大型属性网络中的语义社区识别

王晓[1,5]，金迪[1]，曹晓春[2]，杨亮[2,3]，张伟雄[4,5]
[1]天津大学计算机科学与技术学院，天津 300072，中国
[2]中国科学院IIE信息安全国家重点实验室，北京，100093[3]
天津商业大学信息工程学院，天津，300134
[4]江汉大学数学与计算机科学学院，系统生物学研究所，湖北武汉，430056，中国[5] Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA
{wangxiao cv, jindi}@tju.edu.cn, {caoxiaochun, yangliang}@iie.ac.cn, weixiong.zhang@wustl.edu

摘要

识别网络的模块或社区结构是理解网络语义和功能的关键。虽然已经开发了许多网络社区检测方法，这些方法主要是探索网络工作的拓扑结构，但它们很少提供所发现的社区的语义信息。虽然结构和语义密切相关，但很少有人将这两个基本的网络属性放在一起进行发现和分析。通过整合网络拓扑结构和节点的语义信息，如节点属性，我们研究了社区的检测和语义推断的问题。我们提出了一种新型的非负矩阵因子化（NMF）模型，该模型有两组参数，即社区成员矩阵和社区属性矩阵，并提出了有效的更新规则来评估具有收敛保证的参数。节点属性的使用提高了社区检测的效率，并为产生的网络社区提供了语义上的解释。在合成网络和真实世界网络上的大量实验结果不仅显示了新方法比最先进的方法更优越的性能，而且还证明了其对社区进行语义注释的能力。

简介

复杂系统可以用网络或图来表示。这类网络最突出的特征之一是群落结构，即一个群落内的节点是密集连接的，而不同群落的节点是稀疏连接的（Girvan和Newman
2002）。群落结构有助于揭示复杂系统的组织结构和功能成分。因此，群落检测是对复杂系统进行特征分析的一个重要步骤。

网络拓扑结构是一种重要的网络描述，已经被大多数现有的社区检测方法广泛利用。然而，网络拓扑结构仅仅反映了网络的一个方面，而且往往是有噪声的。因此，仅使用网络拓扑结构不一定能得到满意的网络分区。例如，属于同一社区的两个节点不直接连接的情况并不少见，而一个节点由于不同的原因连接到多个社区是很难的。

仅仅依靠网络拓扑结构就能正确分配到正确的社区。因此，仅仅使用网络拓扑结构来准确地确定社区结构是不充分的。除了网络拓扑结构外，语义信息，如节点属性的信息，往往也是可用的。例如，社交网络中的一个节点（即一个人）通常由个人简介来注释，包括教育背景、朋友圈和职业等信息；引文网络中的一个节点（即一篇论文）通常由标题、摘要和关键词来注释。与网络拓扑结构不同，节点语义捕捉单个节点的特征，并提供与网络拓扑结构信息正交的宝贵信息。网络拓扑和语义信息的整合为社区识别提供了巨大潜力。

然而，要有效地结合这两种有价值的、尽管是正交的信息，在技术上是具有挑战性的。特别是，为了正确地整合这两类信息，需要解决两个障碍。首先，如何充分地描述一个社区。大多数现有的社区检测方法主要依赖于网络拓扑结构。然而，缺失的、无意义的甚至错误的边缘在真实的网络中是无处不在的，这使人们对仅仅基于网络拓扑结构而发现的网络社区的准确性和/或正确性产生了怀疑。虽然社区中的节点是高度关联的，但它们也应该有类似的特征，由属性来体现。因此，节点属性可以携带社区的重要信息，是对网络拓扑信息的补充。因此，即使两个节点没有直接联系，但如果它们有相同的特征，它们也可能属于同一个社区，使用节点属性可以提高社区的识别能力。第二，如何充分解释或从语义上指出社区。网络社区的功能分析通常是社区检测之后的独立后处理任务。除了网络拓扑结构之外，社区发现的结果往往不能提供关于为什么一组节点来自社区、其语义或潜在功能的信息。为了对社区进行语义注释，需要补充信息，例如背景信息和/或领域知识。

边缘，通常是需要的。即使这样的领域信息
如何充分利用这些信息？

语义识别是一种具有挑战性的、特定应用的、耗时的方法。为了解决上述两个问题，我们在本文中提出并开发了一种名为"语义社区识别"（SCI）的方法，用语义注释来识别网络社区。SCI方法整合了网络拓扑和节点语义信息；它在非负矩阵分解（NMF, (Se- ung and Lee 2001)）的框架内将基于拓扑的社区成员和基于节点属性的社区属性（或语义）结合起来。SCI背后的关键直觉来自于两个观察：如果两个节点的社区成员资格相似，那么它们很可能是相连的；如果两个节点的属性与要学习的基础社区属性一致，那么它们很可能属于同一个社区。为了使新的SCI方法有效，我们引入了稀疏性惩罚，以便为每个社区选择最相关的属性，并设计了一个具有收敛保证的乘法上升规则。我们在合成网络和真实网络上进行了广泛的实验，并与几个最先进的方法进行了比较，以评估该方法的有效性。SCI的表现。

相关工作

Xie, Kelley, and Szymanski 2013）中回顾了一些群落检测方法，这些方法被开发出来，用于检测。

探索网络拓扑结构，包括众所周知的拓扑结构
基于非负矩阵分解（NMF）（Wang等人，2011；Yang和Leskovec，2013）和随机块模型（SBM）（Karrer和Newman，2011）。在这些方法中，有一些是结合了网络拓扑结构和节点属性（内容或特征）的。特别是，有人提出了一种单一的方法来结合条件模型
在拓扑结构分析方面，有一个利用节点属性的判别模型（Yang等人，2009）。然而，这
该方法侧重于社区检测，而没有推断出每个社区的最相关属性。边缘检测也被用来改善社区检测过程（Qi, Aggarwal, and Huang 2012）。然而，这种方法是专门为检测链接的社区而设计的，而不是节点的社区。有人提出在节点属性创建的边和边的拓扑信息之间进行启发式的线-
耳组合，以创建一个新的图，用于图的聚类（Ruan, Fuhry, and Parthasarathy 2013）。然而，这种策略在推断社区主题时并没有使用属性的语义信息。一个能够捕捉到社区和属性之间关系的概率模型被开发出来（Yang, McAuley, and Leskovec
2013），它只是给整个网络工作而不是每个社区添加一个稀疏项。此外，学习模型参数的更新规则并不保证能收敛。一个启发式的算法来优化

为恢复群落而进行群落评分，为推断多样化的群落而将描述的复杂性降至最低
这个启发式方法报告了太多相对较小的社区，其中一些有两到三个节点。一个基于非负矩阵三因素化的聚类框架和图形正则化被提出来，以结合这样的方法。

在社交网络中的社会关系和用户生成的内容（Pei, Chakraborty, and Sycara 2015）。然而，这种方法主要是利用额外的内容信息来检测社区，而没有研究社区和这些内容之间的关系。

SCI：网络模型

考虑一个无向网络$G = (V, E)$，有n个节点V和e条边E，用二值邻接矩阵$\mathbf{A}$ R表示$^{n\times n}$ $\in$ 与每个节点i相关的是其at
贡品$\mathbf{S}_i$，这可能是语义上的特征。
节点。一个节点的属性是m维二元值向量的形式，所有节点的属性可以用一个节点属性矩阵来表示
$\mathbf{S}$ $\in$ $\mathsf{R}^{n\times m}$。社区识别的问题是将网络G划分为K个社区，以及推断每个社区的相关属性或语义。对**网络拓扑结构进行建模。**我们将节点i属于社区j的倾向性定义为U_{ij}。然后，网络中所有节点的社区成员资格是$\mathbf{U} = (U_{ij})$，其中$i = 1, 2, ..., n$和$j = 1, 2, ..., k$。因此，$U_{ir}\ U_{pr}$ 提出了预期的边数共同体中的节点i和p之间的关系，对所有共同体进行求和，i和p之间的预期边数是p是$^{k}_{r=1}\ U_{ir}\ U_{pr}$。这个产生边缘的过程意味着如果两个节点具有相似的社区成员资格，它们就有很高的联系倾向。成对的节点之间的预期边数应尽可能地与$\mathbf{A}$表示的网络拓扑结构相一致，这就产生了矩阵表述中的以下函数。

$$\min_{\mathbf{U}\geq\mathbf{o}} /IA - \mathbf{U}\mathbf{U}^T/I2_F 。 \quad (1)$$

建立节点属性模型。我们将社区r拥有属性q的倾向性定义为C_{qr}。因此，对于所有的社区，我们有一个社区属性矩阵$\mathbf{C} = (C_{qr})$，对于$q = 1, 2, ..., m$和$r = 1, 2, ..., k$，其中$第r$列，$\mathbf{C}_r$，是社区r的属性成员。
一个节点的属性与一个社区的属性高度相似，该节点可能有很高的倾向于在社区中。因此，具有类似属性的节点，在$\mathbf{S}_i$
中描述，可以形成一个社区，它可以由节点的共同属性来描述。具体来说，节点i属于社区r的倾向性可以被表述为
为$U_{ir} = \mathbf{S}_i \mathbf{C}_r$。注意，如果节点$i$和社区$r$的属性完全不一致，则节点$i$一定不属于社区$r$，即$U_{ir} = 0$。由于所有节点$\mathbf{U}$的社区成员资格为梳理节点和社区的属性提供了指导，我们有以下优化函数。

$$\min_{\mathbf{C}\geq\mathbf{o}} /IU - \mathbf{SC}/\mathbf{I2}_F 。 \quad (2)$$

为了选择每个社区最相关的属性，我们给ma- trix $\mathbf{C}$的每一列添加一个l_1 norm sparsity。此外，为了防止$\mathbf{C}$的某些列的值过大，这意味着每个社区都有一些有意义的属性，我们对$\mathbf{C}$有一个约束条件

$\sum_{j=1}^{k} \|C(:,j)\|_1^2$，这就产生了以下目标与(2)一起的ive函数。

$$\min_{\mathbf{C}\geq\mathbf{o}} \|U - \mathbf{SC}\|_F^2 + \alpha \sum_{j=1}^{k} \|C(:,j)\|_1^2 。 \quad (3)$$

其中α是一个非负参数，用于在第一个误差项和第二个稀疏项之间进行权衡。

统一的模型。通过结合网络拓扑结构建模的目标函数（1）和节点属性建模的目标函数（3），我们有以下的总体函数。

$$\text{最小值} \quad L = \|U - \mathbf{SC}\|_F^2 + \alpha \sum_{j=1}^{k} \|C(:,j)\|_1^2 + \beta \|A - \mathbf{UU}^T\|_F^2 \quad (4)$$

$$\mathbf{U}\geq\mathbf{o},\mathbf{C}\geq\mathbf{o}$$

其中

。

β是一个正参数，用于调整网络拓扑结构的贡献。

优化

由于（4）中的目标函数不是凸的，所以它是im-实际的，以获得最佳解决方案。局部最小值的(4)可以通过Majorization-Minimization框架实现（Hunter and Lange 2004）。在这里，我们描述了一种算法。迭代更新**U**的算法，**C是固定的**，然后是**C**在**U不变的**情况下，保证不增加对象的数量。每次迭代后，都会有一个新的函数。具体的公式是显示为以下两个子问题。

U-**子问题：**当更新**U**时，**C是固定的**，我们需要解决以下问题。

$$\min_{\mathbf{U}\geq\mathbf{o}} L(\mathbf{U}) = \|U - \mathbf{SC}\|_F^2 + \beta \|A - \mathbf{UU}^T\|_F^2 。 \quad (5)$$

为此，我们为**U**上的非负约束引入了拉格朗日乘数矩阵Θ=（Θ_{ij}），从而得到以下等效的目标函数。

$$L(\mathbf{U}) = tr(\mathbf{UU}^T - \mathbf{UC}^T\mathbf{S}^T - \mathbf{SCU}^T + \mathbf{SCC}^T\mathbf{S}^T) + \beta tr(\mathbf{AA}^T - \mathbf{AUU}^T - \mathbf{UU}^T\mathbf{A} + \mathbf{UU}^T\mathbf{UU}^T) + tr(\Theta\mathbf{U}^T)。 \quad (6)$$

设$L(\mathbf{U})$相对于**U**的导数为0，我们有。

$$\frac{\partial L}{\partial \mathbf{U}} = -2U + 2SC + 4\beta AU - 4\beta UU^T\mathbf{U}。 \quad (7)$$

根据**U**的非负性的Karush-Kuhn-Tucker（KKT）条件，我们有以下方程。

$$(-2U + 2SC + 4\beta AU - 4\beta UU^T\mathbf{U})_{ij} U_{ij} = \Theta_{ij} U_{ij} = 0。 \quad (8)$$

这是收敛时解决方案必须满足的固定点方程。考虑到**U**的初始值，**U**的连续更新是。

为了保证**U**为非负的特性，我们假设矩阵**A**中的对角线元素要大于2β。更新准则**U**的摄取规则满足以下定理，它保证了这是该规则的正确性。

定理1.如果**U**的更新规则是收敛的，那么最终的解决方案满足KKT最优条件。*(证明在附录A1中)*。

我们现在证明更新规则的收敛性。与（Seung和Lee 2001）一样，我们使用一个辅助函数来实现这一目标。

Definition 1. (Seung and Lee 2001) 如果$Q(\mathbf{U}, \mathbf{U}^t)$$L(\mathbf{U})$, $Q(\mathbf{U}, U)$=$L(\mathbf{U})$，对于任何$\mathbf{U}, \mathbf{U}^t$，函数$Q(U, \mathbf{U}^t)$是函数$L(\mathbf{U})$的一个辅助函数。

辅助函数之所以有用，是因为以下几点

lemma:

定理1：（Seung和Lee 2001）如果Q是L的一个辅助函数，那么L在更新规则下是不增加的。

$$\mathbf{U}^{(t+1)} = arg\min_{\mathbf{U}} Q(\mathbf{U}, \mathbf{U}^{(t)}).$$

现在我们有了问题(5)中目标函数$L(\mathbf{U})$的辅助函数$Q(\mathbf{U}, \mathbf{U}^t)$的具体形式，其依据是以下的定律。

悖论2.函数

$$Q(\mathbf{U}, \mathbf{U}^t) = tr(\mathbf{SCC}^T\mathbf{S}^T + \beta AA) + \beta tr(\mathbf{RU}^T\mathbf{U}^T\mathbf{U}) - tr(\mathbf{U}^T\mathbf{A}^T\mathbf{Z}) - tr(\mathbf{Z}^T\mathbf{A}\mathbf{U}_t) - tr(\mathbf{U}^T\mathbf{A}\mathbf{U}_t^T) - 2tr(\mathbf{C}^T\mathbf{S}^T\mathbf{Z}) - 2tr(\mathbf{C}^T\mathbf{S}^T\mathbf{U}^t) \quad (10)$$

是问题（5）中L（**U**）的一个辅助函数，其中$R_{ij} = \frac{U_{ij}^t}{U_{ij}^t} Z_{ij} = U_{ij}^t \ln \frac{U_{ij}}{U_{ij}^t}$, $\mathbf{A}^t = 2\beta A - \mathbf{I}$，而**I**是一个身份矩阵。(*证明见附录A2*）。

根据定理1和2，我们可以证明更新规则的收敛性。

定理2.问题（5）在下列条件下是不增加的迭代更新规则（9）。*(证明见附录A)*。

C-*子问题*：当更新**C**时，**U固定不变**，我们需要解决以下问题。

$$\min_{\mathbf{C}\geq\mathbf{o}} L(\mathbf{C}) = \|U - \mathbf{SC}\|_F^2 + \alpha \sum_{j=1}^{k} \|C(:,j)\|_1^2. \quad (11)$$

这相当于以下优化问题（Kim和Park，2008）。

$$\min_{\mathbf{C}\geq\mathbf{o}} L(\mathbf{C}) = \left\| \begin{pmatrix} \sqrt{\alpha}\mathbf{e}_{1\times m} \\ U \end{pmatrix} - \begin{pmatrix} \sqrt{\alpha}\mathbf{e}_{1\times m} \\ \mathbf{S} \end{pmatrix} \mathbf{C} \right\|_F^2, \quad (12)$$

其中$\mathbf{e}_{1\times m}$是一个行向量，所有分量都等于1，$\mathbf{0}_{1\times k}$是一个零向量。所以更新规则及其收敛分析可以在（Seung and Lee 2001）中找到。

在收敛时，由于**U**表示社区的软成员分布，我们可以直接使用**U**或**U**=**SC**来得到最终的不相交或重叠的社区。**C**的每一列表示一个社区与属性之间的关系，其中较大的值代表了

$$U_{ij} \leftarrow U_{ij} \left(\frac{(\mathbf{SC} + 2\beta \mathrm{AU} - \mathbf{U})_{ij}}{2\beta (\mathbf{UU}\,\mathbf{U})_{ij}} \right)^{\frac{1}{4}}. \quad (9)$$

如果是这样，那么相应的属性与社区的相关性就越大。

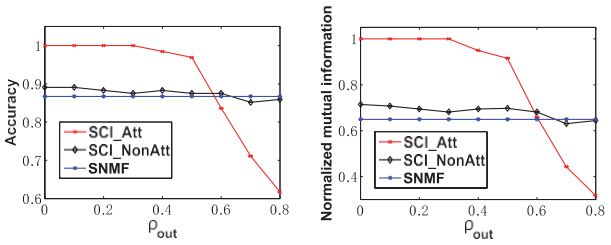
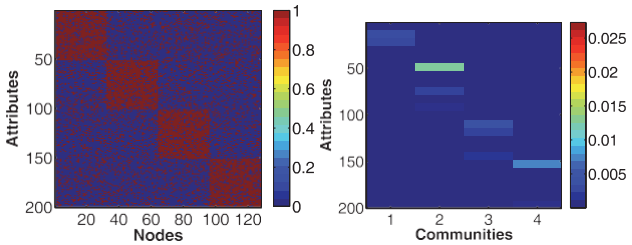


图1：直接使用**U**的SCI（SCI NonAtt）、使用**U** = **SC**的SCI（SCI Att）和SNMF的性能比较。



实验评价

合成网络

我们首先在一个使用广泛采用的纽曼模型（Girvan and Newman 2002）构建的合成网络上评估了SCI。该网络由128个节点组成，分为4个互不相连的社区。每个节点平均有z_{in}条边将其与同一社区的成员联系起来。nity和z_{out}边给其他社群的成员，$z_{in} + z_{out} = 16$。这里我们将z_{in}和z_{out}设置为8，这对大多数方法来说是一个具有挑战性的问题，因为没有明显的社区结构（Yang等人，2014）。然后，我们为每个节点生成一个4h_{in}-维的二元属性，如下所示。对于*第i*个社区内的每个节点，我们使用均值为ρ的二项分布$_{in}$，生成一个h_{in}-二维向量作为其$((i-1)h_{in}+1)$-th to (ih_{in})-th属性，并使用平均值为ρ_{out}的二项分布生成其余属性。我们对每个节点总共有(4h_{in})-dimensional属性。请注意，$\rho_{in} > \rho_{out}$，这意味着这些生成的h_{in}-维属性与这个社区相关，具有高概率，而其余的是不相关的（或嘈杂的）属性。如前所述，新方法推断出两个参数**U**和**C**，因此我们可以直接使用推断出的**U**或属性**C**来推导出一个新的**U**，如**U**=**SC**来重新覆盖社区结构。为方便起见，我们将其命名为这两种方案分别为SCI - - NonAtt和SCI Att。我们的实验首先是为了研究这两种方案之间的差异。我们设定$h_{in} = 50$，$\rho_{in} = 0.8$，ρ_{out}从0到0.8变化，增量为0.1。我们采用SNMF（Wang等人，2011），使用网络拓扑结构单独作为基线方法进行比较。我们使用accuracy（AC）（Liu等人，2012）和normalized mutual informa- tion（NMI）（Liu等人，2012）作为performance评估的质量度量。如图1所示，SCI Att和SCI NonAtt都优于SNMF，除了当ρ_{out}almost达到0.8。该结果表明，识别的质量随着节点信息的增加，已被识别的社区在------中得到改善。贡品。此外，在$\rho_{out} - = 0.5$之前，SCI Att通常明显优于SCI NonAtt。当ρ_{out}增加到0.5以上时，SCI -Att的性能就会恶化。这部分是因为当ρ_{out}大于0.5时，节点属性提供的网络共性的鉴别信息较少，也就是说，与社区相关的特定属性较少。这也意味着，节点属性可能如果它们的质量不高，则可能扭曲结果。然而，在

图2：**左：**节点属性矩阵。右图：递归社区属性矩阵
。

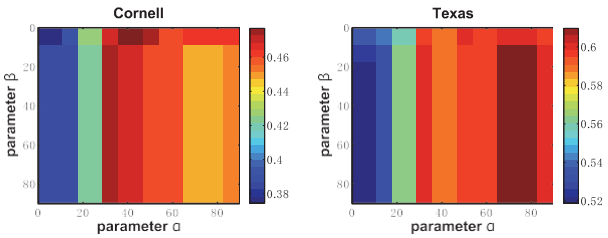


图3：参数α和β的影响。不同的颜色表示不同的精确度，接近红色的颜色表示高精确度。

一般来说，节点属性具有潜在的判别能力，对区分社区很有好处。因此，我们没有直接使用**U**，而是将**U**=**SC**作为以下所有实验的最终社区成员。

此外，我们研究了由SCI推断的社区属性**C**。我们把$\rho_{in}=0.8$，$\rho_{out}=0.2$，$h_{in}=50$。生成的节点属性矩阵显示在图2的左图中。如图所示，每个社区的节点都有50维的相关属性，其余的属性都是不相关的。我们注意到，每个社区的属性都非常不同，如图2的右图所示，这意味着每个社区都有独特的属性。此外，社区属性与社区内节点的相关属性是一致的。
社区。简而言之，新方法能够识别网络模块结构，并推断出社区属性，提供社区的语义信息。

真实的网络

我们考虑了三个具有节点属性和地面真实社区标签的真实网络。Citeseer网络[1]
（6个社区）由3312个科学出版物和4732条边组成，而Cora[1]
网络（7个社区）由2708个科学出版物和5429条边组成。Citeseer和Cora中的出版物分别与3703维和1433维的二值词属性相关联，表明一个相应的词是否出现在出版物中。WebKB网络[1]
，包括4个子网，分别来自4所大学（康奈尔大学、德克萨斯大学、华盛顿大学、哥伦比亚大学）。

表1：不相干社区的性能比较（粗体数字代表最佳结果）。

度量衡	方法	康奈尔大学	德州	华盛顿	威斯康星州	Cora	馨香坊
交流	PCL-DC	0.3487	0.3690	0.4087	0.3547	**0.5543**	**0.6525**
	SNMF	0.3179	0.3583	0.2783	0.3283	0.4173	0.2539
	SBM	0.3436	0.3743	0.2826	0.2981	0.3833	0.2844
	CAN	0.4154	0.4706	0.5087	0.4717	0.3021	0.2129
	SMR	0.3179	0.5401	0.4565	0.4226	0.3002	0.2111
	SCI	**0.4769**	**0.6096**	**0.5435**	**0.5245**	0.4169	0.3442
NMI	PCL-DC	0.0813	0.0686	0.1031	0.0719	**0.3830**	**0.3816**
	SNMF	0.0332	0.0476	0.0211	0.0803	0.1994	0.0403
	SBM	0.0543	0.0839	0.0211	0.0428	0.2047	0.0512
	CAN	0.0614	0.0908	0.1175	0.0702	0.0132	0.0079
	SMR	0.0845	0.1150	0.0381	0.0777	0.0078	0.0032
	SCI	**0.1520**	**0.2197**	**0.2096**	**0.1852**	0.1780	0.0922

廷顿和威斯康星州）。每个子网络被划分为5 个社区。有877 个 网页，1608条边。每个网页都由1703 维的二进制值的单词属性进行注释。

我们将SCI与三种基于拓扑结构的方法进行比较。SNMF（Wang等人，2011），SBM（Karrer和Newman，2011），BIGCLAM（Yang，McAuley和Leskovec，2013）；两种基于节点属性的方法。CAN（Nie、Wang和Huang，2014）和SMR（Hu等人，2014）；三种结合网络拓扑结构和节点属性的方法。PCL
DC（Yang等人，2009），CESNA（Yang，McAuley和Leskovec，2013），DCM（Pool，Bonchi和Leeuwen，2014）。所比较的方法可能提供不相干的或重叠的社区，因此我们选择了不同的评价指标。对于不联合的社群，我们采用了准确度（AC）（Liu等人，2012）和归一化互信息（NMI）（Liu等人，2012）。对于重叠的社群，我们采用了广义的归一化互信息（GNMI）（Lancichinetti, Fortu- nato, and Kerte´sz 2009）。此外，我们将检测到的一组群落M与地面真实的群落M进行比较。
munities M^* 如（Yang, McAuley, and Leskovec 2013）。

$$\frac{1}{2|M|}\sum_{M_i\in M}\max_{M_j\in M}\delta(M^*_j, M_j)+\frac{1}{2|M|}\sum_{M_j\in M}\max_{M^*_j\in M^*}\delta(M^*_j, M_j)$$

, 其中$\delta(M^*_i, M_j)$是社区M^* 和M_j 之间的相似度（F-score和Jaccard相似度）。

我们验证了SCI在不相交和相交的情况下的有效性。重叠的社区结果，分别见表1和表2。如表1所示，SCI在6个网络实例中的4个（科内尔、德克萨斯、华盛顿和威斯康星）上优于其他方法。如表2所示，当用F-score和Jaccard指标测量时，SCI在所有测试的网络上都取得了最好的表现；在GNMI方面，它在六个网络中的四个上超过了其他方法，进一步证明了SCI的有效性。请注意，在不同的网络中，准确率有很大的不同（即使使用同一组方法），这可能反映了所分析的网络的不同特点。

我们测试了SCI的参数α和β对真实网络的影响，也就是说，α和β是调整稀疏项和网络拓扑结构贡献的参数。

[1]http://linqs.cs.umd.edu/projects/projects/lbc/

分别。我们将每个参数从1到100变化，增量为10。由于不同网络的结果有相似的趋势，这里我们只在图3中显示了两个网络（Cor-
nell和Texas）。请注意，随着参数β的变化，SCI是相对稳定的，而它的变化是显著的。
受α的影响，表明了稀疏性项的重要性。因此，我们建议将β设置为1或10至100之间的数值，并对α进行微调，以达到高性能。

由于SCI收敛到局部最优，我们在康奈尔、德克萨斯、华盛顿和威斯康星州的数据集上测试了它的稳健性。我们用10个不同的初始化来重复SCI。损失函数的平均值为81.9509 0 .2844, 87.8448 0 .1728, 102.4153 0.2852，和105.53 0.4075，分别。这些变化都小于0.4%，表明SCI的稳定性。
SCI的主要计算是用于更新**U**和**C**.复杂度为$O(T(mnk + n^2k))$，T次迭代收敛。我们还报告了SCI在康奈尔、德克萨斯、华盛顿、威斯康辛、科拉和Citeseer数据集上的运行时间。在一台装有

"内存：8G；CPU。英特尔I7；平台。Matlab"，运行中的
时间分别为0.4509s, 0.1917s, 0.3234s, 0.4571s, 88.6821s 和69.8382s，分别是。

对检测到的社区进行分析

我们仔细研究了一些由SCI检测到的社区。在这里，我们使用了LASTFM数据集[2]
，该数据集来自在线音乐系统*Last.fm*，其1892名用户在*Last.fm* "朋友"关系产生的社会网络中连接。每个用户都有11946个维度的属性，包括最常听的音乐艺术家列表和标签分配。由于该网络没有真实的标签，我们在上一节中没有对其进行定量评估。我们使用Louvain方法（Blondel等人，2008）将社区的数量设定为38个。四个例子的社区属性在图4中显示为词云。一个词的大小与它的社区属性值成正比，也就是说，一个属性越相关，它在图中就越大。

对于每个社区，我们选择了前十个属性。我们观察到这四个社区有其独特的.....

[2]http://ir.ii.uam.es/hetrec2011/datasets.html

表2：重叠社区的性能比较（粗体数字代表最佳结果）。

度量衡	方法	康奈尔大学	德州	华盛顿	威斯康星州	Cora	馨香坊
辽宁大学	BIGCLAM	0.0051	0.0034	0.0028	0	0.0244	5.551e-17
	CESNA	0.0704	0.0008	**0.1151**	**0.1573**	0.0179	0
	DCM	1.110e-16	0.0090	0.0062	1.110e-16	2.220e-16	0
	SCI	**0.0901**	**0.0955**	0.0859	0.0879	**0.1039**	**0.0506**
F-score	BIGCLAM	0.2267	0.2097	0.2002	0.2399	0.2927	0.1386
	CESNA	0.3368	0.2352	0.3527	0.4393	0.3160	0.1360
	DCM	0.1438	0.0908	0.1127	0.1052	0.0345	0.0245
	SCI	**0.4766**	**0.4740**	**0.4718**	**0.5063**	**0.3835**	**0.3651**
雅卡德	BIGCLAM	0.1294	0.1190	0.1120	0.1380	0.1797	0.0829
	CESNA	0.2120	0.1406	0.2551	0.3164	0.1940	0.0794
	DCM	0.0795	0.0484	0.0607	0.0563	0.0177	0.0125
	SCI	**0.3225**	**0.3413**	**0.3303**	**0.3642**	**0.2519**	**0.2275**

贡品。特别是，图4（a）中的社区显示，这是一个由"重金属"乐队或music的粉丝组成的群体。例如，"metallica"、"queensryche"、"backyard babies"、"sound garden"和"skid row"都是重型精神乐队。此外，"Slash"和"Nikki Sixx"的音乐类型也包括重型精神乐队。特别是，这里出现了"重型精神"和"华丽朋克"的标签。图4（b）中社区的主题应该与歌手"Rihanna"或流行音乐有关，因为"Rihanna"这个词是最大的，她是有史以来最畅销的艺术家之一，在全球范围内的点击率很高。她的歌曲"We Found Love"被Billboard列为美国Billboard第24大排行榜。有史以来最热门的100首歌曲。"下雨的男人"是她的歌曲之一。和"rated r"是她的第四张录音室专辑。"xtina"是另一张流行歌手"Christina Aguilera"。对于图4（c）中的社区，它主要与摇滚乐队"杜兰朵"和摇滚乐有关。此外，"新浪漫主义"、"合成摇滚"和"新浪潮"都是他们的流派。另外，根据维基百科[3]，"超级组合"通常用于摇滚和流行音乐，"杜兰杜兰"就是其中之一。对于图4（d）中的社区，其主题主要是关于社会、民生或政治问题。特别是，"deutsche welle"是一个德国的国际广播公司，向德国以外的受众广播新闻和信息。与之前的音乐社区不同，它谈论的是"女性赋权"和其他话题，如生活、"疾病"和"治疗"。总之，这四个社区有其独特的属性；通过利用这些属性，我们能够解释和理解这些社区。

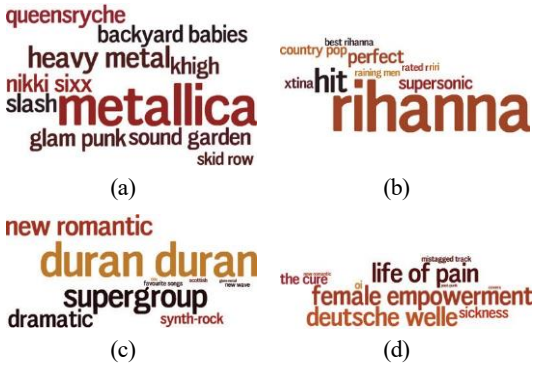


图4：不同社区的词云。这里显示了四个社区的前十个属性。一个词的大小与它的社区属性值成正比。

结论性意见

我们开发了一种新的语义社区识别方法--SCI，用于检测网络社区结构并同时推断其语义。SCI的一个突出特点是它能够从语义上或功能上对每个被识别的社区进行注释。SCI的关键思想是在非负矩阵分解（NMF）的框架下充分整合网络拓扑信息和节点属性信息。我们将SCI表述为NMF中的一个优化问题，并在此基础上进行了优化。设计了具有收敛性保证的高效更新规则。大量的实验结果表明，SCI在准确识别网络社区结构方面表现出色，同时推断出社区语义或属性来理解社区结构。

[3]https://en.wikipedia.org/wiki/Supergroup(音乐)

附录

A1.定理1的证明

在收敛时，$\mathbf{U}^{(\infty)} = \mathbf{U}^{(t+1)} = \mathbf{U}^{(t)} = \mathbf{U}$，其中$t$表示*第$t$次*迭代，即。

$$U_{ij} = U_{ij}\left(\frac{(\mathbf{SC} + 2\beta \mathrm{AU} - \mathbf{U})_{ij}}{2\beta(\mathbf{UU}^{T}\mathbf{U})_{ij}}\right)^{\frac{1}{4}}, \quad (13)$$

这相当于

$$(-2\mathrm{U} + 2\mathrm{SC} + 4\beta \mathrm{AU} - 4\beta \mathrm{UU}^{T}\mathbf{U})_{ij} U_{ij}^{t} = 0。 \quad (14)$$

这等同于（8）。□

A2.推理2的证明。

$$L(\mathbf{U}) = tr(\mathbf{UU}^{T} - \mathbf{UC}^{T}\mathbf{S}^{T} - \mathbf{SCU}^{T} + \mathbf{SCC}^{T}\mathbf{S}^{T}) + \beta\,\mathrm{tr}(\mathbf{AA}^{T} - 2\mathrm{auu}^{T} + \mathbf{uu}^{T}\mathbf{uu}^{T})。$$

(15)

根据(Wang et al. 2011)的定理6和7，我们有

$$tr(\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T) \le tr(\mathbf{P}\mathbf{U}^{t\,T}\mathbf{U}^t) \le tr(\mathbf{R}\mathbf{U}^{t\,T}\mathbf{U}^t\mathbf{U}^{t\,T})。 \quad (16)$$

其中 $\frac{\vec{A}\vec{A}}{\vec{A}}$ $\frac{(\mathbf{U}^T\mathbf{U})^2_{ij}}{(\mathbf{U}^T\mathbf{U}^t)_{ij}}$, $Ri = \frac{U_{ij}^4}{U_{ij}^3}$

根据（Wang等人，2011）的第4条定理，我们有

$$-tr[(2\beta\mathrm{A} - \mathbf{I})\mathbf{U}\mathbf{U}^T] = -tr(\mathbf{A}\mathbf{U}\mathbf{U})^T \le -tr(\mathbf{U}^{tT}\mathbf{A}^t\mathbf{Z}) - tr(\mathbf{Z}^T\mathbf{A}^t\mathbf{U}) - tr(\mathbf{U}^{tT}\mathbf{A}^t\mathbf{U}^t) \quad (17)$$

）。

根据（Wang等人，2011）的Lemma 2，我们有

$$-tr(\mathbf{U}\mathbf{C}\mathbf{S}^T)^T \le -tr(\mathbf{C}\mathbf{S}^T\mathbf{Z}) - tr(\mathbf{C}\mathbf{S}^T\mathbf{U})^T。 \quad (18)$$

对于（17）和（18），$= U_{ij}^t \ln\frac{U_{ij}}{U_{ij}}$。通过结合 Z_{ij} (16)、(17)和(18)，我们有最终的辅助函数在悖论2。□

A3.定理2的证明。

定理2提供了问题（5）中L（$\mathbf{U}$）的辅助函数Q（$\mathbf{U}$，$\mathbf{U}^t$）的具体形式。通过以下KKT条件，我们可以得到min $_{\mathbf{U}} Q(\mathbf{U}, \mathbf{U}^t)$的解决方案

$$\frac{\partial Q(\mathbf{U}, \mathbf{U}^t)}{\partial U_{ij}} = 4\beta(\mathbf{U}^t\mathbf{U}_{ij}^{tT}) \frac{U_{ij}^3}{U_{ij}^t} - \frac{U_{ij}^t}{U_{ij}^t}(2(\mathbf{a}\mathbf{u})_{ij} + 2(\mathbf{sc})_{ij}) = 0 \quad (19)$$

。

这就产生了（9）中的更新规则。根据定理1，在这个更新规则下，（5）的目标函数L（$\mathbf{U}$）将是不增加的。□

鸣谢。曹晓春先生是该书的作者。

本文作者。该工作得到了国家高技术研究发展计划（2014BAK11B03）、国家基础研究计划（2013CB329305）、国家自然科学基金（No.61422213, 61503281, 61502334, 61303110, 61572226）、"战略性先导专项"（No.

中国科学院

"国家重点基础研究发展计划"（XDA06010701）、天津商业大学青年学者基金（150113）、武汉市人才发展计划、武汉市政府（2014070504020241）、武汉市江汉大学内部研究基金（R01GM100364）和美国国家卫生研究院（R01GM100364）。中国湖北武汉市人才发展计划（2014070504020241），中国武汉市江汉大学内部研究经费，美国国家卫生研究院（R01GM100364）和美国国家科学基金会（DBI-0743797）。

参考文献

Blondel, V.; Guillaume, J.; Lambiotte, R.; and Lefebvre, E. 2008.大型网络中群落的快速展开。*Journal of Sta- tistical Mechanics:Theory and Experiment* 2008(10):P10008.

Girvan, M., and Newman, M. 2002.社会和生物网络中的社区结构。*国家科学院院刊》*99（

Karrer, B., and Newman, M. E. 2011.随机区块模型和网络中的社区结构。*Physical Review E* 83(1):016107.

Kim, J., and Park, H. 2008.稀疏的非负矩阵因子集群化。*技术报告GT-CSE-08-01，乔治亚理工学院。*

Lancichinetti, A.; Fortunato, S.; and Kerte´sz, J. 2009。检测复杂的群落结构的重叠性和层次性网络。*New Journal of Physics* 11(3):033015.

Liu, H.; Wu, Z.; Li, X.; Cai, D.; and Huang, T. S. 2012。Con-紧张的非负矩阵分解的图像代表。*Pattern Analysis and Machine Intelligence, IEEE Transac- tions on* 34(7):1299-1311.

Nie, F.; Wang, X.; and Huang, H. 2014.带有自适应邻居的聚类和预测性聚类。In *Proceedings of the*

第20届ACM

*SIGKDD国际会议，关于知识识别和数据挖掘，*977-986。ACM.

Pei, Y.; Chakraborty, N.; and Sycara, K. 2015.社会网络中社区检测的非负矩阵三要素化与图正则化。In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2083- 2089.AAAI出版社。

Pool, S.; Bonchi, F.; and Leeuwen, M. v. 2014。描述驱动的社区检测。*ACM Transactions on Intelligent Systems and Technology (TIST)* 5(2):28.

Qi, G.-J.; Aggarwal, C. C.; and Huang, T. 2012。社区在社交媒体网络中用边缘内容进行检测。*在数据*

12）：7821-7826。

Hu, H.; Lin, Z.; Feng, J.; and Zhou, J. 2014。Smooth represen-tation clustering.In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 3834-3841.IEEE.

Hunter, D. R., and Lange, K. 2004.mm算法教程》。*美国统计学家》*58(1):30-37。

工程（ICDE），2012年IEEE第28届国际会议，534-545。IEEE.

Ruan, Y.; Fuhry, D.; and Parthasarathy, S. 2013.使用内容和链接在大型网络中进行高效的社区检测。在*第22届世界互联网国际会议论文集中*，1089-1098。国际万维网会议指导委员会。

Seung, D., and Lee, L. 2001.Algorithms for non-negative matrix factorization.*Advances in neural information processing systems* 13:556-562.

Wang, F.; Li, T.; Wang, X.; Zhu, S.; and Ding, C. 2011。使用非负矩阵分解的社区发现.*数据挖掘和知识发现*22（3）：493-521.

Xie, J.; Kelley, S.; and Szymanski, B. K. 2013。网络中的重叠社区检测。最先进的技术和comparative研究。*ACM计算调查（CSUR）*45（4）：43。

Yang, J., and Leskovec, J. 2013.规模化的重叠社区检测：非负矩阵分解方法。在*第六届ACM网络搜索和数据挖掘国际会议上*，587-596。ACM.

Yang, T.; Jin, R.; Chi, Y.; and Zhu, S. 2009.结合链接和内容进行社区检测：一种判别方法。在*第15届ACM SIGKDD知识发现和数据挖掘国际会议上*，927-936。ACM.

Yang, L.; Cao, X.; Jin, D.; Wang, X.; and Meng, D. 2014.一个使用潜在空间图正则化的统一半监督社区检测框架。*Cybernetics, IEEE Transac- tions on* 45(11):2585-2598.

Yang, J.; McAuley, J.; and Leskovec, J. 2013.具有节点属性的网络中的社区检测。In *Data Mining (ICDM), 2013 IEEE 13th international conference on*, 1151-1156.IEEE.