31251 – Data Structures and Algorithms

Week 5 - Graphs Part I, Autumn 2020

Xianzhi Wang

Does anyone read these titles?

- Graphs
- Traversing Graphs
- Greedy Algorithms
  - e.g., Prim's Algorithm, Kruskal's Algorithm

Graphs

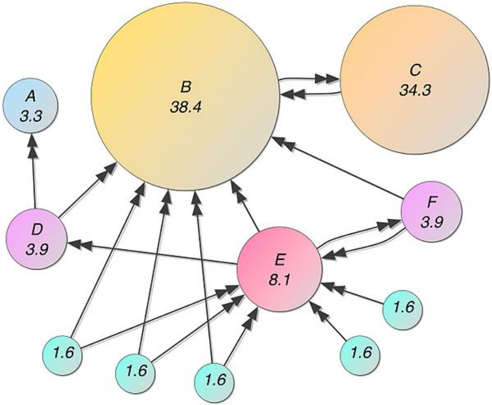
Graphs as Mathematical Objects

- Graphs are an incredibly useful modelling tool.
- A simple graph consists of:
  - A set of elements called *vertices*.
  - A set of unordered pairs of distinct vertices called *edges*.
  - There is only one edge between a pair of vertices.
- Other types of graph come from altering these conditions:
  - Edges with directions gives *directed graphs*.
  - More than one edge per pair gives *multi-graphs*.
  - More than two vertices per edge gives *hypergraphs*.
  - Edges (and vertices) can be weighted, labelled, etc.

Graphs as Mathematical Objects

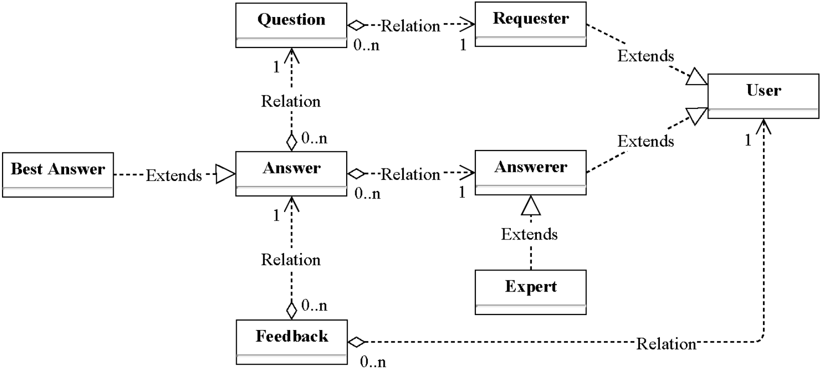
- The vertices model interesting things:
  - Computers
  - Processes
  - Proteins
  - Production facilities
  - Websites
- The edges model relationships between interesting things:
  - Network links
  - Shared resources
  - Protein interactions
  - Transport links
  - Hyper links
- Used somehow in virtually every part of computer science.

- Webpages with Authoritativeness as vertices.
- Directed Links among webpages.
- https://en.wikipedia.org/wiki/PageRank



Another Example: Stack Overflow

- Heterogeneous Attributed Entities
- Heterogeneous Multiple Relations
- Multimodal information (e.g., text, images, code, numeric)



- If two vertices have an edge between them, they are *adjacent*.
- If a vertex is one of the pair that forms an edge, it is *incident* to that edge.
- The number of edges incident to a vertex is the *degree* of that vertex.
- Graphs will be denoted with uppercase letters like G, H.
- Vertices will be denoted by lowercase letters like u, v.
- Edges will be denoted by lowercase letters like e, d. or by their incident vertices:
    - In undirected graphs uv.
    - In directed graphs, edges are called *arcs*: (u, v) – u is the *tail*, v is the *head*.

As an abstract data structure, a graph needs to support a lot of basic operations:

- `void addVertex(Vertex v)`
- `void removeVertex(Vertex v)`
- `void addEdge(Vertex u, Vertex v)`
- `void removeEdge(Vertex u, Vertex v)`
- `bool adjacent(Vertex u, Vertex v)`
- `size_t degree(Vertex u)`
- return vertices in a graph
- return edges incident to a vertex
- return vertices adjacent to a vertex

Graphs as Data Structures – Adjacency Matrix

- Simplest form:
  - Edges are stored as a two-dimensional matrix (e.g. `vector<vector<bool> > edges` or `bool edges[][]`).
  - `edge[i][j] == true` means vertex i is adjacent to j.
- Some enhancements:
  - Can use a numeric (`int`, `double`, etc.) matrix to give weighted edges.
  - Easily supports directed and undirected graphs—asymmetric/symmetric matrices.
  - If vertices have associated data, we can store them separately (another array would make matching indices easy).
- Quick access—$O(1)$, space—$O(n^2)$ (with n vertices).

Graphs as Data Structures – Adjacency List

- Each vertex has associated with a list of its adjacent vertices, forming an array of linked lists or similar.
  - Slower to determine adjacency – $O(n)$, but faster to return all adjacent vertices – $O(1)$.
  - Most compact space representation $O(m + n)$ where m is the number of edges in the graph – we have to store something for each vertex and edge anyway, so this is the best we can do.

- Easy to modify for more complex edge and vertex data structures.

- Works best for sparse graphs (few edges per vertex).

Graphs as Data Structures – Object Oriented

- The extreme version:
  - We have classes for `Vertex`, `Edge` and `Graph`.
  - `Vertex` contains a list of its `Edge`s.
  - Each `Edge` knows its endpoints.
  - the `Graph` knows about everything.
  - *Lots* of references to keep track of.
  - Tends to be the slow way to do things, but has a nice match to the conceptual version.

- Evaluation Criteria:
  - Search Time
  - Space
  - Support for modification, etc.

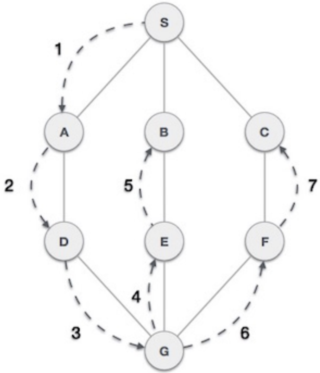
Traversing Graphs

Moving About in Graphs

- Many algorithms rely on being able to explore the graph.
- Visiting history: We need to be able to keep track of which vertices we've been to.
- Visiting order: We also need a way of picking which neighbour to move to next:
  - We can use an inherent order on the vertices (by label, number, etc.), or pick an arbitrary order.

- Recursively pick an unvisited neighbour to visit in a depthward motion.
- Backtrack to previous recorded unvisited neighbours when a dead end occurs.
- Can be implemented recursively, or iteratively using a stack.

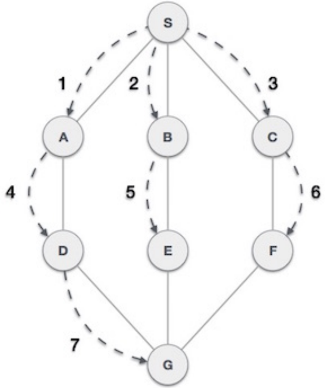
Function *DFT(Vertex v)*
    mark v as visited;
    visit(v);
    for *each neighbour u of v* **do**
        if *u is unmarked* **then**
            DFT(u);



https://www.tutorialspoint.com/data_structures_algorithms/depth_first_traversal.htm

```
Function DFT()
    pick starting vertex v;
    Stack unprocessed = new Stack();
    unprocessed.push(v);
    while !unprocessed.isEmpty() do
        Vertex u = unprocessed.pop();
        if u is unmarked then
            visit(u);
            mark(u);
            for each neighbour w of u do
                unprecessed.push(w);
```

- Pick a starting vertex and put it in a queue.
- Iteratively take a vertex from the queue, visit it and place all its neighbours in the queue.
- It's inherently iterative, it's not impossible to implement recursively, just silly.



```
Function BFT()
    pick starting vertex v;
    Queue unprocessed = new Queue();
    unprocessed.push(v);
    while !unprocessed.isEmpty() do
        Vertex u = unprocessed.pop();
        if u is unmarked then
            visit(u);
            mark u;
            for each neighbour w of u do
                unprocessed.push(w);
```

Breadth First vs. Depth First

- Some graphs produce the same traversal order for both.
- Which one to use will depend on the application.
- The iterative versions of both are actually identical – just swap the stack and the queue.
- Both $O(n + m)$. (Why?)

Greedy Algorithms

Greedy Algorithms – in narrow sense

- “A greedy algorithm is any algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage with the intent of finding a global optimum.”
- Greedy Algorithms usually work when the problem satisfies two properties:
  - Optimal Substructure: “An optimal solution to the problem contains the optimal subsolutions to the subproblems.”
  - Greedy Choice Property: You “make whatever choice seems best at the moment and then solve the subproblems that arise later.” (you don’t have to come back and fix things).

https://en.wikipedia.org/wiki/Greedy_algorithm

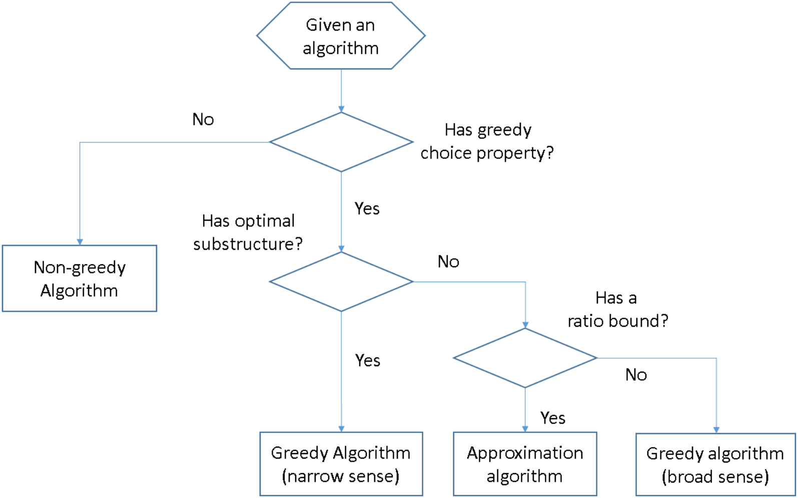
Greedy is ~~good~~cheap – in broad sense

"In many problems, a greedy strategy does not usually produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a globally optimal solution in a reasonable amount of time."

- Greedy Algorithms are based on picking what looks good now.
- One of the simplest algorithmic paradigms.
- Usually easy to implement.
- Only really works for certain types of problems.
- For some problems, a greedy approach may produce the worst solution!

https://en.wikipedia.org/wiki/Greedy_algorithm

Is My Algorithm Greedy?



Spanning Trees of Graphs

A Motivating Problem

Consider the following problem:

- *A company has to connect cities with fibre optic cable such that each city has a (possibly multi-hop) link to every other city. The company knows the cost of linking each pair of cities, and wants to accomplish its task with a minimum cost.*
- We can use a weighted graph to model this problem, but what are we looking for?
  - A set of edges that connects all the vertices.
  - No unneeded edges.
- So we want a subgraph that includes all the vertices and has the minimum total edge weight, i.e., a tree.

- A subgraph is a subset of the vertices and edges of a graph that form a graph.
- A subgraph is spanning if it includes all the vertices of the original graph.
- A spanning subgraph is a spanning tree if it contains no cycles.
- A spanning tree is a minimum spanning tree if it has minimum total (edge) weight over all possible spanning trees of that graph (is it unique?).

In unweighted graphs (or a graph where all edge weights are the same), *any* spanning tree is a minimum spanning tree.

- We can compute one from a depth-first or breadth-first traversal.

Function *df_spanning_tree(Vertex v, Tree t)*

```
mark v as visited;
for each neighbour u of v do
    if u is unmarked then
        add edge vu to t;
        df_spanning_tree(u,t);
```

- If we have different weights on the edges, a simple traversal is not enough.
- There are two main algorithms:
  - Prim's Algorithm
  - Kruskal's Algorithm
- If we have time, we'll look at Borůvka's (Sollin's) Algorithm.
- These are all greedy algorithms, with similar but slightly different approaches.

- Given a *connected* graph G

  1. Pick a starting vertex v (however you want), add v to the partially complete tree T.
  2. While $|T| < n$
     1. Let E' be the set of edges uv where $u \in T$ and $v \in G \setminus T$.
     2. Let uv be the edge of smallest weight in E'.
     3. Add uv to the edges of T and v to the vertices of T.
  3. Return T.

- Keep track of which edges and vertices are in the tree.
- Pick the next smallest edge that extends the tree, add it.
- Keep going until the whole graph is spanned.

Function *prim_spanning_tree(Graph G, Tree T)*

    add to T a random vertice v_0 from G;
    while $|T| < |G|$ **do**
        find $v, u = \underset{v \in T \wedge u \in G \setminus T}{\arg\min}\ weight(vu)$;
        add vu to T;

Prim's Algorithm Complexity

- If we use an adjacency matrix, and search for the edges: $O(n^2)$.
- Putting the edges into a binary heap, with the graph stored as an adjacent list: $O((n+m)\log n) = O(m \log n)$.
- Using Fibonacci heap (don't worry about what this is) and adjacency lists: $O(m + n \log n)$.

- Prim's algorithm grows the spanning tree by greedily picking the best next edge.
- Kruskal's algorithm approaches the problem more globally - start with a lot of trees (a *forest*), and pick the best edge to connect two components.

- Given a *connected* graph G, start with n trees $\{T_i\}$, each with one vertex.
  1. While there is more than one tree
     1. Pick the smallest edge uv such that u is in one tree T_i, and v is in another T_j.
     2. Merge T_i and T_j by adding uv.
  2. Return the final tree T.

Function *Bruskal_spanning_tree(Forest F)*

    regard each vertice to F as a tree;
    while $|F| > 1$ **do**
        find $v, u, T_x, T_y = \underset{v \in T_x \wedge u \in T_y \wedge T_x \in F \wedge T_y \in F \wedge T_x != T_y}{\arg\min} weight(vu)$;
        add vu to T_x, thereby merging T_x and T_y;

- By labelling vertices with which component they're in, we can get $O(n \cdot m)$ – not great.
- If we sorted the edges, and employ an efficient disjoint set data structure (haven't seen one in the course): $O(m \log m) \leq O(m \log n)$ – about the same as the normal Prim implementation.
- If the edges can be sorted efficiently by counting sort or radix sort or similar, we can get $O(m \cdot \alpha(n))$, where $\alpha(n)$ is the inverse of the single valued Ackermann function (look it up some time).

- Invented in 1926 by Otakar Borůvka – see computers aren't necessary for *computer science*.
- Reinvented three more times, lastly by Sollin in 1965.
- Works like a cross between Prim's and Kruskal's algorithms

- Given a graph G, initialise n tree $\{T_i\}$, each containing one vertex.

1. While there is more than one tree
    1. For each component tree
        1. Pick the smallest outgoing edge (connecting this component to another)
        2. Add this edge to the trees, merging them.
2. Return the single remaining tree T

Borůvka's Algorithm – Complexity

- The outer loop only needs to execute $O(\log n)$ times – we halve the number of components at each step.
- Along with search for the edges at each iteration, we get $O(m \log n)$ without too much fiddling.
- A similar approach as used for Kruskal's can be used to get $O(m \cdot \alpha(n))$.
- A randomised version exists with $O(m)$ expected running time – remember this is linear in the size of the graph, about as fast as possible.

Reverse-Delete – Kruskal's Other Algorithm

- Appears in the same paper as Kruskal's algorithm.
- Sort of like a backwards Kruskal's.
- We remove edges, instead of adding them, and see what we're left with at the end.

Reverse-Delete – Kruskal's Other Algorithm

Given a weighted graph G:

1. Sort the edges by decreasing weight.
2. While edges remain to be processed
   1. Take the next biggest edge e.
   2. Check if deleting e will create more components than you already have.
   3. If not, delete it, otherwise keep it.
3. Return the remaining graph.

- We can sort the edges in $O(m \log m)$.
- We can check the connectivity in $O(\log n (\log \log n)^3)$.
- So in total we get $O(m \log n (\log \log n)^3)$ time.

What about disconnected graphs?

- If the graph has several disconnected components, we can't get a spanning tree (trees have to be connected).
- We can get a spanning forest.
- The algorithms we have seen so far won't work (why not? can they be fixed?).

Further Reading

Lemma

Given a connected, weighted graph G, Prim's algorithm produces a minimum spanning tree of G.

Proof:

- As G is connected, it is (or at least should be) clear that Prim's algorithm produces a tree that spans the graph. Thus we need only argue that it is a *minimum* spanning tree.
- Let T_P be the tree produced by Prim's algorithm.
- Assume for contradiction that there exists a minimum spanning tree T_M og G and that the weight of T_M is less than the weight of T_P.
- Let e be the first edge added to T_P that is not in T_M, and let $S \subset V$ be the vertices in the partial tree at the point e is added.
- Note that e has one endpoint in S and the other in $V \setminus S$.

Correctness of Prim's Algorithm II

- As T_M is a spanning tree, there must be a path in the tree between the endpoints of e.
- On this path there must be some edge f in T_M with one endpoint in S and the other not.
- Then at the point of adding e, f must've been a condidate edge, that the algorithm didn't pick, hence the weight of f is at least the weight of e.
- If the weight of f is strictly greater than that of e, we can construct a new tree $T_{M'} = T_M - f + e$ with smaller weight that T_M, contradicting the assumption that T_M is minimum.
- If the weight of f is the same as that of e, we can construct $T_{M'}$, then repeat the argument with $T_{M'}$ in place of T_M – either we get a contradiction as before, or we progressively modify T_M to be T_P and therefore T_P must also be minimum.

Lemma

Given a connected, weighted graph G, Kruskal's algorithm produces a minimum spanning tree of G.

Proof:

- This time we use an inductive proof.
- What we will show is that if F is the set of edges chosen at any point in the algorithm, then there is some minimum spanning tree that contains F.
  1. **Base Case:**
     - $F = \emptyset$. The proposition is trivially true in this case as $\emptyset$ is a subset of any set.
  2. **Inductive Assumption:**
     - Assume the algorithm (to this point) has produced edge set F', and F' can be extended to some MST T.
  3. **Inductive Step:**

Correctness of Kruskal's Algorithm II

- If the next chosen edge e is also in T, then our proposition holds for $F + e$.
- If it is not, then $T + e$ contains a cycle, and there is some edge f that is in the cycle, but not in F.
- The weight of f must be at least the weight of e – otherwise the algorithm would choose f at this point.
- Then $T - f + e$ is a minimum spanning tree that contains $F + e$, and we're done.

- Then by induction the final set of edges can be 'extended' to an MST – as it spans the graph, this extension is 'do nothing', so the algorithm produces an MST.

Lemma

If e is the unique smallest weight edge in a connected, weighted graph G, then e is in every MST of G.

Proof: Assume for contradiction there is some MST T that does not contain e, then the graph $T + e$ contains a cycle, we can pick any edge from this cycle (other than e) and remove it to obtain a new spanning tree. As e has weight less than every other edge, this new tree must have smaller weight that T, therefore T was not an MST.

What happens if there's more than one minimum weight edge – are they all in every MST?

This can be inductively extended:

Lemma

If G is a connected, weighted graph where all edge weights are distinct, G has a unique minimum spanning tree.

And similar proofs give (assume G is connected):

Lemma

Let C be a cycle in G, and e be the largest weight edge in C. e is not in any MST of G.

Lemma

Let D be any cut in G, and e be the smallest weight edge in D. e is in every MST of G.

Let's do more coding (if time permits).