Characterizing and Modeling Social Networks with Overlapping Communities

Dajie Liu*
Norbert Blenn
Piet Van Mieghem

Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands.
E-mail: {d.liu, n.blenn, p.f.a.vanmieghem}@tudelft.nl
*Corresponding author

Abstract Social networks, as well as many other real-world networks, exhibit overlapping community structure. Affiliation networks, as a large portion of social networks, consist of cooperative individuals: Two individuals are connected by a link if they belong to the same organization(s), such as companies, research groups and hobby clubs. Affiliation networks naturally contain many fully connected communities/groups. In this paper, we characterize the structure of the real-world affiliation networks, and 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: overlapping community structure; growing hypergraph; preferential attachment; line graph; graph spectra; eigenvalue.

Reference to this paper should be made as follows: Liu, D., Blenn, N. and Van Mieghem, P. (xxxx) 'Characterizing and Modeling Social Networks with Overlapping Communities', *Int. J. Web Based Communities*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Dajie Liu received a MSc (cum laude) in optical communications and photonic technologies from Polytechnic University of Turin, Turin, Italy, in 2008; and a MSc in access network management from Southeast University, Nanjing, China, in 2009. Currently, he is an PhD student in the group Network Architectures and Services of the faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands. His research interests include the robustness and optimization of complex

networks, the analysis and modeling of social networks and biological networks.

Norbert Blenn received the Diploma (equivalent to the MSc degree) in computer science in media from the Technical University of Dresden, Dresden, Germany, in 2007. He is pursuing the PhD degree in the group Network Architectures and Services of the faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands. His current research interests include content propagation, security, crawling techniques, and the development of online social networks.

Piet Van Mieghem received the Master's and Ph.D. degrees in electrical engineering from the K.U. Leuven, Leuven, Belgium, in 1997 and 1991, respectively. He is a Professor at the Delft University of Technology, Delft, The Netherlands, with a chair in telecommunication networks and Chairman of the section Network Architectures and Services (NAS) since 1998. His main research interests lie in modeling and analysis of complex networks (such as biological, brain, social, intrastructural, etc. networks) and in new Internet-like architectures and algorithms for future communications networks. Before joining Delft, he worked at the IMEC from 1987 to 1991. During 1993 to 1998, he was a member of the Alcatel Corporate Research Center in Antwerp. He was a visiting scientist at MIT (1992-1993) and a visiting professor at UCLA (2005) and at Cornell University (2009).

1 Introduction

Social networks are currently widely studied (Albert and Barabási, 2002; Boccaletti et al., 2006; Girvan and Newman, 2002). Social networks are often defined as networks where nodes are individuals and links are relations between individuals, reflecting acquaintances, friendships, sexual relations, collaboration, common affiliation, etc. Most social networks possess common properties of the real-world networks, such as high clustering coefficient, short characteristic path length and power law degree distribution (Barabási and Albert, 1999; Watts and Strogatz, 1998). Particularly, they possess some special properties like assortative mixture, community and hierarchical structure (Girvan and Newman, 2002; Ahn et al., 2010; Newman, 2003; Van Mieghem et al., 2010). The communities are the subnets, which exhibit relatively higher levels of internal connections. Community structures feature important topological properties that have catalyzed researches on community detection algorithms and on modularity analysis (Fortunato, 2010; Newman and Girvan, 2004; VanMieghem et al., 2010). The communities overlap with each other when nodes belong to multiple communities. The overlap of different communities exists widely in real-world complex networks, particularly in social and biological networks (Palla et al., 2005; Evans and Lambiotte, 2009; McDaid and Hurley, 2010). Human beings have multiple roles in the society, and these roles make people members of multiple communities at the same time, such as companies, universities, families or relationships, hobby clubs, etc.

In the movie actor network, where nodes are the actors and two actors are connected if they have been casted together in one or more movie, we could regard the set of actors in one movie as a community. According to the definition of movie actor network, the communities of all the movies are cliques. These communities overlap with each other if they have actors in common. The similar networks are the science coauthorship networks (nodes represent the scientists and two nodes are connected if they have coauthored one or more articles and the articles are communities), the journal editor 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).

These types of social networks are known as affiliation networks. The affiliation networks, an important and large type of social networks, are the focus of this article. The communities in affiliation networks are called groups. In the rest of this paper, the terms "community" and "group" will be interchangeably used. 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. The clique structure of social networks increases largely the percentage of triangles among the three hops walks, consequently resulting in high clustering coefficient. Besides the statistics of individuals such as clustering coefficient, characteristic path length and nodal degree, we are also interested to answer the following questions: the number of groups, the number of individuals each group has, the groups each individual belongs to, the number of individuals every pair of groups have in common, the number of groups every pair of individuals join together, and the number of groups each group is adjacent to (two groups are adjacent if they have individuals in common).

1.1 Related works

Palla et al. (2005) defined four metrics to describe how the communities of networks overlap with each other: the membership number of an individual, the overlapping depth of two communities, the community degree and the community size. Palla et al. (2005) showed that the communities of real-world networks overlap with each other significantly. They reported that the membership number of an individual and the overlapping depth of two communities and the community size all follow a power law distribution, except that the community degree features a peculiar distribution that consists of two distinct parts: an exponential distribution in the beginning and a power law tail. Poller et al. (2006) proposed a toy model of which both the community size and the community degree follow a power law distribution, by applying preferential attachment to community growth. There have been many efforts devoted to the modeling of social networks (Newman et al., 2002; Skyrms and Pemantle, 2000; Toivonen et al., 2006). The growing networking model proposed by Toivonen et al. (2006) succeeds in reproducing the common characteristics of social networks: community structure, high clustering coefficient and positive assortativity. The degree distribution of this model is somewhat deviating from a power law distribution despite being heavy-tailed.

1.2 Our contributions

We propose a complete set of metrics which can fully characterize the overlapping community structure of networks. We represent social networks by hypergraphs. The hypergraph representation of networks facilitates the computations of the characterizing metrics. We establish a hypergraph-based social network model which exhibits innate tunable overlapping community structure. By comparing simulation results of our model with results of real-world networks, we show that our hypergraph model exhibits the common properties of large social networks: the community (group) size, the community (group) degree and the community (group) overlapping depth all follow a power law distribution, and our model possesses high clustering coefficient, positive assortativity, short average path length. By tuning the input individual membership number to follow a power law distribution, the individual degree and the interest-sharing number also follow a power law distribution.

This paper is organized as follows: Section 2 introduces the hypergraph representation of affiliation networks. In Section 3, we present the analytical properties on the topology and spectra of the social networks. In Section 4, we characterize the overlapping community structure of social networks 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 5, we propose a preferential attachment based growing hypergraph model for social networks. The nodes of the hypergraph model represent the groups of social networks, and the hyperedges, connecting multiple nodes, represent the individuals. Numerical analyses show that our hypergraph model reproduces all the properties of social networks.

2 The representation of social networks with overlapping communities

2.1 Preliminaries

Suppose the 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 1 to exemplify the definitions of d_j, m_j, $\alpha_{i,j}$, s_k, u_k, and $\beta_{k,l}$. The graph in Figure 1 (a) has labeled five nodes which are members of at least two groups. Obviously, $d_1 = 24$, $d_2 = 12$, $d_3 = 10$, $d_4 = 8$ 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 1 (b), 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. Scott (1991) discussed generating affiliation network with simple graphs. Newman et al. (2001) 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 2. Lattanzi and Sivakumar (2009) 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.

2.2 Hypergraph representation

A hypergraph is the generalization of a simple graph. A simple graph is an unweighted, undirected graph containing no self-loops nor multiple links between the same pair of nodes. A hypergraph $H(M, N)$ has M nodes and N hyperedges. We use the term "hyperedge" instead of "hyperlinks" in order not to make confusion with hyperlinks of WWW webs. Its nodes are of the same type as those of a simple graph, as shown in Figure 3 (a). The hyperedges of hypergraphs can connect multiple nodes, like hyperedge A in Figure 3 (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

Table 1 The names and the members of all the communities of the exemplary social network of NAS.

Index	Names of communities	Members (individuals)
I	NAS-TU Delft	A, B, C, D, E, F
II	A research group-MIT	$A, A_1, \cdots, A_5$
III	A research group-Cornell Univ.	$A, A_6, \cdots, A_{10}$
IV	IEEE/ACM ToN editorial board	$A, A_{11}, \cdots, A_{15}$
V	A research group-KSU	$A, A_{16}, \cdots, A_{20}$
VI	A research group-Ericsson	$B, B_1, \cdots, B_4$
VII	A research group-KPN	$C, C_1, \cdots, C_4$
$VIII$	Piano club	$C, C_5, \cdots, C_8$
IX	A research group-TNO	$D, D_1, \cdots, D_4$
X	A rock band	D, D_5, D_6, D_7, G
XI	A soccer team	E, E_1, E_2, E_3, G
XII	Bioinformatics-TU Delft	$F, F_1, \cdots, F_4$

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 other individuals are the members of groups which overlap with the NAS group. Figure 2 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 3 (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 3 (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 3 (b) depicts the line graph $l(H)$ of the hypergraph $H(12, 53)$ in Figure 3 (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 social networks with overlapping communities

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,

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

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

$$\sum_{j=1}^{N} (m_j - 1) = \sum_{k=1,l=1}^{M} \beta_{k,l} \tag{3}$$

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,

$$d_j = \sum_{\substack{\text{All the groups to} \\ \text{which individual } j \text{ belongs}}} (s-1) \tag{4}$$

where s is the group size; And

$$u_k = \sum_{\substack{\text{All the individuals} \\ \text{that group } k \text{ contains}}} (m-1) \tag{5}$$

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 $H(M,N)$

$$A^{l(H)}_{N\times N} = \left(R^T R\right)_{N\times N} - \mathrm{diag}(R^T R) \tag{6}$$

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 $A^{l(H)}$

$$\alpha_{i,j} = a^{l(H)}_{ij} \tag{7}$$

The membership number m_j of an individual j equals,

$$m_j = \sum_{i=1}^{M} r_{ij} = (R^T R)_{jj} \tag{8}$$

The group size s_k of group k is

$$s_k = \sum_{l=1}^{N} r_{kl} = (RR^T)_{kk} \tag{9}$$

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,

$$\beta_{k,l} = w_{kl} \tag{10}$$

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 the line graph 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,

$$A^{l(H_m)}_{N\times N} = R^T R - mI \tag{11}$$

where $R^T R$ is a Gram matrix (Van Mieghem, 2011)(Cvetković et al., 2007).

Lemma 3.1: *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 3.1 and (11) 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 $A_{N\times N}^{l(H_m)}$ has at least $N-M$ eigenvalues $-m$. We have

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

and

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

where $x_{L\times 1}$ is an arbitrary vector. Hence, both $\left(R^TR\right)_{N\times N}$ and $\left(RR^T\right)_{M\times M}$ are positive semidefinite, hence all eigenvalues of $\left(R^TR\right)_{N\times N}$ are non-negative. Due to (11), the adjacency eigenvalues of $A_{N\times N}^{l(H_m)}$ are not smaller than $-m$.

3.2.2 The adjacency spectra of the line graph of non-uniform affiliation networks

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

$$A_{N\times N}^{l(H)}=R^TR+C-m_{\max}I \tag{12}$$

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

Since

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

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

4 Characterizing the real-world social networks with overlapping communities

4.1 ArXiv coauthorship networks

In this section, we use the terms "community" and "group" interchangeably. 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 Leskovec et al. (2007). 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 4 and 5, 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 4 and 5 has a power-law tail. The group overlapping depth β follows a power-law distribution. Most of the pairs of groups have no overlap. We only consider the group pairs which overlaps with each other. The membership number m of an individual here means the number of papers he or she authors and coauthors. It also follows a power-law distribution. The interest-sharing number α, denoting the number of papers in which two individuals participate together, follows a power-law distribution. Only the individual pairs who have nonzero interest-sharing number are considered. The ArXiv coauthorship networks of both subjects possess high clustering coefficient, large assortativity coefficient and short average path length as shown in Table 3.

4.2 IMDB actor 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 movie 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 6 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 as shown in Table 3.

4.3 The SourceForge software collaboration network

SourceForge is a web-based project repository assisting programmers 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 7, 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 3.

5 Modeling of social networks with overlapping communities

5.1 Model description

In this section, we use the terms "community" and "group" interchangeably. 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 than 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: $\Gamma = \left[\bar{m}_1\, \bar{m}_2 \cdots \bar{m}_N\right]$. Note that the membership number vector Γ is the input parameter of our hypergraph model.

3. At growing step j, $j = 1, 2, \cdots, N$, 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 k with probability $p_k = s_k / \sum_{i=1}^{j-1} s_i$, where s_k is the group size of group k and $\sum_{i=1}^{j-1} s_i$ is the sum of group sizes of all the existing groups.

   (b) Repeat 3a) $\bar{m}_j - 1$ times so that the membership number of the hyperedge j increases to the expected membership number $\bar{m}_j$.

4. Repeat 3) until the number of hyperedges increases to $N + N_0$.

The model is also presented with pseudo-codes in Algorithm 1. 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 (6) to (10).

ALGORITHM 1: Growing hypergraph model

Input: A seed hypergraph $H_0(M_0, N_0)$ with M_0 nodes and N_0 hyperedges, The membership numbers for new hyperedges $\Gamma = [\bar{m}_1\, \bar{m}_2\, \cdots\, \bar{m}_N]$

Output: A hypergraph $H(N + M_0, N + N_0)$

$H \leftarrow H_0(M_0, N_0)$
for each $j \in \{1, 2, 3, \cdots, N\}$ **do**
  add a new hyperedge j to H
  $m_j \leftarrow 0$
  add a new node to H and let it be incident to the hyperedge j
  $m_j \leftarrow m_j + 1$
  while $m_j < \bar{m}_j$ **do**
    $k \leftarrow$ a random natural number between 1 and $j - 1$
    $r \leftarrow$ a random real number between 0 and 1
    if $r < s_k / \sum_{i=1}^{j-1} s_i$ **then**
      let the hyperedge j be incident to the node k
      $m_j \leftarrow m_j + 1$

5.2 Properties of the growing hypergraph model

5.2.1 Simulation settings

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.

In the growing process, we first apply the constant membership number $m_j = 2$, $j = 1, 2, \cdots, 5000$, obtaining the uniform hypergraph H_2. In the same way, we construct H_3, H_5, H_7, H_{10} and H_{15}. Then we construct the hypergraph $H_{U[1,100]}$ with a uniformly distributed membership number in the interval $[1, 100]$. We construct these hypergraphs in order to study by comparison the properties of H_{pow} which is obtained by applying the sequence of membership numbers with the pdf $\Pr[\Gamma = m] = m^{-2.02}$. We construct H_{pow} in this way: generating a sequence of natural numbers following a power-law distribution with the pdf $\Pr[\Gamma = m] = m^{-2.02}$, and applying this sequence of natural numbers as the membership numbers in the growing process.

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.

5.2.2 Results and discussion

Due to the principle of preferential attachment (Barabási and Albert, 1999), we expect that the group size of all the generated hypergraphs follow power law distributions, which are confirmed by Figure 8. The exponents of the power laws are shown in Table 2.

Table 2 The exponents γ of power-law fittings $f(x) = x^{-\gamma}$ 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 growing hypergraph model with different sequences of membership numbers.

Network	$\gamma(s)$	$\gamma(u)$	$\gamma(\beta)$	$\gamma(m)$	$\gamma(d)$	$\gamma(\alpha)$
ArXiv GRQC	5.50	2.14	3.93	1.95	1.84	3.56
ArXiv HEP-TH	6.24	1.63	3.56	1.72	1.68	2.86
IMDB actors	2.04/5.35	0.407/3.40	4.80	1.81	1.91	3.62
SourceForge	3.91	2.45	3.76	3.48	2.61	4.60
H_2	2.12	2.39	3.38	n.a.	2.35	n.a.
H_3	2.55	2.46	3.07	n.a.	2.16	n.a.
H_5	2.38	2.09	3.19	n.a.	2.12	n.a.
H_7	3.06	2.81	3.11	n.a.	2.59	n.a.
H_{10}	3.22	2.22	3.53	n.a.	2.38	n.a.
H_{15}	2.90	1.95	3.34	n.a.	2.66	n.a.
$H_{U[1,100]}$	3.66	2.85	3.82	n.a.	3.01	n.a.
H_{pow}	3.91	2.45	3.76	3.48	2.61	4.60

Table 3 The clustering coefficients C, the assortativity coefficients ρ_D and the average path lengths l of the arXiv GR-QC and HEP-TH coauthorship networks, the IMDB actor collaboration network, the SourceForge software collaboration network, and the growing hypergraph model with different sequences of membership numbers.

Network	C	ρ_D	l
ArXiv GRQC	0.637	0.584	6.50
ArXiv HEP-TH	0.289	0.382	4.89
IMDB actors	0.762	0.682	4.29
SourceForge	0.636	0.401	7.06
H_2	0.616	0.508	6.13
H_3	0.581	0.576	6.71
H_5	0.491	0.498	7.85
H_7	0.613	0.644	7.62
H_{10}	0.686	0.519	6.89
H_{15}	0.722	0.478	6.56
$H_{U[1,100]}$	0.566	0.422	7.22
H_{pow}	0.636	0.401	7.06

The group degree of all hypergraphs also follows a power-law distribution, as illustrated in Figure 9, where the proper bin size has been used. The exponents are shown in Table 2. The intriguing thing is if the bin size of 1 is chosen, the oscillation appears in the curves of group degree distribution for H_3, H_5, H_5, H_7, H_{10}, H_{15} and $H_{U[1,100]}$, as shown in Figure 10. Fortunately, not depending on the bin size, the group degree of H_{pow} always follow a power law distribution.

The group overlapping depths of all hypergraphs follow power law distributions with exponents which are relatively larger in absolute values,as depicted in Figure 11. All the exponents are reported in Table 2.

Nacher et al. (2005) and Manka et al. (2010) showed that the nodal degree of line graphs of simple graphs with power law degree distribution follows a power law distribution. The individual degree distribution of H_2 is just the degree distribution of line graphs of scale-free graphs. The first of plot in Figure 12 verifies the conclusion of Nacher et al. (2005) and Manka et al. (2010). Figure 12 shows that the individual degree of H_3 still can be said to follow a power-law and is quite similar to that of H_2. The distributions of the individual degree of H_5, H_7 and H_{10} do not follow any power law. The individual degree of H_{15} and $H_{U[1,100]}$ seem to follow power laws $f(x) = x^{-\gamma}$ with negative γ (the exponent $-\gamma$ would be positive). Above all, the individual degree of H_{pow} perfectly follows a power law distribution, as shown in Figure 12. The distribution transition from H_2 to H_{15} is shown in linear-linear scale in Figure 13, where we see the peak of the curve goes from left to right as the membership number increases from 2 to 15.

The interest-sharing number α of only H_{pow} follows a power-law distribution, as illustrated in the 4th plot on the first row of Figure 14. In H_{15} and $H_{U[1,100]}$, the beginning part is linear and the tail is exponential (insets in the two plots on the second row of Figure 14).

The clustering coefficients C, the assortativity coefficients ρ_D and the average path lengths l of all the generated hypergraphs H_2, H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[1,100]}$ and H_{pow} are reported in Table 3. All the generated hypergraphs exhibit high clustering coefficient, high assortativity coefficient and short average path lengths as what real-world affiliation networks show.

6 Conclusion

Many real-world networks, especially social networks, exhibit an overlapping community structure. 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 with power-law distributed membership numbers 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.

Acknowledgements

This work has been done for the project of "Robustness and Optimization of Complex Networks". The authors would like to acknowledge NGInfra (www.nextgenerationinfrastructures.eu) and TRANS (www.trans-research.nl), and thank the anonymous reviewers for their valuable comments.

References

Albert, R. and Barabási, A.-L. (2002) 'Statistical mechanics of complex networks', *Reviews of modern physics*, Vol. 74, pp.47–96.

Boccaletti, S., Latora, V., Moreno, Y, Chavez, M. and Hwang, D.-U. (2006) 'Complex networks: Structure and dynamics', *Physics Reports*, Vol. 424, pp.175–308.

Newman, M.E.J., Watts, D.J. and Strogatz, S.H. (2002) 'Random graph models of social networks', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp.2566–2572.

Girvan, M. and Newman, M.E.J. (2002) 'Community structure in social and biological networks', *Proc. Natl. Acad. Sci. USA*, Vol. 99(12), pp.7821–7826.

Skyrms, B. and Pemantle, R. (2000) 'A dynamic model of social network formation', *Proc. Natl. Acad. Sci. USA*, Vol. 97(16), pp.9340–9346.

Ahn, Y.-Y., Bagrow, J.P. and Lehmann, S. (2010) 'Link communities reveal multiscale complexity in networks', *Nature*, Vol. 466(7307), pp.761–764.

Newman, W.E.J. (2003) 'Mixing patterns in networks', *Phys. Rev. E*, Vol. 67(2), pp.026126.

Newman, W.E.J. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Phys. Rev. E*, Vol. 69(2), pp.026113.

Van Mieghem, P., Wang, H., Ge, X., Tang, S. and Kuipers, F.A. (2010) 'Influence of assortativity and degree-preserving rewiring on the spectra of networks', *The European Physical Journal B - Condensed Matter and Complex Systems*, Vol. 76(4), pp.643–652.

Newman, M.E.J., Strogatz, S.H. and Watts, D.J. (2001) 'Random graph with arbitrary degree distribution and their applications', *Phys. Rev. E*, Vol. 64, pp.026118.

Lattanzi, S. and Sivakumar, D. (2009) 'Affiliation networks', *Proceedings of the 41st annual ACM symposium on Theory of computing*, pp.427–434.

Van Mieghem, P. (2011), *Graph Spectra for Complex Networks* Cambridge University Press, Cambridge, U.K.

Scott, J. (1991), *Social Network Analysis: A Handbook* SAGE Publications Ltd, London, U.K.

Cvetković, D., Rowlinson, P. and Simić (2007) 'Eigenvalue bounds for the signless laplacians', *Publ. Inst. Math. (Beograd)*, Vol. 81(95), pp.11–17.

Leskovec, J., Kleinberg, J. and Faloutsos, C. (2007) 'Graph evolution: Densification and shrinking diameters', *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, Vol. 1(1), pp.35–49.

Barabási, A.-L. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286(5439), pp.509–512.

Watts, D.J. and Strogatz, S.H. (1998) 'Collective dynamics of 'small-world' networks', *Nature*, Vol. 393(1998), pp.440–442.

Fortunato, S. (2010) 'Community detection in graphs', *Physics Reports*, Vol. 486(3–5), pp.75–174.

Van Mieghem, P., Ge, X., Schumm, P., Trajanovski, S. and Wang, H. (2010) 'Spectral graph analysis of modularity and assortativity', *Phys. Rev. E*, Vol. 82(5), pp.056113.

Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. (2005) 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature*, Vol. 435(7043), pp.814–818.

Evans, T.S. and Lambiotte, R. (2009) 'Line graphs, link partitions, and overlapping communities', *Phys. Rev. E*, Vol. 80(1), pp.016105.

Mcdaid, A. and Hurley, N.J. (2010) 'Detecting highly overlapping communities with model-based overlapping seed expansion', *ASONAM 2010*.

Poller, P., Palla, G. and Vicsek, T. (2006) 'Preferential attachment of communities: the same principle, but a higher level', *Europhys. Lett.*, Vol. 73(3), pp.478–484.

Toivonen, R., Onnela, J.-P., Saramki, J., Hyvnen, J. and Kaski, K. (2006) 'A model for social networks', *Physica A: Statiscal and Theoretical Physics*, Vol. 371(2), pp.851–860.

Nacher, J., Yamada, T., Goto, S., Kanehisa, M. and Akutsu, T. (2005) 'Two complementary representations of a scale-free network', *Physica A: Statistical Mechanics and its applications*, Vol. 349(1–2), pp.349–363.

Manka-Krason, A., Mwijage, A. and Kulakowski, K. (2010) 'Clustering in random line graphs', *Computer Physics Communications*, Vol. 181(1), pp.118–121.

Manka-Krason, A. and Kulakowski, K. (2010) 'Assortativity in random line graphs', *Acta Physica Polonica B Proceedings Supplement*, Vol. 3(10), pp.259–266.

Figure 1 The example graph to illustrate the community structure. The nodes denote individuals. The communities consist of links of the same color and the shared thick black link(s), and the nodes incident to the links of both colors.

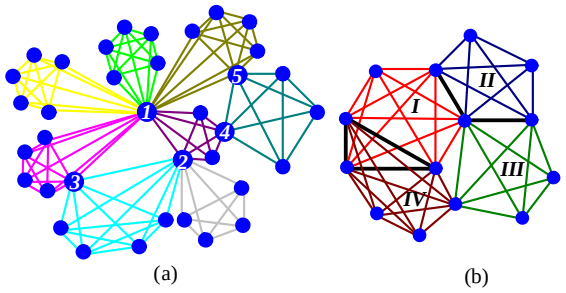

Figure 2 The bipartite graph representation of the affiliation network of the NAS group.

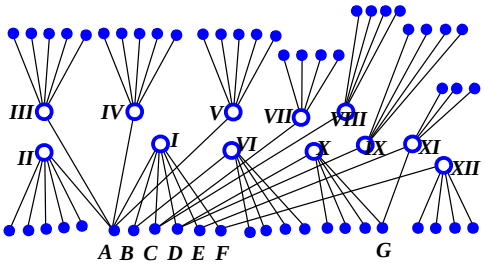

Figure 3 (a) The hypergraph representation of the network described in Table 1. The hyperlinks are the blue ellipse-like closed curves. The nodes are the disks with different colors marked with Roman numerals. A node and a hyperlink are incident if the node is surrounded by the hyperlink. The hyperlinks and nodes represent the individuals and the communities respectively. Individuals participate in multiple communities, implying that the communities 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.

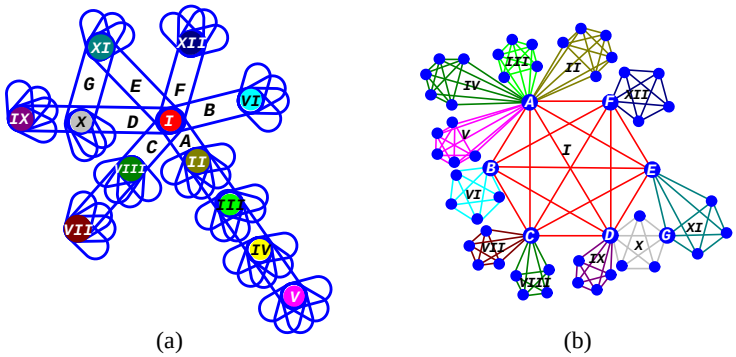

Figure 4 The probability density distribution of group size s, group degree u, group overlapping depth β (the first row from left to right), individual membership number m, individual degree d, individual interest-sharing number α (the second row from left to right) of ArXiv coauthorship networks of "General Relativity and Quantum Cosmology" category.

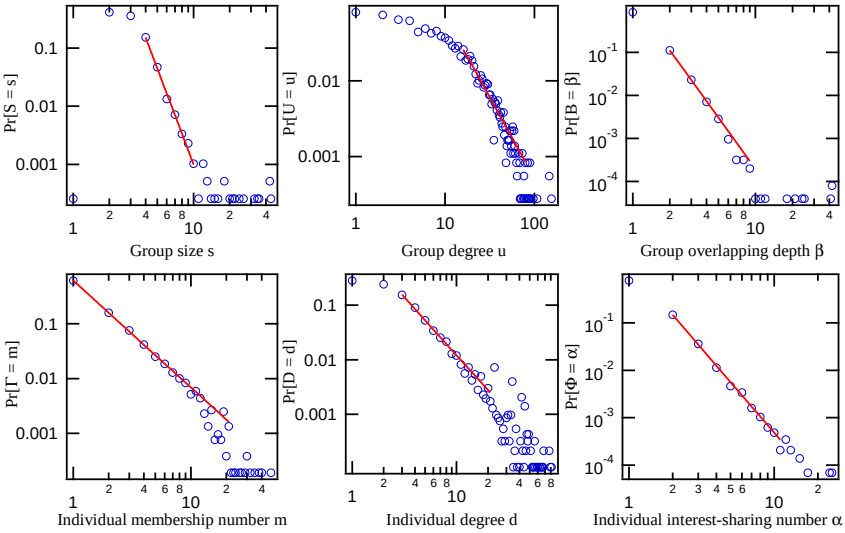

Figure 5 The probability density distribution of group size s, group degree u, group overlapping depth β (the first row from left to right), individual membership number m, individual degree d, individual interest-sharing number α (the second row from left to right) of ArXiv coauthorship networks of "High Energy Physics - Theory" category.

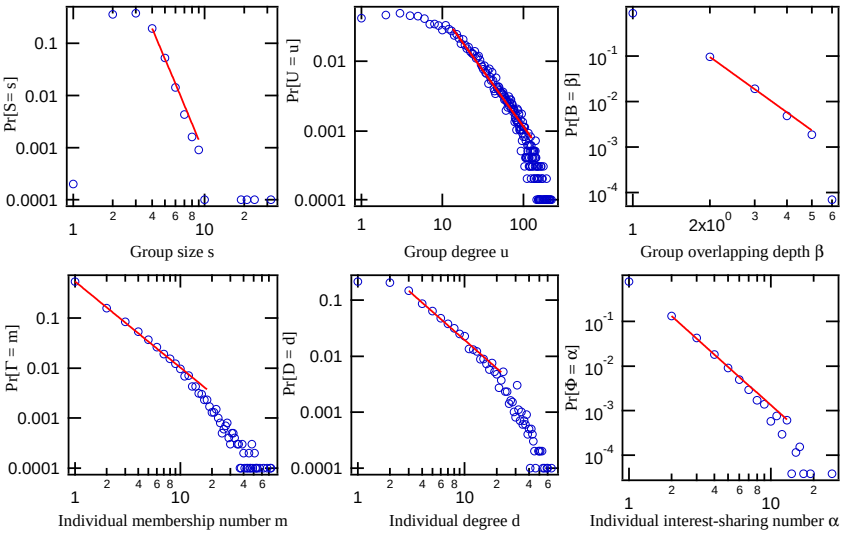

Figure 6 The probability density distribution of group size s, group degree u, group overlapping depth β (the first row from left to right), individual membership number m, individual degree d, individual interest-sharing number α (the second row from left to right) of IMDB movie actors collaboration networks.

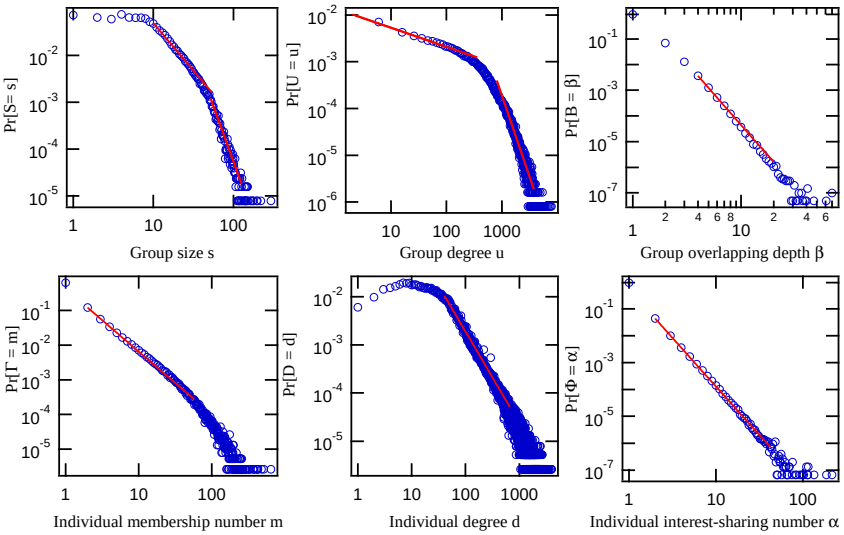

Figure 7 The probability density distribution of group size s, group degree u, group overlapping depth β (the first row from left to right), individual membership number m, individual degree d, individual interest-sharing number α (the second row from left to right) of the SourceForge software collaboration network.

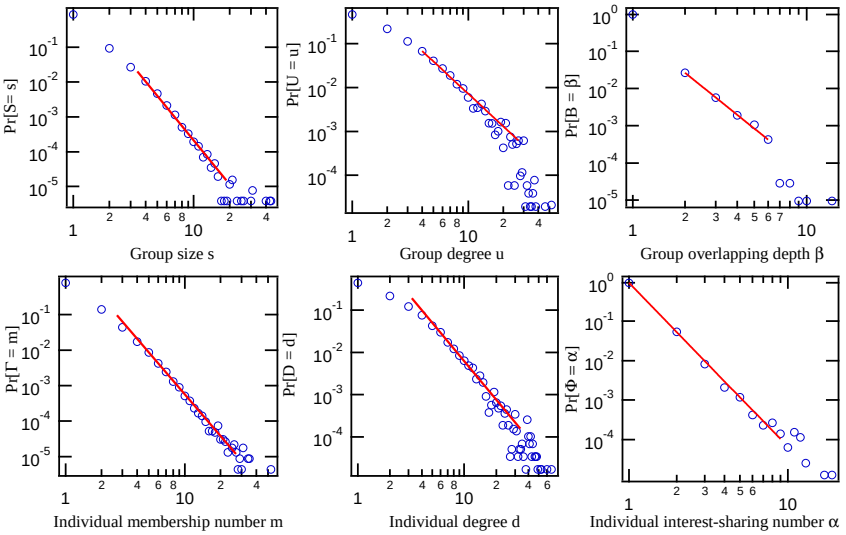

Figure 8 The probability density distribution of group size s for H_2, H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[2,121]}$, and H_{pow}. They all have 5020 groups (nodes) and 5020 hyperedges (individuals).

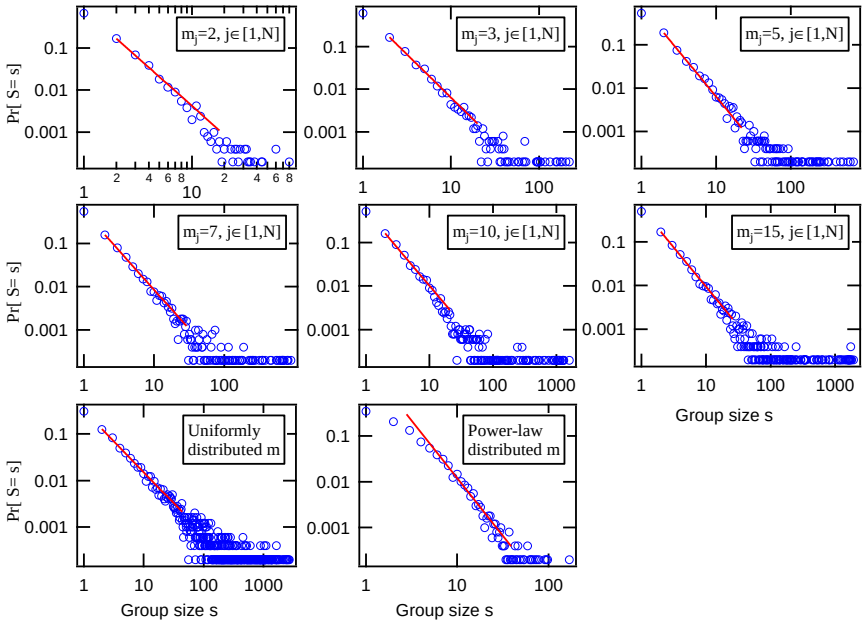

Figure 9 The probability density distribution of group degree u for H_2, H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[2,121]}$, and H_{pow}.

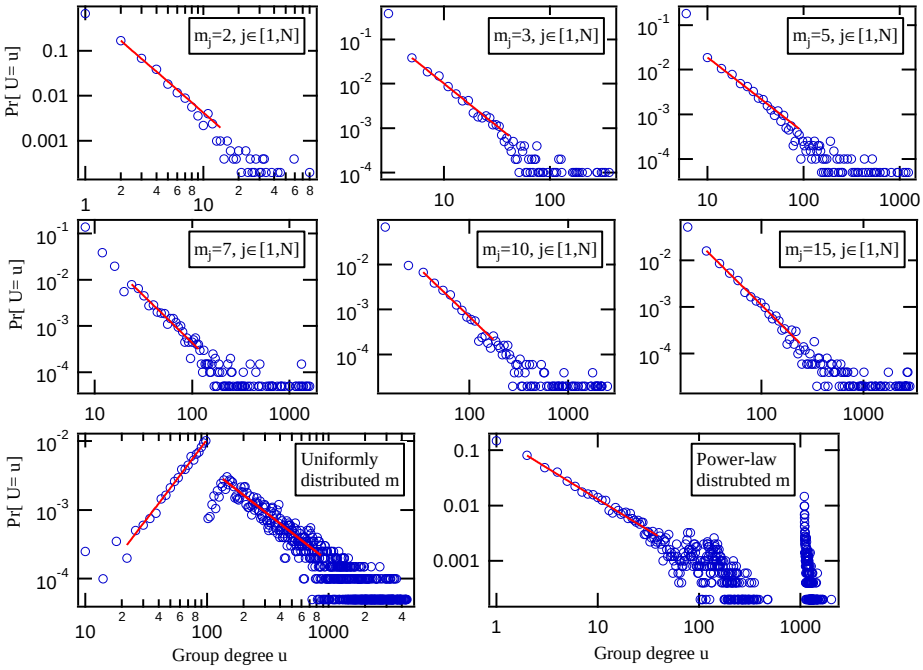

Figure 10 The probability density distribution of group degree u for H_3, H_5, H_7, H_{10}, H_{15}, and $H_{U[2,121]}$ with bin size of 1, showing oscilation.

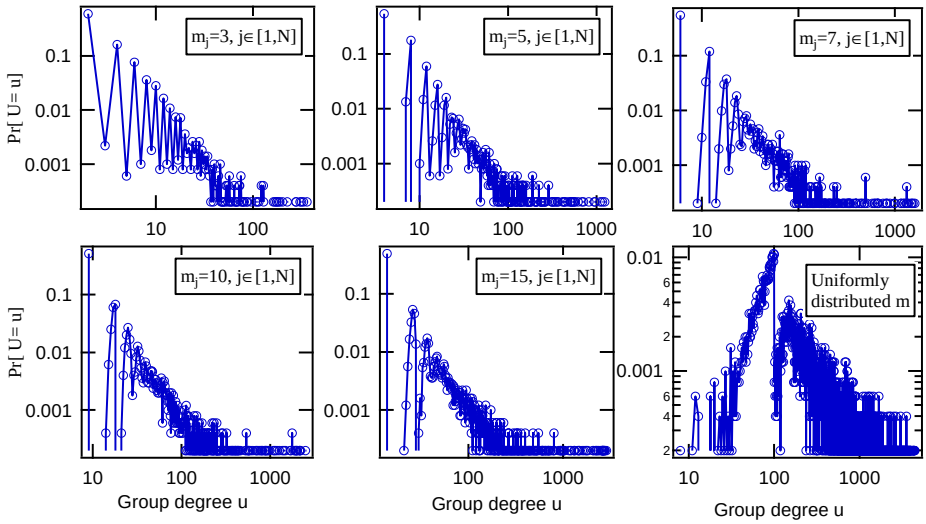

Figure 11 The probability density distribution of group overlapping depth β for H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[2,121]}$, and H_{pow}. They all have 5020 groups (nodes) and 5020 hyperedges (individuals).

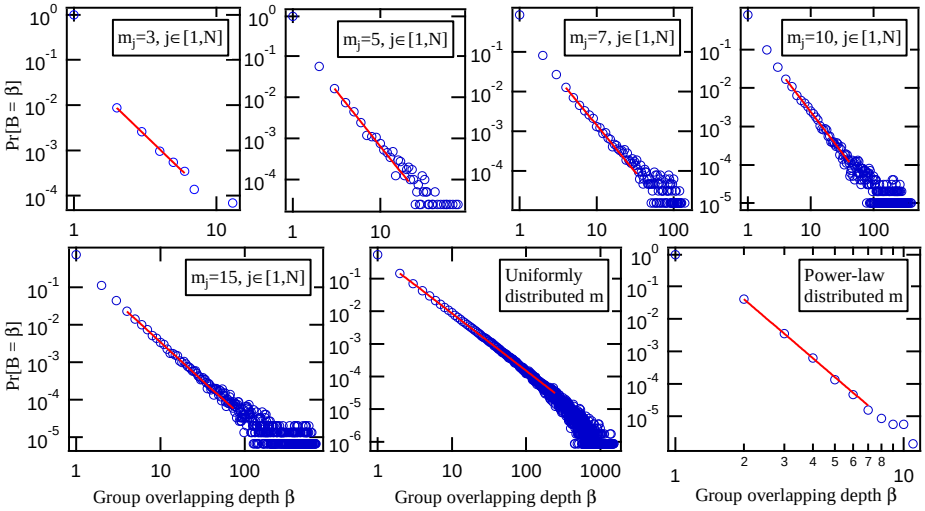

Figure 12 The probability density distribution of individual degree d for H_2, H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[2,121]}$, and H_{pow}. They all have 5020 groups (nodes) and 5020 hyperedges (individuals).

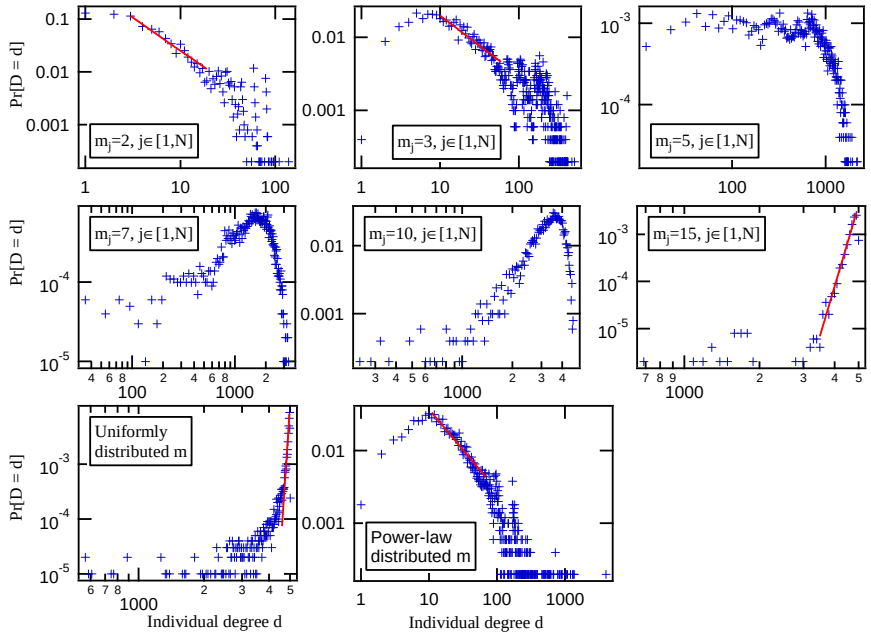

Figure 13 The probability density distribution of individual degree d for H_2, H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[2,121]}$, and H_{pow} in linear-linear scale. They all have 5020 groups (nodes) and 5020 hyperedges (individuals).

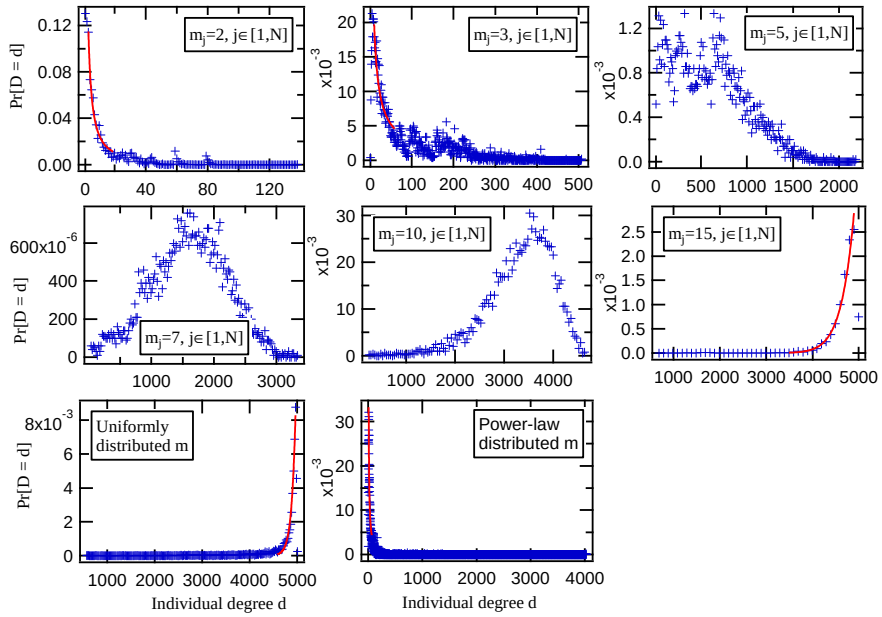

Figure 14 The probability density distribution of individual interest-sharing number α for H_2, H_3, H_5, H_7, H_{10}, H_{15}, $H_{U[2,121]}$, and H_{pow}. They all have 5020 groups (nodes) and 5020 hyperedges (individuals).