Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

https://liacs.leidenuniv.nl/~takesfw/SNACS

Lecture 1 — Introduction and small world phenomenon

Context

Context: Data

- **Data**: facts, measurements or text collected for reference or analysis (Oxford dictionary)
  - Unstructured data: data that does not fit a certain data structure (text, images, audio, video, a list of numeric measurements)
  - Structured data: data that fits a certain data structure (table, **graph/network**, tree, etc.)

Data evolution

- Census data (60s)
- Transaction data (80s)
- Micro event data (00s)
- Social data (10s)

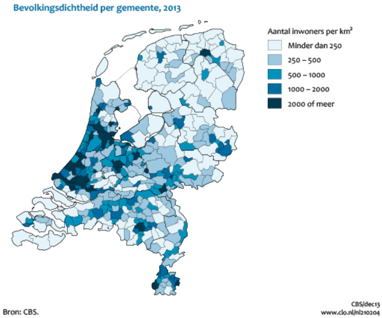
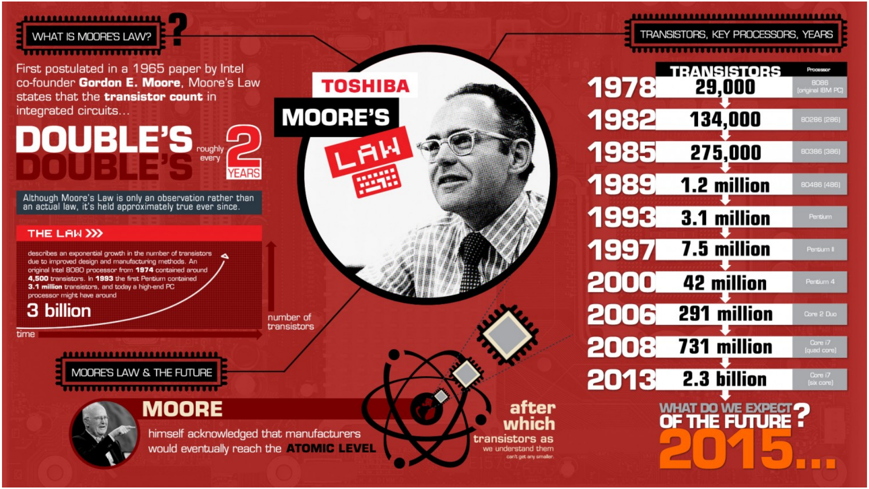


Figure: Census data

Moore's law & Transistors



Source `http://visual.ly`

Moore's law & Data

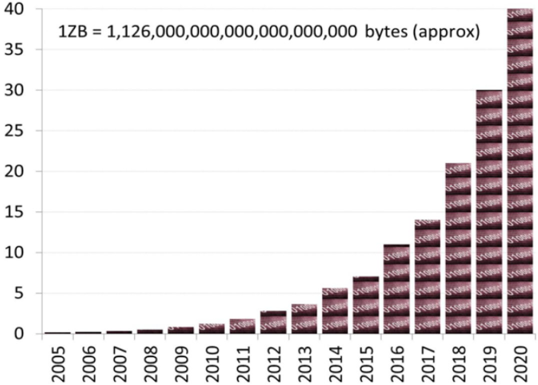
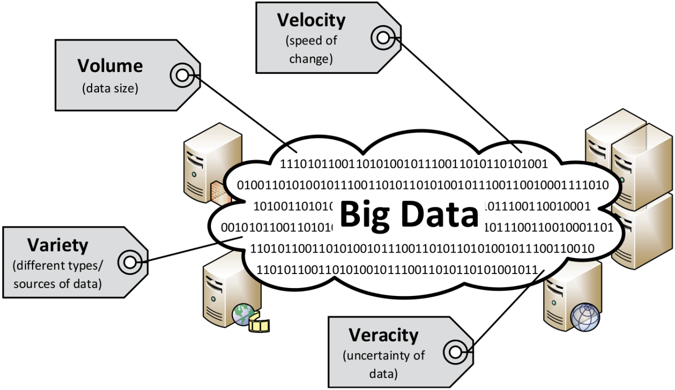


Figure: Zettabytes produced per year

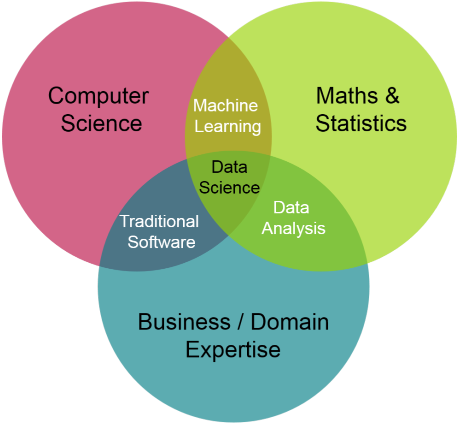
Source: http://www1.unece.org/stats/platform/display/msis/Big+Data

Context: Big data



Source: W. van der Aalst, Process Mining, 2nd edition, 2016.

Context: Data science



Source: https://ion.icaew.com/itcounts/b/weblog/posts/theaccountinganddatascienceworldsmeet

Context: Social media

Source: https://freepik.com

Social media mining

- Social media **platforms**: Facebook, Twitter, LinkedIn, Reddit, YouTube, Blogger, ...
- Platforms generate enormous amounts of (un)structured data
- **Social media mining & analytics**: analyzing this data in order to get insight in user(s), trends, usage patterns, the platform itself, ...
  - Text mining
  - Trend analysis
  - Sentiment mining
  - Topic modelling
  - **Social network analysis**

Context

- **Data**
  - Data analysis
  - Data mining
  - Data science
  - Big data

Context

- **Data**
  - Data analysis
  - Data mining
  - Data science
  - Big data
- **Network/graph data**

Context

- **Data**
  - Data analysis
  - Data mining
  - Data science
  - Big data
- **Network/graph data**
  - Graph mining
  - Network science
  - Complex network analysis
  - Social network analysis

Networks

What is a network?

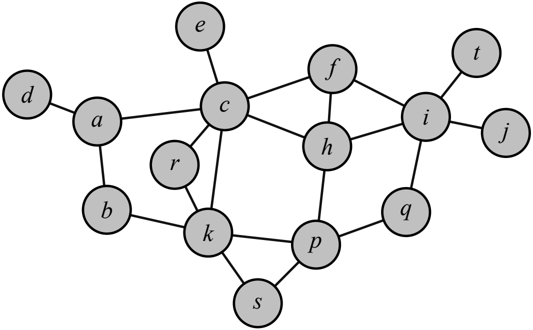


Figure: Visualization of a network with 15 nodes and 21 edges.

What is a network?

Networks, also called **graphs**, consist of:

- **Nodes**, also called objects, vertices, actors or entities, denoting the unit of analysis, and
- **Links**, also called relationships, **edges**, ties, arcs or connections that connect the aforementioned nodes in a particular meaningful way.

In a **social network**, the nodes represent people and the links may, e.g., indicate friendship, acquaintance, frequent proximity or communication.

This course also considers many other types of **real-world networks**.

Real-world networks

Network category	**Examples**
technological networks	webgraphs, information networks (Wikipedia), peer-to-peer networks, software design networks, internet router networks, digital circuit networks, cellular networks, WiFi networks
networks in nature	brain networks, protein interaction networks, neural networks, gene regulatory networks, metabolic networks, drug interaction networks, food webs, ecological networks
social networks	online social networks, human contact networks, playground interaction networks, sexual contact networks
communication networks	telephone call graphs, Twitter mention networks, WhatsApp and text messaging networks, e-mail networks
scientific networks	paper citation networks, co-authorship networks, patent citation networks, legal citation networks
financial & economic networks	money market networks, trade networks, financial transaction networks, cryptocurrency networks, ownership networks, corporate board interlock networks, intra-organizational networks
infrastructure networks	road networks, aiport networks, public transport networks, water transport networks, water distribution grid networks, transport, electricity/power grid networks

Network science

- **Network science**: understanding data by investigating interactions and relationships between individual data objects as a network
- **Networks** are the central model of computation

Network science

- **Network science**: understanding data by investigating interactions and relationships between individual data objects as a network
- **Networks** are the central model of computation
- Branch of data science focusing on network data
- Method in complexity research
- Complex systems approach: the behavior emerging from the network reveals patterns not visible when studying the individuals
- For now assume: network science = social network analysis

Representation and notation

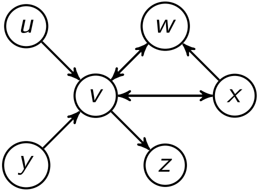
Notation

Concept	**Symbol**
■ Network (graph)	$G = (V, E)$
■ Nodes (objects, vertices, ...)	V
■ Links (ties, relationships, ...)	E
■ Directed — $E \subseteq V \times V$ — "links"	
■ Undirected — "edges"	
■ Number of nodes — $\lvert V \rvert$	n
■ Number of edges — $\lvert E \rvert$	m

- We assume no self-edges (u, u) and no parallel edges

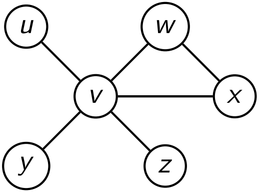
Notation example

- **Directed graph** $G = (V, E)$
- Nodes $V = \{u, v, w, x, y, z\}$
- Edges $E = \{(u, v), (w, v), (v, w)$ $(v, x), (x, v), (x, w), (y, v), (v, z)\}$
- Node count $n = 6$
- Link count $m = 8$



Notation example

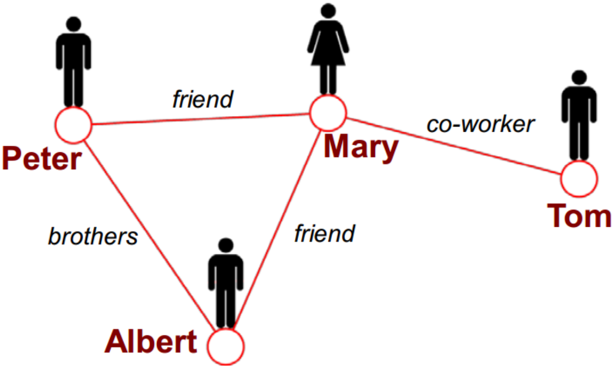
- **Undirected graph** $G = (V, E)$
- Nodes $V = \{u, v, w, x, y, z\}$
- Edges $E = \{\{u, v\}, \{w, v\}, \{v, x\}, \{x, w\}, \{y, v\}, \{v, z\}\}$
- Node count $n = 6$
- Edge count $m = 6$ (counting undirected edges)



Types of networks

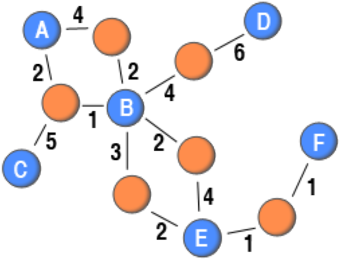
- Directed vs. undirected networks
- Weighted vs. unweighted (binary) networks
- Signed networks with positive and negative links
- Networks with attributed/annotated nodes and/or edges (metadata)
- One-mode (homogenic) vs. multi-mode (heteregenic) networks with different node types. Two-mode networks (bipartite graphs).
- Multiplex or multilayer networks with different edge types
- Static vs. dynamic (temporal/evolving) networks with timestamps on nodes and/or edges
- For now we stick to unweighted static one-mode networks.

One-mode labeled network



Source: `http://web.stanford.edu/class/cs224w`

Two-mode weighted network



Source: `http://toreopsahl.com`

Representation

- Directed **Adjacency Matrix**

	1	2	3	4	5	6
1	0	0	1	0	0	0
2	0	0	1	0	0	1
3	1	1	0	1	1	1
4	0	0	1	0	0	0
5	0	0	1	0	0	0
6	0	1	1	0	0	0

- Directed: $O(n^2)$ memory
- Weighted graphs: integers in cells

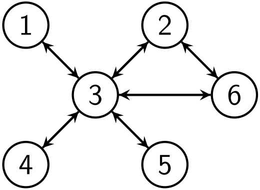


Figure: $n = 6$ and $m = 12$

Representation

- Undirected **Adjacency Matrix**

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 2 & 0 & & & & \\ 3 & 1 & 1 & & & \\ 4 & 0 & 0 & 1 & & \\ 5 & 0 & 0 & 1 & 0 & \\ 6 & 0 & 1 & 1 & 0 & 0 \end{array}$$

- Undirected: $O(\frac{1}{2}n(n-1))$ memory
- Better, but still many zeros

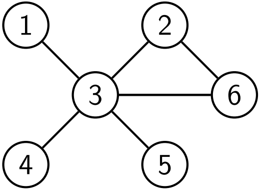


Figure: $n = 6$ and $m = 6$

Representation

- **Adjacency List**

```
1:  3
2:  3 6
3:  1 2 4 5 6
4:  3
5:  3
6:  2 3
```

- $O(n+2m)$ memory

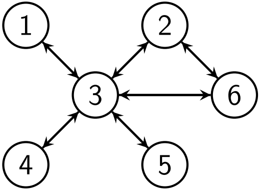


Figure: $n = 6$ and $m = 6$

Representation

- Undirected **Adjacency List**

```
1:  3
2:  3 6
3:        4 5 6
4:
5:
6:
```

- $O(n+m)$ memory

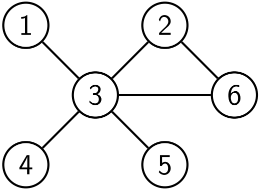


Figure: $n = 6$ and $m = 6$

Representation

- (Undirected) **Edge List**

  1 3
  2 3
  2 6
  3 4
  3 5
  3 6

- Commonly used as an input format
- $O(2m)$ memory

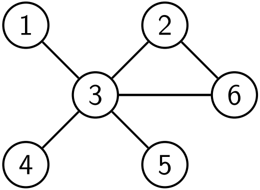
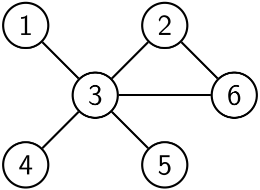
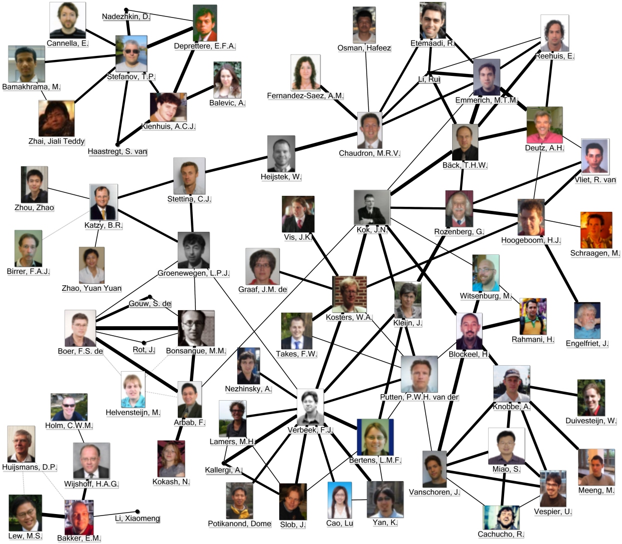


Figure: $n = 6$ and $m = 6$

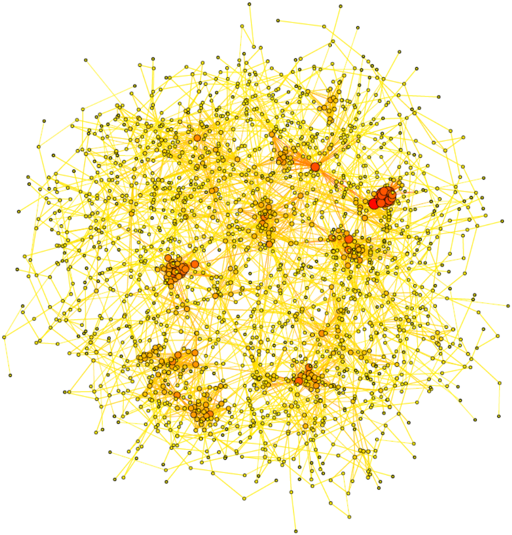
Toy graph: 6 nodes



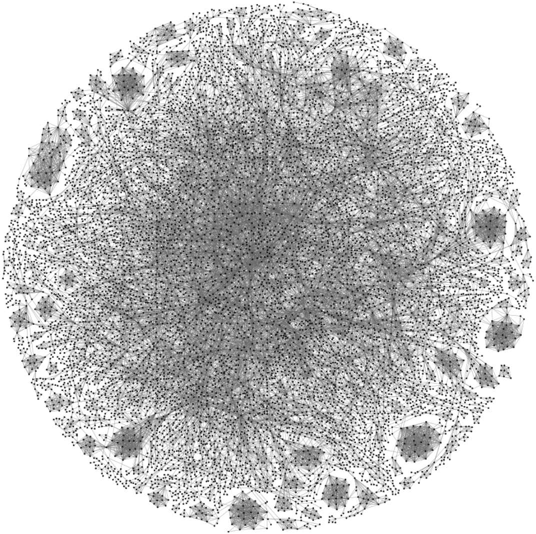
Collaboration network: ∼100 nodes



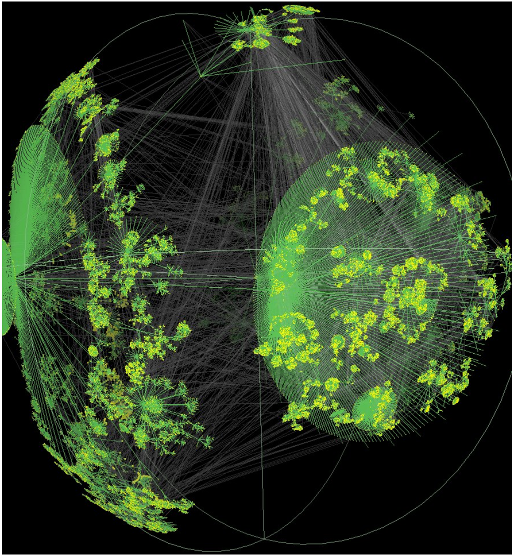
Social network: ∼1,500 nodes



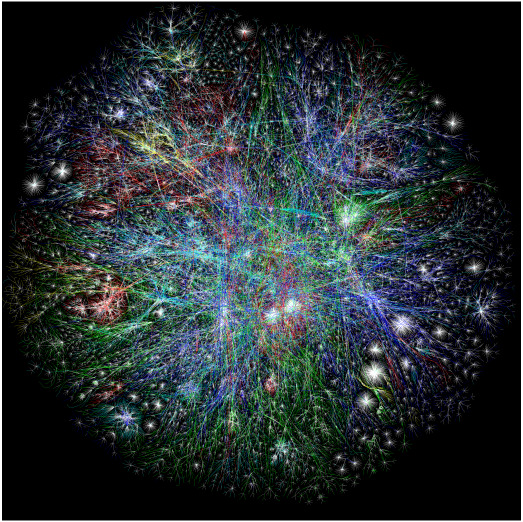
Corporate network: ∼20,000 nodes



Webgraph: ∼500,000 nodes

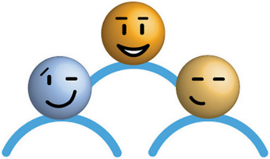


Webgraph: ∼500,000 nodes



Opte, Internet visualization (2005)

Hyves: $\sim$8,000,000 nodes

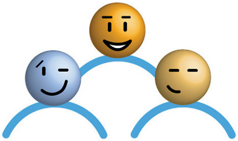


- Online Social Network
- Dutch & pre-Facebook
- Full snapshot
- $n = 8,000,000$ (8 million)
- $m = 1,000,000,000$ (1 billion)

Facebook: 1,000,000,000 nodes

Representing large networks

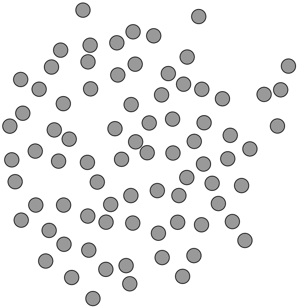
- HYVES online social network
  - $n = 8,000,000$ nodes
  - $m = 1,000,000,000$ links



- Assume 4 bytes per `int` (integer)
- Adjacency Matrix: $n^2 = 8,000,000^2 = 64 \cdot 10^{12}$ `bits` $= \sim 8$TB
- Adjacency List: $n + m = 1,008,000,000$ `ints` $= \sim 4$GB
- Edge List: $2m = 2,000,000,000$ `ints` $= \sim 8$GB
- But "smart" graph compression uses only a few *bits*(!) per edge

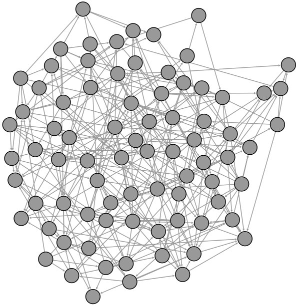
Measuring networks

- We have seen:
  - From 6 to $1,000,000,000$ (1 billion) nodes
  - From 8 to $120,000,000,000$ (120 billion) edges
- Measuring only number of nodes and edges is too simple



Measuring networks

- We have seen:
  - From 6 to $1,000,000,000$ (1 billion) nodes
  - From 8 to $120,000,000,000$ (120 billion) edges
- Measuring only number of nodes and edges is too simple



Real-world network properties

- Measuring only number of nodes and edges is too simple
- Real-world networks are far from random
- Five interesting metrics:
  1. Density
  2. Degree
  3. Components
  4. Distance
  5. Clustering coefficient

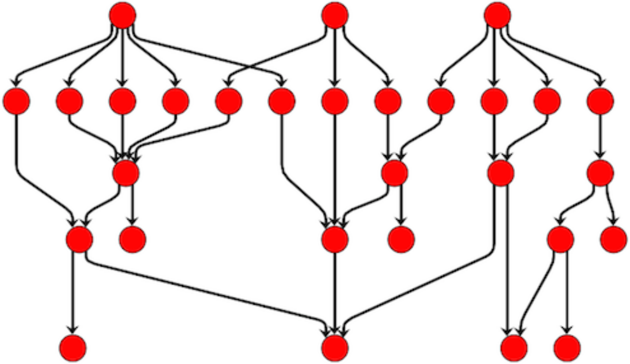
Density

- Maximum number of edges $m_{\max}$
  - $m_{\max} = n(n-1)$ for directed graphs
  - $m_{\max} = \frac{1}{2}n(n-1)$ for undirected graphs
- **Density**: $\frac{m}{m_{\max}}$, so $\frac{m}{n(n-1)}$ or $\frac{m}{\frac{1}{2}n(n-1)}$
- HYVES: $8 \cdot 10^6$ nodes, at most $64 \cdot 10^{12}$ edges.
  But network has "only" $1 \cdot 10^9$ edges, so density 0.0000156.
- Sparse graph if $m \ll m_{\max}$, so low density
- Real-world networks are typically **sparse**
- Density is particularly relevant when comparing networks

Bitcoin network

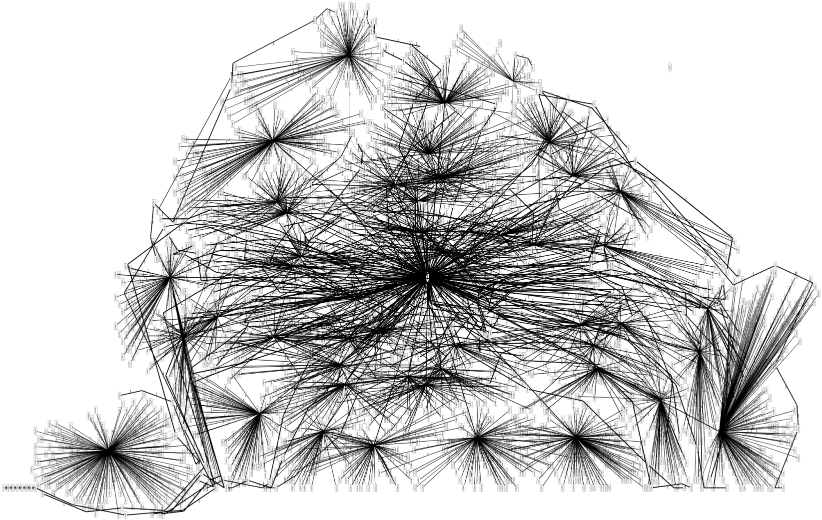
- Bitcoin: digital currency
- Peer-to-peer: no central authority
- Blockchain containing all transactions
- Bitcoin network: nodes are addresses (parts of wallets) and directed links are transactions between addresses
- Sparse: $n = 13,086,528$ nodes and $m = 44,032,115$ links

Bitcoin transaction network



Source: quantabytes.com/articles/a-network-analyst-s-view-of-the-block-chain

Silk Road Bitcoin seizure



Source: reddit.com/r/Bitcoin/comments/1prqpu/what_the_silk_road_bitcoin_seizure_transaction

Degree

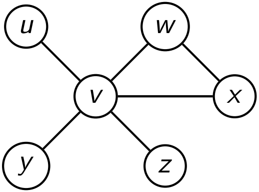


Figure: Undirected graph

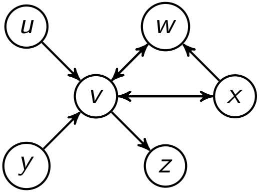


Figure: Directed graph

- Undirected graphs: degree $deg(v) = 5$

Degree

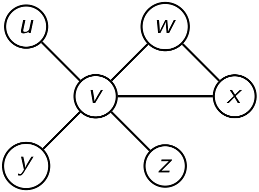


Figure: Undirected graph

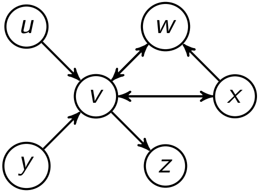


Figure: Directed graph

- Undirected graphs: degree $deg(v) = 5$
- Directed graphs
  - Indegree $indeg(v) = 4$
  - Outdegree $outdeg(v) = 3$

Degree

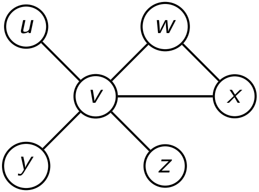


Figure: Undirected graph

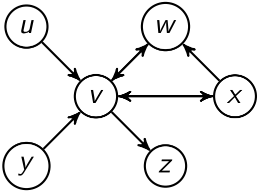
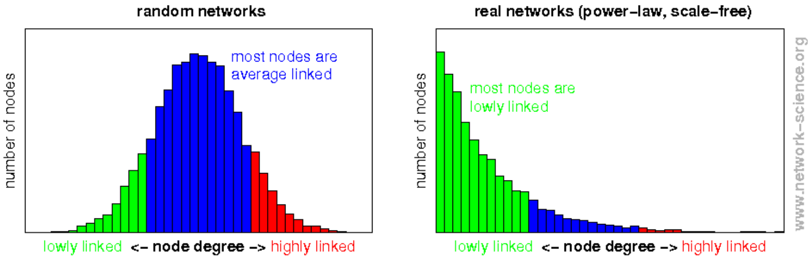


Figure: Directed graph

- Undirected graphs: degree $deg(v) = 5$
- Directed graphs
  - Indegree $indeg(v) = 4$
  - Outdegree $outdeg(v) = 3$
- **Degree distribution**: frequency of each degree value.
  Typically lognormal or power law distribution with "fat tail"

Degree distribution



Degree distribution

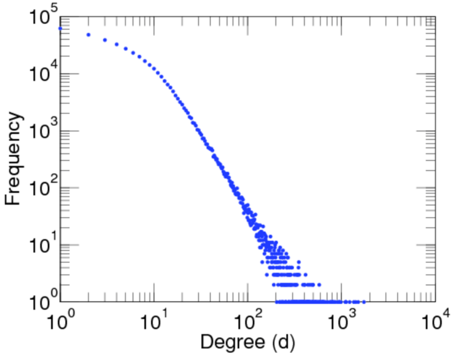
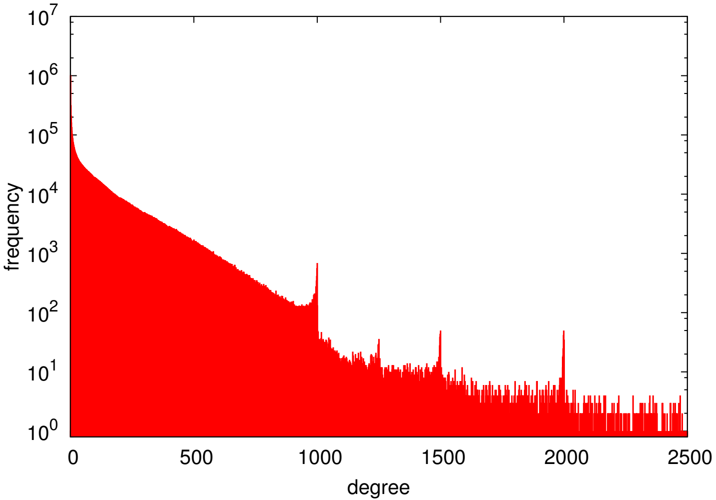


Figure: Degree distribution of Citeseer citation network.

Source: `http://konect.cc/networks/citeseer/`

HYVES degree distribution



Bitcoin network distribution

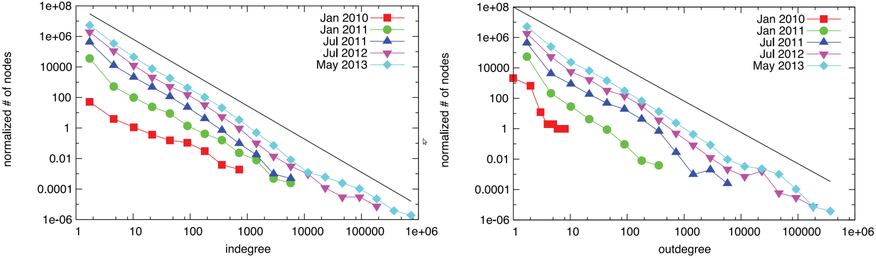
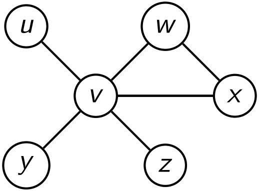


Figure: Scale-free degree distributions

Kondor et al., Do the Rich Get Richer? An Empirical Analysis of the Bitcoin. . . , PLOS ONE 9(2): e86197, 2014

Paths



Concept

- Path
- Path length
- Simple path: no repeated vertices
- Shortest path: path of minimal length
- **Distance**: length of shortest path

Example

$p = (u, v, z, v, w, x)$

$|p| - 1 = 5$

$p' = (u, v, w, x)$

$sp = (u, v, x)$

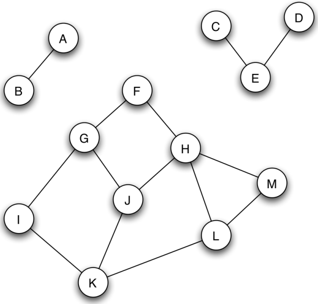
$d(u, x) = |sp| - 1 = 2$

Components in undirected networks

- What if $d(a, c) = \infty$? (so, no path between nodes a and c)

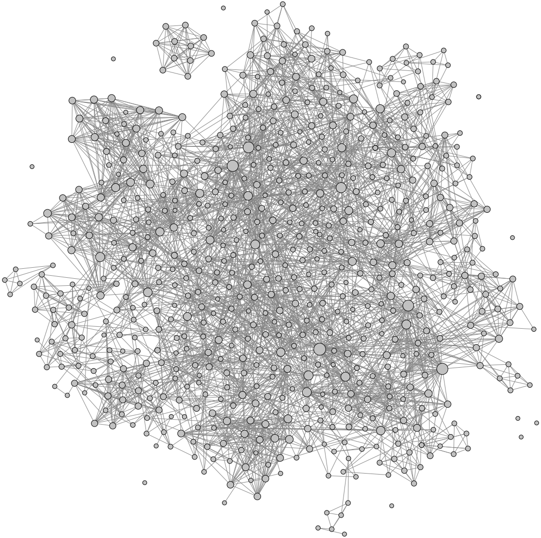
Components in undirected networks

- What if $d(a, c) = \infty$? (so, no path between nodes a and c)
- **Connected component**: subset of nodes (maximal in size) in which each node can form a path to each other node in the subset
- **Giant component**: component containing the largest number of nodes
- Real-world networks typically have one dominant giant component



Connected components
Image source: D. Easley and J. Kleinberg, "Networks, Crowds, and Markets", 2010

Giant component



Components in directed networks

- **Weakly** connected component: subgraph in which there is a path between any pair of nodes, ignoring link direction
- **Strongly** connected component: subgraph in which there is a directed path between any pair of nodes

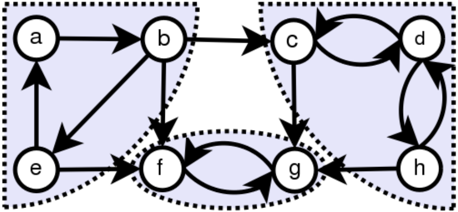


Figure: Directed network with 3 strongly connected components

Source: `https://commons.wikimedia.org/wiki/File:Scc.png`

Component size distribution

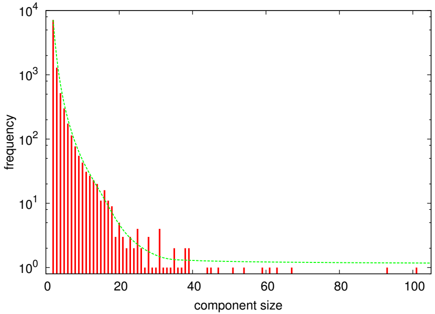


Figure: Component size distribution of HYVES network, excluding the giant component of ~ 8 million nodes.

Small world experiment

- Stanley Milgram
- Starts with 96 random people in Omaha
- Ask them to get a letter to a stock-broker in Boston by passing it through to a closer acquaintance.
- How many steps did it take?

Small world experiment

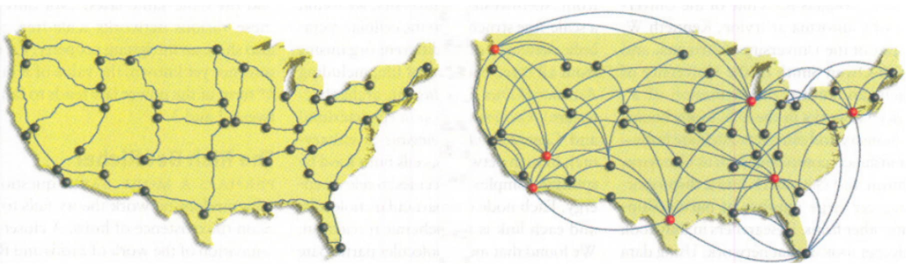
- Stanley Milgram
- Starts with 96 random people in Omaha
- Ask them to get a letter to a stock-broker in Boston by passing it through to a closer acquaintance.
- How many steps did it take?
- Letters arrived after on average 5.9 steps
- Total of 18 chains completed

J. Travers and S. Milgram, "An Experimental Study of the Small World Problem", Sociometry 32(4): 425-443, 1969

Yahoo small world experiment

Core/periphery structure

- Dense **core** containing many hubs
- **Periphery** with many nodes with a small distance to the core



Barabasi, Scientific American, May 2003

Distance

- Average distance $\overline{d} = \frac{1}{n(n-1)} \sum_{v,w \in V} d(v,w)$
- Distance distribution: how often each distance value occurs (computed over all node pairs).

Dataset	Nodes	Links	Average degree	Average distance
AstroPhys	17,903	396K	21	4.15
Enron	33,696	362K	10	4.07
Web	855,802	8.64M	10	6.30
YouTube	1,134,890	5.98M	5.3	5.32
Skitter	1,696,415	22.2M	13	5.08
Wikipedia	2,213,236	23.5M	11	4.81
Orkut	3,072,441	234M	76	4.16
LiveJournal	5,189,809	97.4M	19	5.48
Hyves	8,057,981	871M	112	4.75

F.W. Takes and W.A. Kosters, Determining the Diameter of Small World Networks, In CIKM, pp. 1191-1196, 2011.

Distance distribution

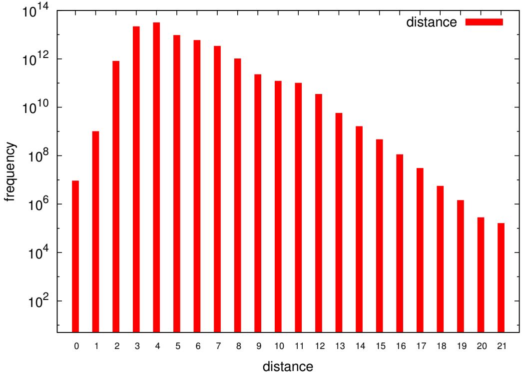


Figure: Distance distribution of the HYVES network (sampled over node pairs)

Erdős number

- Scientific collaboration network
- Edges between scientists who wrote a paper together
- Erdős number: the distance of a scientist (node) to Erdős
- `https://mathscinet.ams.org/mathscinet/collaborationDistance.html`

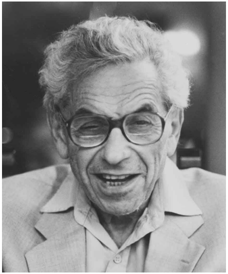
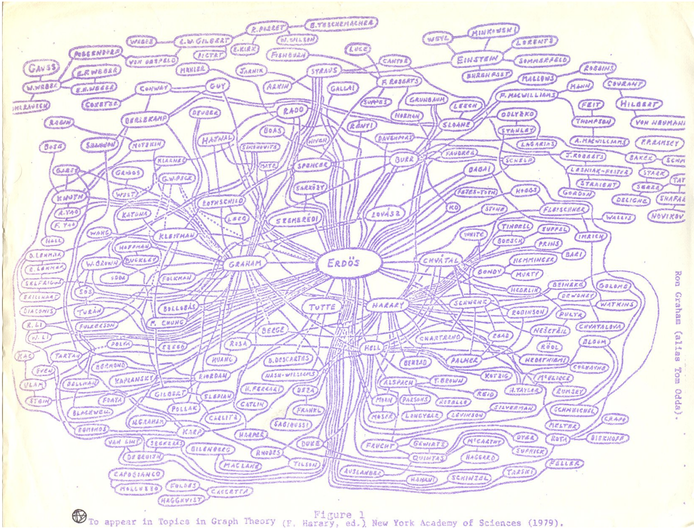
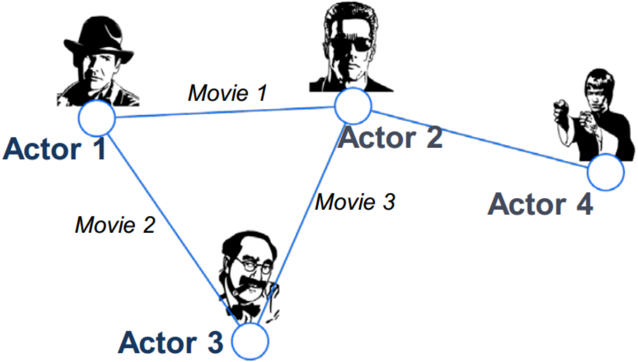


Figure: Paul Erdős (1913-1996)

Erdős number



Movie actor network



Source: `http://web.stanford.edu/class/cs224w`

Six degrees of Kevin Bacon

- Actor collaboration network based on co-starring actors
- Variant of “Six degrees of Separation”
- Edges between actors indicate they played in a movie together
- Try finding a path of length longer than six at `https://oracleofbacon.org`

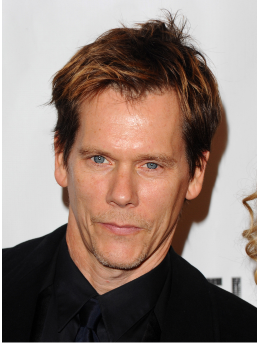
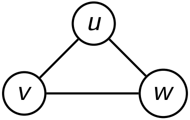


Figure: Kevin Bacon (1958)

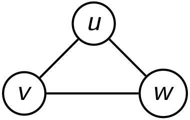
The Wiki Game

Triangles



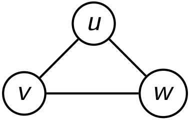
- **Triangle**: for nodes $u, v, w \in V$ we have $(u, v), (v, w), (w, u) \in E$

Triangles



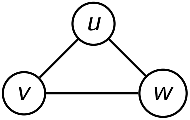
- **Triangle**: for nodes $u, v, w \in V$ we have $(u, v), (v, w), (w, u) \in E$
- Sets of three nodes that might be a triangle: $\binom{n}{3} \approx n^3/6$

Triangles



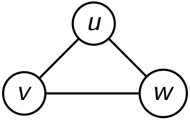
- **Triangle**: for nodes $u, v, w \in V$ we have $(u, v), (v, w), (w, u) \in E$
- Sets of three nodes that might be a triangle: $\binom{n}{3} \approx n^3/6$
- Probability of an edge in a a random graph is $m/\binom{n}{2} \approx 2m/n^2$

Triangles



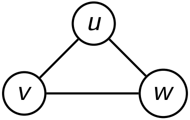
- **Triangle**: for nodes $u, v, w \in V$ we have $(u, v), (v, w), (w, u) \in E$
- Sets of three nodes that might be a triangle: $\binom{n}{3} \approx n^3/6$
- Probability of an edge in a a random graph is $m/\binom{n}{2} \approx 2m/n^2$
- Probability of one triangle is $(2m/n^2)^3 = 8m^3/n^6$

Triangles



- **Triangle**: for nodes $u, v, w \in V$ we have $(u, v), (v, w), (w, u) \in E$
- Sets of three nodes that might be a triangle: $\binom{n}{3} \approx n^3/6$
- Probability of an edge in a a random graph is $m/\binom{n}{2} \approx 2m/n^2$
- Probability of one triangle is $(2m/n^2)^3 = 8m^3/n^6$
- Expected triangles: $(8m^3/n^6)(n^3/6) = \frac{4}{3}(m/n)^3$

Triangles



- **Triangle**: for nodes $u, v, w \in V$ we have $(u, v), (v, w), (w, u) \in E$
- Sets of three nodes that might be a triangle: $\binom{n}{3} \approx n^3/6$
- Probability of an edge in a a random graph is $m/\binom{n}{2} \approx 2m/n^2$
- Probability of one triangle is $(2m/n^2)^3 = 8m^3/n^6$
- Expected triangles: $(8m^3/n^6)(n^3/6) = \frac{4}{3}(m/n)^3$
- For $n = 1000$ and $m = 8000$, we would expect 683 triangles.

Triangles

Network	Nodes	Edges	Expected	Real	Difference
Facebook (WOSN)	63,731	817,035	2,809	3,500,542	1,246×
Epinions	75,879	508,837	402	162,448	404×
Amazon (TWEB)	403,394	3,387,388	789	398,6507	5,049×
Baidu	415,641	3,284,387	658	14,287,651	21,718×
Youtube links	1,138,499	4,942,297	109	3,049,419	27,957×
Flickr	2,302,925	33,140,017	3,973	837,605,842	210,806×
LiveJournal links	5,204,176	49,174,464	1,125	310,876,909	276,367×
Twitter (MPI)	52,579,682	1,963,263,821	69,410	55,428,217,664	798,565×

Table: Expected vs. real triangle counts in real-world networks.

Node clustering coefficient

- **Node clustering coefficient**: extent to which a node v forms triangles with its neighbors
- Measure of transitivity
- Node clustering coefficient for a node $v \in V$:

$$C(v) = \frac{2 \cdot |\{(u,w) \in E : (u,v) \in E \land (v,w) \in E\}|}{deg(v) \cdot (deg(v) - 1)}$$

(where $deg(v) > 1$ is the degree of node v)

Node clustering coefficient

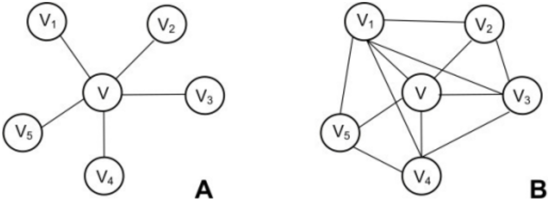
- **Node clustering coefficient**: extent to which a node v forms triangles with its neighbors
- Measure of transitivity
- Node clustering coefficient for a node $v \in V$:

$$C(v) = \frac{2 \cdot |\{(u,w) \in E : (u,v) \in E \wedge (v,w) \in E\}|}{deg(v) \cdot (deg(v) - 1)}$$

(where $deg(v) > 1$ is the degree of node v)

$$C(v) = \frac{2 \cdot \text{edges between neighbors of } v}{\text{maximum number of such edges}}$$

Node clustering coefficient



- Situation A: v has a clustering coefficient of 0
- Situation B: v has a clustering coefficient of $\frac{14}{20} = \frac{7}{10} = 0.7$

Image: G.A. Pavlopoulos et al., "Using graph theory to analyze biological networks", in BioData Mining 4(1), 2011.

Graph clustering coefficient

1. **Average node clustering coefficient** for a graph G:

$$C(G) = \frac{1}{n} \cdot \sum_{v \in V} C(v)$$

Graph clustering coefficient

1. **Average node clustering coefficient** for a graph G:

$$C(G) = \frac{1}{n} \cdot \sum_{v \in V} C(v)$$

2. **Graph clustering coefficient** for a graph G:

$$C'(G) = \frac{3 \cdot \text{number of triangles}}{\text{number of connected triplets of nodes}}$$

- Small world networks: **high clustering coefficients** compared to a random graph with the same number of nodes

Real-world networks

1. Sparse networks — density
2. Fat-tailed power-law degree distribution — degree
3. Giant component — components
4. Low pairwise node-to-node distances — distance
5. Many triangles — clustering coefficient

Real-world networks

1. Sparse networks — density
2. Fat-tailed power-law degree distribution — degree
3. Giant component — components
4. Low pairwise node-to-node distances — distance
5. Many triangles — clustering coefficient

- Many examples: social networks, communication networks, citation networks, collaboration networks (Erdős, Kevin Bacon), protein interaction networks, information networks (Wikipedia), webgraphs, financial networks (Bitcoin) . . .

Other topics

- Centrality, PageRank
- Community detection
- Network motifs
- Graph representation and compression
- Distance approximation
- Graph evolution, link prediction
- Spidering and sampling
- Visualization algorithms
- Virality and influence maximization
- Epidemic spread
- Privacy, anonymity and ethics
- Anomalies in networks
- Resilience and fault tolerance

Upcoming lab session

- From 9:00 to 10:45 in Snellius rooms 302/304 etc.
- Instructions on course website
- Hands-on introduction to `Gephi`
- Get to know the university's (remote) Linux environment (again)
- Start working on Assignment 1

Homework for next week

- Mandatory (de)registration via uSis/Brightspace; see Lecture 0
- Watch the "The Emergence of Network Science" movie at `https://www.cornell.edu/video/emergence-of-network-science` or `https://youtu.be/cf-6qdPerlI?t=1s`

- Ensure you have access to the ULCN Linux environment in, the Snellius computer rooms and/or remotely via `sshgw.leidenuniv.nl`
- Check if you have read access to the files in this folder: `/vol/share/groups/liacs/scratch/SNACS/`
- Solve any IT problems; 8888 or `helpdesk@issc.leidenuniv.nl` or `https://liacs.leidenuniv.nl/ict` (redirect to ISSC portal)