哈尔滨工业大学计算机科学与技术学院

《算法设计与分析》

课程报告

学号	
姓名	
班级	
专业	人工智能
授课教师	
报告日期	2022 年 12 月 16 日

论文题目

Faster Cut Sparsification of Weighted Graphs

作者：Sebastian Forster, Tijn de Vos

刊物：ICALP 2022

出处链接：https://doi.org/10.48550/arXiv.2112.03120

1 问题定义

在许多应用中，图变得越来越大，因此存储和处理这些图成为一个具有挑战性的问题。处理这个问题的一个策略是图的稀疏化，在这里我们用一组保留某些属性的，稀疏的（重新加权的）边来对图进行建模。特别是当我们的目标是处理较大的输入图，这个过程对于图的大小来说应该是有效的。

在不同类型的图稀疏器中，本文重点讨论 Benczúr 和 Karger[BK96, BK15]提出的切割稀疏器。我们说，一个（重新加权的）子图 $H \subseteq G$，对于 G 中每一个切 C 是一个$(1 \pm \epsilon)$-的切割稀疏器，H 中切口边的总权重$w_H(C)$是在 G 中切口边总权重$w_G(C)$的 $1 \pm \epsilon$倍范围内。

对于大小为$O(n\log(n)/\epsilon^2)$的切割稀疏器。Benczúr 和 Karger[BK96, BK15]表明，对于多项式加权图，这些可以在时间$O\big(m\log^2(n)\big)$内计算出来，而对无界权重的图来说，则可以在$O\big(m\log^3(n)\big)$时间内完成。在本文中，我们对具有多项式有界和无界权值为的图的结果进行了改进。

Algorithm	Size	Running time
Unweighted		
[FHHP19]	$O\left(n\log(n)/\epsilon^2\right)$	$O\left(m\right)$
Polynomial weights		
[BK15]	$O\left(n\log(n)/\epsilon^2\right)$	$O\left(m\log^2(n)\right)$
[FHHP19]	$O\left(n\log^2(n)/\epsilon^2\right)$	$O\left(m\right)$
[FHHP19] + [BK15]	$O\left(n\log(n)/\epsilon^2\right)$	$O\left(m+n\log^4(n)/\epsilon^2\right)$
This paper	$O\left(n\log(n)/\epsilon^2\right)$	$O(m\log(n))$
This paper	$O\left(n\log(n)/\epsilon^2\right)$	$O(m\alpha(n)\log(m/n))$
[FHHP19] + this paper	$O\left(n\log(n)/\epsilon^2\right)$	$O\left(m+n\left(\log^2(n)/\epsilon^2\right)\alpha(n)\log(\log(n)/\epsilon)\right)$
Unbounded weights		
[HP10]	$O\left(n\log^2(n)/\epsilon^2\right)$	$O\left(m\log^2(n)/\epsilon^2\right)$
[BK15]	$O\left(n\log(n)/\epsilon^2\right)$	$O\left(m\log^3(n)\right)$
[HP10] + [BK15]	$O\left(n\log(n)/\epsilon^2\right)$	$O\left(m\log^2(n)/\epsilon^2+n\log^5(n)/\epsilon^2\right)$
[LS17]	$O\left(n/\epsilon^2\right)$	$O\left(m\cdot\mathrm{poly}(\log(n),1/\epsilon)\right)$
This paper	$O\left(n\log(n)/\epsilon^2\right)$	$O(m\log(n))$
This paper	$O\left(n\log(n)/\epsilon^2\right)$	$O(m\alpha(n)\log(m/n))$

图 1：计算具有整数权重的无向图的切割疏散器的最新算法概述。算法$A+B$表示算法B是用算法A进行预处理的。

2 算法描述和分析

2.1 算法描述

我们的稀疏化算法的高层设置与Fung等人[FHHP19]的非加权图的算法相似。我们的主要贡献在于通过使用最大生成森林（MSF）指数而非Nagamochi-Ibaraki（NI）指数，展示了如何将该技术推广到加权图；另一方面，我们证明了通过收紧分析，大小和时间界限以高概率成立，而不仅仅是在期望值中。

在本文中，我们使用了MSF指数的独特属性，而NI指数是不够的。我们表明，利用MSF指数，我们可以将非加权图的稀疏化算法推广到加权图的算法中，从而证明MSF指数是加权图中NI指数的自然类似物。我们提供了一种算法来计算一个M-部分MSF包装的算法，时间为$O\left(m \cdot \min\left(\alpha(n)\log(M), \log(n)\right)\right)$。我们还提供了一种算法来计算多项式加权图的部分MSF包装。我们表明，对于无界权值，可以计算一个充分的估计，时间也是$O\left(m \cdot \min\left(\alpha(n)\log(M), \log(n)\right)\right)$。

在多项式加权的输入图的情况下，该算法包括两个主要阶段。在第一阶段，我们计算集合$F_0, F_1, ..., F_T \subseteq E$，其中的边满足对分离其端点的任何切口的权重的一些下限。在第二阶段，我们以相应的概率从每个集合F_i中抽取边。

我们设定一个参数$\rho = \Theta\left(\frac{\ln(n)}{\epsilon^2}\right)$，并从计算$G$的一个$2\rho$-的部分最大生成森林包装开始。我们定义$F_0$是这些$2\rho$森林的并集。我们将$F_0$的边添加到$G_\epsilon$，这将成为我们的稀疏器。我们以$1/2$的概率对剩下的每条边$E \setminus F_0$进行抽样来构建$X_1$。为了平衡抽样，我们将提高每条抽样边的权重，系数为2。现在我们沿着这些思路继续下去，但在每次迭代中我们让F_i由指数级增长的跨越森林组成。F_i被定义为是X_i的一个$(2^{i+1} \cdot \rho)$-的部分MSF包装的联合体。然后，从剩余的边$X_i \setminus F_i$中采样X_{i+1}，其中每条边的概率为$1/2$。我们继续这个过程，直到有足够多的边留在了X_{i+1}，我们将这些剩余的边添加到G_ϵ。

算法的第二阶段是对集合F_i中采样边，并将这些抽样的边添加到G_ϵ。在此，请注意，F_i中的一个边e(对于$i \geq 1$)不属于F_{i-1}的一部分，也就是说，它不属于X_{i-1}的$(2^i \cdot \rho)$-的部分MSF包装中任何生成森林的一部分。这意味着，对于一条边$e \in F_i$，X_{i-1}中包含的任何切口的权重至少是$2^i \cdot \rho \cdot w(e)$。现在我们使用Fung等人[FHHP19]的切口稀疏化的一般框架，它归结为这样一个事实：对切口权重的这种保证意味着我们可以从F_i中抽取概率为$1/\left(2^i w(e)\right)$的边，我们表明，这将生成一个足够稀疏的图。

直观地讲，从$X_i \setminus F_i$中抽取边来形成X_{i+1}似乎是多余的。但这的确是保证所得到的图是一个稀疏图的必要条件。它确保了迭代次数的限制，这保证了对疏散器大小和运行时间的更好约束。由于我们在每个阶段以概率$1/2$对边缘进行抽样，因此在每个阶段，我们需要重复采样$O\left(\log\left(m/(m_0)\right)\right)$次，以使$X_i$缩小到$O(m_0)$。由于这个步骤的数量取决于初始的边的数量$m$，我们可以得到更好的大小和运行时间的界限，如果$m$已经很小了。我们将利用这一点，用[FHHP 19]中的一个算法对图进行预处理，该算法可以在线性时间内得到一个大小为$O(n\log^2(n)/\epsilon^2)$的切

割稀疏器，我们可以证明反复调用我们的算法并不比调用一次更差的渐进时间界限，因为输入图很快就变得稀疏了。这样，我们得到一个大小为$O(n\log(n)/\epsilon^2)$的稀疏器。

2.2 算法分析

在这一节中，我们将介绍计算加权图G的一个$(1\pm\epsilon)$-切疏散器G_ϵ的算法。该算法利用了切割稀疏化的框架和最大生成森林包装。本节致力于介绍以下关于多项式加权图的定理，并将把本节的技术推广到权重不受限制的图上。

定理 2.1. 存在一种算法，即给定一个加权图$G=(V,E)$和自由选择的参数$\epsilon>0$，计算出一个图G_ϵ的算法，这是G的一个高概率的$(1\pm\epsilon)$-的切分疏散器。该算法的运行时间为$O\left(m\cdot\min\left(\alpha(n)\log(m/n),\log(n)\right)\right)$，$G_\epsilon$的边数为$O\left(n(\log(n)/\epsilon^2)\log(m/(n\log(n)/\epsilon^2))\right)$。

准确地说，我们给出了一种算法，在这种算法中，疏散器的运行时间和大小的给定界限都大概率地成立。通过简单地在运行时间超过界限时停止，并在超过大小界限时输出一个空图，这就得到了上述结果。

为了对疏散器的大小实现更好的约束，我们对输入图反复应用这个定理，其精度参数呈指数递减。

2.2.1 算法

为了对图进行稀疏化，我们使用了两种抽样方法。其中之一是切割稀疏化的框架。然而，不是直接将该框架应用于图，而是在它之前有另一个抽样过程。

为了简化方程，我们设定$\rho:=\frac{(7+c)1352\ln(n)}{0.38\epsilon^2}$，如果$|E|\le 4\rho n\log\left(m/(n\log(n)/\epsilon^2)\right)$，我们不做处理，即返回$G_\epsilon=G$；否则，我们从一个初始化步骤开始，继续一个迭代过程，当剩余的图形变得足够小时，这个过程就结束了。

在初始化步骤中，我们定义$X_0:=E$，我们计算出一个$\lfloor 2\rho\rfloor$-的部分最大生成森林包装$T_1,\ldots,T_{\lfloor 2\rho\rfloor}$，并且我们定义$F_0:=\bigcup_{j=1}^{\lfloor 2\rho\rfloor}T_j$，剩余的边$Y_0:=X_0\setminus F_0$进入下一个阶段。

在迭代i中，我们从Y_i对每条边进行概率为 1/2 的抽样来创建X_{i+1}，接下来，我们计算$k_i:=\rho\cdot 2^{i+1}$的最大跨度森林$T_1,\ldots,T_{k_i}$，我们定义$F_i:=\bigcup_{j=1}^{k_i}T_j$，并且$Y_i:=X_i\setminus F_i$。

我们继续，直到Y_i有最多的 $2\rho n$条边，并设Γ为迭代次数。我们保留所有的边在F_0，换句话说：将每条边$e\in F_0$以权重$w(e)$加到G_ϵ，Y_Γ的边也被保留，但它们需要被缩放以抵消$\Gamma-1$ 采样步骤：将每条边$e\in Y_\Gamma$以权重$2^{\Gamma-1}w(e)$添加到G_ϵ。

在X_{i-1}中，任何其他边$e\in F_i$至少是$k_iw(e)$-的权重，例如$e\notin F_{i-1}$。我们利用这种权重，用框架从这些边上取样，对于每个$e\in F_i$，我们：

- 定义$n_e:=2^iw(e)$和$p_e:=\min\left(1,\frac{384}{169}\frac{1}{4^iw(e)}\right)$；
- 生成r_e的二项分布，参数为n_e和p_e；
- 如果r_e为正数，则将e加到G_ϵ以权重r_e/p_e。

在调用二项分布的系数2^i时，可以看作是将边的权重提高了一个系数2^i，这是平衡在创建F_i过程中的i采样步骤所必需的。就 MSF 打包的计算方法而言，所提出的算法对于多项式和超多项式加权的图是相同的，对于无界的情况，我们使用 MSF 指数估计器。

2.2.2 正确性

我们将证明G_ϵ中构建的 SparSIFY$(V, E, w, \epsilon, \mathrm{c})$是一个$(1 \pm \epsilon)$-的切分疏散器。$G$的概率至少为 $1 - 8/n^c$。按照[FHHP19]的证明结构，我们首先定义

$$S := \left(\bigcup_{i=0}^{\Gamma} 2^i F_i\right) \cup 2^\Gamma Y_\Gamma$$

其中Γ是最大的数字，使得$F_i \neq \emptyset$。我们定义$G_S := (V, S)$，证明以下两个定理，它们共同产生了所需的结果。

定理 2.2. G_S是一个G的$(1 \pm \epsilon/3)$-的切分疏散器，G的概率至少为 $1 - 4/n^c$。

推理 2.3. G_ϵ是一个G_S的$(1 \pm \epsilon/3)$-的切分疏散器，G_S的概率至少是 $1 - 4/n^c$。

Algorithm 1: SPARSIFY(V, E, w, ϵ, c)

Input: An undirected graph $G = (V, E)$, with integer weights $w: E \to \mathbb{N}^+$, and parameters $\epsilon \in (0, 1)$, $c \geq 1$.

Output: An undirected weighted graph $G_\epsilon = (V, E_\epsilon)$.

Set $\rho \leftarrow \frac{(7+c)1352\ln(n)}{0.38\epsilon^2}$.
if $|E| \leq 4\rho n \log\left(m/(n\log(n)/\epsilon^2)\right)$ **then**
  return $G_\epsilon = G$.
end
Compute an $\lfloor 2\rho \rfloor$-partial maximum spanning forest packing $T_1, T_2, \ldots, T_{\lfloor 2\rho \rfloor}$ for G.
Set $i \leftarrow 0$.
Set $X_0 \leftarrow E$.
Set $F_0 \leftarrow \bigcup_{j=1}^{\lfloor 2\rho \rfloor} T_j$.
Set $Y_0 \leftarrow X_0 \setminus F_0$.
while $|Y_i| > 2\rho n$ **do**
  Sample each edge in Y_i with probability $1/2$ to construct X_{i+1}.
  $i \leftarrow i + 1$.
  Set $k_i \leftarrow \rho \cdot 2^{i+1}$.
  Compute an k_i-partial maximum spanning forest packing $T_1, T_2, \ldots, T_{k_i}$ for the graph $G_i := (V, X_i)$.
  Set $F_i \leftarrow \bigcup_{j=1}^{k_i} T_j$
  Set $Y_i \leftarrow X_i \setminus F_i$.
end
Set $\Gamma \leftarrow i$. // Γ is the number of elapsed iteration in the previous while-loop.
Add each edge $e \in Y_\Gamma$ to G_ϵ with weight $2^{\Gamma-1}w(e)$.
Add each edge $e \in F_0$ to G_ϵ with weight $w(e)$.
for $j = 1, \ldots, \Gamma$ **do**
  foreach $e \in F_j$ **do**
    Set $p_e \leftarrow \min\left(1, \frac{384}{169} \frac{1}{4^j w(e)}\right)$.
    Generate r_e from Binom$(2^j w(e), p_e)$.
    if $r_e > 0$ **then**
      Add e to G_ϵ with weight r_e/p_e.
    end
  end
end
return $G_\epsilon = (V, E_\epsilon)$.

让我们先来证明定理 2.2。在创建集合F_i时，我们反复使用 MSF 指数。一条边的 MSF 指数确保了该边的某种连接性。下面的定理使这一点变得精确。

定理 2. 4. 设$i \geq 0$和$e \in Y_i$是一条边，并设$k_i := \rho \cdot 2^{i+1}$。那么$e$是$w(e)k_i$ −权重在$G'_{i,e} = \left(V, X'_{i,e}\right)$，其中$X'_{i,e} := \left\{e' \in X_i : w\left(e'\right) \geq w(e)\right\}$，因此，e 也是$w(e)k_i$ − 权重在$G_i = (V, X_i)$中。

接下来，我们在一般情况下表明，某些取样的方式可以保留切割。下面的定理是对[FHHP19]中的定理的推广。

定理 2. 5. 设$R \subseteq Q$是某个顶点集合上的加权边的子集V，对于所有$e \in Q$满足 $0 < w(e) \leq 1$。此外，假设在(V, Q)的每条边R都是π-重的。推断每条边$e \in R$被抽样的概率为$p \in (0,1]$，如果被选中，给定一个权重为$w(e)/p$来形成一个边$\hat{R}$的集合。我们表示，对于每个切C:

$$r^{(C)} := \sum_{e \in R \cap C} w(e), \quad q^{(C)} := \sum_{e \in Q \cap C} w(e), \quad \hat{r}^{(C)} := \sum_{e \in = \hat{R} \cap C} w(e)/p$$

让$\zeta \in \mathbb{N}_{\geq 5}$，并且$\delta \in (0,1]$，这样一来$\delta^2 p\pi \geq \frac{\zeta \ln(n)}{0.38}$，则

$$\left|r^{(C)} - \hat{r}^{(C)}\right| \leq \delta q^{(C)}$$

对于所有切口C的概率至少是 $1 - 4/n^{\zeta - 4}$。

而联合约束给了我们这样的结论：对于所有的切口来说，这条定理都是正确的，$C \in \mathcal{C}_j$的概率至少为 $1 - 2n(4 - \zeta)2^j$，借此可证明上述结论。

我们想把这个定理应用于我们的抽样程序。我们通过分别考虑不同的权重等级来做到这一点。我们定义$X_{i,k} := \left\{e \in X_i : 2^k \leq w(e) \leq 2^{k+1} - 1\right\}$，以及$x_{i,k}^{(C)} = \sum_{e \in X_{i,k} \cap C} w(e)$。我们类似地定义$Y_{i,k}$和$y_{i,k}^{(C)}$。为了确保所有的权重都位于$(0,1]$中，必须进行一些重构，这是定理 2.5 所要求的。对于$A \subseteq E$和$\beta > 0$，我们写成βA来表示我们将边的权重乘以一个系数β。

定理 2. 6. 在概率至少为 $1 - 4/n^{4+c}$的情况下，对于在G_i中每一个切割C,

$$\left|2^{-k} x_{i+1,k}^{(C)} - 2^{-k-1} y_{i,k}^{(C)}\right| \leq \frac{\epsilon/13}{2^{i/2+1}} \sum_{k'=k}^{\infty} 2^{-k'-1} x_{i,k'}^{(C)}$$

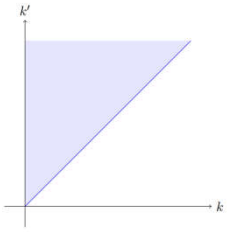


图 2：一个可视化的区域，其覆盖范围为 $\sum_{k=0}^{\infty} \sum_{k'=k}^{\infty} 1 = \sum_{k'=0}^{\infty} \sum_{k=0}^{k'} 1$

因此，我们可以通过引理 2.5 和这些设置来得到：

$$\left|2^{-k}x_{i+1,k}^{(C)} - 2^{-k-1}y_{i,k}^{(C)}\right| \leq \frac{\epsilon/13}{2^{i/2+1}}\sum_{k'=k}^{\infty} 2^{-k'-1}\,x_{i,k'}^{(C)},$$

这对所有概率为 $1 - 4/n^{3+c}$的切割C都成立。

现在我们看一下一般情况，对于这种情况，我们对所有的重量级进行加总。在此，我们定义 $x_i^{(C)} = \sum_{e\in X_i\cap C} w(e), x_{i+1}^{(C)} = \sum_{e\in X_{i+1}\cap C} w(e)$，以及 $y_i^{(C)} = \sum_{e\in Y_i\cap C} w(e).$

推论 2. 7. 对于在G_i中每一个概率至少为 $1 - 4/n^{1+c}$的切割C,

$$\left|2x_{i+1}^{(C)} - y_i^{(C)}\right| \leq \frac{\epsilon/13}{2^{i/2}}\cdot x_i^{(C)}.$$

我们将反复应用这个定理。为了说明累积误差不会超过$\epsilon/3$，我们使用下面的事实。关于证明，我们参考了[FHHP19]。

定理 2. 8. 设$x \in (0,1]$是一个参数。那么对于任何$k \geq 0$,

$$\prod_{i=0}^{k}\left(1+\frac{x/13}{2^{i/2}}\right) \leq 1 + x/3$$

$$\prod_{i=0}^{k}\left(1-\frac{x/13}{2^{i/2}}\right) \geq 1 - x/3$$

作为证明定理 2.2 的最后一步，我们证明了一个定理，该定理集中在我们算法的最后一个$\Gamma - j + 1$ 迭代步骤中发生的稀疏化问题。

定理 2. 9. 设

$$S_j = \left(\bigcup_{i=j}^{\Gamma} 2^{i-j}F_i\right) \cup 2^{\Gamma-j}Y_\Gamma$$

对于任何$j \geq 0$，那么，S_j是一个$G_j = (V, X_j)$的$\left(1 \pm (\epsilon/3)2^{-j/2}\right)$-的切分疏散器，其概率至少为 $1 - 4/n^c$。

定理 2. 10. 每条边$e \in R_{j,k}$在$(V, E_{j,k})$中都是$\pi := \rho \cdot 4^\Gamma 2^\Lambda$-权重。

现在我们把所有的权重类放在一起，找到满足Π-连接性的子图集$\mathcal{G}$。

推论 2. 11. 在$G_i = (V, E_i)$中的每条边$e \in R_i$都是$\rho \cdot 4^\Gamma 2^\Lambda$-权重的，且$E_i$ $:= \bigcup_{j=1}^{\min(\lfloor i/2\rfloor,\Gamma)} E_{j,i-2j}$。

只需证明γ-重合满足的条件。

定理 2. 12. 对于任何切C,

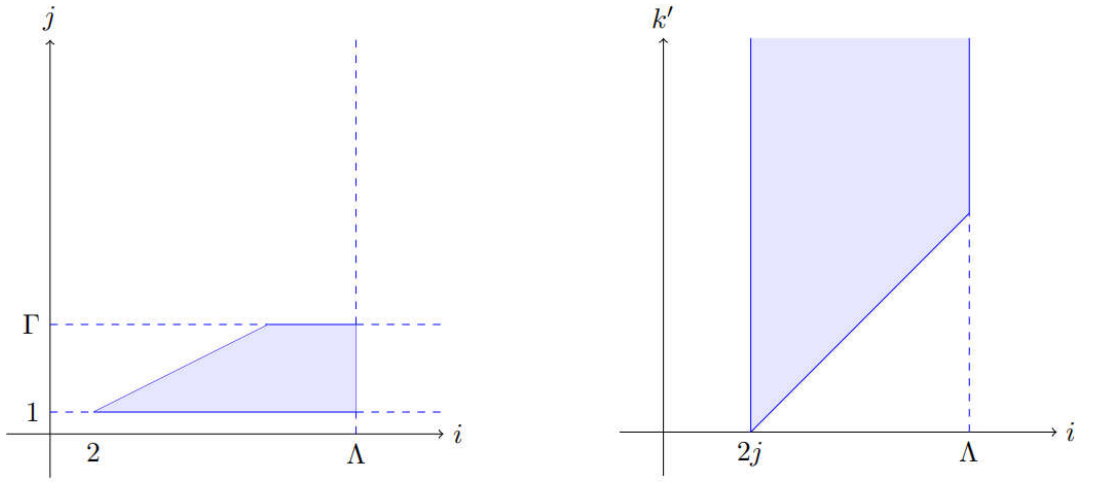
$$\sum_{i=0}^{\Lambda}\frac{e_i^{(C)}2^{i-1}}{\rho\cdot 4^\Gamma 2^\lambda} \leq 64/3\cdot e^{(C)}$$

其中$e^{(C)} = \sum_{e\in C} w_{GS}(e)$和 $e_i^{(C)} = \sum_{e\in C\cap E_i} w_{G_i}(e)$

证明：我们把F_0和Y_Γ加到G_ϵ，所以我们不需要关心切面C与这些集合的交集。这意味着我们只需将一个切口C与F_j相交，其中 $1 \le j \le \Gamma$。因此，我们从$i = 2$开始求和，我们考虑我们需要约束的总和。

$$\begin{aligned}
\sum_{i=2}^{\Lambda} \frac{e_i^{(C)} 2^{i-1}}{\rho \cdot 4^\Gamma 2^\lambda} &= \sum_{i=2}^{\Lambda} \frac{\left(\sum_{e \in C \cap E_i} w_{G_i}(e)\right) 2^{i-1}}{\rho \cdot 4^\Gamma 2^\lambda} \\
&= \sum_{i=2}^{\Lambda} \sum_{j=1}^{\min(\lfloor i/2 \rfloor, \Gamma)} \frac{\left(\sum_{e \in C \cap E_{j,i-2j}} w_{G_i}(e)\right) 2^{i-1}}{\rho \cdot 4^\Gamma 2^\lambda} \\
&= \sum_{i=2}^{\Lambda} \sum_{j=1}^{\min(\lfloor i/2 \rfloor, \Gamma)} \sum_{j'=j-1}^{\Gamma} \sum_{k'=i-2j}^{\infty} \frac{\rho \cdot 4^{\Gamma - j' + 1} 2^{\Lambda - k' + j'} \left(\sum_{e \in C \cap E_{j',k'}} w_G(e)\right) 2^{i-1}}{\rho \cdot 4^\Gamma 2^\lambda} \\
&= \sum_{i=2}^{\Lambda} \sum_{j=1}^{\min(\lfloor i/2 \rfloor, \Gamma)} \sum_{j'=j-1}^{\Gamma} \sum_{k'=i-2j}^{\infty} 2^{-k'-j'+i+1} \left(\sum_{e \in C \cap E_{j',k'}}^{w_G(e)}\right)
\end{aligned}$$

接下来，我们要交换i上的和和j上的和，并相应地改变界限。有关可视化参数，请参见图 3a。



(a) $\sum_{i=2}^{\Lambda} \sum_{j=1}^{\min(\lfloor i/2 \rfloor, \Gamma)} 1 = \sum_{j=1}^{\Gamma} \sum_{i=2j}^{\Lambda} 1$　　(b) $\sum_{i=2j}^{\Lambda} \sum_{k'=i-2j}^{\infty} 1 = \sum_{k'=0}^{\infty} \sum_{i=2j}^{\min(2j+k', \Lambda)} 1$

图 3：双重和所覆盖的区域的两种可视化。

$$\sum_{i=2}^{\Lambda}\sum_{j=1}^{\min(\lfloor i/2\rfloor,\Gamma)}\sum_{j'=j-1}^{\Gamma}\sum_{k'=i-2j}^{\infty}2^{-k'-j'+i+1}\left(\sum_{e\in C\cap E_{j',k'}}w_G(e)\right)$$

$$=\sum_{j=1}^{\Gamma}\sum_{i=2j}^{\Lambda}\sum_{j'=j-1}^{\Gamma}\sum_{k'=i-2j}^{\infty}2^{-k'-j'+i+1}\left(\sum_{e\in C\cap E_{j',k'}}w_G(e)\right)$$

将i和j'上的和互换并不改变界限，因为它们是相互独立的。当交换了i和k'的和时，我们必须更加小心，参见图 3b 的可视化参数

$$\sum_{j=1}^{\Gamma}\sum_{i=2j}^{\Lambda}\sum_{j'=j-1}^{\Gamma}\sum_{k'=i-2j}^{\infty}2^{-k'-j'+i+1}\left(\sum_{e\in C\cap E_{j',k'}}w_G(e)\right)$$

$$=\sum_{j=1}^{\Gamma}\sum_{j'=j-1}^{\Gamma}\sum_{k'=0}^{\infty}\sum_{i=2j}^{\min\left(2j+k',\Lambda\right)}2^{-k'-j'+i+1}\left(\sum_{e\in C\cap E_{j',k'}}w_G(e)\right)$$

$$\leq\sum_{j=1}^{\Gamma}\sum_{j'=j-1}^{\Gamma}\sum_{k'=0}^{\infty}2^{-k'-j'+2j+k'+2}\left(\sum_{e\in C\cap E_{j',k'}}w_G(e)\right)$$

$$=\sum_{j=1}^{\Gamma}\sum_{j'=j-1}^{\Gamma}2^{2j-j'+2}\left(\sum_{k'=0e\in C\cap E_{j',k'}}^{\infty}w_G(e)\right)$$

$$=\sum_{j=1}^{\Gamma}\sum_{j'=j-1}^{\Gamma}2^{2j-j'+2}\left(\sum_{e\in C\cap F_{j'}}w_G(e)\right)$$

接下来，我们要将j的和与j'的和交换，可以在图 4 中找到一个直观的论证。

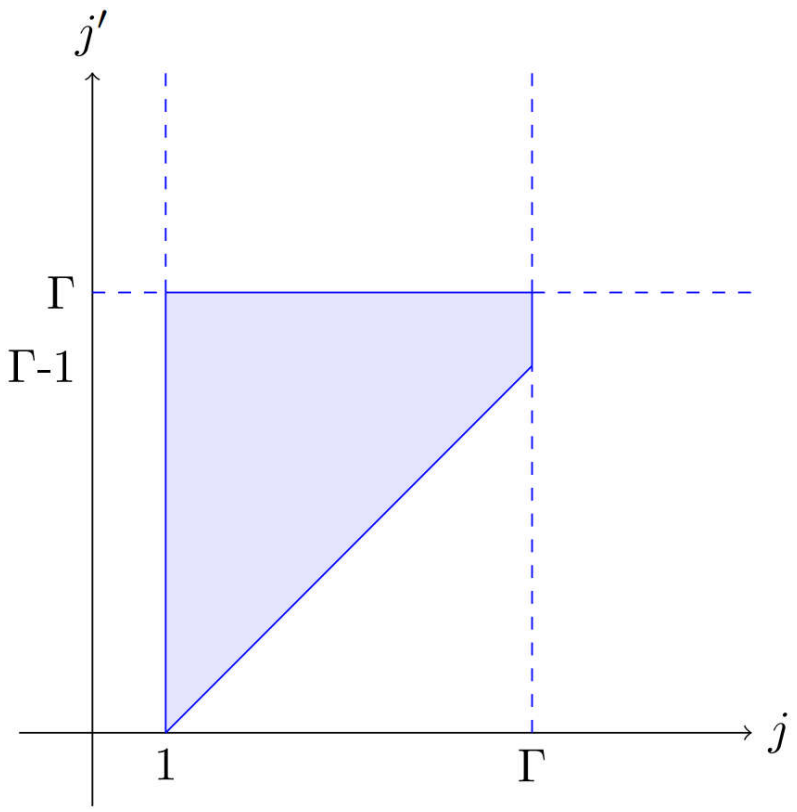


图 4：一个可视化的区域，其覆盖范围为 $\sum_{j=1}^{\Gamma}\sum_{j'=j-1}^{\Gamma} 1 = \sum_{j'=0}^{\Gamma}\sum_{j=1}^{j'+1} 1$.

$$\sum_{j=1}^{\Gamma}\sum_{j'=j-1}^{\Gamma} 2^{2j-j'+2}\left(\sum_{e\in C\cap F_{j'}} w_G(e)\right) = \sum_{j'=0}^{\Gamma} 2^{-j'+2}\sum_{j=1}^{j'+1} 4^{j}\left(\sum_{e\in C\cap F_{j'}} w_G(e)\right)$$

$$\leq \sum_{j'=0}^{\Gamma} 2^{-j'+2}\frac{4^{j'+2}}{3}\left(\sum_{e\in C\cap F_{j'}} w_G(e)\right)$$

$$= \frac{64}{3}\sum_{j'=0}^{\Gamma} 2^{j'}\left(\sum_{e\in C\cap F_{j'}} w_G(e)\right)$$

$$= \frac{64}{3}\sum_{e\in C} w_{Gs}(e)$$

$$= \frac{64}{3}e^{(C)}.$$

推论 2.11 和定理 2.12 表明定理 *2.3* 的条件在给定的参数下得到满足。这证明了引理 2.3，然后是定理 2.1。

2.2.3 疏散器的大小

疏散器G_ϵ由F_0, Y_Γ，和F'组成，其中$F' = \cup_{i=1}^{\Gamma} F_i'$，$F_i$是$F_i'$的抽样边缘。首先，请注意$|F_0| = O(cn\ln(n)/\epsilon^2)$和$|Y_\Gamma| = O(cn\ln(n)/\epsilon^2)$。现在取$e \in F_i$，这条边的结果是在$G_\epsilon$中，如果从二项分布中抽样，参数$n_e = 2^i w(e)$和$p_e = \min\left(1, \frac{384}{169}\frac{1}{4^2 w(e)}\right)$是正的。发生这种情况的概率是

$$\begin{aligned}
\mathbb{P}[\text{Binom}(n_e, p_e) > 0] &= \sum_{k=1}^{n_e} \mathbb{P}\left[\text{Binom}(n_e, p_e) = k\right] \\
&\le \sum_{k=1}^{n_e} k\, \mathbb{P}[\text{Binom}(n_e, p_e) = k] \\
&= \sum_{k=0}^{n_e} k\, \mathbb{P}[\text{Binom}(n_e, p_e) = k] \\
&= \mathbb{E}[\text{Binom}(n_e, p_e)] \\
&= n_e p_e \\
&\le \frac{384}{169} 2^{-i}.
\end{aligned}$$

请注意，这个概率对所有$e \in F_i$都是相等的。由于F_i是$k_i = \rho \cdot 2^{i+1}$生成林的联合，我们知道$|F_i| \le \rho 2^{i+1} n$。因此，F_i'的预期大小，即F_i中取样的边，等于

$$\begin{aligned}
\mathbb{E}\left[\left|F_i'\right|\right] &= \sum_{e \in F_i} \mathbb{P}\left[\text{Binom}(n_e, p_e) > 0\right] \\
&\le \sum_{e \in F_i} \frac{384}{169} 2^{-i} \\
&= |F_i| \frac{384}{169} 2^{-i} \\
&\le \rho 2^{i+1} n \frac{384}{169} 2^{-i} \\
&= \rho \frac{768}{169} n.
\end{aligned}$$

我们有，取样边缘的总数等于

$$\mathbb{E}\left[\left|F'\right|\right] = \sum_{i=1}^{\Gamma} \mathbb{E}\left[\left|F_i'\right|\right] \le \Gamma \rho \frac{768}{169} n,$$

因此，仍然需要约束Γ的数量，也就是F_i's。在此，请注意，第 10-17 行的 while 循环结束时，如果$|Y_i| \le 2\rho n$，我们通过约束X_i的边数来约束Y_i中的边的数

量，其中Y_i是一个子集。$Y_{i-1} \subseteq X_{i-1}$中的每条边都被以概率 *1/2* 采样来形成$X_i$。所以$\mathbb{E}[|X_i|] \leq |X_{i-1}|/2$。现在通过切尔诺夫约束，我们得到：

$$\mathbb{P}\left[|X_i| > \frac{2}{3}|X_{i-1}|\right] \leq \exp\left(-\frac{0.38}{36}|X_{i-1}|\right) > \exp\left(-\frac{cn\ln(n)}{36}\right) = n^{-cn/36},$$

因为$|X_{i-1}| \geq |Y_{i-1}| \geq 2\rho n = 2 \cdot \frac{(7+c)1352\ln(n)}{0.38\epsilon^2} n \geq \frac{cn\ln(n)}{0.38}$。我们最多可以有$n^2$集$X_i$，所以我们可以得出结论：在每一步中，有很大的概率$|X_i| \leq \frac{2}{3}|X_{i-1}|$，通过归纳法$|Y_i| < |X_i| \leq \left(\frac{2}{3}\right)^i m$。我们看到

$$m\left(\frac{2}{3}\right)^{\Gamma} \leq 2\rho n = \frac{21632}{0.38\epsilon^2} cn\ln(n),$$

这相当于

$$\left(\frac{2}{3}\right)^{\Gamma} \leq \frac{\frac{21632}{0.38\epsilon^2} cn\ln(n)}{m},$$

而这就相当于

$$\Gamma \geq \log\left(\frac{m}{\frac{21632}{0.38\epsilon^2} cn\ln(n)}\right)/\log(3/2).$$

因此，我们可以得出结论$\Gamma = O\left(\log\left(\frac{m}{\ln\log(n)/\epsilon^2}\right)\right)$。由此可知，抽样的边的总数预期是，

$$\mathbb{E}\left[\left|F'\right|\right] \leq \Gamma\rho\frac{768}{169}n = O\left(cn\log(n)\log\left(m/(cn\log(n)/\epsilon^2)\right)/\epsilon^2\right).$$

这个压缩过程也可以被看作在$\{1,0\}.^3$中的m取值的独立随机变量之和。我们刚刚计算出，预期值μ是最多为$Bcn\ln(n)\log\left(m/(cn\log(n)/\epsilon^2)\right)/\epsilon^2$，对于某些$B > 0$。利用这一点，我们应用切尔诺夫约束来得到采样边数的上限：

$$\mathbb{P}\left[\left|F'\right| > 2Bcn\ln n\log\left(m/(cn\log(n)/\epsilon^2)\right)/\epsilon^2\right] \leq \exp\left(-0.38Bcn\ln(n)\log\left(m/(cn\log(n)/\epsilon^2)\right)/\epsilon^2\right)$$
$$= n^{-0.38cnB\log\left(m/\left(cn\log(n)/\epsilon^2\right)\right)/\epsilon^2}.$$

我们的结论是，在很大的概率下，抽样的边的数量是

$$O\left(2Bcn\ln(n)\log\left(m/(cn\log(n)/\epsilon^2)\right)/\epsilon^2\right) = O\left(cn\log(n)\log\left(m/(cn\log(n)/\epsilon^2)\right)/\epsilon^2\right).$$

最后，我们得出的结论是，在很大的概率上，G_ϵ的边的数量被限定为$|E(G_\epsilon)| = |F_0| + |Y_\Gamma| + \left|F'\right| = O\left(cn\log(n)\log\left(m/(cn\log(n)/\epsilon^2)\right)/\epsilon^2\right)$。

2.2.4 时间复杂度

首先，如果$m \le 4\rho n\log\left(m/(n\log(n)/\epsilon^2)\right) = O\left(cn\log(n)/\epsilon^2\log\left(m/(n\log(n)/\epsilon^2)\right)\right)$，该算法不做任何处理，并返回原始图。所以在这个分析中，我们可以假设$m > 4\rho n\log\left(m/(n\log(n)/\epsilon^2)\right)$。我们分两个阶段分析该算法的时间复杂度，第一阶段包括计算所有$e \in E$的概率p_e，第二阶段是压缩边缘，给定这些概率。

第一阶段包含 while 循环的i次迭代（第 10-17 行）。在每次迭代中，我们以概率为 *1/2* 从$Y_i \subseteq X_i$采边来形成X_{i+1}，这最多需要时间$O(|X_i|)$。接下来，我们计算图形的最大跨度森林包装$G_{i+1} = (V, X_{i+1})$。我们知道，我们可以在$O(m_0 \cdot \min(\alpha(n)\log(M), \log(n)))$时间内计算含有$n$顶点和$m_0$边图的一个$M$-的部分最大生成森林包装，所以这个迭代最多需要$O\left(|X_{i+1}| \cdot \left(\min\left(\alpha(n)\log(k_{i+1}), \log(n)\right)\right)\right)$时间。如前所述，我们有很高的概率认为$|X_i| \le \left(\frac{2}{3}\right)^i m$，如果$m\alpha(n)\log(m/n) \le m\log(n)$，则我们得出结论：第一阶段的总时间最多为

$$\begin{aligned}\sum_{i=0}^{\Gamma} O\left(|X_i|\right) + O\left(|X_{i+1}|\alpha(n)\log(k_{i+1})\right) &= \sum_{i=0}^{\Gamma}\left(\frac{2}{3}\right)^i O(m) + \left(\frac{2}{3}\right)^{i+1} O\left(m\alpha(n)\log(\rho 2^{i+2})\right)\\ &\le 3O(m) + 3O\left(m\alpha(n)\log(\rho 2^{\Gamma})\right)\\ &= O\left(m\alpha(n)\log(m/n)\right).\end{aligned}$$

而如果$m\log(n) < m\alpha(n)\log(m/n)$，我们就可以知道，在第一阶段中，总时间最多为

$$\begin{aligned}\sum_{i=0}^{\Gamma} O\left(|X_i|\right) + O\left(|X_{i+1}|\log(n)\right) &= \sum_{i=0}^{\Gamma}\left(\frac{2}{3}\right)^i O(m) + \left(\frac{2}{3}\right)^{i+1} O\left(m\log(n)\right)\\ &\le 3O(m) + 3O\left(m\log(n)\right)\\ &= O(m\log n).\end{aligned}$$

在第二阶段，我们从参数为n_e和p_e的二项分布中抽取每条边e，我们将证明这可以用一个时间为$T = O(m)$的过程来完成。在此，我们使用[Dev80]中的二项式抽样算法，算法 2 中给出了该算法的伪代码。

Algorithm 2: BINOM(n, p)

Input: Two parameters n, p.
Output: A random sample from the binomial distribution with parameters n and p.

```
Set k ← −1, S ← 0.
while S < n do
    k ← k + 1.
    Generate u ∼ U(0, 1).
    S ← S + ⌊log(u)/ log(1 − p)⌋ + 1.
end
return k
```

很容易看出，这个算法需要$O(1 + k)$时间，其中k是输出。因此，如果二项分布的样本是k，这需要时间$O(1 + k)$。这意味着，总时间T等于m加上所有样本的总和。

对于每条边$e \in F_i$，我们需要从参数为n_e和p_e的二项分布中抽取。我们用T_e表示我们需要采样e的时间。由上可知，我们有$\mathbb{E}[T_e] = 1 + n_e p_e$。所以，预期的成功次数最多为

$$\mathbb{E}[T] = \sum_i \sum_{e \in F_i} \mathbb{E}[T_e] = \sum_i \sum_{e \in F_i} (1 + n_e p_e)$$

$$= \sum_i |F_i| + O\big(cn\log(n)\log\big(m/(n\log(n)/\epsilon^2)\big)/\epsilon^2\big),$$

如第 2.3 节所示。设$B > 0$，这样$\sum_i \sum_{e \in F_i} n_e p_e \le Bcn\ln(n)\log\big(m/(n\log(n)/\epsilon^2)\big)/\epsilon^2$。我们可以用这些$\sum_i \sum_{e \in F_i} n_e$随机变量的总和的切尔诺夫约束得到：

$$\mathbb{P}\left[T - \sum_i |F_i| > 2Bcn\ln(n)\log\big(m/(n\log(n)/\epsilon^2)\big)/\epsilon^2\right]$$
$$\le \exp\big(-0.38Bcn\ln(n)\log\big(m/(n\log(n)/\epsilon^2)\big)/\epsilon^2\big)$$
$$= n^{-0.38Bcn\log\left(m/\left(n\log(n)/\epsilon^2\right)\right)/\epsilon^2}$$

所以我们可以说，在很大的概率上，我们需要

$$T = \sum_i |F_i| + \left(T - \sum_i |F_i|\right) = O(m) + O\big(2Bcn\ln(n)\log\big(m/(n\log(n)/\epsilon^2)\big)/\epsilon^2\big)$$
$$= O(m)$$

时间来采样。

结论是，该算法对于多项式加权图来说需要时间

$$O(m \cdot \min\big(\alpha(n)\log(m/n), \log(n)\big) + O(m) = O(m \cdot \min\big(\alpha(n)\log(m/n), \log(n)\big)$$

2.3 举例说明

定理 3. 1. 存在一种算法，在给定一个加权图$G = (V, E)$和参数$M > 0$ 的情况下，可在时间$O\big(m\alpha(n)\log(M)\big)$内计算出一个 MSF 指数估计器$\tilde{f}_e$，对于每条边$e \in E'$: = $\{e \in E: w(e) > d(e)/n\}$与$f_e \le M$。

在这一节中，我们将叙述如何将上一节的算法用于权重不受限制的稀疏图。这方面的关键是定理 3.1，它表明，对于无界权重，我们可能无法准确计算 MSF 指数，但我们可以找到$w(e) > d(e)/n$的边e的估计值。回顾一下$d(e)$的定义：计算G的一个最大生成森林F，并定义$d(e)$为F中从u到v在的路径上各条边的最小权重，其中$e = (u, v)$。

对无界权重的唯一调整是，我们在算法 1 中第一次计算最大生成森林时，我们保留了任何$w(e) \le d(e)/n$的边$e \in E$。我们表明，我们可以有效地从这些顶点中取样，因为它们是由F_0很好地连接，在我们的疏散器中保留的初始 MSF。我们将通过使用$\lambda_e = \rho \cdot d(e)$对它们进行抽样。请注意，我们只需要在第一次计算 MSF 打包时把顶点留出来，在这之后，新图中的估计值$d(e)$只会减少，所以如果一个顶点在某个子图中满足了$w(e) \le d(e)/n$，它在初始图中也满足这个条件。

对于剩下的顶点，我们应用上一节中提出的算法。唯一的区别是，我们使用定理 *3.4* 来计算 MSF 指数的估计值。这意味着，如果一条边$e \in E$获得的估计指数是$\tilde{f}_e$对某些图E'而言，我们有在E'中e至少是$f_e w_e(1 - 1/n)$-的权重。为简单起见，我们用 $1 - 1/n \ge 1/2$。我们看到，这影响了两个使用权重的地方的分析：引理 2.6 和引理 2.10。

我们检查一下引理 2.6，我们看到我们应用引理 2.5 与$\delta^2 p\pi \ge \frac{\zeta \ln(n)}{0.38}$，对于某些$\delta, p, \pi$和$\zeta$，我们想应用这个定理，但要有$\tilde{\pi} = \pi/2$，因此我们设定$\tilde{\delta} = \sqrt{2}\delta$，如果我们想最终得到定理 2.6 的原始结果，我们设$\tilde{\epsilon} = \epsilon/\sqrt{2}$。这种常数因素的变化被吸收到算法的大小和运行时间的渐进式符号中。

我们研究的第二个法则是引理 2.10，它是抽样中的Π-连接性。在这里，有一个简单的解决方案：我们将在$E_{j,k}$中的所有边增加 2 倍，这就保证了所期望的Π-连接性。因此，所有的边在E_i中的所有边都被提升了 2 个因子，这也就意味着在$e_i(C)$中的系数 2，如定理 2.12 所表示的，结果是一个γ-的重合，$\gamma = \frac{128}{3}$，而不是$\frac{64}{3}$。

综上所述，我们可以说，当我们调用算法的时候，我们原来的分析是成立的，$\tilde{\epsilon} = \epsilon/\sqrt{2}$和$\tilde{\rho} = \frac{(7+c)2704\ln(n)}{0.38\epsilon^2}$，其中$\rho$的变化是$\gamma$变化的一个直接的结果。.

剩下的最后一件事是表明，当我们取样时，Π-连接性也被满足于边上的$e \in E$与$w(e) \le d(e)/n$。这是对推论 2.11 的一个扩展。

定理 3. 2. 假设 $e \in R_i$和 $w(e) \le d(e)/n$，则 e是 $\pi = \rho \cdot 4^{\Gamma} 2^{\Lambda}$-的权重，$G_i = (V, E_i)$且$E_i = \bigcup_{j=1}^{\min(\lfloor i/2 \rfloor, \Gamma)} E_{j,i-2j}$.

证明。我们知道在F_0中e是$d(e)$-的权重，所以我们要寻找在E_i中的出现：

$$E_i = \bigcup_{j=1}^{\min(\lfloor i/2 \rfloor, \Gamma)} \bigcup_{j'=j-1}^{\Gamma} \bigcup_{k'=i-2j}^{\infty} \rho \cdot 4^{\Gamma - j' + 1} 2^{\Lambda - k' + j'} \left\{ e' \in F_{j'} : 2^{k'} \le \rho \cdot w\left(e'\right) \le 2^{k'+1} - 1 \right\}$$
$$\supseteq \rho \cdot 4^{\Gamma+1} \bigcup_{k'=i-2}^{\infty} 2^{\Lambda - k'} \left\{ e' \in F_0 : 2^{k'} \le \rho \cdot w\left(e'\right) \le 2^{k'+1} - 1 \right\}$$

我们更仔细地观察e在这个特定的集合中的连通性。我们注意到F_0中$w\left(e'\right) \ge d(e)$中的从$u$到$v$任何一条路径上的边$e = (u, v)$，根据定义$d(e)$。所以我们只需要

考虑$e' \in F_0$与$\rho \cdot w\left(e'\right) \geq \rho \cdot d(e) = \lambda_e \geq 2^i$，因为$e \in R_i$。这意味着$e$是$d(e)$-权重的：

$$\bigcup_{k'=i}^{\infty} \left\{e' \in F_0 : 2^{k'} \leq \rho \cdot w\left(e'\right) \leq 2^{k'+1} - 1\right\}$$
$$\subseteq \bigcup_{k'=i-2}^{\infty} \left\{e' \in F_0 : 2^{k'} \leq \rho \cdot w\left(e'\right) \leq 2^{k'+1} - 1\right\}$$

我们可以对其进行重新划分，以充分发掘权重。e是$\bigcup_{k'=i-2}^{\infty} 2^{\Lambda-k'}\left\{e' \in F_0 : 2^{k'} \leq \rho \cdot w\left(e'\right) \leq 2^{k'+1} - 1\right\}$中$2^{\Lambda}$-的权重。将此与方程 1 结合起来，就可以得出$e$在$E_i$中是$\rho \cdot 4^{\Gamma+1} 2^{\Lambda}$-权重的，这比我们需要显示的要多四个要素。

3 讨论和分析

在本文中，我们提出了一种更快的$(1 \pm \epsilon)$-的切割稀疏化算法。我们已经展示了如何在$O\left(m \cdot \min\left(\alpha(n)\log(m/n), \log(n)\right)\right)$时间内计算大小为$O(n\log(n)/\epsilon^2)$中的疏散器，对于整数加权图。这两种算法都采用了抽样技术，其中 MSF 指数被用作连接性估计。

我们已经证明，我们可以在$O\left(m\alpha(m)\log(M)\right)$时间内计算出多义权重的图一个$M$-的部分 MSF 打包。对于具有无界整数权重的图，我们已经证明我们可以在$O\left(m\log(n)\right)$时间内计算出完整的 MSF 打包，并且可以在时间$O\left(m\alpha(m)\log(M)\right)$内计算出来一个$M$-部分 MSF 打包。但是，存在一个悬而未决的问题，是否有可能进行更有效的计算，这将改进我们的稀疏化算法，但在其他应用中也可能是有利的。NI 指数已被证明在各种应用中是有用的，我们相信已经表明 MSF 指数是一个自然的类似物。

为了开发一种计算 MSF 打包的算法，人们可能倾向于建立在一种比 Kruskal 算法更快地计算最小生成树的算法上，例如 Karger、Klein 和 Tarjan[KKT95]的著名的线性时间算法。然而，这个算法和其他许多快速的最小生成树算法都使用了边收缩，如何将其推广到打包中还很不明显：在这种情况下，我们需要同时在多棵树上工作，因此我们不能简单地收缩输入图，以支持任何单一的树。为了使其发挥作用，似乎有必要对数据结构进行更细致的使用，这可能会限制我们算法的应用。

线性时间计算 MSF 指数将是一个最终目标。然而，对于我们的应用来说，稍微宽松的约束就足够了。如果我们能将计算 MSF 指数的运行时间减少到$O\left(m + n\log(n)\right)$，那么我们就可以得到一个$O(m)$的时间界限，这样我们的算法就得到进一步的改进。

参考文献

[1]才力,王永滨,杨莹,巩微.一种基于加权有向图的神经网络稀疏化算法[J].哈尔滨工程大学学报,2006,27(B07):18-21

[2]李涛,王卫卫,翟栋,贾西西.图像分割的加权稀疏子空间聚类方法[J].系统工程与电子技术,2014,36(03):580-585.
[3]许仕杰.基于加权拉普拉斯方法的多层次图分割[D].中国科学技术大学,2020.DOI:10.27517/d.cnki.gzkju.2020.000897.
[4]李小平,王卫卫,罗亮,王斯琪.图像分割的改进稀疏子空间聚类方法[J].系统工程与电子技术,2015,37(10):2418-2424.
[5]岳温川,王卫卫,李小平.基于加权稀疏子空间聚类多特征融合图像分割[J].系统工程与电子技术,2016,38(09):2184-2191.