Systems Biology Across Scales: A Personal View
IV. Networks: Paths & Cycles

Sitabhra Sinha

IMSc Chennai

Networks: directed, weighted or signed

Adjacency matrices tell us about the presence or absence of links (i.e., is A connected with B?)

If the adjacency matrix is symmetric $\Rightarrow$ undirected network,
Otherwise we have directed networks where a direction is associated with each link (A to B or B to A)
Many networks have links with heterogeneously distributed properties.
Connections in such systems can differ

- **quantitatively** by having a distribution of weights (that may for instance represent the strength of interaction) and/or
- **qualitatively** through the nature of their interactions, viz., positive (cooperative or activating) and negative (antagonistic or inhibitory)

Trekking through a network

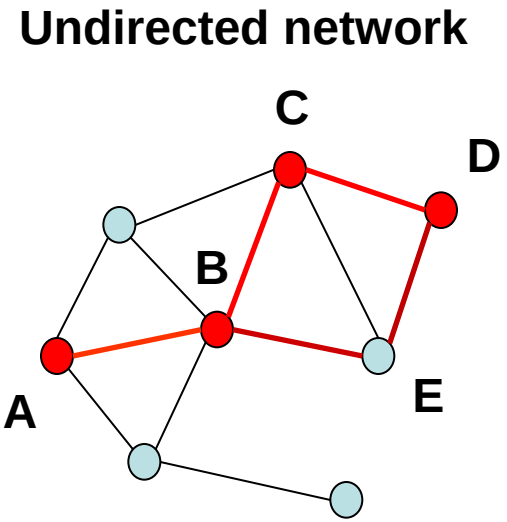
Network **path**: a sequence of nodes such that every consecutive pair is connected by a link in the network,
i.e., a route across the nodes of a network traversing existing links

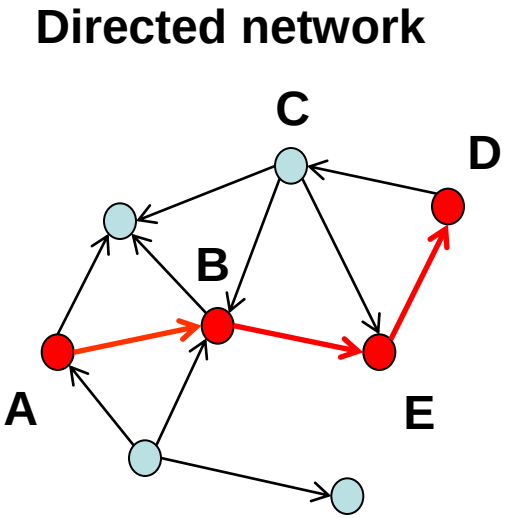
Network **path length**: number of links traversed ("hops") along a path to move from one node to another in the network

Example (Undirected network): A path from A to D having length 3 is {A,B,C,D}
It is non-unique as another path of same length is {A,B,E,D}

Example (Directed network): Unique path from A to D having length 3 is {A,B,E,D}

Typically we focus on *self-avoiding paths* that do not intersect themselves, i.e., visit a node or link more than once (e.g., geodesics and Hamilton paths)

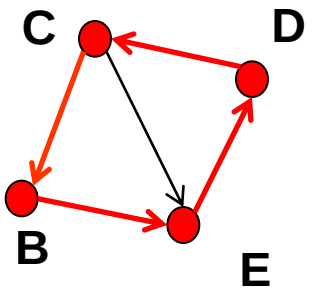




Number of paths of a given length

- For a network, element A_{ij} of the adjacency matrix $\mathbf{A}$ is 1 if there is node *i* and node *j* are connected by a link, and 0 otherwise.
- The product $A_{ik} A_{kj}$ is 1 if there is a path of length 2 from *j* to *i* via *k*, and 0 otherwise.
- The total number of paths of length two from *j* to *i*, via any other vertex, is $N_{ij}^{(2)} = \sum_k A_{ik} A_{kj} = [\, \mathbf{A}^2 \,]_{ij}$
- In general, number of paths of length *r* is $N_{ij}^{(r)} = [\, \mathbf{A}^r \,]_{ij}$
- If *i=j (starting and ending points of a path are same)*, the path is a cycle or loop $\Rightarrow$ total number of cycles of length *r* in a network is $L_r = \sum_i A_{ik} A_{kj} = [\, \mathbf{A}^r \,]_{ii} = \mathrm{Tr}\, \mathbf{A}^r$

it counts separately loops having same nodes but different starting points – i.e., {B,E,D,C,B} is considered different from {E,D,C,B,E}



- A *cycle* in a directed network has arrows on each of its links pointed in same way around the loop.

Acyclic networks

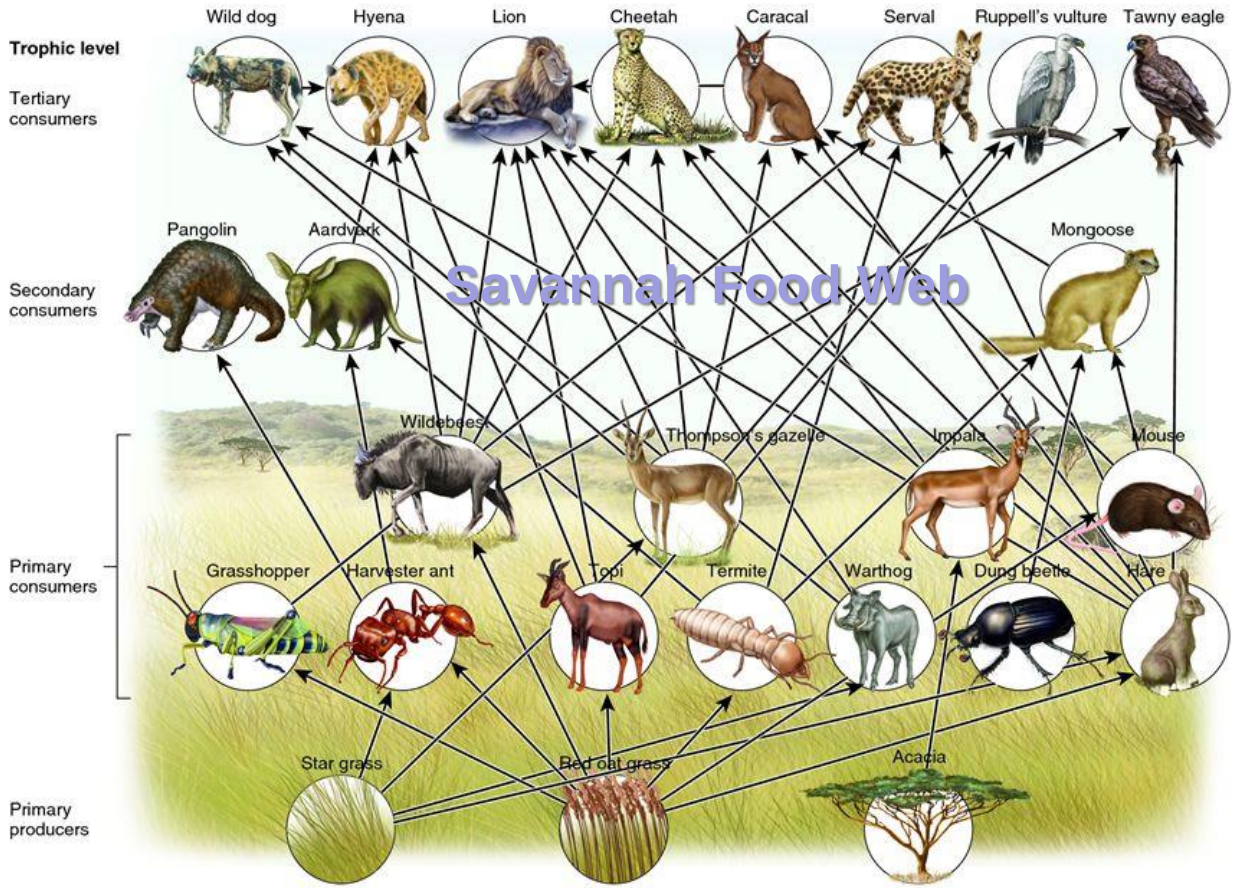
- Number of loops $L_r = \sum_i \lambda_i^r$ where λ are eigenvalues of **A**

- Thus, **acyclic networks** – i.e., networks having no cycles – will have a *nilpotent* adjacency matrix (all eigenvalues zero)

Food webs are approximately directed acyclic networks ⇒ intrinsic hierarchy among species such that, in general, those higher up in the hierarchy prey on those lower down, but not vice versa

Rank of a species in the hierarchy is called its **trophic level**

http://savannabiomeassignment.weebly.com/



Trees

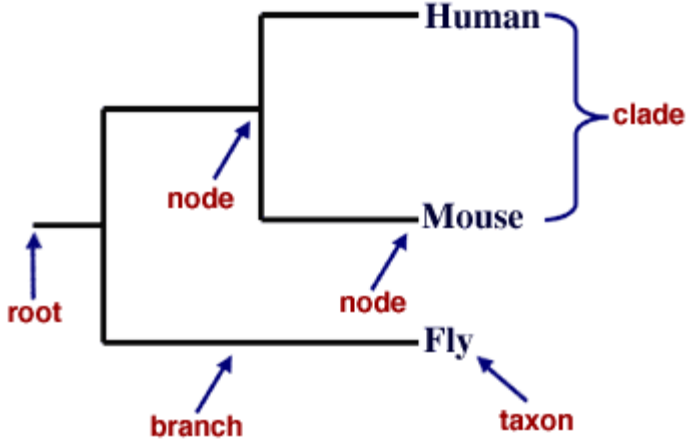
A *tree* is a connected, undirected network that contains no closed loops ("connected" ⇒ every node is reachable from every other via some path through the network)

Represented with **root node** at base, with **branches** appearing from it and terminating in **leaf nodes**
Used for **hierarchical decomposition** of a network as in **dendrogram**

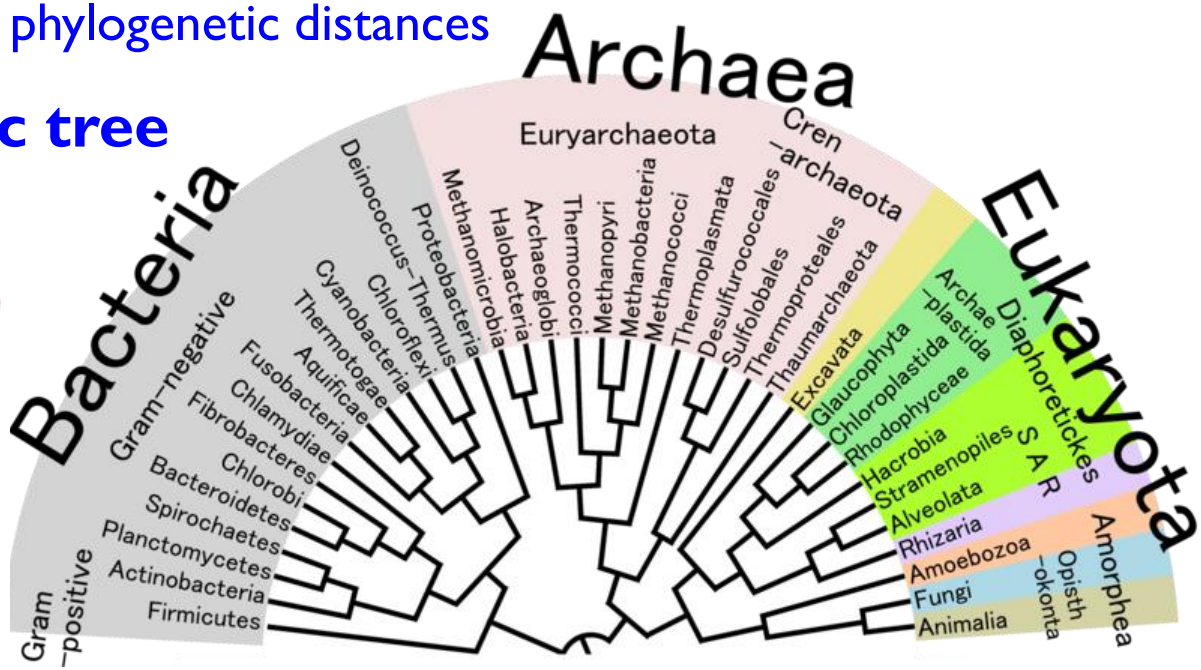
If a network contains two or more disconnected parts – all of whom are trees – it is called a *forest*

Exactly one path between any pair of nodes in a tree
Useful for defining lineage or phylogenetic distances

Example: Phylogenetic tree



http://www.talkorigins.org



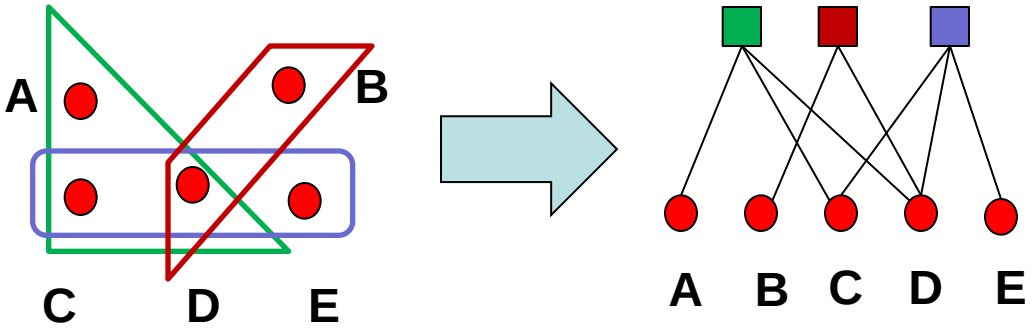
Source: wikimedia.org

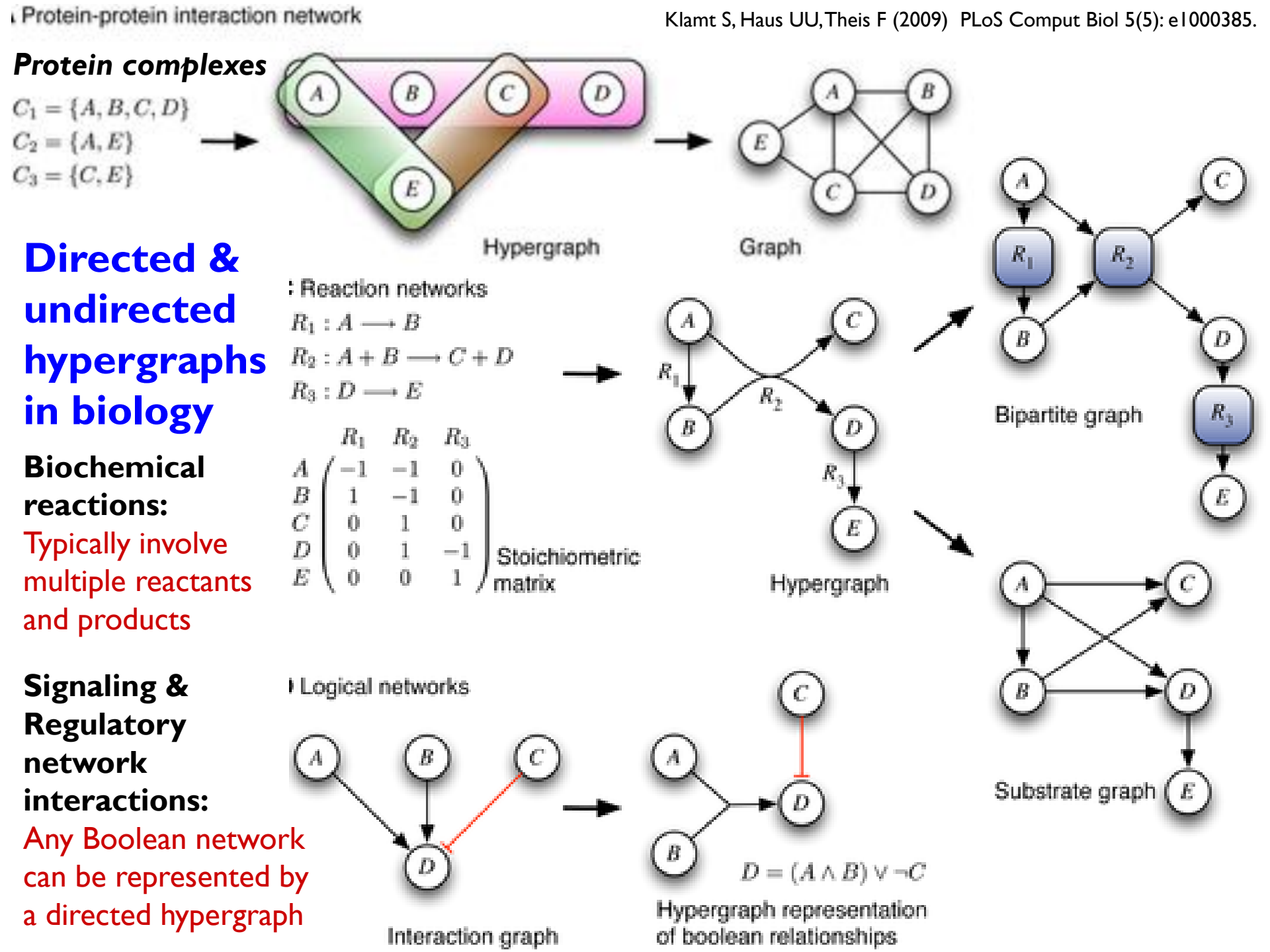
Hypergraphs

Typically, networks are defined by pairwise interactions between nodes.
However, relations may be defined in terms of multilateral rather than just bilateral relations
Many biological processes/reactions involve several components participating together in an interaction, e.g.,

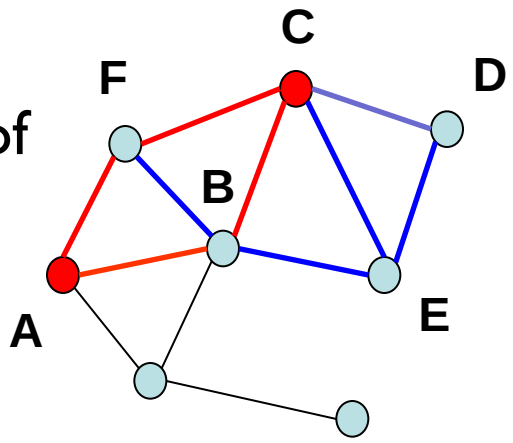
(i) substrate A is converted to product B on coming in contact with enzyme C
(ii) a protein complex that comprises more than 2 proteins

A generalized link connecting more than two nodes is a *hyperedge*
Hypergraph: A network with hyperedges
It is possible to represent a hypergraph by a **bipartite network** – a network consisting of two different types of nodes, with links occurring only between nodes of unlike type




Protein-protein interaction network
Klamt S, Haus UU, Theis F (2009) PLoS Comput Biol 5(5): e1000385.
Protein complexes
$C_1 = \{A, B, C, D\}$
$C_2 = \{A, E\}$
$C_3 = \{C, E\}$
Hypergraph
Graph
Directed & undirected hypergraphs in biology
Reaction networks
$R_1 : A \longrightarrow B$
$R_2 : A + B \longrightarrow C + D$
$R_3 : D \longrightarrow E$
Stoichiometric matrix
Bipartite graph
Hypergraph
Biochemical reactions:
Typically involve multiple reactants and products
Signaling & Regulatory network interactions:
Any Boolean network can be represented by a directed hypergraph
Logical networks
Interaction graph
$D = (A \wedge B) \vee \neg C$
Hypergraph representation of boolean relationships
Substrate graph

Geodesic or Shortest Path



A path between two vertices such that no path of a shorter length exists (necessarily self-avoiding)

Example (Undirected network): A geodesic of length 2 from A to C is {A,B,C}
It is non-unique as another path of same length is {A,F,C}

The length of a geodesic, called *geodesic distance* or *shortest path length*, is the shortest network distance between the nodes at the ends of the path

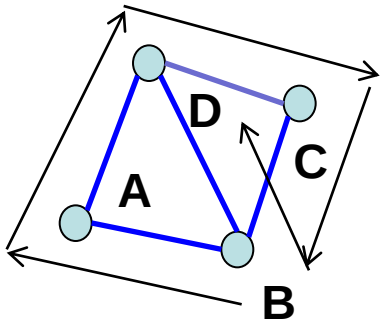
Defn. the smallest value of r such that $[\ \mathbf{A}^r\ ]_{ij} > 0$

If a network has disconnected components, there may be no geodesic between members of one component and those of another $\Rightarrow$ infinite geodesic distance

Diameter of a network: length of the longest geodesic between any pair of nodes in the network for which a path actually exists.

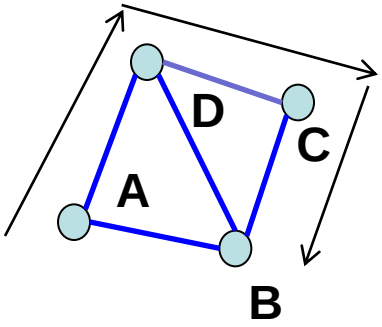
Eulerian path and Hamiltonian Path

An *Eulerian path:* path that traverses each link in a network exactly once.



A *Hamiltonian path*: a path that visits each node in a network exactly once. (by defn, self-avoiding)

A network can have one or many Eulerian or Hamiltonian paths, or none.

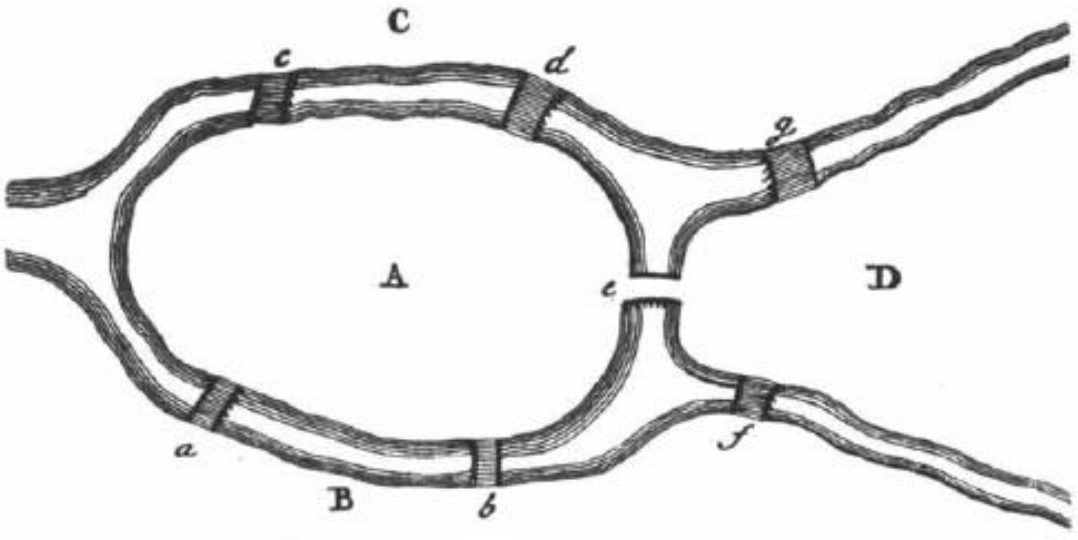


An Eulerian path need not be self-avoiding

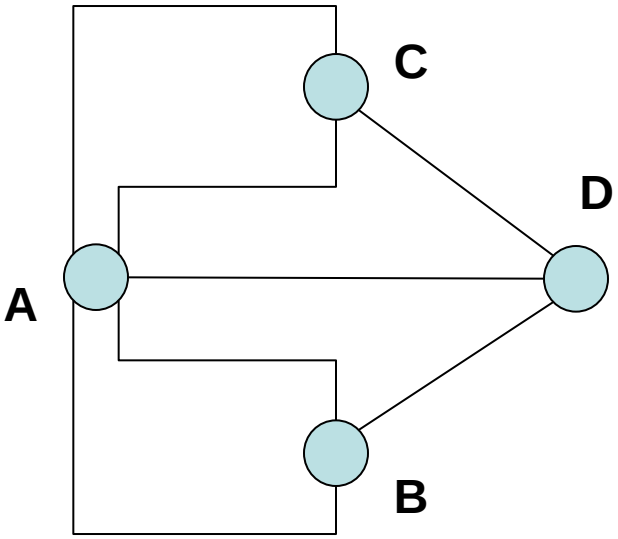
If there are any nodes of degree (number of links) > 2 in a network, an Eulerian path will have to visit those vertices more than once in order to traverse all their links.

Eulerian paths form the basis for solution of the Konigsberg Bridge problem by Euler

Konigsberg Bridge problem



Seven bridges of Koenigsberg crossed the River Pregel



"Does there exist any walking route that crosses all seven bridges exactly once each?" ≡ a problem of finding an Eulerian path on the equivalent network

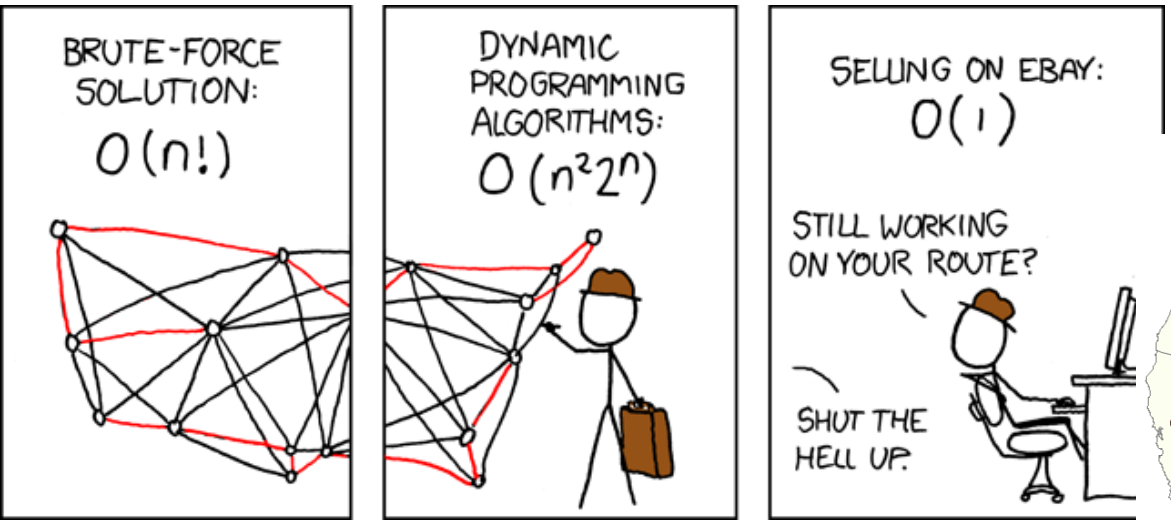
As any Eulerian path must enter as well as exit each node it passes through except the first and last, there can at most be two nodes in the network with odd degree if such a path is to exist. Since all four nodes in the Konigsberg network have odd degree, the bridge problem necessarily has **no solution**.

Traveling Salesman Problem

(or why they created Ebay)

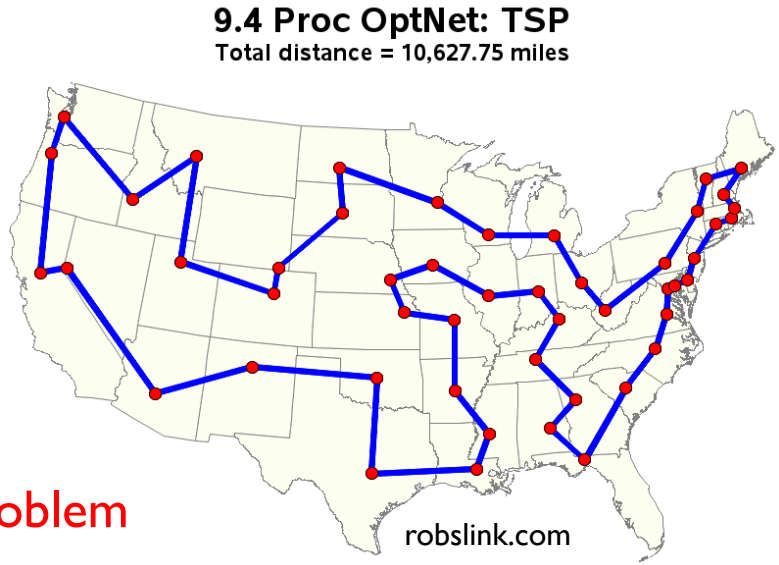
A salesman needs tour a number of cities and get back home. Given the cities (and their locations), the problem is to find the shortest possible route that she can follow to visit each city exactly once and return to the origin city.

Given a complete undirected network *G* that has a nonnegative integer cost (weight) associated with each link, find a Hamilton cycle of *G* with minimum cost.



Source: xkcd.com/399/

Formulated mathematically in 1930s by Merrill Flood trying to solve a school bus routing problem



One of the most studied problems in optimization

A computationally extremely difficult (NP-hard) problem

The Shortest Superstring Problem

can be mapped to Traveling Salesman Problem

see
Finn Rosenbech Jensen,
Masters Thesis,
Aarhus University, 2010

Given a collection of strings $S = (s_1, s_2, \ldots s_n)$ composed of elements from a finite alphabet, determine the shortest string containing each string in S as a (consecutive) substring, i.e., find the shortest superstring for the collection S.

Application in Shotgun Sequencing

When sequencing a large genome, where it is difficult to sequence strands of length > 100-1000 bps, longer strands are divided into smaller pieces and subsequently re-assembled.

In Shotgun Sequencing, first many copies are made of the DNA to be sequenced. These copies are then randomly divided into smaller pieces by physical, enzymatic or chemical means.

The difficulty in assembling the original sequence is because although the smaller pieces constitute overlapping sequences of the original DNA, information about succession of the overlapping sequences is lost when the strands are cut.

The SSP is a theoretical model for reassembling the overlapping pieces into a superstring which is assumed to be a good approximation of the original DNA sequence.