面向带属性复杂网络的鲁棒、强解释性社团发现方法

金 弟[1)] 刘子扬[1)] 贺瑞芳[1)] 王 啸[2)] 何东晓[1)]

[1)](天津大学计算机科学与技术学院 天津 300350)

[2)](北京邮电大学计算机学院 北京 100876)

摘 要 语义社团发现包含两重含义,即提升社团发现质量、精确标注社团语义.传统的社团发现方法通常假设网络拓扑与节点内容共享同一社团结构.然而,在许多真实网络(如社交网络)中,网络拓扑与节点内容所对应的社团结构通常并不一致.例如,在 Twitter 网络中,社交链接能够直接地反映出哪些用户聚集在一个社团之中,而每位用户却会产生迥异的、杂乱无章的内容信息.针对该问题,作者基于非负矩阵分解框架提出了一个鲁棒、强解释性的新的社团发现模型(Robust and Strong Explanatory Community Detection,RSECD).该模型创新地采用一个带先验的转移概率矩阵来刻画网络社团与内容类簇之间的内在关联.在实验中,作者首先采用一组人工网络验证 RSECD 的有效性和鲁棒性;进而在 7 个真实网络上,与 8 种代表性社团发现方法进行量化比较,结果显示 RSECD 的检测精度比性能最高的对比算法高 6%~14%,进一步体现了其优越性;最后,通过一个在线音乐社交网络上的实例分析,作者又验证了 RSECD 对于所发现社团具有很强的可解释能力.

关键词 社交网络;社团发现;语义;非负矩阵分解;转移概率

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2018.01476

A Robust and Strong Explanation Community Detection Method for Attributed Networks

JIN Di[1)] LIU Zi-Yang[1)] HE Rui-Fang[1)] WANG Xiao[2)] HE Dong-Xiao[1)]

[1)](*School of Computer Science and Technology, Tianjin University, Tianjin* 300350)

[2)](*School of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract Semantic community identification contains two aspects, i.e., to improve the accuracy of community detection and to annotate the semantics of communities precisely. Traditional community detection methods always assume that the network topology and node contents share the same community memberships. However, this assumption does not always hold in many real-world networks. For example, in a Twitter network, social links usually directly reflect which users gather into a community, while users may generate diverse and disordered content information. To solve this problem, it is necessary to extract useful content information to assist topology information in detecting out more actual and accurate communities. In this paper, we carefully rethink the relationships between community structure, topic cluster, network topology and node contents, and propose a novel generative model different from the traditional generative model. In the new generative model, we logistically give a more reasonable explanation about the relationships between community structure, topic cluster, network topology and node contents.

收稿日期:2017-10-13;在线出版日期:2018-04-06.本课题得到国家重点研发计划(2017YFC0820106)、国家自然科学基金(61502334,61772361,61702296 和 61472277)、天津大学北洋学者-青年骨干教师项目(2017XRG-0016)资助.**金 弟**,男,1981 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为人工智能、复杂网络分析和网络社团检测. E-mail: jindi@tju.edu.cn. **刘子扬**,男,1994 年生,学士,主要研究方向为复杂网络分析和网络社团检测. **贺瑞芳**,女,1979 年生,博士,副教授,主要研究方向为自然语言处理、信息检索. **王 啸**(通信作者),男,1988 年生,博士,主要研究方向为网络社团检测、网络嵌入. E-mail: wangxiao_cv@163.com. **何东晓**,女,1984 年生,博士,副教授,主要研究方向为网络社团检测、数据挖掘.

Under the drive of the new generative model, we design a new community detection method, referred to as Robust and Strong Explanatory Community Detection (RSECD). To be specific, based on nonnegative matrix factorization (NMF), we are able to obtain the community membership matrix for network topology and cluster membership matrix for node contents. More importantly, there exists some latent relationship between network communities and content clusters (with semantics), thus we innovatively introduce a transition matrix with a suitable prior to describe this relationship. As a result, even though the content information does not match with topology information, our method can still obtain accurate detection results by using the transition matrix with a suitable prior. At last, we put network topology, node content and transition matrix into a unified NMF framework and optimize them altogether by designing effective updating rules in order to achieve an integral balance of them. Furthermore, we analyze RSECD's calculational complexity after taking into account the sparsity of the adjacency matrix and attribute matrix. To justify our approach's effectiveness and robustness, we conduct extensive experiments. Firstly, we employ the Newman's model to generate artificial benchmark networks, then we use the generated networks to analyze the parameter in our objective function and verify that RSECD can solve topology and content's mismatch problem in the network well. The results of artificial networks experiment show our approach's strengths, i. e., effectiveness and robustness. Next, we compare our method with eight state-of-the-art community detection approaches on seven real-world networks. The experimental results show that RSECD achieves up to 6%—14% lift in comparison with the best baseline algorithm under four different kinds of community detection metrics, which further demonstrate RSECD's superior performance. We also report the running time of RSECD and other baseline algorithms spending on seven real-world networks. The results show that RSECD's time cost is far less than the average running time of other baselines on all real-world networks. Finally, in order to validate RSECD's strong interpretability to detected communities, we use a case study on a musical social network to semantically explain the hidden meanings of some topics and tell the 'true stories' behind communities.

Keywords　social networks; community detection; semantics; nonnegative matrix factorization; transition possibility

1　引　言

网络数据分析在许多领域(如社会学与计算机科学等)都是一门非常新颖、非常重要的学科.网络通常可表示为节点以及节点间的连接,且总是出现在各类场景中[1].真实网络通常具有一些共享的统计特性,其中一个典型的特性即社团结构——同一社团内部的节点连接紧密,不同社团之间的节点连接稀疏[2].譬如,在 Facebook 网络中,有相同兴趣爱好的用户通常会聚集在一起构成一个社团,但在这些社团之间,用户的连接就相对稀少.社团结构反映了一个网络的基本功能模块,它可以帮助人们更好地理解和分析网络是如何运行的.

社团发现是挖掘一个网络中社团结构的有效方法,并在近十年内得到了迅速发展.目前已有许多不同类型的网络社团发现方法被提出,譬如:凝聚/分裂算法[3]、基于模块度优化的算法[4]、谱方法[5]、基于标签传播的算法[6]等.另外,众所周知,一个节点可能会属于多个社团(即重叠社团结构[7]).因此,许多用于发现重叠社团的方法也逐渐被提出,譬如:基于团渗理论的方法[8]、基于局部扩张与优化的方法[9]、基于概率社团的方法[10-11]、基于链接划分的方法[12]等.在探究网络特性时,人们通常认为社团结构对网络拓扑的形成起着非常重要的作用,社团发现即是利用网络拓扑信息反推出其中蕴含的社团结构.然而当我们考虑真实网络时,在网络的生成过程中,总会由于一些扰动因素产生与社团结构不一致的拓扑"噪声"信息,这些"噪声"很可能会导致错误的社团划分结果,进而增加了社团发现的难度.因

此,当前社团发现方法存在的普遍问题就是:难以有效规避拓扑"噪声"对社团划分的干扰作用.

利用多元信息进行社团发现可有效弥补仅利用单一信息划分社团的缺陷,所以除网络拓扑以外,一些社团发现方法还将"作为一个节点个体信息的"节点属性(或称节点内容)考虑进来[13-16],通过融合拓扑和内容两类信息更好地进行社团发现,从而降低单一信息中的"噪声"因素对社团发现的干扰作用.当前融合拓扑与内容信息的社团发现方法通常认为网络拓扑和节点内容共享同一社团结构,然而在许多真实网络中,情况并非如此.譬如:在 Twitter 网络中,社交关系通常直接反映了用户群体,而用户却可产生多种多样的内容信息.因此拓扑中所蕴含的网络社团与内容中所蕴含的主题类簇很可能并不一致;在一些特殊情况下,它们甚至可能蕴含着完全相反的冲突信息.然而,融合拓扑与内容进行社团发现的方法通常忽视了网络拓扑和节点内容针对社团结构匹配不一致的情况,因此在任意情形下均可有效地鲁棒融合这两类信息,从而更好地进行社团发现.

最近,研究者们还进一步意识到:融合拓扑与内容的社团发现不仅应精确发现社团结构,还需利用节点内容中富含的语义信息对社团进行解释.因此,融入节点内容的优势不仅在于可弥补单独利用网络拓扑进行社团划分的缺陷,还可对所发现的社团进行语义解释,以拓展其应用范围[17-19].

总而言之,如何解决网络拓扑和节点内容在社团结构上匹配不一致的问题是融合多元信息进行社团发现的关键;此外,在精确发现社团结构的同时,我们还应对所发现的社团进行语义解释.为解决以上问题,本文从一个新的角度重新审视了网络拓扑和节点内容在社团发现中的作用,提出了一个鲁棒、强解释性的社团发现方法(Robust and Strong Explanation Community Detection,RSECD).在该方法中,我们首先将网络拓扑和节点内容视为影响每个节点社团隶属关系的两个相对独立的侧面,其中,网络拓扑直接影响节点的社团划分,而内容信息则直接影响节点的类簇划分(与"社团"含义相类似,同一"类簇"内节点的属性相似,不同"类簇"间节点的属性相异).进而针对拓扑和内容信息,基于非负矩阵分解模型(Nonnegative Matrix Factorization,NMF),分别刻画出每个节点的社团隶属关系矩阵和类簇隶属关系矩阵.这样,我们分别构建了仅利用拓扑信息的社团划分子模型,和仅利用内容信息的类簇划分子模型.更重要的是,网络社团与内容类簇之间存在着一些内在隐关系.为了刻画这种本质联系,我们通过引入带先验的转移概率矩阵,来描述网络社团和主题类簇之间的精准映射关系;基于该映射,我们在社团与类簇具有不同匹配程度的情形下,均可得到鲁棒的社团发现结果.最后,我们将拓扑、内容、转移概率融合到同一个模型中,三者协同更新优化,达到拓扑、内容、转移矩阵整体平衡的状态.在迭代更新中,我们基于 Oja 学习准则[20]推导出有效的乘性更新公式对 RSECD 模型进行优化分析.

在人工网络实验中,我们针对不同网络分析了目标函数中参数的最优取值,并通过与两类典型算法的比较验证了 RSECD 方法的鲁棒性;接着,我们在 7 个真实网络上做了量化实验,并分别使用无重叠和重叠社团评价标准将 RSECD 与 8 个代表性社团发现算法进行比较分析,实验结果显示 RSECD 在所有对比中均可显著提升社团发现性能,进一步验证了该方法的鲁棒性;最后,我们给出了一个定性的实例分析实验,展示了 RSECD 可以对所发现的社团进行精准语义解释.

本文的主要贡献如下:

(1)我们采用非负矩阵分解模型将网络拓扑与节点内容两类信息融入到一个统一的框架中,将社团检测问题公式化并精确求解;

(2)我们在 RSECD 模型中引入带先验的转移概率矩阵,用其表示网络社团和内容类簇之间的内在隐关系.转移矩阵通过引导性的映射作用,很好地解决了网络拓扑与内容信息对于社团结构匹配不一致的问题.使得在拓扑与内容之社团结构在不同匹配程度的情形下,算法的精度都依然保持相对稳定,从而可更加鲁棒地融合这两类信息;

(3)转移矩阵不仅能够更准确地发现社团结构,还可实现对所发现的社团进行完整、丰富的语义解释.具体来说,已有方法[13-18]通常假设社团和主题一一对应,故在解释社团时仅利用单一主题信息;与这些方法不同的是,本文通过转移概率矩阵揭示了网络社团和主题类簇之间更为本质的映射关系,可采用多个主题解释社团语义,因而 RSECD 模型具有更强的解释性.

2 相关工作

目前已有许多仅考虑网络拓扑的社团发现方法被提出[3-10,12,21].譬如:基于凝聚或分裂机制的层次聚

类方法[3]. 此外，模块度优化方法（如 Fanuel 等人[5]提出的谱优化）通过优化模块度函数可发现较为准确的社团结构. Yang 等人[4]针对不同网络图（如符号网络），使用模块度修正的思想解决了复杂图上的社团发现问题，具体做法为：将一个网络映射为拉普拉斯矩阵，并计算该矩阵的特征向量，进而使用谱方法找到每个节点所属的社团.

随着复杂网络研究的深入，人们愈发注意到网络中文本内容所蕴含的有益价值. 一些融合网络拓扑和内容信息的社团发现方法被相继提出. 譬如：van Laarhoven 等人[13]提出一个融合网络结构和内容信息的子图重叠聚类算法，他们使用期望最大化（Expectation Maximization，EM）算法来优化定义好的似然函数，生成固定的候选子图，之后又对边信息进行 k-均值聚类，从而获得重叠社团结构. Pei 等人[14]基于社交网络中的链接信息和由用户产生的内容信息，提出一种基于非负矩阵分解模型的聚类框架，可以捕捉到“用户-单词-信息”的三部图关系；同时，还分别对用户相似度、信息相似度和用户交互进行建模，从而提高了社团发现的精度. Newman 等人[15]考虑到节点属性信息的重要性，设计了一种贝叶斯社团检测模型，能够更加有效地利用节点属性信息来发现网络社团结构.

最近，研究人员也开始意识到社团检测问题不仅应该准确地发现社团结构，还应能够充分利用文本内容所蕴含的语义信息，对已发现的社团进行解释. 社团解释的好处在于：它既能反映出节点形成社团的原因，也能帮助人们更好地理解社团的含义和功能. 社团解释可通过多种形式实现，其中较为常用的是社团词云图和社团画像. 前者通过图像的形式生动地展示了社团中不同属性词的重要程度，以此实现解释社团的目的；后者通过文本的形式将出现频率较高的单词作为社团的语义解释. 譬如：Akbari 等人[17]使用非负矩阵分解模型将社团检测和用户画像两个任务整合到同一个模型中，并通过一个线性算子来整合用户画像，从而获得社团画像. Cai 等人[18]基于用户发表的内容信息和用户之间的链接信息，提出了一个可同时进行社团画像和社团检测任务的统一模型. Wang 等人[22]提出的 SCI 方法使用非负矩阵分解模型将拓扑和内容信息整合到同一个模型框架中，但是它的缺点是忽略了拓扑和内容针对社团结构不匹配的问题. He 等人[23]提出的 NEMBP 方法：设计了一个包含社团结构和语义信息两部分的生成模型，采用协同学习的策略，通过使用 EM 算法和置信传播联合训练模型，实验显示了该方法不仅能准确地检测出社团结构，还能给出合理的社团解释. 总体来说，整合社团检测和社团解释两个任务的社团发现方法具有更高的实用价值.

然而，以上方法在融合网络拓扑和内容信息时，通常假设拓扑与内容共享同一社团结构，但是当拓扑与内容在社团结构上不匹配时，却难以有效地鲁棒融合这两类信息. 譬如：Wang 等人[22]提出的 SCI 方法仅通过选取一组较好的权重系数来融合这两类信息，当拓扑和内容的社团结构不匹配时，这一做法会明显失效. 因此，新的融合方法需要解决的关键问题就是：在网络拓扑和节点内容针对社团结构匹配不一致时，仍能鲁棒地融合这两类信息，并可对所发现的社团进行更加精确地解释.

3　社团发现模型 RSECD

3.1　RSECD 模型的设计思想

网络中拓扑结构与文本内容的匹配程度可分为三类，即完全一致、部分匹配和完全不同. 然而，已有的融合网络拓扑和节点内容的方法（如 SCI 方法[22]）通常忽视了后两种情况，即拓扑和内容信息匹配不一致的情形，单纯假定拓扑和内容所表达的社团信息完全一致. 基于此，我们做出如下改进：将网络拓扑和节点内容视为两个相对独立的要素，分别影响着节点的社团隶属关系和节点的类簇隶属关系；同时借助带先验的转移概率矩阵，刻画出网络社团与内容类簇间的内在关联.

为了更好地说明新模型的设计思想，我们采用逻辑示意图的形式将传统模型与本文思路进行对比. 如图 1(a)所示，融合拓扑和内容的社团发现方法一般认为：存在一个统一的社团结构，蕴含于网络拓扑与节点内容这两类信息之中，即网络中的拓扑结构和节点内容共享同一社团结构. 然而在许多真实网络中，拓扑结构和节点内容中所蕴含的社团结构通常存在着匹配不一致的情形，而图 1(a)中所示的这种逻辑关系难以解决该问题. 因而需要一类新的模型，能够在不同情形下均可鲁棒地融合拓扑和内容信息，有效进行社团检测. 具体来说，当拓扑和内容在社团结构上匹配一致时，它们可通过相互补充，降低单一信息的噪声干扰，以更加准确地发现社团结构；当拓扑和内容的匹配不佳时，内容信息中有益的部分仍可与拓扑信息共同发挥作用，为社团发现提供帮助；当内容与拓扑完全不匹配时，算法可自

动舍弃冲突信息,得到仅利用网络拓扑的社团结果.

这里我们从一个新的角度重新审视了拓扑、内容与社团的内在关系,设计出满足上述鲁棒性需求的 RSECD 模型,其逻辑关系示意图如图 1(b)所示.

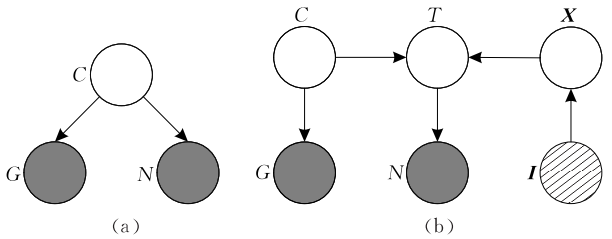


图 1 传统模型与文中 RSECD 模型的示例性比较(这里为了更好地阐明不同模型所描述的数据生成过程,我们采用逻辑图来表示变量间的依赖关系.(a)已有模型的逻辑结构;C 表示社团结构,G 表示网络拓扑,N 表示节点内容;该模型对应了"在真实网络数据下网络拓扑 G 和节点内容 N 共享同一社团结构 C"的情形.(b)新模型 RSECD 的逻辑结构;网络拓扑 G 仍蕴含社团结构 C,而节点内容 N 蕴含主题类簇 T,同时主题类簇 T 由社团结构 C 和转移矩阵 $\boldsymbol{X}$ 共同决定;这里作为先验的单位阵 $\boldsymbol{I}$ 在信息融合中起着关键性的引导作用)

下面 3.2 节给出了网络中的一些符号说明;3.3 和 3.4 节分别介绍了在 RSECD 模型中网络拓扑和节点内容是如何影响节点的社团划分和类簇划分的;3.5 节介绍了如何通过精心设计一个带先验的转移概率矩阵,有效刻画网络社团和内容类簇间的内在隐关系;最后,3.6 节给出了 RSECD 模型的统一形式和鲁棒性分析.

3.2 符号说明

对于一个包含 n 个节点、e 条边的无向无权网络 G,我们用 a_{ij} 表示节点 i 和 j 之间的连接关系:如果节点 i 与 j 之间有边,则 $a_{ij}=1$,否则 $a_{ij}=0$;所有节点对儿之间的 a_{ij} 构成了一个二进制的邻接矩阵 $\boldsymbol{A}\in R^{n\times n}$. 我们将每个节点的属性表示为一个 m 维的二进制向量 $\boldsymbol{s}_i$,整合所有节点的属性信息即形成了一个属性矩阵 $\boldsymbol{S}\in R^{n\times m}$.

此时,网络社团检测与解释的任务可表述为:在网络社团数目 k 和主题类簇数目 k' 给定的情况下,基于网络拓扑和节点内容找到 k 个社团和 k' 个主题,同时推断出社团与主题间的内在关联,以得到每个社团的主题语义. 由于对比算法均假设网络社团与主题类簇一致且相同($k=k'$),因此本文亦假设 $k=k'$. 然而文中算法也同样适用于社团数与类簇数不相等的情形.

3.3 基于网络拓扑的子模型

网络拓扑模型部分应满足以下直观特性:

(1)如果两个节点属于同一社团,那么它们很可能会相连接;

(2)如果两个节点有相似的社团隶属关系,那么它们也很可能会相连接.

我们将节点 i 属于社团 c 的倾向定义为 u_{ic},于是得到了一个覆盖网络中所有节点的社团隶属关系矩阵 $\boldsymbol{U}=(u_{ic})_{n\times k}$. 基于第一个特性,我们可采用 $u_{ic}\times u_{jc}$ 来表示社团 c 中节点 i 与 j 之间连接边数的期望. 基于第二个特性,我们可以得到整个网络中节点 i 与 j 之间连接边数的期望 $\sum_{c=1}^{k}u_{ic}u_{jc}$.

将以上结果推广到所有节点,则可得到下述矩阵形式的损失函数:

$$\min_{\boldsymbol{U}\geq 0}\|\boldsymbol{A}-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\|_F^2 \tag{1}$$

3.4 基于节点属性的子模型

我们将类簇 t 包含属性 q 的倾向定义为 c_{qt},将节点 i 属于类簇 t 的倾向定义为 v_{it},于是得到属性的类簇隶属关系矩阵 $\boldsymbol{C}=(c_{qt})_{m\times k}$ 以及节点的类簇隶属关系矩阵 $\boldsymbol{V}=(v_{it})_{n\times k}$. 此外,我们将节点 i 包含属性 q 的倾向定义为 s_{iq},而 s_{iq} 正是已知属性矩阵 $\boldsymbol{S}$ 中的元素. 我们认为节点属性模型应满足如下直观特性——若节点 i 属于多个类簇 $t_1,t_2,\cdots$,且类簇 $t_1,t_2,\cdots$ 都包含属性 q,则节点 i 也应以更大的概率包含属性 q,即 $s_{iq}=\sum_{t=1}^{k}v_{it}c_{tq}$.

将这一结论推广到所有节点,则可得到如下矩阵形式的损失函数:

$$\min_{\boldsymbol{C}\geq 0,\boldsymbol{V}\geq 0}\|\boldsymbol{S}-\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\|_F^2 \tag{2}$$

3.5 基于转移概率的子模型

转移概率是马尔可夫链中的重要概念,用于表示由一个状态转移到下一个状态的概率. 这里我们引入转移概率来表示"网络社团"与"主题类簇"间的映射关系. 具体来说,我们将"社团"和"类簇"视为两个不同的状态空间,转移过程发生在这两个状态空间之间,则转移概率表示从某一社团转移到某一类簇的概率. 我们将从社团 c 转移到类簇 t 的概率定义为 x_{ct},从社团 c 转移到任一类簇的概率向量定义为 $\boldsymbol{x}_c$($\boldsymbol{x}_c$ 中每一元素非负且元素之和为 1),从任意社团转移到任意类簇的所有概率定义为矩阵 $\boldsymbol{X}$. 此外,为了对拓扑与内容的信息融合产生有效地引导作用,我们采用单位阵 $\boldsymbol{I}$ 作为 $\boldsymbol{X}$ 的先验,则得到以下损失函数:

$$\min_{\boldsymbol{X}\geq 0}\|\boldsymbol{U}\boldsymbol{X}-\boldsymbol{V}\|_F^2+\|\boldsymbol{X}\boldsymbol{1}_k^{\mathrm{T}}-\boldsymbol{1}_k^{\mathrm{T}}\|_F^2+\|\boldsymbol{I}-\boldsymbol{X}\|_F^2 \tag{3}$$

其中,$\boldsymbol{1}_k^{\mathrm{T}}\in R^{k\times 1}$ 且 $\boldsymbol{1}_k^{\mathrm{T}}$ 中的所有元素值为 1,而 $\boldsymbol{I}$ 为单位矩阵.

3.6　统一模型形式

整合式(1)到(3)中的目标，我们得到总的损失函数如下：

$$\min_{\substack{\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0,\\ \boldsymbol{C}\geq 0,\boldsymbol{X}\geq 0}} L=\|\boldsymbol{A}-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\|_F^2+\alpha\|\boldsymbol{S}-\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\|_F^2+\|\boldsymbol{U}\boldsymbol{X}-\boldsymbol{V}\|_F^2+\|\boldsymbol{I}-\boldsymbol{X}\|_F^2+\|\boldsymbol{X}\boldsymbol{1}_k^{\mathrm{T}}-\boldsymbol{1}_k^{\mathrm{T}}\|_F^2 \quad (4)$$

其中，α 是一个平衡网络拓扑和节点内容的参数.

RSECD 模型主要包含 3 部分：生成网络拓扑 $\boldsymbol{A}$(式(1))；生成节点内容 $\boldsymbol{S}$(式(2))；从网络社团到主题类簇的引导性转移过程(式(3)). 其中，前两部分在目标函数中起主导作用，而第三部分则对鲁棒融合拓扑与内容起着关键的制约与引导作用. 由于单位矩阵 $\boldsymbol{I}$ 为转移概率矩阵 $\boldsymbol{X}$ 之先验，因此若不考虑前两部分，则 $\boldsymbol{X}$ 为单位阵时目标函数最优. 若考虑前两部分，当社团与类簇匹配良好时，$\boldsymbol{X}$ 几近单位矩阵，此时只有式(1)和(2)发挥作用，拓扑和内容两类信息相互增强、补充，可起到“1+1>2”的效益，从而提升社团发现质量；当社团与类簇匹配不佳时，仍可借助 $\boldsymbol{X}$ 的映射与牵引作用，尽可能从内容中提取有益的信息，提升社团发现质量；当内容对社团完全无用时，则社团与类簇几近正交，$\boldsymbol{X}$ 将近似于随机矩阵，这时仍可得到近似单独利用拓扑信息的社团划分结果，即仅相当于式(1)在发挥作用. 此外，由于前两部分(生成网络拓扑 $\boldsymbol{A}$ 与生成属性信息 $\boldsymbol{S}$)在目标函数中起主导作用，因此可得到紧凑的社团与主题；又由于我们可学习得出这两者间的隐关系 $\boldsymbol{X}$，故可借助 $\boldsymbol{X}$，采用多个紧凑的主题对社团进行解释. 因此，该方法既具有鲁棒性，又具有对社团的强解释能力.

4　模型优化

由于式(4)中的目标函数是非凸的，因而难以求出全局最优解，针对这一问题，我们的解决方案是：通过使用最大最小化框架(Majorization-Minimization)[24]求出式(4)的局部最优解. 在该框架下，整个目标函数的优化问题被分解为 4 个子问题，即：(1)固定 $\boldsymbol{V}$、$\boldsymbol{C}$ 和 $\boldsymbol{X}$，迭代更新 $\boldsymbol{U}$；(2)固定 $\boldsymbol{U}$、$\boldsymbol{C}$ 和 $\boldsymbol{X}$，迭代更新 $\boldsymbol{V}$；(3)固定 $\boldsymbol{U}$、$\boldsymbol{V}$ 和 $\boldsymbol{X}$，迭代更新 $\boldsymbol{C}$；(4)固定 $\boldsymbol{U}$、$\boldsymbol{V}$ 和 $\boldsymbol{C}$，迭代更新 $\boldsymbol{X}$. 在 Majorization-Minimization 框架下，上述迭代策略可以保证目标函数值在迭代过程中严格非增、且保持所有参数非负(在初始种子为非负的条件下)，下面我们分别给出以上 4 个子问题的具体迭代公式.

4.1　$\boldsymbol{U}$ 的迭代

在更新矩阵 $\boldsymbol{U}$ 时，我们首先给出目标函数(4)中与 $\boldsymbol{U}$ 相关的函数：

$$\min_{\boldsymbol{U}\geq 0} L(\boldsymbol{U})=\|\boldsymbol{A}-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\|_F^2+\|\boldsymbol{U}\boldsymbol{X}-\boldsymbol{V}\|_F^2 \quad (5)$$

由于对任意矩阵 $\boldsymbol{M}$ 都有 $\|\boldsymbol{M}\|_F^2=\mathrm{tr}(\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}})$，因而，式(5)可转换为

$$L(\boldsymbol{U})=\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}+\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}-\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}+\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}) \quad (6)$$

进而，我们基于式(6)对矩阵 $\boldsymbol{U}$ 求偏导，可以得到

$$\frac{\partial L(\boldsymbol{U})}{\partial \boldsymbol{U}}=-2(\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{A})\boldsymbol{U}+2(\boldsymbol{U}\boldsymbol{X}-\boldsymbol{V})\boldsymbol{X}^{\mathrm{T}}+4\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} \quad (7)$$

为了减少计算成本，我们采用基于 Oja 学习准则[20]的乘性更新法则. Oja 准则将梯度 $\nabla_U L(\boldsymbol{U})$ 分解为两个集合：所有正项之和(∇_+)与所有负项之和(∇_-)，即

$$\nabla_U L(\boldsymbol{U})=\nabla_+-\nabla_- \quad (8)$$

这样，$\boldsymbol{U}$ 的更新公式可写为

$$\boldsymbol{U}_{\mathrm{new}}=\boldsymbol{U}_{\mathrm{old}}\frac{\nabla_-}{\nabla_+} \quad (9)$$

在式(7)中，负项包括 $2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}$、$2\boldsymbol{A}\boldsymbol{U}$ 和 $2\boldsymbol{V}\boldsymbol{X}^{\mathrm{T}}$，正项包括 $2\boldsymbol{U}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 和 $4\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$. 因此，我们得到矩阵 $\boldsymbol{U}$ 的更新公式为

$$u_{ij}\leftarrow u_{ij}\left(\frac{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}+\boldsymbol{A}\boldsymbol{U}+\boldsymbol{V}\boldsymbol{X}^{\mathrm{T}}}{\boldsymbol{U}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+2\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}}\right)_{ij} \quad (10)$$

4.2　$\boldsymbol{V}$ 的迭代

我们写出总目标函数中与 $\boldsymbol{V}$ 相关的部分如下：

$$\min_{\boldsymbol{V}\geq 0} L(\boldsymbol{V})=\alpha\|\boldsymbol{S}-\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\|_F^2+\|\boldsymbol{U}\boldsymbol{X}-\boldsymbol{V}\|_F^2 \quad (11)$$

进一步，式(11)可转换为

$$L(\boldsymbol{V})=\alpha\cdot\mathrm{tr}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}-\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{S}+\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}-\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}+\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}) \quad (12)$$

与从式(6)到式(7)的计算相类似，基于式(12)对 $\boldsymbol{V}$ 求偏导，可得到

$$\frac{\partial L(\boldsymbol{V})}{\partial \boldsymbol{V}}=-2\alpha\boldsymbol{S}\boldsymbol{C}-2\boldsymbol{U}\boldsymbol{X}+2\alpha\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}+2\boldsymbol{V} \quad (13)$$

在式(13)中，负项包括 $2\alpha\boldsymbol{S}\boldsymbol{C}$ 和 $2\boldsymbol{U}\boldsymbol{X}$，正项包括 $2\alpha\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}$ 和 $2\boldsymbol{V}$. 因此，我们得到矩阵 $\boldsymbol{V}$ 的更新公式如下：

$$v_{ij}\leftarrow v_{ij}\left(\frac{\alpha\boldsymbol{S}\boldsymbol{C}+\boldsymbol{U}\boldsymbol{X}}{\alpha\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}+\boldsymbol{V}}\right)_{ij} \quad (14)$$

4.3　$\boldsymbol{C}$ 的迭代

我们写出总目标函数中关于 $\boldsymbol{C}$ 的部分如下：

$$\min_{\boldsymbol{C}\geq 0} L(\boldsymbol{C})=\alpha\|\boldsymbol{S}-\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\|_F^2 \quad (15)$$

进而，式(15)可转换为

$$L(\boldsymbol{C})=\alpha\cdot\mathrm{tr}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}-\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{S}+\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}) \quad (16)$$

基于式(16)对 $\boldsymbol{C}$ 求偏导，可得到

$$\frac{\partial L(\boldsymbol{C})}{\partial \boldsymbol{C}}=-2\alpha\boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}+2\alpha\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} \quad (17)$$

在式(17)中,负项包括 $2\alpha \boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}$,正项包括 $2\alpha \boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$. 因此,我们得到矩阵 $\boldsymbol{C}$ 的更新公式为

$$c_{ij} \leftarrow c_{ij}\left(\frac{\boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}}{\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}}\right)_{ij} \tag{18}$$

4.4 $\boldsymbol{X}$ 的迭代

我们写出总目标函数中关于 $\boldsymbol{X}$ 的部分如下:

$$\min_{\boldsymbol{X}\geqslant 0} L(\boldsymbol{X}) = \|\boldsymbol{U}\boldsymbol{X}-\boldsymbol{V}\|_F^2 + \|\boldsymbol{I}-\boldsymbol{X}\|_F^2 + \|\boldsymbol{X}\boldsymbol{1}_k^{\mathrm{T}}-\boldsymbol{1}_k^{\mathrm{T}}\|_F^2 \tag{19}$$

进而,式(19)可转换为

$$\begin{aligned} L(\boldsymbol{X}) = {} & \mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}-\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}+\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}) + \\ & \mathrm{tr}(\boldsymbol{1}_k\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{1}_k^{\mathrm{T}}-\boldsymbol{1}_k\boldsymbol{X}^{\mathrm{T}}\boldsymbol{1}_k^{\mathrm{T}}-\boldsymbol{1}_k\boldsymbol{X}\boldsymbol{1}_k^{\mathrm{T}}+\boldsymbol{1}_k\boldsymbol{1}_k^{\mathrm{T}}) + \\ & \mathrm{tr}(\boldsymbol{I}-\boldsymbol{X}-\boldsymbol{X}^{\mathrm{T}}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) \end{aligned} \tag{20}$$

基于式(20)对 $\boldsymbol{X}$ 求偏导,可得到

$$\frac{\partial L(\boldsymbol{X})}{\partial \boldsymbol{X}} = -2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}-2\boldsymbol{I}-2\boldsymbol{M}+2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}+2\boldsymbol{X}\boldsymbol{M}+2\boldsymbol{X} \tag{21}$$

其中,$\boldsymbol{M}$ 是一个所有元素值为 1 的 $k\times k$ 维矩阵.

在式(21)中,负项包括 $2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}$、$2\boldsymbol{I}$ 和 $2\boldsymbol{M}$,正项包括 $2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}$、$2\boldsymbol{X}\boldsymbol{M}$ 和 $2\boldsymbol{X}$. 因此,我们得到矩阵 $\boldsymbol{X}$ 的更新公式如下:

$$x_{ij} \leftarrow x_{ij}\left(\frac{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}+\boldsymbol{I}+\boldsymbol{M}}{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{M}+\boldsymbol{X}}\right)_{ij} \tag{22}$$

最终,我们得到所有参数的更新公式为

$$\boldsymbol{U}: \quad u_{ij} \leftarrow u_{ij}\left(\frac{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}+\boldsymbol{A}\boldsymbol{U}+\boldsymbol{V}\boldsymbol{X}^{\mathrm{T}}}{\boldsymbol{U}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+2\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}}\right)_{ij} \tag{10}$$

$$\boldsymbol{V}: \quad v_{ij} \leftarrow v_{ij}\left(\frac{\alpha\boldsymbol{S}\boldsymbol{C}+\boldsymbol{U}\boldsymbol{X}}{\alpha\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}+\boldsymbol{V}}\right)_{ij} \tag{14}$$

$$\boldsymbol{C}: \quad c_{ij} \leftarrow c_{ij}\left(\frac{\boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}}{\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}}\right)_{ij} \tag{18}$$

$$\boldsymbol{X}: \quad x_{ij} \leftarrow x_{ij}\left(\frac{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}+\boldsymbol{I}+\boldsymbol{M}}{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{M}+\boldsymbol{X}}\right)_{ij} \tag{22}$$

RSECD 方法的具体实现如算法 1 所示. 其中,T 表示最大迭代次数,n 表示节点数量,k 表示社团(或类簇)数量,m 表示所出现过的属性词数量. 不难发现:矩阵 $\boldsymbol{U}$ 的迭代公式(算法 1 第 7 行)的时间复杂度为 $O(n^2k)$;矩阵 $\boldsymbol{V}$、$\boldsymbol{C}$ 的迭代公式(算法 1 第 8、9 行)的时间复杂度均为 $O(mnk)$;矩阵 $\boldsymbol{X}$ 的迭代公式(算法 1 第 10 行)的时间复杂度为 $O(nk^2)$. 因此,在不考虑数据的稀疏性时,RSECD 算法的时间复杂度为 $O(T(n^2k+2mnk+nk^2))$. 但若考虑到邻接矩阵 $\boldsymbol{A}$ 和属性矩阵 $\boldsymbol{S}$ 的稀疏性,则 RSECD 需要 $O(T(ek+2e'k+nk^2))$ 时间,其中:e 表示网络中边的数量($e\ll n^2$),e' 表示属性矩阵 $\boldsymbol{S}$ 中的非零元素个数($e'\ll mn$). 由于真实世界中的数据通常是稀疏的,因此 RSECD 的时间复杂度与节点数量 n 呈近似线性关系,并具有较高的效率.

算法 1. 基于 Oja 学习准则的优化算法.

输入:社团数量 k,目标函数中的参数 α,
节点数量 n,所出现过的属性词数量 m,
邻接矩阵 $\boldsymbol{A}$,属性词矩阵 $\boldsymbol{S}$,网络拓扑矩阵 $\boldsymbol{U}$,
内容类簇矩阵 $\boldsymbol{V}$,转移矩阵 $\boldsymbol{X}$,最大迭代次数 T

输出:更新后的网络拓扑矩阵 $\boldsymbol{U}$,更新后的转移矩阵 $\boldsymbol{X}$

1. 使用非负矩阵分解方法对 $\boldsymbol{U}$,$\boldsymbol{V}$,$\boldsymbol{X}$ 三个矩阵进行初始化
2. FOR $t=1$ TO T DO
3. IF L 收敛 THEN
4. BREAK;
5. ELSE
6. FOR $u_{ij}\in\boldsymbol{U}$,$v_{ij}\in\boldsymbol{V}$,$c_{ij}\in\boldsymbol{C}$,$x_{ij}\in\boldsymbol{X}$ DO
7. $u_{ij} \leftarrow u_{ij}\left(\frac{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}+\boldsymbol{A}\boldsymbol{U}+\boldsymbol{V}\boldsymbol{X}^{\mathrm{T}}}{\boldsymbol{U}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+2\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}}\right)_{ij}$;
8. $v_{ij} \leftarrow v_{ij}\left(\frac{\alpha\boldsymbol{S}\boldsymbol{C}+\boldsymbol{U}\boldsymbol{X}}{\alpha\boldsymbol{V}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}+\boldsymbol{V}}\right)_{ij}$;
9. $c_{ij} \leftarrow c_{ij}\left(\frac{\boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}}{\boldsymbol{C}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}}\right)_{ij}$;
10. $x_{ij} \leftarrow x_{ij}\left(\frac{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}+\boldsymbol{I}+\boldsymbol{M}}{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{M}+\boldsymbol{X}}\right)_{ij}$;
11. END FOR
12. END IF
13. END FOR

5 实 验

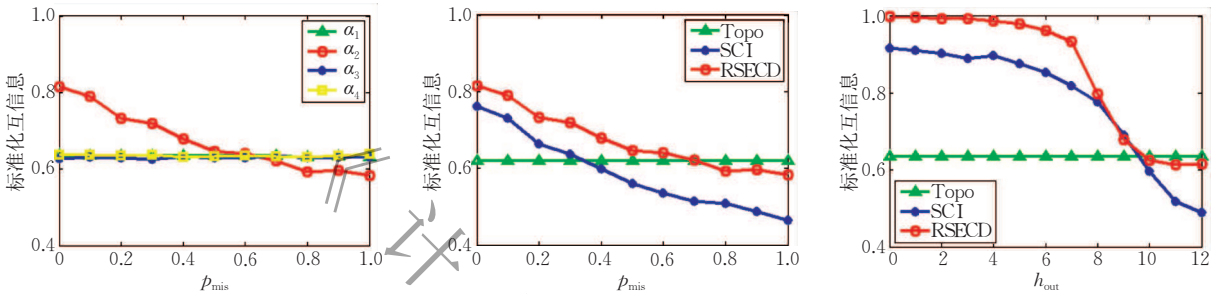
我们首先通过设计合理的人工网络以测试参数 α 取不同值时对算法性能的影响,并通过与两类典型算法的比较来验证本算法可以有效地解决拓扑和内容在社团结构上不匹配的问题以及内容信息所蕴含的类簇结构模糊的问题. 然后我们选取了 4 种常用的评价标准,在 7 个真实网络上,与 8 种代表性社团发现方法进行量化比较. 最后,我们给出了一个定性的实例分析,通过对多个代表性社团的语义分析,验证了本算法对所发现的社团具有强可解释性.

5.1 人工网络

我们使用 Newman 模型[2]生成人工测试网络. 测试网络共有 50 组,每组网络包含 128 个节点. 所有节点被均匀划分为 4 个社团. 每个节点有 z_{in} 条边连接其所属社团内的节点、z_{out} 条边连接其它社团中的节点,且 $z_{\mathrm{in}}+z_{\mathrm{out}}=16$. 同时,所有节点亦被均匀划分到 4 个类簇(每个类簇包含 h 维属性),且每个社团对应一个类簇. 具体来说,对于第 s 个类簇中的每一个节点,我们使用均值为 $p_{\mathrm{in}}=h_{\mathrm{in}}/h$ 的二项式分布生成一个 h 维的二元向量,作为它的第$((s-1)\times h+1)$至$(s\times h)$个属性;再使用一个均值为 $p_{\mathrm{out}}=$

$h_{out}/(3h)$ 的二项式分布生成剩余属性. 为了标识网络中拓扑和内容信息针对社团结构不同程度的匹配情况，我们随机挑选$(100\times p_{mis})\%$个节点并交换它们的属性向量. 可以看出，p_{mis}越大，拓扑和内容信息的社团匹配程度越不一致(其中，p_{mis}的取值范围为$\{0,0.1,0.2,\cdots,1\}$). 实验中，我们设置 $h=50$，$z_{out}=h_{out}=8$ 并使用标准化互信息(Normalized Mutual Information，NMI)[25] 作为实验评价标准(NMI 是一种常用的度量两个社团之间相似程度的非重叠社团评价标准).

在测试参数 α 对算法性能的影响时，我们根据经验考虑了 α 的 4 种取值可能($\alpha=1$，$\alpha=\|\boldsymbol{A}\|_F^2$，$\alpha=1/\|\boldsymbol{S}\|_F^2$ 和 $\alpha=\|\boldsymbol{A}\|_F^2/\|\boldsymbol{S}\|_F^2$)；并分别计算了在 50 组人工网络下，使用这 4 种取值时算法的平均精度. 结果如图 2(a)所示，其中，α_1 对应于 $\alpha=1$，α_2 对应于 $\alpha=\|\boldsymbol{A}\|_F^2$，$\alpha_3$ 对应于 $\alpha=1/\|\boldsymbol{S}\|_F^2$，$\alpha_4$ 对应于 $\alpha=\|\boldsymbol{A}\|_F^2/\|\boldsymbol{S}\|_F^2$. 从图中可以看到，当 p_{mis} 小于 0.6 时，选用 $\alpha=\|\boldsymbol{A}\|_F^2$ 的精度结果明显比选用其它三种取值的结果要好. 考虑到在真实世界网络中，拓扑和内容信息的社团匹配程度往往不会太低. 因此，在接下来的实验中我们选用 $\alpha=\|\boldsymbol{A}\|_F^2$.



(a) 针对α的4种不同参数取值，RSECD算法的NMI精度比较

(b) 在不同的拓扑和内容信息的社团结构匹配程度下，三种不同社团发现算法的NMI精度比较

(c) 当内容信息之类簇结构的清晰程度变化时，三种不同社团发现算法的NMI精度比较

图 2　人工测试网络的实验结果

进一步，为了论证 RSECD 算法可以有效地解决拓扑和内容在社团结构上不匹配的问题，即论证带先验的矩阵 $\boldsymbol{X}$ 具有鲁棒融合拓扑和内容两类信息的功能，我们将 RSECD 与另外两种社团发现方法(Topo 和 SCI[22])进行量化比较. 其中，Topo 是 RSECD 算法在仅使用拓扑信息进行社团发现时的一个变种；SCI 和 RSECD 方法相类似，也是在 NMF 框架下融合拓扑和内容信息进行社团检测的算法；但与 RSECD 不同的是，SCI 并未解决拓扑和内容信息在社团结构上不匹配的问题. 图 2(b)展示了三种算法的比较结果，根据图 2(b)我们可以得到以下信息：

(1) 无论 p_{mis} 如何变化，Topo 始终保持着稳定的检测精度，这是由于存在于该网络中的拓扑信息一直保持不变，而 Topo 仅利用网络拓扑信息；

(2) 当 p_{mis} 小于 0.3 时，由于 SCI 融合了拓扑和内容两类信息，所以它的精度值比仅使用拓扑信息的 Topo 要更高；另一方面，因为 SCI 没有解决拓扑和内容的社团结构不匹配的问题，所以当 p_{mis} 大于 0.34 时，SCI 的性能开始逐渐衰减并劣于 Topo；

(3) 对于引入了带先验的转移概率矩阵的 RSECD 方法：当 p_{mis} 小于 0.4，RSECD 的精度值高于 Topo，这是因为——当社团和类簇匹配良好时，拓扑和内容两类信息相互增强，降低了单一信息的噪声干扰，可找到更为本质的社团结构，提升了社团发现质量；当 p_{mis} 介于 0.4 到 0.7 之间，即社团和类簇匹配不佳时，借助 $\boldsymbol{X}$ 的映射与牵引作用，RSECD 从内容中提取了有益信息，仍可提升社团发现质量，因而 RSECD 的精度略高或持平于 Topo 的精度；当 p_{mis} 大于 0.7，即内容对社团几近无用时，RSECD 的性能仅稍逊于 Topo，却比 SCI 高出很多. 值得注意的是，当 $p_{mis}=0.9$ 时，RSECD 的精度值甚至有所回升，超过了其在 $p_{mis}=0.8$ 时的精度值. 这表明当内容对社团几近无用时，RSECD 能自动减少内容噪声信息对社团检测的影响，从而可鲁棒地检测出社团结构.

在真实的网络数据中，内容信息所蕴含的类簇结构也可能非常模糊，此时的内容信息将会对社团检测产生一定的负面影响；如果社团检测算法可自动忽略掉可能有害的内容信息，就能更加体现出算法的鲁棒特性. 为此，我们设计了第三组人工网络实验. 在这组实验中，我们设置了 $p_{mis}=0$ 不变，用 $h_{out}=0,1,\cdots,12$ 来模拟内容信息的类簇结构从十分清晰到十分模糊的变化趋势. 图 2(c)显示了第三组人工网络实验的结果. 可以看到：当 h_{out} 逐渐增大时，仅使用拓扑信息的 Topo 方法精度值保持不变，而两个融合了拓扑与内容的算法 SCI 和 RSECD 的

精度总体上都呈下降趋势.此外,当内容的类簇结构从十分清晰变为十分模糊时,RSECD 算法的精度值几乎都比 SCI 要好;甚至在内容的类簇结构十分模糊的情况下,RSECD 算法也能与纯拓扑类方法 Topo 近乎持平.这说明 RSECD 可以缓解有害内容信息的干扰,检测出准确的社团划分结果,从而进一步验证了 RSECD 算法的鲁棒性.

综上所述,我们通过 2 组人工网络实验验证了:(1)在拓扑和内容信息针对社团结构匹配一致、不佳、完全不匹配的三种情况下,以及(2)在内容信息所蕴含的类簇结构呈现出清晰、不明显、模糊的三种情形下,RSECD 均可有效地利用拓扑和内容信息进行社团检测.因而,我们说 RSECD 可鲁棒地融合拓扑和内容两类信息进行社团划分.

5.2 真实世界网络

我们选取了 7 个真实网络作为实验数据集,每个网络都包含拓扑结构、节点属性和真实社团标签.具体来说,Citeseer 网络①包含了 6 个社团,3312 个节点(带有 3703 维的二元单词属性)和 4732 条无向边;Uai2010 网络① 包含了 19 个社团,3363 个节点(带有 4972 维的二元单词属性)和 45006 条无向边;Facebook 网络① 包含了 14 个社团,226 个节点(带有 131 维的二元单词属性)和 3417 条无向边;WebKB 网络[26]由 4 个子网络组成,这些子网络数据分别是从 Cornell、Texas、Washington、Wisconsin 四所大学收集而来,每个子网络包含 5 个社团,877 个网页(带有 1703 维的二元单词属性)和 1608 条边.

为了进一步验证 RSECD 算法的性能,我们使用两类评价标准(一类是非重叠社团评价标准,另一类是重叠社团评价标准)来量化分析所检测出的社团结果.针对非重叠社团评价标准,我们选用 Accuracy(AC)[25]和 Normalized Mutual Information(NMI)[25].其中,AC 用于度量社团结果中标签的正确率,NMI 用于度量两个社团划分分布的相似程度.针对重叠社团评价标准,我们选用 F-score[27]和 Jaccard 相似度[27],这两种评价标准常被用于量化真实社团和算法检测出的社团之间的一致程度.

我们共选用 8 种对比算法与 RESCD 进行比较,其中包括 2 个仅使用拓扑信息的算法(DCSBM[28]和 BigCLAM[29])、1 个仅使用内容信息的算法(AP[30])和 5 个融合拓扑和内容信息的算法(CESNA[27], DCM[31], PCL-DC[32], Block-LDA[33]和 SCI[22]).实验过程中,我们结合人工网络参数分析实验所得到的结论,选用 $\|\boldsymbol{A}\|_F^2$ 作为参数 α 的取值;针对每个网络我们将算法重复运行 20 次,从中选取损失函数值最小的结果作为最终的社团划分.比较结果如表 1 至表 4 所示.表中的数字均为百分比数值,每列数据集中最好的结果用粗体标记出来.表 1 和表 2 分别展示了选用 AC 和 NMI 作为评价标准的对比结果,由于 AP 算法不适用 AC 评价标准,CESNA 和 DCM 算法又仅适用于重叠社团评价标准,所以表 1 中没有比较 AP 算法,表 1 和表 2 中也都没有比较 CESNA 和 DCM 算法.由这两个表可知:在使用非重叠社团评价标准下,RSECD 在与所有算法的对比中都展现了最高的性能.表 3 和表 4 分别给出了选用 F-score 和 Jaccard 作为评价标准的实验结果.不难发现,在使用重叠社团评价标准下,RSECD 较之其它算法仍具有最好的性能.

表 1 不同算法的非重叠社团性能比较(采用 AC 度量)

算法/精度(%)		数据集					
算法类型	算法名	Cornell	Texas	Washington	Wisconsin	Citeseer	Uai2010
拓扑	DCSBM	37.95	48.09	31.80	32.82	26.57	2.60
结合	PCL-DC	30.26	38.80	29.95	30.15	24.85	28.82
结合	Block-LDA	46.15	54.10	39.17	49.62	24.35	16.04
结合	SCI	36.92	49.73	46.09	46.42	29.53	29.51
结合	RSECD	**53.85**	**61.50**	**58.70**	**69.43**	**48.67**	**47.21**

表 2 不同算法的非重叠社团性能比较(采用 NMI 度量)

算法/精度(%)		数据集					
算法类型	算法名	Cornell	Texas	Washington	Wisconsin	Citeseer	Uai2010
拓扑	DCSBM	9.69	16.65	9.87	3.14	4.13	31.22
内容	AP	25.27	31.02	31.79	32.48	13.28	41.60
结合	PCL-DC	7.23	10.37	5.66	5.01	2.99	26.92
结合	Block-LDA	6.81	4.21	3.69	10.09	2.42	5.70
结合	SCI	6.80	12.49	6.83	13.28	7.17	23.39
结合	RSECD	**30.24**	**32.67**	**35.10**	**45.32**	**22.34**	**45.73**

① http://snap.stanford.edu

表 3　不同算法的重叠社团性能比较(采用 F-score 度量)

算法/精度(%)		数据集						
算法类型	算法名	Cornell	Texas	Washington	Wisconsin	Facebook	Citeseer	Uai2010
拓扑	DCSBM	34.08	36.14	32.83	29.47	44.92	26.83	30.12
拓扑	BigCLAM	13.23	20.64	13.35	12.84	47.40	9.30	16.99
内容	AP	21.10	23.59	24.11	20.53	23.60	12.92	13.23
结合	CESNA	23.48	23.54	21.91	23.17	52.51	3.38	32.32
结合	DCM	14.38	11.15	12.45	10.45	41.29	2.50	9.65
结合	PCL-DC	32.03	34.30	30.38	27.83	39.49	25.49	29.71
结合	Block-LDA	36.77	32.55	28.95	31.36	39.57	22.49	18.58
结合	SCI	26.94	30.99	28.06	27.06	24.94	26.18	29.66
结合	RSECD	**53.26**	**44.89**	**47.44**	**53.54**	**52.73**	**45.77**	**43.86**

表 4　不同算法的重叠社团性能比较(采用 Jaccard 相似度)

算法/精度(%)		数据集						
算法类型	算法名	Cornell	Texas	Washington	Wisconsin	Facebook	Citeseer	Uai2010
拓扑	DCSBM	21.20	24.14	20.06	17.92	32.18	15.78	18.81
拓扑	BigCLAM	7.18	12.18	7.25	7.01	34.25	5.01	9.87
内容	AP	13.32	16.39	16.26	12.51	13.63	7.39	7.88
结合	CESNA	13.47	13.57	12.40	13.14	39.82	1.73	21.26
结合	DCM	7.95	6.03	6.72	5.54	33.60	1.27	5.77
结合	PCL-DC	19.02	21.56	18.99	16.27	26.99	14.75	19.17
结合	Block-LDA	24.29	22.51	18.20	20.31	26.61	12.80	11.08
结合	SCI	17.10	21.98	18.72	17.15	15.65	15.26	19.11
结合	RSECD	**37.12**	**33.32**	**34.04**	**41.47**	**41.67**	**31.49**	**32.39**

总体来说，RSECD 具有如此优越的性能或许是因为：

(1) 在融合拓扑和内容信息时，算法采用分别迭代拓扑信息矩阵和内容信息矩阵的策略，保证了内容噪声信息不会干扰到拓扑信息对社团划分所起的决定性作用；

(2) 与此同时，带先验的转移概率矩阵在模型中发挥了过滤器的作用. 一方面，转移矩阵从拓扑、内容数据里提取出有益的信息，这些有益信息相互增强、相互补充，使得算法能够发现更为本质的社团结构和主题语义；另一方面，当拓扑和内容信息的社团结构不匹配时，转移矩阵又可自动忽略内容信息的负面影响，仅依据拓扑信息划分社团，因而本算法依然能够取得较高的社团划分精度.

最后，我们记录了在真实网络实验中所有对比算法的运行时间，如表 5 所示. 可以看出，在所有数据集下，基于单一类型数据的算法(仅基于网络拓扑或仅基于节点内容的方法)的运行效率整体都要高于结合拓扑与内容的方法. 尤其是纯基于拓扑的算法 BigCLAM 具有最高的运行效率. 此外，在每个数据集上，文中算法 RSECD 的运行时间都要小于对比算法的平均运行时间；同时，它亦比同类矩阵分解方法 SCI 所花费的时间要小. 这进一步验证了：RSECD 算法具有较高的运行效率.

表 5　不同算法运行时间的比较

(粗体标记了在每个数据集下最快算法的运行时间，“平均值”表示在每个数据集下所有对比算法的平均运行时间)

算法/时间(s)		数据集						
算法类型	算法名	Cornell	Texas	Washington	Wisconsin	Facebook	Citeseer	Uai2010
拓扑	DCSBM	0.35	0.33	0.35	0.61	24.87	93.86	833.91
拓扑	BigCLAM	**0.22**	**0.17**	**0.26**	**0.25**	0.81	**0.78**	15.83
内容	AP	2.11	1.64	1.89	2.34	1.16	769.37	1681.24
结合	CESNA	15.09	12.11	29.27	22.86	60.32	25.63	96.80
结合	DCM	1.70	0.90	1.70	2.60	0.70	28.30	67.80
结合	PCL-DC	2.05	1.15	1.91	2.69	5.49	10.62	120.26
结合	Block-LDA	0.35	1.19	0.27	0.57	2.11	5.87	43.08
结合	SCI	0.27	0.49	1.23	0.60	1.69	85.01	27.11
结合	RSECD	0.30	0.58	0.98	0.54	**0.45**	13.55	**6.00**
平均值		2.49	2.06	4.21	3.67	10.84	114.78	321.34

5.3 Lastfm 实例分析

在实例分析实验中，我们选用 lastfm 数据集①. lastfm 是一个以英国为总部的网络电台和音乐的在线社交网站. 特别是，本实验中采用的 lastfm 数据集包含 1892 个用户，每个用户带有 11946 个属性信息，这些属性信息显示了用户最喜爱的歌曲或者歌手. 由于 lastfm 数据集并没有真实的社团标签信息，这里我们按照 SCI 算法[22]中采用的策略，将网络中的社团数量设置为 38.

我们按照如下步骤进行社团解释：对于某一社团 c，如果我们想给出它的语义解释，首先找到矩阵 $\boldsymbol{X}$ 中对应社团 c 的某一行向量 $\boldsymbol{x}_c$；然后，在此向量中找到转移概率值最高的一个（或多个）类簇 t；进而在矩阵 $\boldsymbol{C}$（所有属性的类簇隶属关系矩阵）中挑选出与类簇 t 最相关的主题词，用于解释类簇 t，进而在解释类簇 t 的同时也就解释了社团 c. 若社团 c 对应了多个起主导作用的类簇，情况类似. 为了直观反映社团中不同属性单词的比重，我们选用词云图作为展示方式. 在一幅词云图中，某一单词和社团越相关，它显示得就越大，从而在视觉上具有较强的社团解释效果. 我们从检测后的社团中挑选了 4 个代表性实例，并选用词云图展示了最终的分析结果，如图 3 至图 5 所示.

(a) 主题类簇36的词云图

(b) 主题类簇13的词云图

图 3 对应单一主题的第 26 号社团和第 28 号社团

(a) 主题类簇1的词云图

(b) 主题类簇32的词云图

图 4 对应两个主要主题的第 30 号社团

第一个例子是序号为 26 的社团，它仅对应于一个序号为 36 的主题类簇. 图 3(a)是主题类簇 36 的词云图，它展示了在该主题下不同属性单词比重的情况. 从图中可知：社团 26 是由一群热爱流行舞曲（尤其是女子流行舞曲）的粉丝组成. 具体来说，“rnb”是“Rhythm and Blues（节奏布鲁斯）”的缩写，代表一种流行舞曲的音乐风格；“disco”是“Discotheque（迪斯科）”的缩写，是一种流行于 20 世纪 70 年代的舞曲；“house（浩室舞曲）”是从迪斯科发展而来的另一种舞曲形式；“latin（拉丁舞曲）”和“trance（迷幻舞曲）”表示不同的舞曲风格；“glam（女孩们要有野心）”是一个韩国女子音乐团队；“sexy（性感的）”描述了女子舞曲的特点；“j-pop（日本流行音乐）”指包含流行舞曲在内的日本流行音乐；此外，代表性单词“dance（舞曲）”也出现在了这里.

(a) 主题类簇13的词云图

(b) 主题类簇24的词云图

(c) 主题类簇36的词云图

图 5 对应三个主要主题的第 16 号社团

第二个例子是序号为 28 的社团，它同样仅对应了一个主题类簇，该类簇的序号为 13，图 3(b)是主题类簇 13 的词云图. 由图 3(b)可知：社团 28 是由一群喜爱以歌剧风格为主的美国流行女歌手的粉丝组成的. 具体来说，“diva（歌剧）”，“rnb（节奏布鲁斯）”和“electropop（流行电子音乐）”是这群女歌手的代表音乐风格；“legend（神话般的）”，“sexy（性感的）”，“hot（热门的）”和“amazing（奇迹的）”描述了这些女歌手在流行音乐史上的突出地位；“female（女性）”，“american（美国）”和“female vocalist（女歌手）”也出现在了这里.

第三个例子是序号为 30 的社团，它包含了两个主要的主题类簇，这两个类簇的序号分别为 1 和 32. 图 4(a)是其主题类簇 1 的词云图，显示了该主题语义为电子流行音乐. 从图中我们可以看到，“electronic（电子音乐）”，“electropop（流行电子音乐）”和“electronica（电子音乐）”在所有属性词中占有很高比重，它们都是电子流行乐的代表性词汇；“australian（澳大利亚风格的）”，“8-bit（芯片音乐）”，“synthpop（流行电音）”，“big beat（摇滚乐）”和“dark pop（黑暗流行乐）”分别代表流行电子音乐的不同风格. 另一方面，图 4(b)是该社团中主题类簇 32 的词云图，它显示该主题的语义信息为合成器

① http://ir.ii.uam.es/hetrec2011/datasets.html

流行音乐.具体来说,合成器流行音乐起源于“new wave(新浪潮)”和“post-punk(后朋克)”,流行于“80s(80 年代)”;“new romantic(新浪漫主义)”是歌手泰勒斯威夫特的一首合成器流行歌曲;“depeche mode(流行尖端)”是一支以另类舞曲和合成器流行乐为风格特色的英国乐队;“electroclash(电子撞击乐)”是“tech pop(科技流行乐)”的另一个名字,而科技流行乐中包含了合成器流行乐风格;此外,“synth(合成器)”和“synth pop(合成器流行乐)”也出现在了这里.值得注意的是,主题 1 和主题 32 所对应的两种音乐风格——电子流行乐和合成器流行乐,它们都属于电子流行音乐,但是,这两个主题语义的区别在于:前者是喜爱各种电子流行乐的粉丝,后者则是酷爱电子流行乐某一个分支(合成器音乐)的粉丝.因此,我们推测社团 30 的成员是一群喜爱电子流行乐、且尤其是更爱合成器流行乐的粉丝.

第四个例子是序号为 16 的社团,它包含了三个主要的主题类簇,这些类簇的序号分别为 13、24 和 36.类似于之前的分析,类簇 13(图 5(a))和一群喜爱以歌剧风格为主的美国流行女歌手的粉丝有关,而类簇 36(图 5(c))和一群喜爱流行舞曲(尤其是女子舞曲)的粉丝有关.图 5(b)是主题类簇 24 的词云图,其显示了该主题的语义信息为以乡村音乐、流行朋克为主要特色的众多美国流行乐女歌手.其中,“Britney spears”,“Lady gaga”,“Ashley tisdale”和“Hilary duff”是美国流行乐女歌手中的一些代表人物;“pop(流行乐)”,“music(音乐)”和“the best(最佳的)”也出现在了这里;而“country(乡村音乐)”和“pop punk(流行朋克)”描述了这些女歌手的部分音乐风格.由上述分析,我们得出以下结论:社团 16 所包含的三个主要的主题类簇(类簇 13、24 和 36)都与流行乐女歌手有关,而它们又各自有精确的含义.具体来说,类簇 36 的主题语义侧重于流行舞曲;类簇 13 的主题语义侧重于歌剧音乐;而类簇 24 的主题语义侧重于乡村音乐和流行朋克.可见,采用多个独立又紧密相关的主题语义,我们可以对社团进行更加精确细致的解释.从而进一步验证了文中方法的强可解释性.

6　总结和讨论

本文提出了一种新的社团发现方法 RSECD,它在检测出社团结构的同时,还能给出社团的语义信息.我们使用了非负矩阵分解模型更加准确地描述节点、主题类簇和网络社团之间的关系.同时,还引入了一个带先验的转移概率矩阵以刻画它们之间的内在关联.由于转移矩阵可以提取内容中的有益信息并规避内容噪声,因而 RSECD 模型可有效解决拓扑和内容信息针对社团结构不匹配的问题,体现了其鲁棒性.在优化过程中,我们使用了一个基于 Oja 迭代学习准则的乘性更新算法,推导出参数迭代公式.通过使用人工网络,我们分析了参数 α 的不同取值对算法性能的影响,并选择最佳参数设置;同时还验证了 RSECD 算法的鲁棒性.在量化实验中,在 4 类社团评价标准下,我们使用 7 个真实网络,对比了 8 种代表性社团发现算法,结果进一步显示了 RSECD 算法的优越性.最后,采用了一个实例分析,展示了对 4 个代表性社团的语义解释,可帮助人们更加准确地理解社团的真实含义.

然而,本文的工作也并不完美,有如下两个方面值得进一步深入探索.首先,与已有模型类社团发现相同,RSECD 算法亦需要事先给定社团数目.这是模型类方法面临的主要问题之一,即模型选择问题.然而,当社团数目未知时,我们亦可采用经典的模型选择方法(如交叉验证、层次贝叶斯等),或者提出新的有效模型选择方法来自动确定社团数目.此外,文中算法 RSECD 假定社团和类簇数目相同,这主要是因为:(1)实验中的对比算法均假设社团数和类簇数相同,所以为公平比较,我们亦采用这一假设;(2)在实验所采用的 7 个真实网络中,实际社团数和类簇数亦相同.然而,在一些复杂的网络数据中,社团数 k 和类簇数 k' 也可能并不相同.但只要对 RSECD 模型稍加修改,文中算法依然适用于这种复杂的情形.我们拟将此类验证作为未来的进一步工作.

参　考　文　献

[1] Fortunato S, Hric D. Community detection in networks: A user guide. Physics Reports, 2016, 659: 1-44

[2] Girvan M, Newman M. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

[3] Jia S, Gao L, Gao Y, et al. Defining and identifying cograph communities in complex networks. New Journal of Physics, 2015, 17(1): 013044

[4] Yang L, Cao X, He D, et al. Modularity based community detection with deep learning//Proceedings of the International Joint Conference on Artificial Intelligence. New York, USA, 2016: 2252-2258

[5] Fanuel M, Alaiz C M, Suykens J A K. Magnetic eigenmaps for community detection in directed networks. Physical Review E, 2017, 95(2-1): 022302

[6] Su C, Jia X, Xie X, Yu Y. A new random-walk based label propagation community detection algorithm//Proceedings of the ACM International Conference on Web Intelligence and Intelligent Agent Technology. Singapore, 2015: 137-140

[7] 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

[8] Hao F, Min G, Pei Z D, Yang L T. *K*-clique community detection in social networks based on formal concept analysis. IEEE Systems Journal, 2017, 11(1): 250-259

[9] Whang J J, Gleich D F, Dhillon I S. Overlapping community detection using neighborhood-inflated seed expansion. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(5): 1272-1284

[10] Jin D, Wang H, Dang J, et al. Detect overlapping communities via ranking node popularities//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 172-178

[11] Liu Shi-Chao, Zhu Fu-Xi, Gan Lin. A label-propagation-probability-based algorithm for overlapping community detection. Chinese Journal of Computers, 2016, 39(4): 717-729(in Chinese)
(刘世超，朱福喜，甘琳. 基于标签传播概率的重叠社区发现算法. 计算机学报，2016，39(4)：717-729)

[12] He D, Liu D, Jin D, Zhang W. A stochastic model for the detection of heterogeneous link communities in complex networks//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 130-136

[13] van Laarhoven T, Marchiori E. Local network community detection with continuous optimization of conductance and weighted kernel *k* -means. Journal of Machine Learning Research, 2016, 17(147): 1-28

[14] Pei Y, Chakraborty N, Sycara K. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 2083-2089

[15] Newman M, Clauset A. Structure and inference in annotated networks. Nature Communications, 2015, 7(2-3): 11863

[16] He Chao-Bo, Tang Yong, Liu Hai, et al. Method for community mining integrating link and attribute information. Chinese Journal of Computers, 2017, 40(3): 601-616(in Chinese)
(贺超波，汤庸，刘海等. 一种集成链接和属性信息的社区挖掘方法. 计算机学报，2017，40(3)：601-616)

[17] Akbari M, Chua T S. Leveraging behavioral factorization and prior knowledge for community discovery and profiling//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. Cambridge, UK, 2017: 71-79

[18] Cai H, Zheng V W, Zhu F, et al. From community detection to community profiling. Proceedings of the VLDB Endowment, 2017, 10(7): 817-828

[19] Zhao Shu, Liu Xiao-Man, Duan Zhen, et al. A survey on social ties mining. Chinese Journal of Computers, 2017, 40(3): 535-555(in Chinese)
(赵姝，刘晓曼，段震等. 社交关系挖掘研究综述. 计算机学报，2017，40(3)：535-555)

[20] Oja E. Principal components, minor components, and linear neural networks. Neural Networks, 1992, 5(6): 927-935

[21] Qiao Shao-Jie, Guo Jun, Han Nan, et al. Parallel algorithm for discovering communities in large-scale complex networks. Chinese Journal of Computers, 2017, 40(3): 687-700(in Chinese)
(乔少杰，郭俊，韩楠等. 大规模复杂网络社区并行发现算法. 计算机学报，2017，40(3)：687-700)

[22] Wang X, Jin D, Cao X, et al. Semantic community identification in large attribute networks//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 265-271

[23] He D, Feng Z, Jin D, et al. Joint identification of network communities and semantics via integrative modeling of network topologies and node contents//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 116-124

[24] Hunter D R, Lange K. A tutorial on MM algorithms. American Statistician, 2004, 58(1): 30-37

[25] Liu H, Wu Z, Deng C, Huang T S. Constrained nonnegative matrix factorization for image representation. IEEE Transactions on Software Engineering, 2012, 34(7): 1299-1311

[26] Sen P, Namata G, Bilgic M, Getoor L. Collective classification in network data. AI Magazine, 2008, 29(3): 93-106

[27] Yang J, McAuley J, Leskovec J. Community detection in networks with node attributes//Proceedings of the 13th International Conference on Data Mining. Dallas, USA, 2013: 1151-11

[28] Karrer B, Newman M. Stochastic blockmodels and community structure in networks. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2011, 83(2): 016107

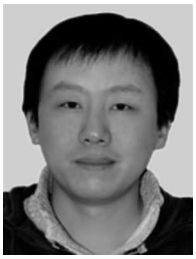
[29] Yang J, Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 587-596

[30] Frey J B, Dueck D. Clustering by passing messages between data points. Science, 2007, 315(5814): 972-976

[31] Pool S, Bonchi F, Leeuwen M V. Description-driven community detection. ACM Transactions on Intelligent Systems and Technology, 2014, 5(2): 1-28

[32] Yang T, Jin R, Chi Y, Zhu S. Combining link and content for community detection: A discriminative approach//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 927-936

[33] Balasubramanyan R, Cohen W W. Block-LDA: Jointly modeling entity-annotated text and entity-entity links//Proceedings of the 11th SIAM International Conference on Data Mining. Mesa, USA, 2011: 450-461



JIN Di, born in 1981, Ph. D., associate professor. His current research interests include artificial intelligence, complex network analysis and network community detection.

LIU Zi-Yang, born in 1994, bachelor. His current research interests include complex network analysis and network community detection.

HE Rui-Fang, born in 1979, Ph. D., associate professor. Her current research interests include natural language processing and information retrieval.

WANG Xiao, born in 1988, Ph. D. His current research interests include network community detection and network embedding.

HE Dong-Xiao, born in 1984, Ph. D., associate professor. Her current research interests include network community detection and data mining.

Background

Network science is a modern and significant discipline in many fields, such as social and computer science. Community detection is one of the most important problems for modeling and analyzing complex networks. Community detection has developed rapidly in recent years and various community detection methods, which mainly focus on network topology, have been proposed, e.g., the agglomerative or divisive algorithm, modularity optimization based methods, spectral algorithm and label propagation based methods. Recently, besides network topology, the text content in complex networks has increasingly attracted people's attention. Some community detection methods combining the content information with network topology have been developed, for example CPD model, SCI method and NEMBP algorithm etc. These methods typically assume that the network topology and node contents share the same community memberships. While, in many real social networks, this assumption does not always hold, because network topology may not match well with node content information. As a result, the robustness and interpretability of these methods are hindered. In addition, community detection should not only detect communities, but also use rich verbal information in the text to give semantic description of communities. The description information reveals why some nodes generate into a community as well as help people better understand the functions or implications of communities.

Thus semantic community identification has two main problems: (1) how to combine topology and content robustly, especially when network communities and content semantics do not match well; (2) how to solve the issue coming from the fact that the communities' description is often difficult to understand. To tackle the above problems, we rethink the functions of network topology and node contents in community detection from a new point of view. Based on this consideration, we design a robust and strong explanatory community detection model called RSECD. Research ideas and innovations in the paper are as follows. Firstly, we put network topology and node contents into a unified framework and use nonnegative matrix factorization model to formalize and solve the problem of community detection. Then a probability transition matrix with prior is introduced and can solve the topology and content's mismatch problem in the network well. Next we derive the effective updating rules based on Oja's iterative learning rule. Finally, to make the detected communities more explainable, we use more than one topic to interpret each community. Experimental results indicate that, RSECD can significantly improve the performance in all comparisons, which illustrates our approach's robustness. The case study semantically explains the hidden meanings of some topics as well as tells the 'actual stories' behind communities.

This work is mainly supported by the National Key R&D Program of China under Grant No. 2017YFC0820106, the Natural Science Foundation of China under Grant Nos. 61502334, 61772361, 61702296 and 61472277, and the Elite Scholar Program of Tianjin University under Grant No. 2017XRG-0016. The first is about "Hidden Community Identification on Heterogeneous Relationship Networks". The second is about "the Studies on Overlapping Community Detection by Combining Network Topology with Node and Link Attributes". The third is about "the Research on Accurate Semantic Community Detection in Large-Scale Complex Networks with Content". The fourth is about "the Research on Complex Network Representation Learning with Heterogeneous Information". And the last is about "Semantic Community Dection in large attributed networks".