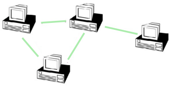
graphology & *networkology*

introduction to *network analysis* (*ina*)

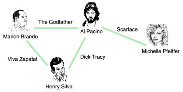
Lovro Šubelj
University of Ljubljana
spring 2021/22

terminology *graphs & networks*

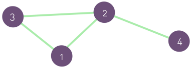
- *synonyms* perspective
- *graph theory* perspective
  - *graph* is formal *mathematical object*
  - *network* is *graph with real data*
- *network science* perspective
  - *network* is some *real-world system*
  - *graph* is *representation of network*
- but Web graph, Internet map



network



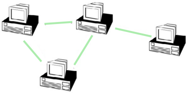
another network



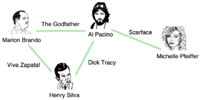
graph

terminology *nodes & links*

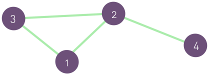
- *graph theory* terminology
  - *vertices* and *edges/relations*
- *network science* terminology
  - *nodes* and *edges/links*
- *social science* terminology
  - *agents/brokers/units* and *ties*



nodes & links



agents & ties

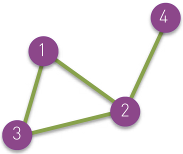


vertices & edges

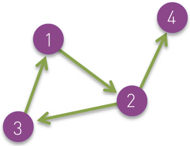
terminology *classes*

- *social* networks
  - nodes are *people or animals*, links are relations or interactions
  - Facebook, offline, online, affiliation, author/actor collaboration
- *information* networks
  - nodes are information sources, links show *information flow*
  - Web, Twitter, citation, communication, peer-to-peer
- *technological* networks
  - human-made infrastructure with *technological constraints*
  - Internet, telephone, transportation, power grid, software
- *biological* networks
  - interaction between genes, cells, neurons in *living beings*
  - gene regulatory, metabolic, protein interaction, neural
- *ecological*, *lexical*, *financial*, *sports* etc. networks

- simple *graph* G is defined by
  - set of nodes $N = \{1, 2 \ldots n\}$
  - set of links L where $m = |L|$
- if G is *undirected* then $L \subseteq \{\{i, j\} | \ i, j \in N\}$
- if G is *directed* then $L \subseteq \{(i, j) | \ i, j \in N\}$



undirected graph



directed graph

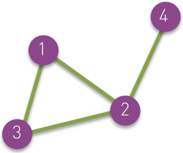
graphology *adjacency*

- *adjacency matrix* A is $n \times n$ matrix defined as
  - $A_{ij} = 1$ if there is link *from* j *to* i
  - $A_{ij} = 0$ if $i = j$ or *otherwise*
- if G is *undirected* then $A = A^T$ and $\sum_{ij} A_{ij} = 2m$
- if G is *directed* then $A \neq A^T$ and $\sum_{ij} A_{ij} = m$

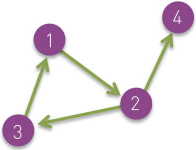
$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

undirected graph   directed graph

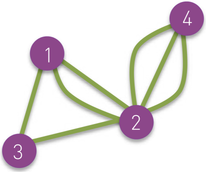
graphology *multigraphs*



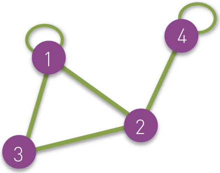
simple undirected
$A_{ij} = A_{ji} \in \{0, 1\}$



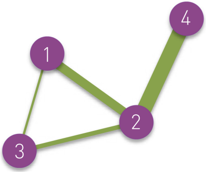
simple directed
$A_{ij} \neq A_{ji} \in \{0, 1\}$



multigraph $A_{ij} \in \mathbb{N}_0$



loops $A_{ii} \in \{2, 4, \dots\}$

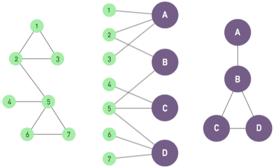


weighted $W_{ij} \in \mathbb{R}_{\geq 0}$

graphology *multipartite*

- undirected *bipartite graph* G_B is defined by
  - sets of nodes $N_1 = \{1, 2 \ldots n_1\}$ and $N_2 = \{1, 2 \ldots n_2\}$
  - set of m links $\subseteq N_1 \times N_2$
- *incidence matrix* B is $n_2 \times n_1$ matrix defined as
  - $B_{ij} = 1$ if there is link *between j and i*
  - $B_{ij} = 0$ *otherwise*
- (*one-mode*) *projections* are multigraphs with

$$A = B^T B - D_1 \qquad A = B B^T - D_2$$

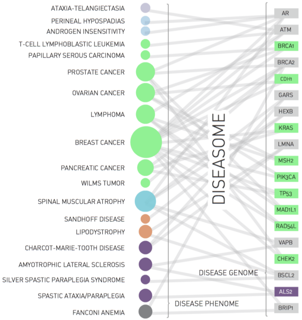


bipartite graph & projections

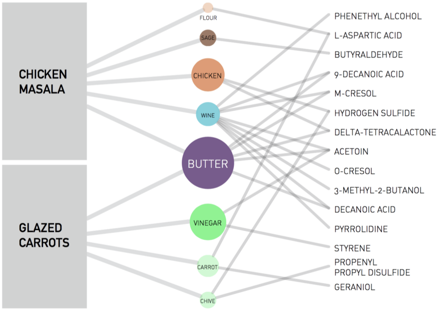
$$B = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

incidence matrix

networkology *multi-mode*



bipartite or *two-mode network*



tripartite graph or *three-mode network*

graphology *degrees*

— for *undirected* G *degree* k_i of i is number of *incident links*

$$k_i = \sum_j A_{ij} = \sum_j A_{ji}$$

— for *directed* G *degree* $k_i = k_i^{in} + k_i^{out}$

- *in-degree* k_i^{in} of i is number of *incoming links*

$$k_i^{in} = \sum_j A_{ij}$$

- *out-degree* k_i^{out} of i is number of *outgoing links*

$$k_i^{out} = \sum_j A_{ji}$$

— thus (*network*) *average degrees* $\langle k \rangle$ and $\langle k^{\cdot} \rangle$ are

$$\langle k \rangle = 2m/n \qquad \langle k^{in} \rangle = \langle k^{out} \rangle = m/n$$

networkology *degrees*

- *average degree* $\langle k \rangle$ of real networks [Bar16]
- mostly $\langle k \rangle \leq 10$ despite *very different n*

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

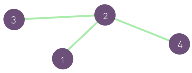
- but $\langle k \rangle = 190.5$ for *Facebook* friendships [BBR+12]

— for *undirected* G *density* ρ is defined as

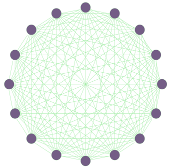
$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)} = \frac{\langle k \rangle}{n-1}$$

— for *directed* G *density* $\rho^{\cdot}$ is defined as

$$\rho^{\cdot} = \frac{m}{n(n-1)} = \frac{\langle k^{in} \rangle}{n-1} = \frac{\langle k^{out} \rangle}{n-1}$$



tree $m = n - 1$



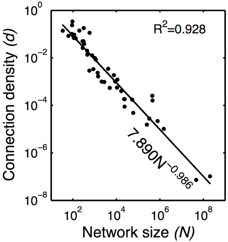
complete $m = \binom{n}{2}$

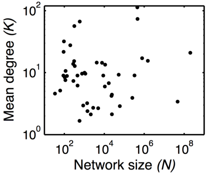
— G is *dense* if $\rho \to c > 0$ as $n \to \infty$ thus $\langle k \rangle = \mathcal{O}(n)$

— G is *sparse* if $\rho \to 0$ as $n \to \infty$ thus $\langle k \rangle \neq \mathcal{O}(n)$

networkology *density*

- *density* ρ and *degree* $\langle k \rangle$ of real networks [LJT+11]
- real networks are *sparse* $\rho \approx \mathcal{O}(n^{-1})$ and $\langle k \rangle \ll n$





- $\rho \approx \frac{2{\cdot}69{\cdot}10^9}{721^2{\cdot}10^{12}} < 10^{-6}$ for *Facebook* friendships [BBR+12]
- A of real networks contains *almost only zeros* $m \approx \mathcal{O}(n)$

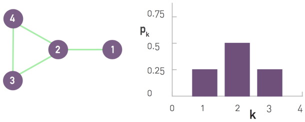
graphology *degree distribution*

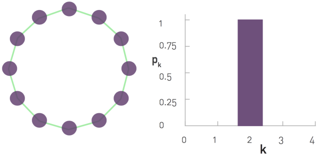
- for *undirected* G *degree distribution* p_k is defined as
  - n_k is number of *degree-k* nodes

$$p_k = n_k/n \qquad \sum_k p_k = 1 \qquad \langle k \rangle = \sum_k k p_k$$

- for *directed* G *in-/out-degree distributions* p_k^{in} and p_k^{out}
  - n_k^{in} and n_k^{out} are numbers of *in-/out-degree-k* nodes

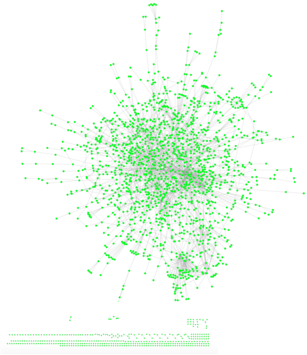
$$p_k^{in} = n_k^{in}/n \qquad \sum_k p_k^{in} = 1 \qquad \langle k^{in} \rangle = \sum_k k p_k^{in}$$

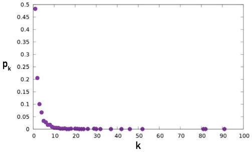


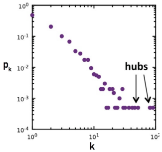


networkology *degree distribution*

- *heavy-tail distribution* p_k of protein network [Bar16]
- nodes with *very high* $k \gg \langle k \rangle$ are called *hubs*

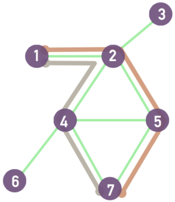




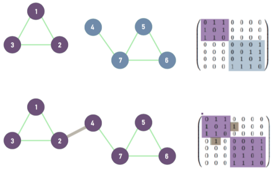


pathology *connectivity*

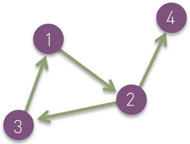
- for *undirected* G *path* P_{ij} is sequence of *links between* i *and* j
  - (*connected*) *component* is *maximal subset* thus $\forall i, j : \exists P_{ij}$
  - *giant component* contains *nontrivial fraction* of nodes
  - *connected* G has *exactly one* connected component
- for *directed* G *path* $\overrightarrow{P_{ij}}$ is seq. of *directed links from* i *to* j
  - *weak*/*strong connectivity* defined through P and $\overrightarrow{P}$



paths P_{17}



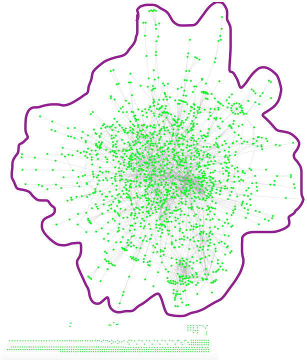
(dis)connected with *bridge*



one *WCC*/two *SCC*

networkology *connectivity*

— *giant/largest component* of protein network [Bar16]



— *giant* $> 99,7\%$ for *Facebook* friendships [BBR+12]

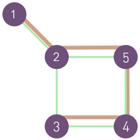
— could real network have *two giant components*?

pathology *distances*

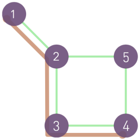
- *length* of path P or $\overrightarrow{P}$ is number of *links*/*hops*
- *geodesic path* G_{ij} or $\overrightarrow{G_{ij}}$ is any *shortest* P_{ij} or $\overrightarrow{P_{ij}}$
- *distance* d_{ij} between i and j is *length* of G_{ij} or $\overrightarrow{G_{ij}}$

- (*network*) *diameter* d_{max} or D is *maximum* d_{ij}
- (*network*) *average distance* $\langle d \rangle = \ell$ and ℓ^{-1} defined as
  - $d_{ij} = 0$ and $d_{ij} = \infty$ for i and j in different components

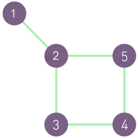
$$\langle d \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij} \qquad \ell^{-1} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}}$$



$P_{13} \neq G_{13}$



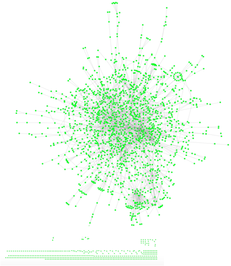
$d_{14} = 3$

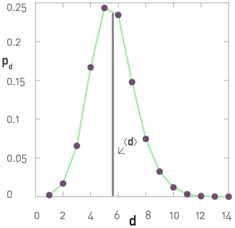


$\langle d \rangle = 1.6$

— *distance distribution* p_d of protein network [Bar16]

— most nodes are on *similar distance* $d \approx \langle d \rangle$





— $\langle d \rangle = 4.74$ for *Facebook* friendships [BBR+12]

— real networks have *surprisingly small* $\langle d \rangle \ll n$

graphology *clustering*

- for *undirected* G *node clustering coefficient* C_i of i is
  - t_i is number of *linked neighbors* or *triangles* of i

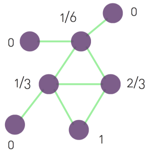
$$C_i = \frac{2t_i}{k_i(k_i-1)} \qquad C_i = 0 \text{ for } k_i \leq 1$$

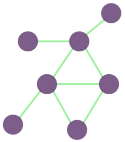
- *average clustering coefficient* $\langle C \rangle$ [WS98] is defined as

$$\langle C \rangle = \frac{1}{n} \sum_i C_i$$

- *network clustering coefficient* C [NSW01] is defined as

$$C = \frac{3 \times \text{number of } \textit{triangles}/\textit{closed triads}}{\text{number of } \textit{linked triples}/\textit{connected triads}}$$



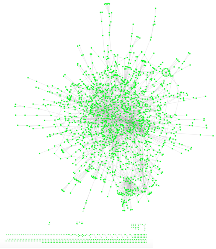


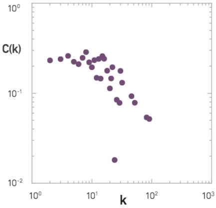
$\langle C \rangle = \frac{13}{7 \cdot 6} = 0.31$ $\qquad$ $C = \frac{3 \cdot 2}{16} = 0.38$

networkology *clustering*

— *clustering distribution* $C(k)$ of protein network [Bar16]

— hubs have *much lower* C than nodes with $k \approx \langle k \rangle$





— $\langle C \rangle = 0.61$ for *Facebook* social circles [ML12]

— real (social) networks have *significant* $\langle C \rangle \gg 0$

A.-L. Barabási.
Network Science.
Cambridge University Press, Cambridge, 2016.

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna.
Four degrees of separation.
In *Proceedings of the ACM International Conference on Web Science*, pages 45–54, Evanston, IL, USA, 2012.

Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj.
Exploratory Social Network Analysis with Pajek: Expanded and Revised Second Edition.
Cambridge University Press, Cambridge, 2011.

David Easley and Jon Kleinberg.
Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, Cambridge, 2010.

Ernesto Estrada and Philip A. Knight.
A First Course in Network Theory.
Oxford University Press, 2015.

Paul J. Laurienti, Karen E. Joyce, Qawi K. Telesford, Jonathan H. Burdette, and Satoru Hayasaka.
Universal fractal scaling of self-organized networks.
Physica A, 390(20):3608–3613, 2011.

Seth A. Myers and Jure Leskovec.
Clash of the contagions: Cooperation and competition in information diffusion.
In *Proceedings of the IEEE International Conference on Data Mining*, 2012.

Mark E. J. Newman.
Networks.
Oxford University Press, Oxford, 2nd edition, 2018.

M. E. J. Newman, S. H. Strogatz, and D. J. Watts.
Random graphs with arbitrary degree distributions and their applications.
Phys. Rev. E, 64(2):026118, 2001.

D. J. Watts and S. H. Strogatz.
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442, 1998.