The Design and Analysis of Algorithms

Dexter C. Kozen
Cornell University

December 1990

To my wife Fran
and my sons
Alexander, Geoffrey, and Timothy

Preface

These are my lecture notes from **CS681: Design and Analysis of Algorithms**, a one-semester graduate course I taught at Cornell for three consecutive fall semesters from '88 to '90. The course serves a dual purpose: to cover core material in algorithms for graduate students in computer science preparing for their PhD qualifying exams, and to introduce theory students to some advanced topics in the design and analysis of algorithms. The material is thus a mixture of core and advanced topics.

At first I meant these notes to supplement and not supplant a textbook, but over the three years they gradually took on a life of their own. In addition to the notes, I depended heavily on the texts

- A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms.* Addison-Wesley, 1975.
- M. R. Garey and D. S. Johnson, *Computers and Intractibility: A Guide to the Theory of NP-Completeness.* W. H. Freeman, 1979.
- R. E. Tarjan, *Data Structures and Network Algorithms.* SIAM Regional Conference Series in Applied Mathematics 44, 1983.

and still recommend them as excellent references.

The course consists of 40 lectures. The notes from these lectures were prepared using scribes. At the beginning of each lecture, I would assign a scribe who would take notes for the entire class and prepare a raw LaTeX source, which I would then doctor and distribute. In addition to the 40 lectures, I have included 10 homework sets and several miscellaneous homework exercises, all with complete solutions. The notes that were distributed are essentially as they appear here; no major reorganization has been attempted.

There is a wealth of interesting topics, both classical and current, that I would like to have touched on but could not for lack of time. Many of these, such as computational geometry and factoring algorithms, could fill an entire semester. Indeed, one of the most difficult tasks was deciding how best to spend a scant 40 lectures.

I wish to thank all the students who helped prepare these notes and who kept me honest: Mark Aagaard, Mary Ann Branch, Karl-Friedrich Böhringer, Thomas Bressoud, Suresh Chari, Sofoklis Efremidis, Ronen Feldman, Ted

Fischer, Richard Huff, Michael Kalantar, Steve Kautz, Dani Lischinski, Peter Bro Miltersen, Marc Parmet, David Pearson, Dan Proskauer, Uday Rao, Mike Reiter, Gene Ressler, Alex Russell, Laura Sabel, Aravind Srinivasan, Sridhar Sundaram, Ida Szafranska, Filippo Tampieri, and Sam Weber. I am especially indebted to my teaching assistants Mark Novick (fall '88), Alessandro Panconesi (fall '89), and Kjartan Stefánsson (fall '90) for their help with proofreading, preparation of solution sets, and occasional lecturing. I am also indebted to my colleagues László Babai, Gianfranco Bilardi, Michael Luby, Keith Marzullo, Erik Meineche Schmidt, Bernd Sturmfels, Éva Tardos, Steve Vavasis, Sue Whitesides, and Rich Zippel for valuable comments and interesting exercises. Finally, I wish to express my sincerest gratitude to my colleague Vijay Vazirani, who taught the course in fall '87 and who was an invaluable source of help.

I would be most grateful for any suggestions or criticism from readers.

Cornell University
Ithaca, NY

Dexter Kozen
December 1990

Contents

I Lectures

Lecture 1 Algorithms and Their Complexity

This is a course on the design and analysis of algorithms intended for first-year graduate students in computer science. Its purposes are mixed: on the one hand, we wish to cover some fairly advanced topics in order to provide a glimpse of current research for the benefit of those who might wish to specialize in this area; on the other, we wish to introduce some core results and techniques which will undoubtedly prove useful to those planning to specialize in other areas.

We will assume that the student is familiar with the classical material normally taught in upper-level undergraduate courses in the design and analysis of algorithms. In particular, we will assume familiarity with:

- sequential machine models, including Turing machines and random access machines (RAMs)
- discrete mathematical structures, including graphs, trees, and dags, and their common representations (adjacency lists and matrices)
- fundamental data structures, including lists, stacks, queues, arrays, balanced trees
- fundamentals of asymptotic analysis, including $O(\cdot)$, $o(\cdot)$, and $\Omega(\cdot)$ notation, and techniques for the solution of recurrences
- fundamental programming techniques, such as recursion, divide-and-conquer, dynamic programming
- basic sorting and searching algorithms.

These notions are covered in the early chapters of [3, 39, 100].

Familiarity with elementary algebra, number theory, and discrete probability theory will be helpful. In particular, we will be making occasional use of the following concepts: linear independence, basis, determinant, eigenvalue, polynomial, prime, modulus, Euclidean algorithm, greatest common divisor, group, ring, field, random variable, expectation, conditional probability, conditional expectation. Some excellent classical references are [69, 49, 33].

The main emphasis will be on *asymptotic worst-case complexity*. This measures how the worst-case time or space complexity of a problem grows with the size of the input. We will also spend some time on probabilistic algorithms and analysis.

1.1 Asymptotic Complexity

Let f and g be functions $\mathcal{N} \to \mathcal{N}$, where $\mathcal{N}$ denotes the natural numbers $\{0, 1, \ldots\}$. Formally,

- f is $O(g)$ if

$$\exists c \in \mathcal{N} \quad \overset{\infty}{\forall} n \; f(n) \leq c \cdot g(n) \; .$$

  The notation $\overset{\infty}{\forall}$ means "for almost all" or "for all but finitely many". Intuitively, f grows no faster asymptotically than g to within a constant multiple.

- f is $o(g)$ if

$$\forall c \in \mathcal{N} \quad \overset{\infty}{\forall} n \; f(n) \leq \frac{1}{c} \cdot g(n) \; .$$

  This is a stronger statement. Intuitively, f grows strictly more slowly than any arbitrarily small positive constant multiple of g. For example, n^{347} is $o(2^{(\log n)^2})$.

- f is $\Omega(g)$ if g is $O(f)$. In other words, f is $\Omega(g)$ if

$$\exists c \in \mathcal{N} \quad \overset{\infty}{\forall} n \; f(n) \geq \frac{1}{c} \cdot g(n) \; .$$

- f is $\Theta(g)$ if f is both $O(g)$ and $\Omega(g)$.

There is one cardinal rule:

Always use O and o for upper bounds and Ω for lower bounds. *Never* use O for lower bounds.

There is some disagreement about the definition of Ω. Some authors (such as [43]) prefer the definition as given above. Others (such as [108]) prefer: f is $\Omega(g)$ if g is not $o(f)$; in other words, f is $\Omega(g)$ if

$$\exists c \in \mathcal{N} \quad \overset{\infty}{\exists}\, n \; f(n) > \frac{1}{c} \cdot g(n) \; .$$

(The notation $\overset{\infty}{\exists}$ means "there exist infinitely many".) The latter is weaker and presumably easier to establish, but the former gives sharper results. We won't get into the fray here, but just comment that neither definition precludes algorithms from taking less than the stated bound on certain inputs. For example, the assertion, "The running time of mergesort is $\Omega(n \log n)$" says that there is a c such that for all but finitely many n, there is some input sequence of length n on which mergesort makes at least $\frac{1}{c} n \log n$ comparisons. There is nothing to prevent mergesort from taking less time on some other input of length n.

The exact interpretation of statements involving O, o, and Ω depends on assumptions about the underlying model of computation, how the input is presented, how the size of the input is determined, and what constitutes a single step of the computation. In practice, authors often do not bother to write these down. For example, "The running time of mergesort is $O(n \log n)$" means that there is a fixed constant c such that for any n elements drawn from a totally ordered set, at most $cn \log n$ comparisons are needed to produce a sorted array. Here nothing is counted in the running time except the number of comparisons between individual elements, and each comparison is assumed to take one step; other operations are ignored. Similarly, nothing is counted in the input size except the number of elements; the size of each element (whatever that may mean) is ignored.

It is important to be aware of these unstated assumptions and understand how to make them explicit and formal when reading papers in the field. When making such statements yourself, always have your underlying assumptions in mind. Although many authors don't bother, it is a good habit to state any assumptions about the model of computation explicitly in any papers you write.

The question of what assumptions are reasonable is more often than not a matter of esthetics. You will become familiar with the standard models and assumptions from reading the literature; beyond that, you must depend on your own conscience.

1.2 Models of Computation

Our principal model of computation will be the unit-cost random access machine (RAM). Other models, such as uniform circuits and PRAMs, will be introduced when needed. The RAM model allows random access and the use

of arrays, as well as unit-cost arithmetic and bit-vector operations on arbitrarily large integers; see [3].

For graph algorithms, arithmetic is often unnecessary. Of the two main representations of graphs, namely *adjacency matrices* and *adjacency lists*, the former requires random access and $\Omega(n^2)$ array storage; the latter, only linear storage and no random access. (For graphs, *linear* means $O(n + m)$, where n is the number of vertices of the graph and m is the number of edges.) The most esthetically pure graph algorithms are those that use the adjacency list representation and only manipulate pointers. To express such algorithms one can formulate a very weak model of computation with primitive operators equivalent to **car**, **cdr**, **cons**, **eq**, and **nil** of pure LISP; see also [99].

1.3 A Grain of Salt

No mathematical model can reflect reality with perfect accuracy. Mathematical models are abstractions; as such, they are necessarily flawed.

For example, it is well known that it is possible to abuse the power of unit-cost RAMs by encoding horrendously complicated computations in large integers and solving intractible problems in polynomial time [50]. However, this violates the unwritten rules of good taste. One possible preventative measure is to use the log-cost model; but when used as intended, the unit-cost model reflects experimental observation more accurately for data of moderate size (since multiplication really does take one unit of time), besides making the mathematical analysis a lot simpler.

Some theoreticians consider asymptotically optimal results as a kind of Holy Grail, and pursue them with a relentless frenzy (present company not necessarily excluded). This often leads to contrived and arcane solutions that may be superior by the measure of asymptotic complexity, but whose constants are so large or whose implementation would be so cumbersome that no improvement in technology would ever make them feasible. What is the value of such results? Sometimes they give rise to new data structures or new techniques of analysis that are useful over a range of problems, but more often than not they are of strictly mathematical interest. Some practitioners take this activity as an indictment of asymptotic complexity itself and refuse to admit that asymptotics have anything at all to say of interest in practical software engineering.

Nowhere is the argument more vociferous than in the theory of parallel computation. There are those who argue that many of the models of computation in common use, such as uniform circuits and PRAMs, are so inaccurate as to render theoretical results useless. We will return to this controversy later on when we talk about parallel machine models.

Such extreme attitudes on either side are unfortunate and counterproductive. By now asymptotic complexity occupies an unshakable position in our computer science consciousness, and has probably done more to guide us in

improving technology in the design and analysis of algorithms than any other mathematical abstraction. On the other hand, one should be aware of its limitations and realize that an asymptotically optimal solution is not necessarily the best one.

A good rule of thumb in the design and analysis of algorithms, as in life, is to use common sense, exercise good taste, and always listen to your conscience.

1.4 Strassen's Matrix Multiplication Algorithm

Probably the single most important technique in the design of asymptotically fast algorithms is *divide-and-conquer*. Just to refresh our understanding of this technique and the use of recurrences in the analysis of algorithms, let's take a look at Strassen's classical algorithm for matrix multiplication and some of its progeny. Some of these examples will also illustrate the questionable lengths to which asymptotic analysis can sometimes be taken.

The usual method of matrix multiplication takes 8 multiplications and 4 additions to multiply two 2×2 matrices, or in general $O(n^3)$ arithmetic operations to multiply two $n \times n$ matrices. However, the number of multiplications can be reduced. Strassen [97] published one such algorithm for multiplying 2×2 matrices using only 7 multiplications and 18 additions:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} s_1 + s_2 - s_4 + s_6 & s_4 + s_5 \\ s_6 + s_7 & s_2 - s_3 + s_5 - s_7 \end{bmatrix}$$

where

$$\begin{aligned} s_1 &= (b-d)\cdot(g+h) \\ s_2 &= (a+d)\cdot(e+h) \\ s_3 &= (a-c)\cdot(e+f) \\ s_4 &= h\cdot(a+b) \\ s_5 &= a\cdot(f-h) \\ s_6 &= d\cdot(g-e) \\ s_7 &= e\cdot(c+d)\ . \end{aligned}$$

Assume for simplicity that n is a power of 2. (This is not the last time you will hear that.) Apply the 2×2 algorithm recursively on a pair of $n \times n$ matrices by breaking each of them up into four square submatrices of size $\frac{n}{2} \times \frac{n}{2}$:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} S_1 + S_2 - S_4 + S_6 & S_4 + S_5 \\ S_6 + S_7 & S_2 - S_3 + S_5 - S_7 \end{bmatrix}$$

where

$$S_1 = (B-D)\cdot(G+H)$$

$$\begin{aligned} S_2 &= (A+D)\cdot(E+H) \\ S_3 &= (A-C)\cdot(E+F) \\ S_4 &= H\cdot(A+B) \\ S_5 &= A\cdot(F-H) \\ S_6 &= D\cdot(G-E) \\ S_7 &= E\cdot(C+D)\ . \end{aligned}$$

Everything is the same as in the 2×2 case, except now we are manipulating $\frac{n}{2} \times \frac{n}{2}$ matrices instead of scalars. (We have to be slightly cautious, since matrix multiplication is not commutative.) Ultimately, how many scalar operations $(+, -, \cdot)$ does this recursive algorithm perform in multiplying two $n \times n$ matrices? We get the recurrence

$$T(n) = 7T(\frac{n}{2}) + dn^2$$

with solution

$$\begin{aligned} T(n) &= (1 + \frac{4}{3}d)n^{\log_2 7} + O(n^2) \\ &= O(n^{\log_2 7}) \\ &= O(n^{2.81\ldots}) \end{aligned}$$

which is $o(n^3)$. Here d is a fixed constant, and dn^2 represents the time for the matrix additions and subtractions.

This is already a significant asymptotic improvement over the naive algorithm, but can we do even better? In general, an algorithm that uses c multiplications to multiply two $d \times d$ matrices, used as the basis of such a recursive algorithm, will yield an $O(n^{\log_d c})$ algorithm. To beat Strassen's algorithm, we must have $c < d^{\log_2 7}$. For a 3×3 matrix, we need $c < 3^{\log_2 7} = 21.8\ldots$, but the best known algorithm uses 23 multiplications.

In 1978, Victor Pan [83, 84] showed how to multiply 70×70 matrices using 143640 multiplications. This gives an algorithm of approximately $O(n^{2.795\ldots})$. The asymptotically best algorithm known to date, which is achieved by entirely different methods, is $O(n^{2.376\ldots})$ [25]. Every algorithm must be $\Omega(n^2)$, since it has to look at all the entries of the matrices; no better lower bound is known.

Lecture 2 Topological Sort and MST

A recurring theme in asymptotic analysis is that it is often possible to get better asymptotic performance by maintaining extra information about the structure. Updating this extra information may slow down each individual step; this additional cost is sometimes called *overhead*. However, it is often the case that a small amount of overhead yields dramatic improvements in the asymptotic complexity of the algorithm.

To illustrate, let's look at *topological sort*. Let $G = (V, E)$ be a directed acyclic graph (dag). The edge set E of the dag G induces a *partial order* (a reflexive, antisymmetric, transitive binary relation) on V, which we denote by E^* and define by: uE^*v if there exists a directed E-path of length 0 or greater from u to v. The relation E^* is called the *reflexive transitive closure* of E.

Proposition 1 *Every partial order extends to a total order (a partial order in which every pair of elements is comparable).*

Proof. If R is a partial order that is not a total order, then there exist u, v such that neither uRv nor vRu. Extend R by setting

$$R \;:=\; R \cup \{(x, y) \mid xRu \text{ and } vRy\} \; .$$

The new R is a partial order extending the old R, and in addition now uRv. Repeat until there are no more incomparable pairs. □

In the case of a dag $G = (V, E)$ with associated partial order E^*, to say that a total order $\leq$ extends E^* is the same as saying that if uEv then $u \leq v$. Such a total order is called a *topological sort* of the dag G. A naive $O(n^3)$ algorithm to find a topological sort can be obtained from the proof of the above proposition.

Here is a faster algorithm, although still not optimal.

Algorithm 2 (Topological Sort II)

1. Start from any vertex and follow edges backwards until finding a vertex u with no incoming edges. Such a u must be encountered eventually, since there are no cycles and the dag is finite.
2. Make u the next vertex in the total order.
3. Delete u and all adjacent edges and go to step 1.

Using the adjacency list representation, the running time of this algorithm is $O(n)$ steps per iteration for n iterations, or $O(n^2)$.

The bottleneck here is step 1. A minor modification will allow us to perform this step in constant time. Assume the adjacency list representation of the graph associates with each vertex two separate lists, one for the incoming edges and one for the outgoing edges. If the representation is not already of this form, it can easily be put into this form in linear time. The algorithm will maintain a queue of vertices with no incoming edges. This will reduce the cost of finding a vertex with no incoming edges to constant time at a slight extra overhead for maintaining the queue.

Algorithm 3 (Topological Sort III)

1. Initialize the queue by traversing the graph and inserting each v whose list of incoming edges is empty.
2. Pick a vertex u off the queue and make u the next vertex in the total order.
3. Delete u and all outgoing edges (u, v). For each such v, if its list of incoming edges becomes empty, put v on the queue. Go to step 2.

Step 1 takes time $O(n)$. Step 2 takes constant time, thus $O(n)$ time over all iterations. Step 3 takes time $O(m)$ over all iterations, since each edge can be deleted at most once. The overall time is $O(m + n)$.

Later we will see a different approach involving depth first search.

2.1 Minimum Spanning Trees

Let $G = (V, E)$ be a connected undirected graph.

Definition 4 A *forest* in G is a subgraph $F = (V, E')$ with no cycles. Note that F has the same vertex set as G. A *spanning tree* in G is a forest with exactly one connected component. Given weights $w : E \to \mathcal{N}$ (edges are assigned weights over the natural numbers), a *minimum (weight) spanning tree (MST)* in G is a spanning tree T whose total weight (sum of the weights of the edges in T) is minimum over all spanning trees. □

Lemma 5 *Let $F = (V, E)$ be an undirected graph, c the number of connected components of F, $m = |E|$, and $n = |V|$. Then F has no cycles iff $c + m = n$.*

Proof.
($\rightarrow$) By induction on m. If $m = 0$, then there are n vertices and each forms a connected component, so $c = n$. If an edge is added without forming a cycle, then it must join two components. Thus m is increased by 1 and c is decreased by 1, so the equation $c + m = n$ is maintained.

($\leftarrow$) Suppose that F has at least one cycle. Pick an arbitrary cycle and remove an edge from that cycle. Then m decreases by 1, but c and n remain the same. Repeat until there are no more cycles. When done, the equation $c + m = n$ holds, by the preceding paragraph; but then it could not have held originally. □

We use a *greedy algorithm* to produce a minimum weight spanning tree. This algorithm is originally due to Kruskal [66].

Algorithm 6 (Greedy Algorithm for MST)

1. Sort the edges by weight.
2. For each edge on the list in order of increasing weight, include that edge in the spanning tree if it does not form a cycle with the edges already taken; otherwise discard it.

The algorithm can be halted as soon as $n - 1$ edges have been kept, since we know we have a spanning tree by Lemma 5.

Step 1 takes time $O(m \log m) = O(m \log n)$ using any one of a number of general sorting methods, but can be done faster in certain cases, for example if the weights are small integers so that bucket sort can be used.

Later on, we will give an almost linear time implementation of step 2, but for now we will settle for $O(n \log n)$. We will think of including an edge e in the spanning tree as taking the union of two disjoint sets of vertices, namely the vertices in the connected components of the two endpoints of e in the forest being built. We represent each connected component as a linked list. Each

list element points to the next element and has a back pointer to the head of the list. Initially there are no edges, so we have n lists, each containing one vertex. When a new edge (u, v) is encountered, we check whether it would form a cycle, *i.e.* whether u and v are in the same connected component, by comparing back pointers to see if u and v are on the same list. If not, we add (u, v) to the spanning tree and take the union of the two connected components by merging the two lists. Note that the lists are always disjoint, so we don't have to check for duplicates.

Checking whether u and v are in the same connected component takes constant time. Each merge of two lists could take as much as linear time, since we have to traverse one list and change the back pointers, and there are $n - 1$ merges; this will give $O(n^2)$ if we are not careful. However, if we maintain counters containing the size of each component and always merge the smaller into the larger, then each vertex can have its back pointer changed at most $\log n$ times, since each time the size of its component at least doubles. If we charge the change of a back pointer to the vertex itself, then there are at most $\log n$ changes per vertex, or at most $n \log n$ in all. Thus the total time for all list merges is $O(n \log n)$.

2.2 The Blue and Red Rules

Here is a more general approach encompassing most of the known algorithms for the MST problem. For details and references, see [100, Chapter 6], which proves the correctness of the greedy algorithm as a special case of this more general approach. In the next lecture, we will give an even more general treatment.

Let $G = (V, E)$ be an undirected connected graph with edge weights $w : E \to \mathcal{N}$. Consider the following two rules for coloring the edges of G, which Tarjan [100] calls the *blue rule* and the *red rule*:

> **Blue Rule:** Find a *cut* (a partition of V into two disjoint sets X and $V - X$) such that no blue edge crosses the cut. Pick an uncolored edge of minimum weight between X and $V - X$ and color it blue.
> **Red Rule:** Find a *cycle* (a path in G starting and ending at the same vertex) containing no red edge. Pick an uncolored edge of maximum weight on that cycle and color it red.

The greedy algorithm is just a repeated application of a special case of the blue rule. We will show next time:

Theorem 7 *Starting with all edges uncolored, if the blue and red rules are applied in arbitrary order until neither applies, then the final set of blue edges forms a minimum spanning tree.*

Lecture 3 Matroids and Independence

Before we prove the correctness of the blue and red rules for MST, let's first discuss an abstract combinatorial structure called a *matroid*. We will show that the MST problem is a special case of the more general problem of finding a minimum-weight maximal independent set in a matroid. We will then generalize the blue and red rules to arbitrary matroids and prove their correctness in this more general setting. We will show that every matroid has a dual matroid, and that the blue and red rules of a matroid are the red and blue rules, respectively, of its dual. Thus, once we establish the correctness of the blue rule, we get the red rule for free.

We will also show that a structure is a matroid if and only if the greedy algorithm always produces a minimum-weight maximal independent set for any weighting.

Definition 8 A *matroid* is a pair $(S, \mathcal{I})$ where S is a finite set and $\mathcal{I}$ is a family of subsets of S such that

(i) if $J \in \mathcal{I}$ and $I \subseteq J$, then $I \in \mathcal{I}$;

(ii) if $I, J \in \mathcal{I}$ and $|I| < |J|$, then there exists an $x \in J - I$ such that $I \cup \{x\} \in \mathcal{I}$.

The elements of $\mathcal{I}$ are called *independent sets* and the subsets of S not in $\mathcal{I}$ are called *dependent sets*. □

This definition is supposed to capture the notion of *independence* in a general way. Here are some examples:

1. Let V be a vector space, let S be a finite subset of V, and let $\mathcal{I} \subseteq 2^S$ be the family of linearly independent subsets of S. This example justifies the term "independent".

2. Let A be a matrix over a field, let S be the set of rows of A, and let $\mathcal{I} \subseteq 2^S$ be the family of linearly independent subsets of S.

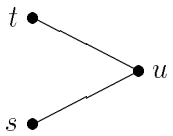
3. Let $G = (V, E)$ be a connected undirected graph. Let $S = E$ and let $\mathcal{I}$ be the set of forests in G. This example gives the MST problem of the previous lecture.

4. Let $G = (V, E)$ be a connected undirected graph. Let $S = E$ and let $\mathcal{I}$ be the set of subsets $E' \subseteq E$ such that the graph $(V, E - E')$ is connected.

5. Elements $\alpha_1, \ldots, \alpha_n$ of a field are said to be *algebraically independent* over a subfield k if there is no nontrivial polynomial $p(x_1, \ldots, x_n)$ with coefficients in k such that $p(\alpha_1, \ldots, \alpha_n) = 0$. Let S be a finite set of elements and let $\mathcal{I}$ be the set of subsets of S that are algebraically independent over k.

Definition 9 A *cycle* (or *circuit*) of a matroid $(S, \mathcal{I})$ is a setwise minimal (*i.e.*, minimal with respect to set inclusion) dependent set. A *cut* (or *cocircuit*) of $(S, \mathcal{I})$ is a setwise minimal subset of S intersecting all maximal independent sets. □

The terms *circuit* and *cocircuit* are standard in matroid theory, but we will continue to use *cycle* and *cut* to maintain the intuitive connection with the special case of MST. However, be advised that cuts in graphs as defined in the last lecture are *unions* of cuts as defined here. For example, in the graph



the set $\{(s, u),\ (t, u)\}$ forms a cut in the sense of MST, but not a cut in the sense of the matroid, because it is not minimal. However, a moment's thought reveals that this difference is inconsequential as far as the blue rule is concerned.

Let the elements of S be weighted. We wish to find a setwise maximal independent set whose total weight is minimum among all setwise maximal independent sets. In this more general setting, the blue and red rules become:

Blue Rule: Find a cut with no blue element. Pick an uncolored element of the cut of minimum weight and color it blue.
Red Rule: Find a cycle with no red element. Pick an element of the cycle of maximum weight and color it red.

3.1 Matroid Duality

As the astute reader has probably noticed by now, there is some kind of duality afoot. The similarity between the blue and red rules is just too striking to be mere coincidence.

Definition 10 Let $(S, \mathcal{I})$ be a matroid. The *dual matroid* of $(S, \mathcal{I})$ is $(S, \mathcal{I}^*)$, where

$$\mathcal{I}^* = \{\text{subsets of } S \text{ disjoint from some maximal element of } \mathcal{I}\} .$$

In other words, the maximal elements of $\mathcal{I}^*$ are the complements in S of the maximal elements of $\mathcal{I}$. □

The examples 3 and 4 above are duals. Note that $\mathcal{I}^{**} = \mathcal{I}$. Be careful: it is *not* the case that a set is independent in a matroid iff it is dependent in its dual. For example, except in trivial cases, $\emptyset$ is independent in both matroids.

Theorem 11

1. *Cuts in $(S, \mathcal{I})$ are cycles in $(S, \mathcal{I}^*)$.*
2. *The blue rule in $(S, \mathcal{I})$ is the red rule in $(S, \mathcal{I}^*)$ with the ordering of the weights reversed.*

3.2 Correctness of the Blue and Red Rules

Now we prove the correctness of the blue and red rules in arbitrary matroids. A proof for the special case of MST can be found in Tarjan's book [100, Chapter 6]; Lawler [70] states the blue and red rules for arbitrary matroids but omits a proof of correctness.

Definition 12 Let $(S, \mathcal{I})$ be a matroid with dual $(S, \mathcal{I}^*)$. An *acceptable coloring* is a pair of disjoint sets $B \in \mathcal{I}$ (the *blue elements*) and $R \in \mathcal{I}^*$ (the *red elements*). An acceptable coloring B, R is *total* if $B \cup R = S$, *i.e.* if B is a maximal independent set and R is a maximal independent set in the dual. An acceptable coloring B', R' *extends* or *is an extension of* an acceptable coloring B, R if $B \subseteq B'$ and $R \subseteq R'$. □

Lemma 13 *Any acceptable coloring has a total acceptable extension.*

Proof. Let B, R be an acceptable coloring. Let U^* be a maximal element of $\mathcal{I}^*$ extending R, and let $U = S - U^*$. Then U is a maximal element of $\mathcal{I}$ disjoint from R. As long as $|B| < |U|$, select elements of U and add them to B, maintaining independence. This is possible by axiom (ii) of matroids. Let $\hat{B}$ be the resulting set. Since all maximal independent sets have the same cardinality (Exercise 1a, Homework 1), $\hat{B}$ is a maximal element of $\mathcal{I}$ containing B and disjoint from R. The desired total extension is $\hat{B}, S - \hat{B}$. □

Lemma 14 *A cut and a cycle cannot intersect in exactly one element.*

Proof. Let C be a cut and D a cycle. Suppose that $C \cap D = \{x\}$. Then $D - \{x\}$ is independent and $C - \{x\}$ is independent in the dual. Color $D - \{x\}$ blue and $C - \{x\}$ red; by Lemma 13, this coloring extends to a total acceptable coloring. But depending on the color of x, either C is all red or D is all blue; this is impossible in an acceptable coloring, since D is dependent and C is dependent in the dual. □

Suppose B is independent and $B \cup \{x\}$ is dependent. Then $B \cup \{x\}$ contains a minimal dependent subset or cycle C, called the *fundamental cycle*[1] of x and B. The cycle C must contain x, because $C - \{x\}$ is contained in B and is therefore independent.

Lemma 15 (Exchange Lemma) *Let B, R be a total acceptable coloring.*

(i) Let $x \in R$ and let y lie on the fundamental cycle of x and B. If the colors of x and y are exchanged, the resulting coloring is acceptable.

(ii) Let $y \in B$ and let x lie on the fundamental cut of y and R (the fundamental cut of y and R is the fundamental cycle of y and R in the dual matroid). If the colors of x and y are exchanged, the resulting coloring is acceptable.

Proof. By duality, we need only prove (i). Let C be the fundamental cycle of x and B and let y lie on C. If $y = x$, there is nothing to prove. Otherwise $y \in B$. The set $C - \{y\}$ is independent since C is minimal. Extend $C - \{y\}$ by adding elements of $|B|$ as in the proof of Lemma 13 until achieving a maximal independent set B'. Then $B' = (B - \{y\}) \cup \{x\}$, and the total acceptable coloring $B', S - B'$ is obtained from B, R by switching the colors of x and y. □

A total acceptable coloring B, R is called *optimal* if B is of minimum weight among all maximal independent sets; equivalently, if R is of maximum weight among all maximal independent sets in the dual matroid.

Lemma 16 *If an acceptable coloring has an optimal total extension before execution of the blue or red rule, then so has the resulting coloring afterwards.*

Proof. We prove the case of the blue rule; the red rule follows by duality. Let B, R be an acceptable coloring with optimal total extension $\widehat{B}, \widehat{R}$. Let A be a cut containing no blue elements, and let x be an uncolored element of A of minimum weight. If $x \in \widehat{B}$, we are done, so assume that $x \in \widehat{R}$. Let C be the fundamental cycle of x and $\widehat{B}$. By Lemma 14, $A \cap C$ must contain another

[1] We say "the" because it is unique (Exercise 1b, Homework 1), although we do not need to know this for our argument.

element besides x, say y. Then $y \in \widehat{B}$, and $y \notin B$ because there are no blue elements of A. By Lemma 15, the colors of x and y in $\widehat{B}, \widehat{R}$ can be exchanged to obtain a total acceptable coloring $\widehat{B}', \widehat{R}'$ extending $B \cup \{x\}, R$. Moreover, $\widehat{B}'$ is of minimum weight, because the weight of x is no more than that of y. □

We also need to know

Lemma 17 *If an acceptable coloring is not total, then either the blue or red rule applies.*

Proof. Let B, R be an acceptable coloring with uncolored element x. By Lemma 13, B, R has a total extension $\widehat{B}, \widehat{R}$. By duality, assume without loss of generality that $x \in \widehat{B}$. Let C be the fundamental cut of x and $\widehat{R}$. Since all elements of C besides x are in $\widehat{R}$, none of them are blue in B. Thus the blue rule applies. □

Combining Lemmas 16 and 17, we have

Theorem 18 *If we start with an uncolored weighted matroid and apply the blue or red rules in any order until neither applies, then the resulting coloring is an optimal total acceptable coloring.*

What is really going on here is that all the subsets of the maximal independent sets of minimal weight form a submatroid of $(S, \mathcal{I})$, and the blue rule gives a method for implementing axiom (ii) for this matroid; see Miscellaneous Exercise 1.

3.3 Matroids and the Greedy Algorithm

We have shown that if $(S, \mathcal{I})$ is a matroid, then the greedy algorithm produces a maximal independent set of minimum weight. Here we show the converse: if $(S, \mathcal{I})$ is not a matroid, then the greedy algorithm fails for some choice of integer weights. Thus the abstract concept of matroid captures exactly when the greedy algorithm works.

Theorem 19 ([32]; see also [70]) *A system $(S, \mathcal{I})$ satisfying axiom (i) of matroids is a matroid (i.e., it satisfies (ii)) if and only if for all weight assignments $w : S \to \mathcal{N}$, the greedy algorithm gives a minimum-weight maximal independent set.*

Proof. The direction ($\rightarrow$) has already been shown. For ($\leftarrow$), let $(S, \mathcal{I})$ satisfy (i) but not (ii). There must be A, B such that $A, B \in \mathcal{I}$, $|A| < |B|$, but for no $x \in B - A$ is $A \cup \{x\} \in \mathcal{I}$.

Assume without loss of generality that B is a *maximal* independent set. If it is not, we can add elements to B maintaining the independence of B; for

any element that we add to B that can also be added to A while preserving the independence of A, we do so. This process never changes the fact that $|A| < |B|$ and for no $x \in B - A$ is $A \cup \{x\} \in \mathcal{I}$.

Now we assign weights $w : S \to \mathcal{N}$. Let $a = |A - B|$ and $b = |B - A|$. Then $a < b$. Let h be a huge number, $h \gg a, b$. (Actually $h > b^2$ will do.)

Case 1 If A is a maximal independent set, assign

$$\begin{aligned} w(x) &= a + 1 \quad \text{for } x \in B - A \\ w(x) &= b + 1 \quad \text{for } x \in A - B \\ w(x) &= 0 \qquad \text{for } x \in A \cap B \\ w(x) &= h \qquad \text{for } x \notin A \cup B \ . \end{aligned}$$

Thus

$$\begin{aligned} w(A) &= a(b+1) = ab + a \\ w(B) &= b(a+1) = ab + b \ . \end{aligned}$$

This weight assignment forces the greedy algorithm to choose B when in fact A is a maximal independent set of smaller weight.

Case 2 If A is not a maximal independent set, assign

$$\begin{aligned} w(x) &= 0 \quad \text{for } x \in A \\ w(x) &= b \quad \text{for } x \in B - A \\ w(x) &= h \quad \text{for } x \notin A \cup B \ . \end{aligned}$$

All the elements of A will be chosen first, and then a huge element outside of $A \cup B$ must be chosen, since A is not maximal. Thus the minimum-weight maximal independent set B was not chosen. □