Margin based Transductive Graph Cuts using Linear Programming

K. Pelckmans(1), **J. Shawe-Taylor**(2), **J.A.K. Suykens**(1), **B. De Moor** (1)
(1) K.U.Leuven - ESAT - SCD/SISTA, Kasteelpark Arenberg 10, B-3001, Leuven (Heverlee), Belgium
(2) University College London, Dept. of Computer Science, Gower Street, London WC1E 6BT, UK
kristiaan.pelckmans@esat.kuleuven.be

Abstract

This paper studies the problem of inferring a partition (or a graph cut) of an undirected deterministic graph where the labels of some nodes are observed - thereby bridging a gap between graph theory and probabilistic inference techniques. Given a weighted graph, we focus on the rules of weighted neighbors to predict the label of a particular node. A maximum margin and maximal average margin based argument is used to prove a generalization bound, and is subsequently related to the classical MINCUT approach. From a practical perspective a simple and intuitive, but efficient convex formulation is constructed. This scheme can readily be implemented as a linear program which scales well till a few thousands of (labeled or unlabeled) data-points. The extremal case is studied where one observes only a single label, and this setting is related to the task of unsupervised clustering.

Keywords: Graph Cuts, Transductive Inference, Statistical Learning, Clustering, Combinatorial and Convex Optimization

1 Introduction

The problems of minimal graph cuts (MINCUT) and related algorithms have an interesting history which can be traced back to the earlier work on linear and integer programming (see [14, 19] for a history), and appears often in a context of NP hardness results. Recent research witnessed a renewed surge of interest in the MINCUT problem, culminating in the theoretical seminal paper [13], and the paper [22] which is of great practical interest.

The theory of learning without reference to a parametric class of underlying stochastic models was advanced by the seminal work of Vapnik, see e.g. [23] for an overview. Its relevance in practical situations was the topic of the earlier work on maximal margin methods, see e.g. [2] and structural risk minimization in terms of data-dependent quantities [20]. A key achievement was booked through the practical and theoretical analysis of the Support Vector Machine (SVM), see e.g. [21]. The benefits of transductive inference were pinpointed e.g. in [23, 9], and integrated in practical methods as the transductive SVMs [4] and SDP relaxations as in [11]. The transduction of labels on graphs resulted in established methods as e.g. the so-called Spectral Graph Transducer (SGT) [15], label propagation [24], and other related approaches as described e.g. in [8]. Transductive inference on graphs triggered research in machine learning in different ways, illustrated e.g. by the work on learning convex combinations of subgraphs, see e.g. [3].

This paper considers the specific problem of transductive inference on a deterministic graph, i.e. the graph topology (and the corresponding weighting terms) is fully observed, i.e. not governed by any probabilistic rules. Given the labels of a few nodes, transductive inference concerns the prediction of the labels of the remaining nodes. A key element for starting an analysis was put into play by considering a random mechanism on the *sample* of nodes whose labels are observed. From a theoretical point of view, the contribution of this paper is that we indicate the importance of the role of a predictor rule (even in the case of deterministic transduction) for restricting the hypothesis space. This idea is exemplified via the adoption of an all-neighborhood rule, and the mechanism of maximal margin and maximal average margin. This paper however works on the above deterministic assumption, and does not yet consider the challenging problem considering inference on random graphs and related convergence issues when the number of nodes increases (i.e. the semi-supervised and inductive case).

From an algorithmical point of view, the main insight

of this paper is as follows. Let a graph with n nodes be divided in a set of positive and negative nodes, say $q \in \{-1,1\}^n$. Let the fixed weights of all undirected edges between different nodes be denoted as $\{w_{ij} \geq 0\}$. Then the weight of the cut corresponding to q can be formalized as

$$\mathrm{CUT}(q) = \sum_{ij|\ q_i \neq q_j} w_{ij} = \frac{1}{4}\sum_{i,j=1}^{n} w_{ij}(q_i - q_j)^2 \quad (1)$$

$$= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(1 - q_i q_j), \quad (2)$$

resulting in the eigenvalue relaxation (also called spectral relaxation, see e.g. [12, 22]) and the Semi-Definite Programming (SDP) relaxation respectively (see e.g. [13, 11]). This paper rewrites the weight of a cut q instead in terms of absolute values as follows:

$$\mathrm{CUT}(q) = \sum_{ij|q_i \neq q_j} w_{ij} = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left|q_i - q_j\right|, \quad (3)$$

resulting in a linear programming relaxation. Advantages are found in the flexibility of this approach (e.g. it becomes straightforward to incorporate additional prior knowledge, modified cost criteria or extra constraints), the connection with one of the most thoroughly studied algorithmical fields (namely linear programming), and its implication for hardness results of combinatorial optimization. Moreover, the relaxation yields in most cases (at least in practice) solutions which satisfy the original integer constraints exactly, thereby avoiding the need for an extra post-processing step (as thresholding, K-means or random projections) as are used in other relaxations.

This paper is organized as follows. Section 2 studies the theoretical properties of the class of neighborhood rules, the maximal margin derivation and the relationship with MINCUT approaches. Section 3 discusses the convex approach for learning based on the above principles and its implications for clustering. Section 4 gives a proof of concept based on an artificial example.

2 Transductive Inference on a Deterministic Graph

2.1 Transductive Graph Cuts

Let $\mathcal{G}_n \subset (\mathcal{V}, \mathcal{W})$ be a fixed observed graph with n nodes $\mathcal{V} = \{v_i\}_{i=1}^n$ and corresponding edges $\mathcal{W} = \{w_i = (w_{i1}, \ldots, w_{in})^T \in \mathbb{R}_+^n\}_{i=1}^n$ with positive terms $w_{ij} \geq 0$ denoting the weight of the connection between v_i and v_j. An undirected and loopless graph is assumed, such that $w_{ij} = w_{ji}$ for all $i, j = 1, \ldots, n$, and $w_{ii} = 0$ for all $i = 1, \ldots, n$. Let $\mathcal{S}$ denote the set containing the indices of nodes having an observed label y_i. Let the degrees d_i be defined as $\sum_{j=1}^n w_{ij} = d_i$ for all $i = 1, \ldots, n$. Let a general labeling of all n nodes be denoted as $q \in \{-1,1\}^n$. Let the hypothesis set $\mathbb{H}^n$ be defined as

$$\mathbb{H}^n = \{q \in \{-1,1\}^n\}, \quad (4)$$

containing a finite number - namely 2^n - of different hypotheses. Note that this is essentially different from the inductive setting as an hypothesis does not represent a predictor function, or a parameter set. Assume there is a unique binary vector $y \in \{-1,1\}^n$ denoting the *true* (but not necessarily observed) labels of each node. Transductive inference of all labels of the deterministic graph $\mathcal{G}_n$ picks a single element $\hat{q}$ from the hypothesis class $\mathbb{H}^n$ which agrees maximally with the given partial labels $y_s \in \{-1,1\}^{n_s}$ associated with the nodes $\{v_j\}_{j \in \mathcal{S}}$ and where $n_s = |\mathcal{S}|$. The working assumption of transductive inference is that a proper restriction of the hypothesis space $\mathbb{H}^n$ will allow one to infer a good matching of $\hat{q}$ with the complete vector y based on a few observations $\mathcal{S}$.

Formally, let (y_K, q_K, w_K) denote the actual label, the hypothetical label and the connections associated to the Kth (unspecified) node $v_K \in \{v_1, \ldots, v_n\}$. One way to formalize the risk is as follows [23, 17]:

$$\mathcal{R}(q) = E[I(y_K q_K < 0)], \quad (5)$$

where $I(u < 0)$ equals 1 if $u < 0$ and zero otherwise, and where the expectation E is taken over the uniform choice of K, denoting which node v_K is considered. The empirical counterpart becomes

$$\mathcal{R}_{\mathcal{S}}(q) = \sum_{k \in \mathcal{S}} I(y_k q_k < 0) = \frac{n_{\mathcal{S}} - \sum_{k \in \mathcal{S}} y_k q_k}{2}, \quad (6)$$

where $\mathcal{S}$ is the iid sample of labeled nodes and $n_{\mathcal{S}} = |\mathcal{S}|$. The following probabilistic bound holds:

Theorem 1 (Generalization Bound) *Let $\mathcal{S} \subset \{1, \ldots, n\}$ be uniformly sampled without replacement. Consider a set of hypothetical labelings $\mathbb{H}' \subset \mathbb{H}^n$ having a cardinality of $|\mathbb{H}'| \in \mathbb{N}$. Then the following inequality holds with probability higher than $(1 - \delta) < 1$.*

$$\sup_{q \in \mathbb{H}'} \mathcal{R}(q) - \mathcal{R}_{\mathcal{S}}(q) \leq \sqrt{\frac{2(n - n_s + 1)}{n_s n}\left(\log(|\mathbb{H}'|) - \log(\delta)\right)}. \quad (7)$$

Proof: This statement follows directly from Serfling's inequality, used similarly as in [17], Theorem 14.

□

The following two subsections study how to construct an appropriate restriction of the hypothesis space based on the adoption of a suitable prediction rule, and the use of a maximal margin principle.

2.2 A Maximal Margin Approach

We now consider the construction of an appropriate hypothesis set $\mathbb{H}'$. At first, consider the All-Neighbors Rule r_q (ANR) defined for a given vector $q \in \{-1,1\}^n$, and evaluated on node v_* as

$$r_q(v_*) = \text{sign}\left(\sum_{j=1}^{n} w_{*j} q_j\right) = \text{sign}(w_*^T q). \quad (8)$$

Note the relation with the common K-NN rules. The key element is to restrict attention to those hypothetical labelings $q \in \mathbb{H}$ which are to a certain degree *consistent* with themselves: a label q_i is consistent with the corresponding ANR rule if

$$r_q(v_i) = q_i \Leftrightarrow q_i(w_i^T q) \geq 0, \quad (9)$$

and thus can be predicted accurately based solely on its neighborhood. Remark that the property of $w_{ii} = 0$ hints at a leave-one-out setting, as explored in e.g. [5, 6]. The margin of the classifier on the ith node $\text{sign}(w_i^T q)$, and the corresponding label $q_i \in \{-1,1\}$ is defined as

$$\begin{aligned} m_i(q) &= \frac{q_i(w_i^T q)}{\sqrt{q^T q}} \\ &= \frac{1}{\sqrt{n}} \sum_{j|\ q_i = q_j} w_{ij} - \frac{1}{\sqrt{n}} \sum_{j|\ q_i \neq q_j} w_{ij} \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} w_{ij} - \frac{2}{\sqrt{n}} \sum_{j|\ q_i \neq q_j} w_{ij} \\ &= \frac{d_i}{\sqrt{n}} - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} w_{ij}|q_i - q_j|, \end{aligned} \quad (10)$$

since $q^T q = n$ by construction. Restricting the hypothesis set to all hypothetical labelings which are consistent with the ANR rule with at least margin $\rho > 0$ gives

$$\mathbb{H}_\rho = \left\{ q \in \{-1,1\}^n \ \middle|\ m_i(q) \geq \rho, \quad \forall i = 1, \ldots, n \right\}, \quad (11)$$

where the rule r_q acts as a restriction of the hypothesis space, not as a predictor. From a practical perspective this gives the following learning problem for a fixed ρ

$$\begin{aligned} \hat{q} = \underset{q \in \{-1,1\}^n}{\arg\min}\ \mathcal{J}_\rho(q) &= -\sum_{k \in \mathcal{S}} q_k y_k \\ \text{s.t.} \quad \frac{1}{\sqrt{n}} \sum_{j=1}^{n} w_{ij}|q_i - q_j| &\leq \frac{d_i}{\sqrt{n}} - \rho, \quad \forall i, \end{aligned} \quad (12)$$

which can be solved as an integer programming problem. A relaxation in terms of a linear programming problem (LP) gives: $\hat{q} = \arg\min_{q \in [-1,1]^n} \mathcal{J}'_\rho(q) = -\sum_{k \in \mathcal{S}} q_k y_k$ s.t. $\frac{1}{\sqrt{n}} \sum_{j=1}^{n} w_{ij}|q_i - q_j| \leq \frac{d_i}{\sqrt{n}} - \rho$, $\forall i = 1, \ldots, n$. It follows that the solution is unique when ρ is taken high enough. Note that the predicted labels $\hat{q}$ can also be derived from the consistent ANR rule with parameters $\hat{q}$. Numerical case-studies however indicate that this relaxation is not behaving very well in practice. Instead of advancing to dedicated integer programming approaches as based on cutting plane algorithms (see e.g. [19]), a slightly different formulation is proposed in the following subsection. From a theoretical point of view, it turns out that the maximal number of hypotheses q in $\mathbb{H}_\rho$ can be approximated well as indicated in the paper [18]. Therefor, the authors introduced a measure denoted as the Kingdom-capacity $\vartheta(\rho)$ of a deterministic graph. This capacity is based on the analogy with a strategy game asking the following: "what is the maximum number of kings which have a large enough kingdom to enforce their will." This definition is closely related to coloring capacities and Shannon capacity of a graph [16], and resembles closely the definition of information capacity of a network (see e.g. [1]). A key difference with this work on capacities is however the notion of margin, which appears to be new to the discussion. The Kingdom-capacity of a deterministic graph actually equals the classical VC dimension of the described hypothesis class $\mathbb{H}_\rho$ equipped with the ANR rule. Naming convention is however kept different to discriminate with *the* VC dimension of a graph as discussed e.g. in [10]. The number of elements in the set $\mathbb{H}_\rho$ can then be bounded using a simple combinatorial argument as in [18], namely $|\mathbb{H}_\rho| \leq \sum_{d=0}^{\vartheta(\rho)} \binom{n}{d}$. Note the correspondence with Sauer's Lemma, see e.g. [2].

2.3 A Maximal Average Margin Approach

An alternative construction of the hypothesis class $\mathbb{H}'$ is considered: we restrict attention to all labelings having margins which are on the average larger than a certain pre-specified value $\overline{\rho} > 0$. The average margin can be written as follows

$$\begin{aligned} \overline{\rho} \leq \overline{m}(q) &= \frac{1}{n} \sum_{i=1}^{n} m_i(q) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{d_i}{\sqrt{n}} - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} w_{ij}|q_i - q_j| \right) \\ &= \frac{1}{n\sqrt{n}} \sum_{i=1}^{n} d_i - \frac{1}{n\sqrt{n}} \sum_{i<j} 2w_{ij}|q_i - q_j|, \end{aligned} \quad (13)$$

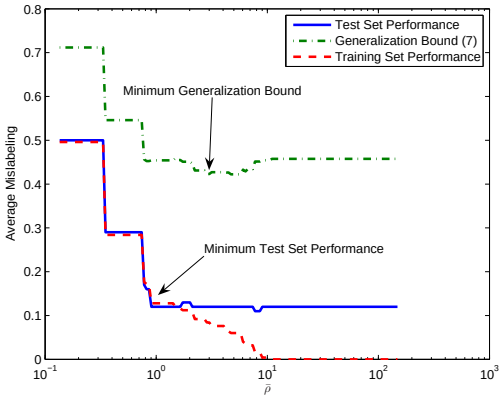

Figure 1: *Numerical results of the problem (15) for 20 different values of $\bar{\rho}$. Results on the Ripley data with 250 labeled and 100 unlabeled (test) samples. The Ripley samples are generated from 2 overlapping distributions, making a tuning for $\bar{\rho}$ crucial. The figure displays the test set performance on the 100 unlabeled samples, training set performance $\mathcal{R}_n(\hat{q})$ as in (6), and the generalization bound (7) corrected for the 20 different prespecified values $\bar{\rho}$.*

and the hypothesis class $\mathbb{H}_{\overline{\rho}}$ can be written as

$$\mathbb{H}_{\overline{\rho}} = \Big\{ q \in \{-1,1\}^n \;\Big|\; \sum_{i<j} 2w_{ij}|q_i - q_j| \leq \sum_{i=1}^{n} d_i - n\sqrt{n\overline{\rho}} \Big\}. \quad (14)$$

Such sets form by construction a properly nested structure of hypothesis classes such that one can write $\mathbb{H}_{\overline{\rho}_k} \subseteq \mathbb{H}_{\overline{\rho}_l}$ whenever $\overline{\rho}_k \geq \overline{\rho}_l$. This enables structural risk minimization in this context. Neglecting all constants in expression (14), it becomes clear that minimizing the cardinality of this set can be done by minimizing $\sum_{i<j} 2w_{ij}|q_i - q_j|$. Minimizing the volume of $\mathbb{H}_{\overline{\rho}}$ can as such be done by using the MINCUT algorithm, as e.g. in [13, 22]. We approach the related problem of finding the hypothesis q from $\mathbb{H}_{\overline{\rho}}$ which coincides optimally with the observed labels. The maximal consistent hypothesis $q \in \{-1,1\}^n$ which belongs to the minimal set $\mathbb{H}_{\overline{\rho}}$ can be found by solving the following integer programming problem:

$$\hat{q} = \operatorname*{arg\,min}_{q\in\{-1,1\}^n} \mathcal{J}_{\overline{\rho}}(q) = -\sum_{k\in\mathcal{S}} q_k y_k$$
$$\text{s.t.} \quad \sum_{i<j} 2w_{ij}|q_i - q_j| \leq \sum_{i=1}^{n} d_i - n\sqrt{n\overline{\rho}}. \quad (15)$$

Note that here the labels $\hat{q}_i$ do not necessarily have the same sign as the corresponding rule $\text{sign}(w_i^T\hat{q})$ for any $i = 1,\ldots,n$. The following section discusses a practical approach. We proceed by deriving a bound on the cardinality $|\mathbb{H}_{\overline{\rho}}|$ based on the eigenvalue spectrum of the Laplacian of the graph.

Theorem 2 (Cardinality of $\mathbb{H}_{\overline{\rho}}$) *Let $\{\sigma_i\}_{i=1}^n$ denote the eigenvalues of the graph Laplacian $L = D - W$ where $D = \text{diag}(d_1,\ldots,d_n) \in \mathbb{R}^{n\times n}$. The cardinality of the set $\mathbb{H}_{\overline{\rho}}$ can then be bounded as*

$$|\mathbb{H}_{\overline{\rho}}| \leq \sum_{d=0}^{n_\sigma(\overline{\rho})} \binom{n}{d} \leq \left(\frac{en}{n_\sigma(\overline{\rho})}\right)^{n_\sigma(\overline{\rho})}, \quad (16)$$

where $n_\sigma(\overline{\rho})$ is defined as

$$n_\sigma(\overline{\rho}) = \left| \left\{ \sigma_k : \quad \sigma_k \leq 2\left(\sum_{i=1}^{n} d_i - n\sqrt{n\overline{\rho}}\right) \right\} \right|. \quad (17)$$

Proof: Let for notational convenience ρ' be defined as $\rho' = \sum_{i=1}^n d_i - n\sqrt{n\overline{\rho}}$. At first, the MINCUT criterion $\sum_{i<j} 2w_{ij}|q_i - q_j|$ is written in terms of the graph Laplacian L as follows:

$$\sum_{i<j} 2w_{ij}|q_i - q_j| = \sum_{i<j} w_{ij}(q_i - q_j)^2 = \frac{1}{2}q^T(D-W)q \triangleq \frac{1}{2}q^T L q.$$

The reasoning of the proof goes as follows: if for a specific $q \in \{-1,1\}^n$ the inequality $q^T L q \leq 2\rho'$ holds, then such a q can always be found as a signed version of an element in the smallest eigenspace of L. Formally, let $I_n \in \mathbb{R}^{n\times n}$ be the identity matrix of size n. Let $L = U\Sigma U^T$ denote the Singular Value Decomposition (SVD) of the Lagrangian, such that $UU^T = U^TU = I_n$, $\Sigma = \text{diag}(\sigma_1,\ldots,\sigma_n)$ and $0 = \sigma_1 \leq \cdots \leq \sigma_n$. It follows that any $q \in \{-1,1\}^n$ can be decomposed in terms of the singular vectors $q = \sum_{i=1}^n s_i U_i$, where $s = (s_1,\ldots,s_n)^T \in \mathbb{R}^n$.

Given the definition of $n_\sigma(\overline{\rho})$, one can write that $\sigma_i > 2\rho'$ for all $i = n_\sigma(\overline{\rho})+1,\ldots,n$. Assume $q^T L q \leq 2\rho'$, then the following inequality follows

$$2\rho' \geq q^T L q = \sum_{i=1}^{n} s_i^2\sigma_i \geq \sum_{i=n_\sigma(\overline{\rho})+1}^{n} s_i^2\sigma_i > 2\rho' \sum_{i=n_\sigma(\overline{\rho})+1}^{n} s_i^2, \quad (18)$$

and hence $\sum_{i=n_\sigma(\overline{\rho})+1}^n s_i^2 < 1$. Moreover, given $1 \leq j \leq n$ fixed, one has $\sum_{i=n_\sigma(\overline{\rho})+1}^n U_{ij}^2 \leq 1$ since ${U^j}^T U^j = 1$ where $U^j \in \mathbb{R}^{1\times n}$ denotes the jth row of the matrix U. Thus, the following inequality holds

$$\left\| \sum_{i=n_\sigma(\overline{\rho})+1}^{n} s_i U_{ij} \right\|_2 \leq \sum_{i=n_\sigma(\overline{\rho})+1}^{n} s_i^2 \sum_{i=n_\sigma(\overline{\rho})+1}^{n} U_{ij}^2 < 1, \quad (19)$$

for all $j = 1,\ldots,n$. Thus one can write any element q satisfying $q^T L q \leq 2\rho'$ as a signed version of a vector in the $n_\sigma(\overline{\rho})$-dimensional subspace, i.e. omitting the

expansion in the principal eigenvalues will not yield a pointwise difference larger than one. Thus:

$$q_j = \text{sign}\left(\sum_{i=1}^{n_\sigma(\bar{\rho})} s_i U_{ij}\right), \quad \forall j = 1, \ldots, n, \tag{20}$$

where sign(z) equals -1 if $z < 0$ and 1 otherwise. Now the result follows from Rado's theorem - stating that the VC-dimension of a linear threshold rule in a $n_\sigma(\bar{\rho})$-dimensional subspace is at most $n_\sigma(\bar{\rho})$ - combined with Sauer's Lemma.

□

This theorem directly motivates the combinatorial learning problem (15) where the average margin $\bar{\rho}$ is fixed a priori. Figure 1 displays the training error, test error and generalization bound based on the Ripley dataset, for a finite number of a priori fixed constants $\bar{\rho}$, illustrating the use of the generalization bound to pick a proper $\bar{\rho}$. We would like to point out that if $\bar{\rho}$ is also to be found from the data, one needs an extra correction of the above derivation, making the union bound for all hypothesis sets $H_{\bar{\rho}}$ with varying cardinality determined through $n_\sigma(\bar{\rho})$, see e.g. [20].

3 A Convex Algorithm

3.1 A Linear Programming Approach

This section considers a simple but powerful and flexible relaxation to the combinatorial problem (15) - in terms of a linear programming problem. Here we opt to show the version where the regularization-accuracy trade-off is made as a bi-criterion loss function in terms of μ because of practical considerations. We note again that the theoretical analysis of this version based on an empirical margin would require a correction w.r.t. the fixed margin case as described e.g. in [20].

Definition 1 (Linear Programming TGC) *Let $\mu > 0$ be a fixed constant, then the TGC follows from solving the following LP:*

$$\hat{q} = \underset{q \in [-1,1]^n}{\arg\min}\, \mathcal{J}'_\mu(q) = -\sum_{i \in \mathcal{S}} y_i q_i + \mu \sum_{i<j} w_{ij}|q_i - q_j|. \tag{21}$$

An apparent disadvantage of this formulation is the fact that one needs $\frac{1}{2}n(n-1)$ slack variables to translate all terms in the sum $\sum_{i<j} w_{ij}|q_i - q_j|$. Remark however that algorithms which are based on SDP-relaxations scale similarly. This is because one parameterizes the problem there as a function of the squared matrix $\Lambda \in \mathbb{R}^{n \times n}$, representing $\Lambda = qq^T$ [12, 13]. A major advantage of the LP formulation however is that any sparseness in the weights results directly in a reduction of computational complexity as the terms $w_{ij}|q_i - q_j|$ become obsolete when $w_{ij} = 0$. In case every node is on the average connected to d neighboring nodes, the problem can be solved with a complexity $O((nd)^3)$. It is furthermore to be expected that structure can be exploited to find a more efficient algorithm using a graph labeling algorithm as common in the literature on combinatorial optimization.

Practice suggests that solutions $\hat{q}$ will often satisfy $\hat{q}_i \in \{-1, 1\}$ for any $i = 1, \ldots, n$. This is a consequence of the box constraints in the LP. We can however not guarantee this property, as easily seen when μ is taken much too high: in that case all values $\hat{q}_i$ will roughly equal one single value strictly between -1 and 1. This observation however makes an additional thresholding step as common in spectral or SDP approaches often obsolete.

3.2 The Dual Minimal Overflow Problem

The dual problem can be written as follows. Let $Y \in \{-1, 0, 1\}^n$ denote the labels if given such that $Y = (y'_1, \ldots, y'_n)^T$, $y'_i = y_i$ if $i \in \mathcal{S}$ and 0 otherwise. Let the matrix $\Delta^w \in \mathbb{R}^{M \times n}$ denote the weighted first order difference matrix such that for each $m = 1, \ldots, M$, there exists a unique combination (i, j) with $1 \leq i < j \leq n$ where $\Delta^w_m q = w_{ij}(q_i - q_j)$. Note that the matrix Δ corresponds with the incidence matrix in case of an unweighted graph. The Laplacian $\mathcal{L}(q, \xi; \alpha^+, \alpha^-, \beta^+, \beta^-)$ - abbreviated as $\mathcal{L}(q, \xi; \cdot)$ - of (21) becomes $\mathcal{L}(q, \xi; \cdot) = -\sum_{i \in \mathcal{S}} y_i q_i + \mu \sum_{m:i<j} w_{ij}\xi_m$ $- \sum_{i=1}^n \left(\alpha_i^+(1 + q_i) + \alpha_i^-(1 - q_i)\right)$ $- \sum_{m=1}^M (\beta_m^+(\Delta_m^w q + \xi_m) + \beta_m^-(\xi_m - \Delta_m^w q))$, with multipliers $\alpha_i^+, \alpha_i^- \geq 0$ for all $i = 1, \ldots, n$ and $\beta_m^+, \beta_m^- \geq 0$ for all $m = 1, \ldots, M$. The first order conditions for optimality $\frac{\partial \mathcal{L}(q, \xi; \cdot)}{\partial q_i} = 0$ and $\frac{\partial \mathcal{L}(q, \xi; \cdot)}{\partial \xi_i} = 0$ give the equalities

$$\begin{cases} y_i = (\alpha_i^+ - \alpha_i^-) + (\beta^+ - \beta^-)^T \Delta^{w,i} & i \in \mathcal{S} \\ 0 = (\alpha_i^+ - \alpha_i^-) + (\beta^+ - \beta^-)^T \Delta^{w,i} & i \notin \mathcal{S} \\ (\beta_m^+ + \beta_m^-) = \mu w_{ij} & \forall m. \end{cases} \tag{22}$$

As Slater's condition holds [7], the duality gap becomes zero, and the dual problem can be written as $\max_{\alpha^+, \alpha^-, \beta^+, \beta^- \geq 0} \min_{q, \xi} \mathcal{L}(q, \xi; \alpha^+, \alpha^-, \beta^+, \beta^-)$, can be written as

$$\min_{\alpha, \beta} \sum_{i=1}^n |\alpha_i| \quad \text{s.t.} \quad \begin{cases} \alpha_i + \beta^T \Delta^{w,i} = y_i & i \in \mathcal{S} \\ \alpha_i + \beta^T \Delta^{w,i} = 0 & i \notin \mathcal{S} \\ |\beta_m| \leq w_{ij}\mu & \forall m, \end{cases} \tag{23}$$

where we define $\alpha_i = (\alpha_i^+ - \alpha_i^-)$ for all $i = 1, \ldots, n$, and $\beta_m = (\beta_m^+ - \beta_m^-)$ for all $m = 1, \ldots, M$. By eliminating

α, the problem can be rewritten as

$$\min_{|\beta_m| \leq \mu w_{ij}} \|Y - \Delta^{w^T} \beta\|_1. \tag{24}$$

This dual problem turns out to give the solution to a similar problem. Consider the problem of establishing an optimal flow between a set of source nodes, and a set of sink nodes. Let $\nu > 0$ be a fixed constant. Let $\mathcal{G}$ be a loopless graph with nodes $\{v_i\}_{i=1}^n$, but let $\{\nu w_{ij}\}_{i \neq j}$ denote the maximal capacity of a flow from node v_i to v_j, in either direction (i.e. $|f_{ij}| \leq w_{ij}$ for all $i \neq j$). Now the problem of generalized max flow is to look for a configuration of flows $\{f_{ij}\}_{i \neq j}$ redirecting the flow from all sources to all sinks, i.e. as far as can be handled by the graph. Therefor, let the vector $z = (z_1, \ldots, z_n)^T \in \{-1, 0, 1\}^n$ be defined as

$$\forall i = 1, \ldots, n: \quad \begin{cases} z_i = +1 & \text{iff } v_i \text{ is a source node} \\ z_i = -1 & \text{iff } v_i \text{ is a sink node} \\ z_i = 0 & \text{otherwise.} \end{cases} \tag{25}$$

In case a node is a sink nor a source, the sum of the flows has to be zero as any overhead causes flooding the graph. This yields the following formulation

$$\exists \{|f_{ij}| \leq \nu w_{ij}\}_{i \neq j}: \quad \sum_{j \neq i} f_{ij} = z_i, \quad \forall i = 1, \ldots, n. \tag{26}$$

Allowing for small deviations for handling the case when the graph cannot handle the total flow properly yields the minimal overflow problem which we define as follows

Definition 2 (The minimal Overflow Problem) *Let $\nu > 0$ be fixed, then the flows wich will cause minimal overflow are given as the solution of a optimization problem as follows*

$$\hat{f} = \underset{|f_{ij}| \leq \nu w_{ij}}{\arg\min} \mathcal{J}_\nu(f) = \sum_{i=1}^n \left| z_i - \sum_{j \neq i} f_{ij} \right|. \tag{27}$$

Theorem 3 (Duality MOP - LPcut) *By comparison of problem (24) and problem (27), the LP formulation (21) is seen to be the dual to the minimal overflow problem where $\mu = \nu$.*

Note the direct relationship of this duality result with the well-known Max-Flow Min-Cut theorem by Ford and Fulkerson, see e.g. [19].

3.3 Transductive Graph Cuts with One-Class Labels

The idea of balancing the unsupervised labels (i.e. imposing that there are roughly a fixed amount of (unsupervised) nodes corresponding with each label) was already explored in various publications, see e.g. [9, 15].

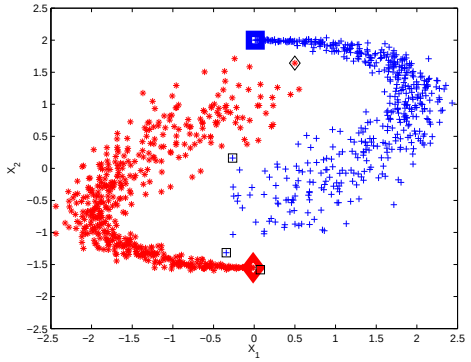

Figure 2: *A toy example of 1000 samples, iid sampled from 2 stochastic* Yin-Yang *like intertwined sources (stars and crosses). The extremal nodes (big marks) were labeled with $+1$ and -1 respectively, the labels of the remaining samples (denoted as square or diamond) were found, satisfying exactly $\hat{q} \in \{-1, 1\}^{1000}$. The algorithm assigns a wrong label to only 3 out of 998 samples due to the overlapping distributions (false positive predictions are indicated by a diamond, true negative predictions by a square).*

The same idea is used here to construct an algorithm for datasets with only positive observed labels. The technique of balancing here is used to counteract the effect of the positive labels, avoiding the trivial solution where all labels equal $+1$. A sufficient counterweight to those positive samples is found in the constraint of having at least a certain amount of labels to be negative. Note that this translates the intuition that one wants to find a class of limited size. Incorporating this constraint gives the following formulation:

Definition 3 (TGC with Balancing) *Let B be a positive known constant. A graph cut belonging to the hypothesis set $\mathbb{H}_{\bar{\rho}}$ containing at least a portion of $Bn \leq n$ negative samples can be found (if it exists) by solving the following integer programming problem:*

$$\min_{q \in \{-1,1\}^n} \mathcal{J}_{\rho,B}(q) = \sum_{i \in \mathcal{S}} -y_i q_i$$
$$s.t. \quad \begin{cases} \sum_{i=1}^n (q_i - 1) \leq -2Bn \\ \sum_{i<j} 2w_{ij} |q_i - q_j| \leq 2\rho'. \end{cases} \tag{28}$$

As previously, we suggest to approximate this problem by a linear programming problem through relaxing the constraints $q_i \in \{-1, 1\}$ as $q_i \in [-1, 1]$. Note that in (28), the regularization term $\sum_{i<j} 2w_{ij}|q_i - q_j|$ is written as a hard constraint $\sum_{i<j} 2w_{ij}|q_i - q_j| \leq 2\rho'$ in order to emphasize the different components.

At this stage, it is instructive to discuss the implications for a clustering algorithm based on MINCUT. The idea is that one can perform this transductive inference algorithm for the graph with each node labeled

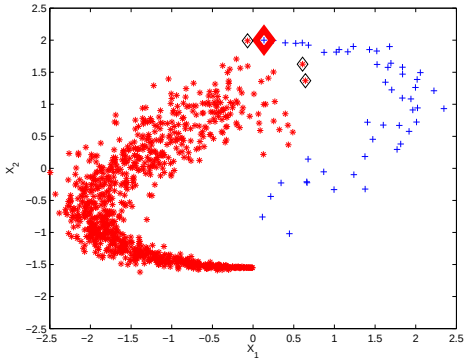

Figure 3: *Similar toy example with 1000 samples, with 50 points positive ('+'), and 950 negative ('*'). Now only one (!) label is given (big diamond), and the learning machine looks for a minimal cut such that one has at least $nB = 900$ negative labels. The figure displays the result: remarkably, most labels are predicted well (except the three as indicated by the diamonds).*

$+1$ exactly once. Nodes (v_i, v_j) which are closely related will have an associated solution $\hat{q}^i$ and $\hat{q}^j$ which coincides. Indeed, the regularization terms based on $\hat{q}^i$ and $\hat{q}^j$ will coincide (if v_i has the same class label as v_j for both runs). If the nodes v_i and v_j are loosely connected, the balancing constrained will enforce that v_i has a different class assigned in both $\hat{q}^i$ and $\hat{q}^j$ than v_j. It becomes clear that the balancing constraint regulates implicitly the size of each cluster: the higher the required amount of negative samples, the smaller the classes of positive points. There remain a couple of issues to be resolved for this implementation to be a practical successful strategy. The first is that previously, we relied on the fact that the solution $\hat{q}$ satisfies exactly the integer constraints $\hat{q} \in \{-1, 1\}^n$. Although occurring remarkably often, it appears not to be guaranteed a priori for all n runs. Related to this fact is that the original integer program can have multiple optima, disturbing somewhat the reasoning. The third point is that it becomes computational difficult to perform the n tasks if $n \gg 1000$. However, it is observed that the CPLEX implementation can effectively exploit the sparse structure of the matrices based on a classical labeling algorithm (see e.g. [19] and references). From a theoretical point of view, it remains a challenge to apply the generalization bound as in Theorem 1 based on Serfling's inequality. The main difficulty is found in the apparent disagreement of the implied sampling of only postive nodes, versus the assumption of random sampling the observed nodes.

	Class 1 (500)	Class 2 (500)
SGT [15]	11.18	10.96
SDP [11]	4.21	5.30
tSVM [8]	10.23	12.12
TGC eq. (21)	5.13	4.45
	Class 1 (950)	Class 2 (50)
SGT [15]	3.80	20.20
SDP [11]	5.01	1.00
tSVM [8]	5.3	17.23
TGC eq. (21)	4.55	0.88

Table 1: Numerical results of a benchmark study - expressed in the number of nodes in a class which are mispredicted. In the first case a 500-500 partition was generated, The second case considers datasets with a true unbalanced 950-50 partition. The algorithms were in both cases provided with exactly 2 opposite lables, as in this case the proposed algorithm performs clearly better than the remaining algorithms.

4 Experiments

Figure 2 gives a visual example of the TGC algorithm of eq. (21) at work on a two dimensional artificially constructed dataset of 1000 nodes. Only two nodes were assigned the labels 1 and -1. The graph between nodes was constructed as follows: two different nodes v_i and v_j were connected ($w_{ij} = 1$) when they belong to the 20 closest neighbors of either, and the value w_{ij} was set to zero otherwise. The algorithm found a global optimum where $\hat{q}_i$ was either 1 or -1 for all $i = 1, \ldots, n$. Figure 3 was constructed analogously, but using only 50 labeled nodes of the positive class. Imposing a balancing of 90% against the single (!) provided positive sample gave the displayed result. Table 1 gives results on both datasets in terms of number of misclassified labels *per class*. Three other existing algorithms for transductive inference were used for benchmarking purposes. At first, the medium size algorithm based on an SDP relaxation as discussed in [11] was used. Secondly, the results of Joachims graph transducer [15] based on a spectral relaxation was reported. Thirdly, we used a large scale refinement as in [8] based on the transductive SVM formulation [4].

5 Conclusions

This paper discusses a novel approach towards the task of transductive inference of the labels of a deterministic weighted graph. The derivation follows from the definition of an appropriate hypothesis, implementing the maximum margin principle. The relationship with a MINCUT approach, and a suitable generalization bound are developed. From a practical perspective, an efficient and intuitive convex approach is formulated, which is capable for handling datasets with over thou-

sand data-points. Extensions towards tasks with only positive labels, and fully unsupervised clustering problems are discussed. An current open question concerns the extension of the method to newly emerging graph nodes, and the handling of empirical observed graphs. We currently investigate the application and tuning of this approach in a large-scale task of information retrieval and in a specific task of gene prioritization.

Acknowledgments. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. Research supported by BOF PDM/05/161, FWO grant V 4.090.05N, IPSI Fraunhofer FgS, Darmstadt, Germany. (Research Council KUL): GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; (Flemish Government): (FWO): PhD/postdoc grants, projects, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, G.0302.07. research communities (ICCoS, ANMMM, MLDM); (IWT): PhD Grants,GBOU (McKnow), Eureka-Flite2 - Belgian Federal Science Policy Office: IUAP P5/22,PODO-II,- EU: FP5-Quprodis; ERNSI; - Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard. JS is a professor and BDM is a full professor at K.U.Leuven Belgium. This publication only reflects the authors' views.

References

[1] Y.S. Abu-Mostafa and J.-M. St. Jacques. Information capacity of the Hopfield model. *IEEE trans. on Inf. Theory*, 31:461–464, 1985.

[2] M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[3] Andreas Argyriou, Mark Herbster, and Massimilano Pontil. Combining graph laplacians for semi–supervised learning. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 67–74. MIT Press, Cambridge, MA, 2006.

[4] K.P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems 10*. MIT Press, Cambridge, MA, 1998.

[5] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 19–26. Morgan Kaufmann Publishers, 2001.

[6] A. Blum, J. Lafferty, M.R. Rwebangaria, and R. Reddy. Semi-supervised learning using randomized mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers, 2004.

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[8] O. Chapelle, B. Schölkopf, and A. Zien(Eds.), editors. *Semi-supervised Learning (In Press)*. MIT Press, Cambridge, MA, 2006.

[9] O. Chapelle, V. Vapnik, and J. Weston. Transductive inference for estimating values of functions. In S. Thrun, editor, *Advances in Neural Information Processing Systems 13*. MIT Press, Cambridge, MA, 2001.

[10] C. Cooper, M. Anthony, and G. Brightwell. The Vapnik-Chervonenkis dimension of a random graph. *Discrete Mathematics*, 138:43–56, 1995.

[11] T. De Bie and N. Cristianini. Convex methods for transduction. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[12] M. Fiedler. A property of eigenvectors of nonegative symmetric matrices and its application to graph theory. *Czech. Math. J.*, 25(100):619–633, 1975.

[13] M.X. Goemans and D.P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

[14] M. Grötschel, L. Lovasz, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 1988.

[15] T. Joachims. Transductive learning via spectral graph partitioning. *23e International Conference on Machine Learning (ICML)*, 2003.

[16] L. Lovasz. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25:1–7, 1979.

[17] R. El-Yaniv P. Derbeko and R. Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22:117–142, 2004.

[18] K. Pelckmans, J.A.K. Suykens, and B. De Moor. The kingdom-capacity of a graph: On the difficulty of learning a graph labelling. In *in Proc. of the workshop on Machine Learning on Graphs*, pages 1–8. TBA, Berlin, Germany, 2006.

[19] A. Schrijver. *Theory of Linear and Integer programming*. Wiley, 1988.

[20] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940., 1998.

[21] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE transactions on Pattern Recognition and Machine Intelligence*, 22(8), aug. 2000.

[23] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.

[24] X. Zhu and Z. Gharamani. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU*, 2002.