

# 基于图割和局部算子的图子集选取

陈丹丹,王 健

(复旦大学 大数据学院,上海 200433)

**摘要:**图子集选取问题旨在从图节点集中采样少部分代表性节点,利用观测的节点信号值去重构原始图信号。在资源有限的情况下,可以降低数据维度和计算复杂度,提高对复杂多变图结构的适应性,从而为网络数据的传输处理提供高效的技术支撑。现有的确定性算法大多采用贪心优化,后序采样点的选择依赖于前序已采样节点,对初始值敏感,且可能陷入局部最优;同时,大多数频域算法没有考虑顶点域内采样集节点的空间关系。该文提出基于局部算子的两步采样算法,通过构建节点局部算子的内积完全图来度量采样节点的距离,首先求解标准图割,将节点集按距离划分指定个数簇;其次,在各个簇内依据稀疏性度量选择最优点,从而生成最终的采样集。该算法同时结合了频域与节点域的信息,并使得采样可并行执行。在多种图场景下与多种代表性算法相比,该算法都可以取得最优或相近的重构效果。

**关键词:**图信号处理;图信号采样;图子集选取;局部算子;图割

**中图分类号:**TP301.6;TN911.7;O157.6 **文献标识码:**A **文章编号:**1673-629X(2023)06-0001-07

**doi:**10.3969/j.issn.1673-629X.2023.06.001

## Graph Subset Selection via Graph Cut and Localization Operators

CHEN Dan-ran, WANG Jian

(School of Data Science, Fudan University, Shanghai 200433, China)

**Abstract:** Graph subset selection aims to select a small set of representative nodes to recover the original graph signal. In the case of limited resources, the data dimension and computational complexity can be reduced, and the adaptability to complex and changeable graph structures can be improved, so as to provide efficient technical support for the transmission and processing of network data. Existing deterministic methods mostly use a greedy serial selection framework, in which node selection depends on previous sampled nodes. It can be sensitive to initialization and may be trapped into local optimum. Meanwhile, most spectrum methods do not consider the spatial relationship of sampled nodes in the vertex domain. We propose a two-step algorithm based on localization operators, which measures distances of sampled nodes by constructing a complete inner graph of localization operators. The first step is to find the normalized graph cut so that the vertex set can be divided into a specified number of clusters based on node distances. Then, an optimal vertex is selected according to a sparsity measure in each cluster, which makes up the final sampling set. This method takes into account both spectral and vertex domain information and realizes the parallel execution in sampling. The proposed algorithm achieves the best or competitive reconstruction performance compared to existing methods in various scenarios.

**Key words:** graph signal processing; graph signal sampling; graph subset selection; localization operator; graph cut

## 0 引言

在统计学、信号处理、计算机视觉等学科中传统关注的信号多为结构化的,如音频、雷达信号、生物医学信号、图像和视频等。随着信息技术的高速发展,当今社会万物互联,数据来源分散且多存在于不规则拓扑结构,如社交网络<sup>[1]</sup>、传感器网络<sup>[2]</sup>、生物网络<sup>[3]</sup>等。这些网络结构数据,大多存在与节点相关的特征,如社交网络中用户的年龄等属性,传感器网络中的各个传

感器的观测值,这些节点数据可以视作一种图上的信号。对于分布式的网络信号,传统的信号处理理论无法很好地描述其节点之间的相互关系,因此,图信号处理(Graph Signal Processing, GSP)应运而生,通过引入图上的傅里叶变换(Graph Fourier Transform, GFT),将时间信号推广至基于图网络表征非规则域中的信号。

图网络的拓扑结构复杂多变,同时,较高的数据维度会使得计算复杂度增加,故在资源有限的情况下,能

收稿日期:2023-02-28

修回日期:2023-04-28

基金项目:国家自然科学基金(61971146)

作者简介:陈丹丹(1998-),女,硕士,研究方向为图神经网络与图信号采样;通讯作者:王 健(1983-),男,副教授,研究方向为图神经网络、视觉与自动驾驶、统计与深度学习、稀疏与低秩优化等。

够直接观测到的图信号个数变得十分有限。如何利用少部分最具有信息量的节点的观测值与图结构信息去准确快速地估计未观测的节点信号,进而为图结构数据的传输处理提供高效的技术支撑,是图信号处理中的核心问题,即图信号采样与重构。该问题依赖于一个隐含假设——信号的平滑性,即图上邻近的节点具有相似的值。通常,图信号处理中的平滑性假设采用图傅里叶基的(近似)带宽限制(Band-limitedness)形式,从而使得原始信号能够被部分采样信号重构。

目前,对图信号采样与重构的研究大致可分为局部观测<sup>[4-5]</sup>与图子集选取(Graph Subset Selection, GSS)<sup>[1,6-14]</sup>两类。前者利用采样节点邻域的聚合观测值去重构信号,更适用于存在聚集性质的图结构。图子集选取则从图节点集中采样最具有信息量的节点子集,将这些节点的(带噪)信号值作为观测值去重构原始信号,可应用于多个领域,如环境监测<sup>[15]</sup>、半监督节点分类<sup>[16]</sup>、矩阵补全<sup>[17]</sup>等。

由于图结构具有离散性质,图子集选取本质是一个组合优化问题,已被证明为 NP 难<sup>[18]</sup>,可以近似求解。现有方法可以归为两类:(1)随机算法;(2)确定性算法<sup>[19]</sup>。前者通过在图节点集上定义一个概率分布函数,依据分布随机地采样节点<sup>[6-8]</sup>。后者则通过优化预先定义的损失函数,寻找最优采样集<sup>[9-14]</sup>。研究表明,在采样点数相同的情况下,随机算法的重构效果很难超过确定性算法<sup>[7]</sup>。虽然确定性算法效果更优,现有方法大多从谱域设计优化准则,并未考虑节点域内采样集节点的距离关系,而在平滑假设下,采样距离相近的点对重构效果带来的增益可能十分有限。同时,这类算法通常使用贪心优化,即每一步迭代时,选择最大化或最小化目标函数的局部最优点。然而,贪心方法可能会陷入局部最优解,并不能保证全局最优,并且由于采样点的选择依赖前序已采样节点,采样只能串行执行,在处理大规模图时效率较低。

针对以上不足,该文提出以图割(Graph Cut)的视角建模图子集选取问题,既考虑了采样集节点在节点域和谱域的信息,又可实现一次性采样,并行运行。具体地,首先利用局部算子构建一个新的内积完全图,再利用谱聚类算法生成标准图割,得到节点集的划分,使得同一子集内的节点距离尽可能近,不同子集间的节点距离足够远;其次,在每个子集内依据稀疏性度量采样最优节点,即可组成最终的采样集。在多种图上的实验结果表明,提出的算法(Cut Sampling)优于现有文献中的几种代表性算法。

该文的贡献总结如下:

(1)提出利用图割为图子集选取划分采样可行集,利用局部算子构建内积完全图,可控制采样集节点

的间隔距离,同时结合了谱域的局部性;

(2)提出的算法不需要贪心优化,可以在采样步骤并行运行,具有可扩展性;

(3)相比于几种现有文献中的代表性算法,提出的方法实验效果更优或相近。

## 1 相关研究

现有的图子集选取方法可分为两类:随机算法和确定性算法。随机算法根据定义的概率分布在节点集随机采样从而生成采样集。研究表明,独立于图结构的均匀分布,当采样足够多个节点时,对于 Erdős-Rényi 图(ER 图)可高概率实现完美重构<sup>[1]</sup>,而非均匀的随机采样则根据基于图结构的节点重要性设计概率分布<sup>[7-8]</sup>。虽然与图结构无关的随机采样计算高效,利用压缩感知中约束等距性质(Restricted Isometry Property, RIP)的理论研究表明,在采样点数相同的情况下,与图结构相关的随机采样方法效果更优<sup>[7]</sup>,文献[6]展示了具体对比。同时,随机方法很难保证采样集中的节点距离足够远,而在图信号的平滑假设下,距离相近的采样点带来的信息增益可能有限<sup>[12]</sup>。故该文沿确定性算法展开研究。

确定性算法大多为谱方法,通过松弛优化<sup>[20]</sup>或贪心优化去最大化或最小化某种定义的损失函数。例如基于实验设计的优化准则,E-optimal<sup>[1]</sup>,A-optimal 和 D-optimal<sup>[9]</sup>等。虽然这类贪心算法通常可以获得较好的重构效果,但依赖特征分解,计算复杂度较高。为避免特征分解,可以对谱域优化准则做近似处理,如 MaxCutoff<sup>[10]</sup>利用图谱代理(Graph Spectral Proxies),去近似最大化截止频率;Fen Wang 等<sup>[21]</sup>利用诺伊曼级数(Neumann Series)优化 A-optimal 准则等。另有一些方法虽无法避免准备工作阶段的特征分解,但将贪心算法中加入一个节点后比较全局指标,转换为利用每个节点的增量误差衡量节点质量,可以避免节点选择步骤的特征分解,如利用 QR 分解<sup>[13]</sup>,借鉴压缩感知中正交匹配追踪算法的思路,通过投影算子定义残差<sup>[14]</sup>等。但上述方法大多只实现了频域内采样的局部性,并没有考虑图拓扑结构以及采样节点的空间关系。还有些确定性算法只考虑节点域信息,如传感器设置方法<sup>[22]</sup>,可有效避免对图傅里叶基的依赖。

近年来,结合谱域与节点域信息的方法被相继提出<sup>[11-12]</sup>,Jayawant 等<sup>[11]</sup>提出一个两步算法,首先寻找和已采样节点距离足够远的节点可行集,其次在可行集内根据准则贪心寻找最优点。Sakiyama 等<sup>[12]</sup>想法类似,利用定义在节点域的局部算子<sup>[23]</sup>合并上述两步,提出 Ed Free 算法。同时,Sakiyama 等还证明了许多确定性算法可用局部算子统一表示。但是这些算法

同样使用贪心优化去一个个选择最优节点,串行采样,效率较低。鉴于结合节点域与谱域的图子集选取方法展示出优秀的节点选择能力,同时考虑到局部算子的泛化性及其可以结合两域信息的特性,该文沿用局部算子并从节点距离出发,提出用新的视角图割<sup>[24]</sup>去建模图子集选取问题,可在一定程度上克服现有谱域贪心算法的不足。

## 2 预备知识

### 2.1 图信号处理

该文考虑无向联通图  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ , 其中  $\mathcal{V}$  表示大小为  $N$  的节点集,  $\mathcal{E}$  表示边集。  $\mathbf{W} \in \mathbb{R}^{N \times N}$  表示图  $\mathcal{G}$  的带权邻接矩阵, 其中  $\mathbf{W}_{ij} = e_{ij}$ ,  $e_{ij}$  表示边权。若节点  $i$  和  $j$  之间无边连接, 即  $(i, j) \notin \mathcal{E}$ , 则  $e_{ij} = 0$ ; 若有边相连,  $(i, j) \in \mathcal{E}$ , 则  $e_{ij} > 0$ 。对于无向图,  $\mathbf{W}$  是对称阵。图拉普拉斯矩阵定义为  $\mathbf{L}_\mathcal{G} = \mathbf{D} - \mathbf{W}$ , 其中度矩阵  $\mathbf{D}$  为对角阵, 对角线元素为节点对应的度  $D_{ii} = \sum_{i \neq j} \mathbf{W}_{ij}$ , 即与节点  $i$  相连的边数。归一化的图拉普拉斯矩阵可表示成  $\hat{\mathbf{L}}_\mathcal{G} = \mathbf{D}^{-1/2} \mathbf{L}_\mathcal{G} \mathbf{D}^{-1/2}$ 。

由于拉普拉斯矩阵是半正定的实对称阵, 因此具有完备的特征基, 其特征分解为  $\mathbf{L}_\mathcal{G} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H$ , 其中  $\mathbf{U} = [\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{N-1}]$ ,  $\mathbf{u}_i$  为  $\mathbf{L}_\mathcal{G}$  的特征向量,  $(\cdot)^H$  表示共轭转置,  $\mathbf{\Lambda} = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_{N-1})$  为对角阵,  $\lambda_i$  为对应的特征值。不失一般性, 可以假设  $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{N-1}$ 。由于  $\{\mathbf{u}_i\}_{i=0}^{N-1}$  是一组正交特征向量, 通常可用来定义图傅里叶变换。

图信号是定义在图节点集的映射, 可以看作一个向量  $\mathbf{x} \in \mathbb{R}^N$ , 其中  $x_i$  表示节点  $i$  处的信号值。图傅里叶变换将  $\mathbf{L}_\mathcal{G}$  的特征向量作为图信号的傅里叶基, 将特征值作为频点, 展开信号记频率系数为  $\hat{\mathbf{x}}$ , 则图傅里叶变换 (GFT) 及其逆变换 (Inverse-GFT) 的数学形式如下:

$$\begin{aligned} \text{GFT: } \hat{\mathbf{x}} &= \mathbf{U} \mathbf{x} \\ \text{Inverse-GFT: } \mathbf{x} &= \mathbf{U}^H \hat{\mathbf{x}} \end{aligned} \quad (1)$$

GFT 之所以可利用  $\mathbf{L}_\mathcal{G}$  的特征分解定义, 可以从  $\mathbf{L}_\mathcal{G}$  的二次型理解, 它可表示为图上有边相连的节点信号值变化的加权平方和, 反映了图信号的平滑程度。

$$\mathbf{x}^T \mathbf{L}_\mathcal{G} \mathbf{x} = \frac{1}{2} \sum_{(i,j) \in \mathcal{E}} \mathbf{W}_{ij} (x_i - x_j)^2 \quad (2)$$

将  $\mathbf{u}_i$  视为图上的信号, 则当  $\lambda_i$  越小,  $\mathbf{u}_i$  会越平滑。传统离散信号处理中, 随时间变化越缓慢的信号频率越低, 类似地, 沿图中边变化越平滑的图信号频率越低, 因此 GFT 可将特征值看作信号的频率, 低频代表图信号沿边的变化平滑, 高频代表变化剧烈。

在图信号处理中, 通常假设图信号满足基于图结

构的某些条件, 例如平滑、带限等。该文研究带限图信号, 定义如下:

定义 1 ( $K$ -带限 ( $K$ -Bandlimited)): 图信号  $\mathbf{x} \in \mathbb{R}^N$  称为是  $K$ -带限的, 当且仅当存在  $K \in \{0, 1, \dots, N-1\}$ , 使得图傅里叶系数  $\hat{\mathbf{x}}$  满足:

$$\hat{x}_i = 0, \quad \forall i \geq K \quad (3)$$

局部算子 (Localization Operators) 是传统信号处理中的平移算子在图信号处理领域的拓展, 可以同时结合节点域和谱域信息, 定义如下:

定义 2 (局部算子<sup>[23]</sup>):  $\forall n \in \{0, 1, \dots, N-1\}$ , 节点  $i \in \mathcal{V}$  处局部算子的第  $n$  个分量元素定义如下:

$$\begin{aligned} T_{g,i}(n) &:= \sqrt{N} (g * \delta_i)(n) = \\ &\sqrt{N} \sum_{l=0}^{N-1} g(\lambda_l) u_l^H(i) u_l(n) \end{aligned} \quad (4)$$

其中,  $(*)$  表示图卷积,  $g$  为谱域核,  $\delta_i$  为节点  $i$  处的脉冲信号。局部算子的矩阵形式如下:

$$\begin{aligned} \mathbf{T}_g &= [\mathbf{T}_{g,0}, \mathbf{T}_{g,1}, \dots, \mathbf{T}_{g,N-1}] = \\ &\sqrt{N} \mathbf{U} g(\mathbf{\Lambda}) \mathbf{U}^H \end{aligned} \quad (5)$$

常用谱域核有热核  $g(\lambda_i) = e^{-\pi \lambda_i}$ , 此时  $\mathbf{T}_{g,i}$  具有围绕节点  $i$  移动窗口的作用<sup>[23]</sup>。另一常用的为理想核, 当  $i < K$ ,  $g(\lambda_i) = 1$ , 否则为 0, 此时重构信号可以写成  $\mathbf{T}_g$  的线性组合<sup>[12]</sup>。

### 2.2 问题形式

图子集选取问题可分为两步: 采样和重构。假设想要找到一个大小为  $M$  的采样集  $\mathcal{M} \subseteq \mathcal{V}$ , 其中  $\mathcal{M} = (\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_{M-1})$  表示采样序列, 并且  $\mathcal{M}_i \in \{0, 1, \dots, N-1\}$  是非重复的。只有采样节点的信号值是可以观测到的。采样矩阵  $\Psi: \mathbb{R}^N \mapsto \mathbb{R}^M$  定义如下:

$$\Psi_{i,j} = \begin{cases} 1, & j = \mathcal{M}_i \\ 0, & \text{其他情况} \end{cases} \quad (6)$$

令  $\mathbf{w} \in \mathbb{R}^M$  为观测中引入的噪声, 假设服从  $i.i.d.$  高斯分布, 则采样信号为  $\mathbf{x}_\mathcal{M} = \Psi \mathbf{x}$ , 观测模型为  $\mathbf{y}_\mathcal{M} = \Psi \mathbf{x} + \mathbf{w}$ , 对于带限图信号, 观测模型可写作:

$$\mathbf{y}_\mathcal{M} = \Psi \mathbf{U}_{\mathcal{V}\mathcal{K}} \hat{\mathbf{x}}_\mathcal{K} + \mathbf{w} = \mathbf{U}_{\mathcal{M}\mathcal{K}} \hat{\mathbf{x}}_\mathcal{K} + \mathbf{w} \quad (7)$$

其中,  $\mathbf{U}_{\mathcal{M}\mathcal{K}} = \Psi \mathbf{U}_{\mathcal{V}\mathcal{K}}$ ,  $\mathbf{U}_{\mathcal{V}\mathcal{K}}$  表示  $\mathbf{U}$  按  $\mathcal{K} = \{0, 1, \dots, K-1\}$  列索引的  $K$  列,  $\hat{\mathbf{x}}_\mathcal{K}$  表示  $\hat{\mathbf{x}}$  对应的  $K$  个系数。

当采样集  $\mathcal{M}$  固定后, 真实信号值可由最优线性估计 (Best Linear Unbiased Estimation)<sup>[25]</sup> 重构得到:

$$\hat{\mathbf{x}}_\mathcal{K} = \mathbf{U}_{\mathcal{K}\mathcal{V}} \hat{\mathbf{x}}_\mathcal{K} = \mathbf{U}_{\mathcal{V}\mathcal{K}}^\dagger \mathbf{U}_{\mathcal{M}\mathcal{K}}^\dagger \mathbf{y}_\mathcal{M} \quad (8)$$

其中,  $\mathbf{U}_{\mathcal{M}\mathcal{K}}^\dagger = (\mathbf{U}_{\mathcal{M}\mathcal{K}}^H \mathbf{U}_{\mathcal{M}\mathcal{K}})^{-1} \mathbf{U}_{\mathcal{M}\mathcal{K}}^H$  是  $\mathbf{U}_{\mathcal{M}\mathcal{K}}$  的伪逆。图子集选取问题的核心就是找到一个最优采样集  $\mathcal{M}$ , 使得重构误差  $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$  最小化。

### 2.3 图割

图割是对图节点集的划分, 标准图割 (Normalized Graph Cut, N-CUT) 是一类经典的图割, 通常具有优秀



的聚类效果,其定义如下:

定义 3 (标准图割(N-CUT)<sup>[24]</sup>): 图  $\mathcal{G} = (\mathcal{V}, \varepsilon, \mathbf{W})$  大小为  $M$  的标准图割为节点集  $\mathcal{V}$  的一个分割  $\{A_1, A_2, \dots, A_M\}$ , 使得下式目标函数最小化。

$$\text{N-CUT}(A_1, A_2, \dots, A_M) = \frac{1}{2} \sum_{i=1}^M \frac{\sum_{u \in A_i, v \in \bar{A}_i} w_{uv}}{\sum_{u \in A_i, v \in \mathcal{V}} w_{uv}} \quad (9)$$

其中,  $\bar{A}_i$  为  $A_i$  的补集。

N-CUT 通常作用于相似性图,用于将节点集划分为多个簇,使得簇间节点具有低相似度,簇内节点具有高相似度。这是一个 NP 难问题,可以用谱聚类算法放缩求解<sup>[26]</sup>。

### 3 Cut Sampling 算法介绍

#### 3.1 信号重构

由公式(8),利用带理想核的局部算子,重构信号可表示为局部算子的加权线性组合形式:

$$\begin{aligned} \mathbf{x}' &= \mathbf{U}_{\mathcal{V}\mathcal{K}} \mathbf{U}_{\mathcal{M}\mathcal{K}}^H (\mathbf{U}_{\mathcal{M}\mathcal{K}} \mathbf{U}_{\mathcal{M}\mathcal{K}}^H)^{\dagger} \mathbf{y}_{\mathcal{M}} = \\ &\mathbf{U} \text{diag}(\mathbf{1}_{\mathcal{K}}) \mathbf{U}_{\mathcal{M}\mathcal{V}}^H (\mathbf{U}_{\mathcal{M}\mathcal{V}} \text{diag}(\mathbf{1}_{\mathcal{K}}) \mathbf{U}_{\mathcal{M}\mathcal{V}}^H)^{\dagger} \mathbf{y}_{\mathcal{M}} = \\ &\mathbf{T}_{\mathcal{V}\mathcal{M}} (\mathbf{T}_{\mathcal{M}\mathcal{M}})^{\dagger} \mathbf{y}_{\mathcal{M}} = \end{aligned}$$

$$\mathbf{T}_{\mathcal{V}\mathcal{M}} \mathbf{c} = \sum_{i \in \mathcal{M}} c_i \mathbf{T}_{g,i} \quad (10)$$

其中,  $\mathbf{c} = (\mathbf{T}_{\mathcal{M}\mathcal{M}})^{\dagger} \mathbf{y}_{\mathcal{M}}$ 。

由于信号重构本质是利用采样值去插值估计未观测值,由公式(10),未观测节点的值亦可以由采样节点处  $\mathbf{T}_{g,i}$  的带权线性组合插值得到。故  $\mathbf{T}_{g,i}$  的非零分量,即覆盖区域,可被视作节点  $i$  对重构信号有贡献的节点区域。为使得重构误差尽可能小,最优采样集应具有尽可能大的覆盖区域,即单个采样节点的覆盖区域尽可能大,同时,采样节点间的覆盖区域交集尽可能小。

具体地,  $\|\mathbf{T}_{g,i}\|_1$  可表示节点  $i$  的信号覆盖域,内积  $\langle \|\mathbf{T}_{g,i}\|_1, \|\mathbf{T}_{g,j}\|_1 \rangle$  可代表节点  $i$  和  $j$  信号覆盖区域的交集大小。例如,若  $\langle \|\mathbf{T}_{g,i}\|_1, \|\mathbf{T}_{g,j}\|_1 \rangle = 0$ ,则表示节点  $i$  和  $j$  的信号覆盖域无交集。

基于上述性质,该文利用局部算子的内积构建一张完全图,边权为信号覆盖域交集大小的度量,并在此基础上提出一个两步的图子集选取算法 Cut Sampling: (1) N-CUT 聚类; (2) 在每个子集内根据稀疏性准则选择最优点。

完整算法流程如图 1 所示。

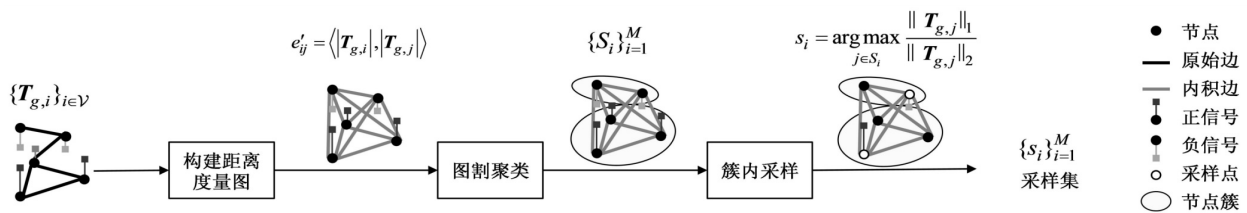


图 1 Cut Sampling 算法流程

#### 3.2 图割聚类

对于原始图  $\mathcal{G} = (\mathcal{V}, \varepsilon, \mathbf{W})$ , 算法首先通过构建一个对应的完全图  $\mathcal{Q} = (\mathcal{V}, \varepsilon')$ , 以边权去度量图上任意两个节点的空间距离关系。具体地,图  $\mathcal{Q}$  中的边权为相连两个节点的局部算子的内积。

$$e'_{ij} = \langle \|\mathbf{T}_{g,i}\|_1, \|\mathbf{T}_{g,j}\|_1 \rangle, \quad \forall i \neq j \in \mathcal{V} \quad (11)$$

由前述局部算子的性质可知,图  $\mathcal{Q}$  即为原始图中节点距离的度量图,可通过求解 N-CUT 将节点集按节点的距离关系划分为指定个数簇。由公式(9),图  $\mathcal{Q}$  的 N-CUT 的最小化目标函数可写作:

$$\begin{aligned} \text{N-CUT}(S_1, S_2, \dots, S_M) &= \frac{1}{2} \sum_{i=1}^M \frac{\sum_{u \in S_i, v \notin S_i} e'_{uv}}{\sum_{u \in S_i, v \in \mathcal{V}} e'_{uv}} = \\ &\frac{1}{2} \sum_{i=1}^M \frac{\sum_{u \in S_i, v \notin S_i} e'_{uv}}{\sum_{u, v \in S_i} e'_{uv} + \sum_{u \in S_i, v \notin S_i} e'_{uv}} = \\ &\frac{1}{2} \sum_{i=1}^M \frac{1}{1 + \sum_{u, v \in S_i} e'_{uv} / \sum_{u \in S_i, v \notin S_i} e'_{uv}} \quad (12) \end{aligned}$$

注意到  $\sum_{u \in S_i, v \notin S_i} e'_{uv}$  和  $\sum_{u, v \in S_i} e'_{uv}$  分别表示不同子集间和同一子集内节点信号覆盖域的交集大小。故通过求解公式(12),可以得到  $M$  个节点子集,控制子集内节点距离足够近,子集间节点距离足够远。利用谱聚类算法,可近似求解 N-CUT 问题,算法流程如算法 1 所示。

算法 1: N-CUT 谱聚类算法<sup>[26]</sup>。

输入: 距离图  $\mathcal{Q} = (\mathcal{V}, \varepsilon')$ , 采样个数  $M$ , 特征向量矩阵  $\mathbf{U}$ ;  
输出: 节点子集  $\{S_1, S_2, \dots, S_M\}$ 。

1. 取  $\mathbf{U}$  的前  $M$  列的第  $i$  行,  $\mathbf{z}_i = \mathbf{U}_{i:M} \in \mathbb{R}^M, \forall i \in \{0, 1, \dots, N-1\}$ ;
2. 利用 k-means 算法对  $\{\mathbf{z}_i\}_{i=0}^{N-1}$  聚类, 得到节点簇  $\{C_i\}_{i=1}^M$ ;
3. 返回节点子集  $\{S_i\}_{i=1}^M$ , 其中  $S_i = \{j \mid \mathbf{z}_j \in C_i\}$ 。

#### 3.3 簇内采样

图割步骤为后续采样划分了可行集范围。由于图割已控制不同子集间的点距离足够远,为了选出最具信息量的采样集,算法 Cut Sampling 第二步在各子集  $S_i$  内分别选取最有代表性的点,共同组成最终的采样集。具体地,定义如下采样准则,其中  $\mathbf{T}_{g,j}$  表示节点  $j$

处的局部算子:

$$s_i = \arg \max_{j \in S_i} \frac{\| \mathbf{T}_{g,j} \|_1}{\| \mathbf{T}_{g,j} \|_2}, i = 1, 2, \dots, M \quad (13)$$

Ed Free<sup>[12]</sup>中利用了局部算子的  $\ell_1$  范数,而最大化  $\ell_1$  范数等价于寻找覆盖域最大的信号节点。该文在此基础上增加了分母对  $\ell_2$  范数的限制,这也是一种稀疏性度量<sup>[8]</sup>。从信号重构的角度不难理解,由于信号重构本质是对观测值的插值,过于稀疏的局部算子代表当前节点可能与其他节点有较小的关联,不具有代表性,可能带来较大的重构误差。算法希望采样信号在具有较大覆盖域的同时,不会过于稀疏,即避免非零分量过于集中导致难以提供足够信息。例如,假设  $\mathbf{T}_{g,1} = [1, 0, 0, \dots, 0]$ ,  $\mathbf{T}_{g,2} = [1/2, 1/2, 0, \dots, 0]$ , 则  $\mathbf{T}_{g,1}$  和  $\mathbf{T}_{g,2}$  的  $\ell_1$  范数均为1,但  $\mathbf{T}_{g,2}$  的  $\ell_2$  范数更小。由公式(13),该文的采样准则将选择  $\mathbf{T}_{g,2}$  而非更稀疏的  $\mathbf{T}_{g,1}$ 。完整算法总结如算法2所示。

算法2:Cut Sampling 算法。

输入:图  $G = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ , 采样个数  $M$ , 特征向量矩阵  $\mathbf{U}_{\mathcal{V} \times K}$ , 局部算子  $\mathbf{T}_g$ ;

输出:采样集  $\mathcal{M} \subseteq \mathcal{V}$ ,  $|\mathcal{M}| = M$ 。

1. 构建完全图  $Q = (\mathcal{V}, \mathcal{E}')$ ,  $\forall i \neq j \in \mathcal{V}, e_{ij}' \in \mathcal{E}'$ ;

$$e_{ij}' = \langle |\mathbf{T}_{g,i}|, |\mathbf{T}_{g,j}| \rangle$$

2. 由算法1对图  $Q$  谱聚类求解  $N$ -CUT, 得到节点集  $\{S_i\}_{i=1}^M$ ;

3. 在各节点子集  $S_i$  中,选择最优节点,  $\forall i = 1, 2, \dots, M$ ;

$$s_i = \arg \max_{j \in S_i} \frac{\| \mathbf{T}_{g,j} \|_1}{\| \mathbf{T}_{g,j} \|_2}$$

4. 返回采样集  $\mathcal{M} = \{s_i\}_{i=1}^M$ 。

## 4 数值实验

### 4.1 实验设置

实验利用 GSPBOX<sup>[27]</sup>,在如下四类图上对比不同算法的重构效果:

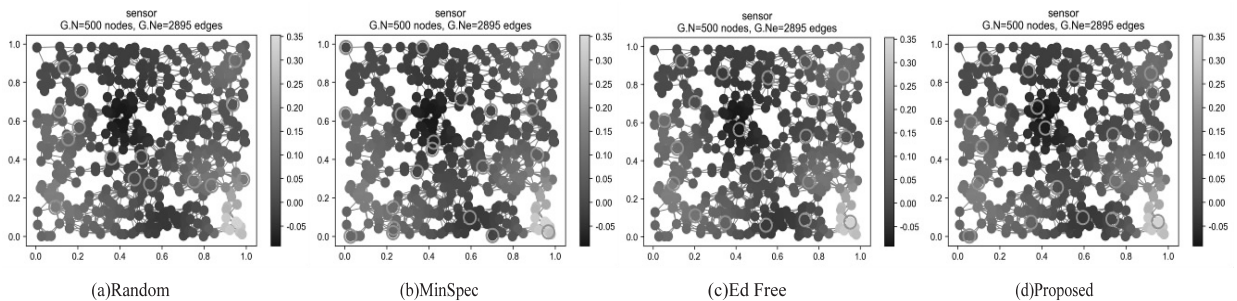


图2 随机 Sensor 图的采样节点集

### 4.3 重构误差比较

图3展示了四种图场景下,四种算法的重构误差随采样集大小的变化。实验通过计算重构信号与原始信号的均方误差(Mean Squared Error, MSE)来评价重构效果。为了获得相对稳定的实验结果,对于每张图,

(1)随机 Sensor 图:为带权稀疏图;

(2)Minnesota 交通图:为等权重图,边权均为1;

(3)基于 Barabási-Albert 模型的随机图(BA图);

(4)基于 Erdős-Rényi 模型的随机图(ER图):边的连接概率  $p$  分别设置为 0.05、0.15 和 0.5。

实验设置 Minnesota 图的总节点数为  $N = 2642$ ,其他图均为  $N = 500$ 。所有实验中,均假设图信号是  $K$ -带限的,  $K = 10$ ,假设图傅里叶系数的非零分量服从高斯分布  $\hat{\mathbf{x}} \sim \mathcal{N}(0, 0.5)$ ,观测噪声的分量  $\mathbf{w} \sim \mathcal{N}(0, 5 \times 10^{-3})$ 。

将 Cut Sampling 与下列三种方法进行比较:

(1)随机采样<sup>[7]</sup>:概率分布为均匀的;

(2)MinSpec<sup>[1]</sup>:确定性算法,贪心优化 E-Optimal 准则,每次迭代需特征分解;

(3)Ed Free<sup>[12]</sup>:确定性算法,无需特征分解,利用局部算子,考虑了节点间的空间关系。

Cut Sampling 与 Ed Free 算法中的局部算子均使用热核函数  $g(\lambda) = e^{-a\lambda}$ ,其中  $a = vp_e p_m p_k / \lambda_{\max}$ ,  $v \in \mathbb{R}_+$  为可调参数,设置为 75,  $p_e = |\mathcal{E}|/N$  为连边概率,  $p_m = M/N$  为采样比例,  $p_k = K/N$  为带宽。

### 4.2 采样集比较

图2展示了  $N = 500$  的 Sensor 图上四种算法的采样结果,圈出节点构成采样集,大小为  $M = 15$ 。如图2(a)和图2(b),随机采样和 MinSpec 生成的采样集中,节点距离可能很近,分布并不均匀。这两种方法在采样时,随机或根据谱域准则优化,并未考虑图的拓扑结构与节点间的位置关系,这也导致了较大的重构误差。Ed Free 和 Cut Sampling 在采样时都考虑了节点的位置关系,最终的采样集分布较为均匀,采样节点间保持了一定距离,如图2(c)和图2(d)所示。同时,这两种方法都利用了局部算子,可以观察到采样集有部分重合。

实验均生成 100 次信号,并取均方误差的平均值。评价指标如下:

$$\text{MSE} = \frac{\| \mathbf{x}' - \mathbf{x} \|^2}{N} \quad (14)$$

其中,  $\mathbf{x}'$  表示重构信号。

实验结果表明,该算法在几乎所有场景下重构效果最优,算法在四种图上都取得了最小或与其他方法接近的重构误差。特别地,在 Sensor 图和 Minnesota 图上,Cut Sampling 稳定优于其他三种方法。对于 BA

图,当采样集足够大时,Cut Sampling 效果更好。对于 ER 图,当连边的概率较小时,Cut Sampling 重构效果接近其他方法,随着连边概率提高,Cut Sampling 相比于 Ed Free 的提升也明显增加。

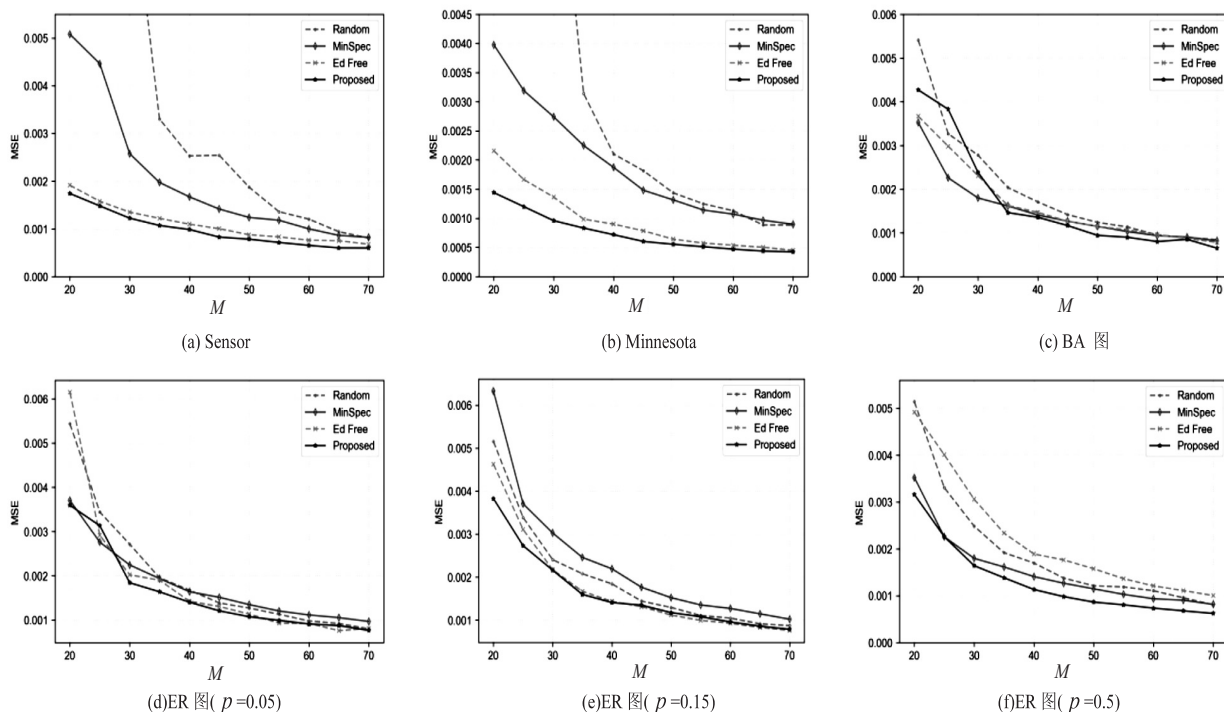


图 3 重构误差 MSE 随采样集大小  $M$  的变化(平均 100 次)

同时,可以观察到,对于考虑了图拓扑结构的 Cut Sampling 和 Ed Free,这两种方法的重构表现基本优于没有考虑节点位置关系的随机采样和 MinSpec,特别是在 Sensor 图和 Minnesota 图上。这也反映了将图拓扑信息与频域信息相结合,可有效提高采样算法的重构效果。

## 5 结束语

该文提出了一种两步图子集选取方法,先图割聚类,再簇内采样。首先利用局部算子的内积生成一张度量节点空间距离的完全图,再利用谱聚类算法求解该图的  $N$ -CUT,得到节点集依距离的划分。最后,在各个子集内,根据局部算子的稀疏性准则,分别选择最优点,组成最终的采样集。相比于大多数谱域算法,该方法结合了顶点域的空间信息去控制采样集内节点的距离,使得采样集分布相对均匀。同时,区别于常见的贪心优化框架,该文利用图割实现采样步骤可以并行选择全部节点,具有一定可扩展性。在多种图场景下与多种代表性算法相比,该方法能达到最优或接近的效果。

考虑到谱聚类算法的计算复杂度较高,如何更加高效准确的聚类是未来的一个研究方向。同时,对该算法进一步的理论分析也是未来工作之一。

## 参考文献:

- [1] CHEN S, VARMA R, SANDRYHAILA A, et al. Discrete signal processing on graphs; sampling theory [J]. IEEE Transactions on Signal Processing, 2015, 63 (24): 6510–6523.
- [2] EGILMEZ H E, ORTEGA A. Spectral anomaly detection using graph-based filtering for wireless sensor networks [C]//2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). Florence: IEEE, 2014:1085–1089.
- [3] HUANG W, BOLTON T A, MEDAGLIA J D, et al. A graph signal processing perspective on functional brain imaging [J]. Proceedings of the IEEE, 2018, 106(5): 868–885.
- [4] WANG X, CHEN J, GU Y. Local measurement and reconstruction for noisy bandlimited graph signals[J]. Signal Processing, 2016, 129:119–129.
- [5] VALSESIA D, FRACASTORO G, MAGLI E. Sampling of graph signals via randomized local aggregations [J]. IEEE Transactions on Signal and Information Processing over Networks, 2019, 5(2): 348–359.
- [6] CHEN S, VARMA R, SINGH A, et al. Signal recovery on graphs; random versus experimentally designed sampling [C]//2015 international conference on sampling theory and applications (SampTA). Washington: IEEE, 2015: 337–

- 341.
- [7] PUY G, TREMBLAY N, GRIBONVAL R, et al. Random sampling of bandlimited signals on graphs[J]. *Applied and Computational Harmonic Analysis*, 2018, 44(2): 446–475.
  - [8] PERRAUDIN N, RICAUD B, SHUMAN D I, et al. Global and local uncertainty principles for signals on graphs[J]. *APSIPA Transactions on Signal and Information Processing*, 2018, 7: e3.
  - [9] TSITSVERO M, BARBAROSSA S, DI LORENZO P. Signals on graphs; uncertainty principle and sampling[J]. *IEEE Transactions on Signal Processing*, 2016, 64(18): 4845–4860.
  - [10] ANIS A, GADDE A, ORTEGA A. Efficient sampling set selection for bandlimited graph signals using graph spectral proxies[J]. *IEEE Transactions on Signal Processing*, 2016, 64(14): 3775–3789.
  - [11] JAYAWANT A, ORTEGA A. A distance-based formulation for sampling signals on graphs[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). Calgary: IEEE, 2018: 6318–6322.
  - [12] SAKIYAMA A, TANAKA Y, TANAKA T, et al. Eigendecomposition-free sampling set selection for graph signals[J]. *IEEE Transactions on Signal Processing*, 2019, 67(10): 2679–2692.
  - [13] KIM Y H. Qr factorization-based sampling set selection for bandlimited graph signals[J]. *Signal Processing*, 2021, 179: 107847.
  - [14] HASHEMI A, SHAFIPOUR R, VIKALO H, et al. Towards accelerated greedy sampling and reconstruction of bandlimited graph signals[J]. *Signal Processing*, 2022, 195: 108505.
  - [15] JANSSEN S, DUMONT G, FIERENS F, et al. Spatial interpolation of air pollution measurements using corine land cover data[J]. *Atmospheric Environment*, 2008, 42(20): 4884–4903.
  - [16] GADDE A, ANIS A, ORTEGA A. Active semi-supervised learning using sampling theory for graph signals[C]//KDD'14: proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2014: 492–501.
  - [17] CANDÈS E J, RECHT B. Exact matrix completion via convex optimization[J]. *Found. Comput. Math.*, 2009, 9(6): 717–772.
  - [18] CIVRIL A, MAGDON-ISMAIL M. On selecting a maximum volume sub-matrix of a matrix and related problems[J]. *Theor. Comput. Sci.*, 2009, 410(47–49): 4801–4811.
  - [19] TANAKA Y, ELDAR Y C, ORTEGA A, et al. Sampling signals on graphs; from theory to applications[J]. *IEEE Signal Processing Magazine*, 2020, 37(6): 14–30.
  - [20] 谢 焯, 冯 辉, 胡 波, 等. 带限图信号的最优采样集设计[J]. *系统工程与电子技术*, 2022, 44(2): 357–364.
  - [21] WANG F, WANG Y, CHEUNG G. A-optimal sampling and robust reconstruction for graph signals via truncated neumann series[J]. *IEEE Signal Processing Letters*, 2018, 25(5): 680–684.
  - [22] KRAUSE A, SINGH A, GUESTRIN C. Near-optimal sensor placements in Gaussian processes: theory, efficient algorithms and empirical studies[J]. *J. Mach. Learn. Res.*, 2008, 9: 235–284.
  - [23] SHUMAN D I, RICAUD B, VANDERGHEYNST P. Vertex-frequency analysis on graphs[J]. *Applied and Computational Harmonic Analysis*, 2016, 40(2): 260–291.
  - [24] SHI J, MALIK J. Normalized cuts and image segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22(8): 888–905.
  - [25] KAY S M. Fundamentals of statistical signal processing, volume i: estimation theory[M]. NJ: Prentice-Hall, Inc., 1993.
  - [26] NG A, JORDAN M, WEISS Y. On spectral clustering: analysis and an algorithm[C]//NIPS'01: proceedings of the 14th international conference on neural information processing systems; natural and synthetic. Cambridge: MIT Press, 2001: 849–856.
  - [27] PERRAUDIN N, PARATTE J, SHUMAN D I, et al. GSP-BOX: a toolbox for signal processing on graphs[J]. *arXiv*: 1408.5781, 2014.