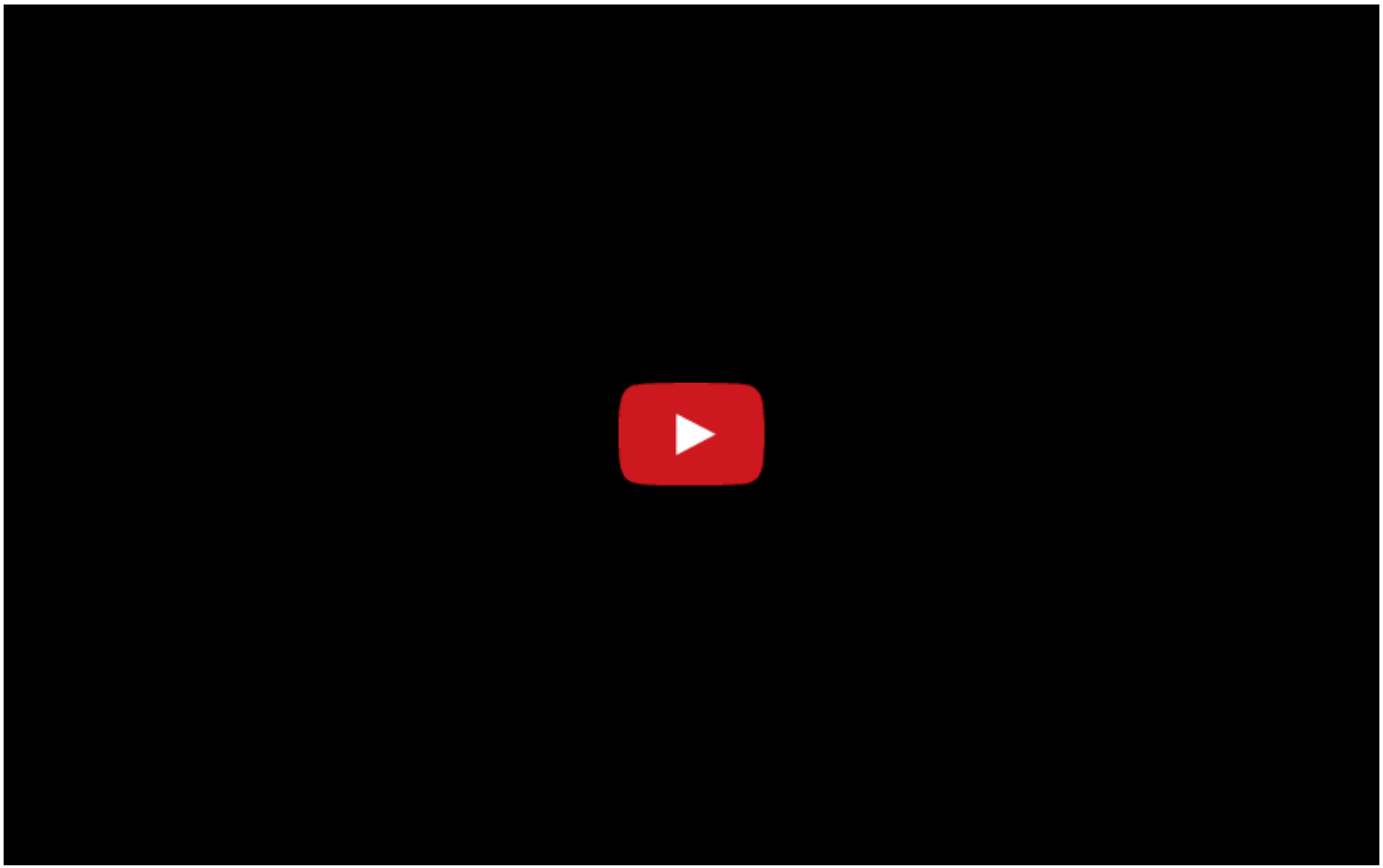
network of data

info-20002: foundations of informatics

network: of connections, people, and places



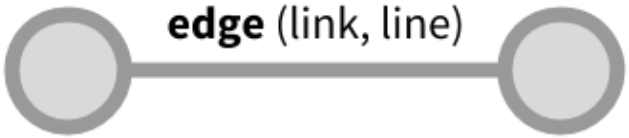
The Guy Who Brings Your Lunch in Mumbai Now Also Brings You Anything From a New Phone to a Fresh Shirte
Dabbawalas - Inforgraphics

about a network

- Who are the most important members of a network?
- Which members are located in the periphery?
- Which members have more possibility to connect to others?
- Which members are the most crucial to maintain the connectedness in a group?
- Can the network broken down into smaller sub-networks?

networks as graphs

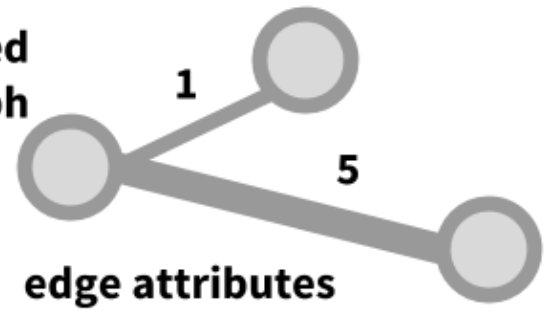
- A collection of nodes (vertices) and edges (lines, links, ties) connecting them.
- In social network, a node is a person and an edge is a relationship between two persons
- *Directed graph*: edges own directionality between two nodes (vs undirected).
- *Weighted graph*: edges have values that indicate a property of the connection.
- *Acyclic*: no loop/cycle
- *Degree* of a node: the number of edges of a node (indegree/outdegree)



node (vertex)
e.g.
place, person,
tweet, web pages

node attributes
e.g.
population of a place,
gender of a person,
timestamp of a tweet

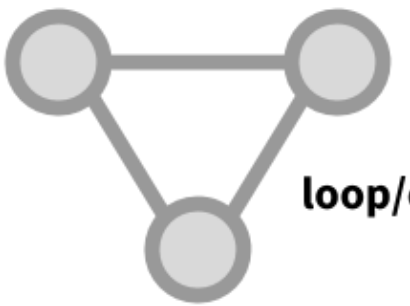
weighted graph



edge attributes
e.g.
lanes of roads,
strength of ties,
no of common friends

directed graph
e.g.
hyperlinks, roads

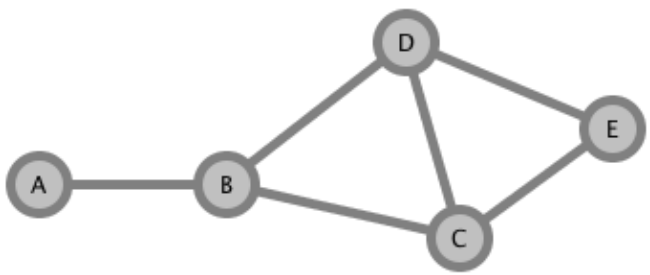
undirected graph
e.g.
Facebook friendship



loop/cycle

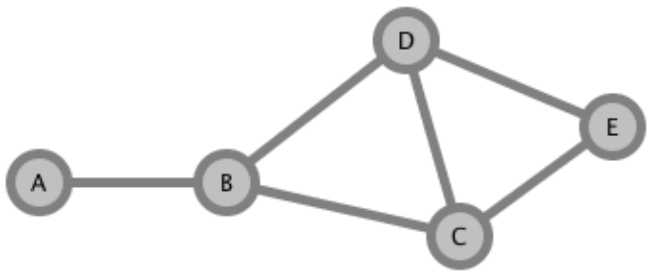
networks as graphs

- *Path*: a sequence of nodes that can be traversed by traveling through edges from a starting node to an ending node.
- *Geodesic path*: the shortest path between two nodes, path with minimum number of edges.



- ABCE, ABDE are geodesic paths between A and E
- ABCDE is a path between A and E but geodesic one.

graph as adjacency matrix



	A	**B**	**C**	**D**	**E**
A	0	1	0	0	0
B	1	1	1	1	0
C	0	1	0	1	1
D	0	1	1	0	1
E	0	0	1	1	0

Graph formal definition

- Graph $G = (V, E)$, where V is a set of nodes and E is a set of edges.
- An edge is defined as v, w, where $v, w \in V$
- The order of a graph = the number of vertices = $|G|$
- The size of a graph = the number of edges = $\|G\|$
- The degree of node $deg(v)$ is the number of edges incident to it
- The geodesic distance between v and w = number of edges in the geodesic path between v and w = $\delta(v, w)$

GEXF (Graph Exchange XML Format)

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" xmlns:xsi="http://www
  <meta lastmodifieddate="2014-05-20">
  <description>A Web of Hyperlinks</description>
  </meta>
  <graph defaultedgetype="directed">
    <attributes class="node">
      <attribute id="0" title="url" type="string"/>
    </attributes>
    <nodes>
      <node id="0" label="Gephi">
        <attvalues>
          <attvalue for="0" value="http://gephi.org"/>
        </attvalues>
      </node>
      <node id="1" label="Webatlas">
        <attvalues>
```

[GEXF Schema]

GML (Graph Modelling Language)

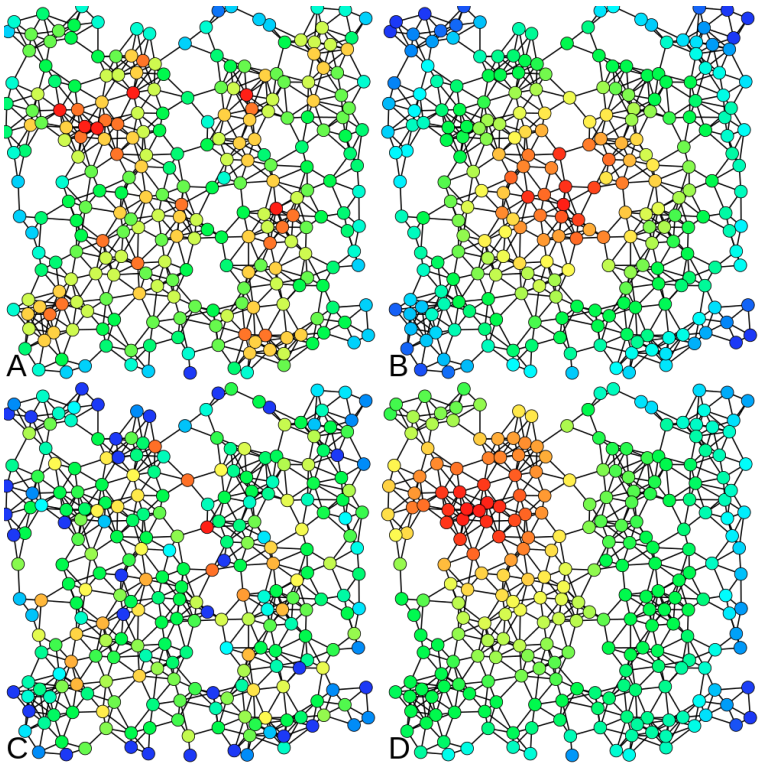
```
graph [
  directed 1
  id 888
  label "Social graph of my friends"
  node [
    id 1
    label "Me"
    ageAttribute 42
  ]
  node [
    id 2
    label "Jack"
    ageAttribute 43
  ]
  node [
    id 3
    label "Jill"
```

GML Specification

analysing graph

- Route Problems:
  - Minumm spanning tree
  - Shortest Path
  - Traveling salesman problem
- Graph coloring
- General properties
  - Radius
  - Density
  - Diameter
- Graph drawing/layout
  - Radial layout
  - Force-directed
- Partition/Community detection
- **Node Centralities**

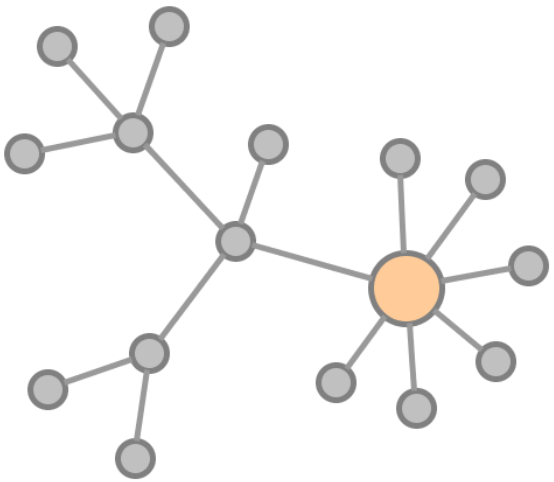
centralities



- A = Degree centrality,
  many neighbours
- B = Closeness centrality,
  close to other nodes
- C = Betweenness centrality,
  bridging nodes
- D = Eigenvector centrality,
  connecte to well-connected nodes

Image source: Claudio Rocchini, Wikimedia

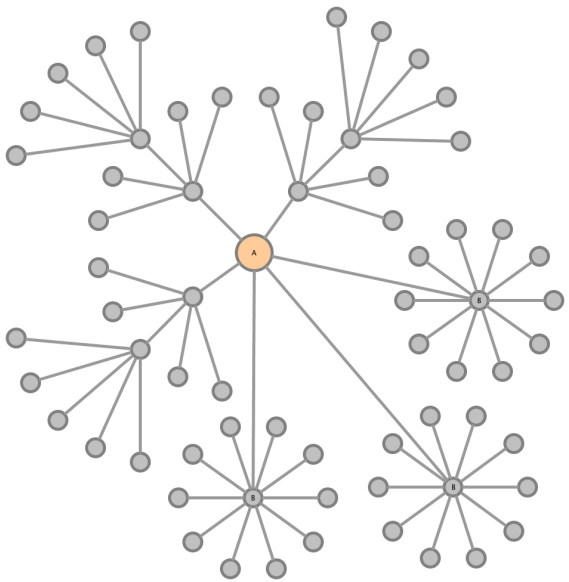
degree centrality



- "An important person typically knows a large number of people"
- Based on the degree of a node: the number of edges attached to it.

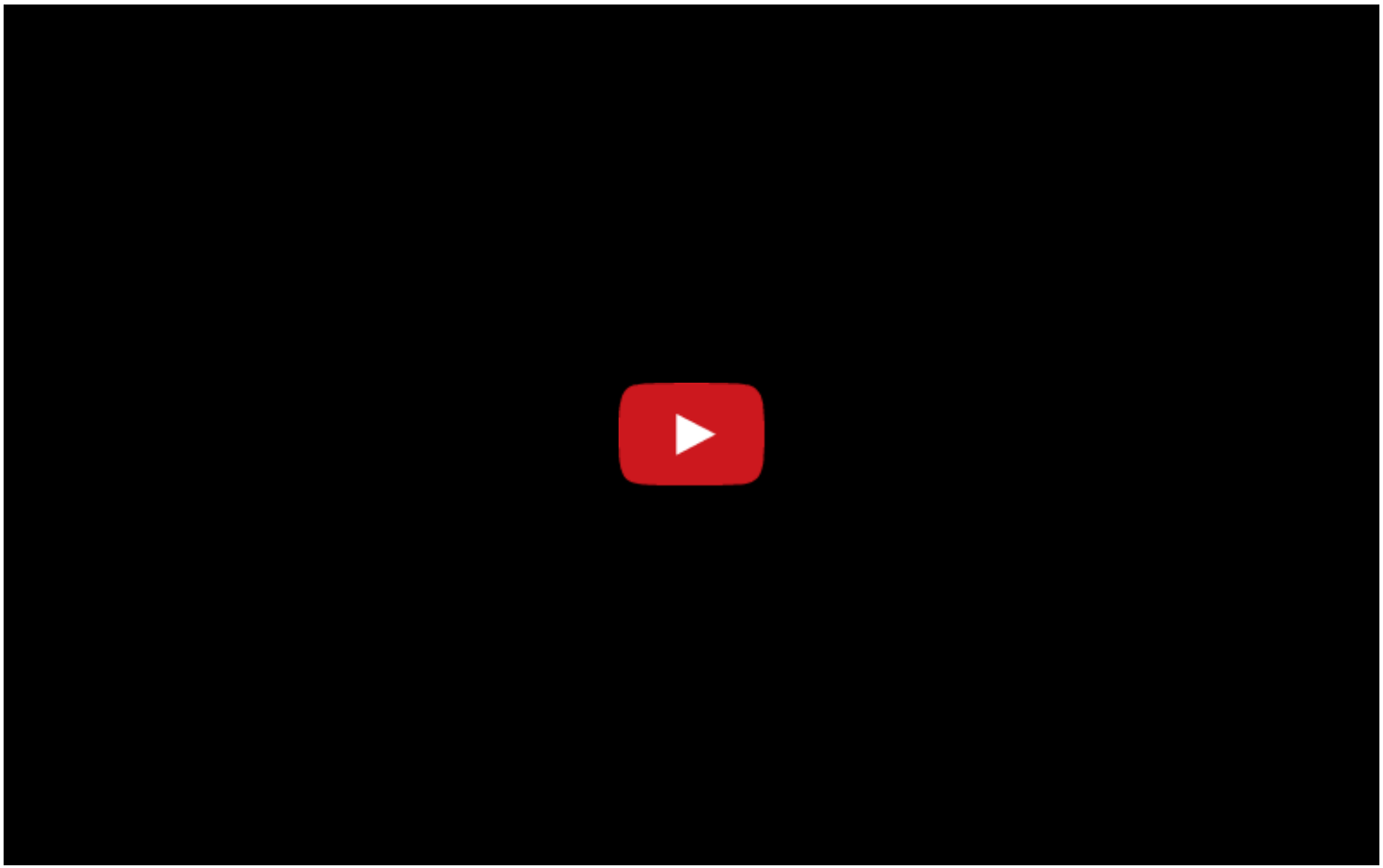
$C_d(v) = deg(v)$

eigenvector centrality



- "It's not simply how many persons you know, but also who they are."
- Similar to degree centrality, but take into consideration the degree of the neighbors. The importance of a node is enhanced by its connectivity to the more important nodes.
- Google PageRank,SocialRank

six degrees of separation



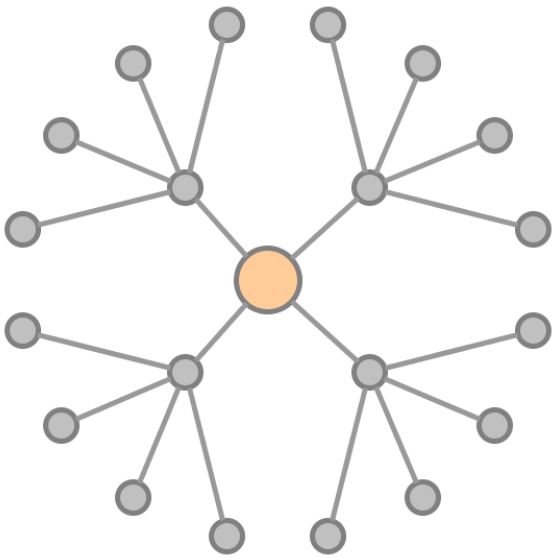
distance in graphs

- Bacon number to Hollywood is **Erdős number** to academia.
- Bacon number for actor X is the length of the "geodesic path" from Kevin Bacon to X
- Kevin Bacon himself has a Bacon number of 0
- Actors who have appeared together with Kevin Bacon in a movie have a Bacon number of 1. If N is the smallest Bacon number of any actor with whom Y has worked together, Actor Y's Bacon number is N+1.
- Average Bacon numbers is about 3.
- Average clicks from one page to any page in the Internet is **19**.

Oracle of Bacon, based on imdb.

R. Albert, H. Jeong, A.-L. Barabási, **Diameter of the world wide web**

closeness centrality



- "How close you are in general with other people in the network."
- The reciprocal of the average length of geodesic paths between a node and all other nodes

$$C_s(v) = \frac{1}{\sum_{w \in V} \delta(v,w)}$$

betweenness centrality

- "An important node lies on most of the paths between other nodes in the network", bridging the connectivity between other nodes.
- The total proportion of geodesic paths between any two nodes in which the node of interest lies.

$$C_b(v) = \sum_{v \neq s \in V} \sum_{v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

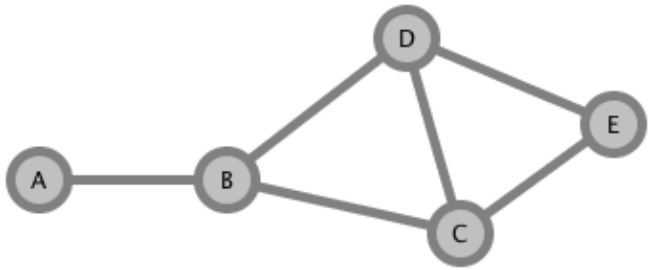
where:

- σ_{st} is the number of geodesic paths from node s to node t
- $\sigma_{st}(v)$ is the number of those paths that pass through v
- $\frac{\sigma_{st}(v)}{\sigma_{st}}$ is the proportion of geodesic paths from node s to node t that pass through v

betweenness centrality

list of geodesic paths

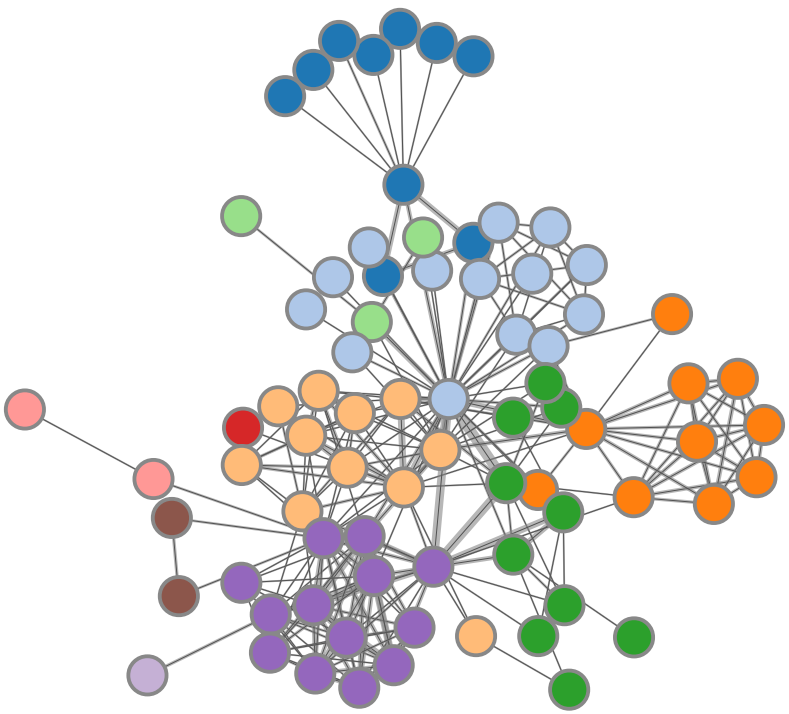
node	A	B	C	D	E
A		AB	ABC	ABD	ABCE, ABDE
B			BC	BD	BDE, BCE
C				CD	CE
D					DE



node	in between	proportion	centralities (normalised)
A	none	0	0
B	A-C, A-D, A-E	1/1+1/1+2/2 = 3	1
C	A-E, B-E	1/2+1/2 = 1	0.33
D	A-E, B-E	1/2+1/2 = 1	0.33
E	none	0	0

graph drawing

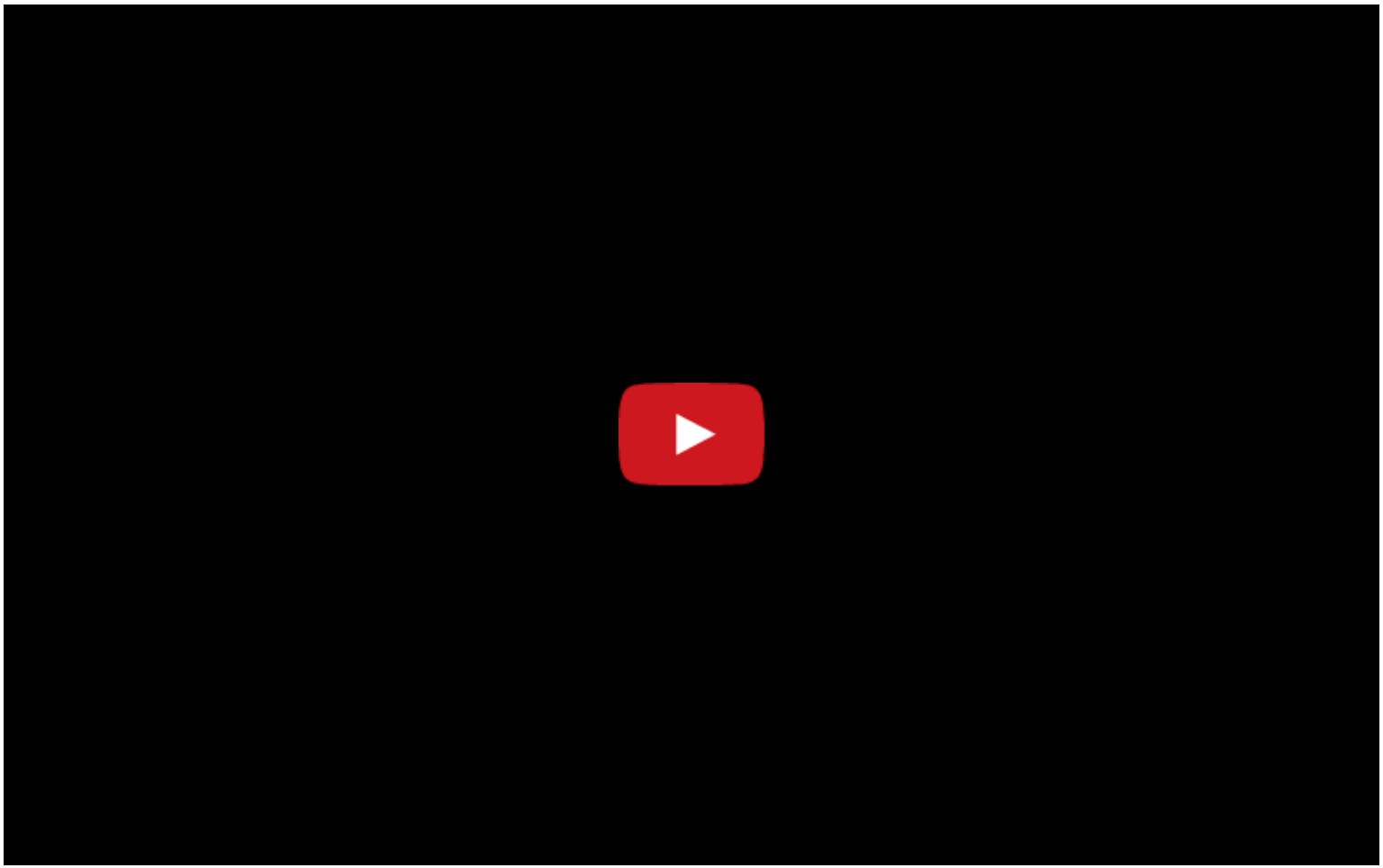
- Various graph "layouts": Radial, Circular,Force-directed, ...
- Force-directed: iterative, nodes as magnets (repulsive), edges as spring (pull)



database systems

- Relational: PostgreSQL, MySQL, MS SQL, Oracle
- Document/NoSql: MongoDB, CouchDB, Cassandra
- **Graph**: neo4j

The Power of Networks



Resources

Books:

- S. Wasserman & K. Faust (1994). Social Network Analysis.
- A.-L. Barabási. **Network Science**.
- D. Easley and J. Kleinberg. **Networks, Crowds, and Markets: Reasoning About a Highly Connected World**.
- R. Hanneman & M. Riddle (2005). **Introduction to social network methods**.

Tools:

- **NetworkX**
- **NetworkX port in JS**
- **Gephi - The Open Graph Viz Platform**
- **Gephi - Tutorial**
- **SigmaJS - JS Graph Drawing**
- **yEd Graph Editor**