Approximation Algorithms for Classification Problems with Pairwise Relationships: Metric Labeling and Markov Random Fields

Jon Kleinberg* Éva Tardos†

Abstract

In a traditional classification problem, we wish to assign one of k labels *(or classes) to each of n* objects*, in a way that is consistent with some observed data that we have about the problem. An active line of research in this area is concerned with classification when one has information about* pairwise relationships *among the objects to be classified; this issue is one of the principal motivations for the framework of Markov random fields, and it arises in areas such as image processing, biometry, and document analysis. In its most basic form, this style of analysis seeks a classification that optimizes a combinatorial function consisting of* assignment costs *— based on the individual choice of label we make for each object — and* separation costs *— based on the* pair *of choices we make for two "related" objects.*

We formulate a general classification problem of this type, the metric labeling problem*; we show that it contains as special cases a number of standard classification frameworks, including several arising from the theory of Markov random fields. From the perspective of combinatorial optimization, our problem can be viewed as a substantial generalization of the multiway cut problem, and equivalent to a type of* uncapacitated quadratic assignment problem*.*

We provide the first non-trivial polynomial-time approximation algorithms for a general family of classification problems of this type. Our main result is an $O(\log k \log\log k)$-approximation algorithm for the metric labeling problem, with respect to an arbitrary metric on a set of k labels, and an arbitrary weighted graph of relationships on a set of objects. For the special case in which the labels are endowed with the uniform metric *— all distances are the same — our methods provide a 2-approximation.*

*Department of Computer Science, Cornell University, Ithaca NY 14853. Email: kleinber@cs.cornell.edu. Supported in part by an Alfred P. Sloan Research Fellowship, an ONR Young Investigator Award, and NSF Faculty Early Career Development Award CCR-9701399.

†Department of Computer Science, Cornell University, Ithaca NY 14853. Email: eva@cs.cornell.edu. Research partially supported by NSF grant CCR-9700163 and ONR grant N00014-98-1-0589.

1 Introduction

Classification Problems with Pairwise Relationships. A fundamental issue in statistics, pattern recognition, and machine learning is that of *classification*. At a very high level, we can view a classification problem as consisting of a set P of *objects* to be classified, and a set L of *labels* (the classes); the goal is to assign a label to each object in a way that is consistent with some observed data that we have about the problem [10, 19].

The "observed data" in such a problem can take a wide variety of different forms; our interest here is in classification problems for which this data involves pairwise relationships among the objects to be classified. Such problems have a rich history in statistics, where they arise naturally from formalisms such as the theory of *Markov random fields* [4, 14, 28, 29, 33]; they have formed some of the core analytical frameworks in areas including image processing [5, 15, 20, 23] and biometric analysis [4], and they have found applications in many other fields, including language modeling [18] and the categorization of hypertext documents [12].

To motivate the formulation, we consider, as an illustrative example, the well-studied problem of "restoring" an image that has been degraded by noise [5, 23]. We are given a large grid of pixels; each pixel has a "true" intensity that we are trying to determine, and an "observed" intensity that is the result of corruption by noise. We would like to find the best way to label each pixel with a (true) intensity value, based on the observed intensities. Our determination of the "best" labeling is based on the trade-off between two competing influences: We would like to give each pixel an intensity close to what we have observed; and — since real images are mainly smooth, with occasional boundary regions of sharp discontinuity — we would like spatially neighboring pixels to receive similar intensity values. Below, we give a self-contained formulation of an objective function that naturally captures the trade-off between these issues; we will mention several models that our formulation generalizes, and survey (in Section 2) the connections between

our model and the framework of Markov random fields.

Formulations of this type have been applied to a wide range of other image processing tasks as well; for example, in image segmentation and interpretation, the objects are pixels and the labels are logical "regions" into which we want to partition the image (see e.g. [15, 20]). Labeling problems with pairwise relationships also arise naturally in the other applications discussed above. For example, in biometry, one interprets observations gathered over many points in a geographical region under the assumption that the "true" properties of nearby points will be similar [4]; in hypertext categorization, it is possible to enhance the power of heuristics for classifying individual documents using relationships between pairs of documents that are connected by a hyperlink [12].

The Metric Labeling Problem. One can make these issues precise in the following way. Consider a set P of n objects that we wish to classify, and a set L of k possible labels. A *labeling* of P over L is simply a function $f : P \to L$; we choose a label for each object. The quality of our labeling is based on the contribution of two sets of terms.

- For each object $p \in P$, we have an estimate of its likelihood of having each label $a \in L$. The determination of these likelihoods is assumed to come from some heuristic pre-processing of the data; for our purposes, it results in a non-negative *cost* $c(p, a)$ for assigning label a to object p.

- We also have information about pairwise relationships among the objects; if p and q are deemed to be *related*, then we would like them to have the same (or "similar") labels. Thus, we have a graph G over the vertex set P, with edge set E indicating the objects that are related; each edge $e = (p, q)$ will have a non-negative weight w_e, indicating the strength of the relation. Moreover, as in our example with pixels and intensities, certain pairs of labels are more similar than others; so we impose a metric $d(\cdot, \cdot)$ on the set L of labels, with larger distance values indicating less similarity. If we assign label a to object p and label b to object q — and $e = (p, q)$ is an edge of G — then we pay a cost of $w_e d(a, b)$.

Thus, the *total cost* of a labeling f is given by

$$Q(f) = \sum_{p \in P} c(p, f(p)) + \sum_{e=(p,q) \in E} w_e d(f(p), f(q)).$$

The *metric labeling problem* asks for a labeling of minimum total cost.

Traditionally, standard local search methods have been applied to labeling problems of this form. Recently, combinatorial flow-based algorithms have been playing an increasingly large role in computer vision (see e.g. [7, 25, 32]). Boykov, Veksler, and Zabih [7, 8, 9] have developed flow-based local search heuristics for labeling problems with good performance in practice.

In this paper, we provide a polynomial-time approximation algorithm for the general metric labeling problem. Our main result is an $O(\log |L| \log \log |L|)$-approximation algorithm for this problem, with respect to an arbitrary graph on the set of objects, and an arbitrary metric on the set of labels. For the special case of the *uniform metric* — when all distances are equal to 1 — our methods provide a 2-approximation algorithm.

Related Problems. The reader may notice a connection between the metric labeling problem and the well-studied *multiway cut problem* [11, 16, 17, 21, 26], in which one must find a partition of a graph into k sets in a way that is consistent with an initial assignment of certain vertices to these sets, and which cuts as few edges as possible. In the multiway cut problem, we can view the k sets as intuitively representing the k labels of a classification problem; however, the objective function in multiway cut is too simple to capture the issues raised in the applications above. First, we do not wish simply to measure the number of neighboring objects that receive different labels; the penalty for labeling two objects differently should depend on the identities of the labels they receive. Thus, in image restoration, there should be a smaller penalty for assigning neighboring pixels similar (but unequal) intensities than for assigning them very different intensities. Additionally, we note that in the standard formulation of the multiway cut problem, one assumes that all nodes of the graph are either "pre-classified" into one of the k sets, or totally "un-classified"; in applications, however, we wish to use prior cost estimates to implicitly define "soft" initial assignments of labels to objects. Metric labeling is thus a substantial generalization of the multiway cut problem, designed to incorporate these fundamental issues.

In fact, the metric labeling problem can perhaps better be viewed as the *uncapacitated quadratic assignment* problem. In the quadratic assignment problem, a heavily studied topic in discrete optimization, one must find a matching between a set of n given *activities* to n *locations* in a metric space so as to minimize a sum of *assignment costs* for activities, and *flow costs* for activities that "interact." We will discuss this problem in more detail in the subsequent section; for now, we simply note that our problem can be obtained from the quadratic assignment by dropping the requirement that at most one activity can be sited at a given location. The

activities then correspond to objects, and the locations to labels, in the metric labeling problem.

Finally, how does our formulation of the metric labeling problem relate to the applications discussed above? In the next section, we will show how the optimum of a metric labeling problem, as we have described it here, is equivalent to the optimal configuration for a certain general class of Markov random fields. Metric labeling contains, as a special case, the *Generalized Potts model* used by Boykov, Veksler, and Zabih [7, 31] to obtain strong results for a range of image analysis tasks; it also contains, as special cases, several other basic models proposed in the areas of image processing and biometry [4, 15, 20, 24].

The Present Work: Summary of Results. We can show the metric labeling problem to be NP-complete even when the underlying graph G is an $n \times n$ grid. (See the proof reported in [8].) However, despite the fact that it contains a number of widely used classification frameworks as special cases, there have been no previous algorithms for metric labeling that provided non-trivial approximation guarantees.

Our main result is an $O(\log k \log\log k)$-approximation algorithm for the metric labeling problem, with respect to an arbitrary graph on the set of objects, and an arbitrary metric on the set of labels. Recall that k denotes the size of the label set L, and hence our approximation guarantee is independent of n, the size of the set of objects.

As the first main step in obtaining our approximation algorithm for the general problem, we develop a 2-approximation for a special case of metric labeling that we refer to as the *uniform labeling* problem. This is the case in which the metric $d(\cdot,\cdot)$ on the set L of labels is *uniform*: for labels a and b, $d(a,b)$ is equal to 0 if $a = b$ and equal to 1 otherwise. The costs $\{c(p,a)\}$ remain arbitrary. The uniform labeling problem is of interest in its own right, as it incorporates one of the key issues discussed above: "soft" initial assignments of labels to objects through costs. It is equivalent to the Generalized Potts model [7, 31]; it also generalizes certain other labeling frameworks (e.g. [24]).

For the special case of uniform labeling, we have learned that Sanjeev Arora has independently obtained a constant-factor approximation algorithm [1]; and Julia Chuzhoy [13] has been able to improve on our approximation ratio of 2 for the case of $k = 3$ and $k = 4$ labels to $4/3$ and $11/6$ respectively. Concurrently with the present work, Boykov, Veksler, and Zabih have shown that for one of their flow-based heuristics for the uniform labeling problem, the objective function value at any local optimum is within a factor of 2 of the global optimum [9].

The Present Work: Labeling, Cuts, and LP Relaxations. We have already indicated connections between metric labeling and the multiway cut problem at an informal level. Indeed, it is easy to see that multiway cut is already a special case of the uniform labeling problem: given an instance of the multiway cut problem with a graph $G = (V, E)$ and *terminals* $t_1, \ldots, t_k \in V$ that we wish to disconnect by deleting edges, we can define the following instance of the uniform labeling problem. The objects will be the nodes V, with the graph of relationships equal to G; the label set is $\{1, 2, \ldots, k\}$. All label assignment costs are set to 0, except that we set $c(t_i, j)$ equal to ∞ (or a very large positive quantity) if $i \neq j$. This forces t_i to receive label i, and we thus assign the remaining labels so as to minimize the number of edges with unequal labels at their ends.

The connection between optimization over labelings and optimization over cuts in a graph has been known for some time in work on Markov random fields. Greig, Porteous, and Seheult [5, 24] showed that the case of $k = 2$ labels is polynomially solvable as a two-terminal minimum cut problem. Boykov, Veksler, and Zabih developed the connection between uniform labelings, with $k > 2$ labels, and multiway cuts in a graph, showing a direct reduction from labelings to multiway cuts [7]. However, their reduction required the use of edges of enormously large weight, and so it is not approximation-preserving.

In fact, it is not difficult to see that once we allow highly variable costs $c(\cdot,\cdot)$ (or infinite costs), direct extension of either the *Isolation Heuristic* of Dahlhaus et al. [17], or the algorithm of Calinescu, Karloff, and Rabani [11] no longer provides a reasonable approximation guarantee. These methods do not deal well with objects that *cannot* receive certain labels, and they can produce results that are arbitrarily far from optimal.

For the case of uniform labeling we use a linear programming relaxation naturally adapted from the one developed for the multiway cut problem by Calinescu, Karloff, and Rabani [11], and also studied in [16, 26]. Although the LP formulation in this special case directly generalizes that of [11], the presence of variable label assignment costs forces us to use a different rounding technique; and indeed, the integrality gap of the formulation now converges to 2, matching our upper bound. (In [11] the analogous formulation converged to a bound of at most 1.5.)

The main difficulties for us arise in moving up to the general metric labeling problem. One problem we face here is that we no longer know of any "natural" linear programming relaxation for the problem. Consequently, we focus first on the case in which the metric on L corresponds to a *hierarchically well-separated tree* [2]; here, we formulate a linear programming relaxation and develop a rounding algorithm whose performance guarantee is a constant depending inversely on the "separation parameter" of the underlying tree. Then, using Bartal's result that any finite metric space can be probabilistically approximated by hierarchically well-separated trees [2, 3], we are able to obtain

an algorithm for general metric spaces.

We note another way, not considered here, in which the connection between labelings and cuts has been exploited — for the identification of polynomially solvable special cases of the metric labeling problem. This issue has been investigated in recent work of Karzanov [27]. Boykov et al. [7] and Ishikawa and Geiger [25] show a direct reduction to two-terminal minimum cuts — and hence an efficient algorithm — in the special case of metric labeling when the label set is $L = 0, 1, \ldots, k$, and the distance between two labels i and j is $d(i, j) = |i - j|$.

2 Connections to Other Models

In this section, we summarize the relationship between the metric labeling problem and various well-studied problems in statistics and combinatorial optimization: in particular, we consider Markov random fields, the quadratic assignment problem, and metrical task systems. We begin developing our algorithms in Section 3, and the reader has the option of proceeding directly to this section without loss of continuity.

Markov Random Fields. For a given set of objects P and labels L, the theory of *random fields* provides a systematic means for converting the set of all possible labelings of P over L into a probability space. Thus, for every labeling f, the random field assigns it a probability $\Pr[f]$. We will make the assumption throughout this discussion that $\Pr[f] > 0$ for all labelings f.

As before, we have a graph $G = (P, E)$ on the set of objects. The random field is *Markovian* if the conditional probability of the label assignment at object p depends only on the label assignments at the neighbors of p in G. That is, a Markov random field (MRF) satisfies

$$\Pr[f(p)|\{f(q) : q \neq p\}] = \Pr[f(p)|\{f(q) : (p, q) \in E\}].$$

If we consider our initial example of image restoration, we see that the "local" nature of the dependencies in a Markov random field captures our intuitive notion that the label assigned to a given pixel depends on the labels of its neighboring pixels.

A theorem of Hammersley and Clifford [4] allows us to write the distribution of a Markov random field in essentially a closed form, as follows. Let $\mathcal{C}$ be the set of all cliques in G; for a clique $C \in \mathcal{C}$, we say that a *clique potential* associated with C is a function Γ_C that maps labelings of objects in C to real numbers. Then the Hammersley-Clifford theorem states that for a given Markov random field, there exists a set of clique potentials $\{\Gamma_C : C \in \mathcal{C}\}$ so that, for an arbitrary labeling f, $\Pr[f] = Z^{-1} \exp(-\sum_{C \in \mathcal{C}} \Gamma_C(f|C))$. Here Z is a normalizing constant, and the notation $f|C$ indicates the restriction of the labeling f to the objects in C. We can take the negative logarithm of this probability, obtaining a value that we call the *cost*, $\sum_{C \in \mathcal{C}} \Gamma_C(f|C)) + \ln Z$. Note that for purposes of minimization we can ignore the $(\ln Z)$ term, as it is the same for all labelings.

In the context of classification, suppose that we have data $f'(p)$ for each object $p \in P$, and we wish to decide on the most probable labeling f of P, given the data and an underlying MRF model. A standard approach is to find the f that maximizes the *a posteriori* probability $\Pr[f|f']$, which by Bayes' Law is proportional to $\Pr[f'|f] \cdot \Pr[f]$. Assuming that f' is obtained from f by the introduction of independent random noise at each object, the term $\Pr[f'|f]$ factors into $\prod_p \Pr[f'(p)|f(p)]$. The second term is given by the Hammersley-Clifford expansion in terms of cliques. Taking negative logarithms, we are seeking to minimize the following *energy function*

$$E(f) = \sum_p (-\ln \Pr[f'(p)|f(p)]) + \sum_{C \in \mathcal{C}} \Gamma_C(f|C)). \quad (1)$$

We can now see how the metric labeling problem arises from this formalism. In standard applications, one very frequently works with a restricted type of MRF satisfying two properties:

(i) *Pairwise interactions.* The MRF is defined purely by pairwise interactions among the objects; so Γ_C is a non-zero function only for cliques of size 2 (i.e. edges of G). We will denote these function by Γ_e, for $e \in G$; note that each is a function mapping $L \times L$ to $\mathbf{R}$.

(ii) *Homogeneity.* The MRF is spatially homogeneous, so each function Γ_e can be written as $w_e\Gamma$, where w_e is a non-negative edge weight, and the function Γ is the same for all edges.

We define a *metric Markov random field* to be an MRF that satisfies the above two properties, and for which the function Γ is a metric on the set of labels. Now, for an arbitrary metric MRF, we can define $c(p, a) = -\ln \Pr[f'(p)|a]$ and re-write the problem of minimizing the energy function defined in Equation (1) as follows:

$$\min_f E(f) = \min_f \sum_p c(p, f(p)) + \sum_{e=(p,q)} w_e \Gamma(f(p), f(q)),$$

where Γ is a metric. This is precisely the statement of the metric labeling problem.

Given that pairwise, homogeneous MRF's are ubiquitous in the application of MRF's to classification problems, we feel that our formulation of metric MRF's is quite general. It contains the Potts model [22, 31] as a special case, by setting all $w_e = 1$ and Γ to be the *uniform metric*. It contains

the Generalized Potts model of Boykov et al. [7] by setting Γ to be the uniform metric and allowing the weights w_e to be arbitrary.

Quadratic Assignment Problem. The quadratic assignment problem is one of the most well-studied facility location problems in operations research; see [30] for a comprehensive recent survey. We are given n *activities* and n *locations* in a metric space; we would like to site each activity at a distinct location in a way that minimizes an objective function consisting of the following types of terms.

• If we site activity i at location ℓ, we incur an operating cost $c(i, \ell)$.

• Moreover, the activities interact, and we wish to keep activities that interact closely near to one another. Thus, if $w_{ij} \geq 0$ measures the interaction between activities i and j, and $d(\ell, \ell')$ measures the distance between locations ℓ and ℓ', then we pay a cost of $w_{ij}d(\ell, \ell')$ for siting activity i at ℓ and j at ℓ'.

The traditional quadratic assignment is based on minimizing $\sum_i c(i, f(i)) + \sum_{i,j} w_{ij}d(f(i), f(j))$ over all *bijections* f. The metric labeling problem seeks to minimize precisely the same objective function over all *functions* f, for the locations can be viewed as the labels, and the set of interaction values $\{w_{ij}\}$ can be viewed as a weighted graph on the set of activities. Thus, metric labeling is equivalent to the *uncapacitated quadratic assignment problem*, in which multiple activities may be sited at the same location. In view of the discussion above, this uncapacitated version — like the traditional version — is NP-complete, due to the fact that activities have different costs for operating at different locations.

Metrical Task Systems. There is another, closely related way to view the metric labeling problem in terms of *metrical task systems*, an optimization problem formulated by Borodin, Linial, and Saks [6] in the context of on-line algorithms. In a metrical task system, we are given an ordered sequence of n *tasks* $t_1, \ldots, t_n$, and a set of k possible *states* in which to perform the tasks. There is a metric on the set of states.

For each task t_i and each state ℓ, there is a cost $c(t_i, \ell)$ to perform t_i in state ℓ. If we wish to move from state ℓ to state ℓ' before performing the next task t_{i+1}, we pay a cost of $d(\ell, \ell')$ — the distance between states ℓ and ℓ'. Thus, we have precisely the uncapacitated quadratic assignment problem (i.e. the metric labeling problem) for the special case in which the graph of interactions is a path: w_{ij} is equal to 1 if $|i-j| = 1$ and 0 otherwise. The optimal state sequence can be determined in this special case by dynamic programming if the set of costs $\{c(t_i, \ell)\}$ is known in advance; Borodin et al. were concerned with the *on-line* setting in which these costs only become known incrementally.

Thus, the general metric labeling problem seeks to find the optimal set of states in which to process a set of tasks, given that the interactions among the tasks are arbitrary and not organized in a linear path.

3 The Uniform Labeling Problem

Before considering the general metric labeling problem, we develop an approximation algorithm for the uniform labeling problem. Thus, we consider the case of uniform distances, in which each pair of different labels is at distance 1.

The uniform labeling problem can be expressed as the following integer program. We use nonnegative variables x_{pa} for each object p and label a such that $\sum_{a \in L} x_{pa} = 1$; i.e. for integer-valued x we use $x_{pa} = 1$ to denote that $f(p) = a$ — object p is assigned label a. The label cost of an object p is then expressed as $\sum_{a \in L} c(p, a)x_{pa}$. Consider two objects p and q. The distance between the assigned labels can be expressed as $d(f(p), f(q)) = \frac{1}{2}\sum_{a \in L} |x_{pa} - x_{qa}|$. Hence the separation cost for an edge $e = (p, q)$ is expressed as $\frac{1}{2}w_e \sum_{a \in L} |x_{pa} - x_{qa}|$. We can write this as an integer linear program by introducing variables z_e for an edge $e = (p, q)$ to express the distance between the labels assigned to objects p and q, and using z_{ea} to express the absolute value for $a \in L$.

(UIP)

$$
\begin{array}{lll}
\text{Min} & \sum_{e \in E} w_e z_e & \\
 & + \sum_{p \in P, a \in L} c(p, a)x_{pa} & \\
\text{s.t.} & \sum_{a \in L} x_{pa} = 1 & p \in P \\
 & z_e = \frac{1}{2}\sum_{a \in L} z_{ea} & e \in E \\
 & z_{ea} \geq x_{pa} - x_{qa} & e = (p, q), a \in L \\
 & z_{ea} \geq x_{qa} - x_{pa} & e = (p, q), a \in L \\
 & x_{pa} \in \{0, 1\} & p \in P, a \in L.
\end{array}
$$

We create the linear programming relaxation of (UIP) by replacing the integrality constraints with $x_{pa} \geq 0$; we will call this (ULP). We say that a set of non-negative variables $\{x_{pa} : p \in P, a \in L\}$ is a *fractional labeling* if $\sum_{a \in L} x_{pa} = 1$ for all objects $p \in P$. The *assignment cost* of the fractional labeling x is $\sum_{p \in P, a \in L} c(p, a)x_{pa}$ and the *separation cost* is $\sum_{e=(p,q) \in E} \left[\frac{1}{2}w_e \sum_{a \in L} |x_{pa} - x_{qa}|\right]$.

We give a randomized method for rounding the fractional solution to (ULP) to integers, losing at most a factor of 2 in the objective function. The idea is to use the x_{pa} values as probabilities that object p gets assigned to label a. However, the simplest randomized rounding scheme — assigning each object p independently to a label with probabilities equal to its x_{pa} values — works terribly. For if $e = (p, q)$

and p and q have the same set of values $\{x_{pa}\}, \{x_{qa}\}$, then $z_e = 0$ in (ULP), and so we must assign p and q the same label if we are to use a term-by-term analysis for the approximation guarantee.

We therefore adopt a more careful scheme. We assign labels to vertices in *phases*; initially each vertex has no label assignment. In a single phase, we pick a label a uniformly at random, and a real number α uniformly in $[0, 1]$. For each object p that currently has no label assignment, we give it the label a if $\alpha \leq x_{pa}$. We iterate these phases until each vertex has a label assignment.

Until object p has a label assignment, the probability it is assigned label a in a given phase is precisely x_{pa}/k. Thus we have

Lemma 3.1 *The probability that an unassigned object p receives a label assignment in a single phase is precisely $1/k$. Further, over all phases, the probability that an object p is assigned label a is precisely x_{pa}.*

We say that the two ends of an edge are separated, or "split," by this process if they receive different labels. We now bound the probability that the ends of a given edge are split.

Lemma 3.2 *For an edge $e = (p, q)$, the probability that the two objects p and q are separated by the process is at most $2z_e = \sum_{a \in L} |x_{pa} - x_{qa}|$.*

Proof. Objects p and q will not be split if they are assigned a label in the same phase; thus, we can upper-bound the probability that p and q are split by the quantity

$$\frac{\Pr[\text{exactly one of } p, q \text{ assigned in one iteration}]}{\Pr[\text{at least one of } p, q \text{ assigned in one iteration}]}.$$

The numerator is equal to

$$\frac{1}{k} \sum_{a \in L} |x_{pa} - x_{qa}| = \frac{\sum_{a \in L} z_{ea}}{k} = \frac{2z_e}{k}.$$

The denominator is at least the probability that object p is assigned, which is $\frac{1}{k}$ by Lemma 3.1. So the probability that edge $e = (p, q)$ is split is at most $2z_e$ as claimed. ■

Now we are ready to give the approximation theorem.

Theorem 3.3 *Let x be a solution of the linear program (ULP) with assignment cost c_{LP} and separation cost w_{LP}. The randomized rounding procedure described above finds a labeling whose expected assignment cost is c_{LP}, and whose expected separation cost is at most $2w_{LP}$.*

Proof. By Lemma 3.1 the probability that object p is assigned label a is x_{pa}, hence the expected contribution of q to the objective function is $\sum_{a \in L} c_{pa} x_{pa}$, precisely its contribution in (ULP). Hence the total expected assignment cost over all objects is c_{LP}.

What is the contribution of edge $e = (p, q)$? This is precisely w_e times the probability that its ends are split. By Lemma 3.2 this probability is at most $2z_e$, hence the expected contribution of edge e is at most $2w_e z_e$, which is twice its contribution in (ULP). Hence the total expected assignment cost over all objects is $2w_{LP}$. ■

Using the optimal fractional labeling x we get a labeling with expected value within a factor of 2 of optimal. In the full paper we will show how to derandomize the procedure using the method of conditional probabilities.

Corollary 3.4 *If x is an optimal solution to (ULP) then the rounding provides a 2-approximation for the uniform labeling problem.*

A Tight Example. We now observe that the factor of two provided in the theorem is essentially tight; as noted above, this forms an interesting contrast with the results of [11, 16, 26] for the multiway cut problem. The following example shows that the integrality gap of our LP formulation is close to 2. Let G be the complete graph on k nodes with all edge weights w_e equal to 1. We set all assignment costs equal to 0, except that we set $c(i, i)$ to infinity for all i. The optimum integer solution is to select an arbitrary i, assign all nodes $j \neq i$ label i, and assign node i some other label. The cost of this assignment is $k - 1$ for separating the $k - 1$ edges out of node i. An optimal LP solution is obtained by setting $x_{ij} = \frac{1}{k-1}$ for all $i \neq j$. The separation cost of each edge in this assignment is $\frac{1}{k-1}$, and the total over the $k(k-1)/2$ edges is $k/2$. This is an integrality gap of $2(k-1)/k$.

4 The Metric Labeling Problem

Next we consider the metric labeling problem. Interestingly, we will be able to obtain an approximation algorithm despite the fact that we do not know of any "natural" linear programming relaxation for the problem.

We will make use of Bartal's result that any finite metric space can be probabilistically approximated by hierarchically well-separated tree metrics [2, 3]. Consequently, we focus first on the case in which the metric on L corresponds to an r-hierarchically well-separated tree [2]; here, we formulate a linear programming relaxation and develop a rounding algorithm whose performance guarantee is a constant slightly larger than 2, depending on the separation parameter r of the underlying tree.

An *r-hierarchically well-separated tree metric* (r-HST) is a tree with lengths on its edges satisfying the following properties.

(i) The edge lengths from any node to each of its children is the same.

(ii) The edge lengths along any path from the root to a leaf decrease by at least a factor of r.

(iii) The metric distance between nodes of the tree is the sum of the edge lengths on the unique path between them.

We make the following additional assumptions. First, we will require that $r > 2$; and second, we will require that the labels L correspond to leaves of the tree. Note that the uniform labeling problem discussed in the previous section is the special case of this setting in which the tree is a star with unit-length edges.

For two subtrees we use "$T' < T$" to denote the statement that T' is a subtree of T. For such a subtree T, let ℓ_T denote the length of the edge leading upward from the root of T, and let $L(T)$ denote the set of leaves of the tree T. As our first main step, we give a $\left(2 + \frac{4}{r-2}\right)$-approximation algorithm for the metric labeling problem on an r-HST that repeatedly uses the 2-approximation algorithm of the previous section.

The metric labeling problem for tree metrics can be expressed as an integer program using the same variables x_{pa} that we used in the uniform case. We have nonnegative variables x_{pa} for each object p and label a such that $\sum_{a \in L} x_{pa} = 1$; i.e., for integer-valued x we use $x_{pa} = 1$ to denote that $f(p) = a$. To express the separation costs we define $x_p(T) = \sum_{a \in L(T)} x_{pa}$. When x takes 0-1 values, then $x_p(T)$ is 1 if and only if the object p is assigned to a label in T. The distance $d(f(p), f(q))$ for an edge $e = (p, q)$ of our graph G can be expressed as $d(f(p), f(q)) = \sum_T \ell_T |x_p(T) - x_q(T)|$. The separation cost of an edge $e = (p, q)$ is w_e times this distance. We can write this as a linear program by introducing variables z_e for the distances, and z_{eT} to express each of the absolute values.

(TIP)

$$
\begin{array}{lll}
\text{Min} & \sum_e w_e z_e & \\
 & + \sum_{p \in P, a \in L} c(p,a) x_{pa} & \\
\text{s.t.} & \sum_{a \in L} x_{pa} = 1 & p \in P \\
 & z_e = \sum_T \ell_T z_{eT} & e \in E. \\
 & z_{eT} \geq x_p(T) - x_q(T) & e = (p,q), \text{ and subtree } T. \\
 & z_{eT} \geq x_q(T) - x_p(T) & e = (p,q), \text{ and subtree } T. \\
 & x_{pa} \in \{0, 1\} & p \in P, a \in L.
\end{array}
$$

We create the linear programming relaxation of (TIP) by replacing the integrality constraints with $x_{pa} \geq 0$; we will call this (TLP). For tree metrics we use $\sum_e [w_e \sum_T \ell_T |x_p(T) - x_q(T)|]$ as the *separation cost* of the fractional labeling x. As before, we will use a fractional labeling x as a probability. Note that if x_{pa} is the probability that object p is assigned to label a then the probability that object p is assigned a label in the subtree T is exactly $x_p(T)$.

We define the randomized rounding algorithm as follows. We consider first the top level of the tree. Assume that the root has m children, and let $T^1, \ldots, T^m$ denote the subtrees rooted at the children. We use the uniform labeling algorithm to assign each object to one of the subtrees T^i, where object p has probability $x_p(T^i)$ to be assigned to subtree T^i. Thus, at the end of this step, each object has been assigned to one of the subtrees, but not yet placed *within* this subtree. Once object p has been assigned to subtree T^i, we modify the fractional labeling accordingly by setting $\bar{x}_{pa} = 0$ for all labels a that are not in the subtree T^i, and $\bar{x}_{pa} = x_{pa}/x_p(T^i)$ for all labels $a \in L(T^i)$. The objects assigned to each subtree will now be treated as separate subproblems, and we use the algorithm recursively on each of the subtrees T^i to obtain the labeling.

To analyze the top level of this algorithm, we use the following property that was proved as Lemma 3.2 in the previous section.

Lemma 4.1 *The probability that an object is assigned to the subtree T^i is $x_p(T^i)$. For each edge $e = (p, q)$ the probability that p and q are assigned to different trees is at most $\sum_{i=1}^m |x_p(T^i) - x_q(T^i)|$.*

When we consider a set of objects that have all been assigned to the same subtree, we have the problem that their fractional separation costs have changed. We bound this in Lemma 4.3. We first need the following useful fact; the proof is based on summing over all levels of the tree, and using the property that edge lengths decrease exponentially as we go down the tree.

Lemma 4.2 *Let x be a fractional labeling over a tree, let p be an object, and let T be a subtree. Then*

$$\sum_{T' < T} \ell_{T'} \, x_p(T') \leq \frac{1}{r-1} \cdot \ell_T x_p(T).$$

Proof. If t is the root of T, we say that a subtree is at *level* j if its root is reachable by a path leading j edges downward from t. Note that we have

$$\sum_{T' \in level \; j} x_p(T') \leq x_p(T)$$

for all j. Thus,

$$\sum_{T' < T} \ell_{T'} \, x_p(T') = \sum_j \sum_{T' \in level \; j} \ell_{T'} \, x_p(T')$$

$$\leq \sum_j r^{-j}\, \ell_T \sum_{T' \in level\ j} x_p(T')$$
$$\leq \sum_j r^{-j}\, \ell_T\, x_p(T)$$
$$\leq \frac{1}{r-1} \cdot \ell_T\, x_p(T). \qquad \blacksquare$$

Lemma 4.3 *Let $e = (p,q)$ be an edge and assume that $x_p(T^i) \leq x_q(T^i)$. If both p and q are assigned to subtree T^i then the new fractional separation cost is at most*

$$\frac{w_e}{x_p(T^i)} \left[\sum_{T < T^i} \ell_T |x_p(T) - x_q(T)| + \frac{1}{r-1} \ell_{T^i} |x_p(T^i) - x_q(T^i)| \right].$$

Proof. The new fractional separation cost for edge e is $w_e \sum_T \ell_T |\bar{x}_p(T) - \bar{x}_q(T)|$, and if p and q are both assigned to T^i then the sum can be written as

$$\begin{aligned}
&\sum_{T<T^i} \ell_T |\bar{x}_p(T) - \bar{x}_q(T)| \\
&= \sum_{T<T^i} \ell_T \left| \frac{x_p(T)}{x_p(T^i)} - \frac{x_q(T)}{x_q(T^i)} \right| \\
&= \sum_{T<T^i} \ell_T \left| \left(\frac{x_p(T)}{x_p(T^i)} - \frac{x_q(T)}{x_p(T^i)} \right) + \left(\frac{x_q(T)}{x_q(T^i)} - \frac{x_q(T)}{x_p(T^i)} \right) \right| \\
&\leq \sum_{T<T^i} \ell_T \frac{1}{x_p(T^i)} |x_p(T) - x_q(T)| + \sum_{T<T^i} \left[\frac{1}{x_p(T^i)} - \frac{1}{x_q(T^i)} \right] \ell_T\, x_q(T) \\
&\leq \frac{1}{x_p(T^i)} \left[\sum_{T<T^i} \ell_T |x_p(T) - x_q(T)| \right] + \frac{1}{r-1} \ell_{T^i} x_q(T^i) \left[\frac{1}{x_p(T^i)} - \frac{1}{x_q(T^i)} \right] \\
&= \frac{1}{x_p(T^i)} \left[\sum_{T<T^i} \ell_T |x_p(T) - x_q(T)| + \frac{1}{r-1} \ell_{T^i} (x_q(T^i) - x_p(T^i)) \right]. \qquad \blacksquare
\end{aligned}$$

We can now prove the main theorem about labeling with hierarchically well-separated tree metrics.

Theorem 4.4 *Let x be a fractional labeling of objects P over an r-hierarchically well-separated tree metric (with $r > 2$) whose leaves are the labels. Let c_{LP} and w_{LP} be respectively the assignment and separation cost of the fractional assignment x. The randomized rounding procedure described above finds a labeling whose expected assignment cost is c_{LP}, and whose expected separation cost is at most $\left(2 + \frac{4}{r-2}\right) w_{LP}$.*

Proof. We prove both statements about the expected cost by induction on the depth of the tree. As before, let $T^1, \ldots, T^m$ denote the subtrees rooted at the children of the root.

First consider the statement about the expected assignment costs. We claim that the probability that an object p is assigned a label a is exactly x_{pa}. This will imply the statement about the expected assignment cost. Consider a top level subtree. For an object p the probability that p gets assigned to a subtree T^i is $x_p(T^i)$, and if p is assigned to the subtree T^i then the fractional assignments are rescaled to be $\bar{x}_{pa} = \frac{1}{x_p(T^i)} x_{pa}$ for each label a in the subtree T^i. By induction, once an object p was assigned to subtree T^i it gets a label $a \in L(T)$ with probability $\bar{x}_{pa}$, and hence the probability that object p is assigned a label a in $L(T^i)$ is $x_p(T^i)\bar{x}_{pa} = x_{pa}$ as claimed.

Next consider the expected separation cost for an edge $e = (p,q)$. The fractional separation cost is $w_e \sum_T \ell_T |x_p(T) - x_q(T)|$. We compute the expected separation cost by considering separately the event $\mathcal{E}'$ that p and q are separated at the top level, and the events $\mathcal{E}_i$ that p and q are both assigned to the same subtree T^i. Let $h(p,q)$ be the random variable so that $w_e h(p,q)$ is the separation cost of edge $e = (p,q)$. Then the expected separation cost is $w_e \mathrm{E}[h(p,q)]$, and we can write $\mathrm{E}[h(p,q)]$ as follows.

$$\mathrm{E}[h(p,q)] = \mathrm{E}[h(p,q)|\mathcal{E}'] \cdot \Pr[\mathcal{E}'] + \sum_{i=1}^{m} \mathrm{E}[h(p,q)|\mathcal{E}_i] \cdot \Pr[\mathcal{E}_i].$$

The probability $\Pr[\mathcal{E}']$ that the edge gets separated at the top level is at most $\sum_{i=1}^m |x_p(T^i) - x_q(T^i)|$. The separation cost in this case is bounded by the diameter of the tree, which — using the hierarchical separation property — is at most $2r/(r-1)\ell_{T^i}$ for any i. So the first term of the above sum is bounded by

$$\mathrm{E}[h(p,q)|\mathcal{E}'] \cdot \Pr[\mathcal{E}'] \leq \frac{2r}{r-1} \sum_{i=1}^{m} \ell_{T^i}\, |x_p(T^i) - x_q(T^i)|.$$

If both objects are assigned to the same subtree T^i then by induction the expected separation cost is at most $2r/(r-2)$ times the fractional separation cost of the scaled solution. The probability that both objects are assigned to the same subtree T^i is at most $\min(x_p(T^i), x_q(T^i))$, and so using Lemma 4.3 we can bound $\mathrm{E}[h(p,q)|\mathcal{E}_i] \cdot \Pr[\mathcal{E}_i]$ as follows.

$$\mathrm{E}[h(p,q)|\mathcal{E}_i] \cdot \Pr[\mathcal{E}_i]$$

$$\leq \frac{2r}{r-2}\left[\sum_{T<T^i}\ell_T|x_p(T)-x_q(T)| + \frac{1}{r-1}\ell_{T^i}|x_p(T^i)-x_p(T^i)|\right].$$

The final bound on $\mathrm{E}[h(p,q)]$ then follows.

$$\begin{aligned}\mathrm{E}[h(p,q)] &\leq \frac{2r}{r-1}\sum_{i=1}^{m}\ell_{T^i}|x_p(T^i)-x_q(T^i)| \\ &\quad + \frac{2r}{r-2}\sum_{i=1}^{m}\left[\sum_{T<T^i}\ell_T|x_p(T)-x_q(T)| \right. \\ &\quad \left. + \frac{1}{r-1}\ell_{T^i}|x_p(T^i)-x_p(T^i)|\right] \\ &= \frac{2r}{r-2}\sum_{T}\ell_T|x_p(T)-x_q(T)|. \quad \blacksquare\end{aligned}$$

We now apply Bartal's result on probabilistic approximation of metric spaces to address the labeling problem with general metrics. Bartal's result states that for every metric d there is a set of r-hierarchically well-separated tree metrics $\{d_T\}$ and probability distribution $\{p_T\}$, so that $d_T(p,q) \geq d(p,q)$ for all T and all pairs of objects p,q; and yet the expected p-q distance in a randomly selected tree is not much bigger than $d(p,q)$. More precisely, $\sum_T p_T d_T(p,q) \leq O(r\log k\log\log k)d(p,q)$. We will use this result to obtain an $O(\log k\log\log k)$-approximation algorithm for the labeling problem in general metric spaces.

We first show that at an additional loss of a factor of $2+1/r$ we may assume that the well-separated trees in Bartal's theorem have all labels at the leaves — that is, the internal nodes of the trees are not labels.

Lemma 4.5 *Let d_T denote an r-HST distance function generated by the tree T. We can modify the tree T to obtain an r-HST distance function generated by the tree T' where all labels are at leaves, and for any two labels a and b we have $d_T(a,b) \leq d_{T'}(a,b) \leq (2+1/r)d_T(a,b)$.*

Proof. For each internal node a of T that is a label we replace this node by a new node a', and slide the label a down one level to become a new child of a'. The biggest change in the distance occurs if a node a and a child b of node a are both labels. Let $\ell = d_T(a,b)$. Then a is pushed down to be at a distance ℓ from the new node a', and b is pushed down from a new node b'. If the tree is an r-HST then the new distance $d_{T'}(b,b') \leq \ell/r$, and so we have that $d_{T'}(a,b) \leq 2\ell + \ell/r = (2+1/r)d_T(a,b)$. $\blacksquare$

It is now easy to show that the following randomized algorithm provides an expected performance guarantee of $O(\log k\log\log k)$: We probabilistically approximate the metric using 3-HST's, choose a random tree T according to the distribution $\{p_T\}$, and run our algorithm above on the metric induced by this tree. In the full version of the paper, we give a derandomized version of the overall algorithm.

Theorem 4.6 *There is a randomized polynomial time $O(\log k\log\log k)$-approximation algorithm for the metric labeling problem where k is the number of labels.*

Proof. Let d be a metric on the set of labels L. By Bartal's tree-metric approximation, and the above lemma we can find a set of 3-hierarchically well-separated tree-metrics that approximate the metric d. Let $\mathcal{T}$ denote the set of tree metrics, and p_T for $T \in \mathcal{T}$ be a probability distribution on the trees so that for each pair of labels $a,b \in L$ and each tree $T \in \mathcal{T}$ we have that $d(a,b) \leq d_T(a,b)$ and $\sum_{T\in\mathcal{T}} p_T d_T(a,b) \leq O(\log k\log\log k)d(a,b)$.

The algorithm is as follows. We choose a tree metric $T \in \mathcal{T}$ choosing the tree T with probability p_T, and run the algorithm for the tree metric defined by T.

Let $\mathcal{O}$ denote the cost of an optimal labeling for the metric d. Let $\mathcal{O}_T$ for $T \in \mathcal{T}$ denote the cost of the same labeling under the tree metric T. By the definition of the tree-metric approximation we have that $\sum_{T\in\mathcal{T}} p_T\mathcal{O}_T \leq O(\log k\log\log k)\mathcal{O}$.

Let the random variable $\mathcal{A}$ denote the cost using metric d of the labeling produced by the algorithm, and let $\mathcal{A}_T$ denote the cost of the solution found in the tree metric T. By the definition of the approximation we have that $\mathcal{A} \leq \mathcal{A}_T$ for all $T \in \mathcal{T}$.

We claim that the expected cost of the resulting labeling $\mathrm{E}[\mathcal{A}]$ is at most $O(\log k\log\log k)\mathcal{O}$. Let $\mathcal{E}_T$ denote the event that tree T is selected. We have that $\Pr(\mathcal{E}_T) = p_T$ and that $\mathrm{E}[\mathcal{A}_T|\mathcal{E}_T] \leq O(1)\mathcal{O}_T$. Using this we get the following.

$$\begin{aligned}\mathrm{E}[\mathcal{A}] &= \sum_{T\in\mathcal{T}}\mathrm{E}[\mathcal{A}|\mathcal{E}_T]\cdot Pr(\mathcal{E}_T) \\ &\leq \sum_{T\in\mathcal{T}} p_T\mathrm{E}[\mathcal{A}_T|\mathcal{E}_T] \\ &\leq O(1)\sum_{T\in\mathcal{T}} p_T\mathcal{O}_T \\ &= O(1)\sum_{T\in\mathcal{T}} p_T\mathcal{O}_T \\ &= O(\log k\log\log k)\mathcal{O}. \quad \blacksquare\end{aligned}$$

Acknowledgments. We would like to thank Yuri Boykov, Olga Veksler and Ramin Zabih for introducing us to the relation of MRF's and multiway cuts, and for numerous discussions on the topic.

References

[1] S. Arora, personal communication, March 1999.

[2] Y. Bartal, "Probabilistic approximations of metric spaces and its algorithmic applications." *Proc. IEEE Foundations of Computer Science*, 1996.

[3] Y. Bartal, "On Approximating Arbitrary Metrics by Tree Metrics," *Proc. ACM Symp. on Theory of Computing*, 1998.

[4] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Royal Stat. Soc. B*, 36(1974).

[5] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Statistical Society B*, 48(1986).

[6] A. Borodin, N. Linial, M. Saks, "An optimal on-line algorithm for metrical task systems," *J. ACM* 39(1992).

[7] Y. Boykov, O. Veksler, R. Zabih, "Markov random fields with efficient approximations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998.

[8] Y. Boykov, O. Veksler, R. Zabih, "Energy minimization with discontinuities," submitted for publication.

[9] Y. Boykov, O. Veksler, R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *Intl. Conf. on Computer Vision*, 1999.

[10] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks/Cole, 1984.

[11] G. Calinescu, H. Karloff, Y. Rabani. "Improved Approximation Algorithms for Multiway Cut," *Proc. ACM Symp. on Theory of Computing*, 1998.

[12] S. Chakrabarti, B. Dom, P. Indyk, "Enhanced hypertext categorization using hyperlinks," *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, 1998.

[13] J. Chuzhoy, personal communication, March 1999.

[14] R. Chellappa, A. Jain, *Markov Random Fields: Theory and Applications*, Academic Press, 1993.

[15] F.S. Cohen, "Markov random fields for image modeling and analysis," in *Modeling and Application of Stochastic Processes*, U.B. Desai, ed., Kluwer, 1986.

[16] W.H. Cunningham, L. Tang, "Optimal 3-terminal cuts and linear programming," *Proc. MPS Conf. on Integer Programming and Combinatorial Optimization*, 1999.

[17] E. Dahlhaus, D. Johnson, C. Papadimitriou, P. Seymour, M. Yannakakis, "The complexity of Multiterminal cuts," *SIAM J. Computing* 23(1994).

[18] S. Della Pietra, V. Della Pietra, J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(1997).

[19] T.G. Dietterich, "Machine learning research: Four current directions," *AI Magazine*, 18(1997).

[20] R. Dubes, A. Jain, "Random field models in image analysis," *J. Applied Statistics*, 16(1989).

[21] P.L. Erdös, L.A. Székely, "Evolutionary trees: An interger multicommodity max-flow min-cut theorem," *Adv. Appl. Math* 13(1992).

[22] P. Ferrari, A. Frigessi, P. de Sá, "Fast approximate maximum *a posteriori* restoration of multi-color images, *J. Royal Statistical Society B*, 57(1995).

[23] S. Geman, D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(1984).

[24] D. Greig, B. Porteous, A. Seheult, "Exact maximum *a posteriori* estimation for binary images," *J. Royal Statistical Society B*, 48(1989).

[25] H. Ishikawa, D. Geiger, "Segmentation by grouping junctions," *IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp. 125-131.

[26] D. R. Karger, P. Klein, C. Stein, M. Thorup, and N. E. Young, "Rounding Algorithms for a Geometric Embedding of Multiway Cut" *Proc. ACM Symp. on Theory of Computing*, 1999.

[27] A. V. Karvanov, "Minimum 0-Extension of Graph Metrics", *European J. Combinatorics*, 19(1998).

[28] R. Kinderman, J. Snell, *Markov Random Fields and their Applications*, AMS Press, 1980.

[29] S. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, 1995.

[30] P. Pardalos, H. Wolkowicz, "Quadratic assignment and related problems," in *DIMACS Series in Discrete Math. and Theoretical Computer Sci.*, vol. 16, 1994.

[31] R. Potts, "Some generalized order-disorder transformations," *Proc. Cambridge Phil. Soc.* 48(1952).

[32] S. Roy, I. Cox, "A Maximum-Flow Formulation of the N-Camera Stereo Correspondence Problem," *Proc. 6th Intl. Conference on Computer Vision*, 1998.

[33] J.W. Woods, "Two-dimensional discrete Markovian fields," *IEEE Trans. Information Theory*, 18(1972).