Physical Review E **76**, 036106 (2007)

# 检测大规模网络中社区结构的近线性时间算法

Usha Nandini Raghavan,[1] Réka Albert,[2] and Soundar Kumara[1]
[1]*宾夕法尼亚州立大学工业工程系，宾夕法尼亚州大学公园，16802，美国*
[2]*美国宾夕法尼亚州立大学物理系，宾夕法尼亚州大学公园，16802，美国*
(2007年4月9日收到；2007年9月11日发表)

群落检测和分析是理解各种现实世界网络组织的重要方法，并应用于社会群体中的共识形成或生化网络中功能模块的识别等各种问题。目前使用的识别大规模现实世界网络中群落结构的算法需要*先验*信息，如群落的数量和大小，或者计算成本很高。在本文中，我们研究了一种简单的标签传播算法，该算法仅使用网络结构作为其指导，既不需要优化预定义的目标函数，也不需要关于群落的先验信息。在我们的算法中，每个节点都被初始化为一个独特的标签，在每一步，每个节点都采用其大多数邻居目前拥有的标签。在这个迭代的过程中，密集连接的节点群在一个独特的标签上形成共识，从而形成社区。我们通过将该算法应用于社区结构已知的网络来验证该算法。我们还证明，该算法几乎需要线性时间，因此它的计算成本比目前可能的要低。

DOI: 10.1103/PhysRevE.76.036106 PACS编号: 89.75.Fb, 89.75.Hc, 87.23.Ge, 02.10.Ox

## I. 简介

各种各样的复杂系统都可以用网络来表示。例如，万维网（WWW）是由超链接相互连接的网页组成的网络；社会网络以人作为节点，以边表示他们之间的关系；生物网络通常以生化分子作为节点，以边表示它们之间的关系。最近的研究大多集中在了解这类网络的演变和组织化，以及网络拓扑结构对系统的动态和行为的影响[1-4]。寻找网络中的群落结构是理解它们所代表的复杂系统的另一个步骤。

网络中的社区是一组节点，它们彼此相似，与网络的其他部分不同。它通常被认为是一个节点密集互联、与网络其他部分稀疏连接的群体[4-6]。对社区没有一个公认的定义，但众所周知，大多数现实世界的网络都显示出社区结构。最近在定义、检测和识别现实世界网络中的社群方面有很多努力[5,7-15]。社群检测算法的目标是在一个给定的网络中找到感兴趣的节点群。例如，WWW网络中的社区表明该组中的节点具有相似性。因此，如果我们知道一小部分网页所提供的信息，那么它就可以被排除在同一社区的其他网页之外。社会网络中的社区可以提供关于人们之间的共同特征或信仰的洞察力，使他们与其他社区不同。在生物分子相互作用网络中，将节点分离成功能模块有助于确定单个分子的作用或功能[10]。此外，在许多大规模的现实世界的网络工作中，社区可以有独特的属性，而这些属性在综合分析中会丢失[1]。

社区检测与已研究的网络划分问题相似[16-18]。一般来说，*网络划分*问题被定义为将网络划分为$c$（固定常数）个大小近似的组，使组间的边数最小。这个问题是*NP-hard*，多年来已经开发了高效的启发式方法来解决这个问题[16-20]。这项工作的动机是工程应用，包括超大规模集成（VLSI）电路布局设计和并行计算的映射。Thompson[21]表明，影响芯片中特定电路的最小布局面积的重要因素之一是其分切宽度。另外，为了提高计算算法的性能，节点代表计算，边代表通信，节点在处理器之间平均分配，使它们之间的通信量最小。

网络划分算法的目标是将任何给定的网络划分为大约相等大小的组，而不是各自的节点相似性。而社区检测则是找到具有固有的或外部指定的组内节点相似性概念的组。此外，网络中社区的数量
工作和它们的大小是事先不知道的，它们是由社区检测算法确定的。

已经提出了许多算法来寻找网络中的社区结构。层次化的方法根据不同的测量方法将网络划分为不同的社区，导致一系列从整个网络工作到单子社区的划分[5,15]。同样地，我们也可以根据相似度来连续地将较小的社区组合在一起，再次导致一系列的分区[22,23]。由于分区的范围很广，衡量社区结构强度的结构指标被用于确定最相关的分区。基于模拟的方法也经常被用来寻找具有强社群结构的分区[10,24]。频谱法[17,25]和流量最大化（切割最小化）方法[9,26]已经被用于

 

成功用于将网络划分为两个或更多的社区。

在本文中，我们提出了一种基于标签传播的本地化社区检测算法。每个节点的初始化都有一个独特的标签，在算法的每一次迭代中，每个节点都采用一个其邻居最多的标签，并以随机方式统一打破联系。当标签以这种方式在网络中传播时，密集连接的节点组就其标签形成共识。在算法的最后，具有相同标签的节点被分组为社区。正如我们所展示的，这种算法比其他方法的优势在于它的简单性和时间效率。该算法使用网络结构来指导其进展，并不优化任何特定的社区强度措施。此外，社群的数量和大小并不是*事先就*知道的，而是在算法结束时确定的。

的计算方法。我们将表明，获得的社区结构
通过在以前考虑过的网络中应用该算法，如Zachary的空手道俱乐部友谊网络和美国大学足球网络，与这些网络中存在的实际社区一致。

## II. 定义和以前的工作

如前所述，社区并没有唯一的定义。社区的一个最简单的定义是Clique，也就是每一对节点之间都有一条边的节点群。悬崖捕获了社区的直观概念[6]，其中每个节点都与其他每个节点相关，因此彼此之间有很强的相似性。Palla等人在[14]中使用了这一定义的扩展，他们将一个社区定义为相邻的Cliques链。他们定义两个*k*个cliques（*k*个节点上的cliques）是相邻的，如果它们共享*k*-1个节点。这些定义是严格的，因为即使没有一条边，也意味着一个clique（以及社区）不再存在。*K* clans和*K* clubs是更宽松的定义，同时仍然保持社区内边的高密度[14]。如果任何一对节点之间的最短路径长度，或群体的直径，最多为*k*，则称一组节点形成*k*族。*k*俱乐部的定义与此类似，只是由节点组引起的子网工作是网络中直径为*k*的最大子图。

基于组内节点的度（边的数量）相对于组外节点的度的定义
是由Radicchi等人给出的[15]。如果$d^{in}$和$d^{out}$是

*i*在其组内和*组外*的节点的度数，那么如果$d_i^{in} > d_i^{out}$，则称*U*形成了一个*强*社群。$i \in U$.
如果$\sum_{iEU\,i} d^{in} > \sum_{iEU\,i} d^{out}$，则*U*是一个社区，*i*在*软弱*意义。其他基于节点度数的定义可以在[6]中找到。

网络中可能存在许多不同的节点分区，以满足社区的特定定义。在大多数情况下[4,22,26-28]，由社群检测算法发现的节点组被认为是社群，而不考虑它们是否满足特定的定义。为了找到其中的最佳社群结构，我们需要

一个可以量化社区强度的措施。衡量社区强度的方法之一是将社区内观察到的边的密度与整个网络中的边的密度进行比较[6]。如果在一个社区*U*内观察到的边的数量是$e_U$，那么在网络中的边在节点对之间均匀分布的假设下，我们可以计算出*U*内预期的边的数量大于$e_U$的概率*P*。如果*P*很小，那么社区内观察到的密度就大于预期的值。最近Newman[13]也采用了类似的定义，其中比较的是社区内边的观察密度和随机网络中相同社区内边的预期密度，但仍保持每个节点的度。这被称为*模块化度量Q*，其中$Q = \sum_i (e_{ii} - a_i^2)$，即$e_{ii}$是观察到的边缘部分。组内*i*和一个$a_i^2$，是预期的边缘内的比例。请注意，如果$e_{ij}$是网络中运行在*i*组和*j*组之间的边的比例，那么$a_i = \sum_j e_{ij}$。$Q = 0$意味着在一个给定的分区中，组内边缘的密度不超过随机机会所预期的程度。*Q*值越接近于1，表明社区结构越强。

给定一个有*n*个节点和*m*条边的网络*N*(*n*, *m*)，任何社区检测算法都能找到节点的子群。让$C_1$, $C_2$, …, $C_p$是发现的社区。在大多数算法中，找到的社区满足以下条件：(i) $C_i \cap C_j = 0$ for $i \neq j$ and (ii) $\cup C_i$跨越*N*的节点集。

一个值得注意的例外是Palla等人[14]，他们将communities定义为相邻*k*个cliques的链，并允许community重叠。在网络中找到所有这样的群体需要指数级的时间。他们用这些集合来研究社会和生物网络中社群的重叠结构。通过形成另一个网络，其中一个社区由一个节点代表，节点之间的边表示重叠的存在，他们表明这种网络在其节点度分布中也是异质性的（肥尾）。此外，如果一个社区与其他两个社区有重叠的区域，那么相邻的社区也极有可能重叠。

一个网络*N*(*n*, *m*)的不同分区的数量是$2^n$，并且随着*n*的增加而呈指数增长。因此，我们需要一个快速的方法来找到相关的分区。Girvan和Newman[5]提出了一种基于*边际中心*概念的分割算法--"*边际中心*"。*的*最短路径数，也就是在所有成对的
网络中经过该边缘的节点。主要的
这里的想法是，在社区之间运行的边缘有
比起那些位于社区范围内的，有更高的间性值。

的关系。通过连续地重新计算和删除具有最高介值的边，网络被分解成互不相连的部分。该算法一直持续到网络中的所有边都被移除。该算法的每一步都需要$O(mn)$时间，由于有*m*条边要被移除，最坏情况下的运行时间是$O(m^2 n)$。随着算法的进行，人们可以构建一个树状图（见图1），描述网络被分解成不同的部分。

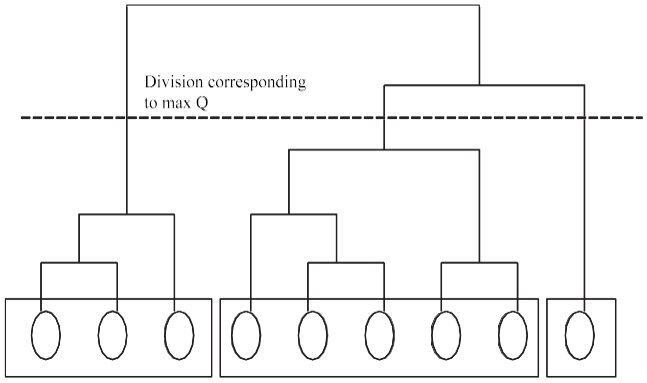


图1.树状图是一个树状图，代表了节点被分隔成不同群体或社区的顺序。

联合连接组件。因此，对于任何给定的$h$，如$1<h<n$，最多只能找到一个网络的分区，即$h$个不相交的子群。树状图中所有这样的分区都被描述出来，而不考虑每个分区中的子群是否代表一个社区。Radicchi等人[15]提出了另一种划分算法，其中树状图被修改为只反映那些满足社区特殊定义的组。此外，他们使用一种叫做*边缘聚类系数的*局部测量方法来代替边缘的中心性，作为去除边缘的标准。边缘聚类系数被定义为一个给定的边缘所参与的三角形的数量占可能的此类三角形总数的比例。一个边的聚类系数被认为是在社区之间运行的边的最小值，因此该算法通过移除聚类系数低的边来进行。总的运行时间
这个除法的算法是$O(\frac{m^2}{n})$。

同样，我们也可以在节点之间定义一个拓扑相似性，并进行聚类分层[23,29]。在这种情况下，我们从$n$个不同社区的节点开始，将最相似的社区组合起来。Newman[22]提出了一种使用模数$Q$*的*合并方法（类似于聚类方法），在每一步中，那些引起$Q$*值*最大增加或最小减少的两个社群被分组在一起。这个过程也可以用树状图来表示，人们可以穿过树状图来找到对应于$Q$值最大值的分区（见图1）。在算法的每一步，人们最多比较$m$对组，最多需要$O(n)$时间来更新$Q$值。该算法一直持续到所有的n个节点都在一个组中，因此该算法的最坏情况下的运行时间是$O[n(m+n)]$。Clauset等人[30]的算法是对这种聚集式分层方法的改编，但使用了一种巧妙的数据结构来存储和检索更新$Q$*值*所需的信息。
效果，它们将算法的时间复杂度降低至
$O(md\log n)$，其中$d$是树状图的深度。
在具有多尺度群落的层次结构的网络中，d ~log n。在具有分层结构的网络中，在许多尺度上都有群落，$d$ ~log $n$。还有其他启发式和基于模拟的方法，可以找到给定网络的部分，使模块化措施$Q$最大化[10,24]。

标签泛滥算法也被用于检测网络中的社区[27,28]。在[27]中，作者提出了一种本地社区的检测方法，其中一个节点被初始化为一个标签，然后通过邻居一步一步地传播，直到它到达社区的末端，从社区向外的边的数量下降到阈值以下。在找到网络中所有节点的本地社区后，形成一个$n$ X $n$ matrix，如果节点$j$属于从$i$开始的社区，则*第ij*项为1，否则为0。然后，矩阵的行被重新排列，使相似的行相互靠近。然后，从第一行开始，他们陆续将所有的行纳入一个社区，直到两个连续的行之间的距离很大，超过一个阈值。在这之后，一个新的社区被形成，这个过程继续进行。形成矩阵的行并重新排列需要$O(n^3)$的时间，因此该算法很耗时。

Wu和Huberman[26]提出了一种线性时间[$O(m+n)$]算法，可以将一个给定的网络分成两个社区。假设可以找到两个属于两个不同社区的节点（$x$和$y$），那么它们就分别以1和0的值初始化。所有其他节点的初始化值为0。然后在算法的每一步，所有的节点（除了$x$和$y$）都按以下方式更新它们的值。如果$z_1$，$z_2$，……，$z_k$是一个节点$z$的邻居，那么值$V_z$是更新为$\frac{Vz1+Vz2+\cdots+Vzk}{k}$。这个过程一直持续下去，直到收敛-证据。作者表明，迭代过程会达到一个唯一的值，而且算法的收敛性不取决于网络的大小$n$。一旦获得了所需的收敛性，就在0和1之间对数值进行排序。按降序浏览数值谱，在两个社区的边界会有一个突然的下降。这个差距被用来识别网络中的两个社区。Flake等人[9]使用了类似的方法来寻找WWW网络中的社区。在这里，给定一小部分节点（源节点），他们形成了一个网页网络，这些网页与源节点的距离在一定范围内。然后通过指定（或人为地引入）汇节点，他们解决从源到汇的最大流量。这样，我们就可以找到与最大流量相对应的*最小切口*。移除切割集后，网络中包含源节点的连接部分就是所需的社区。

频谱分割方法[25]已被广泛用于将一个网络分成两组，使组间的边数最小化。分割过程中使用了给定网络的拉普拉斯矩阵（$L$）的特征向量。可以证明，$L$只有真实的非负的特征值（0 <A1 <A2 <An），并且最小化组间的边数与最小化组间的边数相同
ing的正线性组合$M=\sum_i s_i^2 A_i$，其中$s_i$

$=u_i^T z$ 和 $u_i$ 是$L$的特征向量 对应于 $i$ 艾. $Z$是决策向量，其第1项可以是1或-1，表示节点$i$*属于*哪个组。为了使$M$最小化，$z$被选择为与第二小的特征值对应的特征向量尽可能平行。(最小的特征值是0，选择$z$平行于相关的

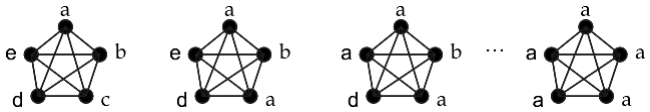


图2.当我们从左到右移动时，节点被逐一更新。由于边的密度很高（在这种情况下是最高的），所有的节点都获得相同的标签。

对应的特征向量给出一个微不足道的解决方案。)这种分割方法已经被扩展到寻找网络中的社区，使模块化程度$Q$最大化[25]。$Q$可以写成矩阵$B$的特征值的正线性组合，其中$B$被定义为两个矩阵$A$和$P$的差。$A_{ij}$是节点$i$和$j$之间的观察到的边数，$P_{ij}$是如果边随机落在节点之间，同时保持每个节点的程度，$i$和$j$之间的预期边数。由于$Q$必须是最大化的，所以$z$被选择为尽可能平行于最大特征值所对应的特征向量。

由于许多现实世界中的复杂网络规模很大，社区检测算法的时间效率是一个重要的考虑因素。当没有关于给定网络中可能存在的社群的*先验*信息时，通常使用寻找优化所选择的社群强度的分区的方法。我们在本文中的目标是开发一个简单的时间效率算法，该算法不需要任何先验信息（如社区的数量、大小或中心节点），仅使用网络结构来指导社区检测。下一节将详细介绍这种算法的拟议机制，它不对任何特定的测量或功能进行优化。

### III. 使用标签传播的社区检测

我们的标签传播算法的主要思路如下。假设一个节点$x$有邻居$x_1$，$x_2$，…，$x_k$，并且每个邻居都有一个标签，表示他们所属的社区。那么$x$就根据其邻居的标签来确定其社区。我们假设网络中的每个节点都选择加入其最大数量的邻居所属的社区，并均匀地随机打破联系。我们用唯一的标签初始化每个节点，并让这些标签通过网络。随着标签的传播，密集连接的群体的节点迅速就一个独特的标签达成共识（见图2）。当整个网络中形成了许多这样密集的（共识）群体时，它们继续向外扩展，直到有可能为止。在propagation过程结束时，具有相同标签的节点被分组为一个社区。

我们迭代地执行这个过程，在每一步，每个节点根据其邻居的标签来更新其标签。这个更新过程可以是同步的，也可以是异步的。在同步更新中，节点$x$在*第t次*迭代中根据其邻居在*第t-1*次迭代中的标签更新其标签。因此，$C_x(t)=f(C_{x_1}(t-1),\ldots,C_{x_k}(t-1))$，其中$c_x(t)$是节点$x$在时间$t$的标签。问题是。

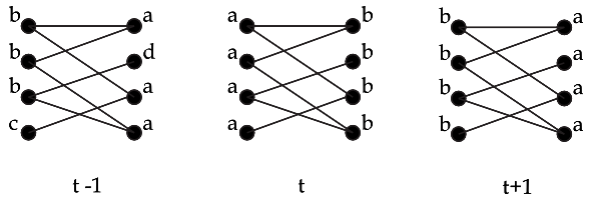


图3.一个两部分网络的例子，其中两部分的标签集是不相交的。在这种情况下，由于节点在步骤$t$所作的选择，节点上的标签在$a$和$b$之间摇摆不定。

然而，网络中的子图在结构上是二方或接近二方的，会导致标签的震荡（见图3）。在社区采取星形图形式的情况下，这种情况尤其真实。因此，我们使用异步更新，其中$C_x(t)=f(C_{x_{i1}}(t),\ldots,C_{x_{i(m+1)}}(t),C_{x_{ik}}(t-1),\ldots,C_x(t-1))$和$x_{i1}$，…，$x_{im}$是$x$的邻居，在当前迭代中已经被更新，而$x_{i(m+1)}$，…，x是在当前迭代中尚未更新的邻居。，$x_{ik}$是在当前迭代中还没有更新的邻居。网络中所有n个节点在每次迭代中被更新的顺序是随机选择的。请注意，虽然我们在算法开始时有$n$个不同的标签，但标签的数量会随着迭代的进行而减少，结果是有多少个独特的标签就有多少个共性。

理想情况下，这个迭代过程应该持续到网络中没有节点改变其标签。然而，网络中可能有一些节点在两个或更多社区中拥有相等的最大邻居数。由于我们在可能的候选者中随机地打破平局，即使其邻居的标签保持不变，这类节点的标签也可能在迭代中发生变化。因此，我们执行迭代过程，直到网络中的每个节点都有一个标签，其邻居的最大数量属于这个标签。通过这样做，我们得到了一个网络的分区，将其分为不相干的社区，其中每个节点在其社区内拥有的邻居数量至少与其他社区相同。如果$C_1$，…，$C_p$是当前网络中活跃的标签，$d_i^{C_j}$是节点$i$与标签$C_j$的节点的邻居数量，那么当每个节点$i$都是这样时，算法就停止了。

如果$i$有标签$C_m$，那么$d_i^{C_m}>d_i^{C_j}$。

近乎线性时间的算法来检测。

在迭代过程结束时，具有相同标签的节点被分组为社群。我们对所获得的社区的停止标准与Radicchi等人[15]提出的*强社区的*定义相似（但不完全相同）。虽然强社区要求每个节点在其社区内的邻居数量严格多于社区外的邻居数量，但通过标签传播过程获得的社区要求每个节点在其社区内的邻居数量至少与它在其他社区的邻居数量相同。我们可以用以下步骤来描述我们提出的标签传播算法。

(i) **初始化网**络中所有节点的标签。对于一个给定的节点$x$，$C_x(0)=x$。

(ii) 设置$t=1$。

(iii) 将网络中的节点按随机顺序排列，并将其设置为$X$。

(iv) 对于按该特定顺序选择的每个$x \in X$，让$C_x(t) = f(C_{i1}(t), \ldots, C_{i(m-1)}(t), C_{i(m+1)}(t-1), \ldots, C_{ik}(t-1))$。这里返回在邻居中出现频率最高的标签，并且均匀地随机打破平局。

(v) 如果每个节点的标签都是其邻居的最大数量，那么就停止算法。否则，设置$t = t + 1$，并转到③。

由于我们在算法开始时，每个节点都有一个独特的标签，头几次迭代的结果是各种小块（密集区域）的节点形成一个共识（需要相同的标签）。然后，这些共识组获得动力，并试图获得更多的节点以加强该组。然而，当一个共识组到达另一个共识组的边界时，它们开始竞争成员。如果组间边缘少于组内边缘，那么组内节点的相互作用可以抵消来自外部的压力。当群组之间达到全局一致时，算法就会收敛，并确定最终的社区。请注意，即使网络作为一个单一的社区满足停止标准，在具有潜在社区结构的异质网络的情况下，这个群体形成和竞争的过程不鼓励所有节点获得相同的标签。在诸如Erdős-Rényi随机图[31]等没有社区结构的同质网络的情况下，标签传播算法将这些图的Giant连接部分识别为一个单一社区。

我们的停止准则只是一个条件，而不是一个被最大化或最小化的措施。因此，不存在唯一的解决方案，而且将网络划分为多个不同的组就能满足停止准则（见图4和图5）。由于该算法是均匀地随机打破联系，在迭代过程的早期，当联系的可能性很高时，一个节点可能投票支持一个随机选择的社区。因此，从相同的初始条件下，可以达到多种社区结构。

如果我们知道网络中可能作为各自社区的吸引中心的节点集合，那么用唯一的标签初始化这些节点就足够了，剩下的节点就没有标签了。在这种情况下，当我们应用所提出的算法时，未被标记的节点将有倾向于从其最接近的吸引者那里获得标签并加入该社区。此外，限制初始化标签的节点集将减少算法所能产生的可能解决方案的范围。由于在确定社区本身之前，通常很难确定哪些节点是社区的中心，因此在这里，我们在算法的开始阶段给予所有节点同等的重要性，并为它们提供独特的标签。

我们将我们的算法应用于以下网络。第一个是Zachary的空手道俱乐部网络，这是一个空手道俱乐部34名成员之间的友谊网络[32]。在一段时间内，由于领导权问题，俱乐部分裂成两个派别，每个成员都加入了两个派别中的一个。我们考虑的第二个网络是美国的空手道俱乐部。

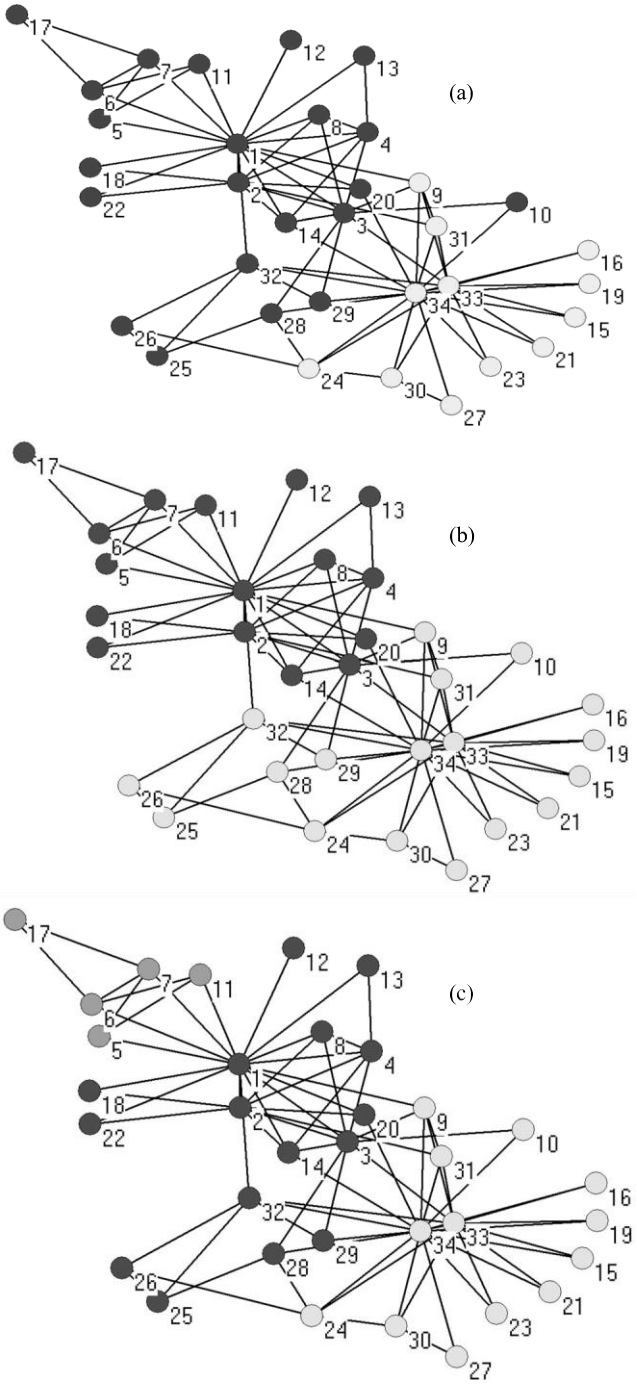


图4.(a)-(c)是算法在Zachary的空手道俱乐部网络上确定的三种不同的社区结构。这些社区可以通过其灰色的深浅来识别。

该网络由115个大学球队组成，作为节点，并在2000年常规赛期间相互比赛的球队之间有边[5]。这些球队被划分为会议（社区），每支球队在自己的会议内进行的比赛比会议间的比赛多。接下来是由16个726名科学家组成的合著网，他们在www.arxiv.org凝聚态物质档案中发布了预印本；边连接了合著论文的

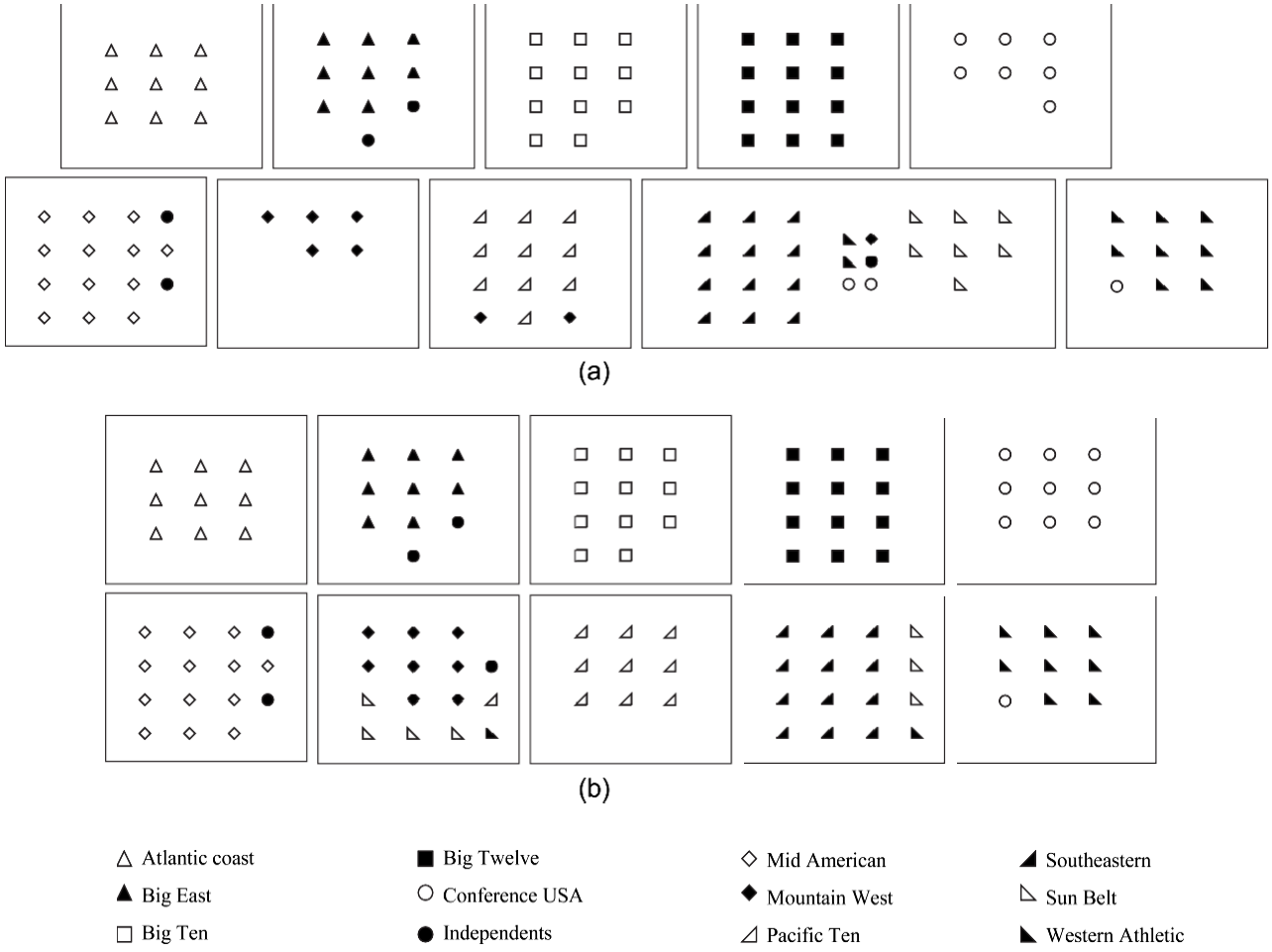


图5.美国大学足球队的分组情况见（a）和（b）。每个解决方案[(a)和(b)]都是在大学橄榄球网络上应用该算法得到的五个不同解决方案的总和。

显示，合著网络中的社区是由在同一领域工作的研究人员组成的，或者是研究小组[22]。沿着类似的思路，我们可以期待一个行动者
协作网络有包含类似类型的演员的社区。在这里，我们考虑一个由374
511个节点组成的演员合作网络，以及在一起演过至少一部电影的演员之间的边[3]。我们还考虑了一个由2115个节点组成的蛋白质-蛋白质交互网络[34]。
社区可能反映了这个网络的功能分组。最后，我们考虑了WWW的一个子集（nd.edu域内的325729个网页和连接它们的超链接[2]。这里的社区被认为是类似主题的网页群。

### A. 多种社区结构

图4显示了为扎克尔空手道俱乐部网络获得的三种不同的解决方案，图5显示了为美国大学足球网络获得的两种不同的解决方案。我们将表明，尽管我们获得了不同的解决方案（社区结构），但它们彼此之间是相似的。为了找到两个不同的解决方案中被归入同一群体的节点的百分比，我们形成一个矩阵$M$，其中$M_{ij}$是一个解决方案中的社区$i$和另一个解决方案中的社区$j$共有的节点数。然后我们计算出$f_{\text{same}}=\frac{1}{2}(L_i \max_j\{M_{ij}\}+L_j \max_i\{M_{ij}\})^{100}$。给定一个网络在已经知道社区的情况下，社区检测算法通常根据被归入正确社区的节点的百分比（或数量）进行评估[22,26]。$f_{\text{same}}$是类似的，通过固定一个方案，我们评估其他方案与固定方案的接近程度，反之亦然。虽然$f_{\text{same}}$可以识别一个解决方案与另一个解决方案的接近程度，但是它对错误的严重性不敏感。例如，当一个解决方案中几个不同社区的节点在另一个解决方案中被融合为一个社区时，$f_{\text{same}}$的值不会有太大变化。因此，我们也使用Jaccard指数，该指数已被证明对解决方案之间的这种差异更加敏感[35]。如果$a$代表在两个解决方案中被归入同一社区的节点对，$b$代表在第一个解决方案中属于同一社区而在第二个解决方案中不同的节点对，$c$反之亦然，那么Jaccard指数被定义为$\frac{a}{a+b+c}$。它的数值在0到1之间，数值越大，说明两个解决方案之间的相似度越高。图6显示了在同一网络上应用该算法五次所得到的解决方案之间的相似性。对于一个

| | | | | |
|---|---|---|---|---|
| -- | 83.8235 | 86.7647 | 97.0588 | 92.6470 |
| 0.5545 | -- | 91.1764 | 86.7647 | 91.1764 |
| 0.6274 | 0.7084 | -- | 89.7058 | 94.1176 |
| 0.8813 | 0.6098 | 0.6955 | -- | 89.7058 |
| 0.8188 | 0.7102 | 0.7908 | 0.7272 | -- |

| | | | | |
|---|---|---|---|---|
| -- | 91.3043 | 92.1739 | 90 | 94.7826 |
| 0.7294 | -- | 89.5652 | 84.7826 | 89.5652 |
| 0.6971 | 0.6223 | -- | 83.9130 | 88.6956 |
| 0.6572 | 0.5903 | 0.5213 | -- | 95.2173 |
| 0.7586 | 0.6787 | 0.5871 | 0.8477 | -- |

Karate club friendship network

| | | | | |
|---|---|---|---|---|
| -- | 86.5248 | 85.8156 | 85.7210 | 85.3901 |
| 0.8569 | -- | 86.2884 | 86.4302 | 85.7683 |
| 0.8316 | 0.8398 | -- | 86.0520 | 83.9007 |
| 0.8430 | 0.8551 | 0.8373 | -- | 85.4137 |
| 0.7998 | 0.8159 | 0.7869 | 0.8124 | -- |

US College football network

| | | | | |
|---|---|---|---|---|
| -- | 84.6735 | 84.7213 | 83.9680 | 85.1488 |
| 0.6019 | -- | 84.4553 | 84.8050 | 85.7683 |
| 0.6082 | 0.5985 | -- | 84.5958 | 84.5689 |
| 0.5909 | 0.6082 | 0.5974 | -- | 85.8543 |
| 0.6036 | 0.6042 | 0.6036 | 0.6254 | -- |

Protein-protein interaction network

Q : 0.738 – 0.745

Co-authorship network

Q : 0.720 – 0.722

| | | | | |
|---|---|---|---|---|
| -- | 91.9743 | 91.9000 | 92.2589 | 92.0706 |
| 0.5431 | -- | 92.0527 | 92.2535 | 92.3695 |
| 0.5358 | 0.5013 | -- | 91.9867 | 92.3490 |
| 0.5454 | 0.5576 | 0.4883 | -- | 92.4904 |
| 0.5425 | 0.5921 | 0.5056 | 0.5703 | -- |

World Wide Web

Q : 0.857 – 0.864

| | | |
|---|---|---|
| -- | 94.5946 | 91.8218 |
| 0.8284 | -- | 92.2034 |
| 0.6026 | 0.6810 | -- |

Actor collaboration

Q : 0.437 – 0.528

图6.每个网络获得的五个不同的解决方案之间的相似性被列成表格。每个表格的下三角的第i行和*第j*列中的条目是相应网络的解决方案*i*和*j的*Jaccard相似性指数。表格上部三角形的第i行和*第j*列中的条目是各自网络中的解决方案*i*和*j的*测量值$f_{\text{same}}$。还给出了每个网络中五个不同解决方案的模块化值*Q的*范围。

给定的网络，表格下部三角形的*第ij个*条目是解决方案*i*和*j*的Jaccard指数，而上部三角形的*第ij个*条目是解决方案*i*和*j*的衡量标准$f_{\text{same}}$。我们可以看到，从五种不同的解决方案中得到了

运转情况相似，这意味着所提出的标签传播算法可以有效地识别任何给定网络的社群结构。此外，五种解决方案的模块化程度*Q的*范围很窄且数值很高（图6），表明这些分区表示重要的社群结构。

### B. 总数

很难在七种不同的解决方案中挑选出一种最好的解决方案。此外，一个解决方案可能能够识别一个在其他解决方案中没有发现的群落，反之亦然。因此，所有不同解决方案的集合可以提供一个包含最有用信息的社区结构。在我们的案例中，一个解决方案是网络中节点上的一组标签，所有具有相同标签的节点构成一个社区。给

将它们合并如下；让$C^1$表示解决方案1中节点的标签，$C^2$表示解决方案中节点的标签。

2.然后，对于一个给定的节点$x$，我们定义一个新的标签为$C_x = (C_x^1, C_x^2)$（见图7）。从一个初始化的网络开始定两个不同的解决方案，我们将

我们对标签C进行迭代处理，直到网络中的每个节点都在一个社区中，其邻居的数量达到最大。当新的解决方案出现时，它们会被逐一与聚合解决方案结合起来，形成一个新的聚合解决方案。请注意，当我们聚合两个解决方案时，如果一个解决方案中的社区$T$被分成两个（或更多）不同的社区$S_1$和$S_2$，那么通过定义上述的新标签，我们显示出对较小的社区$S_1$和$S_2$的偏好。这只是不同解决方案可以被聚合的众多方式之一。关于社区检测中使用的其他聚合方法，请参考[26,36,37]。

图8显示了在不同类型的产品之间的相似性。实验。该算法在每个网络上应用了30次，并记录了解决方案。*第ij*项是前5i个解决方案总和的Jaccard指数，其ag---。

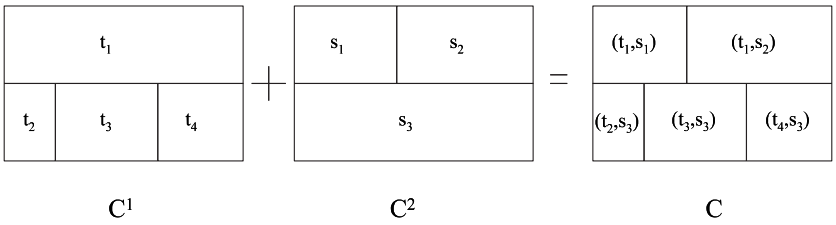


图7.两种群落结构方案汇总的例子。$t_1$ ， $t_2$ ， $t_3$ ， $t_4$ 是方案一中得到的网络中节点的标签，表示为$C^1$。该网络被划分为具有相同标签的节点组。$s_1$ ， $s_2$ ， $s_3$ 是在解决方案2中得到的同一网络中的节点的标签，表示为$C^2$。所有在解决方案1中具有标签$t_1$ 的节点被分成两组，每组分别具有标签$s_1$ 和$s_2$ ， 而所有在解决方案1中具有标签$t_3$ ， $t_4$ ， 或$t_5$ 的节点在解决方案2中具有标签$s_3$。$C$代表由$C^1$ 和$C^2$ 定义的新标签。

前5j个解决方案的集合。我们观察到，这些集合的解决方案在本质上是非常相似的，因此，一小部分解决方案（在这种情况下是5个）可以提供关于网络社区结构的洞察力，就像一个更大的解决方案集一样。特别是WWW网络，单个解决方案之间的相似度很低（Jaccard指数范围为0.4883-0.5931），但总体解决方案之间的相似度明显提高（Jaccard指数范围为0.6604-0.7196）。

### IV. 社区检测算法的验证

由于我们知道Zachary的karate俱乐部和美国足球网络中存在的社区，我们明确验证通过对这些网络的应用，我们对该算法的准确性进行了评估。我们发现，该算法可以有效地挖掘出各自网络中的潜在社区结构。图4显示了在Zachary的空手道俱乐部网络上使用我们的算法得到的社区结构。虽然这三种解决方案都是应用于网络的算法的结果，但图4（b）反映了真正的解决方案[32]。

图5给出了美国大学足球的两种解决方案网络。该算法在这个网络上应用了10次，两个解决方案是前五个和其余五个解决方案的总和。在图5(a)和5(b)中，我们可以看到该算法可以有效地识别除Sunbelt以外的所有会议。造成这种差异的原因如下：在七支球队中

US college football network

| | | | | | |
|---|---|---|---|---|---|
| -- | 0.7455 | 0.7713 | 0.7504 | 0.8851 | 0.7053 |
| 0.7805 | -- | 0.9277 | 0.6853 | 0.7585 | 0.6417 |
| 0.8777 | 0.8814 | -- | 0.6867 | 0.7817 | 0.6508 |
| 0.8777 | 0.8814 | 1 | -- | 0.8256 | 0.6512 |
| 0.8777 | 0.8814 | 1 | 1 | -- | 0.7888 |
| 0.7805 | 1 | 0.8814 | 0.8814 | 0.8814 | -- |

Karate club friendship network

Co-authorship network

| | | | | | |
|---|---|---|---|---|---|
| -- | 0.7691 | 0.7291 | 0.7368 | 0.7349 | 0.7578 |
| 0.8926 | -- | 0.7560 | 0.7561 | 0.7597 | 0.7722 |
| 0.8927 | 0.8827 | -- | 0.7360 | 0.7322 | 0.7604 |
| 0.9002 | 0.8887 | 0.8942 | -- | 0.7717 | 0.7712 |
| 0.9003 | 0.8803 | 0.8885 | 0.8966 | -- | 0.7642 |
| 0.9011 | 0.8864 | 0.8852 | 0.9062 | 0.8966 | -- |

Protein-protein interaction network

| | | |
|---|---|---|
| -- | | |
| 0.6545 | -- | |
| 0.7196 | 0.6604 | -- |

World Wide Web

图8.每个网络获得的总体解决方案之间的相似度。表中第i行和第j列的条目是前5i个解决方案和前5j个解决方案集合之间的Jaccard相似度指数。空手道俱乐部友谊网络和蛋白质-蛋白质相互作用网络的解决方案之间的相似性表现在前两个表格的下方三角形中，而这两个表格的上方三角形中的条目分别是美国大学足球网络和共同著作网络。第三张表的下三角中给出了WWW的聚合解决方案之间的相似性。

在Sunbelt会议中，四支球队（$Sunbelt_4$={北德克萨斯州，阿肯色州，爱达荷州，新墨西哥州}都互相交过手，三支球队（$Sunbelt_3$={路易斯安那门罗，中田纳西州，路易斯安那拉斐特}）再次互相交手。只有一场比赛连接着$Sunbelt_4$和$Sunbelt_3$，即North-Texas和Louisiana-Lafayette之间的比赛。然而，来自阳光地带会议的四支球队（阳光地带$_4$和阳光地带$_3$，各两支）与东南地区会议的七支不同的球队一起比赛。因此，在图5(a)中，我们将阳光地带会议和东南地区会议归为一组。在图5(b)中，Sunbelt会议一分为二，$Sunbelt_3$与东南大学分组，$Sunbelt_4$与一支独立球队（犹他州立大学）、一支来自西大西洋的球队（博伊西州立大学）和Moun-tain West会议分组。后者的分组是由于$Sunbelt_4$的每个成员都与犹他州立大学和博伊西州立大学打过比赛，他们一起与Mountain West的四个不同的球队打过五场比赛。还有五支独立球队不属于任何特定的会议，因此被算法分配到一个他们打过最多比赛的会议。

## V. 时间复杂度

该算法需要近乎线性的时间来运行到完成。用唯一的标签初始化每个节点需要$O(n)$的时间。标签传播算法的每一次迭代都需要边的数量的线性时间[$O(m)$]。在每个节点$x$，我们首先根据他们的标签对邻居进行分组[$O(d_x)$]。然后，我们挑选最大的一组，并将其标签分配给$x$，最坏情况下需要$O(d_x)$的时间。这个过程在所有节点上重复，因此每次迭代的总时间为$O(m)$。

随着迭代次数的增加，被正确分类的节点数量也在增加。这里我们假设，如果一个节点的标签是其邻居的最大数量，那么该节点就被正确分类。根据我们的经验，我们发现无论$n$是多少，95%以上的节点在迭代5结束时都被正确分类。即使是$n$在100到10 000之间、平均度数为4的Erdős-Rényi随机图[31]，也没有社区结构，到迭代5结束时，95%以上的节点被正确分类。在这种情况下，该算法将巨型连接部件中的所有节点确定为属于一个社区。

当算法终止时，有可能两个或更多的断开连接的节点组有相同的标签（这些组在网络中通过其他不同标签的节点连接）。这种情况发生在一个节点的两个或多个邻居收到它的标签，并以不同的方向传递标签，这最终导致不同的社区采用相同的标签。在这种情况下，在算法终止后，我们可以在每个单独组的子网络上运行一个简单的广度优先搜索，以分离不相干的社区。这需要的总体时间为$O(m+n)$.然而，在汇总解决方案时，我们很少发现社区内有不相连的群体。

## VI. 讨论和结论

建议的标签传播过程只使用网络结构来指导其进展，不需要外部参数设置。每个节点根据其近邻的社区做出自己的决定，重新确定其所属的社区。这些本地化的决定导致了特定网络中社区结构的出现。我们用Zachary的空手道俱乐部和美国大学足球网络验证了该算法所发现的社区结构的准确性。此外，模数指标$Q$对所有得到的解决方案都很重要，表明该算法的有效性。每次迭代需要线性时间$O(m)$，尽管人们可以观察到算法在大约五次迭代后开始明显收敛，但数学上的收敛性很难证明。其他以类似时间尺度运行的算法包括Wu和Huberman的算法[时间复杂度为$O(m+n)$]和Clauset等人的算法[30]，其运行时间为$O(n \log^2 n)$。

Wu和Huberman的算法被用来打破一个给定的网络只分为两个社区。在这个迭代过程中，两个选定的节点被初始化为标量值1和0，每个节点将其值更新为其邻居值的平均值。在收敛过程中，如果一个节点的邻居的最大数量的值高于一个给定的阈值，那么该节点也会如此。因此，一个节点倾向于被归入其最大数量的邻居所属的社区。同样，如果在我们的算法中，我们选择相同的两个节点，并为它们提供两个不同的标签（让其他节点没有标签），标签传播过程将产生与Wu和Huberman算法类似的社区。然而，为了在网络中找到两个以上的社群，吴和胡伯曼算法需要*先验地*知道网络中有多少个社群。此外，如果知道网络中有$c$个社区，Wu和Huberman提出的算法只能找到规模大致相同的社区，即$n/c$，而不可能找到规模不同的社区。与Wu和Huberman的算法相比，我们提出的标签传播算法的主要优点是，我们不需要关于给定网络中社区数量和大小的*先验*信息；事实上，这种信息在现实世界的网络中通常是不可用的。另外，我们的算法不对社区的大小进行限制。它仅通过使用网络结构来确定社区的这些信息。在我们的测试网络中，标签传播算法发现社区的大小近似于幂律分布$P(S > s) \sim s^{-\nu}$，指数$\nu$在0.5和2之间（图9）。这意味着网络中不存在特征性的社群大小，这与以前的观察结果一致[22,30,38]。虽然WWW和合著者网络的社群规模分布大致遵循幂律，其指数分别为1.15和1.98，但对于WWW和合著者网络来说，有一个明显的从一种比例关系到另一种比例关系的交叉。

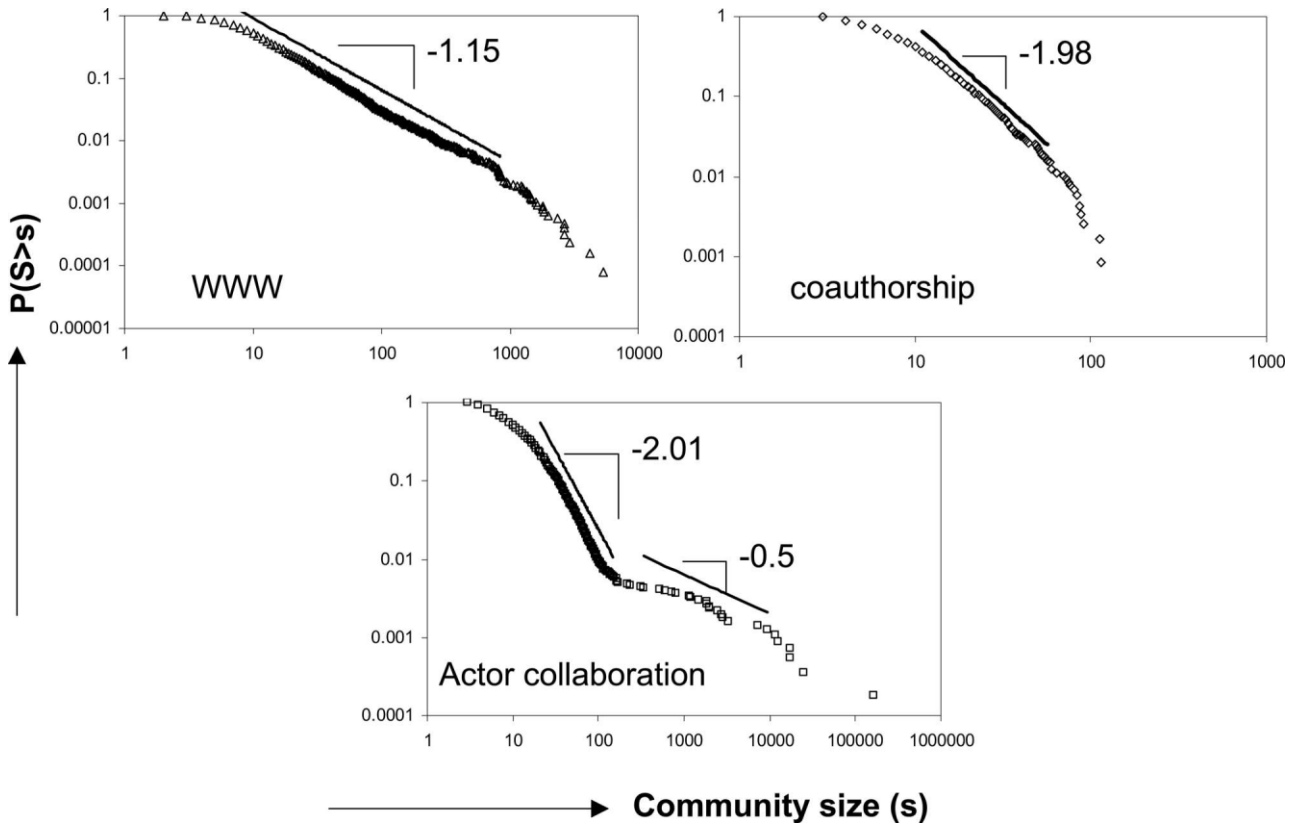


图9.图中显示了WWW、合著和演员合作网络的社区规模（*s*）的累积概率分布。它们大致遵循幂律，其指数如图所示。

行为者协作网络。行为体协作网络的社区规模分布，在164个节点以下的规模中，幂律指数为2，在164至7425个节点之间为0.5（见图9）。

在Clauset等人[30]的分层聚类算法中，对应于最大*Q值*的分区被认为是最能说明网络中社区结构的分区。其他具有高*Q*值的分区将具有与最大*Q值*分区类似的结构，因为这些解决方案是通过每次逐步聚集两个群体而获得的。另一方面，我们提出的标签传播算法可以找到多个明显的模块化解决方案，这些解决方案具有一定的不相似性。特别是对于WWW网络来说，五个不同的这五种解决方案都很低，Jaccard指数在0.4883和0.5921之间，但所有这五种都是显著的模块化，*Q值*在0.857和0.864之间。这意味着所提出的算法不仅可以找到一个，而且可以找到多个重要的社区结构，支持许多现实世界网络中存在的重叠社区[14]。

## 鸣谢

作者要感谢美国国家科学基金会（拨款号：SST 0427840，DMI 0537992，和CCF 0643529）。其中一位作者（R.A.）感谢斯隆基金会的支持。

[1] R.Albert和A.-L. Barabási，Rev. Mod.**74**, 47 (2002).
[2] R.Albert, H. Jeong, and A.-L. Barabási, 自然（伦敦）**401**, 130 (1999).
[3] A.-L.Barabási和R. Albert，《科学》**286**，509（1999）。
[4] M.Newman, SIAM (Soc. Ind. Appl. Math. ) Rev. **45**, 167 (2003).
[5] M.Girvan和M. Newman, Proc.Natl. Acad.Sci. U.S.A. **99**, 7821 (2002).
[6] S.Wasserman和K.Faust，*社会网络分析*（Cambridge大学出版社，英国剑桥，1994年）。
[7] L.Danon, A. Díaz-Guilera, and A. Arenas, J. Stat.Mech.: Theor.Exp. **2006** P11010 (2006).
[8] J.Eckmann and E. Moses, Proc.Natl. Acad.Sci. U.S.A. **99**, 5825 (2002).
[9] G. Flake, S. Lawrence, and C. Giles, Proceedings of the 6th ACM SIGKDD, 2000, pp.150-160.
[10] R.Guimerà和L. Amaral，《自然》（伦敦）**433**，895（2005）。
[11] M.Gustafsson, M. Hornquist, and A. Lombardi, Physica A **367**, 559 (2006).
[12] M.B. Hastings, Phys. Rev. E **74**, 035102(R) (2006).

[13] M.E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[14] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, 自然（伦敦）杂志 **435**, 814 (2005).
[15] F.Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc.Natl. Acad.Sci. U.S.A. **101**, 2658 (2004).
[16] D.Karger, J. ACM **47**, 46 (2000).
[17] B.Kernighan and S. Lin, Bell Syst.Tech.J. **29**, 291 (1970).
[18] C.Fiduccia and R. Mattheyses, Proceedings of the 19th Annual ACM IEEE Design Automation Conference, 1982, pp.175-181.
[19] B.Hendrickson and R. Leland, SIAM (Soc. Ind. Appl. Math. ) J.Sci. Comput.**16**, 452 (1995).
[20] M.Stoer和F. Wagner, J. ACM **44**, 585 (1997).
[21] C.Thompson, Proceedings of the 11th Annual ACM Symposium on Theory of Computing, 1979, pp. 81-88.
[22] M.E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[23] P.Pons and M. Latapy, e-print arXiv:physics/0512106.
[24] J.Duch和A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[25] M.E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[26] F.Wu和B. Huberman， Eur.Phys. J. B **38**, 331 (2004).
[27] J.P. Bagrow和E. Bollt, Phys. Rev. E **72**, 046108 (2005).
[28] L.Costa, e-print arXiv:cond-mat/0405022.
[29] M.E. J. Newman, Eur.Phys. J. B **38**, 321 (2004).
[30] A.Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111（2004）。
[31] B.Bollobás, *Random Graphs* (Academic Press, Orlando, FL, 1985).
[32] W.Zachary, J. Anthropol.Res. **33**, 452 (1977).
[33] M.Newman, Proc.Natl. Acad.Sci. U.S.A. **98**, 404 (2001).
[34] H.Jeong, S. Mason, A.-L. Barabási, and Z. Oltvai, Nature (London) **411**, 41 (2001).
[35] G. Milligan和D. Schilling, Multivariate Behavior.Res. **20**, 97 (1985).
[36] D.Gfeller, J. C. Chappelier, and P. De Los Rios, Phys. Rev. E. **72**, 056135 (2005).
[37] D.Wilkinson and B. Huberman, Proc.Natl. Acad.Sci. U.S.A. **101**, 5241 (2004).
[38] A.Arenas, L. Danon, A. Díaz-Guilera, P. Gleiser, and R. Guimerà， Eur.Phys. J. B **38**, 373 (2004).