Advanced Algorithmics (6EAP)
Graphs I

Jaak Vilo
2012 Spring
CLRS: Chapter 22
Elementary Graph Algorithms

Some slides from:
http://www.cc.nctu.edu.tw/~claven/course/Algorithm/
Graphs

Graphs are one of the unifying themes of computer science. A graph $G = (V, E)$ is defined by a set of vertices $V$, and a set of edges $E$ consisting of ordered or unordered pairs of vertices from $V$. 
Introduction

• \( G=(V, E) \)
  – \( V = \) vertex set (nodes)
  – \( E = \) edge set (arcs)

• Graph representation
  – Adjacency list
  – Adjacency matrix

• Graph search
  – Breadth-first search (BFS)
  – Depth-first search (DFS)
    • Topological sort
    • Strongly connected components

http://www.cc.nctu.edu.tw/~claven/course/Algorithm/
Road Networks

In modeling a road network, the vertices may represent the cities or junctions, certain pairs of which are connected by roads/edges.
Electronic Circuits

In an electronic circuit, with junctions as vertices as components as edges.
Find a shortest path from station A to station B.

-need serious thinking to get a correct algorithm.
A Simple Metabolic Pathway

Metabolic Regulation - Methionine Biosynthesis in E. coli
Evolutionary relationship among organisms based on similarity of the primary sequences of their CYTOCHROME c proteins
Common ancestral community of primitive cells

- Bacteria
- Archaea
- Eukaryotes
- Plastids
- Mitochondria
Green arrows - upregulation  
Red arrows - downregulation  
Thickness of arrow represents certainty of direction (up/down)
A complete graph
Filter

- choose a list of genes (MATING, marked in red)
- filter for these genes plus neighbouring genes from the graph

Mutation network $\Delta_{\gamma}=4$
Mutation network $\Delta_{\gamma}=2$
Probability network $\Pi(\gamma=2.0, \tau=0.8, \xi=10)$, underlayed in green are groups of genes which are more interconnected. The genes are coloured according to annotation in YPD (“cellular role”). The genes which are more interconnected are involved in the same cellular processes, like mating behaviour (mat, green), aminoacid metabolism (aam, red), cos gene family (cos, light blue), mitochondrial function (mitochondrial, dark blue), ribosome (ribo, purple) and a group of genes of unknown function (unknown, grey).
Graphs

• Set of nodes $|V| = n$
• Set of edges $|E| = m$
  – Undirected edges/graph: pairs of nodes $\{v, w\}$
  – Directed edges/graph: pairs of nodes $(v, w)$
• Set of neighbors of $v$: set of nodes connected by an edge with $v$ (directed: in-neighbors, out-neighbors)
• **Degree of a node**: number of its neighbors (indegree, outdegree)
• Path: a sequence of nodes such that every two consecutive nodes constitute an edge
• Length of a path: number of nodes minus 1
• Distance between two nodes: the length of the shortest path between these nodes
• Diameter: the longest distance in the graph
Lines, cycles, trees, cliques

Line

Cycle

Clique

Tree

Dariusz Kowalski
Choose

• The boss wants to produce programs to solve the following two problems
  – **Euler circuit problem:**
    • given a graph $G$, find a way to go through each edge exactly once.
  – **Hamilton circuit problem:**
    • given a graph $G$, find a way to go through each vertex exactly once.

• The two problems seem to be very similar.
• Person A takes the first problem and person B takes the second.
• **Outcome:** Person A quickly completes the program, whereas person B works 24 hours per day and is **fired** after a few months.
Euler Circuit: The original Konigsberg bridge

Konigsberg graph
Euler circuit

- Every vertex of this graph has an even degree, therefore there exists an Eulerian graph. Following the edges in alphabetical order gives an Eulerian circuit/cycle.
Hamilton Circuit

Traveling salesman problem (TSP),
A joke (continued):

• **Why?** no body in the company has taken Algorithmics class.

• **Explanation:**
  
  – Euler circuit problem can be easily solved in polynomial time.
  
  – Hamilton circuit problem is proved to be **NP-hard**.
  
  – So far, no body in the world can give a polynomial time algorithm for a NP-hard problem.
  
  – Conjecture: there does not exist polynomial time algorithm for this problem.
"I can't find an efficient algorithm, I guess I'm just too dumb."
"I can't find an efficient algorithm, because no such algorithm is possible!"
"I can't find an efficient algorithm, but neither can all these famous people."
Flavors of Graphs

The first step in any graph problem is determining which flavor of graph you are dealing with. Learning to talk the talk is an important part of walking the walk. The flavor of graph has a big impact on which algorithms are appropriate and efficient.
Directed vs. Undirected Graphs

A graph $G = (V, E)$ is undirected if edge $(x, y) \in E$ implies that $(y, x)$ is also in $E$.

Road networks between cities are typically undirected. Street networks within cities are almost always directed because of one-way streets. Most graphs of graph-theoretic interest are undirected.
Weighted vs. Unweighted Graphs

In *weighted* graphs, each edge (or vertex) of $G$ is assigned a numerical value, or weight.

The edges of a road network graph might be weighted with their length, drive-time or speed limit. In *unweighted* graphs, there is no cost distinction between various edges and vertices.
Simple vs. Non-simple Graphs

Certain types of edges complicate the task of working with graphs. A *self-loop* is an edge \((x, x)\) involving only one vertex.

An edge \((x, y)\) is a *multi-edge* if it occurs more than once in the graph.

Any graph which avoids these structures is called *simple*. 
Sparse vs. Dense Graphs

Graphs are *sparse* when only a small fraction of the possible number of vertex pairs actually have edges defined between them.

![Sparse vs. Dense Graphs](image)

Graphs are usually sparse due to application-specific constraints. Road networks must be sparse because of road junctions. Typically dense graphs have a quadratic number of edges while sparse graphs are linear in size.
Cyclic vs. Acyclic Graphs

An *acyclic* graph does not contain any cycles. *Trees* are connected acyclic *undirected* graphs.

Directed acyclic graphs are called *DAGs*. They arise naturally in scheduling problems, where a directed edge \((x, y)\) indicates that \(x\) must occur before \(y\).
Implicit vs. Explicit Graphs

Many graphs are not explicitly constructed and then traversed, but built as we use them.

A good example arises in backtrack search.
Embedded vs. Topological Graphs

A graph is *embedded* if the vertices and edges have been assigned geometric positions.

Example: TSP or Shortest path on points in the plane.
Example: Grid graphs.
Example: Planar graphs.
Labeled vs. Unlabeled Graphs

In *labeled* graphs, each vertex is assigned a unique name or identifier to distinguish it from all other vertices.

An important graph problem is *isomorphism testing*, determining whether the topological structure of two graphs are in fact identical if we ignore any labels.
The Friendship Graph

Consider a graph where the vertices are people, and there is an edge between two people if and only if they are friends.

This graph is well-defined on any set of people: SUNY SB, New York, or the world. What questions might we ask about the friendship graph?
If I am your friend, does that mean you are my friend?

A graph is *undirected* if \((x, y)\) implies \((y, x)\). Otherwise the graph is directed. The “heard-of” graph is directed since countless famous people have never heard of me! The “had-sex-with” graph is presumably undirected, since it requires a partner.
Perception of/and experience

- Simple example of 5 entities (persons) and their relationships
- Who would you prefer to be?
- Who wouldn’t you want to be?
- And what if the relationship means “company A sells to company B”?
- What if relationship means “love”?
Am I my own friend?

An edge of the form \((x, x)\) is said to be a *loop*. If \(x\) is \(y\)'s friend several times over, that could be modeled using *multiedges*, multiple edges between the same pair of vertices.

A graph is said to be *simple* if it contains no loops and multiple edges.
Am I linked by some chain of friends to the President?

A path is a sequence of edges connecting two vertices. Since Mel Brooks is my father’s-sister’s-husband’s cousin, there is a path between me and him!

Steve  Dad  Aunt Eve  Uncle Lenny  Cousin Mel
How close is my link to the President?

If I were trying to impress you with how tight I am with Mel Brooks, I would be much better off saying that Uncle Lenny knows him than to go into the details of how connected I am to Uncle Lenny. Thus we are often interested in the shortest path between two nodes.
Is there a path of friends between any two people?

A graph is *connected* if there is a path between any two vertices.
A directed graph is *strongly connected* if there is a directed path between any two vertices.
Who has the most friends?

The *degree* of a vertex is the number of edges adjacent to it.
Graph terminology

This graph has 55 vertices, 70 edges, and 3 connected components. One of the connected components is a tree (right). The graph has many cycles, one of which is highlighted in the large connected component (left). The diagram also depicts a spanning tree in the small connected component (center). The graph as a whole does not have a spanning tree, because it is not connected.
Line graph

Given a graph $G$, its line graph $L(G)$ is a graph such that

• each vertex of $L(G)$ represents an edge of $G$; and
• two vertices of $L(G)$ are adjacent if and only if their corresponding edges share a common endpoint ("are adjacent") in $G$. 
Complete graph

- Every node is connected to every other node
- Clique – fully connected subgraph of a graph
Subgraph

• Subset of vertices (m) \( V' \) is a subset of \( V \)
• Subset of edges (n) \( E' \) is a subset of \( E \), s.t. \( \{u,v\} \in E' \) if \( u,v \) both in \( V \)

• Nr of different possible graphs of size \( m,n \) is huge
How many different subgraphs does a complete graph have?

• How many?
  – $2^m$ different subsets of vertices (= many!)
  – Likewise, nr of edges is any subset of set of edges...

• 5 nodes => 5*4/2 different possible undirected edges without self-loops
  • Calculate the possibility of each edge being present or not...
  • Directed: ->, <-, <->, none (4 options)
Figure 22.1  Two representations of an undirected graph. (a) An undirected graph $G$ having five vertices and seven edges. (b) An adjacency-list representation of $G$. (c) The adjacency-matrix representation of $G$. 
Representation of Graphs

• Adjacency list: $\Theta(V+E)$
  – Preferred for **sparse** graph
    • $|E| \ll |V|^2$
    – Adj[u] contains all the vertices v such that there is an edge $(u, v) \in E$
    – Weighted graph: $w(u, v)$ is stored with vertex v in Adj[u]
    – No quick way to determine if a given edge is present in the graph

• Adjacency matrix: $\Theta(V^2)$
  – Preferred for **dense** graph
    – Symmetry for undirected graph
    – Weighted graph: store $w(u, v)$ in the $(u, v)$ entry
    – Easy to determine if a given edge is present in the graph
Representation For A Directed Graph

(a)

(b)

(c)

1 bit per entry

Figure 22.2 Two representations of a directed graph. (a) A directed graph $G$ having six vertices and eight edges. (b) An adjacency-list representation of $G$. (c) The adjacency-matrix representation of $G$. 
Tradeoffs Between Adjacency Lists and Adjacency Matrices

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Winner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster to test if ((x, y)) exists?</td>
<td>matrices</td>
</tr>
<tr>
<td>Faster to find vertex degree?</td>
<td>lists</td>
</tr>
<tr>
<td>Less memory on small graphs?</td>
<td>lists ((m + n)) vs. ((n^2))</td>
</tr>
<tr>
<td>Less memory on big graphs?</td>
<td>matrices (small win)</td>
</tr>
<tr>
<td>Edge insertion or deletion?</td>
<td>matrices (O(1))</td>
</tr>
<tr>
<td>Faster to traverse the graph?</td>
<td>lists (m + n) vs. (n^2)</td>
</tr>
<tr>
<td>Better for most problems?</td>
<td>lists</td>
</tr>
</tbody>
</table>

Both representations are very useful and have different properties, although adjacency lists are probably better for most problems.
Traversing a Graph

One of the most fundamental graph problems is to traverse every edge and vertex in a graph. For *efficiency*, we must make sure we visit each edge at most twice. For *correctness*, we must do the traversal in a systematic way so that we don’t miss anything. Since a maze is just a graph, such an algorithm must be powerful enough to enable us to get out of an arbitrary maze.
Marking Vertices

The key idea is that we must mark each vertex when we first visit it, and keep track of what have not yet completely explored.

Each vertex will always be in one of the following three states:

- *undiscovered* – the vertex in its initial, virgin state.
- *discovered* – the vertex after we have encountered it, but before we have checked out all its incident edges.
- *processed* – the vertex after we have visited all its incident edges.
**Breadth-First Search (BFS)**

- Graph search: given a source vertex $s$, explores the edges of $G$ to discover every vertex that is reachable from $s$
  - Compute the distance (smallest number of edges) from $s$ to each reachable vertex
  - Produce a breadth-first tree with root $s$ that contains all reachable vertices
  - Compute the shortest path from $s$ to each reachable vertex
- BFS discovers all vertices at distance $k$ from $s$ before discovering any vertices at distance $k+1$
Simple BFS from n

enqueue( Q, n )

while Q not empty
    u = dequeue (Q)
    process u
    for each v in Adjacency( u )  // discover neighbours
        if v not yet discovered
            then enqueue( Q, v )
Data Structure for BFS

- Adjacency list
- color[u] for each vertex
  - WHITE if u has not been discovered
  - BLACK if u and all its adjacent vertices have been discovered
  - GRAY if u has been discovered, but has some adjacent white vertices
  - Frontier between discovered and undiscovered vertices
- d[u] for the distance from (source) s to u
- π[u] for predecessor of u
- FIFO queue Q to manage the set of gray vertices
  - Q stores all the gray vertices
BFS($G, s$)

1. for each vertex $u \in V[G] - \{s\}$
2. do $\text{color}[u] \leftarrow \text{WHITE}$
3. $d[u] \leftarrow \infty$
4. $\pi[u] \leftarrow \text{NIL}$
5. $\text{color}[s] \leftarrow \text{GRAY}$
6. $d[s] \leftarrow 0$
7. $\pi[s] \leftarrow \text{NIL}$
8. $Q \leftarrow \emptyset$
9. $\text{ENQUEUE}(Q, s)$
10. while $Q \neq \emptyset$

11. do $u \leftarrow \text{DEQUEUE}(Q)$
12. for each $v \in \text{Adj}[u]$
13. do if $\text{color}[v] = \text{WHITE}$
14. then $\text{color}[v] \leftarrow \text{GRAY}$
15. $d[v] \leftarrow d[u] + 1$
16. $\pi[v] \leftarrow u$
17. $\text{ENQUEUE}(Q, v)$
18. $\text{color}[u] \leftarrow \text{BLACK}$
Example (BFS)
(Courtesy of Prof. Jim Anderson)

Q: s 0
Example (BFS)

Q: w r
   1 1
Example (BFS)
Example (BFS)

Q: \[ t \times v \\
2 \quad 2 \quad 2 \]
Example (BFS)

Q: x v u
   2 2 3
Example (BFS)

Q: v  u  y
   2  3  3
Example (BFS)
Example (BFS)
Example (BFS)

Q: $\emptyset$
Example (BFS) – BF Tree

BF Tree
Analysis of BFS

- $O(|V| + |E|) = O(n+m)$
  - Each vertex is en-queued ($O(1)$) at most once $\Rightarrow O(n)$
    - No vertex is re-painted white
      - $\Rightarrow$ vertex is inserted into queue and retrieved from there only once
  - Each adjacency list is scanned at most once $\Rightarrow O(m)$
Shortest path

- Print out the vertices on a shortest path from s to v

```plaintext
PRINT-PATH(G, s, v)
1   if v = s
2      then print s
3   else if π[v] = NIL
4      then print "no path from" s "to" v "exists"
5   else PRINT-PATH(G, s, π[v])
6      print v
```
PRINT-PATH Illustration

PRINT-PATH(G, s, u)  
PRINT-PATH(G, s, t)  
PRINT-PATH(G, s, w)  
PRINT-PATH(G, s, s)  
    print s  
    print w  
    print t  
    print u  

Output: s w t u
Connected Components

The *connected components* of an undirected graph are the separate “pieces” of the graph such that there is no connection between the pieces. Many seemingly complicated problems reduce to finding or counting connected components. For example, testing whether a puzzle such as Rubik’s cube or the 15-puzzle can be solved from any position is really asking whether the graph of legal configurations is connected. Anything we discover during a BFS must be part of the same connected component. We then repeat the search from any undiscovered vertex (if one exists) to define the next component, until all vertices have been found:
Two-Coloring Graphs

The *vertex coloring* problem seeks to assign a label (or color) to each vertex of a graph such that no edge links any two vertices of the same color.

A graph is *bipartite* if it can be colored without conflicts while using only two colors. Bipartite graphs are important because they arise naturally in many applications. For example, consider the “had-sex-with” graph in a heterosexual world. Men have sex only with women, and vice versa. Thus gender defines a legal two-coloring.
Bipartite graphs

- people and groups
- men-women
- Stable marriage
  - find matching that will not be “broken” by inevitable divorces
- Apples and Oranges
Finding a Two-Coloring

We can augment breadth-first search so that whenever we discover a new vertex, we color it the opposite of its parent.

twocolor(graph *g)
{
    int i;

    for (i=1; i<=(g->nvertices); i++)
        color[i] = UNCOLORED;

    bipartite = TRUE;

    initialize_search(&g);

    for (i=1; i<=(g->nvertices); i++)
        if (discovered[i] == FALSE) {
            color[i] = WHITE;
            bfs(g,i);
        }
}
process_edge(int x, int y)
{
    if (color[x] == color[y]) {
        bipartite = FALSE;
        printf("Warning: graph not bipartite, due to (%d,%d)\n",x,y);
    }

    color[y] = complement(color[x]);
}

complement(int color)
{
    if (color == WHITE) return(BLACK);
    if (color == BLACK) return(WHITE);

    return(UNCOLORED);
}

We can assign the first vertex in any connected component to be whatever color/sex we wish.
Problem of the Day

Prove that in a breadth-first search on a undirected graph $G$, every edge in $G$ is either a tree edge or a cross edge, where a cross edge $(x, y)$ is an edge where $x$ is neither is an ancestor or descendent of $y$. 
**Depth-First Search (DFS)**

- DFS: search deeper in the graph whenever possible
  - Edges are explored out of the most recently discovered vertex $v$ that still has unexplored edges leaving it
  - When all of $v$’s edges have been explored *(finished)*, the search **backtracks** to explore edges leaving the vertex from which $v$ was discovered
  - This process continues until we have discovered all the vertices that are **reachable** from the original source vertex
  - If any undiscovered vertices remain, then one of them is selected as a new source and the search is repeated from that source
  - The entire process is repeated until all vertices are discovered

- DFS will create a **forest of DFS-trees**
Simple DFS from n

push ( Q, n )

while  u = pop (Q)

    process u

    for each  v  in reverse  Adjacency( u )
        push ( Q, v )
Recursive:

DFS( u ) :

  process u
  for each v in Adjacency( u )
    DFS( v ) if v undiscovered
Data Structure for DFS

• Adjacency list
• color[u] for each vertex
  – WHITE if u has not been discovered
  – GRAY if u is discovered but not finished
  – BLACK if u is finished
• Timestamps: $1 \leq d[u] < f[u] \leq 2|V|$
  – d[u] records when u is first discovered (and grayed)
  – f[u] records when the search finishes examining u’s adjacency list (and blacken u)
• $\pi[u]$ for predecessor of u
The Key Idea with DFS

A depth-first search of a graph organizes the edges of the graph in a precise way. In a DFS of an undirected graph, we assign a direction to each edge, from the vertex which discover it:
DFS: initialise and visit all yet unexplored vertices

\[
\text{DFS}(G) \\
1 \text{ for each vertex } u \in V[G] \\
2 \quad \text{do } \text{color}[u] \leftarrow \text{WHITE} \\
3 \quad \pi[u] \leftarrow \text{NIL} \\
4 \quad \text{time} \leftarrow 0 \\
5 \text{for each vertex } u \in V[G] \\
6 \quad \text{do if } \text{color}[u] = \text{WHITE} \\
7 \quad \text{then } \text{DFS-VISIT}(u)
\]
DFS-visit — visit all reachable nodes

DFS-VISIT($u$)
1. $\text{color}[u] \leftarrow \text{GRAY}$  \hspace{1em} ▷ White vertex $u$ has just been discovered.
2. $\text{time} \leftarrow \text{time} + 1$
3. $d[u] \leftarrow \text{time}$
4. for each $v \in \text{Adj}[u]$  \hspace{1em} ▷ Explore edge $(u, v)$.
5. \hspace{1em} do if $\text{color}[v] = \text{WHITE}$
6. \hspace{1em} then $\pi[v] \leftarrow u$
7. \hspace{1em} DFS-VISIT($v$)
8. $\text{color}[u] \leftarrow \text{BLACK}$  \hspace{1em} ▷ Blacken $u$; it is finished.
9. $f[u] \leftarrow \text{time} \leftarrow \text{time} + 1$
Example (DFS)
(Courtesy of Prof. Jim Anderson)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Example (DFS)
Properties of DFS

- Time complexity: $\Theta(V+E)$
  - Loops on lines 1-3 and 5-7 of DFS: $\Theta(V)$
  - DFS-VISIT
    - Called exactly once for each vertex
    - Loops on lines 4-7 for a vertex $v$: $|\text{Adj}[v]|$
    - Total time $= \sum_{v \in V} |\text{Adj}[v]| = \Theta(E)$

- DFS results in a *forest of trees*
- Discovery and finishing times have *parenthesis structure*
Depth-First Search

DFS has a neat recursive implementation which eliminates the need to explicitly use a stack. Discovery and final times are a convenience to maintain.

dfs(graph *g, int v)
{
edgenode *p; (* temporary pointer *)
    int y; (* successor vertex *)

    if (finished) return; (* allow for search termination *)

    discovered[v] = TRUE;
    time = time + 1;
    entry_time[v] = time;

    process_vertex_early(v);

    p = g—>edges[v];
    while (p ! = NULL) {
        y = p—>y;
}

}
if (discovered[y] == FALSE) {
    parent[y] = v;
    process_edge(v,y);
    dfs(g,y);
}
else if (!(processed[y]) || (g->directed))
    process_edge(v,y);

if (finished) return;

p = p->next;
}

process_vertex_late(v);

time = time + 1;
exit_time[v] = time;

processed[v] = TRUE;
Another Example of DFS
Figure 18.13
Depth-first search

This figure illustrates the progress of DFS in a random Euclidean near-neighbor graph (left). The figures show the DFS tree vertices and edges in the graph as the search progresses through 1/4, 1/2, 3/4, and all of the vertices (top to bottom). The DFS tree (tree edges only) is shown at the right. As is evident from this example, the search tree for DFS tends to be quite tall and thin for this type of graph (as it is for many other types of graphs commonly encountered in practice). We normally find a vertex nearby that we have not seen before.
**BFS**

- **Queue**

---

As shallow and broad, and demonstrates a set of facts about the graph being searched different from those shown by DFS. For example,

- There exists a relatively short path connecting each pair of vertices in the graph.
- During the search, most vertices are adjacent to numerous unvisited vertices.

Again, this example is typical of the behavior that we expect from BFS, but verifying facts of this kind for graph models of interest and graphs that arise in practice requires detailed analysis.

DFS wends its way through the graph, storing on the stack the points where other paths branch off; BFS sweeps through the graph, using a queue to remember the frontier of visited places. DFS explores the graph by looking for new vertices far away from the start point, taking closer vertices only when dead ends are encountered; BFS completely covers the area close to the starting point, moving farther away only when everything nearby has been examined. The order in which

---

**Figure 18.24**

**Breadth-first search**

This figure illustrates the progress of BFS in random Euclidean near-neighbor graph (left), in the same style as Figure 18.13. As is evident from this example, the search tree for BFS tends to be quite short and wide for this type of graph (and many other types of graphs commonly encountered in practice). That is, vertices tend to be connected to one another by rather short paths. The contrast between the shapes of the DFS and BFS trees is striking testimony to the differing dynamic properties of the algorithms.
Randomised search

Use:
Randomized Queue
Parenthesis Theorem

**Theorem 22.7**

For all $u$, $v$, exactly one of the following holds:

2. $d[u] < d[v] < f[v] < f[u]$ and $v$ is a descendant of $u$.

- Like parentheses:
  - OK: ( ) [ ] ( [ ] ) [ ( ) ]
  - Not OK: ( [ ) ] [ ( ) ]

**Corollary**

$v$ is a proper descendant of $u$ if and only if $d[u] < d[v] < f[v] < f[u]$.
Parenthesis theorem

In any depth-first search of a (directed or undirected) graph $G = (V, E)$, for any two vertices $u$ and $v$, exactly one of the following three conditions holds:

1. the intervals $[d[u], f[u]]$ and $[d[v], f[v]]$ are entirely disjoint, and neither $u$ nor $v$ is a descendant of the other in the depth-first forest,

2. the interval $[d[u], f[u]]$ is contained entirely within the interval $[d[v], f[v]]$, and $u$ is a descendant of $v$ in a depth-first tree, or

3. the interval $[d[v], f[v]]$ is contained entirely within the interval $[d[u], f[u]]$, and $v$ is a descendant of $u$ in a depth-first tree.
Example (Parenthesis Theorem)

\[(s \ (z \ (y \ (x \ x) \ y) \ (w \ w) \ z) \ s) \ (t \ (v \ v) \ (u \ u) \ t)\]
Proof

We begin with the case in which $d[u] < d[v]$.

- There are two subcases to consider, according to whether $d[v] < f[u]$ or not.
- The first subcase occurs when $d[v] < f[u]$, so $v$ was discovered while $u$ was still gray. This implies that $v$ is a descendant of $u$. Moreover, since $v$ was discovered more recently than $u$, all of its outgoing edges are explored, and $v$ is finished, before the search returns to and finishes $u$. In this case, therefore, the interval $[d[v], f[v]]$ is entirely contained within the interval $[d[u], f[u]]$.
- In the other subcase, $f[u] < d[v]$, and inequality (22.2) implies that the intervals $[d[u], f[u]]$ and $[d[v], f[v]]$ are disjoint.
- Because the intervals are disjoint, neither vertex was discovered while the other was gray, and so neither vertex is a descendant of the other.

- The case in which $d[v] < d[u]$ is similar, with the roles of $u$ and $v$ reversed in the above argument.
Depth-First Trees

• Predecessor subgraph defined slightly different from that of BFS.

• The predecessor subgraph of DFS is $G_\pi = (V, E_\pi)$ where $E_\pi = \{(\pi[v], v) : v \in V \text{ and } \pi[v] \neq \text{NIL}\}$.
  - How does it differ from that of BFS?
  - The predecessor subgraph $G_\pi$ forms a depth-first forest composed of several depth-first trees. The edges in $E_\pi$ are called tree edges.

Definition:
Forest: An acyclic graph $G$ that may be disconnected.
White-path Theorem

**Theorem 22.9**

$v$ is a descendant of $u$ if and only if at time $d[u]$, there is a path $u \leadsto v$ consisting of only white vertices. (Except for $u$, which was just colored gray.)
Classification of Edges

- **Tree edge:** in the depth-first forest. Found by exploring \((u, v)\). -- \(v\) was white
- **Back edge:** \((u, v)\), where \(u\) is a descendant of \(v\) (in the depth-first tree). -- \(v\) was gray
- **Forward edge:** \((u, v)\), where \(v\) is a descendant of \(u\), but not a tree edge. -- \(v\) was black and \(d[u] < d[v]\)
- **Cross edge:** any other edge. Can go between vertices in same depth-first tree or in different depth-first trees. -- \(v\) was black and \(d[u] > d[v]\)

**Theorem:**
In DFS of an undirected graph, we get only tree and back edges. No forward or cross edges.
Types of edges in DFS

- Tree edge
- Back edge
- Forward edge
- Cross edge
Edge Classification for DFS

Every edge is either:

1. A Tree Edge
2. A Back Edge to an ancestor
3. A Forward Edge to a descendant
4. A Cross Edge to a different node

On any particular DFS or BFS of a directed or undirected graph, each edge gets classified as one of the above.
(for undirected graphs)

**DFS: Tree Edges and Back Edges Only**

The reason DFS is so important is that it defines a very nice ordering to the edges of the graph. In a DFS of an undirected graph, every edge is either a tree edge or a back edge.

Why? Suppose we have a forward edge. We would have encountered \( (4, 1) \) when expanding 4, so this is a back edge.
No Cross Edges in DFS

Suppose we have a cross-edge

When expanding 2, we would discover 5, so the tree would look like:
DFS Application: Finding Cycles

Back edges are the key to finding a cycle in an undirected graph.
Any back edge going from $x$ to an ancestor $y$ creates a cycle with the path in the tree from $y$ to $x$.

```c
process_edge(int x, int y)
{
    if (parent[x] != y) { /* found back edge! */
        printf("Cycle from \%d to \%d: ", y, x);
        find_path(y, x, parent);
        finished = TRUE;
    }
}
```
DFS-visit — visit all reachable nodes

DFS-VISIT(u)
1  color[u] ← GRAY \quad \triangleright \text{White vertex } u \text{ has just been discovered.}
2  time ← time + 1
3  d[u] ← time
4  for each v ∈ Adj[u] \quad \triangleright \text{Explore edge } (u, v).
5      \quad \text{do if } color[v] = \text{WHITE}
6      \quad \quad then \pi[v] ← u
7      \quad \quad DFS-VISIT(v)
8  color[u] ← BLACK \quad \triangleright \text{Blacken } u; \text{ it is finished.}
9  f[u] ← time ← time + 1
Types of edges in DFS

- Tree edge
- Back edge
- Forward edge
- Cross edge
Lemma – DAG acyclicity

• **DAG is acyclic if and only if DFS of G yields no back edges**
  
  ➔ Suppose that there is a back edge \((u, v)\). Then vertex \(v\) is an ancestor of vertex \(u\) in the depth-first forest. There is thus a path from \(v\) to \(u\) in \(G\), and the back edge \((u, v)\) completes a cycle

  ✈ Suppose that \(G\) contains a cycle \(c\). We show that a DFS of \(G\) yields a back edge. Let \(v\) be the first vertex to be discovered in \(c\), and let \((u, v)\) be the preceding edge in \(c\). At time \(d[v]\), the vertices of \(c\) form a path of white vertices from \(v\) to \(u\). By the white-path theorem (Theorem 22.9), vertex \(u\) becomes a descendant of \(v\) in the depth-first forest. Therefore, \((u, v)\) is a back edge.
Articulation Vertices

Suppose you are a terrorist, seeking to disrupt the telephone network. Which station do you blow up?

An *articulation vertex* is a vertex of a connected graph whose deletion disconnects the graph. Clearly connectivity is an important concern in the design of any network. Articulation vertices can be found in $O(n(m+n))$ – just delete each vertex to do a DFS on the remaining graph to see if it is connected.
A Faster $O(n + m)$ DFS Algorithm

In a DFS tree, a vertex $v$ (other than the root) is an articulation vertex iff $v$ is not a leaf and some subtree of $v$ has no back edge incident until a proper ancestor of $v$.

The root is a special case since it has no ancestors.

$X$ is an articulation vertex since the right subtree does not have a back edge to a proper ancestor.

Leaves cannot be articulation vertices.
Problems 22.2: Articulation points, bridges, and biconnected components

Let $G = (V, E)$ be a connected, undirected graph. An *articulation point* of $G$ is a vertex whose removal disconnects $G$. A *bridge* of $G$ is an edge whose removal disconnects $G$. A *biconnected component* of $G$ is a maximal set of edges such that any two edges in the set lie on a common simple cycle. Figure 22.10 illustrates these definitions. We can determine articulation points, bridges, and biconnected components using depth-first search. Let $G_p = (V, E_p)$ be a depth-first tree of $G$.

**Figure 22.10:** The articulation points, bridges, and biconnected components of a connected, undirected graph for use in Problem 22.2. The articulation points are the heavily shaded vertices, the bridges are the heavily shaded edges, and the biconnected components are the edges in the shaded regions, with a $bcc$ numbering shown.

a. Prove that the root of $G_p$ is an articulation point of $G$ if and only if it has at least two children in $G_p$.

b. Let $v$ be a nonroot vertex of $G_p$. Prove that $v$ is an articulation point of $G$ if and only if $v$ has a child $s$ such that there is no back edge from $s$ or any descendant of $s$ to a proper ancestor of $v$.

...
TOPOLOGICAL SORTING
Topological Sorting

A directed, acyclic graph has no directed cycles.

A topological sort of a graph is an ordering on the vertices so that all edges go from left to right. DAGs (and only DAGs) has at least one topological sort (here $G, A, B, C, F, E, D$).
Topological Sort

• A topological sort of a directed acyclic graph (DAG) is a linear order of all its vertices such that if G contains an edge \((u, v)\), then \(u\) appears before \(v\) in the ordering
  – If the graph contains cycles, no linear ordering is possible.
  – A topological sort can be viewed as an ordering of its vertices along a horizontal line so that all directed edges go from left to right

• DAG are used in many applications to indicate precedence among events
Ordering?
Topological Sort

- $\Theta(V+E)$

**TOPOLOGICAL-SORT($G$)**

1. call DFS($G$) to compute finishing times $f[v]$ for each vertex $v$
2. as each vertex is finished, insert it onto the front of a linked list
3. **return** the linked list of vertices
Example
(Courtesy of Prof. Jim Anderson)

Linked List:
Example

Linked List:
Example

Linked List:

A → B → C → D → E
Example

Linked List:

A → B → D

C → B → D

D → E

E → D

1/4 → 2/3

2/3 → 1/4
Example

Linked List:

1/4 → 2/3
Example

Linked List:

A → B
C

D → E

1/4 → 2/3

Comp 122, Fall 2004
Example

Linked List:

A → B → D
C → 5/6

Linked List:

6/7 → 1/4 → 2/3
C → D → E
Example

Linked List:

B → C → D → E
Example

Linked List:

A
9/10
6/7
5/8
B

D
1/4
2/3
E

Comp 122, Fall 2004
Example

Linked List:
Theorem: Correctness of Topological sort

• **TOPOLOGICAL-SORT(G)** produces a topological sort of a directed acyclic graph G
  
  – Suppose that DFS is run on a given DAG G to determine finishing times for its vertices. It suffices to show that for any pair of distinct vertices \( u, v \), if there is an edge in G from \( u \) to \( v \), then \( f[v] < f[u] \).
    
    • The linear ordering is corresponding to finishing time ordering
  
  – Consider any edge \((u, v)\) explored by DFS(G). When this edge is explored, \( v \) cannot be gray (otherwise, \((u, v)\) will be a back edge). Therefore \( v \) must be either white or black
    
    • If \( v \) is white, \( v \) becomes a descendant of \( u \), \( f[v] < f[u] \) (ex. pants & shoes)
    
    • If \( v \) is black, it has already been finished, so that \( f[v] \) has already been set \( \Rightarrow f[v] < f[u] \) (ex. belt & jacket)
STRONGLY CONNECTED COMPONENTS
Strongly Connected Components

A directed graph is strongly connected iff there is a directed path between any two vertices. The strongly connected components of a graph is a partition of the vertices into subsets (maximal) such that each subset is strongly connected.

Observe that no vertex can be in two maximal components, so it is a partition.
Strongly Connected Components

- $G$ is strongly connected if every pair $(u, v)$ of vertices in $G$ is reachable from one another.

- A strongly connected component (SCC) of $G$ is a maximal set of vertices $C \subseteq V$ such that for all $u, v \in C$, both $u \sim v$ and $v \sim u$ exist.
Component Graph

- $G^{SCC} = (V^{SCC}, E^{SCC})$.
- $V^{SCC}$ has one vertex for each SCC in $G$.
- $E^{SCC}$ has an edge if there’s an edge between the corresponding SCC’s in $G$.
- $G^{SCC}$ for the example considered:
**Lemma 22.13**

Let $C$ and $C'$ be distinct SCC’s in $G$, let $u, v \in C$, $u', v' \in C'$, and suppose there is a path $u \sim u'$ in $G$. Then there cannot also be a path $v' \sim v$ in $G$.

**Proof:**

- Suppose there is a path $v' \sim v$ in $G$.
- Then there are paths $u \sim u' \sim v'$ and $v' \sim v \sim u$ in $G$.
- Therefore, $u$ and $v'$ are reachable from each other, so they are not in separate SCC’s.
Example graph
Example graph
Transpose of a Directed Graph

- $G^T = \text{transpose}$ of directed $G$.
  - $G^T = (V, E^T), E^T = \{(u, v) : (v, u) \in E\}$.
  - $G^T$ is $G$ with all edges reversed.

- Can create $G^T$ in $\Theta(V + E)$ time if using adjacency lists.

- $G$ and $G^T$ have the same SCC’s. ($u$ and $v$ are reachable from each other in $G$ if and only if reachable from each other in $G^T$.)
Algorithm to determine SCCs

**SCC(G)**
1. call DFS(G) to compute finishing times $f[u]$ for all $u$
2. compute $G^T$
3. call DFS($G^T$), but in the main loop, consider vertices in order of decreasing $f[u]$ (as computed in first DFS)
4. output the vertices in each tree of the depth-first forest formed in second DFS as a separate SCC

**Time:** $\Theta(V + E)$.

**Example:** On board.

**Kosaraju's algorithm**, 1978?
Tarjan – 1972
Gabow - 1999 (Cheriyan, Melhorn 1996)
Example
(Courtesy of Prof. Jim Anderson)

$G$

(Courtesy of Prof. Jim Anderson)
Example

$G^T$

\begin{tikzpicture}[node distance=2cm, thick]
  \node (a) [circle, fill=blue!20] {a};
  \node (b) [circle, fill=blue!20, right of=a] {b};
  \node (c) [circle, fill=blue!20, below of=b] {c};
  \node (d) [circle, fill=blue!20, right of=c] {d};
  \node (e) [circle, fill=blue!20, below of=a] {e};
  \node (f) [circle, fill=blue!20, below of=b] {f};
  \node (g) [circle, fill=blue!20, below of=c] {g};
  \node (h) [circle, fill=blue!20, below of=d] {h};

  \draw[->,red] (a) -- (b);
  \draw[->,red] (b) -- (c);
  \draw[->,blue] (c) -- (d);
  \draw[->,blue] (d) -- (a);
  \draw[->,red] (e) -- (f);
  \draw[->,red] (f) -- (g);
  \draw[->,blue] (g) -- (h);
  \draw[->,blue] (h) -- (f);
\end{tikzpicture}
Example

- abe
- cd
- fg
- h
How does it work?

♦ Idea:

» By considering vertices in second DFS in decreasing order of finishing times from first DFS, we are visiting vertices of the component graph in topologically sorted order.

» Because we are running DFS on $G^T$, we will not be visiting any $v$ from a $u$, where $v$ and $u$ are in different components.

♦ Notation:

» $d[u]$ and $f[u]$ always refer to first DFS.

» Extend notation for $d$ and $f$ to sets of vertices $U \subseteq V$:

» $d(U) = \min_{u \in U} \{ d[u] \}$ (earliest discovery time)

» $f(U) = \max_{u \in U} \{ f[u] \}$ (latest finishing time)
**SCCs and DFS finishing times**

**Lemma 22.14**
Let $C$ and $C'$ be distinct SCC’s in $G = (V, E)$. Suppose there is an edge $(u, v) \in E$ such that $u \in C$ and $v \in C'$. Then $f(C) > f(C')$.

**Proof:**
- **Case 1: $d(C) < d(C')$**
  - Let $x$ be the first vertex discovered in $C$.
  - At time $d[x]$, all vertices in $C$ and $C'$ are white. Thus, there exist paths of white vertices from $x$ to all vertices in $C$ and $C'$.
  - By the white-path theorem, all vertices in $C$ and $C'$ are descendants of $x$ in depth-first tree.
  - By the parenthesis theorem, $f[x] = f(C) > f(C')$. 
SCCs and DFS finishing times

**Lemma 22.14**
Let $C$ and $C'$ be distinct SCC’s in $G = (V, E)$. Suppose there is an edge $(u, v) \in E$ such that $u \in C$ and $v \in C'$. Then $f(C) > f(C')$.

**Proof:**
- **Case 2: $d(C) > d(C')$**
  - Let $y$ be the first vertex discovered in $C'$.
  - At time $d[y]$, all vertices in $C'$ are white and there is a white path from $y$ to each vertex in $C'$ $\Rightarrow$ all vertices in $C'$ become descendants of $y$. Again, $f[y] = f(C')$.
  - At time $d[y]$, all vertices in $C$ are also white.
  - By earlier lemma, since there is an edge $(u, v)$, we cannot have a path from $C'$ to $C$.
  - So no vertex in $C$ is reachable from $y$.
  - Therefore, at time $f[y]$, all vertices in $C$ are still white.
  - Therefore, for all $w \in C, f[w] > f[y]$, which implies that $f(C) > f(C')$. 
SCCs and DFS finishing times

**Corollary 22.15**

Let $C$ and $C'$ be distinct SCC’s in $G = (V, E)$. Suppose there is an edge $(u, v) \in E^T$, where $u \in C$ and $v \in C'$. Then $f(C) < f(C')$.

**Proof:**

- $(u, v) \in E^T \Rightarrow (v, u) \in E$.
- Since SCC’s of $G$ and $G^T$ are the same, $f(C') > f(C)$, by Lemma 22.14.
Correctness of SCC

- When we do the second DFS, on $G^T$, start with SCC $C$ such that $f(C)$ is maximum.
  - The second DFS starts from some $x \in C$, and it visits all vertices in $C$.
  - Corollary 22.15 says that since $f(C) > f(C')$ for all $C \neq C'$, there are no edges from $C$ to $C'$ in $G^T$.
  - Therefore, DFS will visit only vertices in $C$.
  - Which means that the depth-first tree rooted at $x$ contains exactly the vertices of $C$. 
Correctness of SCC

- The next root chosen in the second DFS is in SCC $C'$ such that $f(C')$ is maximum over all SCC's other than $C$.
  - DFS visits all vertices in $C'$, but the only edges out of $C'$ go to $C$, which we've already visited.
  - Therefore, the only tree edges will be to vertices in $C'$.

- We can continue the process.

- Each time we choose a root for the second DFS, it can reach only
  - vertices in its SCC—get tree edges to these,
  - vertices in SCC's already visited in second DFS—get no tree edges to these.
Strongly Connected Components

Example
Why does strongly connected component method work?

• Seel CLRS (2-3 pages)

Advanced Algorithmics (6EAP)
Graphs II

Jaak Vilo
2011 Spring
WEIGHTED GRAPH ALGORITHMS
Weighted Graph Algorithms

Beyond DFS/BFS exists an alternate universe of algorithms for *edge-weighted graphs*. Our adjacency list representation quietly supported these graphs:

typedef struct {
    int y;
    int weight;
    struct edgenode *next;
} edgenode;
Minimum Spanning Tree

- **Definition:** Given an undirected graph, and for each edge \((v, u) \in E\), we have a weight \(w(u, v)\) specifying the cost to connect \(u\) and \(v\). Find an acyclic subset \(T \subseteq E\) that connects all of the vertices and whose total weight is minimized

\[
w(T) = \sum_{(u,v) \in T} w(u, v)
\]

- May have more than one MST with the same weight

- **Two classic algorithms:** \(O(E \lg V)\) \(\Rightarrow\) Greedy Algorithms
  - Kruskal’s algorithm
  - Prim’s algorithm
Minimum Spanning Trees

A tree is a connected graph with no cycles. A spanning tree is a subgraph of $G$ which has the same set of vertices of $G$ and is a tree.

A minimum spanning tree of a weighted graph $G$ is the spanning tree of $G$ whose edges sum to minimum weight.

There can be more than one minimum spanning tree in a graph → consider a graph with identical weight edges.
Equal weights in left fully connected graph (a)
Why Minimum Spanning Trees?

The minimum spanning tree problem has a long history – the first algorithm dates back at least to 1926!.
Minimum spanning tree is always taught in algorithm courses since (1) it arises in many applications, (2) it is an important example where greedy algorithms always give the optimal answer, and (3) Clever data structures are necessary to make it work.
In greedy algorithms, we make the decision of what next to do by selecting the best local option from all available choices – without regard to the global structure.
Applications of Minimum Spanning Trees

Minimum spanning trees are useful in constructing networks, by describing the way to connect a set of sites using the smallest total amount of wire. Minimum spanning trees provide a reasonable way for clustering points in space into natural groups. What are natural clusters in the friendship graph?
Minimum Spanning Trees and TSP

When the cities are points in the Euclidean plane, the minimum spanning tree provides a good heuristic for traveling salesman problems. The optimum traveling salesman tour is at most twice the length of the minimum spanning tree.

The Option Traveling System tour is at most twice the length of the minimum spanning tree.

Note: There can be more than one minimum spanning tree considered as a group with identical weight edges.
Fully connected graph. Find a MST?
MST

Images: http://www.personal.kent.edu/~rmuhamma/
MST-approximation of TSP
Growing a Minimum Spanning Tree (MST)

• Generic algorithm
  – Grow MST one edge at a time
  – Manage a set of edges A, maintaining the following loop invariant:
    • Prior to each iteration, A is a subset of some MST
  – At each iteration, we determine an edge (u, v) that can be added to A without violating this invariant
    • \( A \cup \{(u, v)\} \) is also a subset of a MST
    • (u, v) is called a safe edge for A
GENERIC-MST

**GENERIC-MST**$(G, w)$

1. $A \leftarrow \emptyset$
2. while $A$ does not form a spanning tree
3. do find an edge $(u, v)$ that is safe for $A$
4. $A \leftarrow A \cup \{(u, v)\}$
5. return $A$

- Loop in lines 2-4 is executed $|V| - 1$ times
  - Any MST tree contains $|V| - 1$ edges
  - The execution time depends on how to find a safe edge
How to Find A Safe Edge?

• **Theorem.** Let A be a subset of E that is included in some MST, let (S, V-S) be any cut of G that respects A, and let (u, v) be a light edge crossing (S, V-S). Then edge (u, v) is safe for A
  
  – **Cut** (S, V-S): a partition of V
  – **Crossing edge**: one endpoint in S and the other in V-S
  – A cut respects a set of A of edges if no edges in A crosses the cut
  – A light edge crossing a cut if its weight is the minimum of any edge crossing the cut
Figure 23.2  Two ways of viewing a cut \((S, V - S)\) of the graph from Figure 23.1. (a) The vertices in the set \(S\) are shown in black, and those in \(V - S\) are shown in white. The edges crossing the cut are those connecting white vertices with black vertices. The edge \((d, c)\) is the unique light edge crossing the cut. A subset \(A\) of the edges is shaded; note that the cut \((S, V - S)\) respects \(A\), since no edge of \(A\) crosses the cut. (b) The same graph with the vertices in the set \(S\) on the left and the vertices in the set \(V - S\) on the right. An edge crosses the cut if it connects a vertex on the left with a vertex on the right.
Illustration of Theorem 23.1

- $A=\{(a,b), (c, i), (h, g), \{g, f\}\}$
- $S=\{a, b, c, i, e\}; V-S = \{h, g, f, d\}$ ➔ many kinds of cuts satisfying the requirements of Theorem 23.1

- $(c, f)$ is the light edges crossing $S$ and $V-S$ and will be a safe edge
Proof of Theorem 23.1

• Let $T$ be a MST that includes $A$, and assume $T$ does not contain the light edge $(u, v)$, since if it does, we are done.

• Construct another MST $T'$ that includes $A \cup \{(u, v)\}$ from $T$
  
  – Next slide
  
  – $T' = T - \{(x, y)\} \cup (u, v)$
  
  – $T'$ is also a MST since $W(T') = W(T) - w(x, y) + w(u, v) \leq W(T)$

• $(u, v)$ is actually a safe edge for $A$
  
  – Since $A \subseteq T$ and $(x, y) \notin A \Rightarrow A \subseteq T'$
  
  – $\Rightarrow A \cup \{(u, v)\} \subseteq T'$
Figure 23.3  The proof of Theorem 23.1. The vertices in $S$ are black, and the vertices in $V - S$ are white. The edges in the minimum spanning tree $T$ are shown, but the edges in the graph $G$ are not. The edges in $A$ are shaded, and $(u, v)$ is a light edge crossing the cut $(S, V - S)$. The edge $(x, y)$ is an edge on the unique path $p$ from $u$ to $v$ in $T$. A minimum spanning tree $T'$ that contains $(u, v)$ is formed by removing the edge $(x, y)$ from $T$ and adding the edge $(u, v)$. 
Properties of GENERIC-MST

• As the algorithm proceeds, the set $A$ is always acyclic
• $G_A=(V, A)$ is a forest, and each of the connected component of $G_A$ is a tree
• Any safe edge $(u, v)$ for $A$ connects distinct component of $G_A$, since $A \cup \{(u, v)\}$ must be acyclic
• Corollary 23.2. Let $A$ be a subset of $E$ that is included in some MST, and let $C = (V_C, E_C)$ be a connected components (tree) in the forest $G_A=(V, A)$. If $(u, v)$ is a light edge connecting $C$ to some other components in $G_A$, then $(u, v)$ is safe for $A$
The Algorithms of Kruskal and Prim

• Kruskal’s Algorithm
  – A is a forest
  – The safe edge added to A is always a least-weight edge in the graph that connects two distinct components

• Prim’s Algorithm
  – A forms a single tree
  – The safe edge added to A is always a least-weight edge connecting the tree to a vertex not in the tree
Prim’s Algorithm

- The edges in the set $A$ always forms a single tree
- The tree starts from an arbitrary root vertex $r$ and grows until the tree spans all the vertices in $V$
- At each step, a light edge is added to the tree $A$ that connects $A$ to an isolated vertex of $G_{A}=(V, A)$
- Greedy since the tree is augmented at each step with an edge that contributes the minimum amount possible to the tree’s weight
Prim’s Algorithm

If $G$ is connected, every vertex will appear in the minimum spanning tree. If not, we can talk about a minimum spanning forest.

Prim’s algorithm starts from one vertex and grows the rest of the tree an edge at a time.

As a greedy algorithm, which edge should we pick? The cheapest edge with which can grow the tree by one vertex without creating a cycle.
**Prim’s Algorithm (Pseudocode)**

During execution each vertex \( v \) is either in the tree, *fringe* (meaning there exists an edge from a tree vertex to \( v \)) or *unseen* (meaning \( v \) is more than one edge away).

Prim-MST(G)

- Select an arbitrary vertex \( s \) to start the tree from.
- While (there are still non-tree vertices)
  - Select the edge of minimum weight between a tree and node
  - Add the selected edge and vertex to the tree \( T_{prim} \).

This creates a spanning tree, since no cycle can be introduced, but is it minimum?
Prim’s Algorithm in Action

G

Prim(G,A)

Kruskal(G)
Key idea of Prim’s algorithm

Select a vertex to be a tree-node

while (there are non-tree vertices)
{
    if (there is no edge connecting a tree node with a non-tree node)
        return “no spanning tree”

    select an edge of minimum weight between a tree node and a non-tree node

    add the selected edge and its new vertex to the tree

}  
return tree
Prim’s Algorithm (Cont.)

• How to efficiently select the safe edge to be added to the tree?
  – Use a min-priority queue Q that stores all vertices not in the tree
    • Based on key[v], the minimum weight of any edge connecting v to a vertex in the tree
      – Key[v] = ∞ if no such edge
  • π[v] = parent of v in the tree
• A = {(v, π[v]): v∈V-{r}-Q} ➔ finally Q = empty
Prim's Algorithm

1. for each $u \in V$
2. do $D[u] \leftarrow \infty$
3. $D[r] \leftarrow 0$
4. MH $\leftarrow$ make-heap($D, V, {}$) // No edges
5. $T \leftarrow \emptyset$

7. while $MH \neq \emptyset$ do
8. $(u,e) \leftarrow MH$.extractMin()
9. add $(u,e)$ to $T$
10. for each $v \in \text{Adjacent}(u)$
11. do if $v \in MH$ && $w(u,v) < D[v]$
12. then $D[v] \leftarrow w(u,v)$
13. MH.decreaseDistance($D[v], v, (u,v)$)
14. return $T$ // $T$ is a MST

Lines 1-5 initialize the min-heap (MH) to contain all vertices. Distances for all vertices, except $r$, are set to infinity. $r$ is the starting vertex of the $T$. The $T$ so far is empty.

Add the closest vertex and edge to current $T$.

Get all adjacent vertices $v$ of $u$, update $D$ of each non-tree vertex adjacent to $u$.

Store the current minimum weight edge and updated distance in the MH.
MST-PRIM(G, w, r)
   1  for each \( u \in V[G] \)
   2      do key[u] \( \leftarrow \infty \)
   3         \( \pi[u] \leftarrow \text{NIL} \)
   4  key[r] \( \leftarrow 0 \)
   5  Q \( \leftarrow V[G] \)
   6  while \( Q \neq \emptyset \)
   7      do u \( \leftarrow \text{EXTRACT-MIN}(Q) \)
   8         for each \( v \in Adj[u] \)
   9         do if \( v \in Q \) and \( w(u, v) < \text{key}[v] \)
   10            then \( \pi[v] \leftarrow u \)
   11                   key[v] \( \leftarrow w(u, v) \)
Illustration of MST-PRIM

\[ u=a, \text{adj}[a] = \{b, h\} \]
\[ \pi[b] = \pi[h] = a \]
\[ \text{key}[b] = 4; \text{key}[h] = 8 \]

\[ u=b, \text{adj}[b] = \{a, c, h\} \]
\[ \pi[h] = a; \pi[c] = b \]
\[ \text{key}[h] = 8; \text{key}[c] = 8 \]

\[ u=c, \text{adj}[c] = \{b, i, d, f\} \]
\[ \pi[d] = \pi[i] = \pi[f] = c \]
\[ \text{key}[h] = 8; \text{key}[d] = 7; \text{key}[i] = 2; \text{key}[f] = 4 \]
Properties of MST-PRIM

• Prior to each iteration of the while loop of lines 6—11
  – \( A = \{(v, \pi[v]): v \in V-\{r\}-Q\}\)
  – The vertices already placed into the MST are those in \( V-Q \)
  – For all vertices \( v \in Q \), if \( \pi[v] \neq NIL \), then \( key[v] < \infty \) and \( key[v] \) is the weight of a light edge \((v, \pi[v])\) connecting \( v \) to some vertex already placed into the MST

• Line 7: identify a vertex \( u \in Q \) incident on a light edge crossing \((V-Q, Q)\) \( \rightarrow \) add \( u \) to \( V-Q \) and \((u, \pi[u])\) to \( A \)

• Lines 8—11: update key and \( \pi \) of every vertex \( v \) adjacent to \( u \) but not in the tree
Performance of MST-PRIM

• Use binary min-heap to implement the min-priority queue Q
  – BUILD-MIN-HEAP (line 5): $O(V)$
  – The body of while loop is executed $|V|$ times
    • EXTRACT-MIN: $O(lg V)$
  – The for loop in lines 8-11 is executed $O(E)$ times altogether
    • Line 11: DECREASE-KEY operation: $O(lg V)$
  – Total performance = $O(V lg V + E lg V) = O(E lg V)$

• Use Fibonacci heap to implement the min-priority queue Q
  – $O(E + V lg V)$
Why is Prim Correct?

We use a proof by contradiction: Suppose Prim’s algorithm does not always give the minimum cost spanning tree on some graph. If so, there is a graph on which it fails. And if so, there must be a first edge \((x, y)\) Prim adds such that the partial tree \(V'\) cannot be extended into a minimum spanning tree.
But if \((x, y)\) is not in \(MST(G)\), then there must be a path in \(MST(G)\) from \(x\) to \(y\) since the tree is connected. Let \((v, w)\) be the first edge on this path with one edge in \(V'\) replacing it with \((x, y)\) we get a spanning tree with smaller weight, since \(W(v, w) > W(x, y)\). Thus you did not have the MST!!
Kruskal’s Algorithm

Since an easy lower bound argument shows that every edge must be looked at to find the minimum spanning tree, and the number of edges \( m = O(n^2) \), Prim’s algorithm is optimal in the worst case. Is that all she wrote?

The complexity of Prim’s algorithm is independent of the number of edges. Can we do better with sparse graphs? Yes! Kruskal’s algorithm is also greedy. It repeatedly adds the smallest edge to the spanning tree that does not create a cycle.
Why is Kruskal’s algorithm correct?

Again, we use proof by contradiction. Suppose Kruskal’s algorithm does not always give the minimum cost spanning tree on some graph. If so, there is a graph on which it fails. And if so, there must be a first edge \((x, y)\) Kruskal adds such that the set of edges cannot be extended into a minimum spanning tree.

When we added \((x, y)\) there previously was no path between \(x\) and \(y\), or it would have created a cycle. Thus if we add \((x, y)\) to the optimal tree it must create a cycle. At least one edge in this cycle must have been added after \((x, y)\), so it must have a heavier weight.

Deleting this heavy edge leave a better MST than the optimal tree? A contradiction!
How fast is Kruskal’s algorithm?

What is the simplest implementation?

- Sort the $m$ edges in $O(m \lg m)$ time.
- For each edge in order, test whether it creates a cycle the forest we have thus far built – if so discard, else add to forest. With a BFS/DFS, this can be done in $O(n)$ time (since the tree has at most $n$ edges).

The total time is $O(mn)$, but can we do better?
Fast Component Tests Give Fast MST

Kruskal’s algorithm builds up connected components. Any edge where both vertices are in the same connected component create a cycle. Thus if we can maintain which vertices are in which component fast, we do not have test for cycles!

- **Same component** \((v_1, v_2)\) – Do vertices \(v_1\) and \(v_2\) lie in the same connected component of the current graph?

- **Merge components** \((C_1, C_2)\) – Merge the given pair of connected components into one component.
Fast Kruskal Implementation

Put the edges in a heap
\[ \text{count} = 0 \]
while (\text{count} < n - 1) do
    get next edge \((v, w)\)
    if (\text{component}(v) ≠ \text{component}(w))
        add to \(T\)
        \text{component}(v) = \text{component}(w)

If we can test components in \(O(\log n)\), we can find the MST in \(O(m \log m)\)!

Question: Is \(O(m \log n)\) better than \(O(m \log m)\)?
Union-Find Programs

We need a data structure for maintaining sets which can test if two elements are in the same and merge two sets together. These can be implemented by union and find operations, where

- \textit{Find}(i) \quad \text{Return the label of the root of tree containing element } i, \text{ by walking up the parent pointers until there is no where to go.}

- \textit{Union}(i,j) \quad \text{Link the root of one of the trees (say containing } i) \text{ to the root of the tree containing the other (say } j) \text{ so } \text{find}(i) \text{ now equals } \text{find}(j).
This path compression will let us do better than $O(n \log n)$ for $n$ union-finds.

$O(n)$? Not quite ... Difficult analysis shows that it takes $O(n \alpha(n))$ time, where $\alpha(n)$ is the inverse Ackerman function and $\alpha($ number of atoms in the universe $) = 5$. 
Problem of the Day

Suppose we are given the minimum spanning tree $T$ of a given graph $G$ (with $n$ vertices and $m$ edges) and a new edge $e = (u, v)$ of weight $w$ that we will add to $G$. Give an efficient algorithm to find the minimum spanning tree of the graph $G + e$. Your algorithm should run in $O(n)$ time to receive full credit, although slower but correct algorithms will receive partial credit.
Prim vs Kruskal vs Boruvka

Figure 20.10
PFS implementation of Prim’s MST algorithm
With PFS, Prim’s algorithm processes just the vertices and edges closest to the MST (in gray).

Figure 20.13
Kruskal’s MST algorithm
This sequence shows 1/4, 1/2, 3/4, and the full MST as it evolves.

Figure 20.16
Boruvka’s MST algorithm
The MST evolves in just four stages for this example (top to bottom).
Table 20.1 Cost of MST algorithms

This table summarizes the cost (worst-case running time) of various MST algorithms considered in this chapter. The formulas are based on the assumptions that an MST exists (which implies that $E$ is no smaller than $V - 1$) and that there are $X$ edges not longer than the longest edge in the MST (see Property 20.10). These worst-case bounds may be too conservative to be useful in predicting performance on real graphs. The algorithms run in near-linear time in a broad variety of practical situations.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>worst-case cost</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prim (standard)</td>
<td>$V^2$</td>
<td>optimal for dense graphs</td>
</tr>
<tr>
<td>Prim (PFS, heap)</td>
<td>$E \lg V$</td>
<td>conservative upper bound</td>
</tr>
<tr>
<td>Prim (PFS, $d$-heap)</td>
<td>$E \log_d V$</td>
<td>linear unless extremely sparse</td>
</tr>
<tr>
<td>Kruskal</td>
<td>$E \lg E$</td>
<td>sort cost dominates</td>
</tr>
<tr>
<td>Kruskal (partial sort)</td>
<td>$E + X \lg V$</td>
<td>cost depends on longest edge</td>
</tr>
<tr>
<td>Boruvka</td>
<td>$E \lg V$</td>
<td>conservative upper bound</td>
</tr>
</tbody>
</table>
SINGLE-SOURCE SHORTEST PATHS
(CHAPTER 24)
4-letter words, distance 1

LEDA tutorial
Shortest paths between nodes in graph

• Practical applications

• Transportation
  – Cheapest or quickest way to travel from A to B

• Motion planning
  – Most natural way for a cartoon character to navigate between places

• Communications
  – Time to send a message; diameter of a graph,...
Example: Predictive Mobile text Entry Messaging...

What was the message?
INPUT

Blank Recognition

Candidate Construction

Sentence Disambiguating

OUTPUT

GIVE ME A RING.
The weight of each edge is a function of the probability that these two words will be next to each other in a sentence. ‘hive me’ would be less than ‘give me’, for example. The final system worked extremely well – identifying over 99% of characters correctly based on grammatical and statistical constraints.
Problem Definition

• Given a weighted, directed graph $G=(V, E)$ with weight function $w: E \rightarrow \mathbb{R}$. The weight of path $p=<v_0, v_1, ..., v_k>$ is the sum of the weights of its constituent edges:
  
  $$ w(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i) $$

• We define the **shortest-path weight from $u$ to $v$** by
  
  $$ \delta(u, v) = \begin{cases} 
  \min \{ w(p): u \xrightarrow{p} v \} & \text{If there is a path from } u \text{ to } v, \\
  \infty & \text{Otherwise.}
  \end{cases} $$

• A shortest path from vertex $u$ to vertex $v$ is then defined as any path with $w(p)=\delta(u, v)$
Variants

- **Single-source shortest paths** problem – greedy
  - Finds all the shortest path of vertices reachable from a single source vertex s

- **Single-destination shortest-path** problem
  - By reversing the direction of each edge in the graph, we can reduce this problem to a single-source problem

- **Single-pair shortest-path** problem
  - No algorithm for this problem are known that run asymptotically faster than the best single-source algorithm in the worst case

- **All-pairs shortest-path** problem – dynamic programming
  - Can be solved faster than running the single-source shortest-path problem for each vertex
Figure 24.2  (a) A weighted, directed graph with shortest-path weights from source $s$.  (b) The shaded edges form a shortest-paths tree rooted at the source $s$.  (c) Another shortest-paths tree with the same root.
Optimal Substructure of A Shortest-Path

• Lemma 24.1 (Subpath of shortest paths are shortest paths). Let $p=<v_1, v_2, ..., v_k>$ be a shortest path from vertex $v_1$ to $v_k$, and for any $i$ and $j$ such that $1 \leq i \leq j \leq k$, let $p_{ij} = <v_{1i}, v_2, ..., v_j>$ be the subpath of $p$ from vertex $v_i$ to $v_j$. Then $p_{ij}$ is a shortest path from vertex $v_i$ to $v_j$. 
Examples of shortest paths depending on start node
Negative-Weight Edges and Cycles

• Cannot contain a negative-weight cycle
• Of course, a shortest path cannot contain a positive-weight cycle
Relaxation

• For each vertex \( v \in V \), we maintain an attribute \( d[v] \), which is an upper bound on the weight of a shortest path from source \( s \) to \( v \). We call \( d[v] \) a shortest-path estimate.

```
INITIALIZE-SINGLE-SOURCE(G, s)
1 for each vertex \( v \in V[G] \)
2 do \( d[v] \leftarrow \infty \)
3 \( \pi[v] \leftarrow NIL \)
4 \( d[s] \leftarrow 0 \)
```

Predecessor of \( v \) in the shortest path
Relaxation (Cont.)

- Relaxing an edge \((u, v)\) consists of testing whether we can improve the shortest path found so far by going through \(u\) and, if so, update \(d[v]\) and \(\pi[v]\)

\[
\text{RELAX}(u, v, w)
\]

1. \(\text{if } d[v] > d[u] + w(u, v)\)
2. \(\text{then } d[v] \leftarrow d[u] + w(u, v)\)
3. \(\pi[v] \leftarrow u\)

**Figure 24.3** Relaxation of an edge \((u, v)\) with weight \(w(u, v) = 2\). The shortest-path estimate of each vertex is shown within the vertex. (a) Because \(d[v] > d[u] + w(u, v)\) prior to relaxation, the value of \(d[v]\) decreases. (b) Here, \(d[v] \leq d[u] + w(u, v)\) before the relaxation step, and so \(d[v]\) is unchanged by relaxation.

By Triangle Inequality
Bellman-Ford

Bellman-Ford (G, w, s)

1. Initialise-Single-Source(G,S)

2. for i=1 to |G.V|-1 /* n-1 rounds */

3. for each edge (u,v) ∈ G.E

4. RELAX( u, v, w )

5. for each edge (u,v) ∈ G.E

6. if v.d > u.d + w(u,v)

7. return FALSE

8. return TRUE
Figure 24.4  The execution of the Bellman-Ford algorithm. The source is vertex $s$. The $d$ values are shown within the vertices, and shaded edges indicate predecessor values: if edge $(u, v)$ is shaded, then $\pi[v] = u$. In this particular example, each pass relaxes the edges in the order $(t, x), (t, y), (t, z), (x, t), (y, x), (y, z), (z, x), (z, s), (s, t), (s, y)$. (a) The situation just before the first pass over the edges. (b)–(e) The situation after each successive pass over the edges. The $d$ and $\pi$ values in part (e) are the final values. The Bellman-Ford algorithm returns TRUE in this example.
Bellman-Ford

• $O(V E)$

• Just repeatedly relax all edges.
  – Allow $V$ cycles to propagate through the network
Shortest paths on a DAG

1. DAG-Shortest-path(G,w,s)
2. topologically sort vertices
3. Initialise-single-source(G,s)
4. for each vertex u in topological order
   for each vertex v ∈ G.Adj[u]
5. RELAX(u,v,w)

O(V + E)
24.2 Single-source shortest paths in directed acyclic graphs
Dijkstra’s Algorithm

• Solve the single-source shortest-paths problem on a weighted, directed graph when all edge weights are nonnegative

• Data structure
  – S: a set of vertices whose final shortest-path weights have already been determined
  – Q: a min-priority queue keyed by their d values

• Idea
  – Repeatedly select the vertex u∈V-S (kept in Q) with the minimum shortest-path estimate, add s u to S, and relax all edges leaving u
Dijkstra’s Algorithm (Cont.)

\[
\text{DIJKSTRA}(G, w, s)
\]

1. \text{INITIALIZE-SINGLE-SOURCE}(G, s)
2. \( S \leftarrow \emptyset \)
3. \( Q \leftarrow V[G] \)
4. \text{while } Q \neq \emptyset \text{ do}\n5. \quad u \leftarrow \text{EXTRACT-MIN}(Q)\n6. \quad S \leftarrow S \cup \{u\}\n7. \quad \text{for each vertex } v \in \text{Adj}[u] \text{ do}\n8. \quad \text{RELAX}(u, v, w)\]

Note: relax requires updating of min values in \( Q \)
Figure 24.6  The execution of Dijkstra's algorithm. The source \( s \) is the leftmost vertex. The shortest-path estimates are shown within the vertices, and shaded edges indicate predecessor values. Black vertices are in the set \( S \), and white vertices are in the min-priority queue \( Q = V - S \). (a) The situation just before the first iteration of the while loop of lines 4–8. The shaded vertex has the minimum \( d \) value and is chosen as vertex \( u \) in line 5. (b)–(f) The situation after each successive iteration of the while loop. The shaded vertex in each part is chosen as vertex \( u \) in line 5 of the next iteration. The \( d \) and \( \pi \) values shown in part (f) are the final values.
Analysis of Dijkstra’s Algorithm

• Correctness: Theorem 24.6 (Loop invariant)
• Min-priority queue operations
  – INSERT (line 3)
  – EXTRACT-MIN( line 5)
  – DECREASE-KEY(line 8)
• Time analysis
  – Line 4-8: while loop $\rightarrow O(V)$
  – Line 7-8: for loop and relaxation $\rightarrow |E|$
  – Running time depends on how to implement min-priority queue
    • Simple array: $O(V^2+E) = O(V^2)$
    • Binary min-heap: $O((V+E)\log V)$
    • Fibonacci min-heap: $O(V\log V + E)$
Edsger Wybe Dijkstra was one of the most influential members of computing science's founding generation. Among the domains in which his scientific contributions are fundamental are:

- algorithm design
- programming languages
- program design
- operating systems
- distributed processing
- formal specification and verification
- design of mathematical arguments
Dijkstra's Algo

1) Dijkstra is a Greedy based algorithm and similar to Prim's MST algo.
2) Dijkstra doesn't work for negative weight edges.
3) Time complexity of Dijkstra is $O(|E| + |V|\log|V|)$
4) Dijkstra's algorithm is usually the working principle behind link-state routing protocols, OSPF and IS-IS
Examples of shortest paths depending on start node

Figure 21.12
SPT examples
These three examples show growing SPTs for three different source locations: left edge (top), upper left corner (center), and center (bottom).
All pairs shortest paths

• Diameter of a graph (longest shortest path)
• Calculate the shortest path from each source
• Find the longest shortest path...

• Means to estimate/approximate it
Fast Fully Dynamic Landmark-based Estimation of Shortest Path Distances in Very Large Graphs

Konstantin Tretyakov* †
University of Tartu
Estonia
kt@ut.ee

Abel Armas-Cervantes
University of Tartu
Estonia
abel.armas@ut.ee

Luciano García-Bañuelos
University of Tartu
Estonia
luciano.garcia@ut.ee

Jaak Vilo
University of Tartu
Estonia
jaak.vilo@ut.ee

Marlon Dumas
University of Tartu
Estonia
marlon.dumas@ut.ee

ABSTRACT
Computing the shortest path between a pair of vertices in a graph is a fundamental primitive in graph algorithmics. Classical exact methods for this problem do not scale up to contemporary, rapidly evolving social networks with hun-

Keywords
Graph Databases, Shortest Paths, Social Networks, Landmarks, Trees, Dynamic Updates

ACM Conference on Information and Knowledge Management (CIKM) 2011
Fast Fully Dynamic Landmark-based Estimation of Shortest Path Distances in Very Large Graphs
Fast Fully Dynamic Landmark-based Estimation of Shortest Path Distances in Very Large Graphs
Fast **Fully Dynamic** Landmark-based Estimation of Shortest Path Distances in Very Large Graphs
Socially sensitive search
Lisa kontakt

Otsi Skype'i kasutajaid ja lisa neid oma kontaktide hulka. Sisesta nende Skype'i nimi, täisnimi või meiliaadress ja klõpsa Leia.

<table>
<thead>
<tr>
<th>Mary</th>
<th>Leia</th>
</tr>
</thead>
</table>

Saad lisada ka MySpaceIM kontakte. Siseta nende MySpaceIM nimi ja klõpsa "Leia".

Vali isik, keda otsisid ja vajuta 'Lisa kontakt':

<table>
<thead>
<tr>
<th>93 inimest on Skype's</th>
<th>MySpace's on 7 kasutajat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Täisnimi</td>
<td>Skype'i nimi</td>
</tr>
<tr>
<td>Syrtis Mary</td>
<td>syrtis-mary</td>
</tr>
<tr>
<td>Mary</td>
<td>abcmarias14</td>
</tr>
<tr>
<td>Mary</td>
<td>aflickofiner</td>
</tr>
<tr>
<td>MARY</td>
<td>amore4ever4</td>
</tr>
<tr>
<td>Mary</td>
<td>bella_luz</td>
</tr>
<tr>
<td>Mary</td>
<td>bluester88</td>
</tr>
<tr>
<td>mary</td>
<td>colpi67</td>
</tr>
<tr>
<td>mary</td>
<td>cucciolotta2993</td>
</tr>
<tr>
<td>Mary</td>
<td>foreverdepp247</td>
</tr>
<tr>
<td>Mary</td>
<td>gg83-mary</td>
</tr>
</tbody>
</table>
Socially sensitive search
Naïve approach (Breadth-First-Search) requires 5-20 minutes
Landmark-based estimation

Mary Lee

Mary Ann
Landmark-based estimation

Basic Method

Mary Lee

[Diagram with distances and locations labeled]

3 <= d <= 7

3 <= d <= 5

Mary Ann
Landmark-based estimation
Shortest path tree
Least common ancestor
Least common ancestor

Shortcutting
Combining multiple landmarks
Combining multiple landmarks
Combining multiple landmarks

Landmarks-BFS
Combining multiple landmarks

Landmarks-BFS

Given two nodes U and V:
1. Collect all paths from U and V to all landmarks
2. Run a BFS* on the induced subgraph

* or Dijkstra, or A*, or anything else
Landmark-based approximation

- Speed
- Basic Method
- LCA
- Shortcutting
- Landmarks-BFS
- Accuracy
Landmark-based approximation

Dynamic

Basic Method

LCA

Shortcutting

Landmarks-BFS
Insertion of an edge
Deletion – more complicated
## Evaluation - Data

| Dataset   | $|V|$ | $|E|$ | $d$  | $\Delta$ | $|S|/|V|$ | $t_{BFS}$ |
|-----------|-----|-----|------|----------|---------|----------|
| DBLP      | 770K| 2.6M| 6.3  | 23       | 85%     | 345 ms   |
| Orkut     | 3.1M| 117M| 5.7  | 10       | 100%    | 8 sec    |
| Twitter   | 41.7M| 1.2B| 4.2  | 24       | 100%    | 9 min    |
| Skype     | 454M| 3.1B| 6.5  | 59       | 85%     | 20 min   |

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Graph file</th>
<th>Landmark file</th>
<th>Basic</th>
<th>LCA/SC/LBFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP</td>
<td>27M</td>
<td>753K</td>
<td>3.0M</td>
<td></td>
</tr>
<tr>
<td>Orkut</td>
<td>918M</td>
<td>3.0M</td>
<td>12.0M</td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>9.3G</td>
<td>40M</td>
<td>160M</td>
<td></td>
</tr>
<tr>
<td>Skype</td>
<td>27G</td>
<td>433M</td>
<td>1.7G</td>
<td></td>
</tr>
</tbody>
</table>

*per each landmark*
### Timings: Query

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>No. of Landmarks</th>
<th>20</th>
<th>60</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skype</td>
<td>Basic</td>
<td></td>
<td>0.18</td>
<td>0.56</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>LCA</td>
<td></td>
<td>1.06</td>
<td>2.43</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>SC</td>
<td></td>
<td>1.22</td>
<td>2.92</td>
<td>4.85</td>
</tr>
<tr>
<td></td>
<td>LBFS</td>
<td></td>
<td>5.10</td>
<td>13.24</td>
<td>16.25</td>
</tr>
</tbody>
</table>

Time for a batch of 500 queries / 500, in ms

Linux, mmap, 32 cores, 256GB RAM
Timings: Updates

<table>
<thead>
<tr>
<th></th>
<th>DBLP</th>
<th>Orkut</th>
<th>Twitter</th>
<th>Skype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertion</td>
<td>1μs</td>
<td>10μs</td>
<td>10μs</td>
<td>30μs</td>
</tr>
<tr>
<td>Deletion*</td>
<td>100μs</td>
<td>2ms</td>
<td>12ms</td>
<td>11ms</td>
</tr>
</tbody>
</table>

* very non-uniform
Outline

- Improvement to Basic Landmark method
- Dynamic updates
- Landmark selection
- Evaluation
Landmark selection method

- Landmark is good if it covers many shortest paths

  **Highest degree**

  **Best coverage**
Best Coverage

A
C
E

B
D
F

•••
Best Coverage
Best Coverage
Error: \( d' - \frac{d}{d} \)
# Timings: Landmark selection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Highest degree</th>
<th>Best coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP</td>
<td>140 ms</td>
<td>2 min</td>
</tr>
<tr>
<td>Orkut</td>
<td>2 s</td>
<td>15 min</td>
</tr>
<tr>
<td>Twitter</td>
<td>22 s</td>
<td>15 h</td>
</tr>
<tr>
<td>Skype</td>
<td>1 min</td>
<td>54 h</td>
</tr>
</tbody>
</table>
### Summary (Skype graph)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network size</strong></td>
<td>500M nodes 3B edges</td>
</tr>
<tr>
<td><strong>Landmark selection time (HD)</strong></td>
<td>1 min / 54hr</td>
</tr>
<tr>
<td><strong>Landmark computation time</strong></td>
<td>20 min x 100</td>
</tr>
<tr>
<td><strong>Total space for 100 landmarks</strong></td>
<td>170G</td>
</tr>
<tr>
<td><strong>Avg query time (SC/LBFS)</strong></td>
<td>5ms / 16ms</td>
</tr>
<tr>
<td><strong>Avg edge insertion time</strong></td>
<td>0.030 ms</td>
</tr>
<tr>
<td><strong>Avg edge deletion time</strong></td>
<td>11ms</td>
</tr>
<tr>
<td><strong>Avg relative error (SC/LBFS)</strong></td>
<td>18% / 15%</td>
</tr>
</tbody>
</table>
Summary

LCA  Shortcutting  Landmarks-BFS

Dynamic updates  Highest degree  Best coverage
### Questions

<table>
<thead>
<tr>
<th></th>
<th>Precomp.</th>
<th>Space</th>
<th>Query</th>
<th>Insertion</th>
<th>Deletion</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCA</td>
<td>1+100x20min</td>
<td>170G</td>
<td>5ms / 16ms</td>
<td>0.030 ms</td>
<td>11ms</td>
<td>18% / 15%</td>
</tr>
<tr>
<td>Shortcutting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Landmarks-BFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic updates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highest degree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Generalizations

- To weighted graph:
  - Use weighted shortest path trees
  - The dynamic update algorithm becomes slightly more complicated

- To directed graph:
  - Use two SPTs per landmark
Improvements

- Parallelization possible at most stages
- “Evolutionary” on-line selection of landmarks
- Use of landmark-based heuristics with A* for exact path possible (Goldberg et al., Ikeda et al.)
## Timings: Query / Twitter

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>No. of Landmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Twitter</td>
<td>Basic</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>LCA</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>SC</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>LBFS</td>
<td>240</td>
</tr>
</tbody>
</table>
Eucl- Path

We’ll look more closely at this with the A* algorithm (heuristic search)
Euclidean Networks

• In applications where networks model maps, our primary interest is often in finding the best route from one place to another. In this section, we examine a strategy for this problem: a fast algorithm for the source–sink shortest-path problem in Euclidean networks, which are networks whose vertices are points in the plane and whose edge weights are defined by the geometric distances between the points.

• These networks satisfy two important properties that do not necessarily hold for general edge weights. First, the distances satisfy the triangle inequality: The distance from s to d is never greater than the distance from s to x plus the distance from x to d. Second, vertex positions give a lower bound on path length: No path from s to d will be shorter than the distance from s to d. The algorithm for the source–sink shortest-paths problem that we examine in this section takes advantage of these two properties to improve performance.
Calculating paths by matrix operations
Paths of length 2
Paths of length 2
\[ c_{ij} = \sum_{k} a_{ik} b_{kj} \]
\[ c_{st} = \sum_{i} a_{si} b_{it} \]

```plaintext
for (s = 0; s < V; s++)
    for (t = 0; t < V; t++)
        for (i = 0, C[s][t] = 0; i < V; i++)
            C[s][t] += A[s][i]*B[i][t];
```
The textbook algorithm for computing the product of two $V$-by-$V$ matrices computes, for each $s$ and $t$, the dot product of row $s$ in the first matrix and row $t$ in the second matrix, as follows:

```plaintext
for (s = 0; s < V; s++)
    for (t = 0; t < V; t++)
        for (i = 0, C[s][t] = 0; i < V; i++)
            C[s][t] += A[s][i]*B[i][t];
```

In matrix notation, we write this operation simply as $C = A \times B$. This operation is defined for matrices comprising any type of entry for which $0$, $+$, and $\times$ are defined. In particular, if the matrix entries are either `true` or `false` and we interpret $a+b$ to be the logical `or` operation and $a*b$ to be the logical `and` operation, then we have Boolean matrix multiplication. In Java, we can use the following version:

```plaintext
for (s = 0; s < V; s++)
    for (t = 0; t < V; t++)
        for (i = 0, C[s][t] = false; i < V; i++)
            if (A[s][i] && B[i][t]) C[s][t] = true;
```
Diagonal 1 = self-loop
Diagonal 0 or 1

\[ G^*G \]
Transitive closure

• Transitive closure of a digraph $G$ is a graph $G'$ with same vertices, and edge between any $u$ and $v$ from $G$ if there is a path from $u$ to $v$ in $G$
Transitive closure

\[ G \ast G \ast G \ast \ldots \]

\[ G[i][j] \text{ and } G[j][k] \Rightarrow G[i][k] \]

Exists link via \( j \)

1 on diagonal - link to itself

Figure 19.13
Transitive closure

This digraph (top) has just eight directed edges, but its transitive closure (bottom) shows that there are directed paths connecting 19 of the 30 pairs of vertices. Structural properties of the digraph are reflected in the transitive closure. For example, rows 0, 1, and 2 in the adjacency matrix for the transitive closure are identical (as are columns 0, 1, and 2) because those vertices are on a directed cycle in the digraph.
Property 19.6  We can compute the transitive closure of a digraph by constructing the latter’s adjacency matrix $A$, adding self-loops for every vertex, and computing $A^V$.

Proof: Continuing the argument in the previous paragraph, $A^3$ has an edge for every path of length less than or equal to 3 in the digraph, $A^4$ has an edge for every path of length less than or equal to 4 in the digraph, and so forth. We do not need to consider paths of length greater than $V$ because of the pigeonhole principle: Any such path must revisit some vertex (since there are only $V$ of them) and therefore adds no information to the transitive closure because the same two vertices are connected by a directed path of length less than $V$ (which we could obtain by removing the cycle to the revisited vertex).
19.3. Reachability and Transitive Closure

Complexity...

- for \( i=1 \) to \( |V| \) do \( V^{(i)} = V^{(i-1)}*V \)
- \( V^3 \) operations for \( V^2, V^3, \ldots V^v \)
- \( \Rightarrow O(V^4) \)

- Use exponential: \( 2 \Rightarrow 4 \Rightarrow 8 \Rightarrow 16 \ldots \) steps.
- \( V^2*V^2=V^4, V^4*V^4=V^8, \ldots \Rightarrow O( \lfloor \log V \rfloor \times V^3 ) \)
- Can we avoid so many cycles?
closure with just one operation of this kind, building up the transitive closure from the adjacency matrix in place, as follows:

```c
for (i = 0; i < V; i++)
    for (s = 0; s < V; s++)
        for (t = 0; t < V; t++)
            if (A[s][i] && A[i][t]) A[s][t] = true;
```

This classical method, invented by S. Warshall in 1962, is the method of choice for computing the transitive closure of dense digraphs. The code is similar to the code that we might try to use to square a Boolean matrix in place: The difference (which is significant!) lies in the order of the for loops.

**Property 19.7** With Warshall's algorithm, we can compute the transitive closure of a digraph in time proportional to $V^3$. 
Multiply:

```c
for (s = 0; s < V; s++)
    for (t = 0; t < V; t++)
        for (i = 0; C[s][t] = 0; i < V; i++)
            C[s][t] += A[s][i] * B[i][t];
```

Transitive closure:

```c
for (i = 0; i < V; i++)
    for (s = 0; s < V; s++)
        for (t = 0; t < V; t++)
            if (A[s][i] && A[i][t]) A[s][t] = true;
```
for (i = 0; i < V; i++)
    for (s = 0; s < V; s++)
        for (t = 0; t < V; t++)
            if (A[s][i] && A[i][t]) A[s][t] = true;

Paths via 0
Paths via 1 (including 0-1, 1-0)
...
Property 19.7  With Marshall’s algorithm, we can compute the transitive closure of a digraph in time proportional to $V^3$.

Proof: The running time is immediately evident from the structure of the code. We prove that it computes the transitive closure by induction on $i$. After the first iteration of the loop, the matrix has true in row $s$ and column $t$ if and only if the digraph has either the edge $s-t$ or the path $s-0-t$. The second iteration checks all the paths between $s$ and $t$ that include 1 and perhaps 0, such as $s-1-t$, $s-1-0-t$, and $s-0-1-t$. We are led to the following inductive hypothesis: The $i$th iteration of the loop sets the bit in row $s$ and column $t$ in the matrix to true if and only if there is a directed path from $s$ to $t$ in the digraph that does not include any vertices with indices greater than $i$ (except possibly the endpoints $s$ and $t$). As just argued, the condition is true when $i$ is 0, after the first iteration of the loop. Assuming that it is true for the $i$th iteration of the loop, there is a path from $s$ to $t$ that does not include any vertices with indices greater than $i+1$ if and only if (i) there is a path from $s$ to $i+1$ and a path from $i+1$ to $t$, neither of which includes any vertices with indices greater than $i$, in which case $A[s][t]$ was set on a previous iteration of the loop (by the inductive hypothesis); or (ii) there is a path from $s$ to $i+1$ and $A[i+1][t]$ were previously set to true (by hypothesis), so the inner loop sets $A[s][t]$. •
Proof

• Proof: transitive closure by induction on i.
• Iteration 1: either s-t or the path s-0-t.
• Iteration 2: all the paths between s and t that include 1 and perhaps 0, such as s-1-t, s-1-0-t, and s-0-1-t.
• Inductive hypothesis: The ith iteration of the loop sets the bit (s, t) to true iff there is a directed path from s to t in the digraph that does not include any vertices with indices greater than i (except possibly the endpoints s and t).
• Assuming that it is true for the \( i \)th iteration of the loop, there is a path from \( s \) to \( t \) that does not include any vertices with indices greater than \( i+1 \) iff
  
  – (i) there is a path from \( s \) to \( t \) without indices >\( i \), in which case \( A[s][t] \) was set on a previous iteration of the loop (inductive hypothesis)
  
  – (ii) there is a path from \( s \) to \( i+1 \) and a path from \( i+1 \) to \( t \), neither of which includes any vertices with indices greater than \( i \) (except endpoints), in which case \( A[s][i+1] \) and \( A[i+1][t] \) were previously set to true (by hypothesis), so the inner loop sets \( A[s][t] \).
• How to further improve?

• Test for A[s][i] early
Program 19.3 Warshall’s algorithm

The constructor for class GraphTC computes the transitive closure of G in the private data field T so that clients can use GraphTC objects to test whether any given vertex in a digraph is reachable from any other given vertex. The constructor initializes T with a copy of G, adds self-loops, then uses Warshall’s algorithm to complete the computation. We use a DenseGraph object for the transitive closure T because the algorithm needs an efficient implementation of the edge existence test (see Section 17.5).

class GraphTC
{
  private DenseGraph T;

  GraphTC(Graph G)
  {
    T = GraphUtilities.densecopy(G);
    for (int s = 0; s < T.V(); s++)
      T.insert(new Edge(s, s));
    for (int i = 0; i < T.V(); i++)
      for (int s = 0; s < T.V(); s++)
        if (T.edge(s, i))
          for (int t = 0; t < T.V(); t++)
            if (T.edge(i, t))
              T.insert(new Edge(s, t));

    boolean reachable(int s, int t)
    {
      return T.edge(s, t);
    }
  }
}
Table 19.1 Empirical study of transitive-closure algorithms

This table shows running times that exhibit dramatic performance differences for various algorithms for computing the transitive closure of random digraphs, both dense and sparse. For all but the adjacency-lists DFS, the running time goes up by a factor of 8 when we double $V$, which supports the conclusion that it is essentially proportional to $V^3$. The adjacency-lists DFS takes time proportional to $VE$, which explains the running time roughly increasing by a factor of 4 when we double both $V$ and $E$ (sparse graphs) and by a factor of about 2 when we double $E$ (dense graphs), except that list-traversal overhead degrades performance for high-density graphs.

<table>
<thead>
<tr>
<th></th>
<th>sparse (10$V$ edges)</th>
<th></th>
<th>dense (250 vertices)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$V$</td>
<td>$W$</td>
<td>$W^*$</td>
</tr>
<tr>
<td>25</td>
<td>5000</td>
<td>289</td>
<td>203</td>
</tr>
<tr>
<td>50</td>
<td>25000</td>
<td>309</td>
<td>226</td>
</tr>
<tr>
<td>125</td>
<td>50000</td>
<td>315</td>
<td>232</td>
</tr>
<tr>
<td>250</td>
<td>100000</td>
<td>326</td>
<td>246</td>
</tr>
</tbody>
</table>

Key:
- W Warshall’s algorithm (Section 19.3)
- $W^*$ Improved Warshall’s algorithm (Program 19.3)
- A DFS, adjacency-matrix representation (Programs 19.4 and 17.7)
- L DFS, adjacency-lists representation (Program 19.4 and 17.9)
Random walks...

Graph:

- A \rightarrow B \ 0.95
- A \rightarrow C \ 0.05
- B \rightarrow D \ 0.7
- B \rightarrow E \ 0.3
- C \rightarrow E \ 1.0
- D \rightarrow A \ 1.0
- E \rightarrow D \ 0.2
- E \rightarrow A \ 0.8
Matrix:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.95</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
</tr>
</tbody>
</table>
vilo@muhu:~/Algorithmics/MarkovWalk$ perl GraphMarkovChain.pl Weighted_Graph.txt
Matrix:
0   0.95  0.05  0   0
0   0    0    0.7  0.3
0   0    0    0   1.0
1.0 0    0    0   0
0.8 0    0    0.2 0

Random Walk with 100000 steps
FINAL: 0.32589 0.30974 0.01614 0.23799 0.11024
FINAL: 32589 30974 1614  23799 11024
Matrix multiplications with 10000 steps
Matrix:
0.326051516139543 0.309748940332566 0.0163025758069772 0.238669709814146 0.10922
0.326051516139543 0.309748940332566 0.0163025758069771 0.238669709814145 0.10922
0.326051516139543 0.309748940332566 0.0163025758069772 0.238669709814146 0.10922
0.326051516139543 0.309748940332566 0.0163025758069772 0.238669709814146 0.10922
0.326051516139543 0.309748940332566 0.0163025758069772 0.238669709814146 0.10922

vilo@muhu:~/Algorithmics/MarkovWalk$
Finding the modules

Public datasets for *H.sapiens*
- IntAct: Protein interactions (PPI), 18773 interactions
- IntAct: PPI via orthologs from IntAct, 6705 interactions
- MEM: gene expression similarity over 89 tumor datasets, 46286 interactions
- Transfac: gene regulation data, 5183 interactions
Finding the modules

Public datasets for *H. sapiens*

- **IntAct**: Protein interactions (PPI), 18773 interactions
- **IntAct**: PPI via orthologs from IntAct, 6705 interactions
- **MEM**: gene expression similarity over 89 tumor datasets, 46286 interactions
- **Transfac**: gene regulation data, 5183 interactions
Module evaluation

GO: Transforming growth factor beta signaling pw.
  embryonic development, gastrulation
KEGG: Cell cycle, cancers, WNT pw.

GO: JAK-STAT cascade, Kinase inhibitor activity
  Insulin receptor signaling pw.
KEGG: Type II diabetes mellitus

GO: Brain development
  Pigment granule
  Melanine metabolic process
MCL clustering algorithm

- Markov (Chain Monte Carlo) Clustering
  - [http://www.micans.org/mcl/](http://www.micans.org/mcl/)

- Random walks according to edge weights

- Follow the different paths according to their probability

- Regions that are traversed “often” form clusters
With this, the MCL algorithm can be written as

\[
G \text{ is a graph} \\
\text{add loops to } G \\
\text{set } \Gamma \text{ to some value } \quad \# \text{ affects granularity} \\
\text{set } M_1 \text{ to be the matrix of random walks on } G \\
\text{while (change) } \{ \\
\quad M_2 = M_1 \ast M_1 \quad \# \text{ expansion} \\
\quad M_1 = \Gamma(M_2) \quad \# \text{ inflation} \\
\quad \text{change} = \text{difference}(M_1, M_2) \\
\} \\
\text{set CLUSTERING as the components of } M_1 \quad \# \text{ see below}
\]
MAXIMUM FLOW

Max-Flow Min-Cut Theorem (Ford Fukerson’s Algorithm)
What is Network Flow?

Flow network is a directed graph $G=(V,E)$ such that each edge has a non-negative capacity $c(u,v) \geq 0$.

Two distinguished vertices exist in $G$ namely:

- Source (denoted by $s$): In-degree of this vertex is 0.
- Sink (denoted by $t$): Out-degree of this vertex is 0.

Flow in a network is an integer-valued function $f$ defined on the edges of $G$ satisfying $0 \leq f(u,v) \leq c(u,v)$, for every edge $(u,v)$ in $E$. 
What is Network Flow?

- Each edge \((u,v)\) has a non-negative capacity \(c(u,v)\).
- If \((u,v)\) is not in \(E\) assume \(c(u,v)=0\).
- We have source \(s\) and sink \(t\).
- Assume that every vertex \(v\) in \(V\) is on some path from \(s\) to \(t\).

Following is an illustration of a network flow:

![Network Flow Diagram](image)

- \(c(s,v_1)=16\)
- \(c(v_1,s)=0\)
- \(c(v_2,s)=0 \ldots\)
Conditions for Network Flow

For each edge \((u,v)\) in \(E\), the flow \(f(u,v)\) is a real valued function that must satisfy the following 3 conditions:

- **Skew Symmetry**: \(\forall u,v \in V, \ f(u,v) = -f(v,u)\) (inflow = -outflow)
- **Capacity Constraint**: \(\forall u,v \in V, \ f(u,v) \leq c(u,v)\) (flow < capacity)
- **Flow Conservation**: \(\forall u \in V - \{s,t\}, \sum_{v \in V} f(u,v) = 0\) (net flow = 0)

Skew symmetry condition implies that \(f(u,u)=0\).
The Value of a Flow.

The value of a flow is given by:

\[ |f| = \sum_{v \in V} f(s, v) = \sum_{v \in V} f(v, t) \]

The flow into the node is same as flow going out from the node and thus the flow is conserved. Also the total amount of flow from source \( s \) = total amount of flow into the sink \( t \).
Example of a flow

Table illustrating Flows and Capacity across different edges of graph above:

\[ f_{s,1} = 9, c_{s,1} = 10 \text{ (Valid flow since } 10 > 9) \]
\[ f_{s,2} = 6, c_{s,2} = 6 \text{ (Valid flow since } 6 \geq 6) \]
\[ f_{1,2} = 1, c_{1,2} = 1 \text{ (Valid flow since } 1 \geq 1) \]
\[ f_{1,t} = 8, c_{1,t} = 8 \text{ (Valid flow since } 8 \geq 8) \]
\[ f_{2,t} = 7, c_{2,t} = 10 \text{ (Valid flow since } 10 > 7) \]

The flow across nodes 1 and 2 are also conserved as flow into them = flow out.
The Maximum Flow Problem

Given a Graph $G (V,E)$ such that:

- $x_{i,j} =$ flow on edge $(i,j)$
- $u_{i,j} =$ capacity of edge $(i,j)$
- $s =$ source node
- $t =$ sink node

Maximize $v$

Subject To

- $\Sigma_j x_{ij} - \Sigma_j x_{ji} = 0$ for each $i \neq s, t$
- $\Sigma_j x_{sj} = v$
- $0 \leq x_{ij} \leq u_{ij}$ for all $(i,j) \in E.$

In simple terms maximize the $s$ to $t$ flow, while ensuring that the flow is feasible.
A Cut in a network is a partition of $V$ into $S$ and $T$ ($T=V-S$) such that
$s$ (source) is in $S$ and $t$ (target) is in $T$. 

![Diagram of a network with labeled edges and a cut set.](Image)
Capacity of Cut \((S, T)\)

\[
c(S, T) = \sum_{u \in S, v \in T} c(u, v)
\]
Min s-t cut (Also called as a Min Cut) is a cut of minimum capacity.
Flow of Min Cut (Weak Duality)

Let $f$ be the flow and let $(S,T)$ be a cut. Then $|f| \leq \text{CAP}(S,T)$.

In maximum flow, minimum cut problems forward edges are full or saturated and the backward edges are empty because of the maximum flow. Thus maximum flow is equal to capacity of cut. This is referred to as weak duality.

Proof:

$$|f| = \sum_{e \text{ out of } S} f(e) - \sum_{e \text{ in to } S} f(e) \left\{ 0 \right\}$$

$$= \sum_{e \text{ out of } S} f(e)$$

$$= \sum_{e \text{ out of } S} u(e)$$

$$= \text{cap}(S, T)$$
Methods

Max-Flow Min-Cut Theorem

- The Ford-Fulkerson Method
- The Preflow-Push Method
The Ford-Fulkerson Method

- Try to improve the flow, until we reach the maximum value of the flow.

- The residual capacity of the network with a flow $f$ is given by:

The residual capacity ($rc$) of an edge $(i,j)$ equals $c(i,j) - f(i,j)$ when $(i,j)$ is a forward edge, and equals $f(i,j)$ when $(i,j)$ is a backward edge. Moreover, the residual capacity of an edge is always non-negative.

$$c_f(u, v) = c(u, v) - f(u, v)$$

Original Network

Residual Network
Begin

x := 0; // x is the flow.
create the residual network G(x);
while there is some directed path from s to t in G(x) do begin
let P be a path from s to t in G(x);
Δ := δ(P);
send Δ units of flow along P;
update the r's;
end
end {the flow x is now maximum}. 
Augmenting Paths (A Useful Concept)

Definition:

An augmenting path $p$ is a simple path from $s$ to $t$ on a residual network that is an alternating sequence of vertices and edges of the form $s,e_1,v_1,e_2,v_2,...,e_k,t$ in which no vertex is repeated and no forward edge is saturated and no backward edge is free.

Characteristics of augmenting paths:

- We can put more flow from $s$ to $t$ through $p$.
- The edges of residual network are the edges on which residual capacity is positive.
- We call the maximum capacity by which we can increase the flow on $p$ the residual capacity of $p$.

$$c_f(p) = \min \{c_f(u,v) : (u,v) \text{ is on } p\}$$
The Ford-Fulkerson’s Algorithm

\[
\text{FORDFULKERSON}(G,E,s,t)
\]

\[\text{FOREACH } e \in E\]
\[f(e) \leftarrow 0\]

\[G_f \leftarrow \text{residual graph}\]

\[\text{WHILE } (\text{there exists augmenting path } P)\]
\[f \leftarrow \text{augment}(f, P)\]
\[\text{update } G_f\]

\[\text{ENDWHILE}\]

\[\text{RETURN } f\]

\[
\text{AUGMENT}(f,P)
\]

\[b \leftarrow \text{bottleneck}(P)\]

\[\text{FOREACH } e \in P\]

\[\text{IF } (e \in E)\]
\[\quad // \text{ backwards arc}\]
\[\quad f(e) \leftarrow f(e) + b\]

\[\text{ELSE}\]
\[\quad // \text{ forward arc}\]
\[\quad f(e^R) \leftarrow f(e) - b\]

\[\text{RETURN } f\]
Proof of correctness of the algorithm

**Lemma:** At each iteration all residual capacities are integers.

**Proof:** It’s true at the beginning. Assume it’s true after the first k-1 augmentations, and consider augmentation k along path P.

The residual capacity $\Delta$ of P is the smallest residual capacity on P, which is integral.

After updating, we modify the residual capacities by 0 or $\Delta$, and thus residual capacities stay integers.

**Theorem:** Ford-Fulkerson’s algorithm is finite

**Proof:** The capacity of each augmenting path is at least 1. The augmentation reduces the residual capacity of some edge (s,j) and doesn’t increase the residual capacity for some edge (s,i) for any i.

So the sum of residual capacities of edges out of s keeps decreasing, and is bounded below 0.

Number of augmentations is $O(nC)$ where C is the largest of the capacity in the network.
When is the flow optimal?

A flow $f$ is maximum flow in $G$ if:

1. The residual network $G_f$ contains no more augmented paths.
2. $|f| = c(S,T)$ for some cut $(S,T)$ (a min-cut)

Proof:

1. Suppose there is an augmenting path in $G_f$ then it implies that the flow $f$ is not maximum, because there is a path through which more data can flow. Thus if flow $f$ is maximum then residual n/w $G_f$ will have no more augmented paths.

2. Let $v=Fx(S,T)$ be the flow from $s$ to $t$. By assumption $v=\text{CAP}(S,T)$. By Weak duality, the maximum flow is at most $\text{CAP}(S,T)$. Thus the flow is maximum.
The Ford-Fulkerson Augmenting Path Algorithm for the Maximum Flow Problem

15.082 and 6.855J (MIT OCW)
This is the original network, and the original residual network.
Ford-Fulkerson Max Flow

Find any s-t path in G(x)
Ford-Fulkerson Max Flow

Determine the capacity $\Delta$ of the path.

Send $\Delta$ units of flow in the path.
Update residual capacities.
Ford-Fulkerson Max Flow

Find any s-t path
Determine the capacity $\Delta$ of the path.

Send $\Delta$ units of flow in the path.

Update residual capacities.
Find any s-t path
Determine the capacity $\Delta$ of the path.

Send $\Delta$ units of flow in the path.

Update residual capacities.
Find any s-t path
Ford-Fulkerson Max Flow

Determine the capacity $\Delta$ of the path.

Send $\Delta$ units of flow in the path.
Update residual capacities.
Ford-Fulkerson Max Flow

Find any s-t path
Determine the capacity $\Delta$ of the path.

Send $\Delta$ units of flow in the path.
Update residual capacities.
Ford-Fulkerson Max Flow

There is no s-t path in the residual network. This flow is optimal.
Ford-Fulkerson Max Flow

These are the nodes that are reachable from node s.
Here is the optimal flow
Counterexample for termination

Ülesanne 54. Vaatleme voogu järgmisel joonisel.

Olgu \( R = \frac{\sqrt{5} - 1}{2} \) (siis \( R^n = R^{n+1} + R^{n+2} \)) ja ülejäänud servadel suured läbilaskevõimed. Olgu esimene suurendav ahel \( s - a - d - t \) ning järgmised suurendavad ahelad (tsüklis)

1. \( s - c - f - d - a - b - e - t \)
2. \( s - b - e - f - c - a - d - t \)
3. \( s - a - d - e - b - c - f - t \).

Näita, et Ford-Fulkersoni algoritm ei lõpeta tööd.
Distribution & Transportation

<table>
<thead>
<tr>
<th>supply</th>
<th>channels</th>
<th>cost</th>
<th>cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: 4</td>
<td>0-3:</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1: 4</td>
<td>0-7:</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2: 6</td>
<td>1-4:</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>distribution</td>
<td>2-4:</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2-9:</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3-5:</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3-6:</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>4-5:</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>demand</td>
<td>4-6:</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>7: 7</td>
<td>5-7:</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8: 3</td>
<td>5-8:</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>9: 4</td>
<td>6-9:</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>supply</th>
<th>channels</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: 3</td>
<td>0-6:</td>
<td>2</td>
</tr>
<tr>
<td>1: 4</td>
<td>0-7:</td>
<td>1</td>
</tr>
<tr>
<td>2: 6</td>
<td>0-8:</td>
<td>5</td>
</tr>
<tr>
<td>3: 3</td>
<td>1-6:</td>
<td>3</td>
</tr>
<tr>
<td>4: 2</td>
<td>1-5:</td>
<td>1</td>
</tr>
<tr>
<td>demand</td>
<td>2-8:</td>
<td>3</td>
</tr>
<tr>
<td>5: 6</td>
<td>2-9:</td>
<td>4</td>
</tr>
<tr>
<td>6: 6</td>
<td>3-6:</td>
<td>2</td>
</tr>
<tr>
<td>7: 7</td>
<td>3-7:</td>
<td>1</td>
</tr>
<tr>
<td>8: 3</td>
<td>4-9:</td>
<td>6</td>
</tr>
<tr>
<td>9: 4</td>
<td>4-5:</td>
<td>3</td>
</tr>
</tbody>
</table>
Assigning teachers to classes

Teacher likes to teach C1, C4, C6

Every course will need a nr of teachers

Every teacher has a maximal capacity to teach

“Likes” – by weight

How would you solve it?
Job placement:
6 people, 6 jobs, preferences...

<table>
<thead>
<tr>
<th>Alice</th>
<th>Adobe</th>
<th>Bob</th>
<th>Apple</th>
<th>Dave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adobe</td>
<td>Bob</td>
<td>Apple</td>
<td>Alice</td>
<td>Bob</td>
</tr>
<tr>
<td>Apple</td>
<td>Dave</td>
<td>Alice</td>
<td>Bob</td>
<td>Dave</td>
</tr>
<tr>
<td>HP</td>
<td>HP</td>
<td>IBM</td>
<td>Carol</td>
<td>Frank</td>
</tr>
<tr>
<td>HP</td>
<td>IBM</td>
<td>Carol</td>
<td>Eliza</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>Sun</td>
<td>Carol</td>
<td>Eliza</td>
<td></td>
</tr>
<tr>
<td>Sun</td>
<td>Yahoo</td>
<td>Frank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yahoo</td>
<td>Bob</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>Bob</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP</td>
<td>Sun</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sun</td>
<td>Yahoo</td>
<td>Frank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yahoo</td>
<td>Bob</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>Bob</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP</td>
<td>Sun</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sun</td>
<td>Yahoo</td>
<td>Frank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yahoo</td>
<td>Bob</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>Bob</td>
<td>Eliza</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Converting the Matching problem to Network Flow
Converting Matching to Network Flow
Converting Optimal Bipartite Matching to Network Flow