Labels vs. Pairwise Constraints: A Unified View of Label Propagation and Constrained Spectral Clustering

Submitted for Blind Review

***Abstract*—In many real-world applications we can model the data as a graph with each node being an instance and the edges indicating a degree of similarity. Side information is often available in the form of labels for a small subset of instance, which gives rise to two problem settings and types of algorithm. In label propagation style algorithms, the known labels are propagated to the unlabeled nodes. In constrained clustering style algorithms, known labels are first converted to pairwise constraints (Must-Link and Cannot-Link), then a constrained cut is computed as a tradeoff between minimizing the cut cost and maximizing the constraint satisfaction. Both techniques are evaluated by their ability to recover the ground truth labeling, i.e. by 0/1 loss function either directly on the labels or on the pairwise relations generated from labels. These two fields have developed separately, but in this paper, we show that they are indeed related. This insight allows us to propose a novel way to generate constraints from propagated labels, which our empirical study shows outperforms and is more stable than state-of-the-art label propagation and constrained spectral clustering algorithms.**

***Keywords*-label propagation; constrained spectral clustering; semi-supervised learning**

I. Introduction

A. Motivation

A common approach to data mining and machine learning is take a set of instances and construct a corresponding graph where each node is an instance and the edge weight between two nodes is their similarity. A recent innovation has been to add labels to a small set of nodes in the graph. This setting gives rise to semi-supervised learning, which studies how to use this side information to label the unlabeled nodes. Depending on how the side information is encoded, there are two popular categories of approaches to achieve this:

1) **Label Propagation**: Side information is kept in the form of node labels. The known labels are propagated to the unlabeled nodes. The result is usually evaluated by 0/1 loss function (accuracy) directly on the labels.
2) **Constrained Spectral Clustering**: Side information is converted to pairwise constraints. A cut is computed by minimizing the cost of the cut while maximizing the constraint satisfaction. The result is usually evaluated by 0/1 loss functions, e.g. Rand index, on the pairwise relations generated from labels.

Both categories have been proven effective in their respective problem domains, but the relation between the two has not been explored [1], [2]. Exploring this topic gives rise to some interesting questions:

- Given a set of labels as side information, should we use label propagation, or should we first convert the labels into pairwise constraints (which is common practice for many constrained clustering algorithms [3]), then use constrained spectral clustering?
- Since labels are more expressive than pairwise constraints (a arbitrary set of pairwise constraints may not correspond to a unique labeling), is constrained spectral clustering inferior to label propagation?
- In the active learning setting, where we have the chance to query an oracle for ground truth, should we query labels (more expressive but difficult to acquire) or constraints (less expressive but easy to acquire)?

To address these and other questions, we need a unified view of label propagation and constrained spectral clustering.

B. Our Contribution

In this work, we explore the relation between label propagation and constrained spectral clustering. We unify the two areas by presenting a new framework called stationary label propagation. This framework gives us new insights into how side information contributes to recovering the ground truth labeling. It also enables us to propose a novel constraint construction technique which can benefit existing constrained spectral clustering algorithms.

Our contributions are:

- We establish equivalence between label propagation and constrained spectral clustering. Constrained spectral clustering using a non-negative and positive semi-definite constraint matrix is equivalent to finding a stationary labeling under label propagation (Section IV).
- We propose a novel algorithm that combines label propagation and constrained spectral clustering. It uses propagated labels to generate a (better) constraint matrix, and then uses the constraint matrix for constrained spectral clustering (Section V).
- We use empirical results to verify our claims, and demonstrates the advantage of the newly proposed algorithm over a variety of techniques. (Section VI) In particular we show that not only is our method more accurate (Fig. 4), but also more stable (Fig. 5). This addresses the stability issue of the generated constraint sets raised by Davidson et al. [4].

Table I

TABLE OF NOTATIONS

Notation	Meaning
G	An undirected (weighted) graph
A	The affinity matrix
D	The degree matrix
I	The identity matrix
$\bar{L}$	The normalized graph Laplacian
$\bar{Q}$	The normalized constraint matrix
P	The transition matrix
N	The number of nodes
y, f	The class/cluster indicator vector

II. RELATED WORK AND PRELIMINARIES

In this section, we give a brief review of existing work on label propagation and constrained spectral clustering. The notations used throughout the remainder of the paper are summarized in Table I.

A. Label Propagation

The idea of label propagation is that given a graph and a small number of nodes with known labels, we want to find a labeling of all nodes in the graph such that 1) the labeling is *smooth* over the graph and 2) the labels that are given *a priori* are not changed, or by too much. We focus on two popular label propagation techniques, namely Gaussian Fields Harmonic Function (**GFHF**) [5], [6] and Learning with Local and Global Consistency (**LLGC**) [7].

GFHF: Given a graph G of N nodes, whose affinity matrix is A. We assume the first N_1 nodes are labeled and the remaining $N_2 = N - N_1$ nodes are unlabeled. Let $y_l \in \mathbb{R}^{N_1}$ be the known labels[1]. $P = D^{-1}A$ is the transition matrix of the graph, where D is the degree matrix. We partition P into blocks

$$P = \begin{bmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{bmatrix}$$

so that P_{ul} is the transition probability from labeled nodes to unlabeled nodes, P_{uu} is the transition probability between unlabeled nodes. GFHF considers the following iterative propagation rule:

$$y^{t+1} = \begin{bmatrix} y_l^{t+1} \\ y_u^{t+1} \end{bmatrix} = \begin{bmatrix} I & 0 \\ P_{ul} & P_{uu} \end{bmatrix} \begin{bmatrix} y_l^t \\ y_u^t \end{bmatrix}. \qquad (1)$$

An illustration of the GFHF model is shown in Fig. 1.

From Eq.(1) we can see that the given labels y_l will not be changed during the propagation. Zhu and Ghahramani [6] showed that y^t converges to

$$f = \begin{bmatrix} f_l \\ f_u \end{bmatrix} = \lim_{t \to \infty} y^t = \begin{bmatrix} y_l \\ (I - P_{uu})^{-1} P_{ul} y_l \end{bmatrix}, \qquad (2)$$

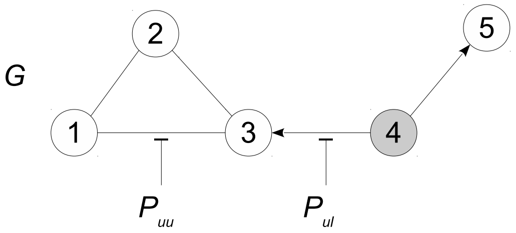


Figure 1. An illustration of the GFHF propagation model ($N = 5, N_1 = 1$). Node 4 is labeled. The propagation from 4 to 3 and 5 is governed by P_{ul} (directed edges); the propagation between 1, 2, and 3 is governed by P_{uu} (undirected edges).

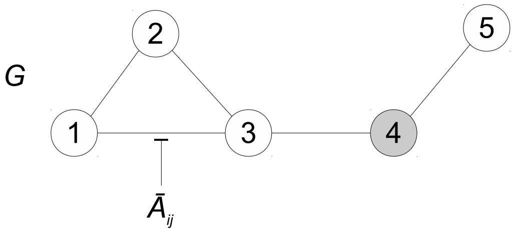


Figure 2. An illustration of the LLGC propagation model ($N = 5$). Node 4 is labeled. The propagation between nodes is governed by $\bar{A}_{ij}$ (all edges are undirected).

Zhu et al. [5] pointed out that f in Eq.(2) is also the solution to the following regularization framework:

$$\operatorname*{argmin}_{f \in \mathbb{R}^N} \frac{1}{2} \left(\sum_{i,j} A_{ij}(f_i - f_j)^2 + \infty \sum_{i=1}^{N-1} (f_i - y_i)^2 \right). \qquad (3)$$

LLGC: A key difference between LLGC and GFHF is that LLGC allows the initially given labels to be changed during the propagation process. Assume we have a graph G whose affinity matrix is A. $\bar{A} = D^{-1/2} A D^{-1/2}$ is the normalized affinity matrix. Let $y^0 \in \mathbb{R}^N$ be the initial labeling. LLGC considers the following iterative propagation rule:

$$y^{t+1} = \alpha \bar{A} y^t + (1 - \alpha) y^0, \qquad (4)$$

$\alpha \in (0, 1)$. An illustration of the LLGC model is shown in Fig. 2.

Zhou el al. [7] showed that y^t converges to:

$$f = \lim_{t \to \infty} y^t = (1 - \alpha)(I - \alpha \bar{A})^{-1} y^0. \qquad (5)$$

f is also the solution to the following regularization framework:

$$\operatorname*{argmin}_{f \in \mathbb{R}^N} \frac{1}{2} \left(\sum_{i,j} A_{ij} \left(\frac{f_i}{\sqrt{D_{ii}}} - \frac{f_j}{\sqrt{D_{jj}}}\right)^2 + \mu \sum_i (f_i - y_i)^2 \right), \qquad (6)$$

where $\mu = \frac{1-\alpha}{\alpha} \in (0, \infty)$ is a regularizer parameter.

GFHF is summarized in Algorithm 1. LLGC is summarized in Algorithm 2.

[1] For the simplicity of notations, for now we limit ourselves to the 2-class/bi-partition problem. We use binary encoding for the class/cluster indicator vector, i.e. $y \in \{-1, 0, +1\}$, where 0 means unknown. After relaxation, $y \in \mathbb{R}^N$ and 0 is used as the boundary of the two classes.

Algorithm 1: Gaussian Fields Harmonic Function [5]

```
Input: Initial labeling $y_l$, affinity matrix $A$, degree matrix $D$;
Output: $f$;
$P \leftarrow D^{-1}A$;
$f_u \leftarrow (I - P_{uu})^{-1} P_{ul} y_l$;
return $f = \begin{bmatrix} y_l \\ f_u \end{bmatrix}$;
```

Algorithm 2: Learning with Local and Global Consistency [7]

```
Input: Initial labeling $y$, affinity matrix $A$, degree matrix $D$, $\alpha \in (0,1)$;
Output: $f$;
$\bar{A} \leftarrow D^{-1/2} A D^{-1/2}$;
$f \leftarrow (I - \alpha \bar{A})^{-1} y$;
return $f$;
```

B. Constrained Spectral Clustering

The idea of constrained spectral clustering is that given a graph G and some pairwise constraints, we want to find a partition f that maximizes constraint satisfaction while minimizing the cost on G.

In particular, Wang and Davidson [8] proposed the following quadratic formulation for constrained spectral clustering:

$$\begin{aligned} \underset{f \in \mathbb{R}^N}{\operatorname{argmin}} \quad & f^T \bar{L} f, \\ \text{s.t.} \quad & f^T \bar{Q} f \geq \beta, \\ & f^T f = 1, \ f \perp D^{1/2}\mathbf{1}. \end{aligned} \tag{7}$$

$\bar{L} = I - \bar{A}$ is the normalized graph Laplacian. $\bar{Q} \in \mathbb{R}^{N \times N}$ is the normalized constraint matrix. Generally speaking, a large positive $\bar{Q}_{ij}$ indicates that node i and j should belong to the same cluster, and conversely large negative entries indicate they should be in different clusters. $f^T \bar{L} f$ is the cost of the cut f; $f^T \bar{Q} f$ measures how well f satisfies constraints in $\bar{Q}$. β is the threshold that lower bounds constraint satisfaction. To solve Eq.(7), [8] first solves the following generalized eigenvalue problem:

$$\bar{L} f = \lambda (\bar{Q} - \beta I) f. \tag{8}$$

Then they pick the eigenvectors associated with non-negative eigenvalues, i.e. these eigenvectors satisfy the constraint $f^T \bar{Q} f \geq \beta$. Among the non-negative eigenvectors, the one that minimizes $f^T \bar{L} f$ is the solution to Eq.(7). The algorithm is summarized in Algorithm 3.

C. Other Related Work

Constrained clustering has been thoroughly studied in the literature [3]. It encodes side information in the form of pairwise constraints. If the side information is originally provided as labels, most of the existing techniques first convert them into Must-Link and Cannot-Link constraints by:

$$\begin{cases} \text{Must-Link}(i,j) & i \text{ and } j \text{ have the same label,} \\ \text{Cannot-Link}(i,j) & i \text{ and } j \text{ have different labels.} \end{cases}$$

To apply the pairwise constraints to graph partition, existing methods either modify the affinity matrix directly [9]–[11], or constrain the underlying eigenspace [12]–[14]. In this work, we focus on the quadratic formulation for constrained spectral clustering proposed in [8] for two reasons: 1) the quadratic formulation matches nicely with the regularization framework for label propagation (referred to as the Generalized Label Propagation framework in [1]); 2) unlike other algorithms, the CSC algorithm in [8] can handle large amount of soft constraints, which is convenient for constraints generated from propagated labels. That being said, the equivalence we are to establish in Section IV is not limited to the formulation in [8], but also valid for other constrained spectral clustering formulation with a regularization framework.

Previous work [15]–[17] studied how to use graph partition to help propagating labels from labeled data to unlabeled data. Their problem setting is fundamentally different from ours. The effectiveness of their algorithms relies on the assumption that the graph partition is coherent with the node labels (often referred to as the *cluster assumption*). More specifically, it is assumed that 1) adjacent nodes in the graph are likely to have the same label; 2) nodes on the same structure (e.g. a well-connected subgraph) are likely to have the same label. On the other hand, the work described in this paper assumes that the side information (be it node labels or pairwise constraints) contradicts the unsupervised partition of the graph and is thus used to help finding a different partition that better conforms to the underlying ground truth.

Algorithm 3: Constrained Spectral Clustering [8]

```
Input: Affinity matrix $A$, degree matrix $D$, $\beta$;
Output: $f$;
$\bar{L} \leftarrow I - D^{-1/2} A D^{-1/2}$;
Solve the generalized eigenvalue problem $\bar{L} f = \lambda (\bar{Q} - \beta I) f$;
Let $\mathcal{F}$ be the set of all generalized eigenvectors;
foreach $f \in \mathcal{F}$ do
    $f \leftarrow f / \|f\|$;
    if $f^T \bar{Q} f < \beta$ then
        Remove $f$ from $\mathcal{F}$;
    end
end
$f \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} f^T \bar{L} f$;
return $f$;
```

III. An Overview of Our Main Results

Here we overview the main results of this work and discuss the relationship between them:

1) We introduce the notion of **Stationary label propagation** in Definition 1, Section IV-A. It can be viewed as label propagation from a latent graph to an observed graph (see Fig. 3 for an example).
2) **The equivalence between CSC and stationary label propagation** is established in Claim 1, Section IV-B. It states that the labeling derived from the constrained cut found by CSC is a stationary labeling, where P_{GH} in Fig. 3 is the constraint matrix $\bar{Q}$ and P_{GG} is the affinity matrix $\bar{A}$. This insight provides us with better understanding of how and why constrained spectral clustering works (Section IV-C). In particular, any labeling that violates the constraints will become non-stationary under propagation.
3) **LLGC is a special case of our stationary label propagation framework**, where P_{GH} is Fig. 3 is an identity matrix (Section IV-D).
4) Given the relationship between label propagation and CSC established above, we propose **a novel constraint construction algorithm**. This algorithm first propagates the labels, and then generate a constraint matrix which is a Gaussian kernel based on the propagated labels (Algorithm 4, Section V).
5) **Empirical evaluation of the new algorithm** with comparison to CSC, GFHF, and LLGC is presented in Section VI. Experimental results indicate that the new algorithm yields better (Fig. 4) and more stable (Fig. 5) results when given the same side information.

IV. The Equivalence Between Label Propagation and Constrained Spectral Clustering

In this section, we explore the relation between label propagation and constrained spectral clustering. We propose a novel concept called *stationary label propagation*, based on a variation of the GFHF propagation framework. This new concept enables us to give the CSC algorithm in [8] a label propagation interpretation. We also define stationary label propagation under the LLGC framework.

A. Stationary Label Propagation: A Variation of GFHF

Given a graph G with N nodes, all unlabeled. To establish the label propagation process, we construct a label-bearing latent graph H, whose nodes have a one-to-one correspondence to the nodes of G. H has no in-graph edges, which means its affinity matrix is I. There are edges from H to G, encoded by the transition matrix P_{GH}. The edges between the nodes of G are encoded by the transition matrix P_{GG}. Fig. 3 is an illustration of our model.

Under the GFHF propagation rule (Eq.(1)), we have:

$$y^{t+1} = \begin{bmatrix} y_H^{t+1} \\ y_G^{t+1} \end{bmatrix} = \begin{bmatrix} I & 0 \\ P_{GH} & P_{GG} \end{bmatrix} y^t. \tag{9}$$

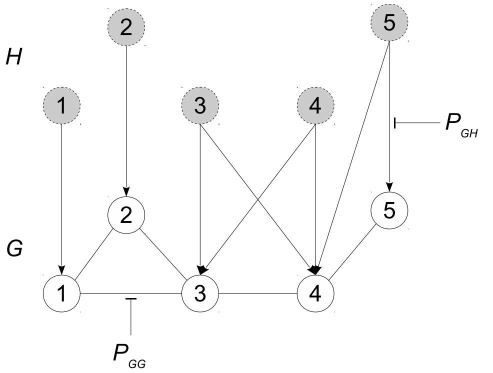


Figure 3. An illustration of our stationary label propagation model ($N = 5$). G is the unlabeled graph we want to propagate labels to. H is a latent node-bearing graph, whose node set matches the node set of G. The label propagation inside G is governed by the transition matrix P_{GG} (undirected edges). The propagation from H to G is governed by the transition matrix P_{GH} (directed edges).

Note that $y_H^t, y_G^t \in \mathbb{R}^N$, $I, P_{GH}, P_{GG} \in \mathbb{R}^{N \times N}$. From [6] we know that the labeling of graph G, which is y_G^t in Eq.(9), converges to:

$$f_G = \lim_{t \to \infty} y_G^t = (I - P_{GG})^{-1} P_{GH} y_H, \tag{10}$$

The traditional GFHF framework assumes that y_H is given and we compute f_G from y_H using Eq.(10). However, we now propose a new setting for label propagation, where we do **not** have a set of known labels to start with, i.e. y_H is unknown. Instead, we want to find a labeling f which will not change after the propagation has converged[2], and we call f a stationary labeling under propagation. Formally:

Definition 1 (Stationary Label Propagation)**.** *Let $f \in \mathbf{R}^N$ be a labeling of the shared node set of graph G and H. Under the propagation rule described in Eq.(9), we call f a stationary labeling if*

$$f = \lambda (I - P_{GG})^{-1} P_{GH} f, \tag{11}$$

where $\lambda \neq 0$ is a constant.

Intuitively speaking, a stationary labeling f is such a labeling on H that after propagation, the labeling on G will be the same as H. Recall the stationary distribution of a transition matrix P reflects the underlying structure of a random walk graph. Similarly, the stationary labeling of P_{GG} and P_{GH} reflects the underlying structure of G and the way labels are propagated from the latent graph H to G.

By re-organizing Eq.(11), we can see that given P_{GG} and P_{GH}, their stationary labeling f can be computed by solving the following generalized eigenvalue problem:

$$(I - P_{GG}) f = \lambda P_{GH} f,$$

[2] Since the nodes of G and H have one-to-one correspondence, f is a labeling of both G and H.

excluding the eigenvector associated with $\lambda = 0$. When $I - P_{GG}$ and P_{GH} are both positive semi-definite and Hermitian matrices, under some mild conditions, above generalized eigenvalue problem guarantees to have N real eigenpairs [18].

B. CSC and Stationary Label Propagation

Above we proposed the concept of stationary label propagation as a variation of the GFHF label propagation scheme. Now we establish the equivalence between CSC and stationary label propagation.

Consider the propagation matrix in Eq.(9). We use $\bar{Q}$ to replace P_{GH} and $\bar{A}$ to replace P_{GG}. $\bar{A} = D^{-1/2}AD^{-1/2}$ is the normalized affinity matrix of G, thus it is non-negative and positive semi-definite. Therefore it can (with proper normalization) serve as the transition matrix within G. Similarly, we require the constraint matrix Q in the CSC framework (Eq.(7)) to be non-negative and positive semi-definite. As a result, $\bar{Q} = D^{-1/2}QD^{-1/2}$ is also non-negative and positive semi-definite, and it serves as the transitive matrix from H to G.

With the new symbols, we rewrite Eq.(11) as follows:

$$f = \lambda(I - \bar{A})^{-1}\bar{Q}f.$$

The stationary labeling f is now the eigenvector of the generalized eigenvalue problem

$$(I - \bar{A})f = \lambda\bar{Q}f.$$

Since $\bar{L} = I - \bar{A}$ ($\bar{L}$ is called the normalized graph Laplacian), we have

$$\bar{L}f = \lambda\bar{Q}f. \tag{12}$$

Comparing Eq.(12) to Eq.(8), we can see that they are equivalent when $\beta = 0$. Note that when $\beta = 0$, since $\bar{Q}$ is now positive semi-definite, the constraint in CSC is trivially satisfied, i.e.

$$f^T\bar{Q}f \geq \beta = 0, \forall f.$$

Hence we have:

Claim 1. *With β set to 0, CSC as formulated in Eq.(7) finds all the stationary labeling f for given $\bar{A}$ and $\bar{Q}$, among which the one with the lowest cost on G will be chosen as the solution to CSC.*

Intuitively speaking, the constraint matrix $\bar{Q}$ in CSC regulates how the labels are propagated from the latent graph H to G. CSC uses the threshold β to rule out the stationary labelings that do not fit $\bar{Q}$ well enough. The solution to CSC is then chosen by graph G from the qualified labelings based on the cost function $f^T\bar{L}f$.

C. Why Constrained Spectral Clustering Works: A Label Propagation Interpretation

The equivalence between CSC and stationary label propagation provides us with a new interpretation to why and how the CSC algorithm works. Assume we have a graph G. According to some ground truth, node i and j should belong to the same cluster. However, in graph G, node i and j are not connected. As a result, if we cut G without any side information, node i and j may be incorrectly assigned to different clusters. Now we assume we have a constraint matrix Q, where $Q_{ij} = 1$. It encodes the side information that node i and j should belong to the same cluster. Under the stationary label propagation framework, $Q_{ij} = 1$ specifies that node i in the latent graph H will propagate its label to node j in graph G. As a result, an incorrect cut f where $f_i \neq f_j$ will become non-stationary under the propagation. Instead, the constrained spectral clustering will tend to find such an f that $f_i = f_j$.

Take Fig. 3 for example, if we cut graph G by itself, the partition will be $\{1, 2, 3|4, 5\}$. However, the constraint matrix (P_{GH} in the figure) specifies that node 3 and 4 should have the same label. As a result, the constrained cut will become $\{1, 2|3, 4, 5\}$.

D. LLGC and Stationary Label Propagation

Under the LLGC propagation scheme, the propagated labels converges to

$$f = (1 - \alpha)(I - \alpha\bar{A})^{-1}y^0,$$

where $\alpha \in (0, 1)$. The stationary labeling under LLGC is:

$$f = \lambda(I - \alpha\bar{A})^{-1}f, \tag{13}$$

where $\lambda \neq 0$ (the term $1 - \alpha$ is absorbed by λ).

Comparing Eq.(13) to Eq.(11), we have:

Claim 2. *Stationary label propagation under the LLGC framework is a special case of Definition 1, where $P_{GG} = \alpha\bar{A}$ and $P_{GH} = I$.*

In other words, when propagating labels from the latent graph H to graph G under the LLGC framework, the label of node i in H will only be propagated to the corresponding node i in graph G.

Combining Claim 1 and 2, we can further establish the equivalence between CSC and LLGC:

Claim 3. *With β set to 0 and $\bar{Q}$ set to I (i.e. a zero-knowledge constraint matrix), CSC finds all the stationary labeling for given $\alpha\bar{A}$ under the LLGC framework, among which the one with the lowest cost will be chosen as the solution to CSC.*

E. Remarks

We introduced above the concept of stationary label propagation as a variation of the GFHF label propagation framework. However, it is important to point out several fundamental differences between the two concepts.

The input of label propagation is a graph G with some known node labels y_l. The input of of stationary label propagation is a graph G and the transition matrix from a latent graph H to G. There are no known labels for stationary label propagation. Rather, we compute the stationary labeling based on the intrinsic characteristics of P_{GG} and P_{GH} (see Definition 1). The output of label propagation is a labeling y, which changes with the input y_l. The output of stationary label propagation is a set of stationary labeling $\{f\}$, which only depends on P_{GG} and P_{GH}.

V. Generating Pairwise Constraints via Label Propagation

Inspired by the equivalence between label propagation and constrained spectral clustering, in this section we propose a novel algorithm that combines the two techniques for semi-supervised graph partition (summarized in Algorithm 4).

Assume we have a graph G with some nodes labeled. One way to construct a constraint matrix Q from the known labels is as follows:

$$Q_{ij} = \begin{cases} 1, & i \text{ and } j \text{ have the same label} \\ 0, & \text{otherwise} \end{cases}. \qquad (14)$$

The problem with this encoding scheme is that it does not encode the Cannot-Link relation, i.e. when node i and j have different labels, Q_{ij} is set to 0, which does not distinguish from the case where the labels of node i and j are unknown. In the original CSC paper [8], Cannot-Link is encoded as $Q_{ij} = -1$. However, in that case, Q will no longer be non-negative and positive semi-definite, and will lose the label propagation interpretation we established in Claim 1.

To overcome this, we propose a new encoding scheme. First we propagate the known labels y_l to the entire graph using the GFHF method (see Algorithm 1). Let $y \in \mathbb{R}^N$ be the propagated labels. Then Q can be encoded as follows:

$$Q_{ij} = \exp(-\frac{\|y_i - y_j\|^2}{2\sigma^2}). \qquad (15)$$

It is easy to see that Q is now non-negative and positive semi-definite. Since y can be viewed as the semi-supervised embedding of the nodes, Q is essentially the similarity matrix (or a kernel) for the nodes under the new embedding.

We choose GFHF for the label propagation step instead of LLGC since in practice the side information is often from domain experts or ground truth. We do not want them to be changed during the propagation process.

```
Algorithm 4: CSC+GFHF
  Input: Initial labeling $y_l$, affinity matrix $A$, degree
         matrix $D$;
  Output: $f$;
  $P \leftarrow D^{-1}A$;
  $y_u \leftarrow (I - P_{uu})^{-1} P_{ul} y_l$;
  $y \leftarrow \begin{bmatrix} y_l \\ y_u \end{bmatrix}$;
  for $i = 1$ to $N$ do
      for $j = 1$ to $N$ do
          $Q_{ij} \leftarrow \exp(-\frac{\|y_i - y_j\|^2}{2\sigma^2})$;
      end
  end
  $D_Q \leftarrow \text{diag}(\sum_{i=1}^N Q_{i1}, \ldots, \sum_{i=1}^N Q_{iN})$;
  $\bar{Q} \leftarrow D_Q^{-1/2} Q D_Q^{-1/2}$;
  $\bar{L} \leftarrow I - D^{-1/2} A D^{-1/2}$;
  Solve the generalized eigenvalue problem $\bar{L}f = \lambda \bar{Q} f$;
  Let $\mathcal{F}$ be the set of all generalized eigenvectors;
  Remove the generalized eigenvector associated with
    $\lambda = 0$ from $\mathcal{F}$;
  $f \leftarrow \text{argmax}_{f \in \mathcal{F}} f^T \bar{Q} f$;
  return $f$;
```

After constructing Q, we normalize it to get $\bar{Q}$. Then we solve the generalized eigenvalue problem in Eq.(12) to get all the stationary labelings. We pick the one that maximally satisfies the given constraints:

$$f^* = \underset{f}{\text{argmax}}\, f^T \bar{Q} f.$$

To derive a bi-partition from f^*, we simply assign nodes corresponding to the positive entries in f^* to one cluster and negative entries the other. Notice that our algorithm, unlike the original CSC algorithm in [8], is parameter-free.

Extension to K-Way Partition: For the simplicity of notations, in this paper we assumed that the graph has 2 classes/clusters. It is straightforward to extend our formulation and algorithm to K-way case. To extend the label propagation step to K-class, we can use the following encoding scheme: Let $y \in \mathbb{R}^{N \times K}$, and

$$y_{ij} = \begin{cases} 1 & \text{node } i \text{ belongs to class/cluster } j \\ 0 & \text{otherwise} \end{cases}.$$

The way we construct Q from y in Eq.(15) remains the same, except that y_i is now a $1 \times K$ row vector. To extend the constrained spectral clustering step to K-way partition, instead of taking the top-1 eigenvector that maximizes $f^T \bar{Q} f$, we take the top-$(K-1)$ eigenvectors:

$$\underset{f_i}{\text{argmax}} \sum_{i=1}^{K-1} f_i^T \bar{Q} f_i.$$

Let F be the $N \times (K-1)$ matrix whose columns are the f_i's. A K-way partition can be derived by performing K-means on the rows of F [19].

Table II
THE UCI BENCHMARKS

Identifier	#Instances	#Attributes
Hepatitis	80	19
Iris	100	4
Wine	119	13
Glass	214	9
Ionosphere	351	34
Breast Cancer	569	30

VI. EMPIRICAL STUDY

Our empirical study aims to answer the following questions:

1) Does label propagation dominate constrained spectral clustering or vice versa?
2) How does our new algorithm compare to state-of-the-art label propagation and constrained spectral clustering techniques in terms of recovering the ground truth partition?
3) Does our new algorithm yield more stable results, i.e. is it able to generate more helpful constraint set from given labels?

A. Experiment Setup

We used six different UCI benchmark dataset, namely Hepatitis, Iris, Wine, Glass, Ionosphere and Breast Cancer Wisconsin (Diagnostic) [20] (see Table II). We removed the SETOSA class from the Iris data set, which is the class that is well separated from the other two. For the same reason we removed Class 3 from the Wine data set. We also removed data instances with missing values. After preprocessing, all datasets have two classes with ground truth labels. We used the RBF kernel to construct the infinity matrix of the graph:

$$A_{ij} = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}),$$

where $\mathbf{x}_i$ is the feature vector of the i-th instance, $i = 1, \ldots, N$.

We implemented five different techniques for our experiments:

- **Spectral**: Spectral clustering [21] on the graph without side information. This serves as the baseline performance.
- **GFHF**: The Gaussian fields harmonic function algorithm for label propagation [5] (Algorithm 1). We propagate the known labels (y_l) on the graph, and partition the graph based on the propagated labels (f) by assigning nodes with positive label values to one cluster and negative the other.
- **LLGC**: The learning with global and local consistency algorithm for label propagation [7] (Algorithm 2). The regularization parameter in LLGC was set to $\alpha = 0.5$ (which means the two terms in Eq.(6) are equally weighted) throughout the experiments.
- **CSC**: The original constrained spectral clustering algorithm in [8]. The constraint matrix was generated directly from given labels where $Q_{ij} = 1$ if the two nodes have the same label, -1 if the two nodes have different labels, 0 otherwise.
- **CSC+GFHF**: The constrained spectral clustering algorithm where the constraint matrix is constructed from propagated labels following Eq.(15) (Algorithm 4).

For each trial, we randomly revealed a subset of ground truth labels as side information. We applied the above algorithms to find a bi-partition, using the given label set. We evaluated the clustering results against the ground truth labeling using adjusted Rand index [22]. Adjust Rand index equal to 0 means the clustering is as good as a random partition and 1 means the clustering perfectly matches the ground truth. This measure is considered more informative when the sizes of the classes are imbalance. For each dataset, we varied the size of the known label set from 5% to 50% of the total size. We randomly sampled 100 different label sets for each size.

B. Results and Analysis

The accuracy of our algorithm: We report the average adjusted Rand index of all five techniques on the UCI benchmark datasets in Fig. 4. The x-axis is the number of known labels. Existing constrained spectral clustering (CSC) and label propagation (GFHF and LLGC) algorithms managed to improve over the baseline method (Spectral) only on three of six datasets. They failed to find a better partition on Wine, Glass, and Breast Cancer, even with a large number of known labels. In contrast, our approach (CSC+GFHF) was able to outperform the baseline method on all six datasets with a small number of labels. More importantly, our algorithm consistently outperformed its competitors (CSC, GFHF and LLGC) on all datasets.

The stability of our algorithm: To examine the stability of our algorithm as compared to existing approaches, we computed their performance gain over the baseline method (Spectral). Specifically, we counted out of 100 random trials how many times the four techniques (GFHF, LLGC, CSC, and CSC+GFHF) can outperform the baseline, respectively. Note that previous work on constrained clustering [4] showed that a given constraint set could contribute either positively or negatively to the clustering, therefore being able to generate constraints that are more likely to be helpful is crucial in practice. In Fig. 5 we report the percentage of trials with positive performance gain for all four techniques. We can see that CSC+GFHF consistently outperformed its competitors on all but one dataset (all four techniques performed comparably on the Hepatitis dataset). It is especially the case when we start with a very small number of labels, which means the constraint matrix for CSC is very sparse and unstable. Label propagation mitigated the problem by constructing a dense constraint matrix (CSC+GFHF). As the number of known labels increased, the results of the two algorithms eventually converged. Fig. 5 suggests that the label propagation step indeed helped to generate better constraint sets.

Comparing existing label propagation and constrained spectral clustering algorithms: From both Fig. 4 and 5 we can see that CSC generated very similar, sometimes identical, results to the two label propagation algorithms over the six datasets, LLGC in particular. This observation supported the equivalence we established between these approaches.

VII. Conclusion and Future Work

In this work we explored the relationship between two popular semi-supervised graph learning techniques: label propagation, which uses node labels as side information, and constrained spectral clustering, which uses pairwise constraints as side information. We related the two approaches by introducing a new framework called stationary label propagation, under which either nodes labels or pairwise constraints can be encoded as the transition matrix from a latent graph to the observed unlabeled graph. A stationary labeling will then simultaneously capture the characteristics of the observed graph and the side information. Inspired by this new insight, we propose a new constraint construction algorithm. Instead of generating pairwise constraints directly from the given labels, our algorithm generates constraints from a kernel based on the propagated labels. Empirical results suggested that our algorithm is more accurate at recovering the ground truth labeling than state-of-the-art label propagation and constrained spectral clustering algorithms. More importantly, its performance is also more stable over randomly chosen label sets.

Given the promising results, in the future we wish to study active learning settings where labels and pairwise constraints are queried collectively in order to maximize the efficiency. The label propagation interpretation to constrained spectral clustering can also facilitate the study of the utility of the constraints given to various constrained clustering algorithms. This will help practitioners to build algorithms with more predictable outcomes.

References

[1] A. Agovic and A. Banerjee, "A unified view of graph-based semi-supervised learning: Label propagation, graph-cuts, and embeddings," University of Minnesota, Tech. Rep. CSE 09-012, 2009.

[2] X. Zhu, "Semi-supervised learning literature survey," University of Wisconsin - Madison, Tech. Rep. CS 1530, 2008.

[3] S. Basu, I. Davidson, and K. Wagstaff, Eds., *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 2008.

[4] I. Davidson, K. Wagstaff, and S. Basu, "Measuring constraint-set utility for partitional clustering algorithms," in *PKDD*, 2006, pp. 115–126.

[5] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003, pp. 912–919.

[6] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnegie Mellon University, Tech. Rep. CMU-CALD-02-107, 2002.

[7] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *NIPS*, 2003.

[8] X. Wang and I. Davidson, "Flexible constrained spectral clustering," in *KDD*, 2010, pp. 563–572.

[9] S. D. Kamvar, D. Klein, and C. D. Manning, "Spectral learning," in *IJCAI*, 2003, pp. 561–566.

[10] B. Kulis, S. Basu, I. S. Dhillon, and R. J. Mooney, "Semi-supervised graph clustering: a kernel approach," in *ICML*, 2005, pp. 457–464.

[11] X. Ji and W. Xu, "Document clustering with prior knowledge," in *SIGIR*, 2006, pp. 405–412.

[12] S. X. Yu and J. Shi, "Grouping with bias," in *NIPS*, 2001, pp. 1327–1334.

[13] ——, "Segmentation given partial grouping constraints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 173–183, 2004.

[14] T. De Bie, J. A. K. Suykens, and B. De Moor, "Learning from general label constraints," in *SSPR/SPR*, 2004, pp. 671–679.

[15] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *ICML*, 2001, pp. 19–26.

[16] A. Blum, J. D. Lafferty, M. R. Rwebangira, and R. Reddy, "Semi-supervised learning using randomized mincuts," in *ICML*, 2004.

[17] T. Joachims, "Transductive learning via spectral graph partitioning," in *ICML*, 2003, pp. 290–297.

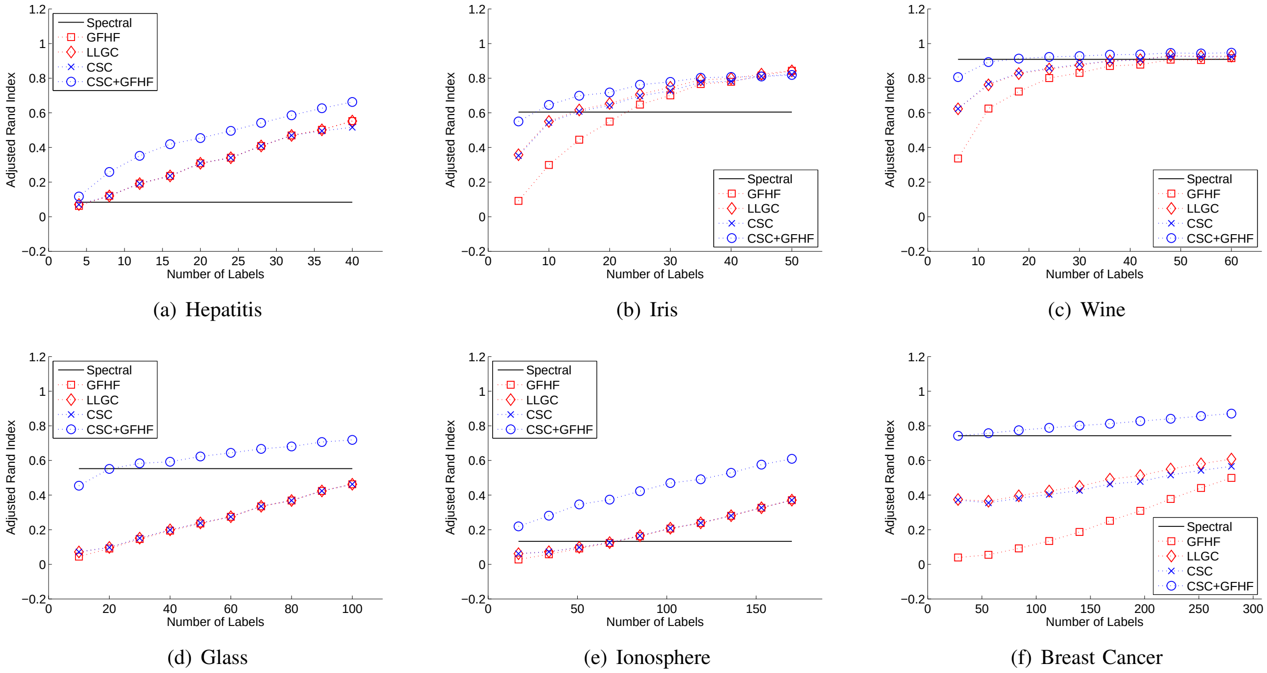
[18] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, Eds., *Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide*. Philadelphia: SIAM, 2000.

[19] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[20] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[21] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

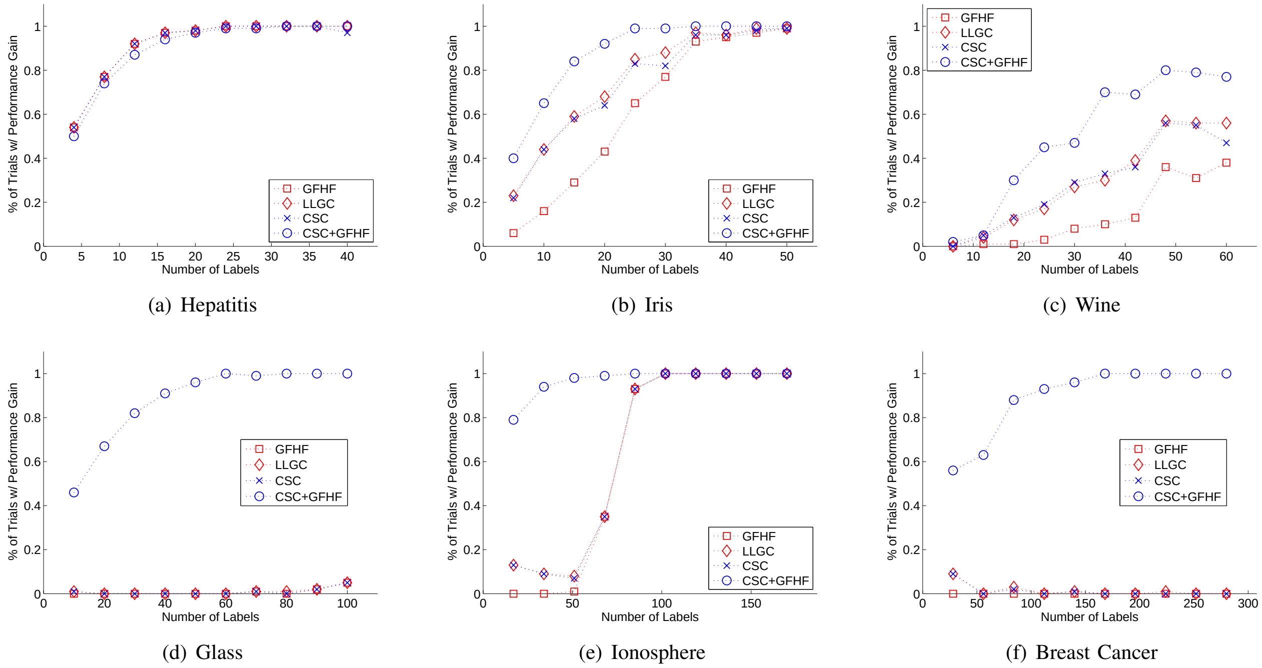
[22] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.



(a) Hepatitis (b) Iris (c) Wine

(d) Glass (e) Ionosphere (f) Breast Cancer

Figure 4. The average adjusted Rand index over 100 randomly chosen label sets of varying sizes. (UCI benchmarks)



(a) Hepatitis (b) Iris (c) Wine

(d) Glass (e) Ionosphere (f) Breast Cancer

Figure 5. The percentage of randomly chosen label sets that lead to positive performance gain with respect to the spectral clustering baseline. (UCI benchmarks)