Constrained fractional set programs and their application in local clustering and community detection

Thomas Bühler TB@CS.UNI-SAARLAND.DE
Saarland University, Saarbrücken, Germany

Syama Sundar Rangapuram SRANGAPU@MPI-INF.MPG.DE
Max Planck Institute for Informatics & Saarland University, Saarbrücken, Germany

Simon Setzer SETZER@MIA.UNI-SAARLAND.DE
Matthias Hein HEIN@CS.UNI-SAARLAND.DE
Saarland University, Saarbrücken, Germany

Abstract

The (constrained) minimization of a ratio of set functions is a problem frequently occurring in clustering and community detection. As these optimization problems are typically NP-hard, one uses convex or spectral relaxations in practice. While these relaxations can be solved globally optimally, they are often too loose and thus lead to results far away from the optimum. In this paper we show that *every* constrained minimization problem of a ratio of non-negative set functions allows a *tight* relaxation into an unconstrained continuous optimization problem. This result leads to a flexible framework for solving constrained problems in network analysis. While a globally optimal solution for the resulting non-convex problem cannot be guaranteed, we outperform the loose convex or spectral relaxations by a large margin on constrained local clustering problems.

1. Introduction

Graph-based data appear in manifold ways in learning problems - either the data have already graph structure as in the case of social networks and biological networks or a similarity graph is constructed using a similarity measure based on features of the data. Several graph-based problems in clustering and community detection can be modelled as the optimization of a ratio of set functions (referred to here as fractional set program). Prominent examples are the normalized cut problem, from which the popular spectral clustering method is derived (Shi & Malik, 2000), and the maximum density subgraph problem, which has applications in community detection (Fortunato, 2010) and bioinformatics (Saha et al., 2010).

It turns out that in practice often additional background or domain knowledge about the learning problem is available. Such prior knowledge can then be incorporated as constraints into the optimization problem. In the case of clustering, Wagstaff et al. (2001) are the first to show how prior information given in the form of must-link and cannot-link constraints between vertices can be integrated into the k-means algorithm. Recently, Rangapuram & Hein (2012) proposed a generalization of the normalized cut problem that can handle must-link and cannot-link constraints. In the recent work of Mahoney et al. (2012), locality constraints in the form of a seed set and volume constraint have been integrated into the normalized cut formulation. Furthermore, Khuller & Saha (2009) and Saha et al. (2010) considered size and distance constraints for the maximum density subgraph problem.

Since the above-mentioned combinatorial problems are NP-hard, the standard approach is to consider convex or spectral relaxations which can be solved globally optimally in polynomial time. Due to its practical efficiency the spectral relaxation is very popular in machine learning, e.g. spectral clustering (Hagen & Kahng, 1991; Shi & Malik, 2000). However, it is often quite loose and thus leads to a solution far away from the optimal one of the original problem. Moreover, spectral-type relaxations (Mahoney et al., 2012)

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

fail to guarantee that the constraints which encode the prior knowledge are satisfied.

In another line of work (Hein & Bühler, 2010; Szlam & Bresson, 2010; Hein & Setzer, 2011; Bresson et al., 2012), it has been shown that *tight continuous relaxations* exist for all balanced graph cut problems and the normalized cut subject to must-link and cannot-link constraints (Rangapuram & Hein, 2012). A tight relaxation means that the continuous and the combinatorial optimization problem are equivalent in the sense that the optimal values agree and the optimal solution of the combinatorial problem can be obtained from the continuous solution. While the resulting algorithms provide no guarantee to yield the globally optimal solution, the standard loose relaxations are outperformed by a large margin in practice.

In this paper we show that *any* constrained minimization problem of a ratio of non-negative set functions allows a tight relaxation into a continuous optimization problem. This result together with our efficient minimization techniques enables the easy integration of prior information in form of constraints into many problems in graph-based clustering and community detection. While the general framework introduced in this paper is applicable to all problems discussed so far, we will focus on two particular applications: local clustering by constrained balanced graph cuts, and community detection via constrained densest subgraph problems. Compared to previous work, the algorithms developed in this paper are the first to guarantee that all given constraints are fulfilled by the obtained solution. Note that in principle our method could also be applied to a setting with soft or noisy constraints, however we will focus here on the case of hard constraints. In the experimental section we will show the superior performance compared to state of the art methods (Andersen & Lang, 2006; Mahoney et al., 2012).

All proofs can be found in the supplementary material.

2. Fractional set programs in clustering and community detection

In the following, $G = (V, W)$ denotes an undirected, weighted graph with a non-negative, symmetric weight matrix $W \in \mathbb{R}^{n \times n}$, where $n = |V|$. Moreover, by assigning a non-negative weight g_i to each vertex i, we can define the general volume of a subset $A \subset V$ as $\mathrm{vol}_g(A) = \sum_{i \in A} g_i$. As special cases, we obtain for $g_i = 1$ the cardinality $|A|$ and for g_i equal to the degree $d_i = \sum_{j \in V} w_{ij}$ the classical volume $\mathrm{vol}(A) = \mathrm{vol}_d(A)$. Furthermore, $\overline{A} = V \backslash A$ denotes the complement of A.

The balanced graph cut problem is a well-known problem in computer science with applications ranging from parallel computing to image segmentation (Pothen et al., 1990; Shi & Malik, 2000). A very popular balanced graph cut criterion is the normalized cut[1],

$$\mathrm{NCut}(C, \overline{C}) = \frac{\mathrm{cut}(C, \overline{C})}{\mathrm{vol}_d(C)\,\mathrm{vol}_d(\overline{C})}, \quad \text{for } C \subset V,$$

where $\mathrm{cut}(C, \overline{C}) := \sum_{i \in C, j \in \overline{C}} w_{ij}$. The spectral relaxation of the normalized cut leads to the popular spectral clustering method (von Luxburg, 2007). A related criterion is the normalized Cheeger cut,

$$\mathrm{NCC}(C, \overline{C}) = \frac{\mathrm{cut}(C, \overline{C})}{\min\{\mathrm{vol}_d(C), \mathrm{vol}_d(\overline{C})\}}, \quad \text{for } C \subset V.$$

More general balanced graph cuts were studied by Hein & Setzer (2011). In practice, often additional information about the desired solution is available which can be incorporated into the problem via constraints. This motivates us to consider a more general class of problems where one optimizes a ratio of set functions[2] subject to constraints. In the following, we discuss two examples of constrained problems in network analysis.

Constrained balanced graph cuts for local clustering. Recently, there has been a strong interest in balanced graph cut methods for local clustering. Starting with the work of Spielman & Teng (2004), initially, the goal was to develop an *algorithm* that finds a subset near a given seed vertex with *small* normalized cut or normalized Cheeger cut value with running time linear in the size of the obtained cluster. The proposed algorithm and subsequent work (Andersen et al., 2006; Chung, 2009) use random walks to explore the graph locally, without considering the whole graph. Algorithms of this type have been applied for community detection in networks (Andersen & Lang, 2006).

In contrast, Mahoney et al. (2012) give up the runtime requirement and formulate the task as an explicit optimization problem, where one aims at finding the *optimal* normalized cut subject to a seed constraint and an upper bound on the volume of the set containing the seed set. Again, the idea is to find a local cluster around a given seed set. Motivated by the standard spectral relaxation of the normalized cut problem, they derive a spectral-type relaxation which is biased towards solutions fulfilling the seed constraint. Their method has been successfully applied in semi-supervised image segmentation (Maji et al., 2011) and

[1] This is up to a constant factor the same as the usual definition, $\mathrm{NCut}(C, \overline{C}) = \mathrm{cut}(C, \overline{C})\big(\frac{1}{\mathrm{vol}_d(C)} + \frac{1}{\mathrm{vol}_d(\overline{C})}\big)$.

[2] A set function $\widehat{S}$ on a set V is a function $\widehat{S} : 2^V \to \mathbb{R}$.

for community detection around a given query set (Mahoney et al., 2012). However, while they provide an approximation guarantee for their relaxation, they cannot guarantee that the returned solution satisfies seed and volume constraints.

In this paper we consider an extended version of the problem of Mahoney et al. (2012). Let J denote the set of seed vertices, $\widehat{S}$ a symmetric balancing function (e.g. $\widehat{S}(C) = \mathrm{vol}_d(C)\,\mathrm{vol}_d(\overline{C})$ for the normalized cut) and let $\mathrm{vol}_g(C)$ be the general volume of set C, where $g \in \mathbb{R}^n_+$ are vertex weights. The general local clustering problem can then be formulated as

$$\min_{C \subset V} \frac{\mathrm{cut}(C, \overline{C})}{\widehat{S}(C)} \tag{1}$$
$$\text{subject to}: \ \mathrm{vol}_g(C) \leq k, \quad \text{and} \quad J \subset C.$$

The choice of the balancing function $\widehat{S}$ allows the user to influence the trade-off between getting a partition with small cut and a balanced partition. One could also combine this with must- and cannot-link constraints (see Rangapuram & Hein, 2012) or add even more complex constraints such as an upper bound on the diameter of C. However, in order to compare to the method of Mahoney et al. (2012), we restrict ourselves in this paper to the normalized cut with volume constraints, that is $\widehat{S}(C) = \mathrm{vol}_d(C)\,\mathrm{vol}_d(\overline{C})$ and $g = d$.

Constrained local community detection. A second related problem is constrained local community detection. In community detection it makes more sense to find a highly connected set instead of emphasizing the separation to the remaining part of the graph by minimizing the cut. Thus, we are searching for a set C which has high association, defined as $\mathrm{assoc}(C) = \sum_{i,j \in C} w_{ij}$. Dividing the association of C by its size yields the density of C. The subgraph of maximum density can be computed in polynomial time (Goldberg, 1984). However, the obtained communities in the unconstrained problem are typically either too large or too small, which calls for size constraints. Note that the introduction of such constraints makes the problem NP-hard (Khuller & Saha, 2009).

A general class of (local) community detection problems can thus be formulated as

$$\max_{C \subset V} \frac{\mathrm{assoc}(C)}{\mathrm{vol}_g(C)} \tag{2}$$
$$\text{subject to}: k_1 \leq \mathrm{vol}_h(C) \leq k_2, \quad \text{and} \quad J \subset C,$$

where $g, h \in \mathbb{R}^n_+$ are vertex weights. This formulation generalizes the above-mentioned density-based approaches by replacing the denominator by a general volume function vol_g. One can use the vertex weights g to bias the obtained community towards one with desired properties by assigning small weights to vertices which one would prefer to occur in the solution and larger weights to ones which are less preferred.

The problem (2) with only lower bound constraints has been considered in team selection (Gajewar & Das Sarma, 2012) and bioinformatics (Saha et al., 2010) where constant factor approximation algorithms were developed. However, in the case of equality and upper bound constraints the problem is very hard even when using only cardinality constraints (i.e., $h_i = 1$), and it has been shown that there is no polynomial time approximation scheme in these cases (Khot, 2006; Khuller & Saha, 2009). Our method can handle such hard upper bound and equality constraints. In the experiments we show results for a community detection problem with a specified query set J and an upper bound on the size for a co-author network.

Note that for the choice of $\mathrm{vol}_g(C) = \mathrm{vol}_d(C)$ in (2) and $\widehat{S}(C) = \mathrm{vol}_d(C)$ in (1), the problem (2) is equivalent to (1) if we choose the same constraints.

Contributions of this paper. We show that **all** constrained non-negative fractional set programs have an equivalent tight continuous relaxation. This general result enables the integration of prior information in form of constraints into clustering and community detection problems. In particular, it allows us to derive efficient algorithms for problems (1) and (2). Our algorithms consistently outperform competing methods (Andersen & Lang, 2006; Mahoney et al., 2012). Moreover, we are not aware of any other methods for the above problems which can guarantee that the solution always satisfies volume **and** seed constraints.

Although the tight relaxation results in Hein & Setzer (2011) and Rangapuram & Hein (2012) encompass a large class of problems, they are not applicable to the problems considered in this paper because of the following limitations: First, tight relaxations were shown by Hein & Setzer (2011) only for a ratio of *symmetric* non-negative set functions, where the numerator is restricted to be *submodular*. We extend the results to *arbitrary* ratios of non-negative set functions without any restrictions concerning symmetry or submodularity. Second, only *equality* constraints for *non-negative* set functions restricted to be either *submodular or supermodular* could be handled by Rangapuram & Hein (2012). We generalize this to *inequality* constraints[3] without any restrictions on the constraint set functions in order to handle the constraints in (1) and (2).

[3] Note that $\widehat{M}(C) = k$ is equivalent to $k \leq \widehat{M}(C) \leq k$.

3. Tight relaxations of fractional set programs with constraints

The problems discussed in the last section can be written in the following general form:

$$\min_{C\subset V} \frac{\widehat{R}(C)}{\widehat{S}(C)} =: \widehat{Q}(C) \tag{3}$$
$$\text{subject to}: \widehat{M_i}(C) \leq k_i, \quad i=1,\dots,K$$

where $\widehat{R}, \widehat{S}, \widehat{M_i} : 2^V \to \mathbb{R}$ are set functions on a set $V = \{1,\dots,n\}$. We assume here that $\widehat{R}, \widehat{S}$ are non-negative and that $\widehat{R}(\emptyset) = \widehat{S}(\emptyset) = 0$. No assumptions are made on the set functions $\widehat{M_i}$, in particular they are not required to be non-negative. Thus also lower bound constraints can be written in the above form. Moreover, the formulation in (3) also encompasses the subset constraint $J \subset C$ in (1) and (2) as it can be written as equality constraint $|J| - |J \cap C| = 0$. Alternatively, we will discuss a direct integration of the subset constraint into the objective in Section 5.

The connection between the set-valued and the continuous space is achieved via thresholding. Let $f \in \mathbb{R}^n$, and we assume wlog that f is ordered in ascending order $f_1 \leq f_2 \leq \dots \leq f_n$. One defines the sets

$$C_i := \{j \in V | f_j \geq f_i\}, \qquad i=1,\dots,n. \tag{4}$$

We frequently make use of this notation in the following. Furthermore, we use $\mathbf{1}_C \in \mathbb{R}^n$ to denote the indicator vector of the set C, i.e. the vector which is 1 at entry j if $j \in C$ and 0 otherwise. A key tool for the derivation of the results of this paper is the Lovasz extension as a way to extend a set function (seen as function on the hypercube) to a function on $\mathbb{R}^n$.

Definition 1 *Let $\widehat{R} : 2^V \to \mathbb{R}$ be a set function with $\widehat{R}(\emptyset) = 0$, and $f \in \mathbb{R}^n$ in ascending order $f_1 \leq f_2 \leq \dots \leq f_n$. The Lovasz extension $R : \mathbb{R}^n \to \mathbb{R}$ of $\widehat{R}$ is defined as $R(f) = \sum_{i=1}^{n-1} \widehat{R}(C_{i+1})\,(f_{i+1} - f_i) + \widehat{R}(V) f_1$.*

Note that $R(\mathbf{1}_C) = \widehat{R}(C)$ for all $C \subset V$, i.e. R is indeed an extension of $\widehat{R}$ from 2^V to $\mathbb{R}^n$. In the following, we always use the hat-symbol ($\widehat{\ }$) to denote set functions and omit it for the corresponding Lovasz extension. A particular important class of set functions are submodular set functions since their Lovasz extension is convex (Bach, 2011).

Definition 2 *A set function $\widehat{R} : 2^V \to \mathbb{R}$ is submodular if for all $A, B \subset V$, $\widehat{R}(A \cup B) + \widehat{R}(A \cap B) \leq \widehat{R}(A) + \widehat{R}(B)$. It is supermodular, if the converse inequality holds true, and modular if we have equality.*

Unconstrained fractional set programs. Using the property of the Lovasz extension that $R(\mathbf{1}_C) = \widehat{R}(C)$ for all $C \subset V$, one can directly observe that the following continuous fractional program is a relaxation of the unconstrained version of problem (3)

$$\inf_{f \in \mathbb{R}^n_+} \frac{R(f)}{S(f)}.$$

The following theorem shows that the relaxation is in fact tight, in the sense that the optimal values agree and the solution of the set-valued problem can be computed from the solution of the continuous problem.

Note that given a vector $f \in \mathbb{R}^n$ for the continuous problem, one can construct a set C' by computing

$$C' = \operatorname*{arg\,min}_{C_i, i=1,\dots,n} \frac{\widehat{R}(C_i)}{\widehat{S}(C_i)},$$

where the sets C_i are defined in (4). We refer to this process as *optimal thresholding.*

Theorem 1 *Let $\widehat{R}, \widehat{S} : 2^V \to \mathbb{R}$ be non-negative set functions and $R, S : \mathbb{R}^n \to \mathbb{R}$ their Lovasz extensions, respectively. Then, it holds that*

$$\inf_{C\subset V} \frac{\widehat{R}(C)}{\widehat{S}(C)} = \inf_{f \in \mathbb{R}^n_+} \frac{R(f)}{S(f)}\ .$$

Moreover, it holds for all $f \in \mathbb{R}^n_+$, $\frac{R(f)}{S(f)} \geq \min_{i=1,\dots,n} \frac{\widehat{R}(C_i)}{\widehat{S}(C_i)}$. Thus a minimizer of the set ratio can be found by optimal thresholding. Let furthermore $\widehat{R}(V) = \widehat{S}(V) = 0$, then all the above statements hold if one replaces $\mathbb{R}^n_+$ with $\mathbb{R}^n$.

Note that no assumptions except non-negativity are made on $\widehat{R}$ and $\widehat{S}$ - *every* non-negative fractional set program has a tight relaxation into a continuous fractional program. The efficient minimization of the continuous objective will be the topic of Section 4. A slightly technical generalization of Theorem 1 can be found in the supplementary material.

Constrained fractional set programs. To solve the constrained fractional set program (3) we make use of the concept of *exact penalization* (Di Pillo, 1994), where the main idea is to transform a given constrained optimization problem into an *equivalent* unconstrained one by adding a penalty term. We use the same idea for our constrained fractional set programs and define the penalty set function for a constraint $\widehat{M_i}(C) \leq k_i$ as

$$\widehat{T_i}(C) = \begin{cases} \max\left\{0, \widehat{M_i}(C) - k_i\right\}, & C \neq \emptyset, \\ 0, & C = \emptyset. \end{cases} \tag{5}$$

The function $\widehat{T}_i(C)$ is zero if C is feasible for the i-th constraint and otherwise increasing with increasing infeasibility. The special treatment of the empty set in the definition of $\widehat{T}_i$ is a technicality required for the Lovasz extension. Defining $\widehat{T}(C) := \sum_{i=1}^K \widehat{T}_i(C)$, we can now formulate a modified problem

$$\min_{C\subset V} \frac{\widehat{R}(C) + \gamma \sum_i^K \widehat{T}_i(C)}{\widehat{S}(C)} =: \widehat{Q}_\gamma(C). \qquad (6)$$

We will show that using a feasible set of (3) one can compute a γ such that (6) is equivalent to the original constrained problem. Once we have established the equivalence, we can then apply Theorem 1, noting that $\widehat{T}$ is a non-negative set function. This leads to the main result of this paper showing a tight relaxation of *all* problems of form (3) where $\widehat{R}, \widehat{S}$ are non-negative set functions. In the following, the constant θ quantifies a "minimum value" of $\widehat{T}_i$ on the infeasible sets:

$$\theta = \min_{i=1,\ldots,K} \big[\min_{\widehat{M}_i(C) > k_i} \widehat{M}_i(C) - k_i \big].$$

For example, if $\widehat{M}(C) = |C|$, then θ is equal to 1. If $\widehat{M}(C) = \mathrm{vol}_g(C)$ and all vertex weights g_i are rational numbers which are multiples of a fraction $\frac{1}{\rho}, \rho \in \mathbb{N}$, then $\theta \geq \frac{1}{\rho}$. Note that in practice, the constant θ and the parameter γ introduced in the following are never explicitly computed (see experimental section).

Theorem 2 *Let $\widehat{R}, \widehat{S} : 2^V \to \mathbb{R}$ be non-negative set functions and R, S their Lovasz extensions. Let $C_0 \subset V$ be feasible and $\widehat{S}(C_0) > 0$. Denote by T the Lovasz extension of $\widehat{T}$. Then, for $\gamma > \frac{\widehat{R}(C_0)}{\theta \widehat{S}(C_0)} \max_{C\subset V} \widehat{S}(C)$,*

$$\min_{\substack{\widehat{M}_i(C) \leq k_i,\\ i=1,\ldots,K}} \frac{\widehat{R}(C)}{\widehat{S}(C)} = \min_{f\in\mathbb{R}^n_+} \frac{R(f) + \gamma\, T(f)}{S(f)} := Q_\gamma(f)$$

Moreover, for any $f \in \mathbb{R}^n_+$ with $Q_\gamma(f) < \widehat{Q}_\gamma(C_0)$ for the given γ, we have $Q_\gamma(f) \geq \min_{i=1,\ldots,n} \widehat{Q}_\gamma(C_i)$, and the minimizing set on the right hand side is feasible.

Note that Theorem 2 implies that the set found by optimal thresholding of the solution of the continuous program is guaranteed to satisfy all constraints. We are not aware of any other method which can give the same guarantee for the problems (1) and (2).

4. Minimization of the tight continuous relaxation

The continuous optimization problems in Theorems 1 and 2 have the form

$$\min_{f\in\mathbb{R}^n_+} \frac{R(f)}{S(f)} := Q(f), \qquad (7)$$

where R and S are non-negative. The fact that they are the Lovasz extensions of set functions $\widehat{R}, \widehat{S}$ also implies that they are one-homogeneous, see Bach (2011). We now apply a slightly modified version of a result from Hein & Setzer (2011).

Proposition 1 *Every set function $\widehat{S}$ with $\widehat{S}(\emptyset) = 0$ can be written as $\widehat{S} = \widehat{S}_1 - \widehat{S}_2$, where S_1 and S_2 are submodular and $\widehat{S}_1(\emptyset) = \widehat{S}_2(\emptyset) = 0$. The Lovasz extension S can be written as difference of convex functions.*

The above result implies that (7) can be written as ratio of differences of convex functions (d.c.), i.e. $R = R_1 - R_2$ with R_1, R_2 convex, and similarly for S. As the proof of Proposition 1 is constructive, the explicit form of this decomposition can be calculated. We can now use a modification of the RatioDCA which has recently been proposed as an algorithm for minimizing a non-negative ratio of one-homogeneous d.c. functions (Hein & Setzer, 2011). This modification is necessary as the problems in Theorem 1 and 2 require optimization over the positive orthant. We report the modified version in order to make the paper self-contained.

RatioDCA Minimization of a non-negative ratio of one-homogeneous d.c functions over $\mathbb{R}^n_+$

```
Initialization: $f^0 \in \mathbb{R}^n_+$, $\lambda^0 = Q(f^0)$
repeat
    $f^{l+1} = \arg\min_{u\in\mathbb{R}^n_+,\, \|u\|_2 \leq 1} \big\{ R_1(u) - \langle u, r_2(f^l)\rangle$
                $+ \lambda^l \big( S_2(u) - \langle u, s_1(f^l)\rangle \big) \big\}$
    where  $r_2(f^l) \in \partial R_2(f^l)$, $s_1(f^l) \in \partial S_1(f^l)$
    $\lambda^{l+1} = Q(f^{l+1})$
until $\frac{|\lambda^{l+1} - \lambda^l|}{\lambda^l} < \epsilon$
```

It is shown in the supplement that properties such as the fact that the sequence $Q(f^l)$ is either strictly decreasing or the sequence terminates carry over from Hein & Setzer (2011). The norm constraint of the inner problem is necessary as otherwise the problem would be unbounded from below. However, the choice of the norm plays no role in the proof and any norm can be chosen. Moreover, in the special case where the one-homogeneous function R is convex and S is concave, the RatioDCA reduces to Dinkelbach's method from fractional programming (Dinkelbach, 1967) and therefore computes the global optimum. In the general case, convergence to the global optimum cannot be guaranteed. However, we can provide a *quality guarantee*: RatioDCA either improves a given feasible set or stops after one iteration.

Theorem 3 *Let A be a feasible set and $\gamma > \widehat{R}(A) \max_{C\subset V} \widehat{S}(C)/(\theta\, \widehat{S}(A))$. Let f^* denote the re-*

sult of RatioDCA after initializing with the vector $\mathbf{1}_A$, and let C_{f^*} denote the set found by optimal thresholding of f^*. Either RatioDCA terminates after one iteration, or C_{f^*} is feasible and $\frac{\widehat{R}(C_{f^*})}{\widehat{S}(C_{f^*})} < \frac{\widehat{R}(A)}{\widehat{S}(A)}$.

The above theorem implies that all constraints of the original constrained fractional set program are fulfilled by the set C_{f^*} returned by RatioDCA.

5. Tight relaxations of constrained maximum density and constrained balanced graph cut problems

The framework introduced in this paper allows us to derive tight relaxations of all problems discussed in Section 2. In the following, we will derive a tight relaxation of the local community detection problem

$$\max_{C\subset V} \frac{\mathrm{assoc}(C)}{\mathrm{vol}_g(C)} \qquad (8)$$
$$\text{subject to}: \mathrm{vol}_h(C) \leq k, \quad \text{and} \quad J \subset C.$$

For the constrained balanced graph cut problem, the tight relaxation can be found in a very similar way and is thus omitted here. Moreover, incorporating constraints of the form $\mathrm{vol}_h(C) \geq k$ in both problems is similar and outlined in the supplementary material.

First, we integrate the volume constraint via a penalty term, see (6), which yields the equivalent problem

$$\min_{\substack{C\subset V\\ \text{s.t.} J\subset C}} \frac{\mathrm{vol}_g(C) + \gamma \widehat{T}_k(C)}{\mathrm{assoc}(C)}, \qquad (9)$$

where $\widehat{T}_k$ is given as $\widehat{T}_k(C) = \max\{0, \mathrm{vol}_h(C) - k\}$ and $\gamma > \frac{\mathrm{vol}_g(C_0)\,\mathrm{vol}(V)}{\theta\,\mathrm{assoc}(C_0)}$ for a feasible set $C_0 \subset V$.

We could reformulate the seed constraint $J \subset C$ as inequality constraint $|J \cap C| - |J| \geq 0$ and add a similar penalty function to the numerator of (9). However, using the structure of the problem, a more direct way to incorporate the seed constraint is possible. It holds that (9) has the equivalent form

$$\min_{A\subset V\backslash J} \frac{\mathrm{vol}_g(A) + \mathrm{vol}_g(J) + \gamma \widehat{T}_{k'}(A)}{\mathrm{assoc}(A) + \mathrm{assoc}(J) + 2\mathrm{cut}(J,A)}, \qquad (10)$$

where $k' = k - \mathrm{vol}_h(J)$. Solutions C^* of (9) and A^* of (10) are related via $C^* = A^* \cup J$. In order to derive the tight relaxation via Theorem 1, we need the Lovasz extension of the set functions in (10). For technical reasons, we replace the constant set functions $\mathrm{vol}_g(J)$ and $\mathrm{assoc}(J)$ by $\mathrm{vol}_g(J)\widehat{P}(A)$ and $\mathrm{assoc}(J)\widehat{P}(A)$, respectively, where $\widehat{P}$ is defined as $\widehat{P}(A) = 1$ for $A \neq \emptyset$ and $\widehat{P}(\emptyset) = 0$. This leads to the problem

$$\min_{A\subset V\backslash J} \frac{\mathrm{vol}_g(A) + \mathrm{vol}_g(J)\widehat{P}(A) + \gamma \widehat{T}_{k'}(A)}{\mathrm{assoc}(A) + \mathrm{assoc}(J)\widehat{P}(A) + 2\mathrm{cut}(J,A)}. \qquad (11)$$

The only difference to (10) lies in the treatment of the empty set. Note that with $\frac{0}{0} := \infty$ the empty set can never be optimal for problem (11). Given an optimal solution A^* of (11), one then either considers either $A^* \cup J$ or J, depending on whichever has lower objective, which then implies equivalence to (10).

The resulting tight relaxation will be a minimization problem over $\mathbb{R}^m$ with $m = |V\backslash J|$ and we assume wlog that the first m vertices of V are the ones in $V\backslash J$. Moreover, we use the notation $f_{\max} = \max_{i=1,\ldots,m} f_i$ for $f \in \mathbb{R}^m$, and $d_i^{(A)} = \sum_{j\in A} w_{ij}$. The following Lovasz extensions are useful:

Set function	Lovasz extension
$\mathrm{cut}(A, \overline{A})$	$\frac{1}{2}\sum_{i,j}^m w_{ij}\lvert f_i - f_j\rvert$
$\mathrm{vol}_g(A)$	$\langle f, (g_i)_{i=1}^m \rangle$
$\mathrm{assoc}(A)$	$\left\langle f, (d_i^{(V\backslash J)})_{i=1}^m \right\rangle - \frac{1}{2}\sum_{i,j}^m w_{ij}\lvert f_i - f_j\rvert$
$\widehat{P}(A)$	$f_{\max}$
$\widehat{T}_{k'}(A)$	$\langle f, (h_i)_{i=1}^m \rangle - T_{k'}^{(2)}(f)$

For the sake of brevity, we do not specify the convex function $T_{k'}^{(2)}$. The formula for a subgradient of $T_{k'}^{(2)}$ which we need in RatioDCA is given in the supplementary material. The above Lovasz extensions lead to the following tight relaxation of (11):

$$\min_{f\in\mathbb{R}_+^m} \frac{R_1(f) - R_2(f)}{S_1(f) - S_2(f)}, \qquad (12)$$

where $R_1(f) = \langle (g_i)_{i=1}^m + \gamma(h_i)_{i=1}^m, f \rangle + \mathrm{vol}_g(J) f_{\max}$, $S_1(f) = \langle (d_i)_{i=1}^m + (d_i^{(J)})_{i=1}^m, f \rangle + \mathrm{assoc}(J)\, f_{\max}$, $R_2(f) = \gamma T_{k'}^{(2)}(f)$ and $S_2(f) = \frac{1}{2}\sum_{i,j}^m w_{ij}|f_i - f_j|$.

Solution via RatioDCA. Observe that both numerator and denominator of the tight relaxation (12) are one-homogeneous d.c. functions and thus we can apply the RatioDCA of Section 4. The crucial step in the algorithm is solving the inner problem (line 3). For both (12) and the tight relaxation of the constrained balanced graph cut problem, it has the form

$$\min_{f\in\mathbb{R}_+^m,\ \|f\|_2\leq 1} \{c_1 f_{\max} + \langle f, c_2\rangle + \lambda^l \frac{1}{2}\sum_{i,j}^m w_{ij}|f_i - f_j|\},$$

for $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}^m$. We solve this problem via the following equivalent dual problem.

Lemma 1 *The above inner problem is equivalent to*

$$-\min_{\substack{\|\alpha\|_\infty\leq 1\\ \alpha_{ij}=-\alpha_{ji}}} \min_{v\in S_m} \frac{1}{2}\left\|P_{\mathbb{R}_+^m}\left(-c_1 v - c_2 - \frac{\lambda^l}{2}A\alpha\right)\right\|_2^2$$

where $(A\alpha)_i := \sum_j w_{ij}(\alpha_{ij} - \alpha_{ji})$*,* $P_{\mathbb{R}^m_+}$ *denotes the projection on the positive orthant in* $\mathbb{R}^m$ *and* S_m *is the simplex* $S_m = \{v \in \mathbb{R}^m \,|\, v_i \geq 0, \sum_{i=1}^m v_i = 1\}$*.*

This dual problem can be solved efficiently using FISTA (Beck & Teboulle, 2009), a proximal gradient method with guaranteed convergence rate $O(\frac{1}{k^2})$ where k is the number of steps. The explicit steps of FISTA for the above problem can be found in the supplementary material. The most expensive part of each iteration of the algorithm is a sparse matrix multiplication, which scales linearly in the number of edges.

6. Experiments

We empirically evaluate the performance of our approach on local clustering and community detection problems. Our goal is to address the following questions: (i) In terms of the original objective of the fractional set program, how does the locally optimal solution of our tight relaxation compare to the globally optimal solution of a loose relaxation? (ii) How good is our quality guarantee (Theorem 3), i.e. how often does our method improve a given sub-optimal solution obtained by another method?

In all experiments we start the RatioDCA with 10 different random initializations and report the result with smallest objective value. Regarding the parameter γ from Theorem 2, it turns out that best results are obtained by first solving the unconstrained case ($\gamma = 0$) and then increasing γ sequentially, until all constraints are fulfilled. In principle, this strategy could also be used to deal with soft or noisy constraints, however we focus here on the case of hard constraints.

Local clustering. We first consider the local normalized cut problem,

$$\min_{\substack{C \subset V \\ s \in C,\ \mathrm{vol}_d(C) \leq k}} \frac{\mathrm{cut}(C, \overline{C})\,\mathrm{vol}\,(V)}{\mathrm{vol}_d(C)\,\mathrm{vol}_d(\overline{C})}, \tag{13}$$

where $s \in V$ is a given seed vertex. We evaluate our approach (denoted as CFSP) against the Local Spectral (LS) method by Mahoney et al. (2012) and the Lazy Random Walk (LRW) by Andersen & Lang (2006) on large social networks of the Stanford Large Network Dataset Collection (Leskovec).

In Mahoney et al. (2012), a spectral-type relaxation is derived for (13) that can be solved globally optimally. The resulting continuous solution is then transformed into a set via optimal thresholding. However, contrary to our method this is not guaranteed to yield a set that satisfies both the seed and volume constraints. Hence Mahoney et al. (2012) suggest, at the cost of losing their approximation guarantees, to perform *constrained* optimal thresholding which considers only thresholds that yield feasible sets. In a recent generalization of their work, Hansen & Mahoney (2012) compute a sequence of locally-biased eigenvectors, the first of which corresponds to the solution of the spectral-type relaxation of Mahoney et al. (2012). We use the code of Hansen & Mahoney (2012) to compute the solution of LS in our experiments. The local clustering technique of Andersen & Lang (2006) explores the graph locally by performing a lazy random walk with the transition matrix $M = \frac{1}{2}\left(I + WD^{-1}\right)$, where D is the degree matrix of the graph and the initial distribution is concentrated on the seed set. Under some conditions on the seed set, it is shown that after a specified number of steps optimal thresholding of the random walk vector yields a set with “good” normalized Cheeger cut. However, they cannot guarantee that the resulting set contains the seed. For a fair comparison, we compute the full sequence of random walk vectors until the stationary distribution is reached, and in each step perform constrained optimal thresholding according to the normalized cut objective.

For each dataset we generate 10 random seeds. In order to ensure that meaningful intervals for the volume constraint are explored, we first solve the local clustering problem only with the seed constraint. Treating this as the “unconstrained” solution C_0, we then repeat the experiment with upper bounds of the form $\mathrm{vol}(C) \leq \alpha\,\mathrm{vol}(C_0)$, where $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$.

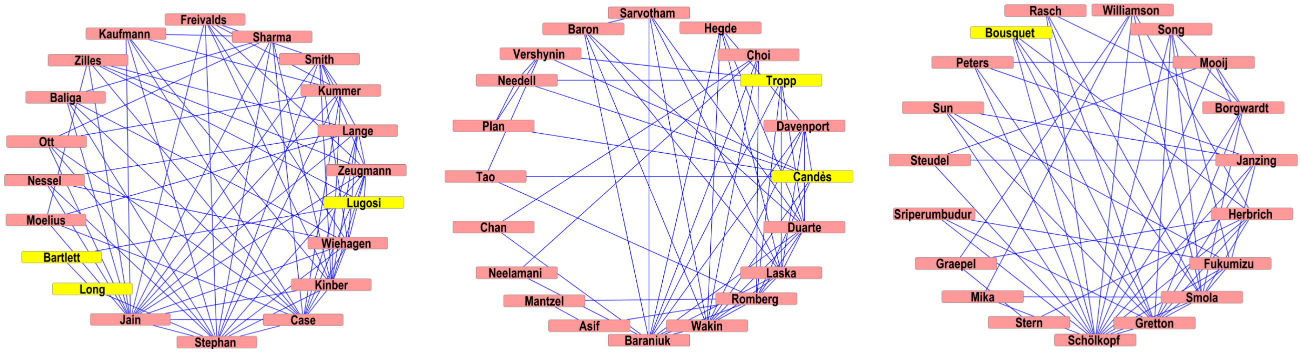
Table 1 shows mean and standard deviation of the normalized cut values averaged over the 10 different random trials (seeds) and average runtime over the different runs and volume constraints. To demonstrate the quality guarantee (Theorem 3) we also initialize CFSP with the solution of LS and LRW. Our method CFSP consistently outperforms the competing methods by large margins and always finds solutions that satisfy all constraints. In some cases CFSP initialized with LS or LRW outperforms CFSP with 10 random initializations. While LRW is very fast, the obtained normalized cuts are far from being competitive. In the supplement we show that CFSP still performs better if one uses for the optimal thresholding the normalized Cheeger cut for which LRW has been designed.

Community detection. We evaluate our approach for local community detection according to (8). The task is to extract communities around given seed sets in a co-author network constructed from the DBLP publication database. Each node in the network represents a researcher and an edge between two nodes indicates a common publication. The weights of the

Table 1. Results for the constrained local normalized cut. Our solutions (CFSP) always satisfy all constraints and have smaller cuts than the two competing methods LS and LRW.

	Method	≤ 20%	≤ 40%	≤ 60%	≤ 80%	≤ 100%	Runtime
CA-GrQc	LRW	0.1311 (0.0686)	0.1005 (0.0542)	0.0984 (0.0543)	0.0920 (0.0439)	0.0773 (0.0341)	2
(4158,13422)	LRW+CFSP	0.1048 (0.0486)	0.0695 (0.0318)	0.0614 (0.0268)	0.0614 (0.0268)	0.0457 (0.0217)	2 + 3
	LS	0.2014 (0.0958)	0.1182 (0.0958)	0.0685 (0.1089)	0.0314 (0.0423)	0.0217 (0.0259)	6
	LS+CFSP	0.1366 (0.0914)	0.0709 (0.0592)	0.0340 (0.0494)	0.0200 (0.0270)	0.0147 (0.0120)	6 + 3
	CFSP	**0.0315 (0.0292)**	**0.0157 (0.0131)**	**0.0138 (0.0115)**	**0.0083 (0.0055)**	**0.0069 (0.0044)**	31
CA-HepTh	LRW	0.2607 (0.0914)	0.2157 (0.0533)	0.2015 (0.0498)	0.1954 (0.0491)	0.1888 (0.0483)	9
(8638,24806)	LRW+CFSP	0.2074 (0.1003)	0.1076 (0.0561)	0.0976 (0.0452)	0.0882 (0.0305)	0.0869 (0.0324)	9 + 8
	LS	0.4125 (0.1079)	0.3439 (0.0631)	0.3089 (0.0839)	0.2926 (0.0913)	0.2778 (0.0923)	13
	LS+CFSP	0.3258 (0.1236)	0.1894 (0.1126)	0.1274 (0.0986)	0.0651 (0.0315)	0.0618 (0.0324)	13 + 9
	CFSP	**0.0518 (0.0226)**	**0.0327 (0.0104)**	**0.0318 (0.0094)**	**0.0263 (0.0082)**	**0.0104 (0.0038)**	58
Cit-HepTh	LRW	0.5052 (0.2208)	0.4697 (0.2010)	0.4373 (0.1962)	0.4067 (0.1998)	0.3807 (0.2224)	15
(27400,352021)	LRW+CFSP	**0.3888 (0.2261)**	**0.3249 (0.2072)**	0.2960 (0.1778)	0.2528 (0.1689)	0.2476 (0.1928)	15 + 368
	LS	0.5430 (0.2617)	0.5099 (0.2524)	0.4737 (0.2586)	0.4290 (0.2773)	0.3997 (0.2834)	175
	LS+CFSP	0.4496 (0.2848)	0.3585 (0.2185)	0.3122 (0.2138)	0.2074 (0.0814)	0.1772 (0.0782)	175 + 190
	CFSP	0.4693 (0.2676)	0.3732 (0.2166)	**0.2683 (0.1494)**	**0.1748 (0.0683)**	**0.0752 (0.0233)**	3704
Cit-HepPh	LRW	0.1784 (0.0541)	0.1466 (0.0503)	0.1234 (0.0256)	0.1079 (0.0120)	0.1048 (0.0062)	19
(34401,420784)	LRW+CFSP	0.1365 (0.0305)	0.1132 (0.0201)	0.1070 (0.0181)	0.0966 (0.0135)	0.0948 (0.0052)	19 + 219
	LS	0.1720 (0.0055)	0.1292 (0.0224)	0.1155 (0.0147)	0.1107 (0.0062)	0.1078 (0.0007)	103
	LS+CFSP	0.1335 (0.0064)	**0.1064 (0.0114)**	**0.0965 (0.0091)**	0.0944 (0.0061)	0.0916 (0.0011)	103 + 102
	CFSP	**0.1181 (0.0143)**	0.1127 (0.0101)	0.1109 (0.0089)	**0.0928 (0.0039)**	**0.0913 (0.0015)**	2666
amazon0302	LRW	0.1768 (0.0833)	0.1465 (0.0749)	0.1336 (0.0601)	0.1221 (0.0504)	0.1120 (0.0429)	336
(262111,899792)	LRW+CFSP	0.1072 (0.0666)	0.0724 (0.0455)	0.0577 (0.0419)	0.0423 (0.0373)	0.0344 (0.0294)	336 + 608
	LS	0.2662 (0.1204)	0.2496 (0.1155)	0.2247 (0.1021)	0.2066 (0.0892)	0.1946 (0.0840)	5765
	LS+CFSP	0.1775 (0.0807)	0.1248 (0.0643)	0.0923 (0.0675)	0.0878 (0.0694)	0.0641 (0.0435)	5765 + 458
	CFSP	**0.0194 (0.0063)**	**0.0095 (0.0043)**	**0.0072 (0.0031)**	**0.0056 (0.0024)**	**0.0050 (0.0022)**	3007
amazon0505	LRW	0.2472 (0.1112)	0.2369 (0.1124)	0.2249 (0.1132)	0.2200 (0.1152)	0.2163 (0.1183)	210
(410236,2439437)	LRW+CFSP	0.1058 (0.0833)	0.0636 (0.0319)	0.0636 (0.0319)	0.0636 (0.0319)	0.0610 (0.0337)	210 + 2061
	LS	0.4124 (0.1751)	0.3704 (0.1864)	0.3653 (0.1878)	0.3576 (0.1919)	0.3529 (0.1956)	20558
	LS+CFSP	0.1300 (0.0935)	0.0903 (0.0545)	0.0782 (0.0587)	0.0782 (0.0587)	0.0782 (0.0587)	20558 + 2900
	CFSP	**0.0227 (0.0076)**	**0.0116 (0.0089)**	**0.0058 (0.0020)**	**0.0048 (0.0011)**	**0.0047 (0.0008)**	13171

Figure 1. Different machine learning communities detected by our algorithm for the highlighted seeds. *Left:* Learning Theory *Middle:* Sparsity *Right:* Kernels



graph are defined as $w_{ij} = \sum_{l \in P_i \cap P_j} \frac{1}{|A_l|}$, where P_i, P_j denotes the set of publications of authors i and j and A_l denote the sets of authors for publication l, i.e. the weights represent the total contribution to shared papers. This normalization avoids the problem of giving high weight to a researcher who has publications that have a large number of authors, which usually does not reflect close collaboration with all co-authors.

To avoid finding a trivial densely connected group of researchers with few connections to the rest of the authors, we further restrict the graph by considering only authors with at least two publications and maximum distance two from the seed set. As volume function in (8), we use the volume of the original graph in order to further enforce densely connected components.

We perform local community detection with the size constraint $|C| \leq 20$ and three different seed sets $J_1 = \{\text{P. Bartlett}, \text{ P. Long}, \text{ G. Lugosi}\}$, $J_2 = \{\text{E. Candes}, \text{ J. Tropp}\}$ and $J_3 = \{\text{O. Bousquet}\}$. J_1 consists of well-known researchers in learning theory, and all members of the detected community work in this area. To validate this, we counted the number of publications in the two main theory conferences COLT and ALT. On average each author has 18.2 publications in these two conferences (see Table 3 in the supplementary material). The seeds J_2 yield a community of key scientists in the field of sparsity such as T. Tao, R. Baraniuk, J. Romberg, M. Wakin, R. Vershynin etc. The third community contains researchers who either are/were members of the group of B. Schölkopf or have closely collaborated with his group.

Acknowledgements

This work has been supported by DFG Excellence Cluster MMCI and ERC Starting Grant NOLEPRO.

References

Andersen, R. and Lang, K. Communities from seed sets. In *WWW*, pp. 223–232, 2006.

Andersen, R., Chung, F., and Lang, K. Local graph partitioning using pagerank vectors. In *FOCS*, pp. 475–486, 2006.

Bach, F. Learning with submodular functions: A convex optimization perspective. *CoRR*, abs/1111.6453, 2011.

Beck, A. and Teboulle, M. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Processing*, 18(11):2419–2434, 2009.

Bresson, X., Laurent, T., Uminsky, D., and von Brecht, J. H. Convergence and energy landscape for Cheeger cut clustering. In *NIPS*, pp. 1394–1402, 2012.

Chung, F. A local graph partitioning algorithm using heat kernel pagerank. In *WAW*, pp. 62–75, 2009.

Di Pillo, G. Exact penalty methods. In Spedicato, E. (ed.), *Algorithms for Continuous Optimization*, pp. 209–253. Kluwer, 1994.

Dinkelbach, W. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967.

Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

Gajewar, A. and Das Sarma, A. Multi-skill collaborative teams based on densest subgraphs. In *SDM*, pp. 165–176, 2012.

Goldberg, A. V. Finding a maximum density subgraph. Technical Report UCB/CSD-84-171, EECS Department, UC Berkeley, 1984.

Hagen, L. and Kahng, A. B. Fast spectral methods for ratio cut partitioning and clustering. In *ICCAD*, pp. 10–13, 1991.

Hansen, T. and Mahoney, M. Semi-supervised eigenvectors for locally-biased learning. In *NIPS*, pp. 2537–2545, 2012.

Hein, M. and Bühler, T. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In *NIPS*, pp. 847–855, 2010.

Hein, M. and Setzer, S. Beyond spectral clustering - tight relaxations of balanced graph cuts. In *NIPS*, pp. 2366–2374, 2011.

Khot, S. Ruling out PTAS for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM J. Comput.*, 36(4), 2006.

Khuller, S. and Saha, B. On finding dense subgraphs. In *ICALP*, pp. 597–608, 2009.

Leskovec, J. Stanford large network dataset collection. URL http://snap.stanford.edu/data/index.html.

Mahoney, M. W., Orecchia, L., and Vishnoi, N. K. A local spectral method for graphs: With applications to improving graph partitions and exploring data graphs locally. *JMLR*, 13:2339–2365, 2012.

Maji, S., Vishnoi, N. K., and Malik, J. Biased normalized cuts. In *CVPR*, pp. 2057–2064, 2011.

Pothen, A., Simon, H. D., and Liou, K.-P. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, 11(3):430–452, 1990.

Rangapuram, S. S. and Hein, M. Constrained 1-spectral clustering. In *AISTATS*, pp. 1143–1151, 2012.

Saha, B., Hoch, A., Khuller, S., Raschid, L., and Zhang, X.-N. Dense subgraphs with restrictions and applications to gene annotation graphs. In *RECOMB*, pp. 456–472, 2010.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 22(8):888–905, 2000.

Spielman, D. A. and Teng, S.-H. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *STOC*, pp. 81–90, 2004.

Szlam, A. and Bresson, X. Total variation and Cheeger cuts. In *ICML*, pp. 1039–1046, 2010.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

Wagstaff, K., Cardie, C., Rogers, S., and Schroedl, S. Constrained K-means clustering with background knowledge. In *ICML*, pp. 577–584, 2001.