watts, Random graph with arbitrary degree distribution and their applications, Phys. Rev. E 64 (2001) 026118.
[9] S. Lattanzi, D. Sivakumar, Affiliation networks, in: Proceedings of the 41st annual ACM symposium on Theory of computing, STOC '09, ACM, New York, NY, USA, 2009, pp. 427–434.
[10] P. Van Mieghem, Graph Spectra for Complex Networks, Cambridge University Press (Cambridge, U.K.), 2011.
[11] D. Cvetković, P. Rowlinson, S. K. Simić, Eigenvalue bounds for the signless laplacians, Publ. Inst. Math. (Beograd) 81 (95) (2007) 11–27.
[12] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Transactions on Knowledge Discovery from Data (ACM TKDD) 1 (1).