

面向复杂网络的图稀疏算法综述

徐丽丽 董一鸿 潘剑飞 陈华辉

(宁波大学信息科学与工程学院 浙江 宁波 315211)

摘 要 大规模数据下复杂网络的算法分析面临复杂度高的挑战,为此引入图稀疏的思想,在保持原始图性质的情况下以一定的精度在稀疏图上实现了高效的算法分析。图稀疏算法是一种保留顶点、对边稀疏采样的方法。按照相应算法分析所需要的原始图性质,提出图稀疏的边度量方式。文中系统回顾了4种边度量下的图稀疏采样方法:生成图稀疏、边连通图稀疏、聚类图稀疏、边传播性图稀疏,归纳了不同边度量方式下图稀疏的优缺点和适应性,并进一步讨论了动态图流稀疏化的最新研究进展。最后,总结了图稀疏领域有待解决的问题并展望了未来的研究方向。

关键词 复杂网络,图结构,图稀疏,边的度量

中图分类号 TP391 文献标识码 A DOI 10.11896/j.issn.1002-137X.2018.05.004

Survey of Graph Sparsification Algorithms for Complex Networks

XU Li-li DONG Yi-hong PAN Jian-fei CHEN Hua-hui

(Institute of Information Science and Technology, Ningbo University, Ningbo, Zhejiang 315211, China)

Abstract With the increasing of data scale, traditional graph algorithms confront the challenge of excessive complexity. The idea of graph sparse was introduced into algorithm, and analytical algorithm was realized efficiently on the sparse graph while preserving the original property with certain accuracy precision. Graph sparsity algorithm is a sampling method which preserves the original vertexes and sparsifies the edges. In this paper, the latest progresses of graph sparse algorithm were reviewed from four aspects, such as sparse spanning algorithm, edge connectivity sparse, clustering sparse and influence propagation sparse. The advantages and disadvantages of the graph sparse algorithms and the adaptability were summarized in different edge metrics. In addition, dynamic graph streaming sparse methods were analyzed. Finally, several important problems were outlined and future research directions in the field of sparse complex network were prospected.

Keywords Complex network, Graph structure, Graph sparsification, Edge measurement

1 引言

对于现实生活中常见的复杂系统,如交通网络、在线社交网络,可以自然地将其建模成复杂网络来进行抽象表达。图作为一种数据结构,被应用到复杂网络的描述中^[1],其中网络实体用图的顶点表示,实体之间的联系用关联边表示。近年来,随着在线社交及网络技术的飞速发展,真实的网络结构包含了数百亿的顶点和数万亿的边,网络的数据规模呈现指数增长趋势,从而产生了海量图数据^[2]。用于分析复杂网络的算法有:最短路径、最大流最小割、社区发现、预测影响力的传播、图的聚类分析等。大规模的结构阻止了上述算法对复杂网络有效的图分析^[3-6]。面对计算复杂度高、效率低的挑战,如何有效解决大规模复杂网络的算法分析已经成为研究热点^[7]。

许多学者提出了图压缩方法来缩减数据规模。图压缩^[8-10]是通过将原始图数据压缩成更小的存储形式来保证对大规模图的分析。常见的形式有:1)将网络中具有相似或相同邻居的节点压缩成超级节点^[11-12],但压缩后的图保留相应原始节点和超级节点之间的边联系;2)网络中节点的数目不变,通过去掉一些冗余边来减小复杂网络的规模。本文主要讨论第二种情况——图稀疏,即顶点全保存而边稀疏的采样。图稀疏算法移除对图结构有影响的边,原始图和稀疏图之间具有很好的相似性,有助于建立更精准的分析模型。本文主要总结了面向复杂网络的图稀疏算法,分析这些算法的优缺点和适应性;并进一步针对近年来逐渐兴起的动态图流稀疏^[13-14]算法进行了分析和讨论;最后总结了图稀疏领域面临的挑战并展望了未来的发展方向。

收稿日期:2017-08-16 返修日期:2017-11-13 本文受国家自然科学基金(61572266),宁波市自然科学基金(2017A610114),浙江省自然科学基金(LY16F020003)资助。

徐丽丽(1993—),女,硕士,主要研究方向为大数据、数据挖掘;董一鸿(1969—),男,博士,教授,主要研究方向为大数据、数据挖掘和人工智能, E-mail: dongyihong@nbu.edu.cn(通信作者);潘剑飞(1991—),男,硕士,主要研究方向为大数据、数据挖掘;陈华辉(1964—),男,博士,教授,主要研究方向为数据流与数据挖掘。

2 基本概念

假设复杂网络 $G=(V, E)$ 是由 $|V|=n$ 个节点和 $|E|=m$ 条边连接成的一个无向原始图 G , 若存在带权子图 $H=(V, E', w)$ 满足条件 $\forall A \subset V, (1-\epsilon)\lambda_A(G) \leq \lambda_A(H) \leq (1+\epsilon)\lambda_A(G)$ (其中 λ_A 是原始图上重要的性质), 则称 H 是图 G 的稀疏图^[15]。

图稀疏算法产生的稀疏图 H 能在 $\epsilon \in (0, 1)$ 误差范围内保持原始图 G 的特征。下面给出常见的 3 种基于相似性而提出的图稀疏算法的定义。

定义 1 (生成图)^[16] 稀疏图 H 保持原始图 G 两顶点之间最短距离的相似性, 是基于计算几何学的图稀疏。

$$(1-\epsilon)\text{dist}_G(u, v) \leq \text{dist}_H(u, v) \leq (1+\epsilon)\text{dist}_G(u, v) \quad (1)$$

定义 2 (割稀疏)^[17] 稀疏图 H 保持原始图 G 边割值的相似性, 是基于边连通度的图稀疏。

$$(1-\epsilon)\lambda_G(U) \leq \lambda_H(U) \leq (1+\epsilon)\lambda_G(U) \quad (2)$$

定义 3 (谱稀疏)^[17] 稀疏图 H 保持原始图 G 的拉普拉斯矩阵的谱值, 其具有相似性, 是基于代数计算的图稀疏。

$$(1-\epsilon)x^T L_G x \leq x^T L_H x \leq (1+\epsilon)x^T L_G x \quad (3)$$

3 面向复杂网络的图稀疏方法

由于稀疏图要保持原始图的性质, 因此图稀疏算法根据不同的边度量方式对原始网络图进行边采样, 形成稀疏图, 以加速图相应的算法分析。本文将已有的图稀疏算法归纳为 4 类, 如图 1 所示。

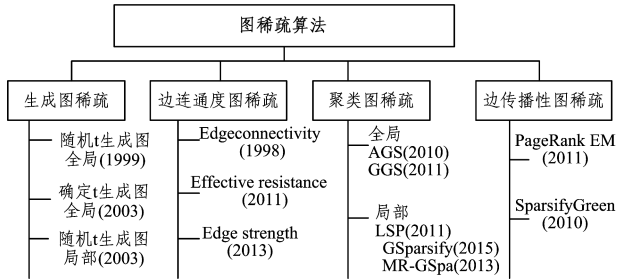


Fig. 1 Classification and appearance time of graph sparse

3.1 生成图稀疏

路径是图的全局特征, 指图连通的情况下顶点之间边最小的数目值。生成图稀疏采用保持原始图中顶点之间距离性质的边度量方式, 典型的算法有 t -spanner。

Cohen^[16]提出了生成图的随机算法来计算 t -spanner。 t -spanner 的衡量关键是延展因子 t 的大小和生成图边的数目, 它们决定了算法的时间复杂度。确定 t -spanner 的方法如下^[20]: 对于有权图的构建, 与最小生成树克鲁斯卡尔算法类似, 即初始化边集为空, 按照边权重递增的形式来处理, 如果增加某边前, 生成图顶点 (u, v) 的最短距离大于 t -spanner, 则加入这条边, 否则将其丢弃。这些算法都需要计算每对顶点之间的最短距离, 是一个全局生成图稀疏算法。Baswana^[21]提出生成图局部随机算法, 该算法仅仅需要探索一个或一系列顶点的关联边, 从而可以通过外存计算和分布式计算构成稀疏生成图。算法分两个阶段进行: 1) 形成聚类, 即独立

地以概率 $\frac{1}{\sqrt{n}}$ 采样, 形成第一批顶点集合, 然后以采样到的顶点为聚类中心点再次计算, 依次处理余下的顶点; 2) 将顶点和相邻的簇相连, 如果顶点 v 不与任意采样点关联, 则把与 v 关联的边都加入生成图, 如果与 v 关联的采样顶点有一个或多个, 则将距离顶点最近的边加入生成图, 并将权重小于上述最近边权重的关联边加入生成图, 最后将 v 变成聚类中心点。算法丢弃了两个顶点都没有被采样的边, 且它们归属于同一类。

生成图稀疏删除对复杂网络的特征路径没有影响的边, 而适当地对原始图边的数目进行压缩, 在保证算法精确度的同时加速了最短路径算法的求解。

3.2 边连通稀疏

边连通稀疏基于图的连通度进行边的度量, 其主要思想是以不同的边连通度 λ 对边赋值, 然后依照边权值设定边稀疏采样概率。3 种度量性质为: 边关联度 (edge connectivity)、边强度 (edge strength) 和边有效电阻 (effective resistance), 如表 1 所列。Hariharan^[17]提出将边的关联度作为边连通度的上界对边进行采样, 并证明了算法在误差范围内的可行性; 割稀疏^[20]按照边强度进行采样; 谱稀疏^[21]算法按照边的有效电阻来进行计算。基本方法是按照度量方式为每条边赋予一个 p_e 值, 并以概率 p_e 独立地对边进行采样, 对采样到稀疏图上的边再按照 $1/p_e$ 为其赋值, 从而形成带权值的子图 H 。

表 1 3 种边连通度的比较

Table 1 Comparison of three kinds of edge connectivity

符号	名称	意义
k_{st}	边关联度	分离边两个顶点最小边的数目 (割值)
k'_{st}	边强度	包含边 (s, t) 最大 k 连通量
$1/C_{st}$	边有效电阻	图中连接顶点对路径的有效电阻

3.2.1 边关联度稀疏

边关联度是分离边 (s, t) 顶点的最小割值。Hariharan^[17]将带权边用 w 条无权边代替, 无权图独立地以概率 $p_e = \min\{\rho/\lambda_e, 1\}$ 进行边采样, 如果被采样到稀疏子图 H 中, 则赋权值 $1/p_e$, 其中 $\rho = \Theta(\log n/\epsilon^2)$, λ_e 的取值即为边连通度。对于任意图 G , 边的关联度都是连通度的上界值。具体实例请参考图 2; 按照表 1 中各个边连通度的定义: 1) 如果将边 (s, t) 的顶点 s 和 t 分离, 则至少删除 $n-1$ 条边, 从而关联度 $k_{st} = n-1$; 2) 包含边 (s, t) 的最大连通 k 值为 2, 边强度 $k'_{st} = 2$; 3) 每条边的电阻值为 1, 顶点 s 和 t 是由 $n-2$ 个 two-path 路径和一条边组成, 则 $c_{st} = (n-2)/2 + 1 = n/2$, 边有效电阻 $1/c_{st} = 2/n$ 。

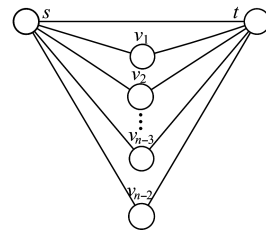


图 2 边 (s, t) 连通度 λ_e 的实例

Fig. 2 Example of connectivity λ_e of edge (s, t)

3.2.2 割稀疏

图上割的大小包含图的结构信息^[24], 例如有权图的顶点

度数是由大小为 $|k_{st}|=1$ 的割所给出的。最大流最小割算法的每一次增广都是一次 BFS, 效率为 $O(m)$, 最坏情况下的时间复杂度为 $O(m * n)$, 由于稀疏图中边的数目大大减少, 因此该算法效率较高。文献[17]中 Benczur 和 Karger 的割稀疏算法如下: 独立地以概率 p 对原始图进行采样, 最后形成的稀疏子图由 $O(pm)$ 条边组成。上述算法为均匀采样, 需要知道最小割值来获得最小误差界限, 当割值较小时, 算法失效。Benczur 和 Karger 进一步提出非均匀采样, 将图分解成 K 个强连通部分, K 连通部分的边数至多为 $K * n$, 且最小割值为 K 。对每个连通部分的内部进行均匀采样, 不同 K 连通部分的采样概率不同, 即与边的强度 k'_{st} 成反比; 同时为采样到的边进行权重赋值。形成的割稀疏子图由 $O(n \log n / \epsilon^2)$ 条边组成。

3.2.3 谱稀疏

1) 拉普拉斯矩阵与图的连通性

首先, 按照不同的连通性(根据 Cheeger 不等式^[25], 图的连通性与其对应规范化后的拉普拉斯矩阵的第二小特征值有关联)对图进行分解, 然后在每个子图内部进行随机采样, 采样参数的设定依赖于顶点对度。子图连通性较高的情况下, 每个图都有一个谱稀疏。

具体地, 假定 $\epsilon \in (0, 1/2)$, $G=(V, E)$ 是无权图且 G 的拉普拉斯矩阵的最小特征值大于或等于 λ , 则根据每条边的采样概率 $p_e = \min(1, C/\min(d_u, d_v))$ (其中 $C = \Theta((\log n)^2 (\epsilon \lambda)^{-2})$, $\tilde{w}=1/p_e$), 可得一个 $(1+\epsilon)$ 谱稀疏图 $\tilde{G}=(V, \tilde{E}, \tilde{w})$, $\tilde{G}$ 的平均度为 $O((\log n)^2 (\epsilon \lambda)^{-2})$ 。

2) 完全图的谱稀疏

根据谱稀疏的定义考虑完全图稀疏图^[26]的过程, 原始图 G 是具有 n 个结点的完全图, 则拉普拉斯矩阵 L_G 的所有非零特征值都等于 n 。从谱图理论来说, Ramanujan 图是最好的稀疏图, 它是一个 d -正规图, 非零特征值范围为 $[d - 2\sqrt{d-1}, d + 2\sqrt{d-1}]$, 为稀疏图 $\tilde{G}$ 的每一条边都赋值 n/d , 则 $x^T L_{\tilde{G}} x \in [n - \frac{2n\sqrt{d-1}}{d}, n + \frac{2n\sqrt{d-1}}{d}]$ 。由谱稀疏的定义得知, $\tilde{G}$ 是原始图 G 的 $(1 - \frac{2\sqrt{d-1}}{d})^{-1}$ 稀疏。图 3(a) 给出一个包含 12 个顶点的完全图, 其稀疏后的 4-正规图如图 3(b) 所示。

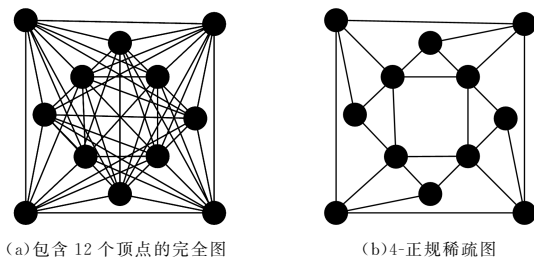


图 3 完全图的谱稀疏

Fig. 3 Spectral sparsity of complete graphs

3) 边有效电阻

与边强度相比, 边有效电阻不仅考虑了路径边的数目, 还考虑了路径长度, 因此谱稀疏是割稀疏的一种扩展, 能更好地存储原始图的相关信息。

边有效电阻的基本性质如下:

①路径的电阻是沿着路径的电阻的总和;

②并联路径的电阻是单个路径电阻的调和平均值;

③将边添加到图形中, 电阻不会增加。

将网络中的每条边识别成一个电阻, $R_{eff}(e)$ 是指顶点之间边的有效阻力, 如图 4 所示。

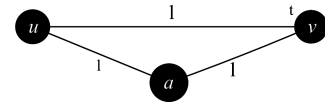


图 4 网络电流图

Fig. 4 Current network diagram

边有效电阻的数值表达(B 关联矩阵)如下:

$$B(e, v) = \begin{cases} +1, & \text{如果 } v \text{ 是边 } e \text{ 的头} \\ -1, & \text{如果 } v \text{ 是边 } e \text{ 的尾} \\ 0, & \text{其他情况} \end{cases} \quad (4)$$

拉普拉斯矩阵: $L = B^T B$

每条边 $e=(u, v)$ 的有效阻力被定义为 $R_{eff}(u, v) = BL^+ B^T$, 其中 L^+ 是图拉普拉斯矩阵的伪逆矩阵, B 是关联矩阵。边排序算法: 首先为图 G 中的每一条边指定属性值 $p_e = w_e R_{eff}(e)$, 然后以一定的概率对原图的边进行采样, 使其满足谱稀疏的要求。用有效电阻对谱稀疏^[28-29]进行随机采样, 可以将有效电阻的计算转化为一些向量的 ℓ_2 范数的计算, 即对称对角占优^[26]问题的求解, 利用线性求解算子来加速谱稀疏。文献[27]提出针对有向图的谱稀疏算法, 仅仅考虑顶点的出度, 构建相应拉普拉斯矩阵执行图稀疏算法, 通过正交映射, 用较少的边向量估计所有向量。

图分割算法在保证子图连通性的基础上去除冗余边, 实现了连通图稀疏在大数据规模分布式^[30]运算上的可行性与有效性。

3.3 聚类图稀疏

聚类图稀疏是指保留按照社区度量的边及即保留社区内部的边及稀疏社区间的边。边度量的标准是顶点相似度, 即如果某条边的顶点具有越多的相同邻居节点, 则该条边的两个顶点属于同一个簇的可能性就越大, 该边越可能为聚类图稀疏内部的边, 需要被保留。图 5 为将聚类稀疏应用到社区检测算法中的模型图。

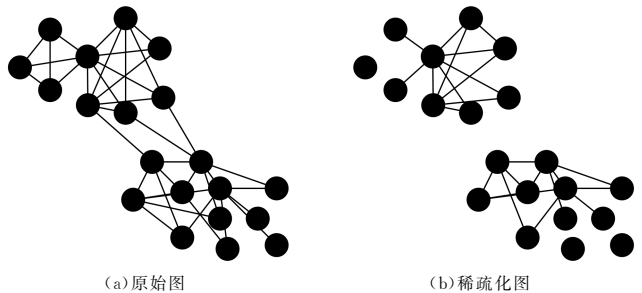


图 5 应用聚类稀疏算法前后的对比

Fig. 5 Comparison of before and after applying clustering sparse algorithm

3.3.1 全局稀疏算法

全局稀疏算法, 即对图中每条边都按照相似度赋值, 首先设定全局阈值 S , 然后按照阈值对相似度高的边采样。

1) 全局稀疏算法(GGS)^[31]。采用 Jaccard 相似度衡量边 $e_{(i,j)}$ 的相似度, 假设 $Adj(i)$ 为顶点 i 的邻居节点集合, $Adj(j)$

为顶点 j 的邻居节点集合,则边相似度 $sim(i, j)$ 的计算式如下:

$$sim(i, j) = \frac{|Adj(i) \cap Adj(j)|}{|Adj(i) \cup Adj(j)|} \quad (5)$$

该算法设定一个比例 $\theta(0 < \theta < 1)$,使得稀疏后的图与原始图之间边的关系满足 $|E_s| = \lceil \theta \times |E| \rceil$,然后根据相似度递减的顺序对边进行排序,最后采样前 $\lceil \theta \times |E| \rceil$ 条边。

2) 自适应全局稀疏(AGS)^[32]。自适应是指在数据处理和分析过程中,特征自动调整方法、顺序、参数、边界,使得特征与处理后数据的统计分布特征和结构特征相适应,以取得最佳的效果。

Tanwistha 和 Huzefa^[32] 针对集合分类提出了一种自适应全局稀疏算法,来解决集合分类中出现的问题。考查带标签(属性值)的顶点,比较两个顶点的属性值的相似性,并按照比例要求采样相似度较高的边;剩下的边若能增加图中连通部分的数目,则将其添加到稀疏图中,否则丢弃。

3.3.2 局部稀疏算法

全局图稀疏算法不适合边密度不均匀的图,因为邻接节点较少可能导致两个节点之间没有边,从而丢失图的一部分结构信息。局部稀疏算法很好地弥补了该缺陷,能保证每部分的拓扑结构信息。图 6 对局部稀疏与全局稀疏进行了对比。

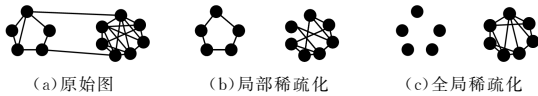


图 6 局部稀疏与全局稀疏的对比

Fig. 6 Comparison of local sparse and global sparse

针对密度不均匀的图,局部稀疏算法主要是在每对顶点之间至少保存一条边。

1) 局部图稀疏(Lsp)^[33]。该类算法避免了全局稀疏密度不均匀的问题,不使用全局阈值。算法扫描图中的每个顶点,然后根据边的 Jaccard 相似度度量和一对顶点之间保持 d^* (其中 d 是考虑稀疏的顶点的度值, e 是事先指定的 0 或 1) 条边的关系进行边采样,确保每一对顶点之间至少有一条关联边被采样,保证了图的拓扑结构信息。

2) 基于小图的局部聚类稀疏算法^[34] (gSparsify)。该类算法对小图使用聚合函数来估计其所包含边的采样率,在局部图稀疏的基础上增加了对 l -clique 频数^[35] 的度量。通常是对图 7 所示的几种类型的小图进行聚类分析。局部稀疏考虑的 Jaccard 相似度是小图的一部分;而 gSparsify 提出的算法综合考虑了几种连通子图,能够很好地保存聚类信息。

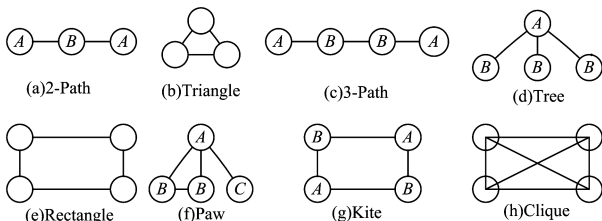


图 7 大图上常见的几类小图

Fig. 7 Several kinds of small templates commonly seen on larger graphs

3) 基于 MapReduce 的局部稀疏算法(MR-GSpar)^[36]。

基于图数据日益普及、规模不断增大的现状,文献[35]提出了一种分布式大规模图稀疏算法 MR-GSpar。该算法在 Map-Reduce 并行计算框架的基础上,对传统的最小哈希算法进行并行化改造,使其可在分布式的集群环境中实现对大规模图数据的高效稀疏化处理。主要步骤如下:Map 任务对每个顶点的邻居节点进行计算;Reduce 任务运行最小哈希算法,计算每个节点的签名矩阵并进行存储,最后设定阈值,进行图稀疏化处理。

对于聚类图稀疏算法,在得到大图聚类之前增加一个稀疏化图的环节,在误差范围内稀疏图能够很好地保持原始图中的聚类结构,算法可在稀疏图上进行图聚类,提高了运行速度。聚类稀疏算法通常分为两大类:全局稀疏和局部稀疏。全局稀疏主要针对密度均匀的图,局部稀疏能保证全局拓扑结构。表 2 比较了各种图稀疏算法的优缺点和适应性。

表 2 聚类稀疏算法的比较

Table 2 Comparison of clustering sparse algorithms

类别	算法	优点	缺点和局限	适用范围
全局稀疏	GGS	提高聚类精度	不适用于密度不均匀的图,计算时需考虑图的密度	小图,密度均匀
	AGS	自适应性 原始图连通性		顶点存储信息结构
局部稀疏	LSP	能够保证全局拓扑结构	仅考虑邻居节点,算法思想简单、不精准	小图,全局拓扑
	GSP	增加小图的频数分布	计算复杂度高,计算时受图规模的限制	小图,簇间差距大
	MRG	利用最小哈希分布式稀疏	存储空间较大,存储图的空间复杂度较高	分布式,大规模图

针对社会网络的社区检测是研究热点。随着数据集规模的增大,噪声边对图结构信息的影响增大,使得聚类稀疏算法的精度有所提高,可有效防止过拟合。

3.4 边传播性图稀疏

社会网络都具有传播的性质。边传播性图稀疏在保持网络传播影响力最大化的前提下,提出边度量方式。假设传播影响力模型 $G=(V, D)$,对于给定的社会网络,其包含 $|V|=n$ 个实体和 $|D|=m$ 个活动 A 的传播路径。

边传播性图稀疏是指找寻对活动 A 传播最重要的 k 条弧组成的稀疏图 $G_s=(V, D_s)$,其中 $|D_s|=k$,以达到影响力最大化。通过随机游走^[37]和 EM 方法求得网络的稀疏传播结构。

3.4.1 随机游走

通过随机游走算法使得稀疏图和原始图结构^[38]相匹配,利用较小的数据集进行分析并将结果传回原始数据集。随机游走算法 PageRank 的最初功能是确定网络结构中顶点的重要性。在处理某些应用时,需要利用 PageRank 向量值度量^[39]边的重要性,其被称为边格林值,它是一种广义的边有效电阻值。典型的随机游走是概率转移矩阵 $P(u, v) = w(u, v) / d(u)$,其中 w 是带权值的邻接矩阵, d 是带权值的度矩阵。通过慢随机游走,格林值稀疏算法首先进行预处理,即通过格林值稀疏原始图,随后解决平衡割问题。稀疏图不需要按

照谱稀疏的方式进行计算,使得谱稀疏的计算速度有所提高。

3.4.2 EM 算法

EM 算法^[40]中原始图的每条路径传播活动 A 的概率为 $p_{e(u,v)}$, 稀疏图与原始图具有相同的传播概率, 即 $p_{D_s}(u, v) = p(u, v)$ 。其主要思想是通过概率模型使得稀疏后的图能够以最大概率传播社会活动。如图 8(a) 所示, Ω 是活动产生的点, 沿着每条弧以概率 $p(u, v)$ 进行延伸。图 8(b) 给出影响的传播路径, 实线是指活动 a 能影响到 v 的顶点集, 定义为 $F_a^+(v)$; 虚线是指活动 a 不能将影响传播到 v , 定义为 $F_a^-(v)$ 。

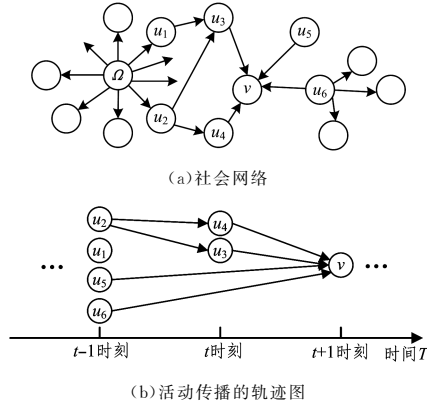


图 8 网络传播影响

Fig. 8 Influence of network communication

传播路径的似然函数为:

$$L_a(G) = \prod_{v \in V} P_a^+(v) P_a^-(v) \quad (6)$$

其中:

$$P_a^+(v) = \begin{cases} 1, & \text{if } F_a^+(v) \neq \emptyset \\ 1 - \prod_{u \in F_a^+(v)} (1 - p(u, v)), & \text{otherwise} \end{cases}$$

$$P_a^-(v) = \prod_{u \in F_a^-(v)} (1 - p(u, v)) \quad (7)$$

最大似然法, 是取对数函数较好的数值解法:

$$\log L_a(G) = \sum_{a \in A} \log L_a(G) \quad (8)$$

以满足式(9)最大为目标:

$$p^{(k+1)}(u, v) = \frac{p^{(k)}(u, v)}{|A_{v|u}^+| + |A_{v|u}^-|} \sum_{a \in A_{v|u}^+} \frac{1}{P_a^+(v)} \quad (9)$$

其中, A 是 a 的集合, $A_{v|u}^+$ 表示某活动中有多少事件 a 对传播 u 产生影响。

作为复杂网络影响力最大化的预处理过程, 边传播稀疏以精简的边进行图分析, 提升了算法效率。

3.5 小结

基于图结构的大数据稀疏算法按照 4 种边的度量形式对原始图进行边的稀疏采样, 移除影响图分析算法的噪声边, 以加快原始图的分析, 同时增强图分析算法的有效性。以上算法都属于静态图稀疏算法, 其中, 生成图稀疏基于路径特性, 能近似求解出原始图中顶点间的距离值; 边连通稀疏基于大图分割, 有利于社会网络的分布式计算; 聚类稀疏基于网络的社区特性, 提高了求解精度; 边传播性稀疏能最大化网络的影响力。表 3 对 4 类静态网络的图稀疏算法存储原始图的重要特性、算法思想及应用场合进行了比较。

表 3 静态复杂网络图稀疏算法的对比

Table 3 Comparison of sparse algorithms for static complex network

稀疏类别	算法	存储性质	主要思想	应用场合
生成图	t 生成图	最短距离相似	边能构成最短距离	特征路径
边连通	割稀疏 谱稀疏	割值相似 谱值相似	边保证图的连通性	图分割
聚类稀疏	全局稀疏 局部稀疏	边的聚类特性	边区分簇间和簇内	社区检测
边传播	格林值图稀疏	边的概率最大化	重要的 k 条边	传播影响

4 动态图流稀疏算法

动态性是社会网络的重要特征, 造成网络算法分析计算复杂度过高。图流稀疏算法^[41-43]使用较少的边能保留动态过程的图属性。下面针对复杂网络动态图流模型的稀疏算法进行总结和比较。单通道半流图稀疏算法是较为典型的动态图流稀疏算法。

4.1 图流模型

假设动态社交网络 $G' = (V, E)$ 由 $|V| = n$ 个节点和 $|E| = m_t$ 条边构成, 它表示仅边发生改变的特定动态网络, 即节点之间边的联系会随着时间动态变化。研究人员提出了动态图流模型来解决动态网络模型的构建问题。图 9 给出了动态图流模型的一些实例。

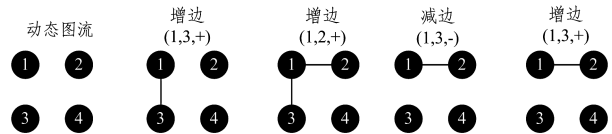


图 9 动态图流模型实例

Fig. 9 Examples of dynamic graph streaming model

按照算法内存大小的限制, 图流模型分为流模型和半流模型两种。图流有不同的形式, 主要分为任意流和关联流两类。任意流中的边是不重复的, 且以任意的顺序出现; 关联流中的边是以顶点的邻边出现的。根据图流通过通道的次数, 又可以将算法分为单通道算法和多通道算法。

1) 流模型

流模型是将图 G 的边以一串流的形式 $(e_1, e_2, \dots)$ 输入, 通过多趟顺序访问数据流的形式解决计算内存有限的问题。流模型主要对滑动窗口的应用进行扩展, 即考虑当前窗口内的任意流对图性质的影响。

2) 半流模型

半流模型的限制较少, 理想情况下每个处理器接受固定的顶点集合, 处理器对边的处理只需要单(少)趟。目前, 研究人员已经将半流模型应用于动态复杂网络最小生成树和最大匹配等相关图拓扑问题的处理中。

半流模型仅与顶点数目有关, 更加适应动态复杂网络, 空间复杂度为 $O(poly \log n)$ 。在动态图流稀疏算法中, 半流模型表示将复杂网络的顶点信息存储在内存中, 将边存储在内存中。下文的稀疏算法都是基于半流模型提出的。

4.2 动态图流稀疏算法

图流稀疏算法^[44-46]是任意流的随机概率模型, 假设独立条件下为每条边分配 p_e 的采样概率。算法的目的是以高概率来近似计算动态边的属性。动态图流稀疏的关键在于对边

的采样概率赋值时须考虑概念漂移,当前的边更新操作形成稀疏图的后续影响,相较于静态图稀疏,边连通度仅仅是估计值。针对复杂网络,如果仅增加边,则是半动态网络;实体之间的边既增加也删除,则是全动态网络。

4.2.1 半动态网络图流稀疏算法

半动态模式中的边以流的形式进入,仅能进行插入操作^[47]。AHN^[48]改进静态割稀疏算法,提出单通道图流割稀疏,其输入 $C = \{e_{i_1}, e_{i_2}, \dots, e_{i_t}\}$ 是图 G 的割,且 $VAL(C) = c$, 边的索引对应着数据流中边到达的顺序。设定一个独立概率事件 A_c 生成稀疏图 H , 使得图流采样是独立事件且具有割值上界,即对于任意的 $\beta(0 < \beta < 2e - 1)$, $P[A_c \wedge (|VAL(C, H) - c| > \beta c)] < 2\exp(-\beta^2 pc/4)$ 。

Kelner^[49]首次提出了单通道图流谱稀疏,即用静态谱稀疏得到稀疏图后,将更新的边集合 Γ 都加入到 H 中,当 H 变大而超出设定的边阈值时,对 $H + \Gamma$ 的所有边进行有效电阻值估计,再稀疏形成图流稀疏。该算法不考虑被丢弃的边。

4.2.2 全动态网络图流稀疏算法

全动态模式解决了边的插入和删除问题。AHN^[50]提出了线性概要结构来描述动态图流,用矩阵形式来分析任意流的变化。定义一个多图(动态流图) $G = (V, E)$, 其中顶点 $V = [n]$, $A(i, j) = |\{k : a_k = (i, j, +)\}| - |\{k : a_k = (i, j, -)\}|$ 表示边 $e_{(i,j)}$ 的插入与删除操作。其中边变化的数据为 $s = \langle a_1, \dots, a_t \rangle$, 其中 $a_k \in [n] \times [n] \times \{-1, 1\}$ 。该线性概要结构可用来评估图上的某些相关属性,如双重性、连通性、 k 连通性、密集子图、最大加权匹配^[51]、最小生成树和生成图^[52]等。线性概要图和稀疏恢复技术^[53]使得图稀疏算法能解决边的删除问题。

GILBERT^[54]首次利用概要图提出边删除的单通道图流割稀疏算法,该算法保存两个数据结构分别用来估计边的连通度和实际采样。其主要思想是通过顶点边关联矩阵 A_G 的行采样估计连通部分,对输入的流以几何概率递减的形式随机采样,在每次采样中寻找连通部分。估计阶段:假设任意边 $(u, v) \in [n]^2$ 都有一个 $h_{(u,v)} \in [0, 1]$ 的均匀随机数,定义一个随机变量 $h_{(u,v)}^b (b = 1, \dots, O(\log n))$ 来估计边的连通度。采样阶段:独立产生随机数 $g_{(u,v)}^b \in [0, 1] (b = 1, \dots, O(\log n))$ 划分顶点便于采样, $g_{(u,v)}^* \in [0, 1]$ 用于决定稀疏图 H 中的采样。如果某条边被采样,则 g, g^*, h 在算法中保持不变,这有助于解决边删除问题,确保某概要图中存在的边被移除。为了压缩线性空间存储的稀疏图,借鉴压缩感知稀疏恢复^[55-56],可以通过概要图稀疏恢复出割稀疏。

KAPRALOV^[55]提出单通道图流谱稀疏算法,鉴于动态图流不能直接求出某条边的有效阻力,借助 ℓ_2 稀疏恢复算法形成谱稀疏。其主要使用线性概要图和稀疏恢复方法。算法分两个部分: MaintainSketches(B, ϵ) 通过 $s \in \{1, \dots, O(\log n)\}$ 和 h_s 对每条边进行 Hash 运算,以概率 $\frac{1}{2^s}$ 独立对原始图关联矩阵 A_G 进行递归采样,构成粗略的线性概要图。RefineSparsifier^[55]在边有效电阻过采样的基础上使用 ℓ_2 heavy hitters 采样方法细化采样,得到谱稀疏。算法使用线性概要图,能简单地处理边删除问题,即将边的删除问题看成负的边插入问题。

4.3 动态图流稀疏算法的比较

动态图流稀疏算法主要是针对特定的动态社会网络,能

够在有限的内存空间解决边采样估计的问题。针对无限图流更新操作采取的方法是:加入滑动窗口,仅考虑当前 ω 条边的更新操作。表4在流模型上对比了动态图流的3种常见稀疏算法。

表4 动态图流中3种常见稀疏算法的比较

Table 4 Comparison of three common sparse algorithms in dynamic graph streaming

类别	半动态模式	全动态模式	滑动窗口 (width ω)
割稀疏	确定	随机	随机
谱稀疏	确定	随机 $\tilde{O}(n^{5/3})$	随机 $\tilde{O}(n^{5/3})$
($2t-1$)生成图	$O(n^{1+1/t})$	仅多通道算法	$O(\sqrt{\omega n^{(1+1/t)}}$)

注:算法的空间复杂度是 $O(n \text{poly} \log n)$

5 研究展望

图稀疏的研究工作刚刚起步,还有以下问题有待解决。

1)网络压缩在保持网络原有性质的前提下,具有良好的结构压缩性,减少了网络的冗余,能运用较少的信息对复杂网络进行信息存储和分析。除了对网络的边进行稀疏采样外,对顶点的压缩和采样也将逐渐为研究人员所关注。

2)伴随着微信、微博、脸谱、腾讯等在线网络的不断发展,网络中包含的信息量呈现爆炸式增长,社交网络的节点数已达到几十亿。压缩和抽样是处理大规模图数据的有效手段,然而,怎样评价图压缩和图采样的效果,如何均衡压缩比和误差的关系使得效果最佳,在限制条件下如何估计估计缩比极限等,还需要更深入的研究和分析。

3)网络中的信息具有动态变化的特性,社交网络中每一时刻都存在着某个用户的增加或者减少,也存在着用户之间关系的变化,这种动态复杂网络的稀疏化和采样是未来研究的热点,同时采样后信息的存储、传递、边权值的更新策略以及关键节点的性质都值得研究人员关注。

4)理论研究方面,原始网络与压缩网络之间在多个尺度上具有层次相似性,图稀疏算法在图压缩方面的理论分析需要进一步深入。

结束语 随着网络规模的不断扩大,个体与个体之间的相互关系越来越难挖掘。对网络进行有效的压缩,将有助于挖掘隐藏在网络内部的特征和价值。在这个背景下,研究人员开始对复杂网络的采样和稀疏表示展开大量的工作^[56-57],针对复杂网络运用不同的图稀疏算法来加速对网络拓扑结构的分析。本文对基于边采样的图稀疏算法进行了归纳和总结,不同的图分析算法可以根据不同的边度量方式进行稀疏化采样;同时展望了未来复杂网络图稀疏可以深入研究的几个方向。

参考文献

- [1] ZHENG X, CHEN J P, SHAO J L, et al. Analysis on topological properties of Beijing urban public transit based on complex network theory [J]. Acta Physica Sinica, 2012, 61(19): 95-105. (in Chinese)
郑啸, 陈建平, 邵佳丽, 等. 基于复杂网络理论的北京公交网络拓扑性质分析[J]. 物理学报, 2012, 61(19): 95-105.
- [2] QIAO S J, HAN N, ZHANG K F, et al. Algorithm for Detecting Overlapping Communities from Complex Network Big Data [J].

- Journal of Software, 2017, 28(3): 631-647. (in Chinese)
- 乔少杰, 韩楠, 张凯峰, 等. 复杂网络大数据中重叠社区检测算法[J]. 软件学报, 2017, 28(3): 631-647.
- [3] LIU X, YI D Y. Complex Network Community Detection by Local Similarity [J]. Acta Automatica Sinica, 2011, 37(12): 1520-1529. (in Chinese)
- 刘旭, 易东云. 基于局部相似性的复杂网络社区发现方法[J]. 自动化学报, 2011, 37(12): 1520-1529.
- [4] ZHOU X, ZHANG F M, LI K W, et al. Finding vital node by node importance evaluation matrix in complex networks[J]. Acta Physica Sinica, 2012, 61(5): 50201. (in Chinese)
- 周漩, 张凤鸣, 李克武, 等. 利用重要度评价矩阵确定复杂网络关键节点[J]. 物理学报, 2012, 61(5): 50201.
- [5] ZHANG K, LI P P, ZHU B P, et al. Based on PageRank Weighted importance evaluation method for weighted weighted complex network nodes[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2013, 45(3): 429-434. (in Chinese)
- 张琨, 李配配, 朱保平, 等. 基于 PageRank 的有向加权复杂网络节点重要性评估方法[J]. 南京航空航天大学学报, 2013, 45(3): 429-434.
- [6] YANG B, LIU D Y, LIU J M, et al. Complex Network Clustering Method[J]. Journal of Software, 2009, 20(1): 54-66. (in Chinese)
- 杨博, 刘大有, LIU J M, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54-66.
- [7] CAI Q. Structure Analytics of Big Data Complex Networks Based on Swarm Intelligence[D]. Xi'an: Xidian University, 2015. (in Chinese)
- 蔡清. 基于群体智能优化的大数据复杂网络结构分析[D]. 西安: 西安电子科技大学, 2015.
- [8] LI M P, GAO H, ZOU Z N, et al. K-Reach Query Processing Based on Graph Compression[J]. Journal of Software, 2014, 25(4): 797-812. (in Chinese)
- 李鸣鹏, 高宏, 邹兆年, 等. 基于图压缩的 k 可达查询处理[J]. 软件学报, 2014, 25(4): 797-812.
- [9] WANG X H, YANG X Y, JIAO L C. Compression of Complex Networks Based on Multiscale Geometric Analysis[J]. Journal of Electronics & Information Technology, 2009, 31(4): 968-972. (in Chinese)
- 王晓华, 杨新艳, 焦李成. 基于多尺度几何分析的复杂网络压缩策略[J]. 电子与信息学报, 2009, 31(4): 968-972.
- [10] ZHANG Y, LIU Y B, XIONG G, et al. Journal of Electronics & Information Technology[J]. Journal of Software, 2014, 25(9): 1937-1952. (in Chinese)
- 张宇, 刘燕兵, 熊刚, 等. 图数据表示与压缩技术综述[J]. 软件学报, 2014, 25(9): 1937-1952.
- [11] LI H B, ZHANG J P, YANG J, et al. Acta Scientiarum Naturalium Universitatis[J]. Acta Scientiarum Naturalium Universitatis Pekinesis, 2013, 49(1): 117-125. (in Chinese)
- 李泓波, 张健沛, 杨静, 等. 基于社区节点重要性的社会网络压缩方法[J]. 北京大学学报(自然科学版), 2013, 49(1): 117-125.
- [12] YANG H L, ZHANG J P, YANG J. Compression online social networks by calibrating structure redundancy [J]. Journal of Computer Research and Development, 2013, 50(12): 2504-2519. (in Chinese)
- 杨海陆, 张健沛, 杨静. 基于结构冗余性校准的在线式社会网络压缩[J]. 计算机研究与发展, 2013, 50(12): 2504-2519.
- [13] QIAN J B, ZHU Q, WANG Y L. Bloom Filter Based Associative Deletion[J]. IEEE Transactions on Parallel and Distributed Systems (TPDS), 2014, 25(8): 1986-1998.
- [14] CHEN Y F, HUANG Z P, CAO P, et al. Subscription Aggregation Query Processing Based on Matrix Summation over DTN [J]. IEICE Transactions on Communications, 2016, 99(4): 812-819.
- [15] BATSON J, SPIELMAN D A, SRIVASTAVA N, et al. Spectral sparsification of graphs: Theory and algorithms[J]. Communications of the ACM, 2013, 56(8): 87-94.
- [16] COHEN E. Fast Algorithms for Constructing t-Spanners and Paths with Stretch t[C] // Symposium on Foundations of Computer Science. IEEE, 1999: 648-658.
- [17] FUNG W S, HARIHARAN R, HARVEY N J A, et al. A general framework for graph sparsification[C] // ACM Symposium on Theory of Computing. 2011: 71-80.
- [18] AHN K J, GUHA S, MCGREGOR A. Graphsketches: sparsification, spanners, and subgraphs[C] // 31st ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. 2012: 5-14.
- [19] ABELLO J, KORN J, FINOCCHI I. Graph Sketches[C] // Proceedings of the ACM Symposium. 2001: 67-70.
- [20] SPIELMAN D A, TENG S H. Smoothed analysis of termination of linear programming algorithms[J]. Mathematical Programming, 2003, 97(1/2): 375-404.
- [21] BASWANA S, SEN S. A simple and lineartime randomized algorithm for computing sparse spanners in weighted graphs[J]. Random Structures & Algorithms, 2010, 30(4): 532-563.
- [22] AHN K J, GUHA S, MCGREGOR A. Spectral Sparsification in Dynamic Graph Streams[J]. Mathematical Programming, 2013, 7(11): 9896-9904.
- [23] MARTIN C. Spectral Sparsification: The Barrier Method and its Applications [D]. Cambridge, Massachusetts, 2014.
- [24] LEFEVRE K, TERZI E. GraSS: Graph Structure Summarization [C] // SIAM International Conference on Data Mining, SDM 2010. 2010: 454-465.
- [25] SPIELMAN D A, TENG S H. Spectral Sparsification of Graphs [J]. SIAM Journal of Computing, 2008, 40(4): 981-1025.
- [26] KOUTIS I, LEVIN A, PENG R. Faster Spectral Sparsification and Numerical Algorithms for SDD Matrices[J]. ACM Transactions on Algorithms, 2012, 12(2): 17.
- [27] HUANG H. Spectral Sparsification for Directed Graphs via Random Sampling [D]. Shanghai: Fudan University, 2014. (in Chinese)
- 黄璜. 有向图的随机采样谱稀疏化方法[D]. 上海: 复旦大学, 2014.
- [28] NEYLON T. Matrix Sparsification and the Sparse Null Space Problem[M]. New York: Springer-Verlag, 2016.
- [29] SPIELMAN D A, TENG S H. Nearly-lineartime algorithms for graph partitioning, graph sparsification, and solving linear systems [C] // Proceedings of IEEE Foundations of Computer Science. New York: IEEE Press, 2004: 81-90.
- [30] KOUTIS I, XU S C. Simple Parallel and Distributed Algorithms for Spectral Graph Sparsification[M] // Algorithms and Models for the Web-Graph. Springer Berlin Heidelberg, 2016: 13-19.