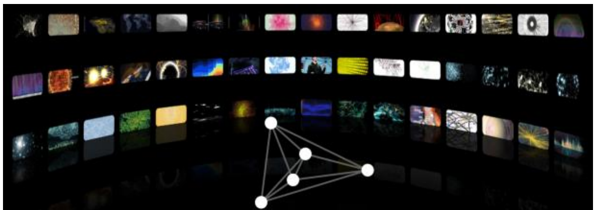


Network Science

Class 2: Graph theory and basic terminology
Learning the language

Jaap Kaandorp (credits Mike Lees)
Slides adapted from: Prof. Albert-László Barabási

Dr. Baruch Barzel, Dr. Mauro Martino
Network Science: Graph Theory 2012

1

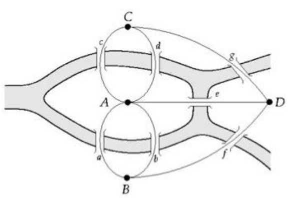
THE BRIDGES OF KONIGSBERG

Can one walk across the seven bridges and never cross the same bridge twice?

Network Science: Graph Theory 2012

2

THE BRIDGES OF KONIGSBERG



Can one walk across the seven bridges and never cross the same bridge twice?

1735: **Leonhard Euler's theorem:**

(a) If a graph has nodes of odd degree, there is no path.

(b) If a graph is connected and has no odd degree nodes, it has at least one path.

Network Science: Graph Theory 2012

3

THE 7 BRIDGES OF KONIGSBERG

- Can one walk across all seven bridges and never cross the same one twice? Despite many attempts, no one could find such path. The problem remained unsolved until 1736, when Leonard Euler, a Swiss born mathematician, offered a rigorous mathematical proof that such path does not exist.
- Euler's insight was to represent each of the four land areas separated by the river as nodes, distinguishing them with letters A, B, C, and D. Next he connected with lines each piece of land that had a bridge between them. He thus built a {graph}, whose {nodes} were the pieces of land and {links} were the bridges.

Network Science: Graphs 4

4

THE 7 BRIDGES OF KONIGSBERG

- Then Euler made a simple observation: if there is a path crossing all bridges, but never the same bridge twice, then nodes with odd number of links must be either the starting or the end point of this path. Indeed, with an odd number of links you can arrive to a node and have no unused link for you to leave it. Yet, a continuous path that goes through all bridges can have only one starting and one end point. Thus such a path cannot exist on a graph that has more than two nodes with an odd number of links. As the K\"{o}nigsberg graph had three such nodes, B, C, D, each with three links, no path could satisfy the problem.

Network Science: Graphs 5

5

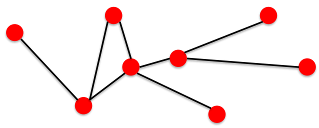
THE 7 BRIDGES OF KONIGSBERG

- Today we remember Euler's proof because it was the first time someone solved a mathematical problem by turning it into a graph. In hindsight the proof has two important messages: The first is that some problems become simpler and more treatable if they are represented as a graph. The second is that the existence of the path does not depend on our ingenuity to find it. Rather, it is a property of the graph. Indeed, given the layout of the K\"{o}nigsberg bridges, no matter how smart we are, we will never find the desired path

Network Science: Graphs 6

6

COMPONENTS OF A COMPLEX SYSTEM



- **components**: nodes, vertices N
- **interactions**: links, edges L
- **system**: network, graph (N,L)

7

NETWORKS OR GRAPHS?

network often refers to real systems
- www,
- social network
- metabolic network.

Language: (Network, node, link)

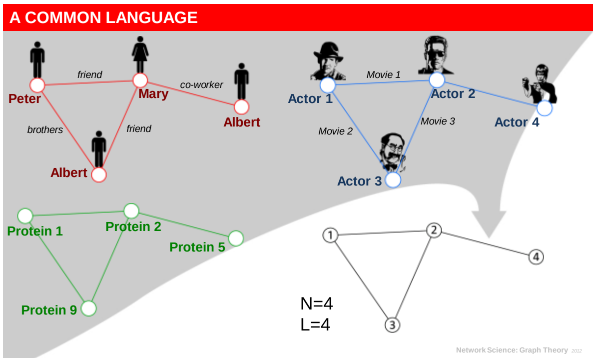
graph: mathematical representation of a network
- web graph,
- social graph (a Facebook term)

Language: (Graph, vertex, edge)

We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably.

8

A COMMON LANGUAGE



9

CHOOSING A PROPER REPRESENTATION

The choice of the proper network representation determines our ability to use network theory successfully.

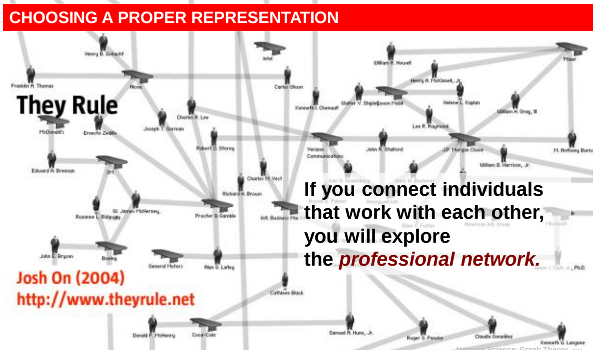
In some cases there is a unique, unambiguous representation. In other cases, the representation is by no means unique.

For example,, the way we assign the links between a group of individuals will determine the nature of the question we can study.\

THIS IS MODELLING!

10

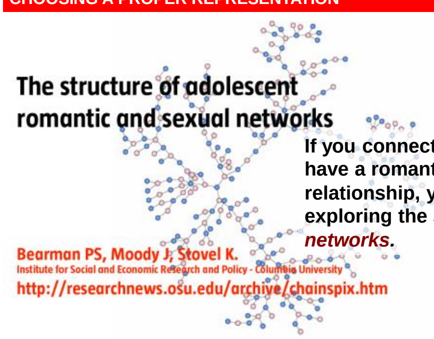
CHOOSING A PROPER REPRESENTATION



If you connect individuals that work with each other, you will explore the *professional network.*

11

CHOOSING A PROPER REPRESENTATION



If you connect those that have a romantic and sexual relationship, you will be exploring the *sexual networks*.

12

CHOOSING A PROPER REPRESENTATION

If you connect individuals based on their first name (*all Peters connected to each other*), you will be exploring what?

It is a network, nevertheless.

13

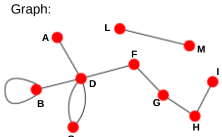
14

UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:

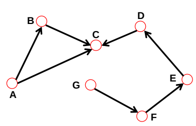


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



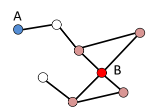
An undirected link is the superposition of two opposite directed links.

Directed links :
URLs on the www
phone calls
metabolic reactions

15

NODE DEGREES

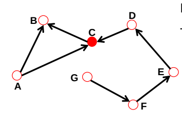
Undirected



Node degree: the number of links connected to the node.

$k_A = 1 \qquad k_B = 4$

Directed



In *directed networks* we can define an in-degree and out-degree.

The (total) degree is the sum of in- and out-degree.

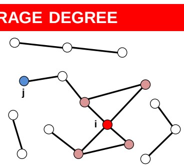
$k_C^{in} = 2 \quad k_C^{out} = 1 \quad k_C = 3$

Source: a node with $k^{in} = 0$; Sink: a node with $k^{out} = 0$.

16

AVERAGE DEGREE

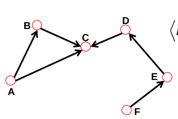
Undirected



$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N – the number of nodes in the graph

Directed



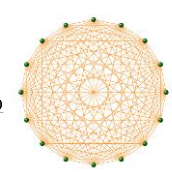
$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

$$\langle k \rangle \equiv \frac{L}{N}$$

18

COMPLETE GRAPH

The maximum number of links a network of N nodes can have is: $L_{max} = \binom{N}{2} = \frac{N(N-1)}{2}$



A graph with degree $L=L_{max}$ is called a complete graph, and its average degree is **<k>=N-1**

19

REAL NETWORKS ARE SPARSE

Most networks observed in real systems are sparse:

$L \ll L_{max}$
or
$\langle k\rangle \ll N-1$.

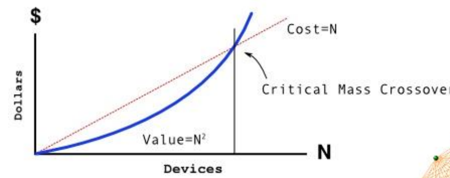
WWW (ND Sample):	N=325,729;	$L=1.4\ 10^6$	$L_{max}=10^{12}$	$\langle k\rangle=4.51$
Protein (*S. Cerevisiae*):	N= 1,870;	L=4,470	$L_{max}=10^7$	$\langle k\rangle=2.39$
Coauthorship (Math):	N= 70,975;	$L=2\ 10^5$	$L_{max}=3\ 10^{10}$	$\langle k\rangle=3.9$
Movie Actors:	N=212,250;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	$\langle k\rangle=28.78$

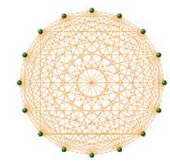
(Source: Albert, Barabasi, RMP2002)

Network Science: Graph Theory 2012

20

METCALFE'S LAW



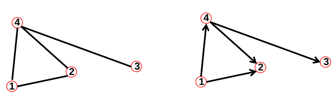


The maximum number of links a network of N nodes can have is: $L_{max} = \binom{N}{2} = \frac{N(N-1)}{2}$

Network Science: Graph Theory 2012

21

ADJACENCY MATRIX



$A_{ij}=1$ if there is a link between node *i* and *j*

$A_{ij}=0$ if nodes *i* and *j* are not connected to each other.

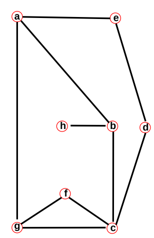
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Note that for a directed graph (right) the matrix is not symmetric.

Network Science: Graph Theory 2012

22

ADJACENCY MATRIX



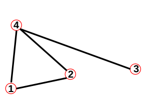
	a	b	c	d	e	f	g	h
a	0	1	0	0	1	0	1	0
b	1	0	1	0	0	0	0	1
c	0	1	0	1	0	1	1	0
d	0	0	1	0	1	0	0	0
e	1	0	0	1	0	0	0	0
f	0	0	1	0	0	0	1	0
g	1	0	1	0	0	1	0	0
h	0	1	0	0	0	0	0	0

Network Science: Graph Theory 2012

23

ADJACENCY MATRIX AND NODE DEGREES

Undirected



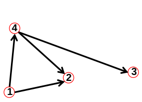
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$k_i = \sum_{j=1}^{N} A_{ij} \qquad k_j = \sum_{i=1}^{N} A_{ij}$$

$$A_{ij} = A_{ji} \qquad A_{ii} = 0$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij}^{N} A_{ij}$$

Directed



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

$$k_i^{out} = \sum_{j=1}^{N} A_{ij} \qquad k_j^{in} = \sum_{i=1}^{N} A_{ij}$$

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j}^{N} A_{ij}$$

$$A_{ij} \neq A_{ji} \qquad A_{ii} = 0$$

Network Science: Graph Theory 2012

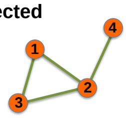
24

Network Science: Graph Theory 2012

25

GRAPHOLOGY 1

Undirected



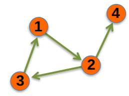
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0 \qquad A_{ij} = A_{ji}$

$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \langle k \rangle = \frac{2L}{N}$

Actor network, protein-protein interactions

Directed



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$

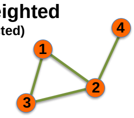
$L = \sum_{i,j=1}^{N} A_{ij} \qquad \langle k \rangle = \frac{L}{N}$

WWW, citation networks

26

GRAPHOLOGY 2

Unweighted
(undirected)



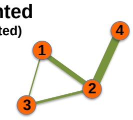
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0 \qquad A_{ij} = A_{ji}$

$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \langle k \rangle = \frac{2L}{N}$

protein-protein interactions, www

Weighted
(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0 \qquad A_{ij} = A_{ji}$

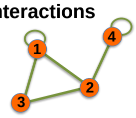
$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \langle k \rangle = \frac{2L}{N}$

Call Graph, metabolic networks

27

GRAPHOLOGY 3

Self-interactions



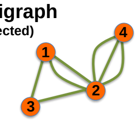
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$

$L = \frac{1}{2}\sum_{i,j=1,i\neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii} \qquad ?$

Protein interaction network, www

Multigraph
(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$A_{ii} = 0 \qquad A_{ij} = A_{ji}$

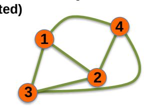
$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad \langle k \rangle = \frac{2L}{N}$

Social networks, collaboration networks

28

GRAPHOLOGY 4

Complete Graph
(undirected)

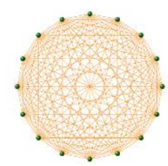


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$A_{ii} = 0 \qquad A_{i\neq j} = 1$

$L = L_{max} = \frac{N(N-1)}{2} \qquad \langle k \rangle = N - 1$

Actor network, protein-protein interactions



29

GRAPHOLOGY: Real networks can have multiple characteristics

WWW > directed multigraph with self-interactions

Protein Interactions > undirected unweighted with self-interactions

Collaboration network > undirected multigraph or weighted.

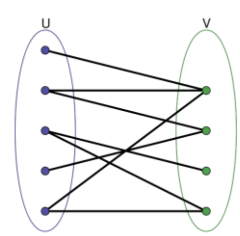
Mobile phone calls > directed, weighted.

Facebook Friendship links > undirected, unweighted.

30

BIPARTITE GRAPHS

bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.

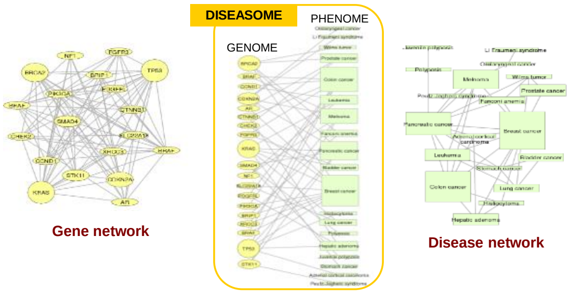


Examples:

Hollywood actor network
Collaboration networks
Disease network (diseasome)

31

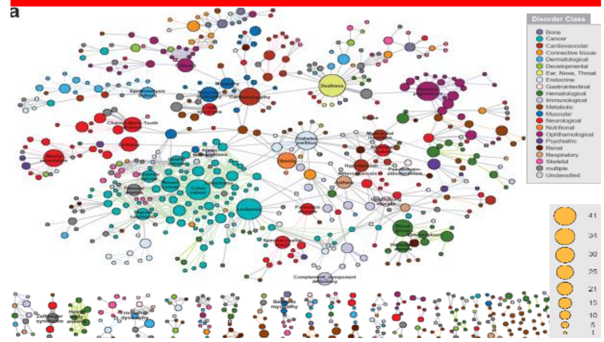
GENE NETWORK – DISEASE NETWORK



Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

32

HUMAN DISEASE NETWORK



33

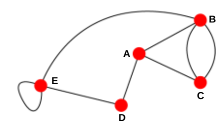
34

PATHS

A *path is* a sequence of nodes in which each node is adjacent to the next one

$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

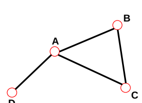
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0,i_1),(i_1,i_2),(i_2,i_3),\ldots,(i_{n-1},i_n)\}$$



- A path can intersect itself and pass through the same link repeatedly. Each time a link is crossed, it is counted separately
- A legitimate path on the graph on the right: **ABCBCADEEBA**
- In a directed network, the path can follow only the direction of an arrow.

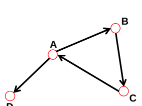
35

DISTANCE IN A GRAPH — Shortest Path, Geodesic Path



The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them.

*If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arrows.

Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

36

NUMBER OF PATHS BETWEEN TWO NODES — Adjacency Matrix

N_{ij}, number of paths between any two nodes *i* and *j*:

Length n=1: If there is a link between *i* and *j*, then A_{ij}=1 and A_{ij}=0 otherwise.

Length n=2: If there is a path of length two between *i* and *j*, then $A_{ik}A_{kj}$=1, and $A_{ik}A_{kj}$=0 otherwise.
The number of paths of length 2:

$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik}A_{kj} = [A^2]_{ij}$$

Length n: In general, if there is a path of length *n* between *i* and *j*, then $A_{ik} \ldots A_{lj}$=1 and $A_{ik} \ldots A_{lj}$=0 otherwise.
The number of paths of length *n* between *i* and *j* is*

$$N_{ij}^{(n)} = [A^n]_{ij}$$

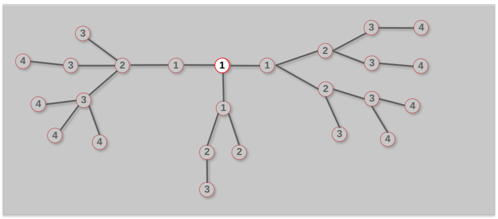
**holds for both directed and undirected networks.*

37

FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 1 and node 4:

1.Start at 1.

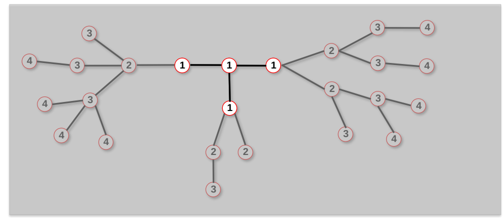


38

FINDING DISTANCES: BREADTH FIRST SEATCH

Distance between node 1 and node 4:

1.Start at 1.
2.Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.

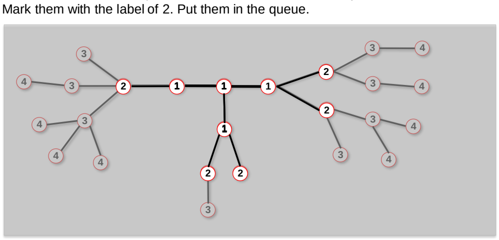


39

FINDING DISTANCES: BREADTH FIRST SEATCH

Distance between node 1 and node 4:

1.Start at 1.
2.Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.
3.Take the first node out of the queue. Find the unmarked nodes adjacent to it in the graph. Mark them with the label of 2. Put them in the queue.

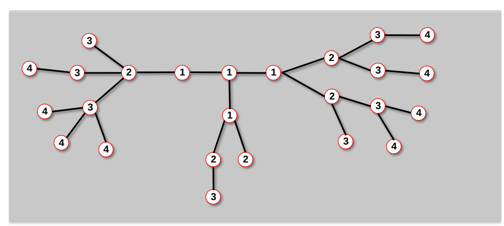


40

FINDING DISTANCES: BREADTH FIRST SEATCH

Distance between node 1 and node 4:

1.Repeat until you find node 4 or there are no more nodes in the queue.
2.The distance between 1 and 4 is the label of 4 or, if 4 does not have a label, infinity.



41

NETWORK DIAMETER AND AVERAGE DISTANCE

Diameter: d_{max} the maximum distance between any pair of nodes in the graph.

Average path length/distance, <d>, for a connected graph:

where d_{ij} is the distance from node *i* to node j

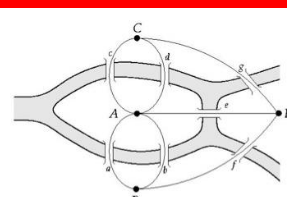
$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

In an *undirected graph* $d_{ij} = d_{ji}$, *so* we only need to count them once:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j>i} d_{ij}$$

42

THE BRIDGES OF KONIGSBERG

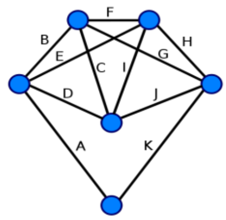


Can one walk across the seven bridges and never cross the same bridge twice?

Euler PATH or CIRCUIT: return to the starting point by traveling each link of the graph once and only once.

43

EULERIAN GRAPH: it has an Eulerian path

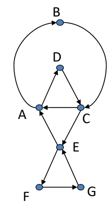


Every vertex of this graph has an even degree, therefore this is an Eulerian graph.
Following the edges in alphabetical order gives an Eulerian circuit/cycle.

http://en.wikipedia.org/wiki/Euler_circuit

44

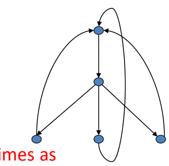
EULER CIRCUITS IN DIRECTED GRAPHS



If a digraph is strongly connected and the in-degree of each node is equal to its out-degree, then there is an Euler circuit

Otherwise there is no Euler circuit.

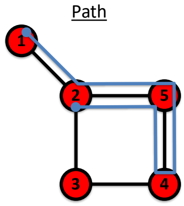
n a circuit we need to enter each node as many times as we leave it.



45

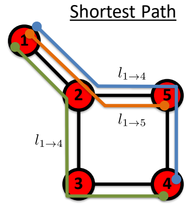
PATHOLOGY: summary

Path



A sequence of nodes such that each node is connected to the next node along the path by a link.

Shortest Path



$l_{1\to4} = 3$

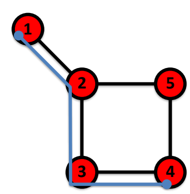
$l_{1\to5} = 2$

The path with the shortest length between two nodes (distance).

46

PATHOLOGY: summary

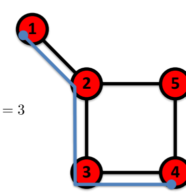
Diameter



$l_{1\to4} = 3$

The longest shortest path in a graph

Average Path Length



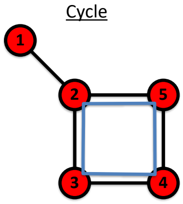
$(l_{1\to2} + l_{1\to3} + l_{1\to4} + l_{1\to5} + l_{2\to3} + l_{2\to4} + l_{2\to5} + l_{3\to4} + l_{3\to5} + l_{4\to5})/10 = 1.6$

The average of the shortest paths for all pairs of nodes.

47

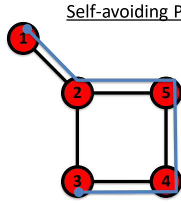
PATHOLOGY: summary

Cycle



A path with the same start and end node.

Self-avoiding Path

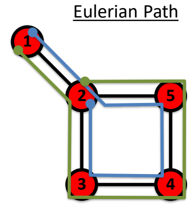


A path that does not intersect itself.

48

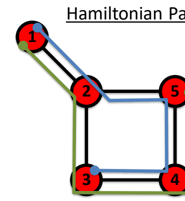
PATHOLOGY: summary

Eulerian Path



A path that traverses each link exactly once.

Hamiltonian Path

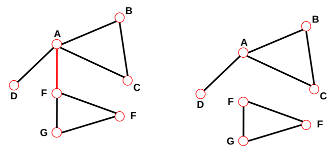


A path that visits each node exactly once.

49

CONNECTIVITY OF UNDIRECTED GRAPHS

Connected (undirected) graph: any two vertices can be joined by a path.
A disconnected graph is made up by two or more connected components.



Largest Component:
Giant Component

The rest: **Isolates**

Bridge: if we erase it, the graph becomes disconnected.

CONNECTIVITY OF UNDIRECTED GRAPHS Adjacency Matrix

The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

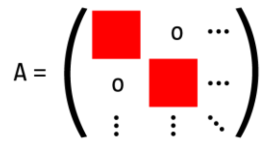
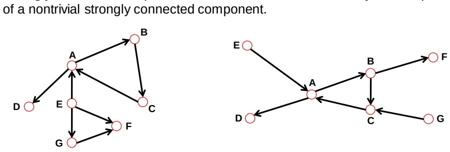


Figure after Newman, 2010

CONNECTIVITY OF DIRECTED GRAPHS

Strongly connected directed graph: has a path from each node to every other node and vice versa (e.g. AB path and BA path).
Weakly connected directed graph: it is connected if we disregard the edge directions.

Strongly connected components can be identified, but not every node is part of a nontrivial strongly connected component.



In-component: nodes that can reach the scc,
Out-component: nodes that can be reached from the scc.

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE

Degree distribution	p_k
Average path length	<d>
Clustering coefficient	C

DEGREE DISTRIBUTION

STATISTICS REMINDER

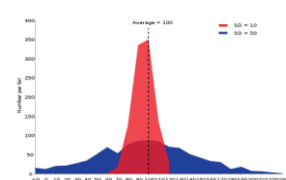
We have a sample of values $x_1, ..., x_N$

Distribution of **x** (a.k.a. PDF): probability that a randomly chosen value is **x**

P(x) = (# values x) / N

$\Sigma_i P(x_i) = 1$ always!

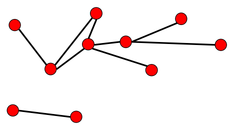
Histograms >>>

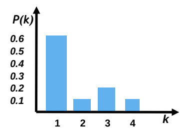


DEGREE DISTRIBUTION

Degree distribution P(k): probability that a randomly chosen vertex has degree k

N_k = # nodes with degree k
P(k) = N_k / N ➒ plot





56

DEGREE DISTRIBUTION

discrete representation: p_k is the probability that a node has degree **k**.

continuum description: **p(k)** is the pdf of the degrees, where

$$\int_{k_1}^{k_2} p(k)dk$$

represents the probability that a node's degree is between k_1 and k_2.

Normalization condition:

$$\sum_{0}^{\infty} p_k = 1 \qquad \int_{K_{min}}^{\infty} p(k)dk = 1$$

where K_{min} is the minimal degree in the network.

57

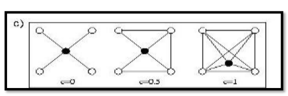
CLUSTERING COEFFICIENT

- **Clustering coefficient:**

  what portion of your neighbors are connected?

  - Node i with degree k_i
  - C_i in [0,1]
  - e_i Number of edges between neighbors

$$C_i = \frac{2e_i}{k_i(k_i-1)}$$



58

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE

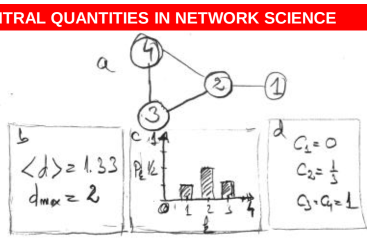
Degree distribution: **P(k)**

Path length: ***<d>***

Clustering coefficient: $C_i = \frac{2e_i}{k_i(k_i-1)}$

59

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE



A. Degree distribution: p_k
B. Path length: ***<d>***
C. Clustering coefficient: $C_i = \frac{2e_i}{k_i(k_i-1)}$

60

63