# 大规模社交网络中最大共处社区搜索

## 摘要

k-truss搜索的问题已被很好地定义和调查，以找到社交网络中高度相关的用户组。 但是，之前没有研究考虑用户空间信息在k-truss搜索中的约束，本文称为共置社区搜索。 共处社区可以为许多实际应用提供服务。 为了有效地搜索最大的共处社区，我们首先使用几种修剪技术开发一种有效的精确算法。 之后，我们进一步开发了一种具有可调精度保证的近似算法，并探索了更有效的修剪规则，这可以显着降低计算成本。 为了加快实时效率，我们还设计了一种新的基于四叉树的索引，以支持区域内用户的有效检索，并根据给定的查询区域优化搜索区域。 最后，我们使用五个真实数据集验证了我们提出的算法和索引的性能。

**1.引言**

随着在线社交网络的日益普及，社交网络数据分析中最重要的任务之一是找到彼此具有紧密结构联系的用户社区。关于寻找社区的广泛研究可分为全球社区检测GCD（例如[19,20,21,34,16,6]），当地社区检测LCD（例如[14,43]），全球社区搜索GCS （例如，[31,30,35]）和当地社区搜索LCS（例如，[18,39,14,13,24,17,44,23]）。社区检测方法通常用于基于预定义的隐式标准（例如模块化[19]）发现社交网络中的社区。 GCD和LCD之间的主要区别在于每个用户在GCD中测量同等重要，而用户的重要性取决于他与LCD中给定查询顶点的相关性。与社区检测不同，社区搜索方法集中于根据用户指定的显式标准从社交网络中查找社区，例如基于k核心的模型中的参数k [38]，基于k-truss的模型[12]和k基于边缘连接的组件模型[4]。与社区检测方法类似，GCS和LCS之间的主要区别在于LCS要求社区包含给定的查询顶点，但GCS没有这样的附加要求。但是，上述大多数作品都没有考虑用户空间信息在社区检测或搜索方法中的影响。

搜索具有社交和空间凝聚力的社区在许多应用中非常重要，例如，事件安排，产品推荐，有针对性的广告，本地激进主义和倡导，以及更有效的内容传播，如商店促销，本地新闻和职位空缺。虽然空间特征在应用中是非常需要的，但在实践中，现有的空间社会社区研究仍然有限。在[17]中，Fang等人。要求返回的k-core社区的所有顶点都在半径最小的最小覆盖圆中，并且结果社区必须包含给定的查询顶点。所以它是一种具有空间约束的LCS。在[16,6]，Expert等人。和陈等人。通过基于两个链接用户的空间距离对链接进行加权来考虑社区检测过程中的空间信息。它是一种具有空间约束的GCD。然而，这两类工作不能保证社区成员的空间亲密度，这将在我们的实验中进一步讨论。在[45]中，张等人。要求返回的基于k核的社区的所有顶点满足相似性约束，其中相似性可以是距离相似性。然而，由于其NP-硬度，在大规模社交网络中找到该社区模型的确切结果是昂贵的。

因此，在本文中，我们研究共处社区搜索问题，揭示具有高社会和空间凝聚力的最大社区，表示为（k，d）-MCCs搜索。 使用最小桁架值k [12]来定义社会凝聚力，并且通过用户指定的距离值d来参数化空间凝聚力。 因此，我们提出的（k，d）-MCC搜索问题可以使用户容易地确认所得社区的质量，这也填补了具有空间约束的GCS类型的研究空白。

给定社交网络G和两个参数k和d，直接的方法是枚举G满足最小桁架值k的所有可能子图，其中子图的数量可以与O（2 n）一样大。 然后，它过滤具有节点对的候选者，其距离高于空间接近度阈值d。 因此，这种方法的时间复杂度至少为O（2 n），其中n是G中的顶点数。显然，使用这种方法支持在线（k，d）-MCC搜索是不可行的，特别是对于大型 规模社交网络。 因此，本文着重于设计有效的算法，以实现理论保证的实时响应。

为了解决效率的挑战，我们首先通过提出新的修剪技术来开发精确（k，d）-MCCs搜索算法。 在搜索过程中，我们通过考虑基于早期终止的上限，启发式搜索顺序以及重用剪枝计算的条件，探索显着修剪搜索空间的技术。 在搜索之前，我们还提出了用于减少输入数据大小的预修剪技术。 为了设计多项式算法，我们开发了一种具有空间精度保证的新近似模式。 注意，我们提出的近似方案可以根据用户对空间精度的要求提供可调节的空间误差率。 为了进一步提高近似算法的性能，我们提出了更多的修剪技术，并设计了新的索引TQ树。 我们工作的主要贡献概括如下。

•我们提出了一种新颖的共存社区模型，并正式定义了（k，d）-MCCs搜索问题。 （第2节）

•我们通过在搜索之前和搜索期间提出有效的修剪技术，开发了一种有效的精确算法，用于查找（k，d）MCC。 （第3节）

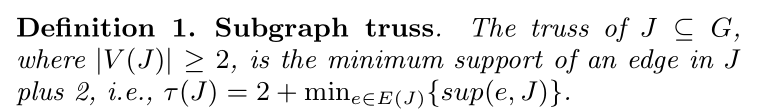
•我们还开发了一种空间近似算法，可提供从2√2+到√2+的可变空间误差率？ 0。 通过提出更有效的修剪技术和新颖的TQ树索引，进一步提高了近似算法的效率。 （第4节）

•我们对五个真实数据集进行了广泛的实验研究，以证明所提算法的效率和有效性。 （第5节）

**2.问题定义**

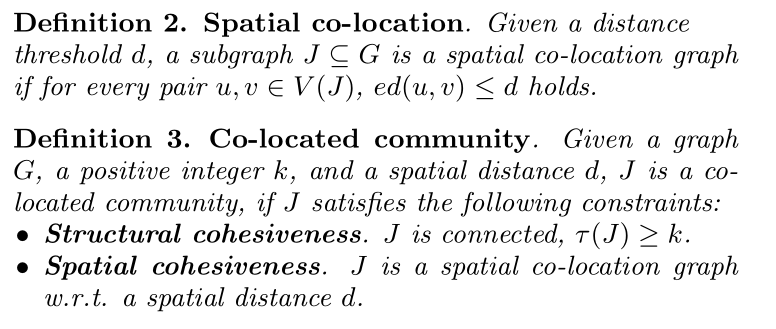
我们考虑社交网络图G =（V，E），它是具有顶点集V（G）和边集E（G）的无向图，其中顶点表示社交用户，边表示他们的友谊。 对于每个顶点v∈V（G），它具有空间属性（v.x，v.y），其中v.x和v.y表示其在二维空间中沿x轴和y轴的空间位置。 共处社区。 共同定位的社区是一个子图J⊆G满足：（1）连通性：J连通，（2）结构凝聚力：J中的所有顶点密集连接，（3）空间凝聚性：J中的所有顶点在空间上都是紧密的 彼此。

结构凝聚力。 我们将桁架视为衡量共处社区结构凝聚力的指标。 桁架测量图中每条边所涉及的三角形数量。 给定J，让我们将包含顶点u，v，w∈V（J）的三角形表示为4 uvw。 由sup（e，J）表示的边e（u，v）∈E（J）的支持是包含e的三角形的数量，即sup（e，J）= | {4 uvw：w∈ N（v，J）∩N（u，J）} |，其中N（v，J）和N（u，J）相应地是V，u的邻域。 接下来，我们定义共处的社区j的桁架如下：

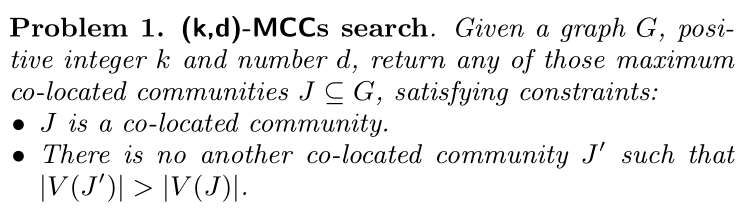


如果它是连接的并且τ（J）≥k，则J是连通的k-truss。 直观地，k-truss是子图，其中每个连接（边缘）（u，v）具有至少k-2个共同邻居。 具有大值k的k-truss表示对成员的强内部连接。 在k-truss中，每个节点应该具有至少k-1的度数，这意味着k桁架必须是（k-1）-core。 连接的k-truss也是（k-1）-edge-connected。

空间凝聚力。 让ed（u，v）表示顶点u和v之间的空间距离。我们首先介绍空间协同定位的概念来测量空间凝聚力。 然后我们正式定义共处社区。

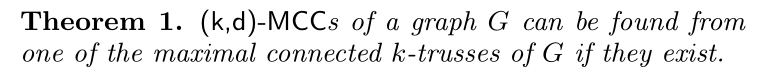


通常，在搜索社区时，用户可能希望在他们修复空间和社会凝聚力参数后最大化社区中包含的成员。 因此，在本文中，给定图G，我们研究找到最大共处社区，表示为（k，d）-MCC，其中k代表k-truss，d代表空间距离，M代表最大值，CC代表co - 社区。 现在我们正式定义（k，d）-MCCs搜索的问题。



例如，在图1（b）中，深蓝色区域中的顶点是共同定位的。 类似地，在图1（a）中，三个可能的共同定位的社区处于蓝色区域，其中k = 4.这里的（4，d）-MCC是包含顶点{d，e，f，g，h的子图， 我有基数6，因为它是最大值。

我们可以通过检查整个图来从G中找到（k，d）-MCC。 但是，为了提高搜索性能，我们只想搜索可能包含（k，d）MCC的G部分。 为此，我们引入下面的定理：



证明是微不足道的，因为不是最大连通k-truss的一部分的顶点显然不能满足定义3中的结构凝聚力要求。

根据定理1，在G中找到（k，d）-MCC的直观步骤包括：（1）计算最大连通k-trusses（注意：这些k-trusses是非重叠的），（2）搜索本地（k） ，d）在这些k-桁架中的每一个中的-MCC，以及（3）通过比较基数来找到来自本地的全局（k，d）-MCC。 分析。 对于图G，可以在O（| E（G）| 3 2）内建立k-truss指数.G的k-truss指数基本上是与由τ（e，G）定义的边缘桁架相关联的边的列表。 ）=maxH⊆G∧e∈E（H）{τ（H）} [24]。 使用k-truss索引，给定k，我们可以在多项式时间内检索G中的所有最大连通k个桁架。 然而，由于以下原因在最大连通k-truss中找到局部（k，d）MCC仍然具有挑战性：（1）k-truss中空间共址子图的总数可以是指数[22]和（22） 2）托管子图的大小与其共处社区的大小之间没有保证的单调关系。