1. Overview

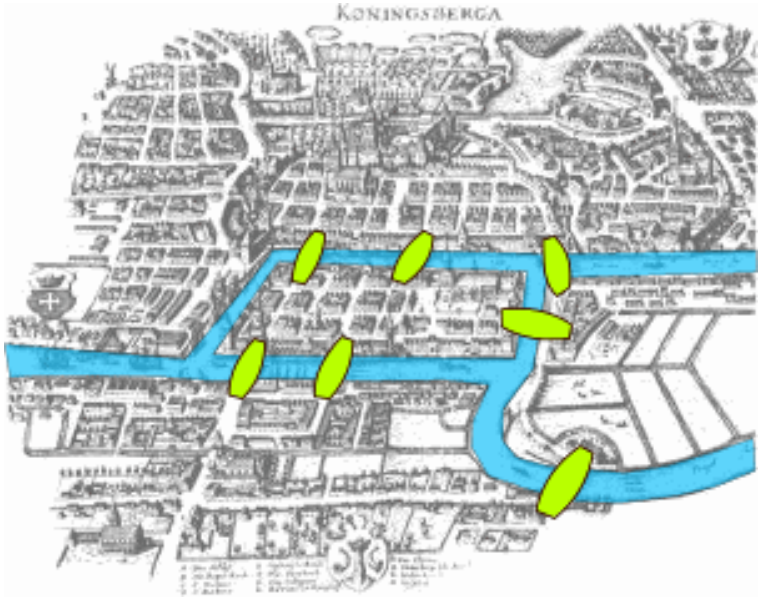
- Eulerian cycles
- Minimum spanning trees
- Matroids

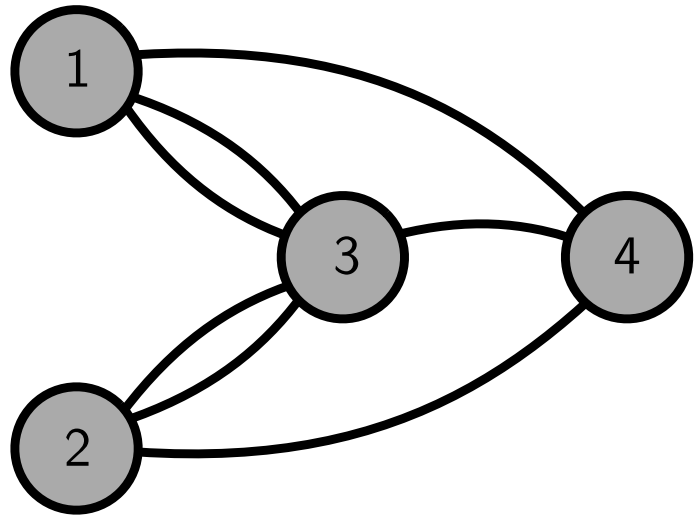
Eulerian cycle

Graph theory

Seven bridges of Königsberg, [Leonhard Euler 1735]

Can you cross all seven bridges exactly once?





Graph $G = (V, E)$

- ★ V: set of vertices $\{1, \ldots, n = |V|\}$
- ★ E: set of edges $\{(i, j), \ldots\}$ or $\{e_1, e_2, \ldots, e_{m=|E|}\}$
- ★ *directed edge* is an ordered pair such that $(i, j) \neq (j, i)$
- ★ *undirected edge* is an unordered pair such that $(i, j) = (j, i)$
- ★ *multigraph* is a graph with multiple edges between a pair of nodes
- ★ *loop* or a self-loop is an edge between a node and itself
- ★ *simple* graph is a graph with no loops and no multi-edges

Graph terminology

- ★ two nodes are said to be *adjacent* or *neighbors* if they are connected by an edge
- ★ *walk*: $v_1 v_2 \cdots v_k$ such that $(v_i, v_{i+1}) \in E$
- ★ *path*: a walk $v_1 v_2 \cdots v_k$ such that $v_i \neq v_j$
- ★ *closed walk*: a walk such that $v_1 = v_k$
- ★ *cycle* is a closed path
- ★ a graph is *connected* if there exists a path from any node i to any j
- ★ *degree* of a node is the number of adjacent nodes

For the seven bridges of Königsberg example,

- ★ Eulerian walk: a walk that includes all the edges exactly once
- ★ Eulerian cycle: an Eulerian walk that is closed (precisely, it should be called Eulerian closed walk)

given a graph, can we decide whether there is an Eulerian cycle or not?

if there is one, how can we find one efficiently?

Theorem. there exists an Eulerian cycle if and only if the graph is connected and every node has an even degree

- "$\Rightarrow$" only if part is easy: it follows from "all closed walks have even degrees"
- "$\Leftarrow$" if part: a constructive proof due to Fleury, 1883

1. algorithm:
   - ★ start at any node
   - ★ at each step, choose the next edge in the path to be the one whose deletion would not disconnect the graph
   - ★ if there is no such edge, pick the remaining edge left at current node
   - ★ move to the chosen node and delete the edge
   - ★ if there is no edge left in the end, the sequence of edges traversed is an Eulerian cycle
   - ★ if there are edges left that cannot be traversed, then the graph has no Eulerian cycle

2. correctness:

fact 1. if the original graph has all even degrees, then all node degrees remain even, except for the start node and current node

fact 2. hence, there is always an edge to move to, and the process only fails to find an Eulerian cycle only if at a certain step all edges of the current node result in disconnected graph

fact 3. this only happens when current node has degree one, in which case disconnected part is a single isolated node.

proof of fact 3. We prove by contradiction. Suppose current node has degree d larger than 2, and removing any of these edges result in a disconnected graph. Then, the graph looks like a star where removing the current node results in d disconnected subgraphs. Let $G_1 = (V_1, E_1), \ldots, G_d = (V_d, E_d)$ denote these subgraphs, with the current node removed. We know that there are even number of odd degree nodes in each of these G_i's. It follows that if we put the current node back in, then there are odd number of odd degree nodes in each V_i's. In particular, there is at least one odd degree node in each subset of nodes $V_1, \ldots, V_d$. Hence, in the graph G (which is the graph with remaining edges at current step of the algorithm), there are at least $d+1$ odd degree nodes counting the current node. This is a contradiction if $d > 2$, since we know that there are only two odd degree nodes from Fact 1.

3. complexity (running time): $|E|$ steps, but even with the best known algorithm for bridge-finding gives $T(G) = O(|E| \log^3 |E| \log\log |E|)$

   a more efficient algorithm: Hierholzer, 1873

   - ★ idea: augmenting closed walks
   - ★ complexity: $O(|E|)$

   Corollary. An undirected graph has an *Eulerian path* iff it is connected and only two nodes have odd degrees.

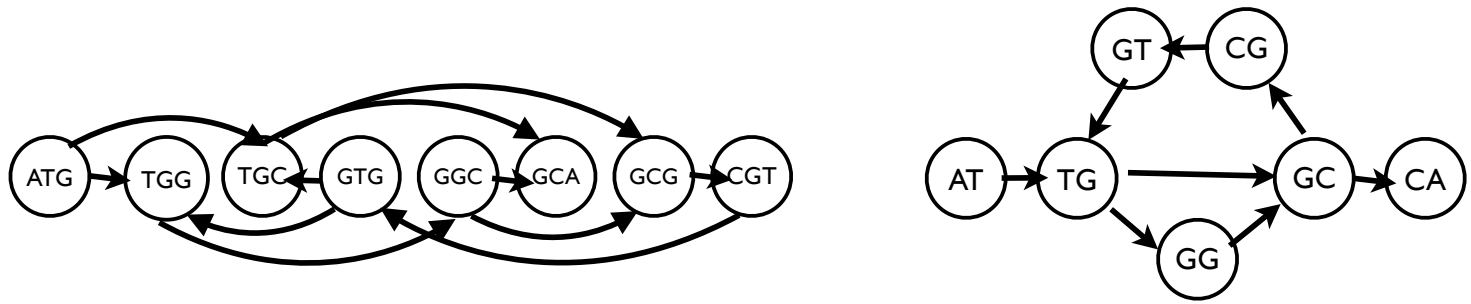
   Theorem. A *directed* graph has an Eulerian cycle off it is connected and the in-degree is equal to the out-degree for all nodes.

   Similarly, a *Hamiltonian path* is a path that passes each node exactly once
   However, the best known algorithm for finding a hamiltonian path has exponential complexity $\Omega(e^n)$ (NP-complete)

Application: DNA sequencing

DNA is a chain of complementary nucleotides. There are 4 different nucleotides: A, G, C, and T. One technique to sequence DNA, introduced by Professor Patrick Brown, is to use a collection of small subsequences of length ℓ. DNA sequence is cut into pieces, and tagged with a fluorescent agent, then exposed to an micro-array with known subsequences. One can detect the presence of particular subsequences.

Given a DNA sequence s and all detected set of length ℓ subsequences $\sigma = \{\sigma_1, \ldots, \sigma_k\}$, we want to reconstruct s from σ.
For example, $\{ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT\}$



- ★ s is a Hamiltonian path (left): node is σ_i, edge if $\ell - 1$ overlap
- ★ s is a Eulerian path (right): node is a length $\ell - 1$ subsequence, edge if σ_i

Graphs, Networks, and Algorithms

overview of the course

- learn many interesting problem on graphs and networks (motivated by real applications)
- identify which problems are solvable and which are not
- algorithms for exactly or approximately solving the problems
- analyze those algorithms: correctness and complexity

examples

- minimum spanning tree
- matching
- maximum flow
- spectral methods
- linear programming

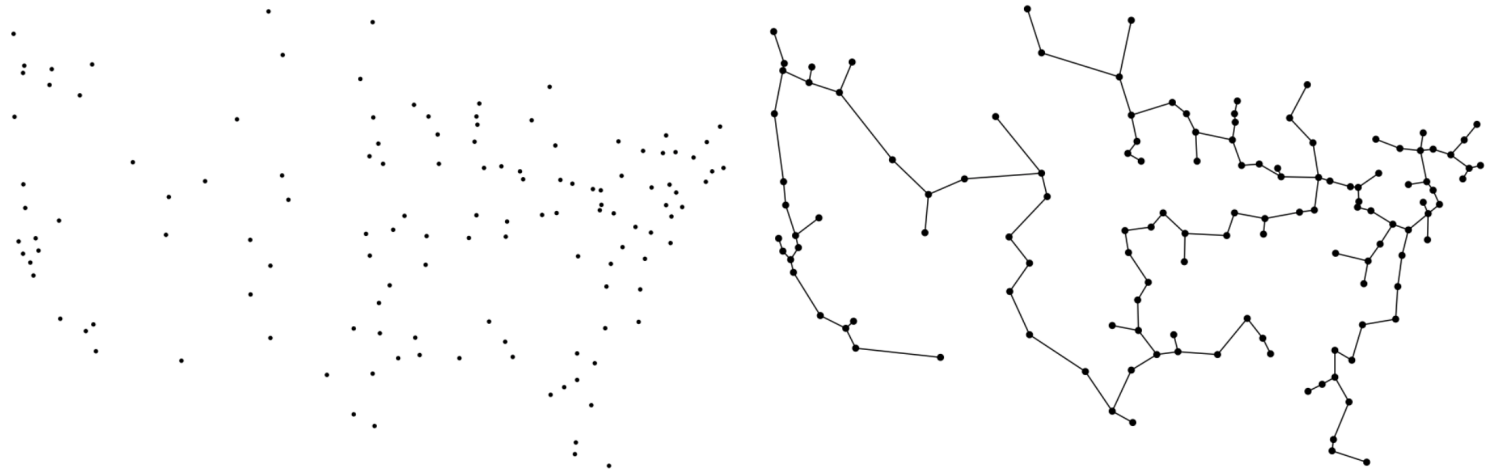
Minimum spanning trees

- tree: a connected graph with no cycle
- leaf: a node in a tree with degree one
- **claim 1** the number of edges in a tree of size n is $n-1$ (Homework 0)
- **claim 2** every finite size tree of size at least two has at least two leaves
- **proposition.** if a graph has two of the following three properties, it has all three
  1. G is connected
  2. G has no cycle
  3. $|E| = |V| - 1$

  Therefore, any graph with any two of the above properties is a tree

  proof.
  - $(1), (3) \Rightarrow (2)$: we prove by contradiction. Suppose there are cycles. Then, we can remove an edge in a cycle and the resulting graph is still connected. we remove edges until we get a connected graph with no cycle. Then by claim 1., there must be $n-1$ edges. However, this is a contradiction. We removed (some positive number of) edges starting from $n-1$ edges. So it is impossible that we end up with $n-1$ edges.
  - $(2), (3) \Rightarrow (1)$: Similarly, assume (2), (3) and suppose (1) is false.

Minimum spanning trees



- spanning tree of a connected undirected graph: a tree composed of all vertices and a subset of the edges
- minimum spanning tree of a weighted connected undirected graph: a spanning tree with minimum weight
- motivation: design of an efficient connected network (e.g. electrical grid, transportation network), phylogeny

- **definition.** *Cut*: a cut (S, S') is a partition of V such that $S \cap S' = \emptyset$ and $S \cup S' = V$. The set of edges between S and S' is also called a cut or a cut-set.
- **definition.** *Forest*: a forest is a collection of disconnected trees
- properties of a minimum spanning tree
  - in general, MST is not unique
  - **Uniqueness**: if each edge has a distinct weight, then MST is unique.
    - ★ Proof by contradiction:

      Assume there are two MSTs T_1 and T_2. Consider one edge e_1 which is included T_1 but not in T_2. Also consider another edge e_2 which is in T_2 but not in T_1, and makes a cycle when added to T_1 such that the cycle includes e_1.

      Without loss of generality, let e_1 be the one with *strictly* smaller weight than e_2. Then, by removing e_2 from T_2 and adding e_1, we can create a new spanning tree that has strictly smaller weight than T_2. This contradicts with our assumption that T_2 is a minimum spanning tree.

- **Cut property**: for any cut C in the graph, if the weight of an edge e of C is smaller than any other edges of C, then e belongs to all MST's.
  - ★ Proof by contradiction:

    Assume that there exists a cut such that the minimum weight edge (i, j) in that cut is not in a MST. Then, there is another edge (k, ℓ) in the cut which is included in the MST. If we create another spanning tree from the MST by eliminating (k, ℓ) and adding (i, j), then the weight of this new spanning tree is smaller than the original MST. This violates the assumption that the original tree was a minimum spanning tree.

- **Minimum-cost edge**: If an edge e is the edge with unique minimum cost in a graph, then e is included in all MST.
  - ★ This follows from the Cut property, since there is a cut that includes this min-cost edge.

How can we find one of MSTs in a weighted undirected graph?

Algorithms for finding a MST

- Prim's algorithm, 1957
  1. Initialize $E_T = \{\}$ and $V_T = \{i\}$ with any single node i
  2. Grow the tree by one edge: of all the edges that connect the tree to nodes not yet in the tree, find the minimum-weight edge, and transfer it to the tree.
  3. Repeat until all nodes are in the tree

  correctness of Prim's algorithm follows from the cut property
  complexity of Prim's algorithm: $O(|V|^2)$ using adjacency matrix and distance array

- Kruskal's algorithm, 1956
  1. Initialize $E_T = \{\}$ and $V_T = V$
  2. Sort the edges such that $c(e_{i_1}) \leq \cdots \leq c(e_{i_m})$
  3. While $|E_T| < |V| - 1$, add the cheapest unused edge that does not create a cycle and discard the chosen edge and those edges that create cycles from the candidate set
- correctness of Kruskal's algorithm follows from the cut property, since we are adding the minimum weight edge in a cut
- complexity of Kruskal's algorithm: $O(|E| \log |E|) = O(|E| \log |V|)$ if we use the best union-find data structure

- Application: clustering in bio-informatics

  DNA arrays measure gene expressions. One can use these gene expressions to compute a distance $d(i,j)$ between a pair of genes g_i and g_j. This distance $d(i,j)$ records how close, or similar, genes g_i and g_j are, based on their expression levels. Given n nodes (=genes) and distance between all pairs of nodes, clustering problem aims to partition the nodes into k groups, such that the nodes in the same group are closer compared with nodes in different groups. Consider a clustering $C = \{C_1, \ldots, C_k\}$, where C partitions the nodes into k groups. Then we can formulate the clustering problem as

  $$\text{maximize}_C \;\; D(C)\,,$$

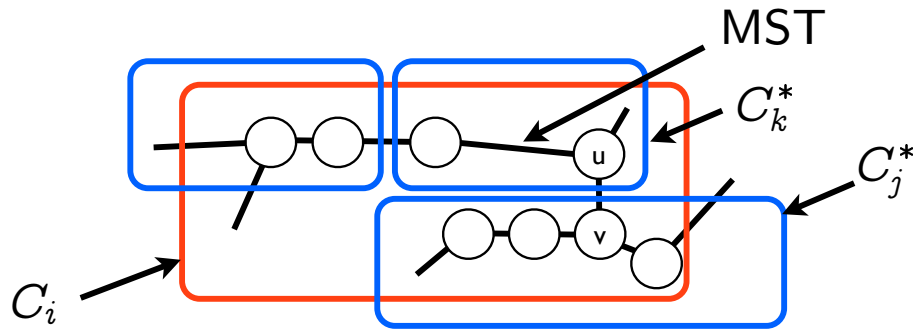
  where $D(C) \triangleq \min_{i,j} D(C_i, C_j)$ and $D(C_i, C_j) \triangleq \min_{u \in C_i, v \in C_j} d(u,v)$.

  - Consider the following MST based clustering. Given a minimum spanning tree, delete the most expensive $k-1$ edges in the MST. Then, let C denote the k partition resulting from the k connected components of the deletion.

  Theorem. The partition C of the MST based clustering is the optimal solution to maximizing $D(C)$.

proof.

We prove by contradiction. Assume there is another clustering $C^* = \{C_1^*, C_2^*, \ldots, C_k^*\}$ such that it achieves larger cost: $D(C^*) > D(C)$.



One fact is that by the cut property of a MST, for a pair of cluster C_i and C_j, the deleted edge of MST that was connecting these two sets has the minimum weight among all edges between C_i and C_j. Consider two nodes u and v, which are adjacent in the MST and in the same cluster C but different clusters in C^*. Since we deleted largest edges in MST to get C, we know

$$d(u, v) \leq \min_{a,b} D(C_a, C_b) = D(C)$$

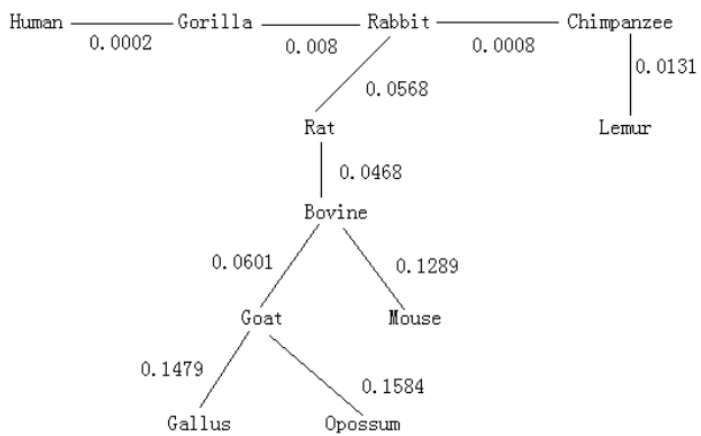
Also, by definition, $D(C^*) \leq D(C_k^*, C_j^*) \leq d(u, v)$. This contradicts the supposition that $D(C^*) > D(C)$.

Application: MST gives a heuristic for phylogeny (tree of life)

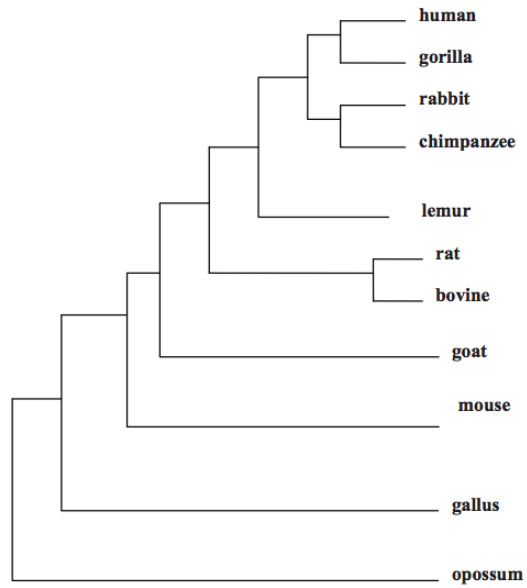
[A method for constructing phylogenic tree based on MST, Yang et al. 2011]

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rabbit	Rat	Bovine	Gorilla	Chimpanzee
Human	0	0.1254	0.4844	0.3571	0.0425	0.0594	0.0108	0.0292	0.0589	0.0002	0.0117
Goat		0	0.1479	0.1584	0.0387	0.2746	0.0789	0.1754	0.0601	0.1172	0.0830
Gallus			0	0.1601	0.2749	0.7214	0.3785	0.5838	0.3635	0.4681	0.3797
Opossum				0	0.2023	0.5831	0.2969	0.3914	0.2528	0.3450	0.2865
Lemur					0	0.1815	0.0137	0.0899	0.0537	0.0368	0.0131
Mouse						0	0.1103	0.0559	0.1289	0.0663	0.1167
Rabbit							0	0.0568	0.0575	0.0080	0.0008
Rat								0	0.0468	0.0316	0.0604
Bovine									0	0.0570	0.0641
Gorilla										0	0.0087
Chimpanzee											0

distance matrix between genes



corresponding MST



Phylogenic tree

Example: minimax path problem

[Network Flows, Ahuja, Magnanti, Orlin, page 513]

On a weighted undirected graph $G = (V, E, \{w_{ij}\})$, define the value of a path $P = p_1 p_2 \cdots p_k$ from node p_1 to node p_k as the maximum weight of an edge in P:

$$V(P) \quad = \quad \max_{i=1,\ldots,k-1} \; w_{p_i, p_{i+1}}$$

The **minimax path problem** is to find, for every pair of nodes i and j, a minimum value path from node i to j. Let T be the minimum spanning tree on G.

Theorem. The unique path between i and j in the MST T is the minimum value path between i and j.

proof. Let us focus on a particular pair of nodes i and j. Let P be the unique path between i and j in T. Let (k, ℓ) be the maximum weight edge in P. Then, removing (k, ℓ) from T creates two partitions S and S', such that $i \in S$ and $j \in S'$. This defines a cut (S, S'). For any edge in the cut (i', j') such that $i' \in S$ and $j' \in S'$, the cut property of a MST implies that

$$w_{k,\ell} \leq w_{i',j'}$$

Since any path P' between i and j must contain one of the nodes from the cut (S, S'), $w_{k,\ell}$ is the value of path P and P is the minimum value path. Hence, this establishes that the unique path in MST T is the minimum value path between all pairs of nodes.

Matroids

perhaps surprisingly, greedy algorithms (Kruskal's and Prim's) find an *optimal* solution for minimum spanning tree problem. This is due to the fact that the MST problem has a special structure.

now, we can ask, when do greedy algorithms find the optimal solution? we can answer this question for a wide range of problems by exploiting the *structure*, namely *matroids*.

- ★ introduced by the mathematician Hassler Whitney in 1935
- ★ it is a combinatorial generalization of linear independence of vectors
- ★ 'matroid' means 'something sort of like a matrix'

a **subset system** (E, I) consists of a set E and a set of subsets of E, which we call I, such that I is closed under inclusion, i.e.

$$\text{if } X \subseteq Y \text{ and } Y \in I \text{ then } X \in I$$

examples

1. $E = \{e_1, e_2, e_3\}$ and $I = \{\emptyset, \{e_1\}, \{e_2\}, \{e_3\}, \{e_1, e_2\}, \{e_2, e_3\}\}$
2. E is any set of vectors in a vector space, I is the set of sets of linearly independent vectors. (I is closed under inclusion)
3. MST problem: E is the set of the edges of an undirected graph, I is the set of all *acyclic* (having no cycles) sets of edges
4. I is the set of sets of edges such that no two edges share the same node (known as *independent set* of edges).

we are interested in finding the **maximum weight set** in a subset system. The algorithm has as input a positive weight for each element of E. We want to find a set $X \in I$ such that X has at least as much total weight as any other set in I.

MST (maximum spanning tree) example

- E is the set of the edges of an undirected graph, I is the set of all *acyclic* (having no cycles) sets of edges
- w_e is the weight of an edge e
- for an acyclic set of edges $S \in I$, $w(S) = \sum_{e \in S} w_e$
- we can turn this into a minimum spanning tree problem by taking weight $M - w_e$ for large enough M

I is called the **independent sets** of the subset system, since in general I will be defined so that it will include those sets that don't have a particular dependence.

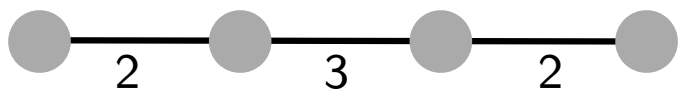
- linearly independent vectors, independent edge set, etc. satisfy the inclusion property

a **generic greedy algorithm** for a *finite* subset system (E, I) to find a set with the maximum weight is

- ★ set $X = \emptyset$
- ★ sort the elements of E in descending order of the weight
- ★ starting from the largest weight element, add to X if and only if the resulting X is in I
- ★ repeat for all elements in E
- ★ output X

examples

1. greedy algorithm finds e_2 and $\max\{e_1, e_3\}$
3. for a connected graph, this is Kruskal's algorithm
4. greedy algorithm can be stuck at a maximal set which is not maximum



fact 1. $X \in I$

fact 2. X is a *maximal set*, no element in E can be added to X without bringing X outside of I

however, we want a *maximum* set with largest total weight over all sets in I

a subset system is a **matroid** if it satisfies the *exchange property*

a subset system has *exchange property* if for all pairs of sets $X, Y \in I$ such that $|X| < |Y|$, there exists an element $i \in Y \setminus X$ such that $X \cup \{i\} \in I$

examples

1. this is a matroid
2. this is not a matroid in general
3. this is a matroid
4. this is not a matroid

theorem. for any subset system (E, I), the greedy algorithm finds the *maximum* weight set in I for (E, I) if and only if (E, I) is a matroid.

theorem. a subset system (E, I) is a matroid if and only if for any set $X \subseteq E$, any two maximal independent subsets of X have the same cardinality

- ★ this is called the *cardinality property*
- ★ Y is maximally independent in X if there is no set $Z \in I$ such that $Y \subset Z \subseteq X$

proof. first we prove the optimality of greedy algorithm for matroids

1. $\Leftarrow$ if case: We prove by contradiction.
   Let X be the set chosen by the greedy algorithm, sorted in a descending order of the weight.
   Let Y be another maximal set sorted in a descending order of the weight, and we suppose Y has a strictly larger total weight than X.
   By the exchange property, two maximal sets must have the same cardinality, since otherwise we can add an element to the set with a smaller cardinality.
   Since Y has a larger total weight, there exists an index k between 1 and n such that Y_k is strictly larger than X_k for the first time.

$$\begin{aligned} X &= \{\overbrace{x_1 \geq x_2 \geq \ldots \geq x_{k-1}}^{S} \geq x_k \geq \ldots \geq x_n\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad \wedge \\ Y &= \{\underbrace{y_1 \geq y_2 \geq \ldots \geq y_{k-1} \geq y_k}_{T} \geq \ldots \geq y_n\} \end{aligned}$$

   Let S and T be the set of $k-1$ largest weighted elements of X and Y respectively. By construction, all elements of T have larger weights than x_k. Then, again by the exchange property, there exists an element in T such that when added to S gives a new set $Z \in I$ with total weight strictly larger than $S \cup x_k$. This is a contradiction, since a greedy algorithm at k-th iteration must have produced Z instead of X.

2. $\Rightarrow$ only if case: We prove by construction.
We want to show that for a subset system (E, I) where the exchange property fails, there exists a set of weights where greedy algorithm fails to find the maximum weight set.
If the exchange property does not hold, then there exists two sets X and Y in I such that $|Y| > |X|$ and no element in $Y \setminus X$ can be added to X while still being inside I.

$$X = \underbrace{\{x_1, x_2, \ldots, x_m\}}_{m+2}$$

$$Y = \underbrace{\{y_1, y_2, y_3, \ldots, y_n\}}_{\geq m+1}$$

We choose the weights as follows

- $w(x_i) = m + 2$, where $m = |X|$
- elements in $Y \setminus X$ have weights $m + 1$
- other elements have weights zero

The greedy algorithm chooses X in the first m steps. Then, since no element in Y can be added to X, only zero weight elements are added, resulting in total weight of $m(m + 1)$.

The optimal solution is Y, whose total weight is at least $n(m + 1) \geq m^2 + 2m + 1$. Hence, there exists a set of weights such that the greedy algorithm fails to find the optimal solution.

proof. next we prove the cardinality theorem

1. $\Leftarrow$ if case: We prove by contraposition.
   We want to show that if (E, I) is not a matroid, then it fails to satisfy the cardinality property.
   Since the system is not a matroid, there exists two sets $Y \in I$ and $Z \in I$ such that $|Y| < |Z|$ and no element of $Z \setminus Y$ can be added to Y and still remain in I.
   Let $X = Y \cup Z$.Then, Y is a maximal set in X, since no other element can be added to Y.
   Z might not be maximal, but there exists a maximal set in X that include Z. Since $|Z| > |Y|$, thus created maximal set has cardinality larger than $|X|$. Hence, we conclude that the cardinality property does not hold.
2. $\Rightarrow$ only if case: We prove by contradiction.
   Suppose there are two maximally independent sets $Y \in I$ and $Z \in I$ of different cardinality such that $|Y| < |Z|$, and $Y, Z \subseteq X$. Since the system is a matroid, exchange property holds, and there exists an element in Z (and thus in X) that can be added to Y and still remain independent, that is inside I. This is a contradiction to the supposition that X is maximally independent.

Maximum (or minimum) spanning tree problem

claim 1. given a graph G the subset system (E, I) where E is the set of edges in an undirected graph, and I is the set of acyclic edges is a matroid

fact 1. a maximal set for such a system corresponds to a spanning tree (spanning forest if the graph is disconnected)

- ⋆ claim 1 follows from the fact that for any subset of E, a spanning tree has cardinality $n - k$, where n is the number of nodes in the graph and k is the number of connected components in the graph (hence satisfies the cardinality property)

Homework 1

- Problem 1. [Network Flows, Ahuja, Magnanti, Orlin, page 513]
  An intelligence service has n agents in a non friendly country. Each agent knows some of the other agents and can exchange messages with them. For each message exchanged between agents i and j, the message will fall into hostile hands with a certain probability p_{ij} that is known to us. The leader wants to pass a message to everyone while maximizing the total probability that the message is not intercepted, where the probability of the message not being intercepted is

$$\mathbb{P}(\text{message not intercepted}) \quad = \quad \prod_{(i,j)\in\mathcal{E}} (1 - p_{ij}) \,,$$

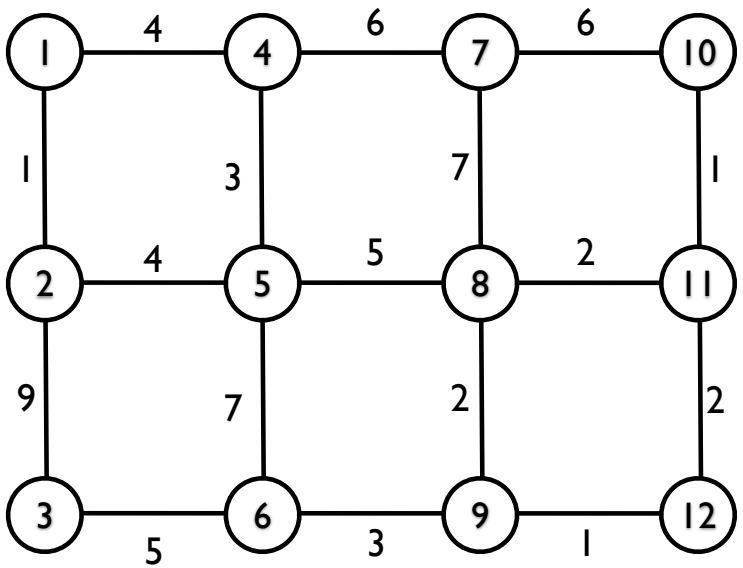
  where $\mathcal{E}$ is the set of all pairs of agents exchanging messages.

  Consider an undirected graph, where each node is an agent and an edge indicates that those two agents can exchange messages. That is two agents who are not connected by an edge cannot exchange messages.

  Explain how to find the set $\mathcal{E}$ of pairs of agents that needs to exchange messages, such that the probability of message being intercepted is minimized and every agent has the messages when no interception occurs. Specifically, formulate this problem as a maximum spanning tree problem, and explain your answer.

Homework 1

- Problem 2. [Network Flows, Ahuja, Magnanti, Orlin, Ex.13.6] Consider the following network of a highway map, and the number on the edge is the maximum elevation encountered in traversing the edge. A traveler plans to drive from node 1 to node 12 on this highway. This traveler dislikes high altitudes and so would like to find a path connecting node 1 to node 12 that minimizes the maximum altitude. Formulate this as a minimum spanning tree problem and find the best path for this traveler.



Homework 1

- Problem 3.
  - We proved in class the **cut property** of a Minimum Spanning Tree (MST). Let cut C be the collection of edges between two partition of vertices (S, S^c). Then, if an edge in C has smaller weight than any other edges in C, it belongs to all MSTs of this graph.
    - (a) In this problem we prove the **cycle property** of a MST. Show that in any cycle C in the graph, if an edge has larger weight than any of the other edges in C, then this edge cannot belong to an MST.
    - (b) In this problem, we prove a sufficient condition for uniqueness of a MST. Show that a graph has a unique minimum spanning tree if, for every cut of the graph, the edge with the smallest weight across the cut is unique. Show that the converse is not true by giving a counter-example.
    - (c) In this problem, we prove a sufficient condition for uniqueness of a Minimum Spanning Tree. Show that a graph has a unique minimum spanning tree if, for every cycle in the graph, the edge with the largest weight in the cycle is unique. Show that the converse is not true by giving a counter-example.

Homework 1

- Problem 4. Prove that at least one of G or $\overline{G}$ is connected. Here, $\overline{G}$ is a graph on the vertices of G such that two vertices are adjacent in $\overline{G}$ if and only if they are not adjacent in G.

- Problem 5. A cographic matroid (E, I) for an arbitrary undirected graph $G = (V, E)$ is defined as the set of edges E and an independent set I. A set of edges $S \subset E$ is independent (i.e. $S \in I$) if the complementary subgraph $G' = (V, E \setminus S)$ is connected.
  - (a) Show that a cographic matroid is a subset system. That is, show that I is closed under inclusion.
  - (b) Show that a cographic matroid is a matroid. That is, show that this system satisfies the exchange property.