Tendency-Aware Directed Graph Clustering based on Bi-directional Connections

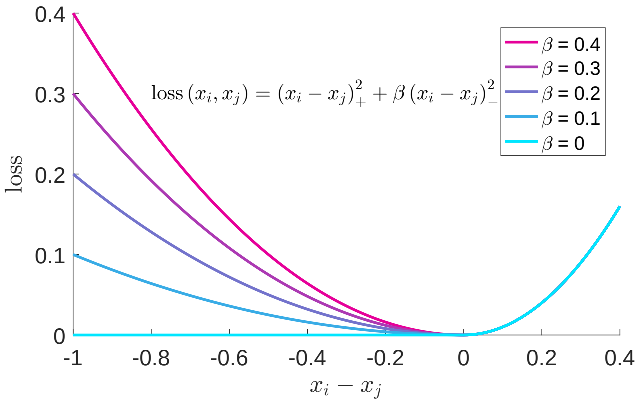

Figure 1: A novel loss function, termed the $Mutual\ Loss$, is proposed by us. The basic element of it with $W_{ij} = 1$ is shown.

I. Model Description

In this section, we introduce our proposed optimization objectives, and the algorithms for digraph clustering. The problems for two clusters are solved at first, and then we generalize it to other cases further.

A. Optimization Objectives

Firstly, we give a supporting lemma on minimizing NMcut.

Lemma 1. *Minimizing* $\mathrm{NMcut}\,(\mathcal{V}_1, \mathcal{V}_2)$ *is NP-hard.*

Proof: Minimizing $\mathrm{NMcut}\,(\mathcal{V}_1, \mathcal{V}_2)$ for undirected graphs is NP-hard, as proven in [1], which is a special case of the NMcut problem. Thus, it is also NP-hard. ■

The optimal solutions of NP-hard problems cannot be found within polynomial time. Our strategy is to find good enough results for it while taking acceptable computation costs. To solve the problem efficiently, we do a skillful transformation on it at first. Before this, some functions are defined: $[x]_+ = \max\{x, 0\}$ and $[x]_- = \min\{x, 0\}$, with respect to a scalar $x \in \mathbb{R}$, and the mutual loss

$$\mathrm{MLoss}\,(\mathbf{x}) := \sum_{i,j=1}^{N} W_{ij} \left([x_i - x_j]_+^2 + \beta\,[x_i - x_j]_-^2 \right) \quad (1)$$

on $\mathbf{x} \in \mathbb{R}^N$ with $\beta + \alpha = 1$, as shown in Figure 1. Then, to transform the NMcut problem, we have the supporting lemma.

Lemma 2. *Minimizing* NMcut *in* (**??**) *is equivalent to*

$$\begin{aligned} &\min_{\{\mathcal{V}_1,\mathcal{V}_2\}} \mathrm{NMcut}\,(\{\mathcal{V}_1, \mathcal{V}_2\}) \Longleftrightarrow \\ &\min_{\mathbf{y}\,\in\,\mathbb{R}^N} \ \frac{\mathrm{MLoss}(\mathbf{y})}{\mathbf{y}^T\mathbf{y}}, \textit{subject to} \end{aligned} \quad (2)$$

the constraints $y_i \in \left\{1, -\frac{\# y_i > 0}{\# y_i < 0}\right\}$. *Note that* $y_i^\star = 1$ *indicates* $v_i \in \mathcal{V}_1^\star$, *and* $y_i^\star \neq 1$, *otherwise.*

Proof: We define an indicator variable $\mathbf{f} \in \mathbb{R}^N$: $f_i = 1$ indicates the vertex v_i belongs to $\mathcal{V}_1$ and -1, otherwise. Then with the function $b\,(\mathbf{x}) = \frac{\# x_i > 0}{\# x_i < 0}$ and $c\,(\mathbf{x}) = \mathrm{MLoss}\,(\mathbf{x}/2)$, we have

$$\begin{aligned} \mathrm{NMcut}\,(\{\mathcal{V}_1, \mathcal{V}_2\}) = &\ \mathrm{Mcut}\,(\mathcal{V}_1, \mathcal{V}_2) \left(\tfrac{1}{\#\mathcal{V}_1} + \tfrac{1}{\#\mathcal{V}_2} \right) \\ = &\ [\alpha \min\{\mathrm{cut}\,(\mathcal{V}_1, \mathcal{V}_2), \mathrm{cut}\,(\mathcal{V}_2, \mathcal{V}_1)\} \\ &+ (1 - \alpha)\,(\mathrm{cut}\,(\mathcal{V}_1, \mathcal{V}_2) + \mathrm{cut}\,(\mathcal{V}_2, \mathcal{V}_1))] \\ &\times \left(\tfrac{1}{\#\mathcal{V}_1} + \tfrac{1}{\#\mathcal{V}_2} \right) \\ = &\ \min_{\mathbf{x}\in\{\mathbf{f},-\mathbf{f}\}} \tfrac{\#(x_i>0)+\#(x_i>0)}{(\# x_i>0)(\# x_i<0)}\, c\,(\mathbf{x}) \\ = &\ \min_{\mathbf{x}\in\{\mathbf{f},-\mathbf{f}\}} \tfrac{(\#(x_i>0)+\#(x_i<0))^2 c(\mathbf{x})}{N(\# x_i>0)(\# x_i<0)} \\ = &\ \min_{\mathbf{x}\in\{\mathbf{f},-\mathbf{f}\}} \tfrac{(\#(x_i>0)+\#(x_i<0))^2 c(\mathbf{x})}{\#(x_i>0)(\#(x_i<0))^2+(\#(x_i>0))^2\#(x_i<0)} \\ = &\ \min_{\mathbf{x}\in\{\mathbf{f},-\mathbf{f}\}} \tfrac{1+2b(\mathbf{x})+b(\mathbf{x})^2}{\#(x_i>0)+b(\mathbf{x})^2\#(x_i<0)}\, c\,(\mathbf{x}) \\ = &\ \min_{\mathbf{x}\in\{\mathbf{f},-\mathbf{f}\}} \tfrac{(2+2b(\mathbf{x}))^2 c(\mathbf{x})}{4(\# x_i>0)+4b(\mathbf{x})^2(\# x_i<0)} \\ = &\ \min_{\mathbf{x}\in\{\mathbf{f},-\mathbf{f}\}} \tfrac{\mathrm{MLoss}((1+\mathbf{x})-b(\mathbf{x})(1-\mathbf{x}))}{4(\# x_i>0)+4b(\mathbf{x})^2(\# x_i<0)}. \end{aligned}$$

The sixth equation holds by dividing $\#\,(x_i < 0)^2$ up and down. Denoting $\mathbf{y} = (1 + \mathbf{x}) - b\,(\mathbf{x})\,(1 - \mathbf{x})$, we have

$$\begin{aligned} &\min_{\{\mathcal{V}_1,\mathcal{V}_2\}} \mathrm{NMcut}\,(\{\mathcal{V}_1, \mathcal{V}_2\}) \Longleftrightarrow \\ \min_{\mathbf{f}} &\min_{\mathbf{y}\in\{[(1\pm\mathbf{f})-b(\pm\mathbf{f})(1\mp\mathbf{f})]\}} \frac{\mathrm{MLoss}(\mathbf{y})}{\mathbf{y}^T\mathbf{y}} \Longleftrightarrow \\ &\min_{y_i\in\left\{1,-\frac{\#(y_i>0)}{\#(y_i<0)}\right\}} \frac{\mathrm{MLoss}(\mathbf{y})}{\mathbf{y}^T\mathbf{y}}. \end{aligned}$$

■

B. Iterative and Divisive Algorithms

Denoting the objective function of the right-hand side problem of (2) as $g\,(\mathbf{y})$, we relax the original problem into $\mathbf{y} \in \mathbb{R}^N$ and keep the constraint $\mathbf{y}^T\mathbf{1} = 0$ (as implied by Lemma 2). Since $g\,(\mathbf{y})$ is positively scale invariant, we can add a constraint $\mathbf{y}^T\mathbf{y} = 1$ to the problem without loss of optimality, leading to the problem

$$\begin{aligned} &\underset{\mathbf{y}\in\mathbb{R}^N}{\text{minimize}} && \mathrm{MLoss}\,(\mathbf{y}) \\ &\text{subject to} && \mathbf{y}^T\mathbf{y} = 1, \mathbf{y}^T\mathbf{1} = 0. \end{aligned} \quad (3)$$

We solve (3) based on MM. At each iteration, a majorized function $\check{f}\left(\mathbf{y}|\mathbf{y}^{(t)}\right)$ around $\mathbf{y}^{(t)}$ is built for $f\left(\mathbf{y}^{(t)}\right) = \mathrm{MLoss}\,(\mathbf{y})$ and the corresponding majorized problem, with the original constraints, needs to be solved. $\check{f}\left(\mathbf{y}|\mathbf{y}^{(t)}\right)$ has to satisfy: $\check{f}\left(\mathbf{y}^{(t)}|\mathbf{y}^{(t)}\right) = f\left(\mathbf{y}^{(t)}\right)$, $\check{f}\left(\mathbf{y}|\mathbf{y}^{(t)}\right) \geq f\,(\mathbf{y})$, and $\nabla_{\mathbf{y}}\check{f}\left(\mathbf{y}^{(t)}|\mathbf{y}^{(t)}\right) = \nabla_{\mathbf{y}} f\left(\mathbf{y}^{(t)}\right)$ [2]. After this, the solution

$\mathbf{y}^{(t+1)}$ to the majorized problem is used to update the problem. This process is iterated until convergence. To construct $\check{f}\left(\mathbf{y}|\mathbf{y}^{(t)}\right)$, we have the following lemma.

Lemma 3. *The following problem is a majorization of* (3)*:*

$$\begin{array}{ll} \underset{\mathbf{y}\in\mathbb{R}^N}{\text{maximize}} & \left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)^T\mathbf{y} \\ \text{subject to} & \mathbf{y}^T\mathbf{y} = 1, \mathbf{y}^T\mathbf{1} = 0, \end{array} \tag{4}$$

where $\lambda\mathbf{I} \succeq \mathbf{P}\mathbf{P}^T$*,* $\mathbf{P} \in \mathbb{R}^{N\times M}$ *is the weighted incidence matrix with binary elements being* $-\sqrt{W_{ij}}$*, and* $\mathbf{s}^{(t)} \in \mathbb{R}^M$ *is* $\mathbf{s}^{(t)} = \alpha\left[\mathbf{P}^T\mathbf{y}^{(t)}\right]_-$.

Proof: The original problem is the problem (3). The original objective function is $\text{MLoss}\left(\mathbf{y}\right) = \sum_{i,j=1}^N W_{ij}\left(\left[y_i - y_j\right]_+^2 + \beta\left[y_i - y_j\right]_-^2\right) = \overset{[}{i}, j = 1]N\sum W_{ij}\left[y_i - y_j\right]_+^2 + (1-\alpha)\overset{[}{i}, j = 1]N\sum W_{ij}\left[y_i - y_j\right]_-^2$. Considering the basic function $[x]_+^2$, for $x^{(t)} > 0$, the majorized function is just x^2, for $x^{(t)} < 0$, the majorized function can be $\left(x - x^{(t)}\right)^2$. And for $(1-\alpha)[x]_-^2$, when $x^{(t)} < 0$, we have the majorized function as $\left(x - \alpha x^{(t)}\right)^2 + const$. Thus, for $W_{ij}\left[y_i - y_j\right]_+^2$, the majorized function around $\mathbf{y}^{(t)}$ is

$$\left\|\mathbf{y}^T\mathbf{P} - \mathbf{s}^{(t)^T}\right\|_2^2, \tag{5}$$

where $\mathbf{P}$ and $\mathbf{s}^{(t)}$ are defined as before. Furthermore, for the majorized function (5), we do majorization for the second time. Based on Lemma 4 given in [3], we have the majorized function,

$$\begin{aligned} & \left(\mathbf{y}^T\mathbf{P} - \mathbf{s}^{(t)^T}\right)^T\left(\mathbf{y}^T\mathbf{P} - \mathbf{s}^{(t)^T}\right) \\ = \; & \text{Tr}\left(\mathbf{P}^T\mathbf{y}\mathbf{y}^T\mathbf{P}\right) - 2\mathbf{s}^{(t)^T}\mathbf{P}^T\mathbf{y} + const \\ = \; & \mathbf{y}^T\left(\mathbf{P}\mathbf{P}^T\right)\mathbf{y} - 2\mathbf{s}^{(t)^T}\mathbf{P}^T\mathbf{y} + const \\ \le \; & \left[-2\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)}\right)^T\mathbf{y} + \lambda\left\|\mathbf{y}\right\|_2^2 + const\right] - \\ & 2\text{Tr}\left(\mathbf{s}^{(t)^T}\mathbf{P}^T\mathbf{y}\right) + const \\ = \; & -2\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)^T\mathbf{y} + const, \end{aligned}$$

where $\lambda\mathbf{I} \geqslant \mathbf{P}\mathbf{P}^T$. Thus finally, we got the final majorized function $-2\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)^T\mathbf{y}$, leading to the problem,

$$\begin{array}{ll} \underset{\mathbf{y}\in\mathbb{R}^N}{\text{minimize}} & -2\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)^T\mathbf{y} \\ \text{subject to} & \mathbf{y}^T\mathbf{y} = 1, \mathbf{y}^T\mathbf{1} = 0 \\ \Longleftrightarrow & \\ \underset{\mathbf{y}\in\mathbb{R}^N}{\text{maximize}} & \left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)^T\mathbf{y} \\ \text{subject to} & \mathbf{y}^T\mathbf{y} = 1, \mathbf{y}^T\mathbf{1} = 0, \end{array} \tag{6}$$

■

Next, we only need to obtain the solution to (4), which is much easier than the original problem 3. For it, we give a supporting lemma.

Algorithm 1 IGC Algorithm

```
Require: $N \in \mathbb{Z}_+$, $0 \le \alpha \le 1$, $\mathbf{W} \in \mathbb{R}^{N\times N}$, $\mathbf{T} \in \mathbb{R}^{N\times(N-1)}$.
Initialize $\mathbf{y}^{(0)} \in \mathbb{R}^N$.
Compute $\mathbf{P}$, $\lambda$ satisfying $\lambda\mathbf{I} \ge \mathbf{P}\mathbf{P}^T$.
repeat
    Update $\mathbf{s}^{(t)} \leftarrow \alpha\left[\mathbf{P}^T\mathbf{y}^{(t)}\right]_-$.
    Update $\mathbf{y}^{(t+1)} \leftarrow \frac{\mathbf{T}\mathbf{T}^T\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)}{\left\|\mathbf{T}^T\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)\right\|_2}$.
until Convergence
Find $p \in \mathbb{R}$ as the splitting point for $\mathbf{y}^{(t)}$.
return Vertex clusters $\{\mathcal{V}_1, \mathcal{V}_2\}$ seperated by $p$.
```

Lemma 4. *The closed-form solution to* (4) *is* $\mathbf{y}^\star = \frac{\mathbf{T}\mathbf{T}^T\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)}{\left\|\mathbf{T}^T\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)\right\|_2}$*, where* $\mathbf{T} \in \mathbb{R}^{N\times(N-1)}$ *is any constant matrix satisfying* $\mathbf{T}^T\mathbf{T} = \mathbf{I}$ *and* $\mathbf{1}^T\mathbf{T} = 0$.

Proof: Firstly, we define a new variable $\mathbf{u} \in \mathbb{R}^{N-1}$ meeting $\mathbf{T}\mathbf{u} = \mathbf{y}, \mathbf{u}^T\mathbf{u} = 1$,where $\mathbf{T} \in \mathbb{R}^{N\times(N-1)}$ satisfies: $\mathbf{T}^T\mathbf{T} = \mathbf{I}$ and $\mathbf{1}^T\mathbf{T} = 0$. Then, (4) becomes

$$\begin{array}{ll} \underset{\mathbf{u}\in\mathbb{R}^{N-1}}{\text{maximize}} & \left[\mathbf{T}^T\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)\right]^T\mathbf{u} \\ \text{subject to} & \mathbf{u}^T\mathbf{u} = 1. \end{array} \tag{7}$$

From Cauchy-Schwarz's inequality, the solution to (4) is $\mathbf{y}^\star = \mathbf{T}\mathbf{u}^\star = \frac{\mathbf{T}\mathbf{T}^T\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)}{\left\|\mathbf{T}^T\left(\lambda\mathbf{y}^{(t)} - \mathbf{P}\mathbf{P}^T\mathbf{y}^{(t)} + \mathbf{P}\mathbf{s}^{(t)}\right)\right\|_2}$. ■

Ideally, the optimal solution $\mathbf{y}^\star$ holds discrete values as given in Lemma 2. After embedding, this may not be the case, but the continuous values of $\mathbf{y}^\star$ can indicate the clustering result, on which we need to choose a splitting point to partition it into two parts. One can take 0 or the median value as the splitting point, or search for the appropriate splitting point so that NMcut is small. We take the former strategy in our work. The empirical experiments show that the latter two methods can improve the performance a little but come with more computation costs. This algorithm is named IGC (short for iterative algorithm for graph clustering).

Now, we consider the clustering with $K \ge 2$ clusters. Our basic idea is to obtain the result by recursively applying IGC on the graph from up to down until we have K partitions. Then the remaining problem is how to choose the cluster to be clustered further at each iteration. The strategy we use is to choose the cluster of the largest cardinality, since large clusters generally tend to contain smaller ones. Finally, we name this divisive algorithm for multiple clusters RGC (short for the recursive algorithm for graph clustering).

In IGC, the larger values of $\mathbf{y}^\star$ indicate that the corresponding cluster is the target cluster (the corresponding cluster tend to hold more out-paths), as implied by the proof of Lemma 2. In RGC, the tendencies between clusters are inferred by the results. At iterations, the addresses of temporary clusters can be stored in a linked list. In detail, at iteration t, the selected cluster $\mathcal{V}^{(t-1)}$ is clustered into two ones by IGC. Then, based on the tendency inferred by $\mathbf{y}^\star$, two new clusters are taken in order and replace $\mathcal{V}^{(t-1)}$ in the linked list. Thus, finally, on K clusters, the pair-wise tendencies are indicated by the positions of clusters in the linked list.

II. Convergence and Complexity Analysis

IGC is based on MM scheme. According to [2], the sequences of objective values evaluated at $\left\{\mathbf{y}^{(t)}\right\}$ generated by IGC are non-increasing and are bounded below by 0. Thus, the sequence of objective values is guaranteed to converge to finite values. The convergence property of the sequences $\left\{\mathbf{y}^{(t)}\right\}$ is proven, of which the supporting theorem is given as following.

Theorem 5. *Let $\left\{\mathbf{y}^{(t)}\right\}$ be the sequence generated by the IGC algorithm. Then every limit point of the sequence $\left\{\mathbf{y}^{(t)}\right\}$ is a stationary point of the problem shown in* (3).

Proof: Denote the objective function of the problem (3) as $f_1\left(\mathbf{y}\right)$ and the corresponding majorized function shown in (4) as $-\check{f}_1\left(\mathbf{y}|\mathbf{y}^{(t)}\right)$. (Here the $-$ operator before $\check{f}_1\left(\mathbf{y}|\mathbf{y}^{(t)}\right)$ is for the simplification of derivation, since for $f_1\left(\mathbf{y}\right)$, there is minimize, and for $\check{f}_1\left(\mathbf{y}|\mathbf{y}^{(t)}\right)$, there is maximize.) Denote the constraint set of (4) as $\mathcal{C}_1$. Since the sequence $\left\{\mathbf{y}^{(t)}\right\}$ is bounded, we know that it has at least one limit point. Consider a limit point $\mathbf{y}^{(\infty)}$ and a subsequence $\left\{\mathbf{y}^{(t_j)}\right\}$ that converges to $\mathbf{y}^{(\infty)}$, we have

$$\begin{array}{rl}\check{f}_1\left(\mathbf{y}^{(t_{j+1})}|\mathbf{y}^{(t_{j+1})}\right) = & f_1\left(\mathbf{y}^{(t_{j+1})}\right) \leq f_1\left(\mathbf{y}^{(t_j+1)}\right) \\ \leq & \check{f}_1\left(\mathbf{y}^{(t_j+1)}|\mathbf{y}^{(t_j)}\right) \leq \check{f}_1\left(\mathbf{y}|\mathbf{y}^{(t_j)}\right), \\ & \forall \mathbf{y} \in \mathcal{C}_1.\end{array}$$

Letting $j \to +\infty$, we obtain

$$\check{f}_1\left(\mathbf{y}^{(\infty)}|\mathbf{y}^{(\infty)}\right) \leq \check{f}_1\left(\mathbf{y}|\mathbf{y}^{(\infty)}\right), \forall \mathbf{y} \in \mathcal{C}_1,$$

i.e. $\mathbf{y}^{(\infty)}$ is a global minimizer of $\check{f}_1\left(\mathbf{y}|\mathbf{y}^{(\infty)}\right)$ over $\mathcal{C}_1$. Then as a necessary condition, we have

$$\nabla\check{f}_1\left(\mathbf{y}^{(\infty)}|\mathbf{y}^{(\infty)}\right)^T \mathbf{z} \geq 0, \forall \mathbf{z} \in T_{\mathcal{C}_1}\left(\mathbf{y}^{\star}\right),$$

Furthermore, from the principles of the majorized function in the MM scheme [2], we have

$$\nabla f_1\left(\mathbf{y}^{(\infty)}\right) = \nabla\check{f}_1\left(\mathbf{y}^{(\infty)}|\mathbf{y}^{(\infty)}\right).$$

Thus we have

$$\nabla f_1\left(\mathbf{y}^{(\infty)}\right)^T \mathbf{z} \geq 0, \forall \mathbf{z} \in T_{\mathcal{C}_1}\left(\mathbf{y}^{\star}\right),$$

implying that $\mathbf{y}^{(\infty)}$ is a stationary point of the problem in the right side of (3). The proof is done. ■

Algorithm 2 RGC Algorithm

```
Require: $N, K \in \mathbb{Z}_+$, $K < N$, $\mathbf{W} \in \mathbb{R}^{N\times N}$, $0 \leq \alpha \leq 1$.
Initialize the set of clusters $\mathcal{C}^{(0)} = \varnothing$, $\mathbf{W}^{(0)} = \mathbf{W}$.
for $t = 1 \to K$ do
    Apply IGC algorithm on $\mathbf{W}^{(t-1)}$ to get $\left\{\mathcal{V}_1^{(t)}, \mathcal{V}_2^{(t)}\right\}$.
    Update $\mathcal{C}^{(t)} \leftarrow \mathcal{C}^{(t-1)} \cup \left\{\mathcal{V}_1^{(t)}, \mathcal{V}_2^{(t)}\right\}$.
    Update $\mathcal{V}^{(t)} \leftarrow \arg\max_{\mathcal{V}\in\mathcal{S}^{(t)}} \mathrm{vol}\left(\mathcal{V}\right)$.
    Update $\mathcal{C}^{(t)} \leftarrow \mathcal{C}^{(t)} - \mathcal{V}^{(t)}$.
    Build $\mathbf{W}^{(t)}$ from $\mathcal{V}^{(t)}$.
end for
return $\mathcal{S}^{(t)}$.
```

RGC recursively invokes IGC within a finite number of times, so its convergence holds if IGC converges.

At each iteration of IGC, the overall time complexity is $\mathcal{O}\left(M\right)$, since $\mathbf{P}$ is sparse. Both IGC and RGC can be greatly accelerated by parallel computing such as on GPUs because of massive matrix multiplications. The In terms of space complexity, IGC needs to store $\mathbf{W}$, $\mathbf{P}$, $\mathbf{T}$ and compute $\mathbf{y}^{(t+1)}$ and $\mathbf{s}^{(t)}$ in each update. Thus, the space of $\mathcal{O}\left(M\right)$ is needed. RGC needs to store the variables of IGC and $\mathcal{S}$, of which the space complexity of it is also $\mathcal{O}\left(M\right)$.

III. Experiments

In this section, we present the results of the synthetic and empirical experiments. The algorithms included are: IGC, RGC, and the benchmark algorithms: Dir-rem, Deg-dis, NGA-LP, Ran-wal, Non-lin, Mar-sta, and Mut-lin. For the algorithms designed for unweighted edges, we perform linear grid searches between 0 and the largest edge weight with 20 intervals to find the optimal threshold, below which the edges are ignored. By default, IGC terminates when $\left\|\mathbf{y}^{(t+1)} - \mathbf{y}^{(t)}\right\|_{fro} < 1 \times 10^{-6}$. All the experiments are done on a computer with an i7-6700 3.4 GHz CPU and 8 GB RAM. The results are evaluated by NMI, of which the value is between 0 and 1, and a larger NMI means better performance [4]. NMI is defined as following:

$$\mathrm{NMI}\left(\Omega, \mathbb{C}\right) = \frac{I\left(\Omega; \mathbb{C}\right)}{\left[H\left(\Omega\right) + H\left(\mathbb{C}\right)\right]/2},$$

where $\Omega = \left\{\omega_1, \ldots, \omega_K\right\}$ and $\mathbb{C} = \left\{c_1, \ldots, c_J\right\}$ are two sets consisting of the clustering results and the ground-truth clusters respectively. Mutual information $I\left(\Omega, \mathbb{C}\right)$ is defined as

$$I\left(\Omega; \mathbb{C}\right) = \sum_k \sum_j P\left(\omega_k \cap c_j\right) \log \frac{P\left(\omega_k \cap c_j\right)}{P\left(\omega_k\right) P\left(c_j\right)},$$

with $P\left(\omega_k\right)$, $P\left(c_j\right)$, $P\left(\omega_k \cap c_j\right)$ being the probability of a vertice belonging to ω_k, c_j, and the intersection of ω_k and c_j. And the entropy $H\left(\Omega\right)$ is defined as

$$H\left(\Omega\right) = -\sum_k P\left(\omega_k\right) \log P\left(\omega_k\right).$$

It is obvious that a higher $I\left(\Omega; \mathbb{C}\right)$ infers a better clustering result. But considering that $K = J$ may not hold, and a larger K tends to make a higher $I\left(\Omega; \mathbb{C}\right)$, the normalized component $\left[H\left(\Omega\right) + H\left(\mathbb{C}\right)\right]/2$ is applied, which also tends to increase with K. It is proven that $\left[H\left(\Omega\right) + H\left(\mathbb{C}\right)\right]/2$ is a tight upper bound for $I\left(\Omega; \mathbb{C}\right)$, so $\mathrm{NMI}\left(\Omega, \mathbb{C}\right)$ is a real value metric between 0 and 1 for evaluting the goodness of the clustering result $\Omega = \left\{\omega_1, \ldots, \omega_K\right\}$ given the ground-truth one $\mathbb{C} = \left\{c_1, \ldots, c_J\right\}$.

References

[1] I. S. Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 269–274.

[2] Y. Sun, P. Babu, and D. P. Palomar, “Majorization-minimization algorithms in signal processing, communications, and machine learning,” *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[3] Z. Wang, P. Babu, and D. P. Palomar, “Design of PAR-Constrained Sequences for MIMO Channel Estimation via Majorization-Minimization.” *IEEE Trans. Signal Processing*, vol. 64, no. 23, pp. 6132–6144, 2016.

[4] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 39.