

# 基于信息瓶颈的社区发现

沈华伟<sup>1),2)</sup> 程学旗<sup>1)</sup> 陈海强<sup>1),2)</sup> 刘悦<sup>1)</sup>

<sup>1)</sup>(中国科学院计算技术研究所 北京 100080)

<sup>2)</sup>(中国科学院研究生院 北京 100049)

**摘 要** 该文提出一种映射方法,把单部网络变换成二部图网络.针对得到的二部图网络,在信息论的框架下,提出了一种基于信息瓶颈的社区发现方法.该方法通过寻找网络的最优压缩表示来发现网络的社区结构,最优压缩表示尽可能多地保留原始网络的拓扑特征.在真实数据集和计算机产生的数据集上的实验表明,该方法能够有效地发现网络的社区结构.另外,对于有向网络的社区发现,现有方法忽略有向网络中边的方向而作为无向网络来处理,损失了有向网络的方向信息,文中提出的社区发现方法能够很好地解决这一问题,并能从有向网络中挖掘出一些现有方法无法发现的知识,这一特点使得该文的方法比现有方法更适用于解决像 WWW 这样的有向网络.同时,真实世界的许多网络本身就是二部图网络,相对于现有的社区发现方法,文中的方法可以直接应用于这类网络.

**关键词** 社区发现;信息瓶颈;聚团性

**中图法分类号** TP391

## Information Bottleneck Based Community Detection in Network

SHEN Hua-Wei<sup>1),2)</sup> CHENG Xue-Qi<sup>1)</sup> CHEN Hai-Qiang<sup>1),2)</sup> LIU Yue<sup>1)</sup>

<sup>1)</sup>(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

<sup>2)</sup>(Graduate University of Chinese Academy of Sciences, Beijing 100049)

**Abstract** This paper proposes a projection method to transform a unipartite network into a bipartite network. As to the obtained bipartite network, it presents an information-bottleneck-based method for community detection under the information-theoretic framework. This method detects the community structure of networks by finding an efficient compression of the network. The efficient compression holds the regularity of the original network as many as possible. Applications on the computer-generated networks and many real-world networks demonstrate that this method is very effective at community detection of networks. As for the community detection of directed networks, existing methods neglect the direction of edges and treat them as undirected networks. The information provided by directionality is lost in this process. Using the projection method proposed in this paper, the direction of edges can be retained and thus the method is more suitable to detect the community structure in directed networks, such as the world-wide-web. And some new knowledge can be found by the method. In addition, the method can be directly applied to the detection of community structure in bipartite networks, which are common in real world.

**Keywords** community detection; information bottleneck; modularity

收稿日期:2007-12-10. 本课题得到国家“九七三”重点基础研究发展规划项目基金(2004CB318109, 2007CB311100)、微软亚洲研究院 IST2007- Web 2.0 社区发现与社区演化研究课题(FY07-RES-THEME-067)资助. 沈华伟,男,1982年生,博士研究生,主要研究方向为社会计算、复杂网络、信息检索. E-mail: shenhuawei@software.ict.ac.cn. 程学旗,男,1971年生,博士,研究员,研究领域包括网络信息安全、大规模信息检索与信息挖掘、P2P 计算等. 陈海强,男,1979年生,博士研究生,主要研究方向为社会计算、复杂网络、信息检索. 刘悦,女,1971年生,博士,副研究员,研究方向包括 Web 搜索和挖掘、复杂网络分析与社会计算、分布式系统等.

## 1 引 言

真实世界中的许多系统可以使用网络来表示,例如,Internet、WWW、科技文献引用关系网络、食物链网络、社会关系网络等.一个网络一般由一个顶点集和一个边集构成.对于 WWW 而言,每一个网页是一个顶点,所有网页构成网络的顶点集,网页间的超链接是边,所有超链接构成这个网络的边集.

近几年,网络统计特征得到研究人员越来越多的关注.网络的一些统计特征,像小世界效应、顶点度分布、顶点的高聚团性等逐渐为研究人员所熟知<sup>[1-2]</sup>.但是,目前的研究主要关注于网络的宏观和微观结构特征,宏观特征是基于网络整体,微观特征是基于顶点,介于宏观和微观之间的特征却没有得到充分的关注和研究.

社区结构是一种介于宏观和微观之间的网络特征,是真实世界中许多复杂网络所具有的一种普遍性质<sup>[3]</sup>.社区结构和网络的功能有着紧密的关系,像鲁棒性、高传播速度等.发现网络社区结构是揭示网络结构和功能之间关系的重要基础.因此,对于网络社区结构的研究开始得到许多领域的学者日益增加的关注和研究<sup>[4]</sup>.

社区没有一个明确的定义,一个被普遍接受的观点是:一个社区是网络的一个子图,子图中的顶点更倾向于和子图内的顶点有边相连.因此,社区内部顶点连边密度较高,不同社区之间顶点连边密度相对较低.对于真实网络,同属于一个社区的顶点更有可能具有相似的性质或相近的功能.在 WWW 网络中,同一个社区的页面通常表达相近的主题;在神经网络中,一个社区通常对应一个功能组,也就是说一个社区的顶点有着相近的功能;在科技文献共同署名网络中,同一个社区内的学者有着相近的研究兴趣.发现网络的社区结构能够发现新的知识和现象,有助于更深刻地理解和认识网络结构和功能之间的关系.

网络社区结构的研究有着很长的历史.作为网络社区发现的一个简化问题,图划分问题几十年来得到了广泛和深入的研究.图划分问题的一个典型应用是并行计算中的任务分配问题.有  $C$  个处理器处理  $N$  个任务,这些任务之间在处理过程中可能需要相互通信.为简单起见,我们假设:(1) 每个任务都需要相同的处理时间;(2) 任何有通信的两个任务之间的通信量都是一样的.这样一来,我们就得到

了一个无权无向简单图, $N$  个顶点代表  $N$  个任务,边连接需要通信的任务.相对于同一处理器上的两个任务之间的通信代价,不同处理器上两个任务之间通信的代价是非常高的.因此,为了提高处理的效率,应减少不同处理器上的任务之间的通信.从而,这样一个任务分配问题就变成了一个典型的图划分问题,把图中的  $N$  个顶点划分成  $C$  个组(社区),在平衡各个处理器的负载的同时,使组间的通信尽可能地少.精确求解图划分问题是困难的,因此多种启发式算法被提出,试图近似求解图划分问题.在很多情况下,这些启发式算法能够得到可以接受的解.在这类算法中,最著名的是由 Kernighan-Lin 提出的算法<sup>[5]</sup>,该算法在稀疏图上的时间复杂度可以达到  $O(n^3)$ .

和图划分问题不同,对于社区发现问题,事先并不知道一个网络中有多少个社区存在,各个社区包含的顶点个数也是未知的.这使得社区发现问题是一个比图划分更加困难的问题.另外,网络的社区结构通常呈现出层次特征,一个社区可以进一步划分成几个子社区;有些顶点出现在不同的社区中,这使得社区之间有重叠.这些现象也增加了社区发现的难度.

目前,已经有许多社区发现方法提出,分别基于连边密度<sup>[6]</sup>、介数<sup>[7]</sup>、信息中心度<sup>[8]</sup>、随机行走<sup>[9]</sup>、谱分析<sup>[10]</sup>、最优化某个目标质量函数<sup>[11]</sup>等等,最有代表性的一类方法是基于优化网络聚团性(modularity)的方法<sup>[11]</sup>.Modularity<sup>[7]</sup>是由 Newman 提出的,最早是衡量网络划分好坏的一种度量.Modularity 值(通常也叫  $Q$  值)的计算方法为

$$Q = \sum_i (e_{ii} - a_i^2) \quad (1)$$

其中, $e_{ij}$  表示社区  $i$  和社区  $j$  之间的边数占总边数的比率; $a_i = \sum_j e_{ij}$  表示有一个端点在社区  $i$  中的边占边总数的比率.Modularity 的基本思想是:完全随机网络没有社区结构,如果一个网络有良好的社区结构,就存在一种对该网络的划分,使得这种划分对应一个较高的  $Q$  值.对于真实世界的网络而言, $Q$  的取值一般介于  $0.3 \sim 0.7$  之间.基于 modularity 的方法大多旨在优化  $Q$  值,希望找到一种网络的划分,使得这种划分对应的  $Q$  值最大.一个网络的所有可能划分数等于第二类 Stirling 数<sup>[11]</sup>,因此枚举所有划分是计算上不可行的.研究人员提出许多启发式方法来优化  $Q$  值,贪婪方法<sup>[12]</sup>、模拟退火算法<sup>[13]</sup>、极值优化方法<sup>[14]</sup>等被用来寻找较优的  $Q$  值,

这些方法在许多网络上取得了不错的结果。

对于社区结构的认识, 基于 modularity 的方法认为社区结构是真实网络和随机网络之间的统计偏差, 社区发现的过程是找一种最优划分使这种偏差最大. 社区结构的另一种认识是基于网络压缩的, 认为社区结构是网络结构规则性的一种体现, 而发现网络中的社区结构是去掉那些不重要的细节而保留网络的整体拓扑特征, 这一过程可以看成是对网络拓扑的有损压缩<sup>[15]</sup>. 基于社区结构的第二种认识, 一些社区发现方法被提出<sup>[15-16]</sup>, 其中, 文献[16]考虑使用信息瓶颈来定义网络模块性并基于这一定义, 提出了一种社区发现方法.

基于社区结构的第二种认识, 本文重新思考社区发现问题. 社区发现转变成寻找网络结构的一种高效压缩表示, 这种压缩能够尽可能多地反映原始网络中的结构规则性. 本文首先提出一种变换方法, 把网络转换成二部图网络; 进而, 针对该二部图, 在信息论的框架下, 提出了一种新的基于信息瓶颈的社区发现方法.

为验证基于信息瓶颈社区发现方法的有效性, 本文展示了该方法在计算机生成网络上的性能, 并把该方法应用到一些真实世界的网络中, 而且与现有方法作了对比. 结果表明, 基于信息瓶颈的社区发现方法能够有效地发现这些网络中的社区结构.

本文第 2 节简要叙述了信息瓶颈方法; 第 3 节给出一种把一般网络变换成二部图网络的方法, 阐述了如何使用信息瓶颈方法发现网络的社区结构, 给出了基于信息瓶颈社区发现方法的一般框架; 第 4 节把该社区发现方法应用到计算机生成的网络和真实网络上, 以检验方法的有效性; 第 5 节总结了全文, 提出一些有待探讨的问题, 并对未来工作进行了展望.

## 2 信息瓶颈方法

在介绍基于信息瓶颈的社区发现方法之前, 本节对信息瓶颈方法做一简单介绍.

信息瓶颈方法由 Tishby 等人提出<sup>[17]</sup>. 信息瓶颈方法基于下面的基本思想: 给定两个随机变量  $X$  和  $Y$  的先验联合分布  $p(x, y)$ , 压缩其中一个随机变量  $X$ , 同时保证尽可能多地维持两个变量的互信息  $I(X, Y)$ . 依据香农信息论, 两个随机变量  $X$  和  $Y$  的互信息表示它们的联合分布  $p(x, y)$  关于各自边缘分布乘积  $p(x)p(y)$  的相对熵, 它反映了变量

$X(Y)$  包含的关于变量  $Y(X)$  的信息量. 假定变量  $X$  压缩后的表示记为  $C$ , 有  $I(C, Y) \leq I(X, Y)$ .

和著名的率失真理论相似, 我们要在尽可能压缩  $X$  的表示长度和尽可能多地维持关于  $Y$  的信息之间寻求一个折中. 每一种压缩对应着一种由  $X$  到  $C$  的赋值  $p(c|x)$ ,  $p(c|x)$  表示  $X$  的一个取值  $x$  对应  $C$  中取值  $c$  的概率. 一般情况下, 每一个  $x$  可以对应  $C$  中多个甚至所有取值  $c$ , 这种情况下的赋值称为软赋值, 如果一个  $x$  只对应唯一一个  $c$ , 这种赋值为硬赋值. 信息瓶颈方法就是找到一种最优的赋值, 最小化

$$L[p(c|x)] = I(C, X) - \beta I(C, Y) \quad (2)$$

这里, 参数  $\beta$  是拉格朗日乘子, 用以保证  $C$  包含足够多地关于  $Y$  的信息.

在对先验联合分布  $p(x, y)$  不作任何假设的情况下, 问题(2)对应着一个精确的最优解. 这个最优解通过 3 个分布给出: 第 1 个分布是  $C$  的分布  $p(c)$ ; 第 2 个分布是由  $X$  到  $C$  的赋值  $p(c|x)$ ; 第 3 个分布是  $p(y|c)$ , 它刻画  $C$  和  $Y$  的关系. 精确解的形式表达如下:

$$\begin{cases} p(c|x) = \frac{p(c)}{Z(\beta, x)} \exp(-\beta D_{\text{KL}}[p(y|x) \| p(y|c)]) \\ p(y|c) = \frac{1}{p(c)} \sum_x p(c|x) p(x) p(y|x) \\ p(c) = \sum_x p(c|x) p(x) \end{cases} \quad (3)$$

这里,  $Z(\beta, x)$  是归一化因子, 参数  $\beta$  决定了赋值的“软硬程度”,  $D_{\text{KL}}[p(y|x) \| p(y|c)]$  是  $p(y|x)$  和  $p(y|c)$  之间的 Kullback-Leibler 距离, 用以度量  $X$  和  $C$  在表示  $Y$  时的偏差.

方程组(3)可以通过迭代的方式来求解, 迭代是收敛的. 当参数  $\beta \rightarrow \infty$  时, 我们得到的解对应一种硬赋值, 也就是说每一个  $x$  只对应于一个  $c$ . 此时, 只考虑(2)中的第 2 项, 而忽略了第 1 项, 进而方程组(3)可以表示成

$$\begin{cases} p(c|x) = \begin{cases} 1, & \text{如果 } x \in c \\ 0, & \text{其它} \end{cases} \\ p(y|c) = \frac{1}{p(c)} \sum_{x \in c} p(x, y) \\ p(c) = \sum_{x \in c} p(x) \end{cases} \quad (4)$$

方程组(4)可以使用一种简单的层次聚类方法来求解<sup>[18]</sup>. 初始时, 每一个  $x$  单独属于一个类别, 也就是说没有对  $X$  进行任何压缩. 每一步, 我们合并两个

类别,选择要合并的两个类别时,依据局部最小化信息损失的原则,选择合并后  $I(C,Y)$  减少最小的两个类别进行合并.

假设要合并的两个类别为  $c_i$  和  $c_j$ , 合并后新生成的类别记为  $c_*$ , 合并过程可以形式化表示成方程组(5).

$$\begin{cases} p(c_*|x) = \begin{cases} 1, & x \in c_i \text{ 或 } x \in c_j \\ 0, & \text{其它} \end{cases} \\ p(y|c_*) = \frac{p(c_i)}{p(c_*)}p(y|c_i) + \frac{p(c_j)}{p(c_*)}p(y|c_j) \\ p(c_*) = p(c_i) + p(c_j) \end{cases} \quad (5)$$

合并代价是指合并所带来的互信息的减少量, 定义为

$$\delta I(c_i, c_j) \equiv I(C_{\text{before}}; Y) - I(C_{\text{after}}; Y) \quad (6)$$

通过一些简单的代数变换, 得到

$$\delta I(c_i, c_j) \equiv (p(c_i) + p(c_j)) \cdot D_{\text{JS}}[p(y|c_i), p(y|c_j)] \quad (7)$$

这里, 函数  $D_{\text{JS}}$  是 Jensen-Shannon(JS) 距离, 计算方法为

$$D_{\text{JS}}[p_i, p_j] = \pi_i D_{\text{KL}}[p_i \parallel \hat{p}] + \pi_j D_{\text{KL}}[p_j \parallel \hat{p}] \quad (8)$$

对于我们的问题, 有

$$\begin{cases} \{p_i, p_j\} \equiv \{p(y|c_i), p(y|c_j)\} \\ \{\pi_i, \pi_j\} \equiv \left\{ \frac{p(c_i)}{p(c_*)}, \frac{p(c_j)}{p(c_*)} \right\} \\ \hat{p} = \pi_i p(y|c_i) + \pi_j p(y|c_j) \end{cases} \quad (9)$$

JS 距离是非负的, 当且仅当两个参数是一致的时候才取值为 0, 上界是 1, 且是对称的.

值得注意的是, 合并代价可以看成参加合并“元素”的权重与其 JS 距离的乘积.

3 基于信息瓶颈的社区发现方法

3.1 图的变换

给定一个无向图  $G=(V,E)$ , 假定变换后的二部图记为  $B$ . 变换规则如下: (1) 对于  $G$  中的任意一个顶点  $v_i$ , 在  $B$  中对应两个顶点  $u_i$  和  $w_i$ ; (2) 如果在  $G$  中顶点  $(v_i, v_j) \in E$ , 对应  $B$  中两条边  $(u_i, w_j)$  和  $(u_j, w_i)$ ; (3) 边  $(u_i, w_j)$  和  $(u_j, w_i)$  的权重等于  $(v_i, v_j)$  的权重. 图 1(a) 是一个无向图变换成二部图的例子.

对于有向图  $G=(V,E)$ , 假定变换后的二部图记为  $B$ . 变换规则为: (1) 对于  $G$  中的任意一个顶点  $v_i$ , 在  $B$  中对应两个顶点  $u_i$  和  $w_i$ ; (2) 如果在  $G$  中顶点  $\langle v_i, v_j \rangle \in E$ , 对应  $B$  中一条边  $\langle u_i, w_j \rangle$ , 这里使用

尖括号表示边是有向边; (3) 边  $\langle u_i, w_j \rangle$  的权重等于  $\langle v_i, v_j \rangle$  的权重. 图 1(b) 是一个有向图变换成二部图的例子.

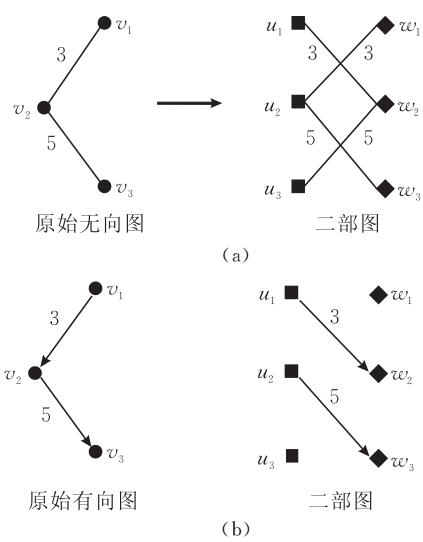


图 1 图变换成二部图的例子

3.2 基于信息瓶颈的社区发现

对于变换得到的二部图网络, 其矩阵表示记为  $M$ , 矩阵的行对应二部图的左部, 列对应二部图的右部. 矩阵元素  $M_{ij}$  表示边  $(u_i, w_j)$  或  $\langle u_i, w_j \rangle$  的权重, 如果  $u_i$  和  $w_j$  之间没有边相连, 权重为 0.

把二部图的左部看成是随机变量  $X$ , 右部看成是随机变量  $Y$ ,  $X$  和  $Y$  的联合分布可以由二部图对应的矩阵表示  $M$  直接构造出来. 构造方法为对  $M$  的所有元素均除以  $TW$ ,  $TW$  是  $M$  中所有元素之和, 也是所有边的权重之和 (Total Weight).

对于无向网络, 原始网络的顶点集和二部图网络任意一部的顶点集是完全一样的. 二部图任意一部的压缩表示均对应着原始网络的一种压缩表示, 而原始网络的压缩表示包含着原始网络的社区结构的全部知识. 而且, 无向网络对应的二部图网络是左、右部对称的. 因此, 我们可以通过压缩二部图网络的任意一部来达到发现原始网络的社区结构的目的.

对于有向网络, 所得二部图的左部和右部从入边和出边两个不同的角度来表现原始网络的结构特征. 和无向网络一样, 压缩该二部图的任意一部可以得到原始网络的社区结构. 和无向网络不同的是, 压缩左部和压缩右部所得到的社区结构是不一样的, 两种压缩分别从入边和出边这两个角度反映原始网络的社区结构特征. 表 1 给出了一个有向网络, 该网络共有 12 个结点, 标号为 1~12, 顶点 1~6 的

出边指向顶点 1~3 和 7~9, 顶点 7~12 的出边指向顶点 4~6 和 10~12, 其它顶点之间没有边相连. 对于这样一个网络, 按照本文的变换方法得到一个二部图网络. 如果压缩其左部, 得到的社区结构反映了原始网络的出边特征, 得到的社区为 (1~6) 和 (7~12); 如果压缩其右部, 得到的社区结构反映了原始网络的入边特征, 得到的社区为 (1~3, 7~9) 和 (4~6, 10~12). 相比于现有的方法, 本文下面提出的基于信息瓶颈的方法可以很好地解决这一问题.

表 1 有向网络的一个例子

Out	In	
	1~3 和 7~9	4~6 和 10~12
1~6	Yes	No
7~12	No	Yes

现在介绍基于信息瓶颈的社区发现方法的一般框架. 假设我们选择压缩二部图网络的左部  $X$ ,  $X$  的压缩会使  $X$  和二部图网络的右部  $Y$  的互信息  $I(X, Y)$  减少, 需要在尽可能压缩  $X$  和尽可能多地维持关于  $Y$  的信息之间做折中. 我们使用第二节介绍的信息瓶颈方法来完成这一折中任务. 本文使用信息瓶颈方法的一种简单实现——层次聚类方法, 这种方法对应参数  $\beta \rightarrow \infty$ , 得到的解是一种硬赋值, 是对原始网络的一种划分.

算法的基本思想是: 初始时, 每个顶点被看成是一个社区, 然后逐步合并各个社区, 每次合并选择合并代价最小的两个社区进行合并, 合并过程直到只有一个社区时停止. 合并代价是基于第二节介绍信息瓶颈方法时所使用的合并代价.

算法输出结果是一个 dendrogram, 这是一个有层次结构的树图(图 2 是一个例子), 从每一层对该树图进行切分都对应原始网络的一种划分. 选择在 哪一层进行切分是一个重要的问题. 我们使用 modularity 值来衡量一种划分的质量, 选择 modularity 值最大的那个划分作为原始网络的最终的划分. 算法 1 给出了算法的具体实现.

**算法 1.** 基于信息瓶颈社区发现方法的具体实现步骤.

- 输入：一个网络
- 输出：网络的一种划分
- 初始化：
1. 把输入网络变换成二部图网络, 并表示成两个随机变量  $X$  和  $Y$  的联合分布  $p(x, y)$ .

2. 构建一个随机变量  $C \equiv X$ .

3. 对于任意的  $i, j = 1 \cdots |X|, i < j$ , 计算

$$d_{ij} \equiv (p(c_i) + p(c_j)) \cdot D_{JS}[p(y|c_i), p(y|c_j)]$$

循环：

For  $m = |X| - 1 \cdots 1$

找出  $d_{ij}$  最小的  $i, j$

把  $c_i, c_j$  合并成  $c_*$

从  $C$  中删除  $c_i, c_j$ , 并把  $c_*$  加入到  $C$  中

更新和  $c_*$  相关的  $d_{ij}$

计算当前的 modularity 值, 记录下来

End

输出结果：

1. 找出最大的 modularity 值

2. 输出对应的各个社区.

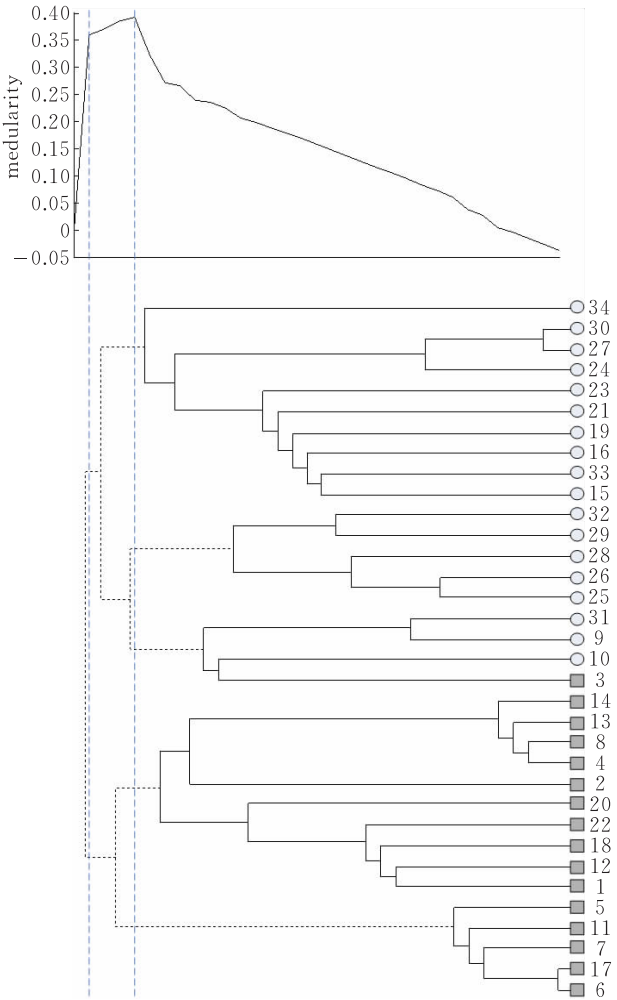


图 2 对 Zachary 空手道俱乐部网络, 使用本文的方法得到的树图(dendrogram)

在计算  $d_{ij}$  时, 只有当  $c_i, c_j$  之间有边相连时, 才需要计算  $d_{ij}$ , 因为合并没有边相连的  $c_i$  和  $c_j$  是没有意义的. 这样大大减少了运算次数, 提高了算法的计算效率.

下面分析算法的时间复杂度. 假设网络的顶点个数为  $N$ , 边的个数为  $M$ . 初始化阶段, 计算一个

$d_{ij}$  的时间复杂度为  $O(N)$ , 由于只需要计算有边相连的  $c_i, c_j$  对应的  $d_{ij}$ , 那么初始化阶段的总体时间复杂度为  $O(M \times N)$ . 循环阶段, 共循环了  $N-1$  次, 每次循环中, 寻找最小  $d_{ij}$  的时间复杂度为  $O(M)$ , 合并和删除的时间复杂度为  $O(N)$ , 更新  $d_{ij}$  时只更新了和  $c_*$  相关的  $d_{ij}$ , 因此时间复杂度不超过  $O(N^2)$ , 计算 modularity 值的时间复杂度也不超过  $O(N^2)$ . 输出结果的时间复杂度小于循环阶段的时间复杂度. 因此整个算法的时间复杂度取决于循环阶段的时间复杂度, 为  $O(N^3)$ . 但这只是最坏情况下的时间复杂度, 通常情况下算法的时间复杂度要比分析的. 在下一节的实验中, 我们把该算法应用到有 5000 多个顶点的网络中, 在几分钟内便得到输出结果. 另外, 这个算法是可以优化的, 这在我们以后的工作中解决.

## 4 实验结果与分析

本节分别在计算机生成的网络和真实世界的网络上验证基于信息瓶颈社区发现方法的有效性和适用性.

### 4.1 计算机生成的网络上的实验

为了测试方法的有效性, 使用计算机生成有已知社区结构的网络, 在这样的网络上实验, 测试我们的方法能否发现或抽取出这些已知的社区结构.

#### 4.1.1 网络的生成

使用计算机生成一系列有已知社区结构的网络<sup>[7]</sup>. 每个网络有 128 个顶点, 这些顶点分别属于 4 个已知的社区中. 顶点之间的边是随机添加的, 同一个社区的两个顶点之间有边相连的概率是  $p_{in}$ , 不同社区的两个顶点之间有边相连的概率是  $p_{out}$ .  $p_{out}$  的选择要保证顶点和其它社区顶点之间边的总数为某个指定的值  $z_{out}$ ,  $p_{in}$  的选择保证顶点的度数  $z_{tot}$  等于 16. 随着  $z_{out}$  由 0 逐渐增大, 社区结构变得越来越模糊, 社区发现变得越发困难.

#### 4.1.2 评测方法

由于网络中的“真实”社区结构是已知的, 给出一个社区划分后, 我们就有可能度量有多少个顶点是正确划分的. 比较发现的社区和网络中已知的社区, 对于发现的社区  $C_F$ , 把它和各个已知社区  $C_R$  进行比较, 找出重叠度最大的那个  $C_R^*$ ,  $C_F$  和  $C_R^*$  重叠的顶点认为是正确划分的顶点. 对发现的各个社区都做类似的处理, 从而得到正确划分的顶点总数占所有顶点数的比率 FVIC (Fraction of Vertices

Identified Correctly). 对于每一个  $z_{out}$ , 我们可以得到一个 FVIC, 从而得到一个 FVIC 关于  $z_{out}$  的函数.

但是, 当发现的社区个数和已知的社区个数不一致时, 上述度量方法存在不小的偏差. 例如, 如果已知社区个数为 4 个, 而实际发现的社区个数远多于 4 个, 且实际发现的社区都是已知社区的子集时, 所有的顶点都被认为是正确划分的, 但这个结果并不是我们希望的.

这里我们还使用了一种更合适的度量方法——归一化后的互信息 (NMI)<sup>[4]</sup>. 定义一个混淆矩阵  $N$ , 矩阵的每一行对应一个“真实”的社区, 每一列对应一个“发现”的社区.  $N$  的元素  $N_{ij}$  表示“真实”社区  $i$  和发现的社区  $j$  重合的顶点个数. 矩阵第  $i$  行的元素之和记为  $N_{i.}$ , 矩阵第  $j$  列的元素之和记为  $N_{.j}$ ,  $N$  的所有元素之和记为  $S$ . 我们用  $C_R$  表示“真实”社区的个数, 用  $C_F$  表示发现的社区个数. 基于信息论, 我们得到发现的社区划分  $F$  和真实的社区划分  $R$  之间相似度的一种度量:

$$NMI(R, F) = \frac{-2 \sum_{i=1}^{C_R} \sum_{j=1}^{C_F} N_{ij} \log \left( \frac{N_{ij} S}{N_{i.} N_{.j}} \right)}{\sum_{i=1}^{C_R} N_{i.} \log \left( \frac{N_{i.}}{S} \right) + \sum_{j=1}^{C_F} N_{.j} \log \left( \frac{N_{.j}}{S} \right)},$$

当发现的划分和网络的真实划分一致时,  $NMI$  取值为 1. 如果发现的划分和网络的真实划分完全独立时,  $NMI$  取值为 0.

在下面的实验中, 我们同时使用 FVIC 和 NMI 两种评测方法来度量一种划分的好坏.

#### 4.1.3 实验结果

实验时,  $z_{out}$  的取值由 0 到 8, 对于  $z_{out}$  的任意一个取值, 为了降低偶然性带来的影响, 我们生成 100 个网络, 对每个网络使用我们的方法发现其中的社区结构, 计算 FVIC 值和 NMI 值, 然后求各自的平均值, 得到对应这个  $z_{out}$  的平均 FVIC 值和平均 NMI 值.

从图 3 中, 我们可以看出, 当  $z_{out}$  不大于 6 时, 我们的方法能正确划分 95% 以上的顶点, 相应的 NMI 值也高达 0.90 以上; 当  $z_{out}$  取值为 7 时, 也能保证有 80% 以上的顶点被正确划分;  $z_{out}$  取 8 时, 此时一个顶点和社区内部顶点的连边数等于它和社区外部顶点的连边数, 此时网络的社区结构变得模糊和难以识别. 另外, 为了说明本文所提方法的有效性, 本文和 Newman Fast 方法<sup>[11]</sup>进行了对比, Newman Fast 方法是目前广泛使用且效果较好的社区发现方法之一. 从图 3 可以看出, 使用 FVIC 和 NMI 中任一种



评测方法,在大部分情况下,我们的方法都比 Newman Fast 算法要好. 这说明,对于有已知社区结构的网络,我们的方法能很好地发现网络的社区结构.

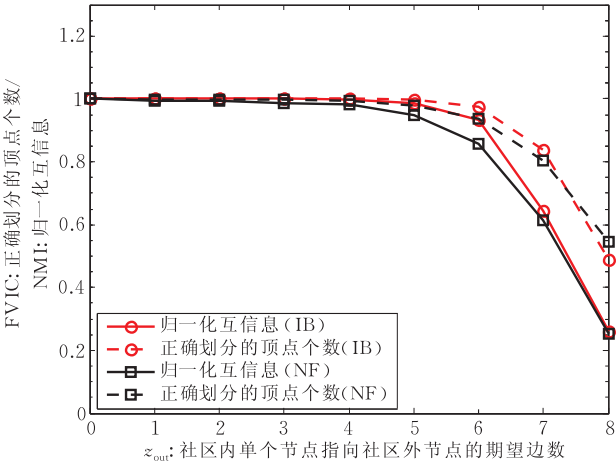


图 3 FVIC 和 NMI 关于  $z_{out}$  的变化曲线图(NF 表示 Newman Fast 算法<sup>[11]</sup>,IB 表示本文的方法)

4.2 在真实网络上的实验

在上一小节中,在计算机产生的网络上测试了本文方法的性能,这一节,我们将该方法应用到一些真实的网络上,验证方法的适用性. 首先,我们在一个小规模的网络上实验,验证方法的有效性;然后在一个有 5000 多个顶点的网络上实验,验证方法的适用性.

4.2.1 Zachary(圣扎伽利)空手道俱乐部网络

Zachary 空手道俱乐部网络(图 4)是社会网络分析的一个经典例子. 该网络取材于美国一所大学的一个空手道俱乐部,俱乐部共有 34 个成员. Zachary Wayne 通过两年的观察,获取了 34 个成员的社会交往关系,从而得到了一个网络. 在这个网络中,每个成员代表一个顶点,如果两个成员在俱乐部内或在俱乐部外有社会交往关系,这两个成员对应的顶点之间有一条边相连. 后来,俱乐部的管理者和

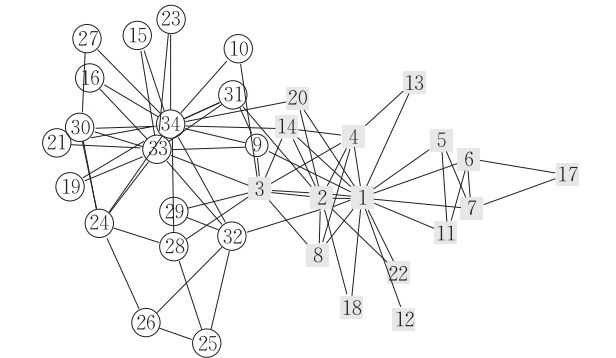


图 4 Zachary 空手道俱乐部网络

老师之间就是否提高俱乐部的费用发生了争论,结果俱乐部分成了两个分别以管理者和老师为中心的俱乐部. 图 4 中,圆形顶点和方形顶点分别代表俱乐部分开后形成的两个新俱乐部的成员.

图 2 展示了我们的方法在 Zachary 空手道俱乐部网络中所发现的社区结构. 图的下半部分是一个 dendrogram,展现了顶点的合并过程,图的上半部分表述了合并过程中相应的 modularity 值的变化趋势. 共发现了 5 个社区,对应的 modularity 值为 0.392,而俱乐部的实际拆分却只有两个部分,但是拆分成两个部分所对应的 modularity 值并不是最优的. 造成算法发现的划分和实际划分不一致另有原因,俱乐部的实际划分是一种指定社区划分个数的社区发现,而我们的算法却是在事先不知道社区个数的情况下进行社区发现的. 从图 2 中,可以看出,如果限定社区个数为 2,我们的算法所发现的划分和实际的划分是几乎一致的,只有一个顶点 3 被错误划分,从图 2 中还可以看出,事实上,顶点 3 和实际划分的每个部分联系紧密程度差不多. 另外,当社区个数限定为 2 时,算法发现的划分所对应的 modularity 值为 0.360,这个值与算法发现的最优值差别并不大.

4.2.2 词联想网络

词联想网络<sup>①</sup>(Word Association Network)是美国南佛罗里达大学(University of South Florida)收集的词自由联想网络,该网络通过问卷调查获得. 共对 5019 个词进行了调查,共有 6000 人参与了此次调查. 调查时,对于每个词,要求参与者写出和该词关系最密切的词. 调查共收到近 75 万个的针对这 5019 个词的应答. 被调查的词称为“提示词”,用户提供的对提示词的应答称为“目标词”. 设  $A$  为任意一个提示词, $B$  为任意一个目标词, $\#G$  表示  $A$  作为提示词出现的次数, $\#P$  表示  $B$  作为提示词  $A$  的目标词而出现的次数,那么提示词  $A$  到目标词  $B$  的前向强度为

$$FSG = \frac{\#P}{\#G},$$

如果  $B$  同时也是一个提示词,那么  $A$  到  $B$  的后向强度(BSG)等于  $B$  到  $A$  的前向强度.

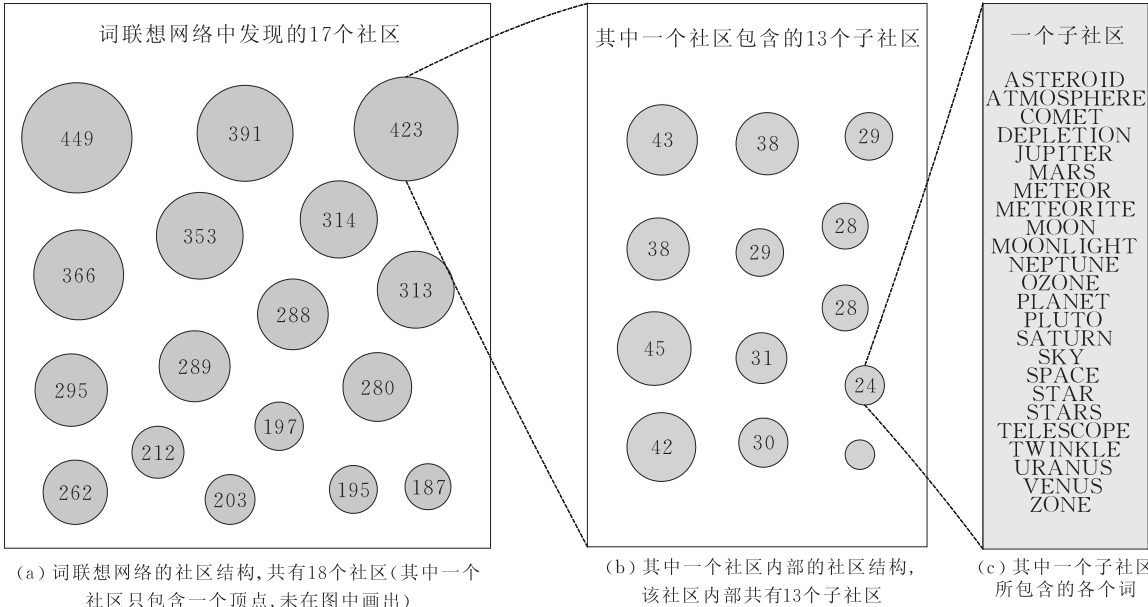
目标词不一定在 5019 个提示词中出现,我们这里只保留在提示词集合中出现的目标词,这样一来,

① The University of South Florida word association, rhyme, and word fragment norms. <http://w3.usf.edu/FreeAssociation/>

每一个词既是提示词,又是目标词. 我们定义词  $A$  和词  $B$  的关联强度为  $A$  到  $B$  的前向强度(等于  $B$  到  $A$  的后向强度)加上  $A$  到  $B$  的后向强度(等于  $B$  到  $A$  的前向强度).

5019 个词看成是 5019 个顶点,如果两个词之间的关联强度大于某个阈值(本文取值为 0.025),这两个词对应的顶点存在一条边. 这样我们就得

到了一个有 5019 个顶点的无向无权简单图. 我们的方法共发现了 18 个社区(图 5(a)),此时对应的 modularity 值为 0.423,这表明词联想网络具有良好的社区结构. 其中有一个社区只包含一个顶点(未在图 5(a)中画出),分析词联想网络,发现这个顶点是一个孤立顶点,因此它单独构成一个社区是合适的.



注:圆圈内的数字表示社区的大小(也就是社区所包含的顶点个数),圆圈的大小也反映了社区的大小.

图 5

为了更清晰地分析各个社区的内部结构,使用第 3 节中的方法,进一步分别对 17 个社区(只包含一个顶点的社区除外)进行了社区发现. 图 5(b)展示了其中一个社区的内部结构,对这个社区进行重新社区发现时,我们得到的 modularity 值高达 0.691,这说明这个社区内部具有非常明显的社区结构. 事实上,17 个社区都具有明显的社区结构,这表明词联想网络的社区结构是有层次性的. 不同层次的社区对应着网络不同粒度的划分. 图 5(c)向我们展示了其中一个社区内部的子社区所包含的各个词,这些词大多是和天文学有关的. 其它各个子社区所包含的词也具有明显的类别特征,这里不一一列举.

值得一提的是,我们发现的社区大小不满足 power-law 规则<sup>[1-2]</sup>. 经过分析,我们发现 5019 个顶点所构成的词联想网络的顶点度分布呈泊松分布,而一般情况下,复杂网络的顶点度分布是满足 power-law 规则的,这可能是造成所发现的社区大小不满足 power-law 规则的原因.

本文提出的方法能够发现词联想网络中不同粒度的社区,且这些社区都是有明确意义的社区,这对我们理解词联想网络的结构是非常重要的,而理解词联想网络的结构对使用词联想网络是非常有意义的.

### 5 结论和下一步的工作

本文从一个新的角度重新思考社区发现问题,认为社区结构是网络结构规则性的一种体现,而发现网络中的社区结构是去掉那些不重要的细节而保留网络的整体拓扑特征,这一认识是本文的基础. 基于上述认识,社区发现问题可以看成是对网络拓扑的有损压缩. 社区发现目标是寻找一种网络结构的高效压缩表示,这种压缩能够尽可能多地反映原始网络中的结构规则性.

进而,本文在信息论的框架下,提出了一种基于信息瓶颈的社区发现方法. 该方法从信息论的角度为社区发现问题提供了一种理论解释,即网络社区



结构的最优表示是原始网络的压缩比和反映原始网络结构规则之间的最优折中。

对于有向网络的社区发现, 现有的社区发现方法是忽略网络中边的方向, 进而使用无向网络社区发现的方法来解决。事实上, 这种做法损失了原始有向网络的许多信息, 边的方向自身向我们提供了网络结构的一种反映。基于本文使用的网络变换方法, 可以方便地把有向网络转换成一个二部图网络, 该二部图网络的左部反映有向网络的出边特征, 右部反映有向网络的入边特征。然后, 使用本文提出基于信息瓶颈的社区发现方法, 可以很方便地发现分别基于出边或入边特征的社区结构。

真实世界中许多网络自身就是一个二部图网络, 和现有方法相比, 本文提出的方法可以很方便地应用到这类网络的社区发现中。

另外, 许多网络中存在不止一类顶点, 顶点之间也存在不止一种关系, 这种网络被称为异质网络。例如, 在文献及其作者构成的网络中, 存在文献和作者两类顶点, 也存在文献和作者之间的署名关系, 文献之间的引用关系。对于这种异质网络, 如何发现其中的社区结构是一个有着重要研究意义和应用价值的问题。我们下一步的工作将主要致力于解决这一问题。

**致 谢** 中国科学院计算技术研究所信息智能与信息安全中心社会计算与 P2P 课题组的吕建明、陈国耀、陈友等同事对本文的完成提出了很多有益的建议, 互联网挖掘与搜索组的杜伟夫对本文的修改提供了帮助, 在此一并表示感谢。最后, 特别感谢评审老师细致耐心的审查!

## 参 考 文 献

- [1] Albert R, Barabási A-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 2002, 74(1): 47-97
- [2] Newman M E J. The structure and function of complex networks. *SIAM Review*, 2003, 45(2): 167-256
- [3] Newman M E J. Detecting community structure in networks. *European Physical Journal B*, 2004, 38(2): 321-330

- [4] Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. *Journal of Statistical Mechanics*, 2005, P09008
- [5] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 1970, 49(2): 291-307
- [6] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, 435(7043): 814-818
- [7] Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, 69(2): 026113
- [8] Fortunato S, Latora V, Marchiori M. A method to find community structures based on information centrality. *Physical Review E*, 2004, 70(5): 056104
- [9] Pons P, Latapy M. Computing communities in large networks using random walks//*Proceedings of the 20th International Symposium on Computer and Information Sciences*. Lecture Notes in Computer Science 3733. Springer, New York, 2005: 284-293
- [10] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 2006, 74(3): 036104
- [11] Newman M E J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, 69(6): 066133
- [12] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Physical Review E*, 2004, 70(6): 066111
- [13] Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Physical Review E*, 2006, 74(1): 016110
- [14] Duch J, Arenas A. Community detection in complex networks using extremal optimization. *Physical Review E*, 2005, 72(2): 027104
- [15] Rosvall M, Bergstrom C T. An information-theoretic framework for resolving community structure in complex networks//*Proceedings of the National Academy of Sciences of the United States of America*, 2007, 104(18): 7327-7331
- [16] Ziv E, Middendorp M, Wiggins C. An information-theoretic approach to network modularity. *Physical Review E*, 2005, 71(4): 046117
- [17] Tishby N, Pereira F C, Bialek W. The information bottleneck method, 1999, *HarXiv:physics/0004057v1*
- [18] Slonim N, Tishby N. Agglomerative information bottleneck//*Proceedings of the Neural Information Processing Systems(NIPS-99)*. The MIT Press, 1999, 12: 617-623



**SHEN Hua-Wei**, born in 1982, Ph.D. candidate. His research interests include social computing, complex networks, information retrieval.

**CHENG Xue-Qi**, born in 1971, Ph.D., professor. His research interests include network information security, large-scale information retrieval and knowledge mining, peer-to-peer computing.

**CHEN Hai-Qiang**, born in 1979, Ph.D. candidate. His research interests include social computing, complex networks, information retrieval.

**LIU Yue**, born in 1971, Ph.D. , associate professor. Her research interests include Web search and mining, anal-

ysis on complex networks and social computing, distributed system.

**Background**

Research on complex networks attracts considerable attentions from many scientific fields. The main focus is on three aspects: The statistical properties of complex networks, the model of the evolution of complex networks, the dynamics of complex networks.

Community structure is a common and important property of complex networks. It forms an intermediate level between the microscopic and macroscopic description of networks. It provides the knowledge about the relation between the function and structure of complex networks. Thus, community detection plays a crucial role in the research on complex networks.

Recent research on community detection devotes most efforts to unipartite undirected networks. Only few works concern the directed networks and multipartite networks. This paper aims to address the problem of community detection for bipartite networks. The method proposed in this paper is also applicable to unipartite networks through a projection used in this paper.

The research is a part of "Research on Community Identification and Community Evolution on the Web 2.0", which is supported by MSRA IST 2007 (FY07-RES-THEME-067). The latter aims to the sociality emerged in the Web 2.0. It is useful to help us understand the rules that govern the networks formed in rapidly emerging Web 2.0 applications. The identification of community structure is a fundamental task of this research. It is the basis of research on community evolution in Web 2.0.

The authors' recent studies mainly focus on the community detection of complex networks. The authors have made some in-depth research on the measure of community structure, the method to uncover the hierarchical and overlapping community structure in large networks, and its application on expert finding.

In addition, a case study on the Douban (www.douban.com) is carried out by the research team. And the corresponding results are published in the 17th International World Wide Web Conference.