CAREER: Approximation Algorithms – New Directions and Techniques

Moses S. Charikar

Project Summary

The theory of NPCompleteness has shown that several naturally occurring problems are unlikely to have polynomial time algorithms. One approach to overcome this fundamental intractability for optimization problems has been to shift the focus from exact solutions to obtaining approximate solutions. The study of approximation algorithms has thus emerged as a rich and exciting field and recent advances have led to good approximation algorithms for several fundamental optimization problems. Despite these developments, several interesting and difficult open problems remain. A primary focus of this career development plan is studying the approximability of fundamental problems and attempting to close such gaps in our understanding.

The broad goals of this career development plan are the following:

- Developing new tools to devise approximation algorithms for problems on directed graphs.
- Developing new techniques for metric approximations and embeddings as building blocks for approximation.
- Investigating a systematic way to obtain and exploit strengthened SDPs as well as the use of strengthened SDPs for graph partitioning minimization problmes and ordering problems.
- Devise techniques to deal with scheduling problems with precedence constraints.
- Extend the machinery of approximation algorithms to new settings such as information theoretic and algebraic problems.
- Involve students at all levels of the research, from undergraduate projects on understanding the quality of LP and SDP relaxations through computational experiments, to the mathematical research suitable for Ph.D. students.
- Develop courses to convey core algorithmic ideas to students outside the theory community, as well as expose theory students to new developments and research issues through advanced courses.

Broader Impacts. The proposed research will attempt to provide new tools and techniques for designing approximation algorithms and enhance our understanding of the approximability of fundamental problems as well as the limitations of our current algorithmic toolkit. The educational component will develop new courses geared towards disseminating new algorithmic ideas outside the theoretical computer science community; the techniques developed in the latest research will be distilled into new graduate and undergraduate courses. Course materials developed for such new courses will be made freely available to enable similar courses to be taught elsewhere.

1 Introduction

Advances in technology and the demand for efficient algorithmic solutions to problems that arise in application domains continually pose new challenges for computer scientists. The theory of NPCompleteness has shown that several naturally occurring problems are unlikely to have polynomial time algorithms. One approach to overcome this fundamental intractability for optimization problems has been to shift the focus from exact solutions to obtaining approximate solutions. The study of approximation algorithms has thus emerged as a rich and exciting field and recent advances have led to good approximation algorithms for several fundamental optimization problems. Despite these developments, several interesting and difficult open problems remain. A primary focus of this career development plan is studying the approximability of fundamental problems and attempting to close such gaps in our understanding.

An optimization problem consists of a mapping of problem instances to a set of a feasible solutions and an objective function that assigns a value to every feasible solution. The goal is to find a solution that minimizes or maximizes the value of the objective function. For an algorithm A and instance I, let $A(I)$ denote the value of the objective function for the solution produced by A on instance I, and let $OPT(I)$ denote the value of the optimal solution for the instance. The quality of the approximation algorithm A is measured by its *approximation ratio*, which is defined as the worst case ratio $\frac{A(I)}{OPT(I)}$, over all instances I. For minimization problems, the approximation ratio is ≥ 1, obtained by maximizing this ratio over all instances I. For maximization problems, the approximation ratio is ≤ 1, obtained by minimizing this ratio over all instances I. (In the remainder of this section, approximation ratios are stated for minimization problems, unless otherwise specified). The NP-Hardness of an optimization problem implies that unless $P = NP$, we cannot achieve an approximation ratio of 1 in polynomial time. For certain problems, it is possible to achieve a $1 + \epsilon$ approximation in polynomial time for any fixed $\epsilon > 0$. Such an algorithm is called a polynomial time approximation scheme (PTAS).

Research in the last two decades has led to great progress in our understanding of many central questions in optimization. This period has seen the emergence of various algorithmic paradigms for the design of approximation algorithms: the use of linear programming (LP) relaxations and randomized rounding, the primal-dual method, the use of semidefinite programming (SDP) relaxations and the use of metric embeddings [LLR95]. In the last 5 years, new techniques have emerged: the use of dynamic programming and structured decomposition theorems to obtain PTASes for geometric problems [Aro98], metric approximations [Bar98], iterative rounding [Jain01], sophisticated dynamic programming approaches to obtain PTASes for scheduling problems [ABC+99], lagrangian relaxation [JV99], and a renewed interest in local search and greedy algorithms [AGK+01].

The connections discovered between the theory of probabilistically checkable proofs and the hardness of approximation for optimization problems have given us techniques to prove lower bounds on the approximability of several problems. It is now standard to show that an optimization problem is MAX-SNP-hard, thus proving that it is hard to approximate beyond a certain constant.

Despite these advances, there are significant gaps in our understanding of the approximability of fundamental optimization problems and we still lack general tools and techniques for broad classes of problems. The primaldual framework, though immensely successful for undirected connectivity problems, has had mostly failure with directed problems and there is little understanding about why this happens. The directed variants of several basic optimization problems such as Steiner tree, TSP and k-center are far less understood than their undirected counterparts; there are huge gaps between the known upper and lower bounds for these problems. SDP relaxations have proved to be a powerful tool for approximation; however, for several basic graph partitioning minimization problems such as multicut and sparsest cut as well as vertex ordering problems such as max acyclic subgraph, we have not managed to harness the power of SDPs despite the knowledge of promising relaxations for these problems. In fact for these problems, any techniques that improve the state of the art would be a breakthrough. Another broad class of problems not very well understood is the class of scheduling problems with precedence constraints; the lack of techniques to deal with precedence constraints could be linked to the lack of general decomposition techniques for directed graphs.

With this in mind, the broad goals of this career development plan are the following:

1. Developing new tools to devise approximation algorithms for problems on directed graphs.
2. Developing new techniques for metric approximations and embeddings as building blocks for approximation.
3. Investigating a systematic way to obtain and exploit strengthened SDPs as well as the use of strengthened SDPs for graph partitioning minimization problmes and ordering problems.
4. Devise techniques to deal with scheduling problems with precedence constraints.
5. Extend the machinery of approximation algorithms to new settings such as information theoretic and algebraic problems.
6. Involve students at all levels of the research, from undergraduate projects on understanding the quality of LP and SDP relaxations through computational experiments, to the mathematical research suitable for Ph.D. students.
7. Develop courses to convey core algorithmic ideas to students outside the theory community, as well as expose theory students to new developments and research issues through advanced courses.

Broader Impacts. The proposed research will provide new tools and techniques for designing approximation algorithms and enhance our understanding of the approximability of fundamental problems as well as the limitations of our current algorithmic toolkit. The educational component will develop new courses geared towards disseminating new algorithmic ideas outside the theoretical computer science community; the techniques developed in the latest research will be distilled into new graduate and undergraduate courses. Course materials developed for such new courses will be made freely available to enable similar courses to be taught elsewhere.

Institutional Context. Princeton University is an excellent place to pursue the career development and teaching program described in this proposal. The department has an extremely strong theory group. I expect to collaborate with several colleagues in the department. In particular, I expect to collaborate with Sanjeev Arora, on problems related to hardness of approximation and analyzing LP and SDP based approximation algorithms. I also expect to collaborate with Paul Seymour, in the Math department, on problems related to analyzing LP relaxations for optimization problems on directed graphs. I will also benefit from having algorithms experts Bob Tarjan and Bernard Chazelle as colleagues and I expect to work with them in the coming years. I have collaborated with Amit Sahai in the past, and I expect to continue working with him on problems related to embeddings and information theoretic notions of network capacity.

Princeton is very supportive of my career development plan. The university has given me a generous start-up package that has allowed me to recruit and support two students in my first year. In addition, Princeton has also given me great freedom in terms of teaching which I have used to design and teach two new courses in my first year.

The budget for this proposal includes one month of summer support and funds to support a graduate student over a five year period. This will support some, but not all the activities of my research group. As mentioned before, I have two students working with me already and I expect to recruit more students in the coming years. One of my current students (Tony Wirth) is working full time on some of the problems outlined in this proposal.

2 Proposed Research

Below we give examples of specific problems we hope to make progress on and outline our approaches for attacking them. We stress that this is not intended to be a comprehensive list of everything we will do in the course of this work. As with any good research, we hope that new discoveries will take us in unexpected directions.

2.1 Optimization Problems on Directed Graphs

Optimization problems on directed graphs have proved to be much harder than their undirected counterparts. For many interesting problems, there are large gaps between the known upper and lower bounds for the approximability of optimization problems on directed graphs. I illustrate this by discussing three such problems I intend to study. Another very interesting problem here is the *minimum directed multicut* problem, for which Cheriyan, Karloff and Rabani [CKR01] recently gave the first non-trivial approximation. Much remains to be done in understanding this problem; we omit a discussion due to space constraints.

Asymmetric (directed) k-center

Given a weighted directed graph, the goal is to choose k centers, and assign vertices to centers so as to minimize the maximum distance of a vertex from its assigned center. Here distance is measured as the *directed* distance from a center to a vertex.

In the undirected setting, the problem is approximable to a factor of 2 [HS86, Gon85] which is tight. Panigrahy and Vishwanathan [PV98] gave an $O(log^*n)$ approximation algorithm for this problem. This was improved slightly by Archer [Arc01], who obtained an $O(log^*k)$ approximation.

Recently, I have been studying a natural LP relaxation for this problem (also used in [Arc01]) , in collaboration with Howard Karloff (AT&T Research), Inge Li Goertze (PhD student, U. Copenhagen), and Tony Wirth (PhD student, Princeton). Without loss of generality, we can think of the input as an unweighted directed graph where the optimal solution covers all vertices with k centers within a distance of 1. The LP has the constraint $\forall j \ \sum_{(i,j)\in E} y_i \geq 1$, and the objective is to minimize $\sum_i y_i$. The basic question here is that if the LP produces a solution with at most k fractional centers that covers all points (within radius 1), can we choose k centers that cover all points within a constant distance, i.e. is the *radius gap* of the LP a constant ? It turns out that the radius gap of the LP formulation is at most $O(\log^* k)$.

We have managed to reduce this problem to a combinatorial question which I believe we should be able to resolve. The question is the following: Supose we have a layered graph with n_i vertices in layer i and edges directed from layer i to layer $i+1$. Further, each vertex in layer $i+1$ has edges from at least $1/k$ fraction of the vertices in layer i. For any such graph with ℓ layers, *is it possible to choose a set S of $O(k)$ vertices in layer 1 such that each vertex in the last layer has a directed path to it from some vertex in S.* I conjecture that this is true for any such layered graph with more than a certain constant number of layers (in fact, I think 4 layers suffice). If true, it would imply that the LP has a constant radius gap.

Directed Steiner tree

Given a weighted directed graph, a root r and a set of terminals S, the goal is to construct a tree rooted at r which spans all the vertices in S such that there is a directed path from the root r to every vertex in S.

The objective is to find such a Steiner tree so as to minimize the sum of the weights of the edges in the tree.[1] The best known result for this problem is an algorithm that was presented in co-authored work [CCC+99], which yields an $O(n^{\epsilon})$ approximation in $O(n^{1/\epsilon})$ time and an $O(\log^3 n)$ approximation in $n^{O(\log n)}$ time. The question of designing a polylogarithmic approximation algorithm that runs in polynomial time is a very interesting open problem.

An important fact used in [CCC+99] is that the optimal Steiner tree can be approximated by an ℓ level tree within a factor of $n^{1/\ell}$. The algorithm constructs a tree of depth ℓ by piecing together trees of depth $\ell - 1$ constructed recursively. Roughly speaking, the running time of this algorithm is n^{ℓ}. The polylogarithmic approximation is obtained by setting $\ell = \log n$, i.e. finding a good approximation to the best $O(\log n)$ level Steiner tree. The goal would be to achieve this approximation bound in polynomial time.

There are two possible approaches we plan to pursue for this problem.

1. Perhaps it is possible to modify the greedy algorithm so that it searches fewer possibilities and hence runs in polynomial time. I/ndeed such an approach was used by Chekuri, Even and Kortsarz [CEK02] for a different problem: the group Steiner tree problem. Here they used geometric bucketing and other techniques to obtain a combinatorial greedy algorithm for the group Steiner tree problem.

2. Another possible approach is to consider an LP relaxation for this problem. This was the approach taken by Zosin and Khuller [ZK02]. They point out a possible problem with this approach for a natural LP relaxation and provide an example where the LP gap is $O(\sqrt{k})$. However, in their example, k is polylogarithmic in n, so this does not rule out an approximation ratio that is polylogarithmic in n, based on the LP. Nevertheless, the example suggests that the LP solution conveys very litle information and one may need to work with a strengthening of this natural LP in order to obtain an upper bound.

Asymmetric TSP

Given a weighted directed graph, the problem is to construct a directed tour that spans all the vertice, so as to minmize the total weight of all the edges in the tour. The best known algorithm for this problem, due to Frieze, Galbiati and Maffioli [FGM82], achieves an $O(\log n)$ approximation. The best lower bound is $117/116$, by Papadimitriou and Vempala [PV00]. Determining whether this problem has a constant factor approximation has been an intriguing problem that has received some attention.

For the symmetric case, the $3/2$-approximation algorithm of Christofides [Chr76] is the best known. The well-known Held-Karp conjecture says that the value of a certain LP relaxation of Held and Karp [HK70] is within a factor of $4/3$ of the optimal traveling salesman tour. This conjecture has been a subject of intense study in the past decade.

Carr and Vempala [CV00] studied a LP relaxation for the asymmetric TSP, which can be viewed as an asymmetric generalization of the Held-Karp relaxation. They showed that a certain strengthening of the Held-Karp conjecture, if true, would imply that the value of this relaxation is at most a factor of $4/3$ away from the optimal asymmetric travelling salesman tour. Achieving such an approximation ratio seems a far cry from the current state of knowledge about this problem.

An interesting aspect of [CV00], from my point of view, is the fact that they identify the structure of certain basic solutions for the asymmetric TSP LP relaxation, which they call fundamental extreme points and ultra-fundamental extreme points. In [CV00], these are used to show the connection between the relaxations for the asymmetric and symmetric problems. It is precisely such fractional solutions that we need to be able to round to integer solutions. In some sense these are the hardest solutions, but they also possess considerable structure which may make dealing with them easier.

[1] This problem is hard to approximate within a factor of $O(\log n)$ since set cover can be reduced to it.

2.2 Approximation of metric spaces and applications

Recently developed techniques to approximate metric spaces by simpler spaces that have found various applications to the design of approximation algorithms. There are several interesting research issues here, related to improving the general tools for approximation of metric spaces and improving the approximation factors for specific problems that currently depend on these general tools and consequently have $\Omega(\log n)$ approximation factors.

Probabilistic approximation of metric spaces via tree metrics

Bartal [Bar96, Bar98], introduced the notion of probabilistic approximation of metric spaces via tree metrics.[2] Several optimization problems on metric spaces are easier to solve when the input is restricted to a tree metric. This technique allows such solutions to be translated to the general metric case with a loss of $O(\log n \log \log n)$.

One interesting question is whether the factor of $O(\log n \log \log n)$ can be improved. There is a lower bound of $\Omega(\log n)$ on this factor. It turns out that the $\log n \log \log n$ term comes about from an application of Seymour's graph partitioning and recursion analysis [Sey95]. This in turn is the basis for several divide and conquer algorithms based on the spreading metric approach [ENRS00], which have approximation factors of $O(\log n \log \log n)$. Thus making an improvement on the approximation factor for probabilistic approximation of metric spaces could potentially lead to improvements in these other algorithms as well. A second issue is whether better approximation factors can be achieved by allowing more complicated structures than trees. e.g. there are several problems that can be solved on graphs of constant treewidth by dynamic programming. How well can general metric spaces be approximated by such constant treewidth graphs ?

Below we describe two interesting problems for which the best known algorithms use this approximation machinery. Another such interesting problem (whose discussion we omit) is *minsum clustering* (See [Ind99, BCR01, DKKR02, GI02]).

Buy at Bulk network design

This problem [SCRS97] models economy of scale in the cost of network connections. Given a concave cost function for links (here the cost per unit length is a concave function of capacity), the goal is to design a minimum cost network by installing/buying capacity on the links of a given graph so as to satisfy a set of given demands. In general the demands are specified by source-sink pairs s_i, t_i and demand d_i. The network constructed should be able to route a flow of value d_i between s_i and t_i simultaneously for all i.

Awerbuch and Azar [AA97] gave an $O(\log n \log \log n)$ approximation based on Bartal's technique [Bar96, Bar98]. Guha, Munagala and Meyerson [GMM01] gave the first constant factor approximation for the single source case. This was simplified and improved recently by Talwar [Tal02].

For the general case, i.e. multiple source sink pairs, the technique based on probabilistic approximation of metric spaces via trees is the only method we know of, but we have no lower bound that rules out a constant factor approximation. A natural place to look for good lower bounds on the optimal solution is to use a suitable LP relaxation. However, obtaining a good LP relaxation for the problem seems tricky because of the concave costs. In the single source case, [GMM01, Tal02] show how to overcome this difficulty. Perhaps their ideas can be combined with new techniques to obtain such LP relaxations for the general case.

Classification with pairwise relationships

[2] The idea is to approximate a metric space by a probability distribution over tree metrics. For any pair of vertices, the distance in each tree metric is greater than the distance in the original metric. Further, the expected distance in the tree metric is at most an $O(\log n \log \log n)$ factor more than the distance in the original metric.

This problem was introduced by Kleinberg and Tardos [KT99], motivated by problems in image reconstruction and vision. It can be viewed as a generalization of the multiway cut problem. We are given a graph and the objective is to assign one of k labels to each vertex. There is a metric specified on the labels. We are given assignment costs for every vertex label pair. Further, if the labels of two adjacent vertices differ, we must also pay a separation cost which is equal to the weight of the edge times the distance between the labels. The objective is to come up with an assignment of labels that minimizes the sum of the assignment and separation costs.

[KT99] obtained an $O(\log k \log \log k)$ approximation for this problem using Bartal's techniques [Bar96, Bar98]. Calinescu, Karloff and Rabani [CKR01] considered a special case, called 0-extension where the assignment costs are zero and k vertices are preassigned each of the k labels; they achieved an $O(\log k)$ approximation. Chekuri, Khanna, Naor and Zosin [CKNZ01] studied a natural LP relaxation for the problem of [KT99] and obtained some improved results for special metrics. We do not know of super constant gap examples for this LP relaxation, and this seems like a promising starting point for better approximations.

In recent work [Cha02b], I showed that the analysis of this LP relaxation is interesting beyond designing an approximation algorithm for the problem of classification with pairwise relationships. Briefly, rounding algorithms for this LP relaxation can be viewed as *sketching* algorithms for a certain distance measure on distributions, called the *Earth Mover Distance* (EMD).[3] It turns out that rounding schemes for the LP relaxation can be viewed as similarity preserving hashing schemes that map distributions to a compact representation so that an estimate of the EMD between two distributions can be made from their compact representations. Further, the quality of the approximation here is exactly the approximation factor obtained from the analysis of the LP relaxation.

2.3 Semidefinite relaxations

In a seminal paper, Goemans and Williamson [GW95] introduced semidefinite programming (SDP) relaxations as a tool for obtaining approximation algorithms. They applied semidefinite programming to get an approximation ratio of roughly 0.878 for the MAX CUT problem. Since then, semidefinite programming has been used to obtain approximation algorithms for several problems, including graph-coloring and constraint satisfaction problems.

In this section, we describe several problems for which we feel a better understanding of the power of SDP relaxations is required. Other than classical optimization problems such as coloring and vertex cover, there are two broad classes of problems for which improved rounding algorithms for SDP solutions can potentially make substantial improvements to the state of the art: graph partitioning minimization problems and vertex ordering problems.

2.3.1 Stronger SDP relaxations

In this section, I discuss SDP based techniques for the classical optimization problems of graph coloring and vertex cover and a general way to obtain a sequence of strengthened SDP relaxations which I hope will be useful for these and other problems.

3-coloring

Feige and Kilian [FK98] showed that it is hard to approximate the chromatic number within ratio $n^{1-\epsilon}$ for any $\epsilon > 0$ unless $NP = ZPP$. However we know relatively little about how well graph coloring can

[3] This is used in vision applications [RGT97, RT98, RTG98a, RTG98b, Rub99, CG97a, CG97b, RT99a, RT99b], where images are represented as distributions over points in a metric space and EMD is used to compute the distance between these distributions. Roughly speaking, the EMD between two distributions is the min cost transportation that transforms one distribution to the other.

be approximated for graphs with a constant chromatic number. In particular the following question has received a lot of attention: *how many colors do we need to color a* 3*-colorable graph in polynomial time ?*

Karger, Motwani and Sudan [KMS98] used semidefinite programming and obtained $\tilde{O}(n^{1-\frac{3}{k+1}})$-colorings for k-colorable graphs, improving on the results of Blum [Blu94]. The result of [KMS98] was subsequently improved by Blum and Karger [BK97] for 3-colorable graphs, and by Halperin, Nathaniel and Zwick [HNZ01] for k-colorable graphs with $k > 3$.

On the algorithmic side, the current best known algorithm for coloring 3-colorable graphs needs $\tilde{O}(n^{3/14})$ colors [BK97]. However, on the hardness side, the only hardness result we know is that it is NP-Hard to 4-color a 3-colorable graph (see Khanna, Linial and Safra [KLS00] as well as Guruswami and Khanna [GK00]) Khot [Kho01] showed recently that a for a sufficiently large constant k, a k colorable graph cannot be colored using $k^{O(\log k)}$ colors.

An interesting question is whether we can achieve an approximation ratio of $n^{o(1)}$ for 3-coloring. Recently, I showed [Cha02a] that such a result cannot be achieved using certain strengthened versions of the SDP relaxations used by Karger, Motwani and Sudan [KMS98].

Very recently, Feige, Langberg and Schechtman [FLS02] demonstrated that the analysis of [KMS98] is almost tight, i.e. it is not possible to improve on the rounding procedure they use for the particular SDP they consider. (Note that to make improvements, [BK97] and [HNZ01] use combinatorial bounds in addition to the bound provided by the SDP). Interestingly, the techniques of [FLS02] use and build on the ideas in the geometric construction of the earlier paper [FS01] on SDPs for MAX-CUT.[4] It would be interesting to see whether such ideas can be used to demonstrate tightness of rounding procedures for other SDPs as well.

The techniques of [FLS02] do not extend to the strengthened SDP relaxations considered in [Cha02a]. New rounding techniques will be needed to demonstrate that these strengthened relaxations are strictly better than those considered in [KMS98]; however they cannot achieve an $n^{o(1)}$ approximation. I intend to investigate whether other, stronger SDP relaxations will lead to such a result.

Vertex Cover

Given a graph, $G(V, E)$, the goal is to pick the smallest posible subset S of vertices so that every edge is adjacent to at least one vertex in S. One can also consider the variant where vertices have weights associated with them and the goal is to find the vertex cover of minimum total weight.

The best known algorithms for this problem achieve an approximation factor of $2 - o(1)$.[5] Obtaining a $2 - \epsilon$ approximation factor is an open problem that has intrigued researchers for a long time. Håstad showed that vertex cover is hard to approximate within a factor of $7/6$. Recently, Dinur and Safra [DS02] obtained a stronger lower bound of approximately 1.36.

Figuring out whether we can beat the 2 barrier for vertex cover is a fascinating question that I intend to investigate. As a first step, I want to understand whether SDP relaxations with additional constraints can achieve a factor better than 2. Kleinberg and Goemans [GK98] showed that a certain SDP relaxation does not give an approximation better than $2 - \epsilon$ for any $\epsilon > 0$, and in [Cha02a], I showed that for a stronger relaxation, the gap approaches 2 as well. Very recently, Arora, Bollobas and Lovasz [ABL02] showed that for a very general class of LP relaxations (even in cases where we do know the LP explicitly), the gap for

[4] The basic approach here is to construct a graph that comes equipped with a vector solution for the SDP; in fact the vector solution is chosen first and the graph is constructed based on it. In most cases, the SDP solution corresponds to an embedding of the graph on the unit sphere with certain constraints. The approach is to start with a dense collection of points on the sphere and build the graph by putting in all edges so that the constraints required of the SDP solution are met. A considerable amount of technical machinery is required to prove the required properties of the graphs constructed thus for the MAX-CUT and k-coloring relaxations.

[5] An approximation factor of $2-\frac{\log\log n}{2\log n}$ was achieved by Monien and Speckenmeyer [MS85] and Bar-Yehuda and Even [BE85]. Halperin [Hal00] improved the factor slightly to $2 - \frac{2\ln\ln n}{\ln n}(1 - o(1))$.

vertex cover is $2-\epsilon$ for arbitrarily small ϵ. I intend to investigate whether an analogous result can be shown for SDP relaxations as well.

Systematic construction of strengthened SDPs

We describe a systematic way of writing down a sequence of progressively stronger SDPs for various optimization problems involving assigning discrete values to variables. We start with the basic SDP formulation and then explain how it can be made stronger. The SDP formulation has a canonical unit vector v_0. All other vectors used in the SDP are *vector analogues* of $\{0,1\}$-variables, each representing the probability of a specific event. Any such vector v satisfies the constraint $v \cdot v_0 = v \cdot v$, and this value is interpreted as the probability of the underlying event. Further, the dot product of any two vectors is non-negative; this value is interpreted as the probability of both events occuring simultaneously. For every variable x in the optimization problem and possible assignment t to x, we have the vector $v_{(x,t)}$ (representing the event that x is assigned the value t). The sum of the vectors $v_{(x,t)}$ over all possible values t is set to be v_0, indicating that x must be assigned exactly one value. If x_1, t_1 and x_2, t_2 are mutually exclusive assignments, then we have the constraint $v_{(x_1,t_1)} \cdot v_{(x_2,t_2)} = 0$.[6]

The optimization objective (if required) is expressed in terms of the dot products of these vectors (interpreted as probabilities). Using this kind of formulation, one can write down an SDP for graph coloring for example. Though very different for the SDPs mentioned before, one can show that this is equivalent to the strongest SDP in [Cha02a].[7]

Further, one can produce a sequence of successively stronger SDPs in the following way: We consider sets of upto k assignments $(x_1, t_1) \ldots (x_k, t_k)$. For each such set S, we have a vector v_S. These vectors satisfy the previously mentioned constraints with v_0, and $v_{S_1} \cdot v_{S_2} = 0$ if S_1 and S_2 are contradictory sets of assignments. Further, for a set S of assignments and x a variable not assigned by S, we have the consistency constraint $v_S = \sum_t v_{S \cup (x,t)}$. Here $S \cup (x,t)$ denote the set of assignments in S together with the new assignment (x,t). Such an SDP enforces that all valid constraints on k variables are satisfied. As we increase k, we get a sequence of progressively stronger SDPs. This has close connections to the lifting procedures proposed by Lovász and Schrijver [LS91] and other authors.[8] I intend to investigate this further and determine whether such strengthened relaxations can be used to derive improved approximations, especially for graph coloring and vertex cover. In collaboration with David Williamson (IBM Almaden), I am looking at such relaxations for multiway cut.

2.3.2 Graph Partitioning

We discuss a number of graph partitioning minimization problems where we know of promising SDP relaxations which could potentially be the basis for substantially improved approximations. The first two problems, minimum multicut and sparsest cut, are particularly interesting since algorithms to solve these problems are used as subroutines in solving a host of other problems, usually in divide and conquer approaches. Making an improvement for these problems is likely to result in improvements for a variety of other optimization problems as well.

Minimum multicut

Given an undirected graph $G(V,E)$ and k pairs of vertices $(s_1,t_1), \ldots (s_k,t_k)$, a multicut is a subset $F \subseteq E$ of edges such that if all the edges in F are removed, then none of the pairs $(s_i,t_i), i = 1 \ldots k$ are in

[6] e.g. this happens if $x_1 = x_2$ and $t_1 \neq t_2$, or if the two assignments are prohibited due to the problem constraints.

[7] Such SDPs have been used by other authors; see Khot [Kho02].

[8] See the article by Laurent [Lau01] for a comparison of various lifting procedures; also see Goemans and Tuncel [GT00b].

the same connected component in the remaining graph $G' = (V, E - F)$, In the minimum multicut problem, we are also given a non-negative cost $c(e)$ associated with each edge $e \in E$ and we wish to find the multicut of minimum total cost.

The best known algorithm for this problem is due to Garg, Vazirani and Yannakakis [GVY96], which achieves an $O(\log k)$ approximation. The algorithm uses an LP relaxation that assigns a length function $\ell(e) \in [0, 1]$ to each edge $e \in E$.[9] The constraint is that the distance from s_i to t_i under this length function is at least 1.[10] [GVY96] show that a fractional solution to this LP relaxation can be rounded to an integral multicut with value at most $O(\log k)$ times the value of the fractional multicut. Such LP relaxations that assign length functions are useful for other graph partitioning problems as well and the region growing technique of [GVY96] has found applications beyond the minimum multicut problem.

One can also write an SDP relaxation for the minimum multicut problem. Here, we have a unit vector x_v corresponding to each vertex $v \in V$. The SDP measures the extent to which a pair of vertices u and v are separated by the squared norm of $(x_u - x_v)/2$. Note that this is 0 if $x_u = x_v$ and 1 if $x_u = -x_v$. For every pair (s_i, t_i), we have the constraint that $x_{s_i} = -x_{t_i}$.[11] Now if we consider any hyperplane through the origin then the edges whose endpoints are on opposite sides of the hyperplane form a valid multicut since x_{s_i} and x_{t_i} must be on opposite sides of the hyperplane (excluding degenerate cases). However, simply taking the multicut produced by a random hyperplane does not yield a good approximation. Consider vertices u, v with corresponding vectors x_u, x_v that subtend a very small angle θ. Then x_u and x_v are on opposite sides of a random hyperplane with probability proportional to θ; however the contribution of the pair (u, v) to the objective function of the SDP is $O(\theta^2)$. Indeed it turns out that the gap of this SDP is huge and it is not hard to construct an example with integrality gap $\Omega(n)$.[12]

Now, one can strengthen this SDP by the addition of the so-called triangle inequality constraints which can be described as follows. Recall that we said that the SDP measures the extent to which u and v are separated by the squared norm of $(x_u - x_v)/2$. This can be thought of as assigning a length function $\ell(u, v)$ to the pair u, v. The triangle inequality constraints stipulate that this length function satisfies triangle inequality.

Notice that the SDP relaxation is at least as strong as the LP relaxation used by [GVY96]. Thus the gap of the SDP relaxation is at most $O(\log k)$. However, unlike the LP relaxation, we do not know of any $\Omega(1)$ gap example for this SDP. Thus it may be possible to obtain a constant factor approximation for the minimum multicut problem by appropriately rounding the vector solution produced by this SDP. Such a result would be extremely interesting, but seems out of reach of currently known SDP rounding techniques.

Sparsest cut

The input is similar to that for the minimum multicut problem discussed above. It consists of an undirected graph $G(V, E)$ with non-negative costs $c(e)$ for each edge $e \in E$ and k pairs of vertices (s_i, t_i) with non-negative demands d_i associated with each pair. The goal is to find a cut that minimizes the ratio of the total cost of the edges removed to the total demand of the edges separated.

An LP relaxation for this problem similar to that for minimum multicut is as follows. We have a fractional length $\ell(e) \in [0, 1]$ associated with each edge e. The extent to which a pair s_i, t_i is separated is measured by y_i, which is the length of the shortest path between s_i and t_i according to this length function. The LP solution is chosen so as to minimize the ratio of $\sum_{e \in E} \ell(e) \cdot c(e)$ to $\sum_i y_i \cdot d_i$.

Improving a series of previous approximation algorithms, Linial, London and Rabinovich [LLR95] and

[9] $\ell(e) = 1$ indicates that the edge is cut, while $\ell(e) = 0$ indicates that the edge is not cut.

[10] This encodes the constraint that at least one edge should be cut on every path from s_i to t_i.

[11] This encodes the constraint that s_i and t_i must be separated.

[12] This is achieved for an instance on a cycle with one terminal pair.

independently, Aumann and Rabani [AR98] gave an $O(\log k)$ approximation for this problem, which is the currently best known approximation factor. Their algorithm is based on rounding the LP solution by embedding the length function into ℓ_1 with an $O(\log k)$ distortion..

Similar to the minimum multicut problem, one can consider an SDP relaxation for this problem. Here, the length function is obtained by an embedding of the vertices on a unit sphere. In particular, $\ell(u, v)$ is the squared norm of $(x_u - x_v)/2$. The triangle inequality constraints stipulate that this length function is indeed a metric. Again, we do not know of an $\Omega(1)$ gap example for this SDP.

Minsum 2-clustering

Given a graph $G(V, E)$ with weights on edges, the goal is to divide the vertices into two disjoint clusters such that the sum of the weights of the intracluster edges is minimized. Note that we discussed minsum clustering earlier, but the assumption previously was that the weight function satisfied triangle inequality. In this case, no such assumption is made. The objective function is exactly the complement of the MAX-CUT objective.

The results of Garg, Vazirani and Yannakakis [GVY96] give an $O(\log n)$ algorithm for this problem. Similar to the proposed SDP relaxations for multicut and sparsest cut, one can write down an SDP relaxation for this problem as well. Again, triangle inequalities are required to ensure that the gap is not $\Omega(n)$. In analyzing the SDP with triangle inequalities, we encounter the very same technical difficulties that we encounter in the analysis of the SDPs for multicut and sparsest cut. (In this case, the difficulty comes in analyzing pairs of vertices where the corresponding vectors subtend angles approaching π; in this case, the probability that a random hyperplane has both vectors on the same side is much more than the contribution of this pair to the objective function). In many ways, this is a simpler problem to analyze and understand. Though arguably not as interesting in its own right, a better understanding of this problem may give valuable clues to attack the more interesting problems of multicut and sparsest cut.

Feige and Schechtman [FS01] constructed gap examples for the MAX-CUT SDP relaxation with triangle inequality. Since this problem is the complement of MAX-CUT, an obvious question that comes to mind is whether the techniques of [FS01] can be used to construct gap examples for this problem. The graphs constructed by [FS01] are not good gap examples for the minsum 2-clustering problem. This is because for the graphs they construct, the angle subtended by vectors corresponding to end points of vertices is bounded away from π. For such graphs, a cut obtained by a random hyperplane gives a constant factor approximation.[13]

2.3.3 Ordering problems

We have a very poor understanding of the use of SDP relaxations for vertex ordering problems.[14] We mention a couple of interesting vertex ordering problems where there is a gap between upper bounds and lower bounds or essentially, the only known upper bounds are trivial. SDP relaxations could prove to be a useful algorithmic tool here. Another interesting problem in this category (not discussed due to space constraints) is *betweenness* studied by Chor and Sudan [CS98a].

[13] It may be possible to modify their construction to obtain gap examples for the complementary objective. A plausible approach for doing this could be the following: Pick points on the unit sphere and construct a graph by connecting points that subtend a very large angle at the center. Finally, eliminate triples of points that do not satisfy the triangle inequality constriants. It will be tricky to pick the right parameters to make this construction work since the triangle inequality constraints and the fact that we need points to subtend large angles seem to conflict.

[14] One exception is the work of Blum, Konjevod, Ravi and Vempala [BKRV00] and Dunagan and Vempala [DV01] to the minimum bandwidth problem.

Maximum acyclic subgraph

Given a directed graph G, the goal is to find an acyclic subgraph of G that contains the largest number of edges. Equivalently, we can formulate the problem as finding an ordering of the vertices $v_1, v_2 \dots v_n$, such that the number of edges of the form (v_i, v_j) for $i < j$ is maximized. In general, edges may have non-negative weights and then the objective is to maximize the weight of the edges picked.

A $1/2$ approximation is trivial for this problem. Simply order the vertices arbitrarily. At least half the edges must go either forward or backward in this ordering. This gives an acyclic subgraph with half the number of edges. The best known result for this problem is due to Berger and Shor [BS97], who obtained a $1/2 + O(1/\sqrt{d_{max}})$ approximation (where d_{max} is the maximum degree).

This problem, (also called the *linear ordering problem*) has been studied in the mathematical programming literature [GJR85a, GJR85b]. LP relaxations for this problem typically use variables $x_{ij} \in [0, 1]$ to indicate whether i precedes j in the linear ordering together with other constraints on the x_{ij} variables. One common family of constraints is the family of *cycle* constraints: for every cycle C, the sum of edge variables does not exceed $|C| - 1$.

Recently, Newman and Vempala [NV01] showed that most of these LP relaxations have integrality gaps approaching $1/2$, i.e. they construct examples where the ratio of the value of the optimal integral solution to the value of the fractional solution approaches $1/2$. This implies that these relaxations do not provide useful bounds that can be used to improve the aproximation factor beyond $1/2$. [NV01] also prove that this problem is hard to approximate beyond a factor of $65/66$.

We propose to investigate an SDP relaxation for this problem. Here, we have a canonical unit vector v_0 and unit vectors v_{ij} for every ordered pair of vertices (i, j). ($v_{ij} = -v_{ji}$.) If $v_{ij} = v_0$, this implies that the edge (i, j) is picked, while $v_{ij} = -v_0$ implies that the edge (i, j) is not picked. The extent to which an edge (i, j) is picked by the SDP solution is measured by the squared norm of $(v_0 + v_{ij})/2$. For every i, j, k, we have the constraint that $v_{ij} + v_{jk} + v_{ki}$ has norm 1.[15]

It is not too hard to show that this SDP relaxation is at least as strong as the LP relaxation with cycle constraints. However, unlike the LP relaxation, we do not know of a gap example for this SDP which shows that we cannot achieve a factor better than $1/2$ using this SDP relaxation. I plan to investigate this problem further and a natural first step would be to obtain an increased understanding of the SDP relaxation. What is interesting about the SDP is that is represents an ordering of the vertices via local information. If such a local representation is indeed good enough to represent a total ordering, it would most likely be useful for other vertex ordering problems as well.

Feedback edge set

Given a directed graph, the objective is to remove the smallest subset of edges such that the remaining graph is acyclic. One can also consider the natural generalization to the setting where edges have weights asociated with them. This is exactly the complement of the maximum acyclic subgraph problem, discussed above.

The best algorithm for this problem, is an $O(\log n \log \log n)$ approximation due to Even, Naor, Schieber and Sudan [ENSS98]. They use the value of an LP relaxation similar to the LP for maximum acyclic subgraph with cycle constraints. Since the objective function is the complement in this case, the cycle constraints now say that the sum of the fractional values of the edges in any directed cycle is at least 1. The result of [ENSS98] implies that the gap of this LP relaxation is at most $O(\log n \log \log n)$. We do not have any hardness results for this problem which suggest that a constant factor approximation is not possible.

I mention a particular special case of this problem that I intend to investigate. The hope is that this will

[15] This encodes the constraint that exactly two of the pairs (i, j), (j, k) and (k, i) must be oriented in the same direction.

shed some light on the general case. The problem is of finding a permutation that is close to a given set of permutations. This arises in finding an aggregate ranking from the rankings obtained from different sources, such as rankings for web pages from different search engines [DKNS01]. More formally, the problem is the following: Given k permutations of n items, the goal is to find a permutation π that minimizes the sum of distances to the given k permutations. Here the distance between two permutations is measured by the number of inversions.[16]

This is a special case of the feedback edge set problem on directed graphs. In this case, an approximation ratio of 2 is easy to achieve. In fact, simply picking the best of the given k permutations achieves an approximation factor of $2(1 - 1/k)$. However, achieving a better bound does not seem easy. I am hopeful that investigating this problem further will lead to techniques to deal with orderings that would be useful for other ordering problems as well.

2.4 Metric Embeddings

Linial, London and Rabinovich [LLR95] introduced the use of metric embeddings for analyzing the max-flow min-cut gap of multicommodity flow problems and designing approximation algorithms for sparsest cut. I mention two interesting problems related to metric embeddings and their relationship to approximation problems.

Negative metrics in ℓ_1

Earlier, we described an SDP relaxation for minimum multicut and sparsest cut. It turns out that the gap of the SDP relaxation can be related to a question about metric embeddings: Given a set of points in R^d such that the square of distances satisfy triangle inequality, what is the minimum distortion that must be incurred in emedding the metric given by squares of distances in L_1 ? It turns out that gap of the SDP relaxation is closely related to the minimum distortion required for such an embedding.

Bourgain's results [Bou85] imply that any metric on n points can be embedded into ℓ_1 with distortion $O(\log n)$. The interesting question is whether any such squared Euclidean metric (also called a negative metric) can be embedded in ℓ_1 with constant distortion. If true, it would imply that the gap of these SDP relaxations we discussed is a constant.

Planar metrics

It has been conjectured that any planar metric (i.e. the shortest path metric on a weighted planar graph) can be embedded into ℓ_1 with constant distortion. By the results of [LLR95], this would imply that the max-flow min-cut gap for multicommodity flow on planar networks is a constant. Rao [Rao99] showed that any planar metric on n points can be embedded into ℓ_1 with distortion $O(\sqrt{(}\log n))$. This is the best known distortion result for general planar metrics. Gupta, Newman, Rabinovich and Sinclair showed that outerplanar metrics can be embedded into ℓ_1 with constant distortion. This was extended to k-outerplanar graphs (for constant k) in recent work by Chekuri, Gupta, Newman, Rabinovich and Sinclair [CGN+02]. However, the question of whether general planar metrics can be embedded in ℓ_1 with constant distortion remains an interesting open problem and I intend to investigate this.

2.5 Scheduling Problems: precedence constraints

Scheduling problems are amongst the oldest problems considered by algorithm designers. A number of major advances have been made in the theory of scheduling algorithms over the past three decades. Yet, some

[16] An inversion is a pair of elements that is ordered differently in the two permutations.

basic scheduling problems have remained open. I mention two problems relating to minimizing maximum completion time (makespan) of a set of jobs under precedence constraints. These have been mentioned as the first two problems in a recent list of 10 open problems in scheduling [SW99]. Making progress on either of these two questions requires a new insight into dealing with precedence constraints. In particular, we may need to devise new techniques to decompose directed acyclic graphs in order to obtain improved algorithms. There could be a synergy between the ideas used to attack optimization problems on directed graphs and these scheduling problems with precedence relationships. I elaborate on the problem of scheduling unit time jobs on uniform machines below. The second problem (discussion omitted) is that of scheduling jobs on *related machines with precedence constraints*, studied by Jaffe [Jaf80], Chudak and Shmoys [CS99] and Chekuri and Bender [CB98].

Unit time jobs with precedence constraints

Given a set of unit time jobs with precedence constraints defined on them, the goal is to assign jobs to m machines so as to minimize the maximum completion time. In other words, the goal is to assign each job to one of T timesteps, where at most m jobs are assigned to one timestep; the objective is to minimize the completion time T. The precedence constraints stipulate that if job i precedes job j, then i must be executed before j.

The best known algorithm, due to Coffman and Graham [CG72] achieves an approximation ratio of $2 - \frac{2}{m}$ for m processors. [17] Surprisingly, the problem is not even known to be NP-hard for a constant number of processors. The scheduling survey by Karger, Stein and Wein [KSW98] mentions this as one of the most famous open problems in scheduling.[18]

In undergraduate research [Cha95], I showed that no level based algorithm[19]. achieves an approximation factor better than $2 - 2/\sqrt{m}$. Note that all the algorithms known for this problem are level based. This suggests that very different ideas are needed to break the 2 barrier. Recently, Ranade [Ran01] devised an algorithm, very different from the Coffman-Graham approach, that achieves an approximation ratio strictly better than 2 in the case when roughly, the precedence graph can be decomposed into large independent pieces. This could be the basis of a divide and conquer algorithm to this problem. I intend to explore this approach further.

Another possible approach is to consider LP relaxations for this problem. LP relaxations have been employed very successfully for various other scheduling problems. In this case, it is natural to formulate an LP with variables for each job and time slot indicating whether the job is scheduled in that time slot. We can formulate precedence constraints by saying that if job i precedes job j, then the extent to which job i is executed upto time t does not exceed the extent to which job j is scheduled upto time $t + 1$. This LP has an integrality gap of $2 - 2/(m + 1)$. However, one can strengthen this LP by adding valid constraints for all sets of at most $k = poly(m)$ variables. [20] Such an LP can be solved for constant m. My preliminary investigations did not yield gap examples for such a strengthened LP with gap approaching 2 as $m \to \infty$. I intend to explore this approach further to determine whether a better algorithm is possible for constant m using such an LP.

[17] The analysis is due to Lam and Sethi [LS77] and an error in their analysis was corrected by Braschi and Trystram [BT94].

[18] The problem is known to be hard to approximate within a factor of $4/3$ for general m, via a reduction from clique.

[19] A level based algorithm is one that assigns levels to each job based on the maximum length of the path starting from the job. Jobs with higher level numbers are scheduled before jobs with lower level numbers

[20] One way to enforce such a constraint is to ensure that the values of the variables corresponding to any k jobs are convex combinations of valid schedules for the k jobs.

2.6 Approximation algorithms in unconventional settings

In addition to studying traditional applications of approximation algorithms, I also want to explore approximation in new scenarios. Here, much of the approximation machinery such as mathematical relaxations do not apply and new techniques are required. For lack of space, I elaborate on only one such class of problems below, related to information theoretic information capacity. Other problems I would like to study include approximation questions in algebraic settings, such as minimizing the number of multiplications required in evaluating a given set of monomials (see Pippenger [Pip80]), and a special case of this problem called the minimum addition chain problem (see Yao [Yao76]).

Network Information Flow

Recently, Ahlswede, Cai, Li and Yeung [ACLY00] introduced an interesting model to give an information theoretic characterization of network capacity. Given a network with capacities on links, they considered the scenario where one transmitter (source) is broadcasting the same information to a set of receivers (sinks). The key difference from the usual multicommodity flow formulations used to model such situations in Computer Science, is that the transmissions on links are considered to be bits carrying information rather than flows. Here, a router is allowed to transmit information on a link that is an arbitrary function of the information it receives on other links (e.g. it could XOR certain bits and so on). It turns out that this additional power allows a greater capacity to be supported than can be achieved by thinking of the transmissions as flows. In this model, [ACLY00] gave an elegant characterization of network capacity for multicast from a single source: *the maximum transmission rate that can be supported for multicast from a source to a set of sinks in a given network is exactly the minimum cut that separates the source from any sink*. Later, Li, Yeung and Cai [LYC02] showed that the network capacity could be achieved using linear codes, i.e. where the transmitted bits are XOR functions of the received bits.

However, the question of characterization of network capacity for multiple source sink pairs is wide open. The problem seems considerably more difficult in this case and an elegant characterization of the kind given by [ACLY00] may not be possible. One of the major hurdles one runs into in the more general setting is that we do not have a good characterization of the entropy function. Given a set of random variables, consider joint entropies of subsets of these variables, conditional entropies and so on. We do not have a way to decide if a given set of entropy function values is feasible (see [ZY97, ZY98, YZ01, CY02]). Recently, Koetter and Meddard [KM02] gave an algebraic characterization of network capacity in the setting where transmissions are restricted to be linear functions,[21] and showed hardness of determining feasibility of network coding strategies.

Here is a basic problem that occurs as a subcase of the 2 sources scenario and seems out of reach of our current techniques: We wish to simultaneously transmit two sources X and Y using n bits (think of these as n channels of capacity 1). We are given sets R_X and R_Y of receivers, each with access to a specified subset of the n bits. Each receiver in R_X must be able to decode X and similarly for R_Y. Clearly, if X and Y have very high rates, such an encoding will not be possible. Can we characterize the rate pairs for which such a scheme is possible ?

This information theoretic model of network flow is very interesting and has not been considered in the Computer Science community. I have begun peliminary investigations of this model in collaboration with Rina Panigrahy (Cisco Systems) and Amit Sahai (Princeton). It would be interesting to see if the techniques developed in the theoretical computer science community can be extended to obtain approximation results for this information theoretic notion network capacity. Very likely, new tools will have to be developed to deal with these problems.

[21] It is not known whether linear functions suffice to achieve network capacity in the general case.

3 Previous Research Accomplishments

A significant portion of my research has been focussed on the study of approximation algorithms. In joint work with Guha, Shmoys and Tardos [CGTS99], we developed the first constant factor approximation for the k-median problem; this problem had been open for a long time. My original conference submission with Sudipto Guha was awarded the best student paper award at STOC '99. In joint work with Guha [CG99], I devised improved combinatorial algorithms for k-median and simple local search heuristics for facility location. In collaboration with other researchers, I also devised algorithms for other clustering objectives [BCR01, CP01] and introduced formulations for clustering problems with outliers [CKMN01].

Together with colleagues at Stanford [CCGG98, CCG+98], I showed that every metric can be approximated by a small number of tree metrics. This yields a very general derandomization technique for algorithms that use the framework of probabilistic approximation of metrics via tree metrics. Other optimization problems I have studied are the directed Steiner tree problem [CCC+99] (mentioned previously) and problems in vehicle routing [CKR98, CR98].

In the past, I have also been interested in online algorithms, having worked on online algorithms for page migration [BCI01], load balancing [BCK00], and models for delayed information and action [ACM01].

Some of my recent research has focused on the design of algorithmic techniques for large data sets, such as (1) algorithms to produce compact *sketches* of data that enable approximate computations to be done using the sketches and (2) algorithms that process large amounts of data in one pass (i.e. the streaming model). In particular, I have worked on designing hash functions for estimating similarity [BCFM00], providing a mathematical analysis of a technique used to eliminate near duplicate documents in AltaVista. This technique, called minwise independent permutations, has been used in several research papers [CDF+00, HGI00, CKKS00, CJK+01, GGK01]. In a recent paper [Cha02b], I showed that techniques for rounding LPs and SDPs in approximation algorithms could be interpreted as constructions of hash functions for estimating similarity, providing new constructions for vector similarity and the earth mover distance.

A recent paper [CCF02] devised an approximate counting data structure for identifying frequent items in one pass over a data stream. I have also been interested in designing streaming algorithms for clustering. An earlier paper [CCFM97] addressed this issue for the p-center objective and a recent paper [CPO02] presented a streaming algorithm for the k-median objective.

4 Education Plan

Education, both at the undergraduate and graduate level, is an integral part of this career development plan. I plan to involve students, both at the graduate and undergraduate level in the research activities outlined in this proposal. Undergraduate research projects involving constructing and analyzing gap examples for LPs and SDPs will make valuable contributions to the proposed study and expose undergraduate students to the excitement of doing research. The involvement of Ph.D. stdudents is vital for the success of this project. I am currently advising two PhD students, one of whom is working on some of the problems outlined in this proposal. The educational component of this proposal also involves a combination of a redesign of existing courses and the introduction of new ones, distilling ideas developed in current research.

Discrete Mathematics

Beginning Fall 2002, I will be teaching the Discrete Mathematics course at Princeton. I intend to revamp the course organization significantly. This has traditionally been a problematic course because of the difficulty in motivating students to learn the course material. My goal will be to bring out the applications of discrete mathematics via examples that students can relate to and appreciate. For the more advanced students, I would like to point to research papers where concepts they learn in the course are used. Topics that I particularly want to emphasize are proof techniques, probability and linear algebra. I think an important concept that people ought to learn from this course is the notion of a rigorous proof and how one goes about writing correct proofs. The course should also familiarize students with thinking abstractly. In teaching subjects such as probability and linear algebra, I plan to draw examples from the web and link analysis to bring out the potential applications and strengths of these techniques. I also plan to incorporate the use of mathematical software such as Mathematica or Maple into the course contents - I think that the use of such software packages can be a very valuable addition to the discrete mathematics toolkit and a useful aid in enhancing ones ability to think about and analyze problems. The goal of the course will be to equip students with the basic tools and techniques from discrete mathematics that all computer science majors should know and can expect to use. At the same time the course should be able to whet the appetitites of the more mathematically inclined students and provide a solid foundation for further study in theoretical computer science.

Algorithms for Large Data Sets

In Spring 2002, I designed and taught an advanced undergraduate course about new algorithmic techniques for large data sets. (`http://www.cs.princeton.edu/courses/archive/spring02/cs493/`). The course covered a combination of classical topics as well as a selection of some of the interesting ideas that have arisen in recent research.

The course came out of my belief that at the heart of many of the recent developments of algorithms for large data sets, lie simple basic principles and techniques that can be understood and appreciated by anyone with a reasonable mathematical background. Indeed, the first offering of the class attracted about 20 students with 10 of them outside of theory. This class was designed to serve as a gentle introduction to advanced algorithmic ideas that they would not learn in standard algorithms classes and that would potentially be useful to them in their individual fields of interest. Course notes are available off the course web page and should serve as a valuable resource. These will be edited, embellished and updated with future offerings of the course and finally be turned into a book.

The course will be developed over several years. I expect the next offering of the course to be a graduate class, oriented around research projects. The goal will be to get students to work on topics that could potentially be turned into research papers.Software developed as a result of these projects will be made freely available to other researchers who may want to use it for experiments.

New Graduate Classes

Together with the other theory faculty at Princeton, I am involved in a redesign of the graduate theory curriculum. The number of theory faculty at Princeton gives us the flexibility to design and offer innovative graduate classes. So far, our theory curriculum does not include graduate classes on specialized topics such as Approximation Algorithms, Randomized Algorithms, Online Algorithms, Communication Complexity and so on. While students are exposed to these topics in various courses, there is no single course that covers a particular topic in detail. These and other sub-areas in theoretical computer science have been around long enough and have matured to the point where an entire course can be devoted towards covering the foundational results in the area. In preparing our graduate students, I believe it is important to expose them to such an exhaustive coverage of specific areas. This will allow them to draw on ideas outside of their immediate research area in conducting their own research. Hopefully, this will also encourage them to work on problems outside their immediate area of focus and broaden their research profile. The addition of all these new courses cannot be done single handedly and will certainly be done with the collaboration of my colleagues. My contribution towards these new classes would be the introduction of a new graduate class on Approximation Algorithms. In later years, I will also teach graduate classes on Randomized Algorithms and On-line Algorithms. Below, I elaborate on my plans for an Approximation Algorithms class.

Approximation Algorithms

In Fall 01, I taught a graduate seminar on Approximation Algorithms (see the webpage at `http://www.cs.princeton.edu/courses/archive/fall01/cs593/`) which consisted of 50% lectures and 50% student presentations. My plan is to alternate offerings of such a graduate seminar with a regular course on Approximation Algorithms. The regular course will build up a basic background in approximation algorithms, covering classical problems and illustrating different techniques through carefully chosen problems. Two optimization problems I plan to use as examples are facility location and k-median. Both these problems have been the subject of active research recently and are excellent examples to illustrate the application of different algorithmic techniques that can be used to develop approximation algorithms, including LP-rounding [STA97, CGTS99], the primal-dual method [JV99], local search and greedy methods [CG99, AGK+01], Lagrangian relaxation [JV99], and dual fitting [MMSV01]. The seminar will focus on more advanced topics and involve student presentations of relatively new research papers. My goal will be to select papers that lead to interesting open problems and encourage students to work on these. The course should be useful for theory students as well as mathematically inclined students from other areas who want to understand the general tools and techniques developed in the field of approximation algorithms. On the other hand, the seminar should be useful for students who are considering doing research in approximation algorithms.

References

[ABC+99] F. N. Afrati, E. Bampis, C. Chekuri, D. R. Karger, C. Kenyon, S. Khanna, I. Milis, M. Queyranne, M. Skutella, C. Stein, and M. Sviridenko. Approximation Schemes for Minimizing Average Weighted Completion Time with Release Dates. *Proc. 40th FOCS*, pp. 32–44, 1999.

[ACLY00] R. Ahlswede, N. Cai, S.-Y. R. Li and R. W. Yeung. Network information flow. *IEEE Transactions on Information Theory*, 46(4):1024–1016, 2000.

[ACM01] S. Albers, M. Charikar, and M. Mitzenmacher. Delayed Information and Action in On-Line Algorithms. *Proc. Information and Computation*, 170(2): 135–152, 2001.

[AK98] N. Alon and N. Kahale. Approximating the Independence Number via the θ-function. *Mathematical Programming*, 80: 253–264, 1998.

[ASZ01] N. Alon, B. Sudakov and U. Zwick. Constructing worst case instances for semidefinite programming based approximation algorithms. *Proc. 12th SODA*, pp. 92–100, 2001.

[Arc01] A. Archer. Two $O(\log^* k)$-Approximation Algorithms for the Asymmetric k-Center Problem. *Proc. IPCO*, pp. 1–14, 2001.

[Aro98] S. Arora. Polynomial Time Approximation Schemes for Euclidean Traveling Salesman and other Geometric Problems. *JACM* 45(5): 753–782, 1998.

[ABL02] S. Arora, B. Bollobas and L. Lovasz. Proving Integrality Gaps without Knowing the Linear Program. To appear in *Proc. 43rd FOCS*, 2002.

[AFK96] S. Arora, A. M. Frieze, and H. Kaplan. A New Rounding Procedure for the Assignment Problem with Applications to Dense Graph Arrangement Problems. *Proc. 37th FOCS*, pp. 21-30, 1996.

[AKK99] S. Arora, D. R. Karger, and M. Karpinski. Polynomial Time Approximation Schemes for Dense Instances of NP-Hard Problems. *JCSS* 58(1): 193–210, 1999.

[ARR98] S. Arora, P. Raghavan, and S. Rao. Approximation Schemes for Euclidean k-Medians and Related Problems. *Proc. STOC*, pp. 106-113, 1998.

[AGK+01] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristic for k-median and facility location problems. *Proc. 33rd STOC*, pp. 21-29, 2001.

[AR98] Y. Aumann and Y. Rabani. An $O(\log k)$ Approximate Min-Cut Max-Flow Theorem and Approximation Algorithm. *SIAM J. Comput.* 27(1): 291–301, 1998.

[AA97] B. Awerbuch and Y. Azar. Buy-at-Bulk Network Design. *Proc. 38th FOCS*, pp. 542–547, 1997.

[Bar96] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic application. *Proc. 37th FOCS*, pp. 184–193, 1996.

[Bar98] Y. Bartal. On approximating arbitrary metrics by tree metrics. In Proc. 30th STOC, pp. 161–168, 1998.

[BCI01] Y. Bartal, M. Charikar, and P. Indyk. On page migration and other relaxed task systems. *TCS* 268(1): 43–66, 2001.

[BCR01] Y. Bartal, M. Charikar, and D. Raz, Approximating min-sum k-clustering in metric spaces. *Proc. 33rd STOC*, pp. 11-20, 2001.

[BE81] R. Bar-Yehuda and S. Even. A Linear-Time Approximation Algorithm for the Weighted Vertex Cover Problem. *J. Algorithms*, 2(2): 198–203, 1981.

[BE85] R. BarYehuda and S. Even. A localratio theorem for approximating the weighted vertex cover problem. *Annals of Discrete Mathematics*, 25: 27-45, 1985.

[BS97] B. Berger and P. W. Shor. Tight Bounds for the Maximum Acyclic Subgraph Problem. *J. Algorithms* 25(1): 1-18, 1997.

[BCK00] P. Berman, M. Charikar, and M. Karpinski. On-Line Load Balancing for Related Machines. *J. Algorithms* 35(1): 108–121, 2000.

[Blu94] A. Blum. New approximation algorithms for graph coloring. *JACM*, 41: 470–516, 1994.

[BK97] A. Blum and D. R. Karger. An $\tilde{O}(n^{3/14})$-Coloring Algorithm for 3-Colorable Graphs. *Information Processing Letters* 61(1): 49–53, 1997.

[BKRV00] A. Blum, G. Konjevod, R. Ravi, and S. Vempala. Semi-definite relaxations for minimum bandwidth and other vertex-ordering problems. *TCS* 235(1): 25–42, 2000.

[Bor33] K. Borsuk. Drei Sätze über die n-dimensionale euklidische Sphäre. *Fundamenta Math.*, 20: 177–190, 1933.

[Bou85] J. Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985.

[BT94] B. Braschi and D. Trystram. A New Insight into the Coffman-Graham Algorithm. *SIAM J. Comput.*, 23(3): 662–669, 1994.

[Bro97] A. Z. Broder. On the resemblance and containment of documents. *Proc. Compression and Complexity of SEQUENCES*, pp. 21–29. IEEE Computer Society, 1997.

[Bro98] A. Z. Broder. Filtering near-duplicate documents. *Proc. FUN 98*, 1998.

[BCFM00] A. Z. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher. Min-wise independent permutations. *Journal of Computer Systems and Sciences*, vol. 60(3), pp. 630-659, 2000 (special issue for STOC '98).

[BGMZ97] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the Web. *Proc. 6th Int'l World Wide Web Conference*, pp. 391–404, 1997.

[CKR00] G. Calinescu, H. J. Karloff, and Y. Rabani. An Improved Approximation Algorithm for MULTIWAY CUT. *JCSS* 60(3): 564–574, 2000.

[CKR01] G. Calinescu, H. J. Karloff, and Y. Rabani. Approximation algorithms for the 0-extension problem. *Proc. SODA*, pp. 8–16, 2001.

[CV00] R. D. Carr and S. Vempala. Towards a $4/3$ approximation for the asymmetric traveling salesman problem. *Proc. 11th SODA*, pp. 116-125, 2000.

[CY02] T. H. Chan and R. W. Yeung. On a relation between information inequalities and group theory. *IEEE Transactions on Information Theory*, vol. 48, no. 7, 2002.

[Cha95] M. Charikar. Approximation Algorithms for Problems in Combinatorial Optimization. B.Tech. Project Report, 1995.

[Cha02a] M. Charikar. On semidefinite relaxations for graph coloring and vertex cover. *Proc. SODA*, 2002.

[Cha02b] M. Charikar. Similarity Estimation Techniques from Rounding Algorithms. *Proc. 34th STOC*, 2002.

[CCMN00] M. Charikar, S. Chaudhuri, R. Motwani, and V. R. Narasayya Towards Estimation Error Guarantees for Distinct Values. *Proc. 19th PODS*, pp. 268-279, 2000.

[CCC+99] M. Charikar, C. Chekuri, T. Cheung, Z. Dai, A. Goel, S. Guha, and M. Li. Approximation Algorithms for Directed Steiner Problems. *J. Algorithms* 33(1): 73–91, 1999.

[CCFM97] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. *Proc. 29th STOC*, pp. 626–635, 1997.

[CCGG98] M. Charikar, C. Chekuri, A. Goel, and S. Guha. Rounding via Trees: Deterministic Approximation Algorithms for Group Steiner Trees and k-Median. *Proc. 30th STOC*, pp. 114-123, 1998.

[CCG+98] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. A. Plotkin. Approximating a Finite Metric by a Small Number of Tree Metrics. *Proc. 39th FOCS*, pp. 379–388, 1998.

[CCF02] M. Charikar, K. Chen and M. Farach-Colton. Finding Frequent Items in Data Streams. *Proc. 29th ICALP*, pp. 693-703, 2002.

[CFG+00] M. Charikar, R. Fagin, V. Guruswami, J. M. Kleinberg, P. Raghavan, and A. Sahai. Query strategies for priced information. *Proc. 32nd STOC*, pp. 582–591, 2000.

[CG99] M. Charikar and S. Guha. Improved combinatorial algorithms for the facility location and k-median problems. *Proc. 40th FOCS*, pp. 378–388, 1999.

[CGTS99] M. Charikar, S. Guha, E. Tardos, and D. Shmoys. A constant factor approximation algorithm for the k-median problem. *Proc. 31st STOC*, pp. 1–10, 1999.

[CGK+00] M. Charikar, V. Guruswami, R. Kumar, S. Rajagopalan, and A. Sahai, Combinatorial feature selection problems. *Proc. 41st FOCS* pp. 631–640, 2000.

[CIP02] M. Charikar, P. Indyk and R. Panigrahy. New Algorithms for Subset Query, Partial Match, Orthogonal Range Searching and Related Problems. *Proc. 29th ICALP*, pp. 451–462, 2002.

[CKMN01] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. *Proc. 12th SODA* pp. 642–651, 2001.

[CKR98] M. Charikar, S. Khuller, and B. Raghavachari. Algorithms for Capacitated Vehicle Routing. *Proc. 30th STOC*, pp. 349–358, 1998.

[CP01] M. Charikar and R. Panigrahy. Clustering to minimize the sum of cluster diameters. *Proc. 33rd STOC*, pp. 1–10, 2001.

[CPO02] M. Charikar, R. Panigrahy and L. O´Callaghan. Better Streaming Algorithms for Clustering Problems. manuscript.

[CR98] M. Charikar and B. Raghavachari. The Finite Capacity Dial-A-Ride Problem. *Proc. 39th FOCS*. pp. 458–467, 1998.

[CB98] C. Chekuri and M. Bender. An efficient approximation algorithm for minimizing makespan on uniformly related machines. *Proc. 6th IPCO*, 1998.

[CEK02] C. Chekuri, G. Even, and G. Kortsarz. An approximation algorithm for the group Steiner problem. Manuscript. (Preliminary version appeared in SODA 2002), 2002.

[CGN+02] Chandra Chekuri, A. Gupta, I. Newman, Y. Rabinovich, and A. Sinclair. Embedding k-Outerplanar graphs into ℓ_1. manuscript, 2002.

[CKNZ01] C. Chekuri, S. Khanna, J. Naor, and L. Zosin. Approximation algorithms for the metric labeling problem via a new linear programming formulation. *Proc. 12th SODA*, pp. 109-118, 2001.

[CJK+01] Z. Chen, H. V. Jagadish, F. Korn, N. Koudas, S. Muthukrishnan, R. T. Ng, and D. Srivastava. Counting Twig Matches in a Tree. *Proc. 17th ICDE* pp. 595-604. 2001.

[CKKS00] Z. Chen, F. Korn, N. Koudas, and S. Muthukrishnan. Selectivity Estimation for Boolean Queries. *Proc. 19th PODS*, pp. 216-225, 2000.

[CS98a] B. Chor and M. Sudan. A Geometric Approach to Betweenness. *SIAM Journal on Discrete Mathematics*, 11(4): 511–523, 1998.

[Chr76] N. Christofides. Worst-case analysis of a new heuristic for the travelling salesman problem. Report 388, Graduate School of Industrial Administration, CMU, 1976.

[Chu98] F. A. Chudak. Improved approximation algorithms for uncapacitated facility location. In E. A. Boyd R. E. Bixby and R. Z. Ríos-Mercado, editors, *Proc. 6th IPCO*, LNCS, 1412:180–194. Springer, 1998.

[CS98b] F. A. Chudak and D. Shmoys. Improved approximation algorithms for the uncapacitated facility location problem. Unpublished manuscript, 1998.

[CS99] F. A. Chudak and D. B. Shmoys. Approximation Algorithms for Precedence-Constrained Scheduling Problems on Parallel Machines that Run at Different Speeds. *J. Algorithms*, 30(2): 323–343, 1999.

[CG72] E. G. Coffman, and R. L. Graham. Optimal Scheduling for Two-Processor Systems. *Acta Informatica*, 1:200–213, 1972.

[CDF+00] E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. D. Ullman, and C. Yang. Finding Interesting Associations without Support Pruning. *Proc. 16th ICDE* pp. 489-499, 2000.

[CG97a] S. Cohen and L. Guibas. The Earth Mover's Distance under Transformation Sets. *Proc. 7th IEEE Intnl. Conf. Computer Vision*, 1999.

[CG97b] S. Cohen and L. Guibas. The Earth Mover's Distance: Lower Bounds and Invariance under Translation. Tech. report STAN-CS-TR-97-1597, Dept. of Computer Science, Stanford University, 1997.

[DKKR02] W. F. de la Vega, M. Karpinski, C. Kenyon and Y. Rabani. Polynomial Time Approximation Schemes for Metric Min-Sum Clustering. *ECCC*, Report TR02-025.

[DS02] I. Dinur and S. Safra. The Importance of Being Biased. *Proc. STOC*, 2002.

[DV01] J. Dunagan and S. Vempala. On Euclidean Embeddings and Bandwidth Minimization. *Proc. RANDOM-APPROX*, pp. 229–240, 2001.

[DKNS01] C. Dwork, S. R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the Web. *Proc. 10th WWW*, pp. 613–622, 2001.

[ENSS98] G. Even, J. Naor, B. Schieber, and M. Sudan. Approximating Minimum Feedback Sets and Multicuts in Directed Graphs. *Algorithmica* 20(2): 151–174, 1998.

[ENRS00] G. Even, J. Naor, S. Rao, B. Schieber. Divide-and-conquer approximation algorithms via spreading metrics. *JACM* 47(4): 585–616, 2000.

[Fei97] U. Feige. Randomized Graph Products, Chromatic Numbers, and the Lovász θ-function. *Combinatorica*, 17(1): 79–90, 1997.

[FK98] U. Feige and J. Kilian. Zero Knowledge and the Chromatic Number. *JCSS*, 57(2): 187–199, 1998.

[FLS02] U. Feige, M. Langberg, and G. Schechtman. Graphs with Tiny Vector Chromatic Numbers and Huge Chromatic Numbers. To appear in *Proc. 43rd FOCS*, 2002.

[FK02] U. Feige and R. Krauthgamer. A polylogarithmic approximation of the minimum bisection. *SIAM Journal on Computing*, 31(4):1090–1118, 2002.

[FS01] U. Feige and G. Schechtman. On the Optimality of the Random Hyperplane Rounding Technique for MAX CUT. *Proc. 33rd STOC*, pp. 433–442, 2001.

[FM88] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Comb. Theory, Ser. B*, 44:355–362, 1988.

[FR87] P. Frankl and V. Rödl. Forbidden Intersections. *Transactions of the American Mathematical Society*, 300: 259–286, 1987.

[FW81] P. Frankl and R. Wilson. Intersection theorems with geometric consequences. *Combinatorica*, 1: 357–368, 1981.

[FK00] A. Freund and H. J. Karloff. A lower bound of $8/(7+(1/k)-1)$ on the integrality ratio of the Calinescu-Karloff-Rabani relaxation for multiway cut. *Information Processing Letters* 75(1-2): 43-50, 2000.

[FGM82] A. M. Frieze, G. Galbiati and F. Maffioli. On the worst-case performance of some algorithms for the asymmetric traveling salesman problem. *Networks*, 12: 23–39, 1982.

[FKR95] A. M. Frieze, R. M. Karp, and B. A. Reed. When is the Assignment Bound Tight for the Asymmetric Traveling-Salesman Problem? *SIAM J. Comput.*, vol 24(3), pp. 484-493, 1995.

[FS01b] A. M. Frieze, G. B. Sorkin. The probabilistic relationship between the assignment and asymmetric traveling salesman problems. *Proc. 12th SODA*, pp. 652-660, 2001.

[FJ97] A. M. Frieze, M. Jerrum. Improved Approximation Algorithms for MAX k-CUT and MAX BISECTION. *Algorithmica*, vol. 18(1), pp. 67-81, 1997.

[FKN69] M. Fuji, T. Kasami and K. Ninomiya. Optimal sequencing of two equivalent processors. *SIAM J. Appl. Math.*, 17(3):784–789, 1969.

[GVY96] N. Garg, V. V. Vazirani, and M. Yannakakis. Approximate Max-Flow Min-(Multi)Cut Theorems and Their Applications. *SIAM J. Comput.*, 25(2): 235–251, 1996.

[GGK01] A. Gionis, D. Gunopulos, and N. Koudas. Efficient and Tunable Similar Set Retrieval. *Proc. SIGMOD Conference* 2001.

[GT00b] M. Goemans and L. Tuncel. When does the positive semidefiniteness constraint help in lifting procedures. *Mathematics of Operations Research*, to appear.

[GW95] M. X. Goemans and D. P. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *JACM* 42(6): 1115-1145, 1995.

[Gon85] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Comput. Sci.* 38: 293–306, 1985.

[Gra66] R. L. Graham. Bounds for certain multiprocessor anomalies. *Bell Systems Technical Journal*, 45:1563–1581, 1966.

[GJR85a] M. Grötschel, M. Jünger, and G. Reinelt. On the Maximum Acyclic Subgraph Polytope. *Mathematical Programming*, 33: 28–42, 1985.

[GJR85b] M. Grötschel, M. Jünger, and G. Reinelt. Facets of the Linear Ordering Polytope. *Mathematical Programming*, 33: 43–60, 1985.

[GLS93] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer Verlag, 1993.

[GK98] S. Guha and S. Khuller. Greedy strikes back: Improved facility location algorithms. *Proc. 9th SODA*, pp. 649–657, 1998.

[GMMO00] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. *Proc. 41st FOCS*, pp. 359–366, 2000.

[GMM01] S. Guha, A. Meyerson, and K. Munagala. A constant factor approximation for the single sink edge installation problems. *Proc. 33rd STOC*, pp. 383–388, 2001.

[GNRS99] A. Gupta, I. Newman, Y. Rabinovich and A. Sinclair. Cuts, Trees and ℓ_1-Embeddings of Graphs. *Proc. FOCS*, pp. 399–409, 1999.

[GT00] A. Gupta and Éva Tardos. A constant factor approximation algorithm for a class of classification problems. *Proc. 32nd STOC*, pp. 652–658, 2000.

[GI02] V. Guruswami and P. Indyk. Embeddings into Normed Spaces and Hardness of Approximation for Geometric Problems. manuscript.

[GK00] V. Guruswami and S. Khanna. On the Hardness of 4-coloring a 3-colorable graph. *Proc. Annual IEEE Conf. Comput. Complexity*, pp. 188–197, 2000.

[Hal00] E. Halperin. Improved approximation algorithms for the vertex cover problem in graphs and hypergraphs. *Proc. 11th SODA*, pp. 329–337, 2000.

[HNZ01] E. Halperin, R. Nathaniel, and U. Zwick. Coloring k-colorable graphs using relatively small palettes. *Proc. 12th SODA*, pp. 319–326, 2001.

[Has01] J. Håstad. Some optimal inapproximability results. *JACM*, 48(4): 798–859, 2001.

[HGI00] T. H. Haveliwala, A. Gionis, and P. Indyk. Scalable Techniques for Clustering the Web. *Proc. 3rd WebDB*, pp. 129-134, 2000.

[HK70] M. Held and R. M. Karp. The traveling salesman problem and minimum spanning trees. *Operations Research*, 18: 1138-1162, 1970.

[Hoc01] D. S. Hochbaum. An efficient algorithm for image segmentation, Markov random fields and related problems. *JACM*, 48(4):686–701, 2001.

[HS85] D. S. Hochbaum and D. B. Shmoys. A best possible approximation algorithm for the k-center problem. *Mathematics of Operations Research*, 10:180–184, 1985.

[HS86] D. S. Hochbaum and D. B. Shmoys. A unified approach to approximation algorithms for bottleneck problems. *JACM*, 33(3): 533–550, 1986.

[Ind99] P. Indyk. A Sublinear Time Approximation Scheme for Clustering in Metric Spaces. *Proc. FOCS*, pp. 154–159, 1999.

[Jaf80] J. Jaffe. Efficient scheduling of tasks without full use of processor resources. *Theoretical Computer Science*, 12:1–17, 1980.

[Jain01] K. Jain. A Factor 2 Approximation Algorithm for the Generalized Steiner Network Problem. *Combinatorica* 21(1): 39–60, 2001.

[JV99] K. Jain and V. Vazirani. Approximation algorithms for metric facility location and k-Median problems using the primal-dual schema and Lagrangian relaxation. *JACM* 48(2): 274-296, 2001 (prelim. version in *Proc. 40th FOCS*, pp. 2–13, 1999.

[JL84] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[KK93] J. Kahn and G. Kalai. A Counterexample to Borsuk's Conjecture. *Bulletin of the American Mathematical Society*, 29(1): 60–62, 1993.

[KVV00] R. Kannan, S. Vempala, and A. Vetta. On Clusterings - Good, Bad and Spectral. *Proc. 41st FOCS*, pp. 367-377, 2000.

[KKS+99] D. R. Karger, P. N. Klein, C. Stein, M. Thorup, and N. E. Young. Rounding Algorithms for a Geometric Embedding of Minimum Multiway Cut. *Proc. STOC*, pp. 668–678, 1999.

[KMS98] D. R. Karger, R. Motwani, and M. Sudan. Approximate Graph Coloring by Semidefinite Programming. *JACM* 45(2): 246–265, 1998.

[KSW98] D. R. Karger, C. Stein and J. Wein. Scheduling Algorithms. In *Algorithms and Theory of Computation Handbook*, CRC Press, 1998.

[KH79] O. Kariv and S. L. Hakimi. An algorithmic approach to network location problems, part ii: p-medians. *SIAM Journal of Appl. Math*, 37:539–560, 1979.

[Kar99] H. Karloff. How Good is the Goemans-Williamson MAX CUT Algorithm ? *SIAM Journal on Computing*, 39(1): 336–350, 1999.

[KLS00] S. Khanna, N. Linial, and S. Safra. On the Hardness of Approximating the Chromatic Number. *Combinatorica*, 20: 393–415, 2000.

[Kho01] S. Khot. Improved Inaproximability Results for MaxClique, Chromatic Number and Approximate Graph Coloring. *Proc. FOCS*, pp. 600-609, 2001.

[Kho02] On the power of unique 2-prover 1-round games. *Proc. STOC*, 2002.

[KPR93] P. Klein, S. PLotkin, and S. Rao. Excluded Minors, Network Decomposition, and Multicommodity Flow. *Proc. STOC*, pp. 682–690, 1993.

[KG98] J. Kleinberg and M. Goemans. The Lovasz Theta Function and a Semidefinite Programming Relaxation of Vertex Cover. *SIAM Journal on Discrete Mathematics*, 11(2): 196–204, 1998.

[KT99] J. M. Kleinberg and Éva Tardos Approximation Algorithms for Classification Problems with Pairwise Relationships: Metric Labeling and Markov Random Fields. *Proc. 40th FOCS*, pp. 14–23, 1999.

[Knu94] D. Knuth. The Sandwich Theorem. *The Electronic Journal of Combinatorics*, 1:1–48, 1994.

[KM02] R. Koetter and M. Medard. An Algebraic Approach to Network Coding. *Proc. INFOCOM*, 2002.

[KPR98] M. Korupolu, C. G. Plaxton, and R. Rajaraman. Analysis of a local search heuristic for facility location problems. *Proc. 9th SODA*, pages 1–10, 1998.

[LS77] S. Lam and R. Sethi. Worst Case Analysis of Two Scheduling Algorithms. *SIAM J. Comput.*, 6(3): 518–536, 1977.

[Lau01] M. Laurent. A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming. Technical Report PNA-R0108, CWI, Amsterdam, 2001.

[LR88] F. T. Leighton and S. Rao. An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. *Proc. 29th FOCS*, pp. 422–431, 1988.

[LR99] F. T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *JACM*, 46(6):787–832, 1999.

[LYC02] S.-Y. R. Li, R. W. Yeung, and N. Cai. Linear network coding. to appear in *IEEE Transactions on Information Theory*.

[LK73] S. Lin and B. W. Kernighan. An effective heuristic algorithm for the traveling salesman problem. *Operations Research*, 21:498–516, 1973.

[LV92b] J.-H. Lin and J. S. Vitter. Approximation algorithms for geometric median problems. *Information Processing Letters*, 44:245–249, 1992.

[LV92a] J.-H. Lin and J. S. Vitter. ϵ-approximations with minimum packing constraint violation. *Proc. 24th STOC*, pp. 771–782, 1992.

[LLR95] N. Linial, E. London, and Y. Rabinovich. The Geometry of Graphs and Some of its Algorithmic Applications. *Combinatorica* 15(2): 215–245, 1995.

[Lov79] L. Lovász. On the Shannon Capacity of a Graph. *IEEE Transactions on Information Theory*, 25: 1–7, 1979.

[LS91] L. Lovász and A. Schrijver. Cones of matrices and set-functions and $0-1$ optimization. *SIAM Journal on Optimization*, 1: 166–190, 1991.

[MMSV01] M. Mahdian, E. Markakis, A. Saberi, and V. V. Vazirani. A Greedy Facility Location Algorithm Analyzed Using Dual Fitting. *Proc. RANDOM-APPROX*, pp. 127–137, 2001.

[MS85] B. Monien and E. Speckenmeyer, Ramsey numbers and an approximation algorithm for the vertex cover problem. *Acta Inf.* 22: 115–123, 1985.

[NZ97] J. Naor and L. Zosin. A 2-Approximation Algorithm for the Directed Multiway Cut Problem. *Proc. 38th FOCS*, pp. 548–553, 1997.

[New01] A. Newman. The Maximum Acyclic Subgraph Problem and Degree-3 Graphs. *Proc. RANDOM-APPROX*, pp. 147–158, 2001.

[NV01] A. Newman and S. Vempala. Fences Are Futile: On Relaxations for the Linear Ordering Problem. *Proc. IPCO*, pp. 333–347, 2001.

[PV00] C. Papadimitriou and S. Vempala. On the approximability of the traveling salesman problem (extended abstract). *Proc. 32nd STOC*. pp. 126–133, 2000.

[PV98] R. Panigrahy and S. Vishwanathan. An $O(\log^* n)$ Approximation Algorithm for the Asymmetric p-Center Problem. *J. Algorithms* 27(2): 259–268, 1998.

[Pip80] N. Pippenger. One the evaluation of powers and monomials. *SIAM Journal on Computing*, 9(2): 230–250, 1980.

[Ran01] A. Ranade. Scheduling Loosely Connected Task Graphs. manuscript, 2001.

[Rao99] S. Rao. Small Distortion and Volume Preserving Embeddings for Planar and Euclidean Metrics. *Proc. Symp. Comp. Geom.*, pp. 300–306, 1999.

[Rub99] Y. Rubner. Perceptual Metrics for Image Database Navigation. Phd Thesis, Stanford University, May 1999

[RGT97] Y. Rubner, L. J. Guibas, and C. Tomasi. The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval. *Proc. of the ARPA Image Understanding Workshop*, pp. 661-668, 1997.

[RT98] Y. Rubner, C. Tomasi. Texture Metrics. *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, 1998, pp. 4601-4607.

[RTG98a] Y. Rubner, C. Tomasi, and L. J. Guibas. A Metric for Distributions with Applications to Image Databases. *Proc. IEEE Int. Conf. on Computer Vision*, pp. 59-66, 1998.

[RTG98b] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. Tech. Report STAN-CS-TN-98-86, Dept. of Computer Science, Stanford University, 1998.

[RT99a] M. Ruzon and C. Tomasi. Color edge detection with the compass operator. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2:160-166, 1999.

[RT99b] M. Ruzon and C. Tomasi. Corner detection in textured color images. *Proc. IEEE Int. Conf. Computer Vision*, 2:1039-1045, 1999.

[SSZ02] M. Saks, A. Samorodnitsky, and L. Zosin. A lower bound on the integrality gap for minimum multicut in directed graphs. *Combinatorica*, to appear.

[SCRS97] F. S. Salman, J. Cheriyan, R. Ravi, and S. Subramanian. Buy-at-Bulk Network Design: Approximating the Single-Sink Edge Installation Problem. *Proc. 8th SODA*, pp. 619–628, 1997.

[Sey95] P. D. Seymour. Packing directed circuits fractionally. *Combinatorica*, 15(2):281–288, 1995.

[Sch79] A. Schrijver. A Comparison of the Delsarte and Lovász Bounds. *IEEE Transactions on Information Theory*, 25:425–429, 1979.

[SW99] P. Schuurman and G. Woeginger. Polynomial time approximation algorithms for machine scheduling: Ten open problems. *Journal of Scheduling*, 2:203–213, 1999.

[Shm97] D. B. Shmoys. Cut problems and their application to divide-and-conquer. In *Approximation Algorithms for NP-hard Problems*, D. S. Hochbaum, ed., PWS, pp. 192–235, 1997.

[STA97] D. B. Shmoys, É. Tardos, and K. I. Aardal. Approximation algorithms for facility location problems. *Proc. 29th STOC*, pp. 265–274, 1997.

[Sze94] M. Szegedy. A Note on the θ Number of Lovász and the Generalized Delsarte Bound. *Proc. 35th FOCS*, pp. 36–39, 1994.

[Tal02] K. Talwar. The Single-Sink Buy-at-Bulk LP Has Constant Integrality Gap. *Proc. 9th IPCO*, pp. 475–486, 2002.

[VY99] S. Vempala and M. Yannakakis. A Convex Relaxation for the Asymmetric TSP. *Proc. 10th SODA*, pp. 975-976, 1999.

[Yao76] A. C. Yao. On the evaluation of powers. *SIAM Journal on Computing*, 5(1): 100–103, 1976.

[YZ01] R. W. Yeung and Z. Zhang. A class of non-Shannon-type information inequalities and their applications. *Communications in Information and Systems*, 1:87–100, 2001.

[ZY97] Z. Zhang and R. W. Yeung. A non-Shannon-type conditional inequality of information quantities. *IEEE Trans. Inform. Theory*, 43:1982–1985, 1997.

[ZY98] Z. Zhang and R. W. Yeung. On characterization of entropy function via information inequalities. *IEEE Trans. Inform. Theory*, 44:1440–1452, 1998.

[ZK02] L. Zosin and S. Khuller. On directed Steiner trees. *Proc. SODA*, 2002.