A Local Algorithm for Structure-Preserving Graph Cut

现如今，大规模的图数据可以从社交网络、共同作者网络、蛋白质交换网络、公共交通网络等应用中产生。许多在图像挖掘方面的工作以一阶马尔科夫链为基础模型，研究重点为图中的节点和边。但是这些工作在应用到更重要的领域中的高阶图的结构时却不行，比如在银行的用户身份鉴定信息网络中，星型的网络结构对应于一组组合身份；在财务传输网络中，环形结构常常意味着洗钱行为的发生。在这篇文章中，我们注重于挖掘***特定用户***的高阶网络结构以及希望找到一种能够在不用破坏很多原有边的基础上将之从原图中分离出来的子图，称为***structure-rich***的子图。

在寻找这样的子图过程中最重要的是避免过高的计算成本。为了解决这个问题，受局部图聚类算法族（用来高效定位低导通性的分割图且不用探索整个图）的启发，我们提出通过规范化这个关键思想来对高阶网络结构建模。In particular，我们从高阶导通性的一般定义开始，定义了高阶扩散核，这是以用户指定的高阶网络结构引发的高阶随机游走为基础。然后我们提出一个全新的High-Order Structure-Preserving LOcal Cut(HOS-PLOC)算法，该算法运行时间为图中的边的数量的polylogarithmic time。算法从一个种子节点开始，迭代地探索它的邻域，直到找到一个具有小的高阶导通性的子图。此外，我们从效率和效率两方面分析了它的表现。在合成图和实际图上的实验结果证明了我们提出的HOSPLOC算法的有效性和有效性。

A Temporally Heterogeneous Survival Framework with Application to Social Behavior Dynamics

社会行为动力学是理解和建模复杂的社会动态现象（如信息传播，意见形成和社会动员）的核心基石之一。尽管近年来提出了多种社会行为动力学模型，但社会行为动力学的基本成分和最小模型仍然没有得到解答。在这里，我们通过探索大规模的社交沟通数据集发现人类互动行为动力学在响应时间维度和自然时间维度上展现了丰富的复杂性。为了应对这一挑战，我们开发了一个temporal Heterogeneous Survival framework，其中响应时间维度和自然时间维度的规律可以有机地结合在一起。将我们的模型应用在两个在线社交沟通数据上，结果表明我们的模型能够成功完成对社交沟通数据集中交互模式的再现，并且所提出的方法可以显着超越其他最先进的baseline。同时，学习的参数和发现的统计规律可能导致多种潜在的应用。

Ego-Splitting Framework: from Non-Overlapping to

Overlapping Clusters

我们提出了ego-splitting，一种新的用于复杂网络中的聚类检测框架，通过利用ego-net来对重叠的聚类进行解耦。Ego-splitting是一个高度可扩展和灵活的框架，具有可证明的理论保证，将复杂的重叠聚类问题简化为一个更简单，更易于处理的非重叠（分区）问题。我们可以将社区检测扩展为具有数百亿边缘的图，并且比以前的基于ego-net的其他解决方案表现更好。

更准确来说，我们的框架按照以下两个步骤运行：一个局部的ego-net分析阶段，和一个全局的图像分区阶段。在局部的阶段，我们首先使用分区算法对节点的ego-net进行分区，然后使用计算的群集将每个节点拆分成代表其社区中节点的实例化的角色节点。最后，在全局的步骤中，我们对新创建的图进行分割，以获得原始图的重叠聚类

FORA: Simple and Effective Approximate Single-Source

Personalized PageRank

给定一个图G，一个源节点s以及一个目标节点t，节点t关于节点s的personalized PageRank(PRR)就是从s开始结束于t的随机游走的概率。一个single-source PPR（SSPPR）查询枚举G中的所有节点，并返回关于源节点s中具有最高PPR值的前k个节点。SSPPR在网络检索核社交网络中具有重要的应用，比如Twitter的“关注谁”的推荐服务。但是，SSPPR的计算代价十分昂贵，应用中无法满足实时性。目前的解决方法要么使用没有保证质量的启发式算法，或者使用十分昂贵的高计算能力的数据中心。

因此，我们提出了FORA，一种简单而有效的基于索引的对SSPPR过程的近似解决方法。FORA的基本思想是以一种简单而又平凡的方式将两种现有的方法Forward Forward（快速但不保证质量）和Monte Carlo Random Walk（准确但缓慢）结合起来，从而产生一种快速和准确的算法。除此之外，FORA包含了一个简单而有效的索引方案以及一个具有高度可裁剪性质的top-k选择模块。大量的实验表明，FORA比其主要竞争对手的效率高出数个数量级。值得注意的是，在一个数十亿的Twitter数据集上，FORA使用一台商用服务器，在5秒钟内的除了top-500的近似SSPPR查询。

Graph Edge Partitioning via Neighborhood Heuristic

我们考虑将输入图的边分割成多个平衡部分的边分割问题，同时最小化重用的顶点总数（一个顶点可能出现在多个分区中）。对于几个大规模分布图计算平台（例如，PowerGraph，Spark GraphX）来说，这个问题对于最小化通信成本和运行时间至关重要。