FFHS
Fernfachhochschule Schweiz
Zürich | Basel | Bern | Brig
Mitglied der SUPSI

Einführung in Data Science - Block2 Netzwerkanalyse

Programm

Thema	Form	Zeit
Besprechung der Semesterarbeit		13:45 – 14:00
Netzwerkanalyse -VB	Diskussion	14:00 – 14:30
Netzwerkanalyse	Vorlesung – Teil 1	14:30 – 15:00
Pause		15:00 – 15:15
Netzwerkanalyse	Vorlesung – Teil 2	15:15 – 16:45
Workshop	Gruppenarbeit	16:45 – 17:00

Netzwerkanalyse
Graph Theorie

Workshop aus den Verständnisfragen

Tisch 1

- Welche Elemente hat GitHub als Netzwerk – was sind Knoten was sind Kanten?
- Welche Eigenschaften hat das Netzwerk?
- Was bedeutet die Aussage, dass GitHub ein Interest Graph sei?

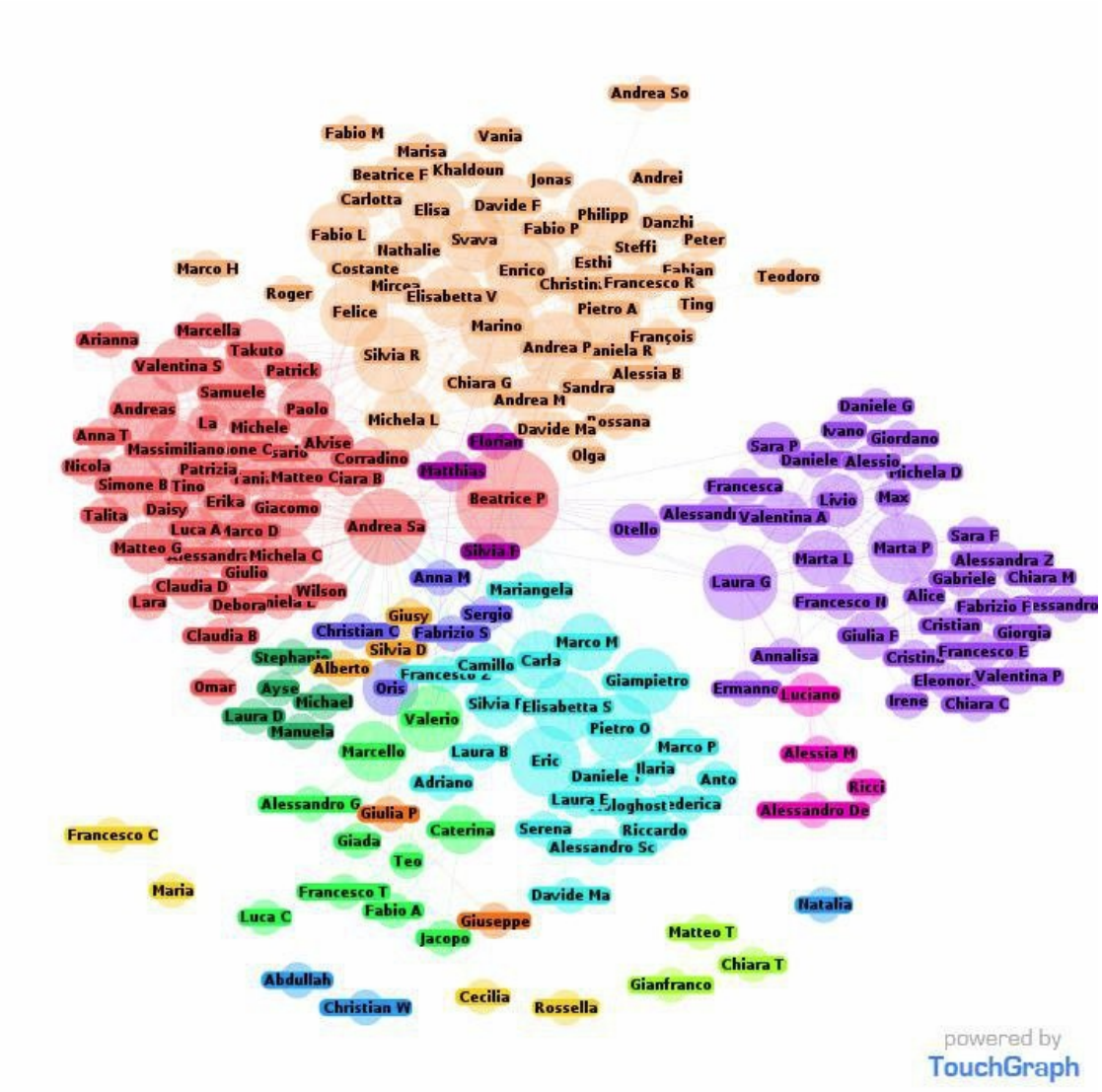
Tisch 2

- Inwiefern ist GitHub ein Social Graph?
- Ist GitHub ein Ego Graph?
- Ist GitHub ein bipartiter Graph?
- Inwiefern beschreibt die Adjazenzmatrix ein Netzwerk?

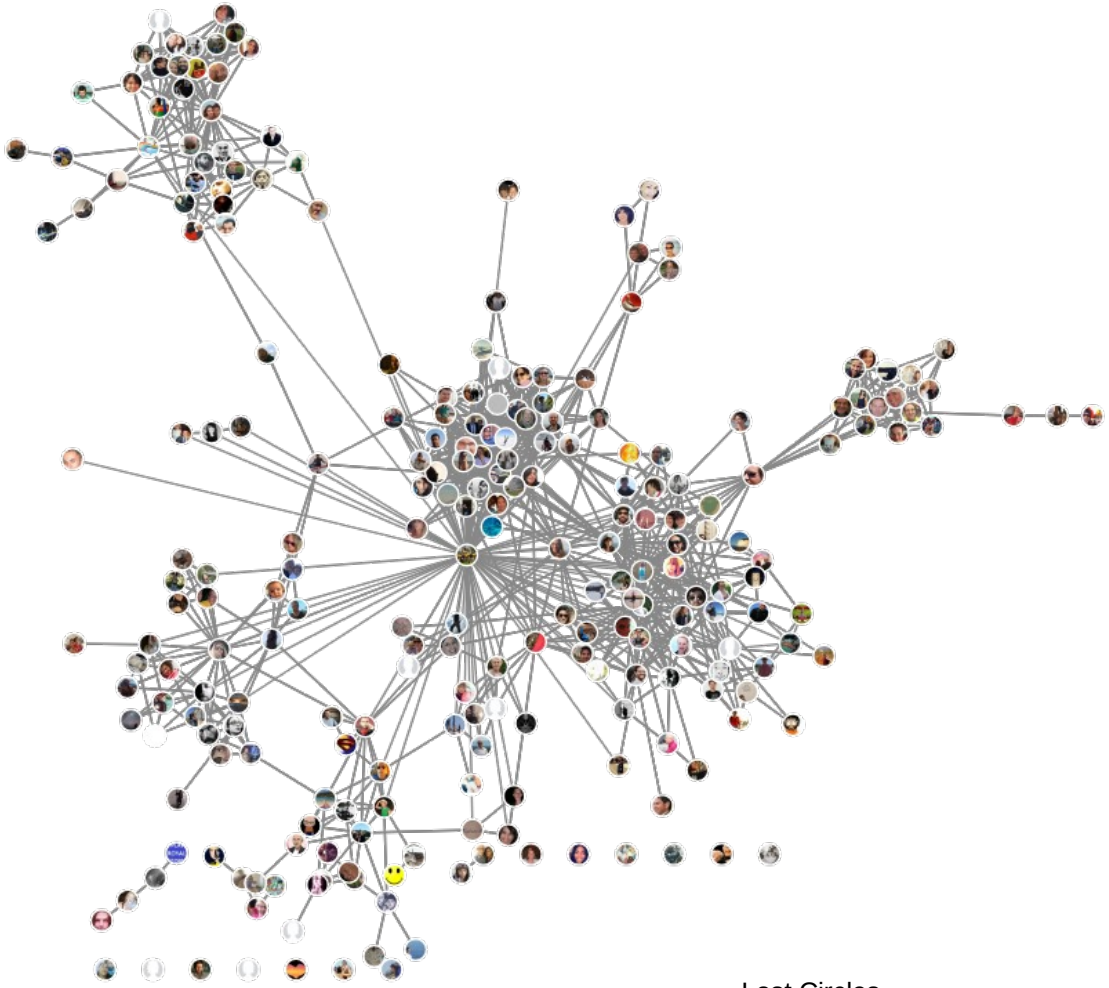
Netzwerke

- Ein Netzwerk ist ein System bestehen aus einzelnen Elemente (Knoten) und Verbindungen (Kanten) zwischen ihnen.
- Beispiele
  - Soziale Netzwerke
  - Kommunikationsnetwerke
  - Informationsnetzwerke
  - Biologische Netzwerke
  - Transportnetze

Soziale Netzwerke: Facebook

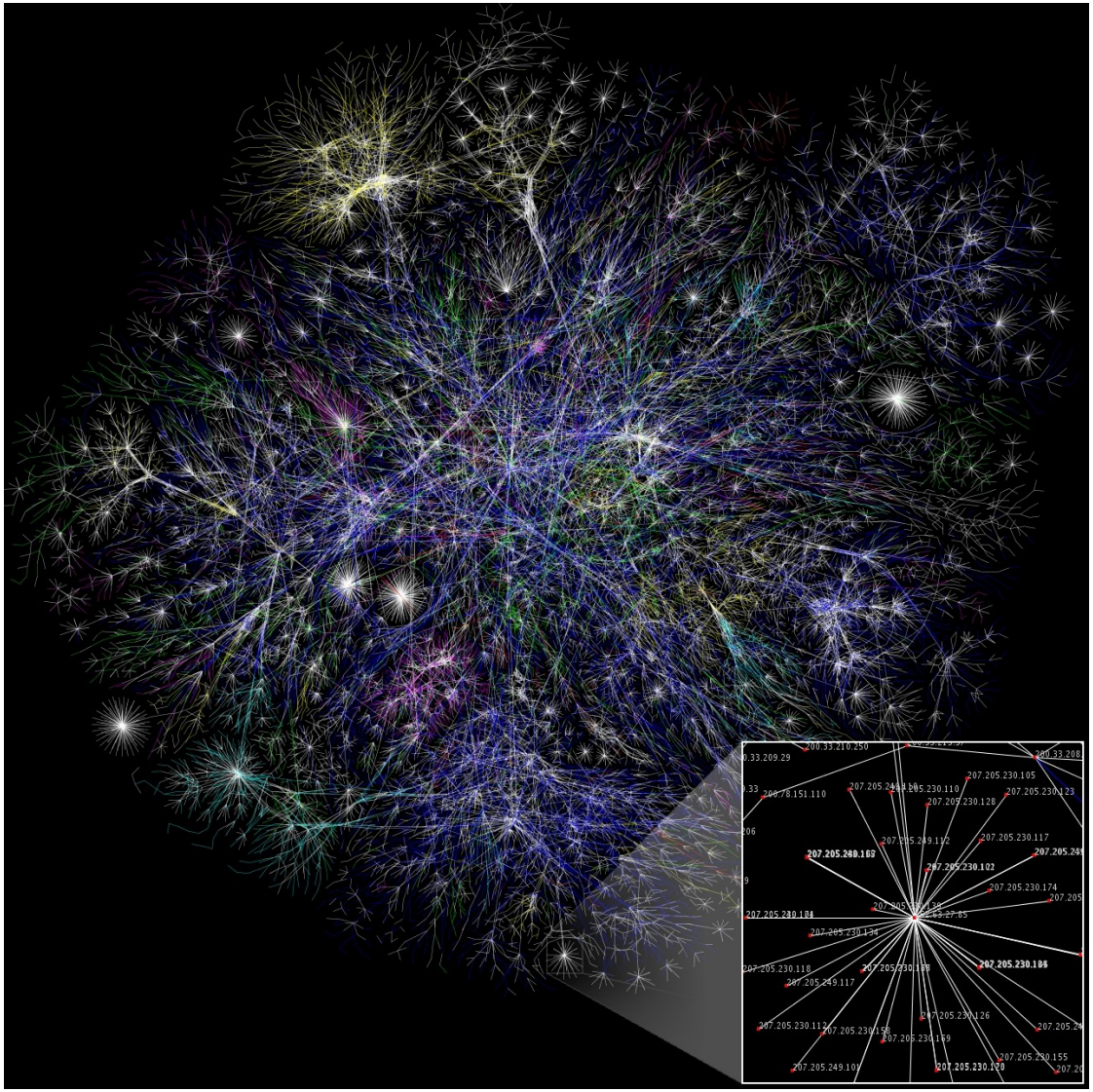


Soziale Netzwerke: Facebook

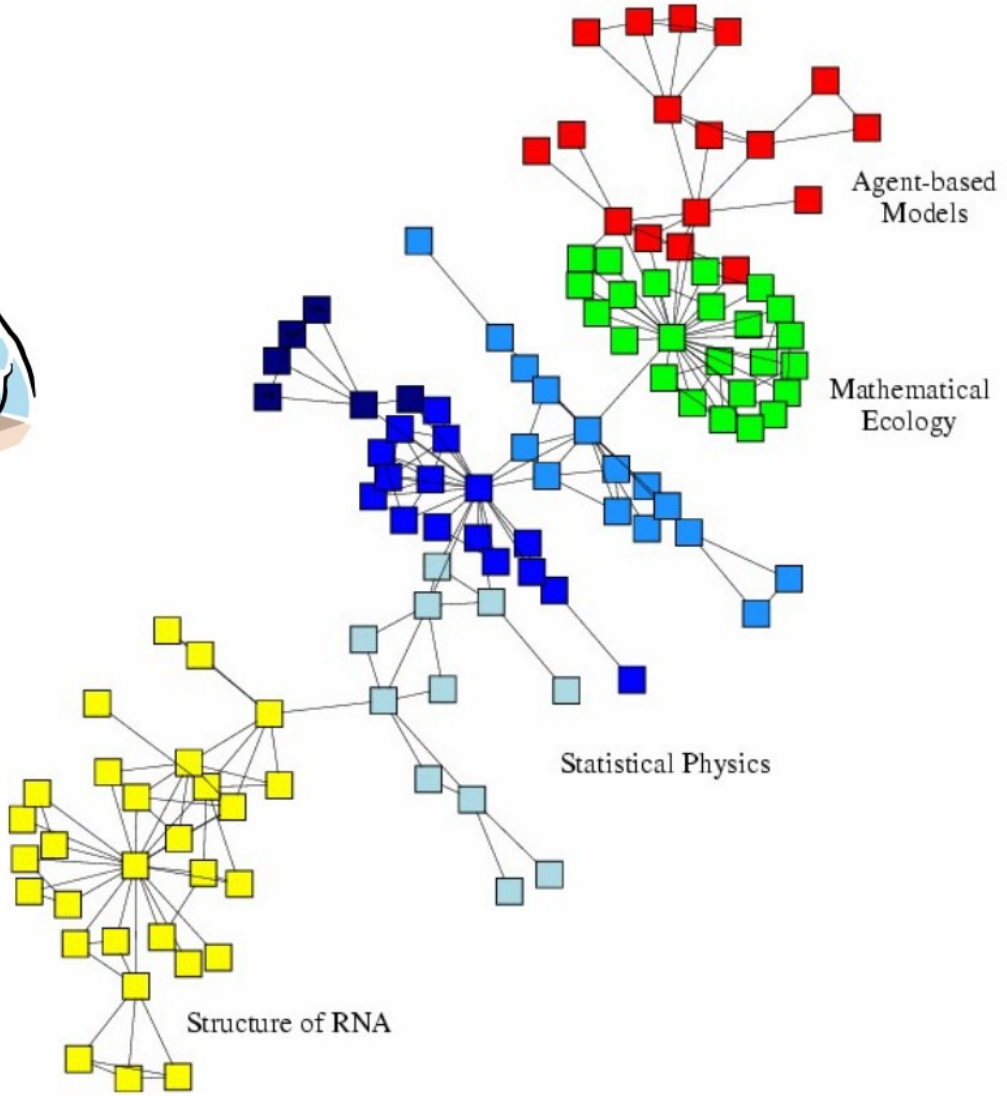


Lost Circles

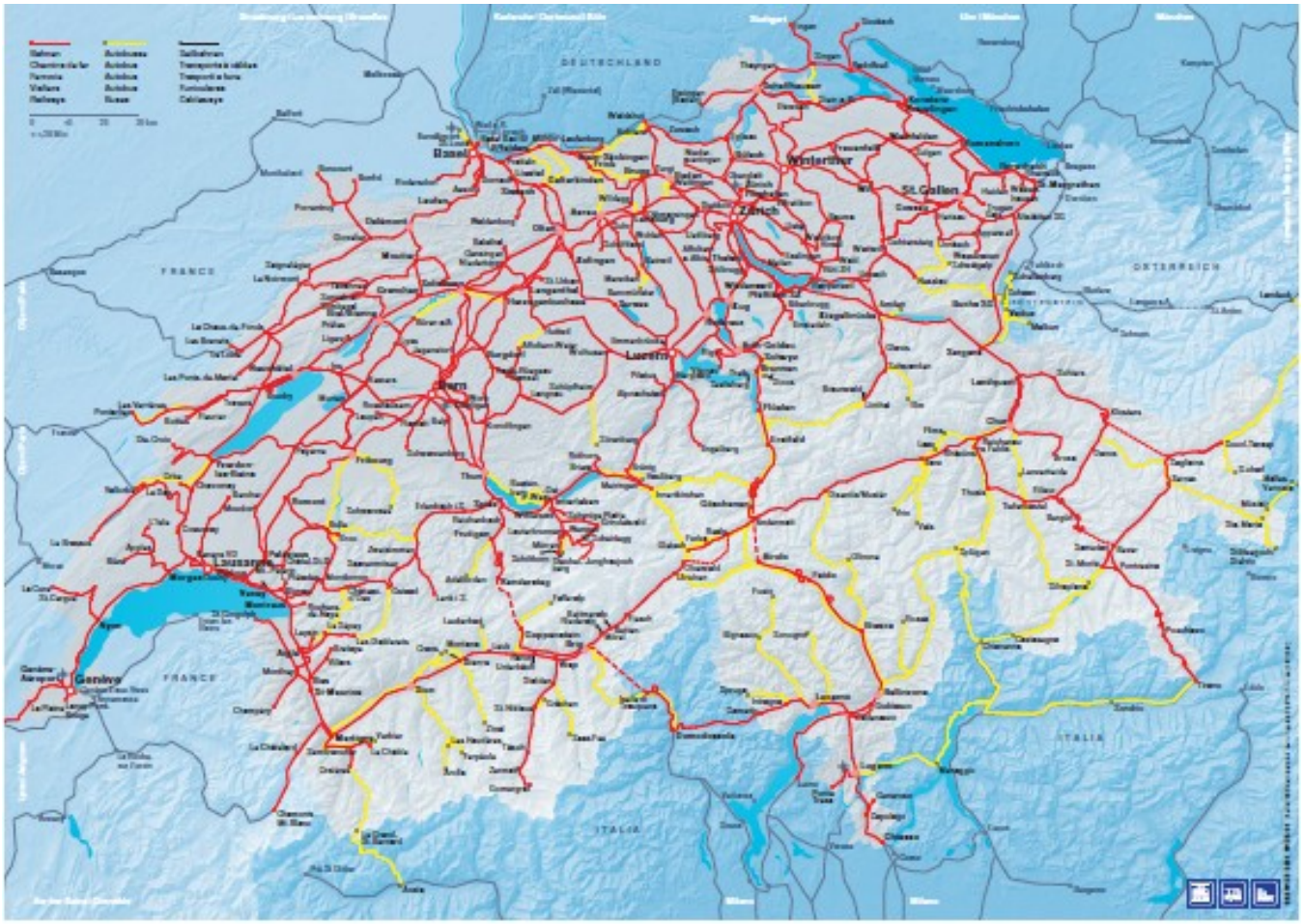
Informationsnetzwerke: Internet



Wissenschaftliche Zusammenarbeit

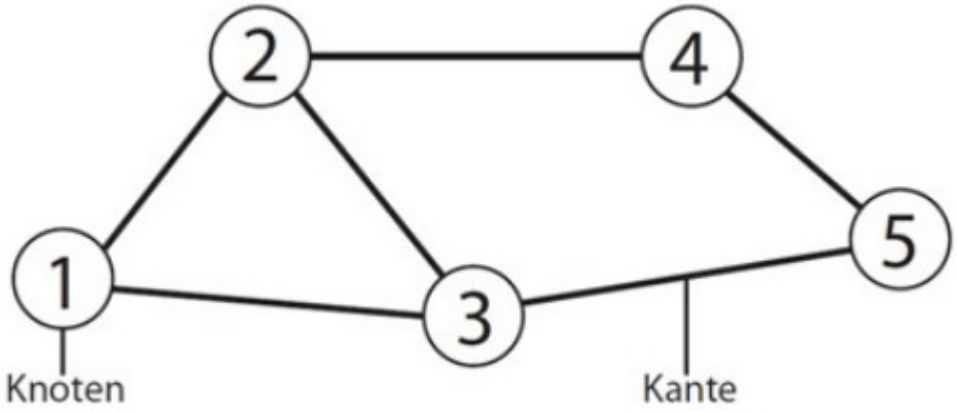


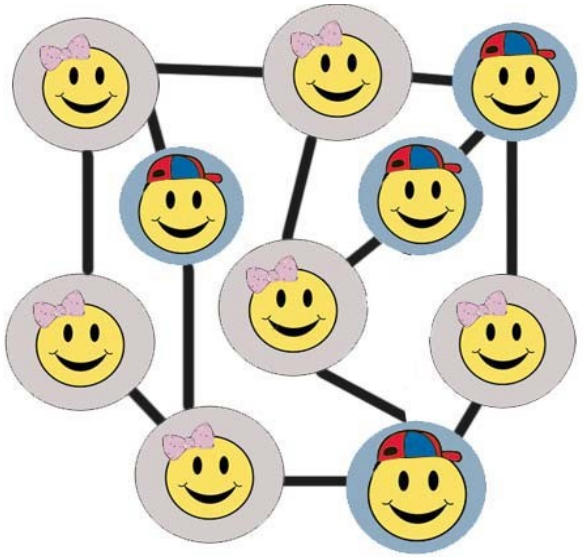
Verkehrsnetze: Schweizer Schienen



Grundbegriffe der Graphentheorie

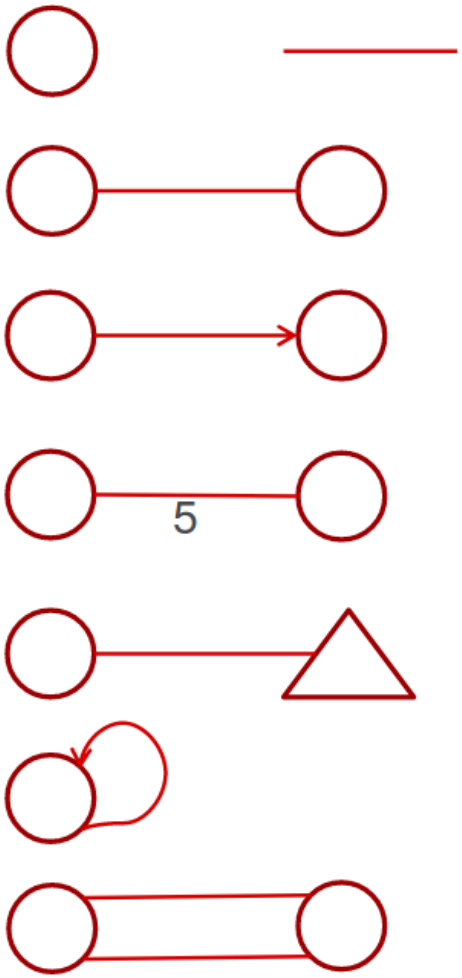
- Netzwerke lassen sich mit Graphen beschreiben
- Mathematisches Modell, bestehend aus Knoten und Kanten
- In einem sozialen Netz entsprechen die Knoten den Individuen einer Gesellschaft, die Kanten repräsentieren ihre Beziehungen





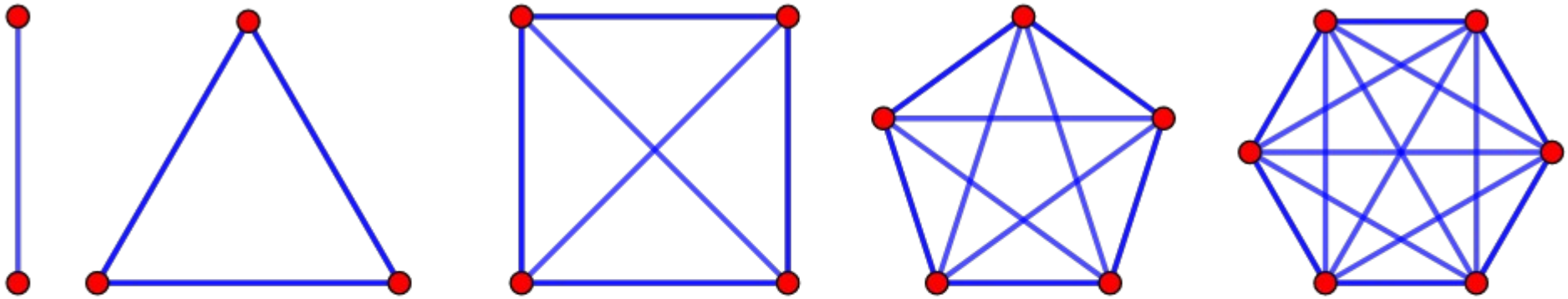
Netzwerke
Graphen

- Knoten und Kanten
- Ungerichtete Graphen
- Gerichtete Graphen (digraph)
- Gewichtete Graphen
- Bipartite Graphen
- Schleife (Loop)
- Mehrfachkanten (multi-edges)

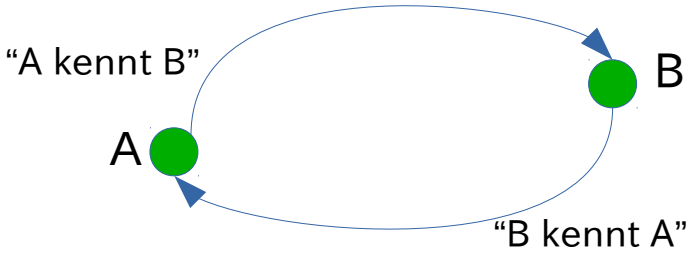


Einfache Graphen

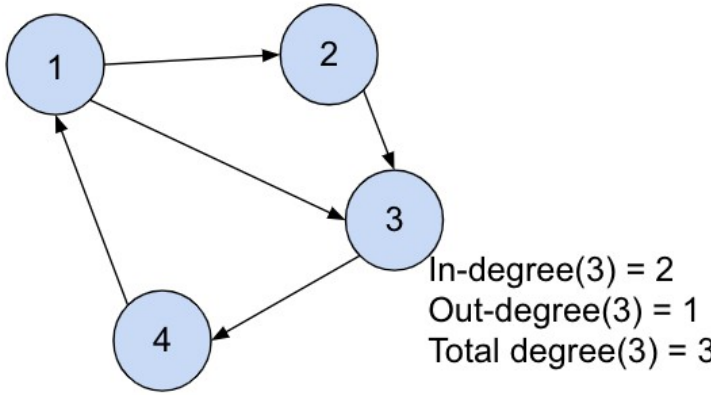
- Ein Graph heisst einfach (simple, strict) wenn er
  - Ungewichtet und
  - Ungerichtet ist und
  - Keine Schleifen hat und
  - Keine Mehrfachkanten aufweist
  - Ein Complete Graph bezeichnet einen einfachen Graph, in dem jeder Knoten mit jedem anderen Knoten durch eine Kante verbunden ist.

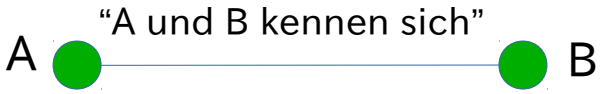


Knotengrade

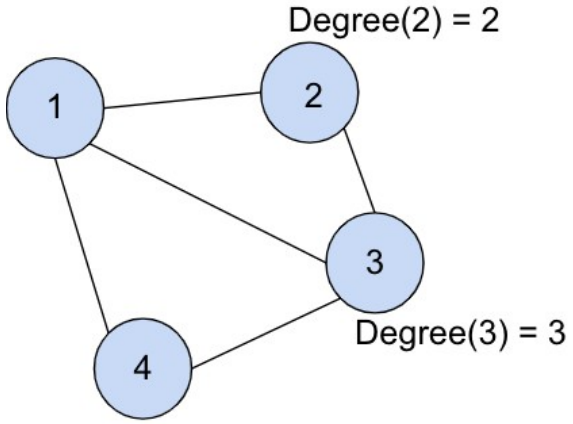


gerichtet





ungerichtet

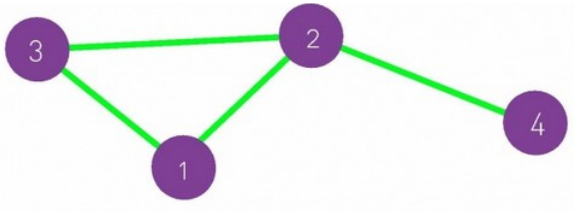


Average Degree

Undirected

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

Anzahl Links = Summe der Knotengrade

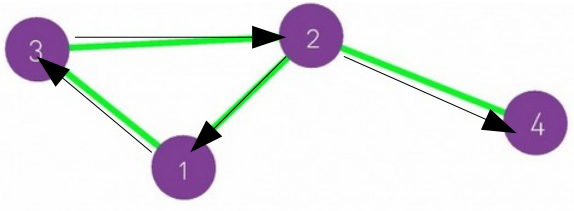


$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

Directed

$$k_i = k_i^{in} + k_i^{out}$$

Total Degree

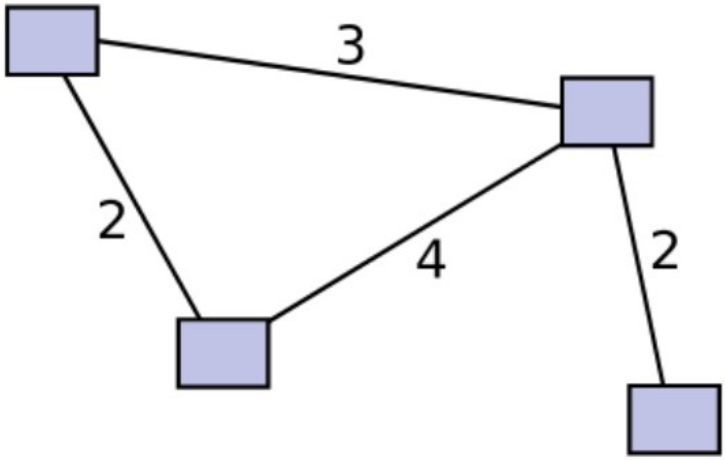


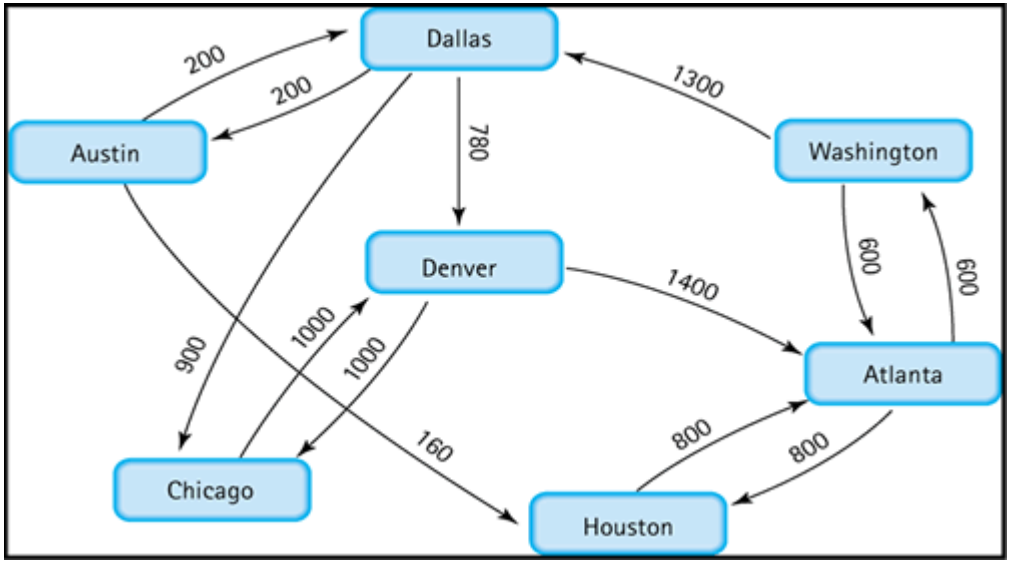
$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$$

Anzahl Links

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

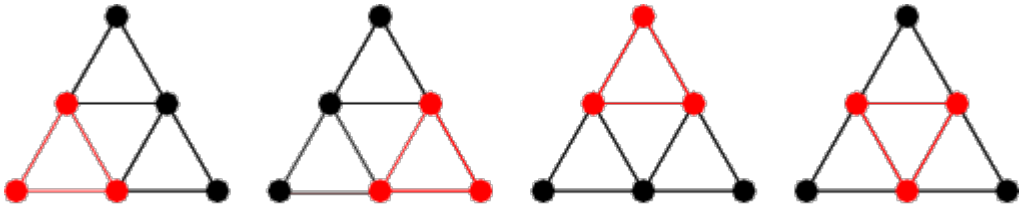
Gewicht


3
2
4
2

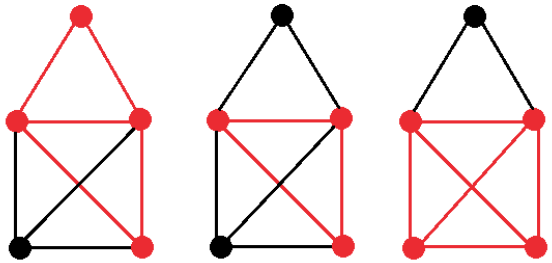

Dallas
Austin
Washington
Denver
Atlanta
Chicago
Houston
200
200
1300
780
600
600
1400
1000
1000
900
160
800
800

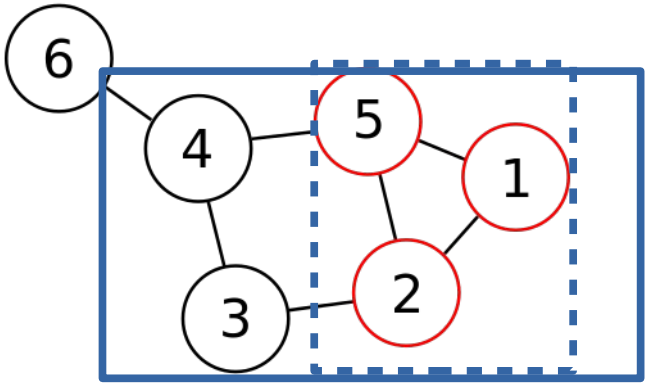
Clique

- Teilmenge von Knoten in einem **ungerichteten Graphen**, bei der jedes Knotenpaar durch eine Kante verbunden ist (vollständiger Teilgraph)



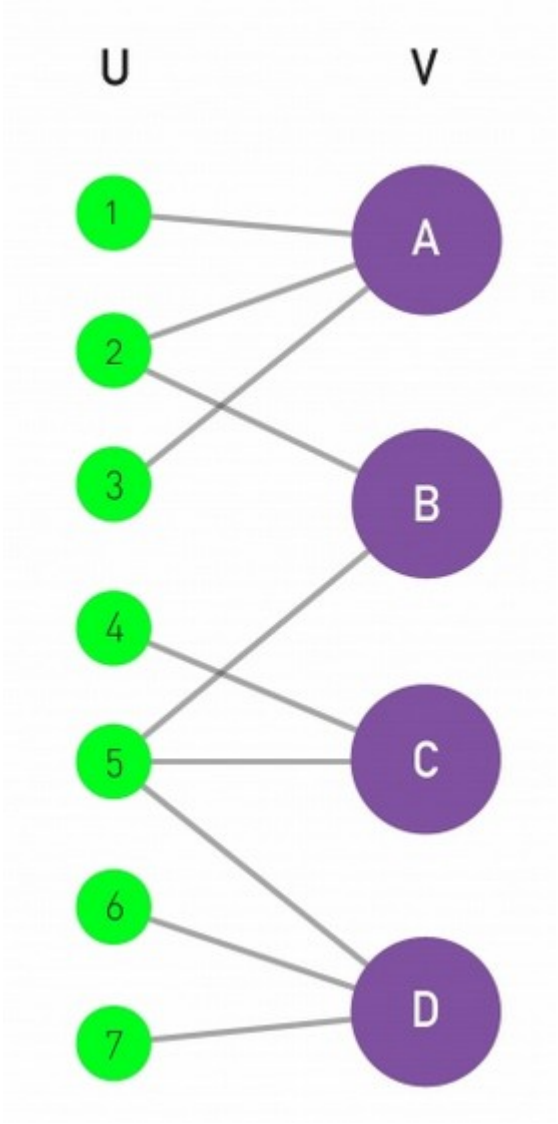
cliques



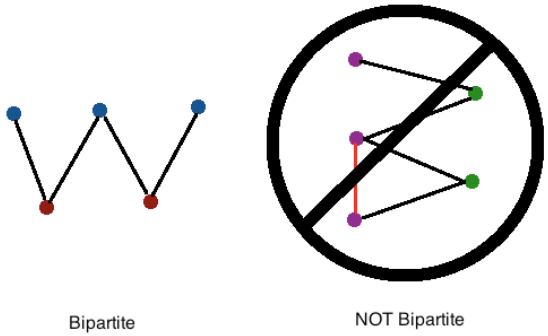


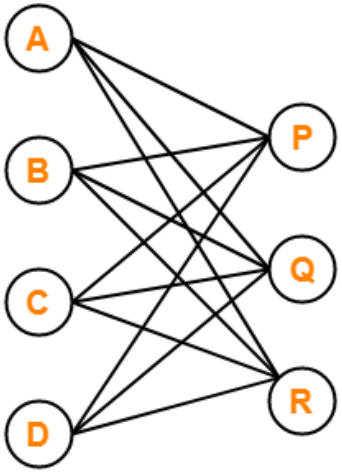
Clique der Grösse 3
+
Clique der Grösse 5

Bipartite Networks



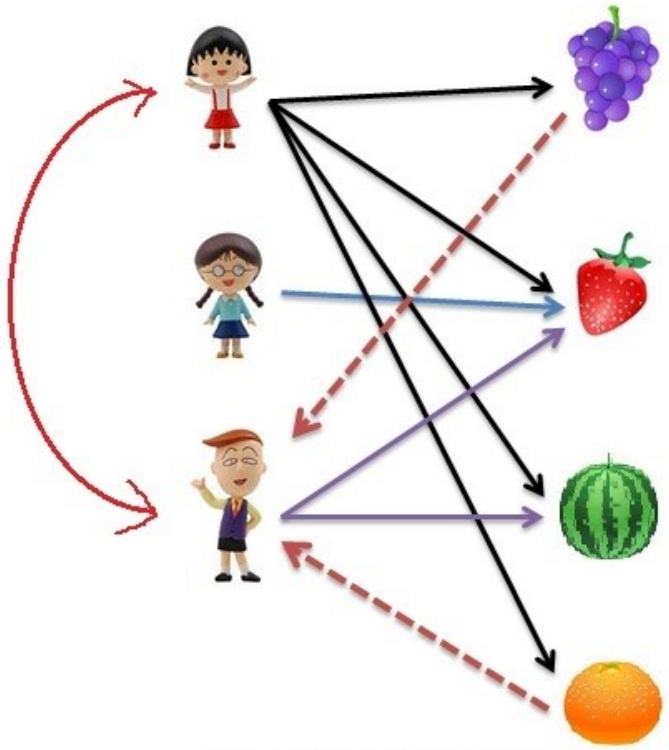
Keine Kanten, zwischen set U und set V



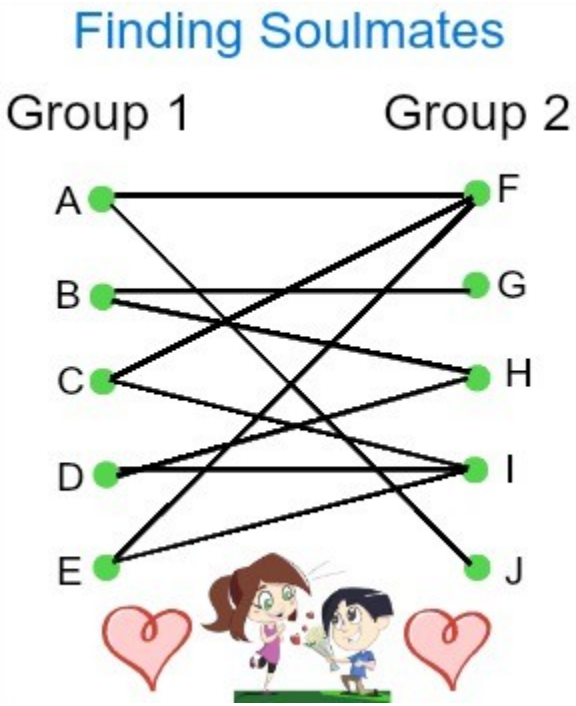


Example of Complete Bipartite Graph

Bipartite Networks - Beispiele

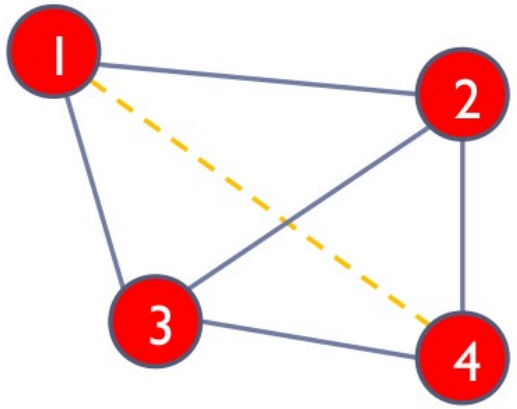


User-based filtering

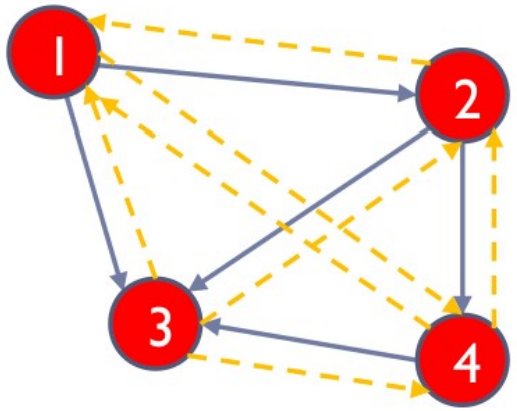


Dichte eines Graph
Density

- Anzahl vorhandener Kanten / Anzahl möglicher Kanten
- Ein perfekt verbundenes Netzwerk = Clique, Dichte = 1
- Ein gerichteter Graph hat die hälfte Dichte seiner ungerichtetes Äquivalent



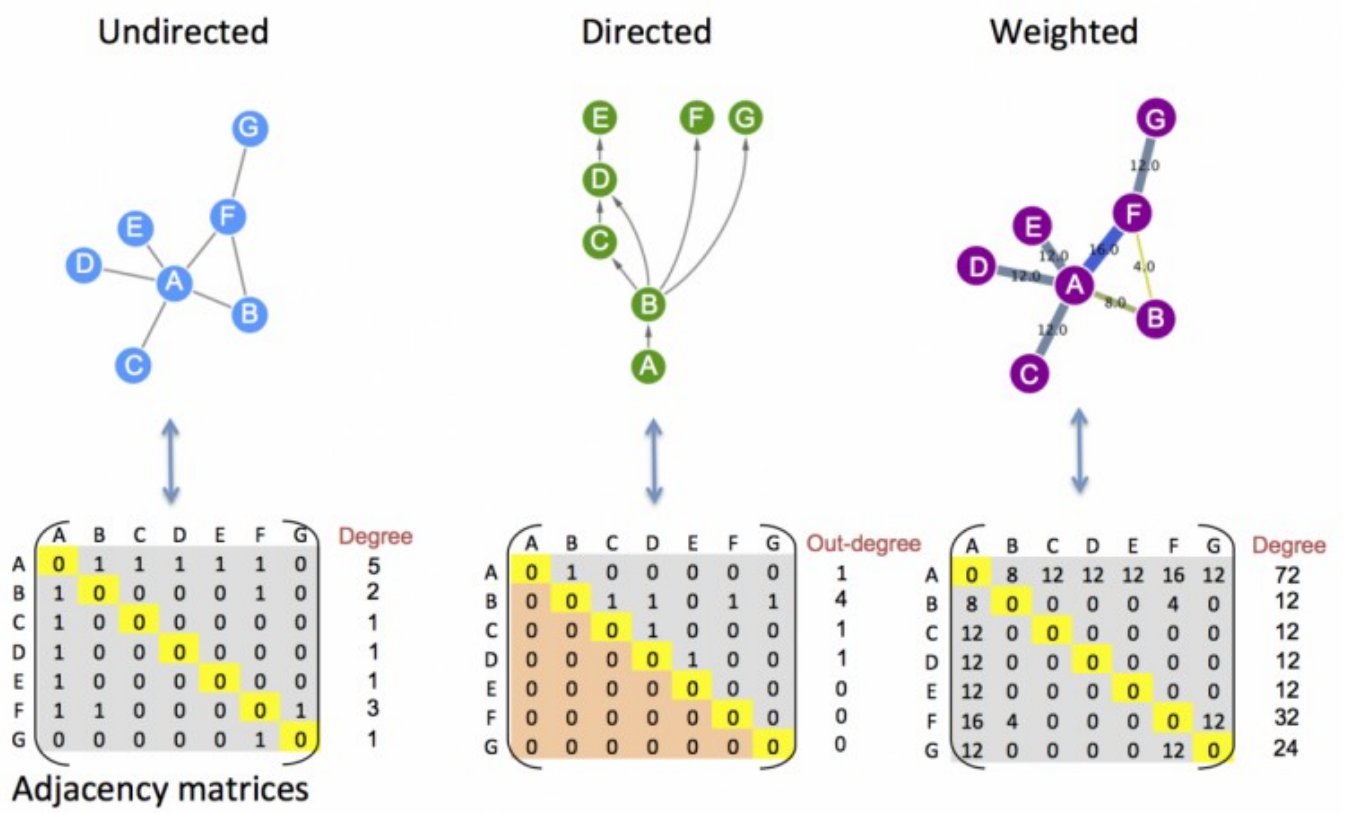
density = 5/6 = 0.83



density = 5/12 = 0.42

Adjazenz-Matrizen

- Darstellung eines Graphs als quadratische Matrix
- Kante zwischen mi und mj → dann ist mij = 1

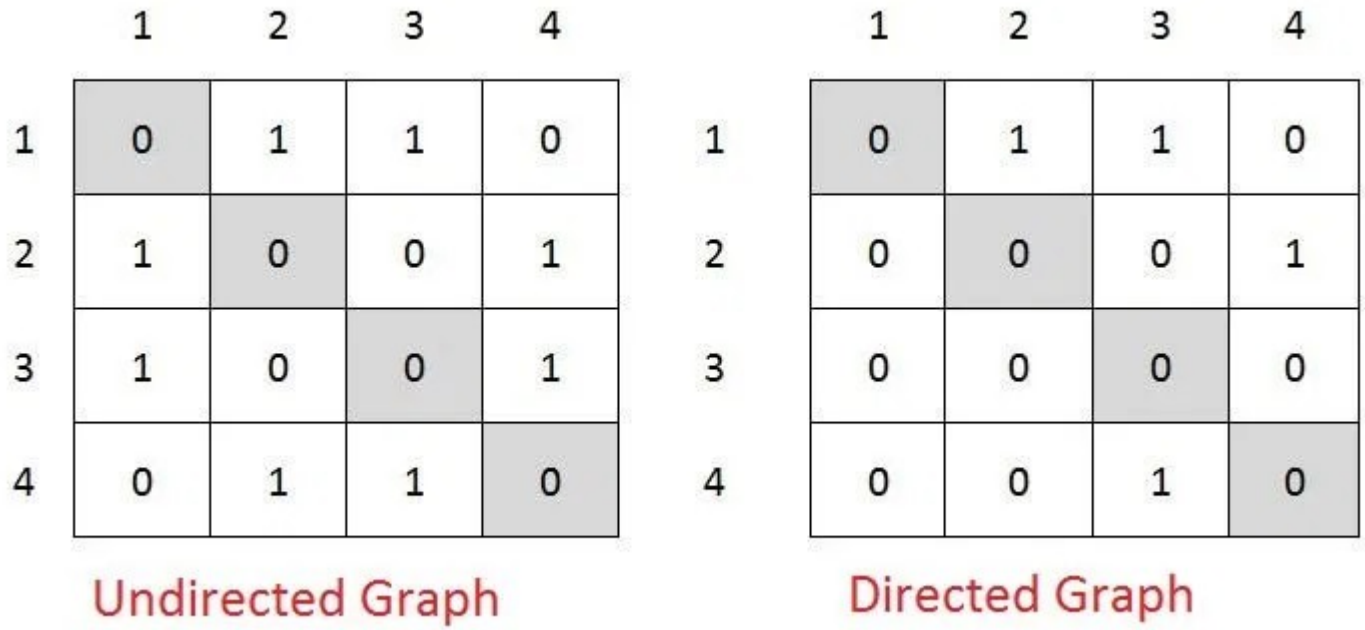


	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

	A	B	C	D	E	F	G	Out-degree
A	0	1	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

Adjazenz-Matrizen



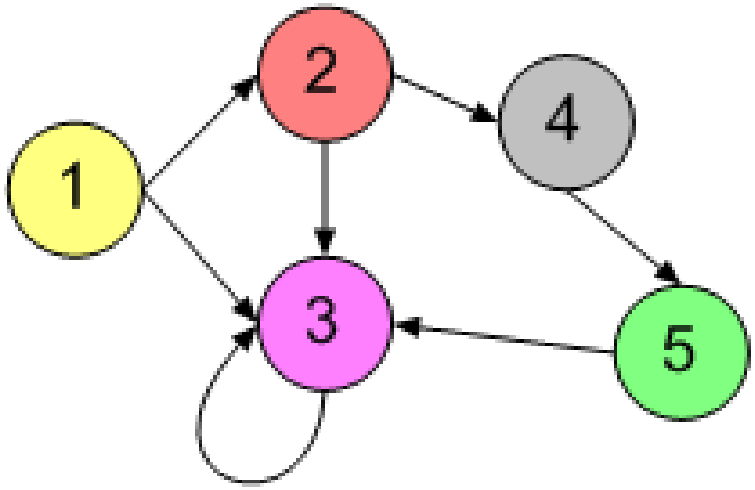
	1	2	3	4
1	0	1	1	0
2	1	0	0	1
3	1	0	0	1
4	0	1	1	0

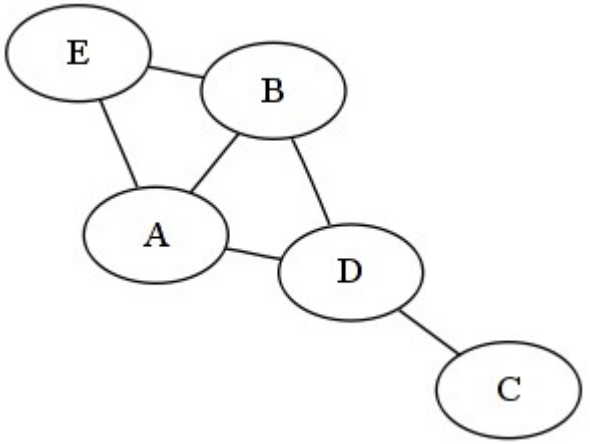
Undirected Graph

	1	2	3	4
1	0	1	1	0
2	0	0	0	1
3	0	0	0	0
4	0	0	1	0

Directed Graph

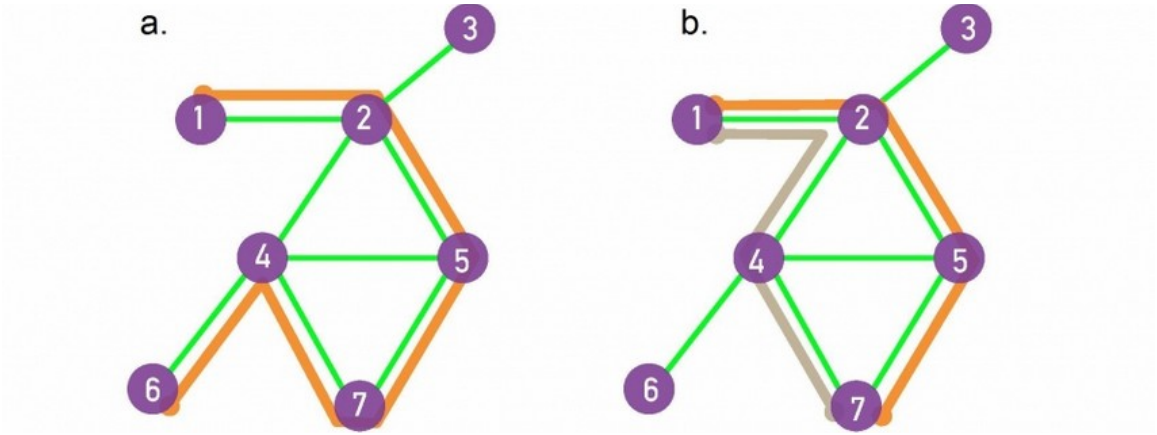
Adjazenz-Matrizen



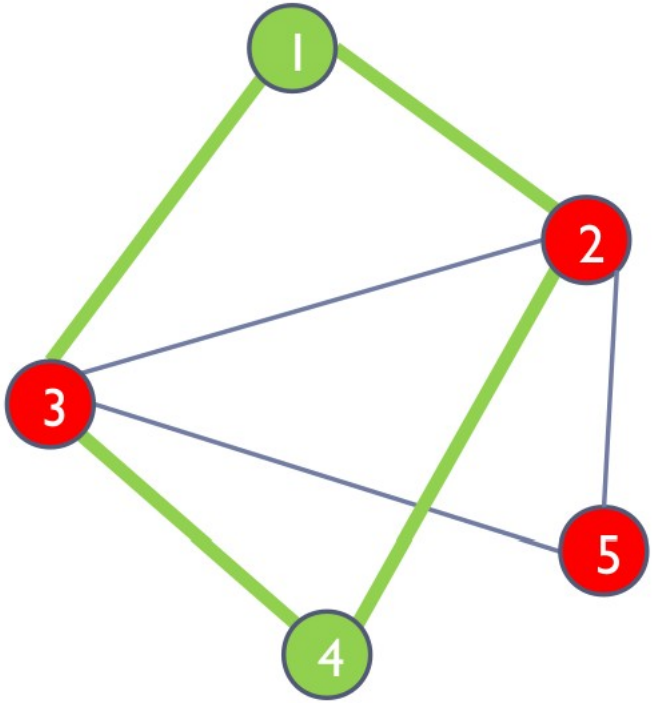


Pfade und Pfadlängen

- Pfad = Verbindung zweier Knoten in einem Graphen (geordnete List von Links)
- Pfadlänge: Anzahl der Kanten in einem Pfad
- Ein kürzester Pfad zwischen zwei Knoten ist ein Pfad mit minimaler Länge.
- Die durchschnittliche Pfadlänge ist die durchschnittliche Distanz zwischen allen Knotenpaaren im Netzwerk.

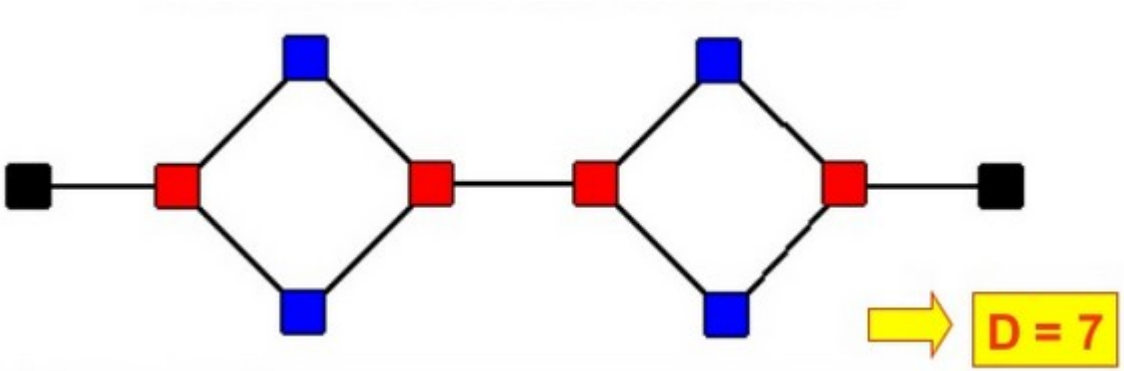


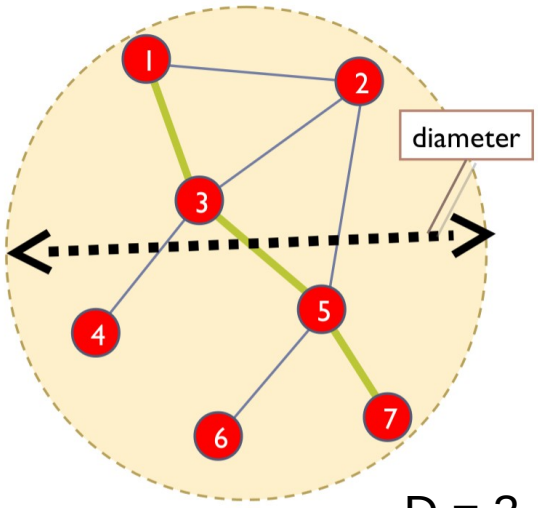
Kürzeste Pfade



Durchmesser

- Der Durchmesser eines Netzwerks ist der längste kürzeste Pfad.

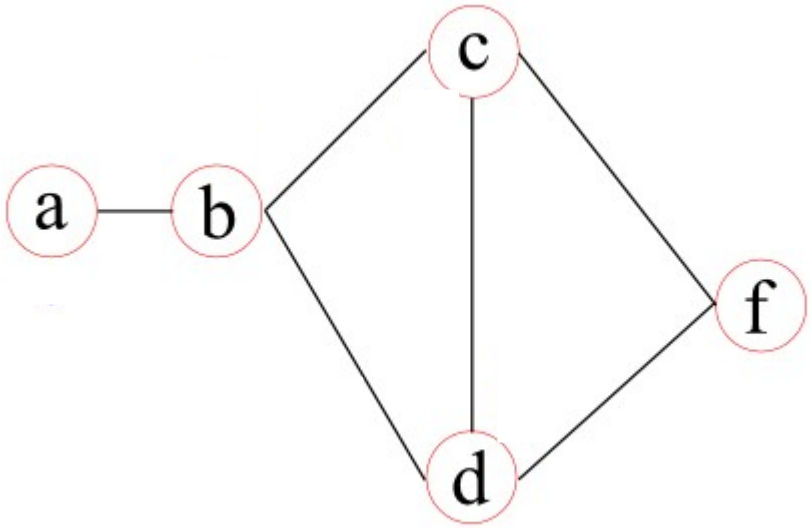




D = 3

Cluster-Coefficient

- Er beschreibt, wie stark Nachbarknoten untereinander vernetzt sind
- Anzahl verbundene Nachbarn /Mögliche Nachbarverbindungen



Cluster-Coefficient

"Unclustered" network

None of Ego's friends know each other*

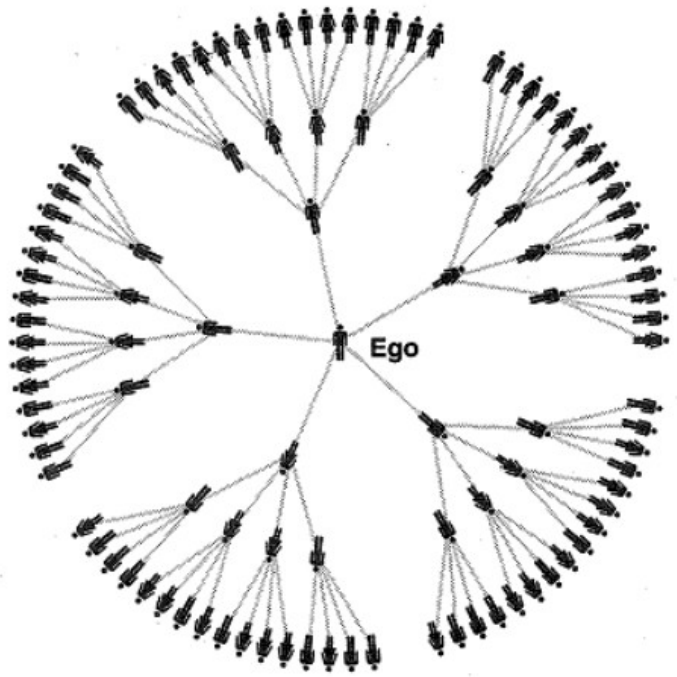


Figure 1.2. A pure branching network. Ego knows only 5 people, but within two degrees of separation, ego can reach 25; within three degrees, 105; and so on.

"Clustered" network

All of Ego's friends know each other

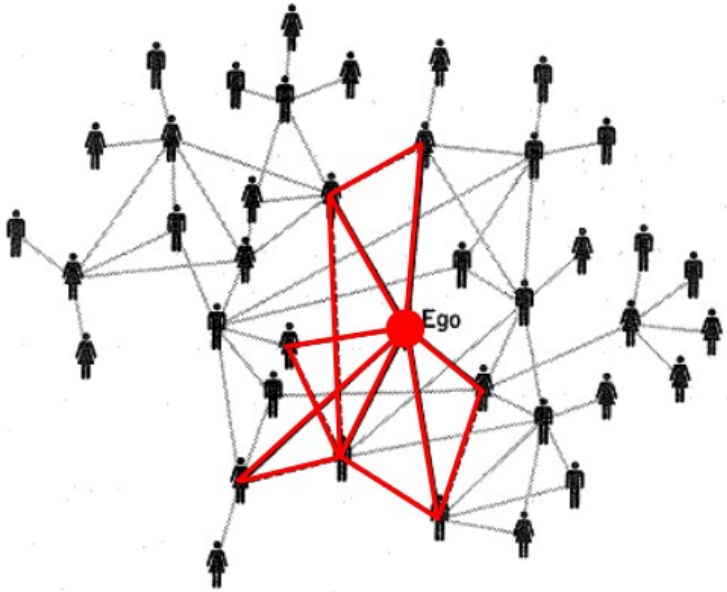


Figure 1.3. Real social networks exhibit clustering, the tendency of two individuals who share a mutual friend to be friends themselves. Here, Ego has six friends, each of whom is friends with at least one other.

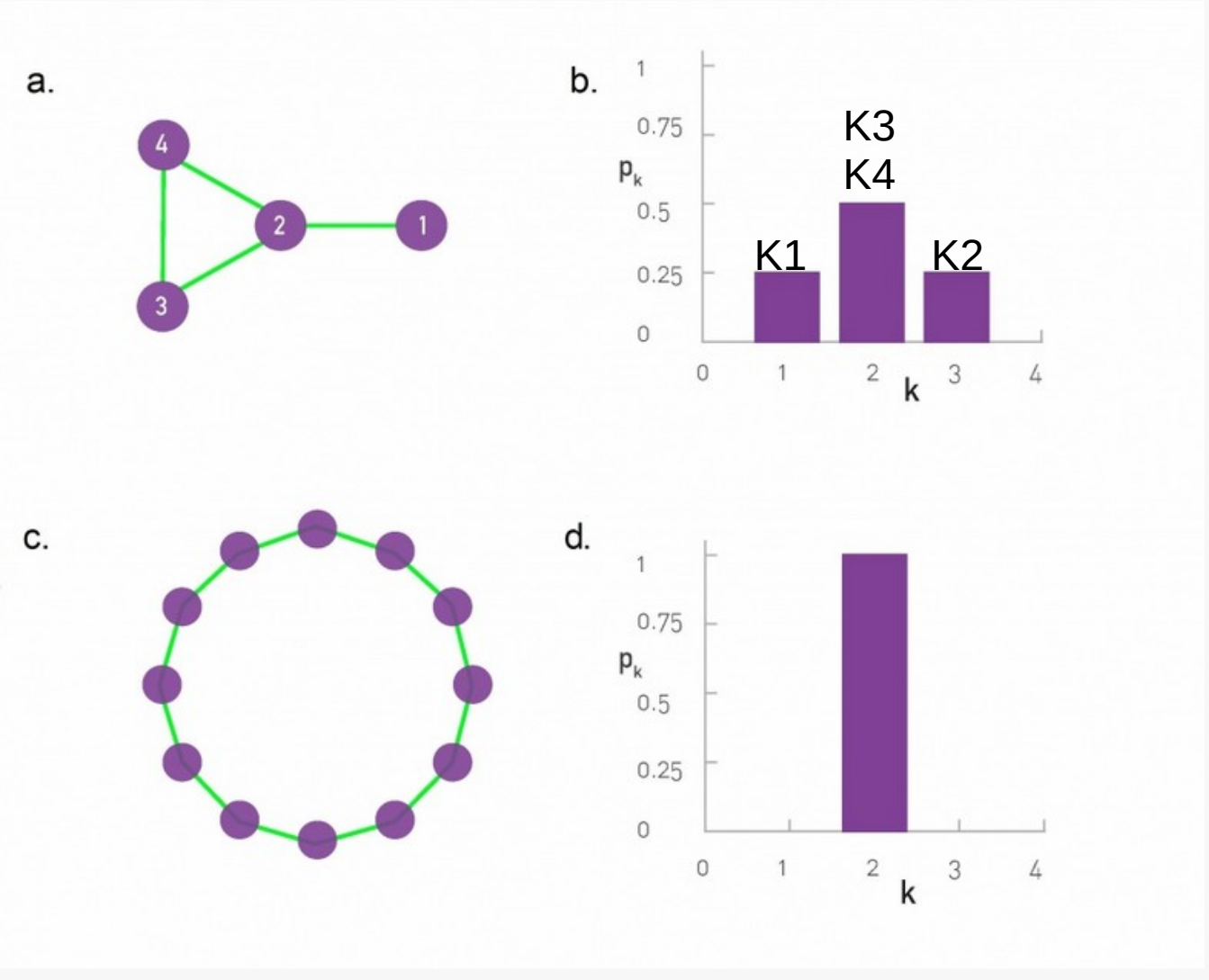
Gradverteilung

- Für ein Netzwerk ist die Gradverteilung P(k) definiert durch:

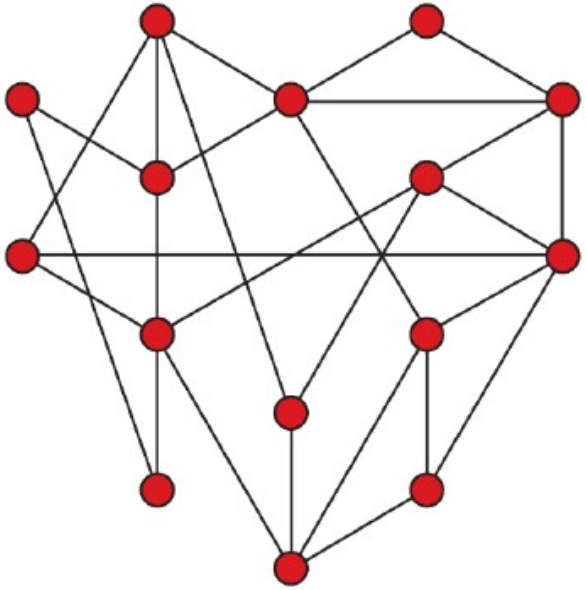
$$P(K) = \text{Prob}(\text{Knoten hat genau Grad } k)$$

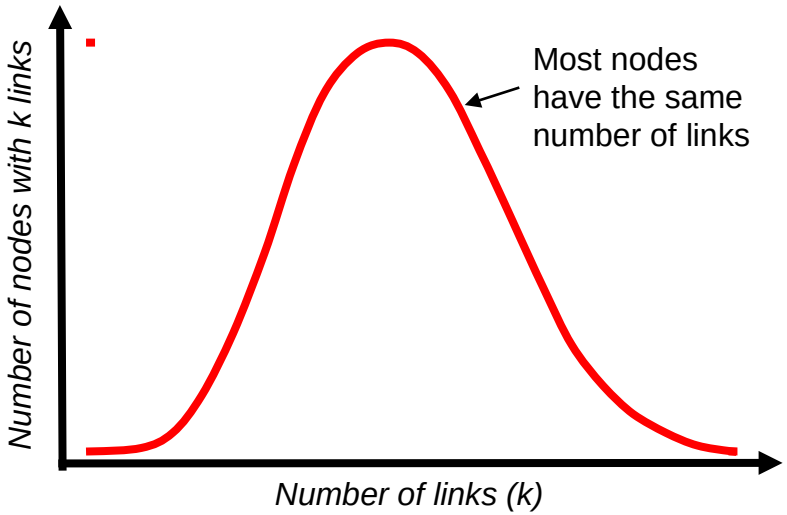
- Anhand der Gradverteilung können wie Aussagen über die Struktur des Netzwerkes treffen:
- Sind alle Knoten gleichartig?
- Wie stark ist die Verbundenheit des Netzes?

Gradverteilung

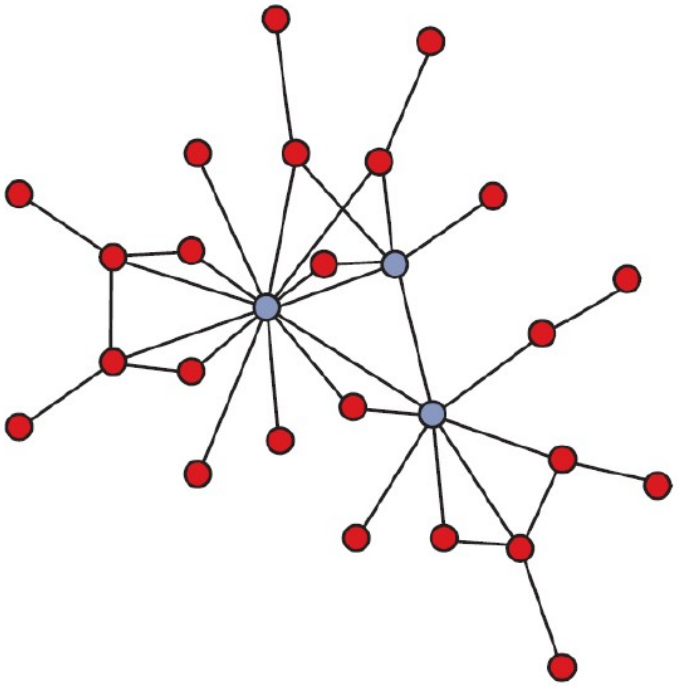


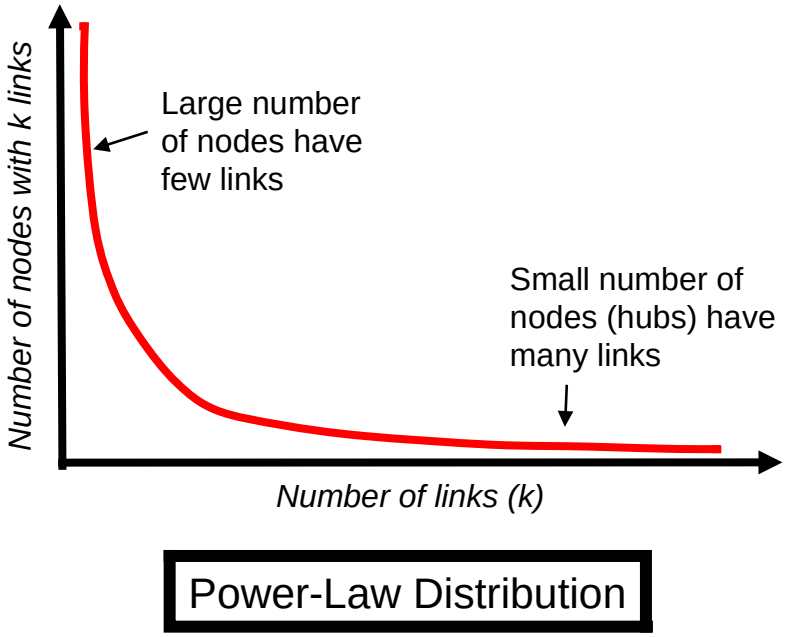
Zufallnetzwerk





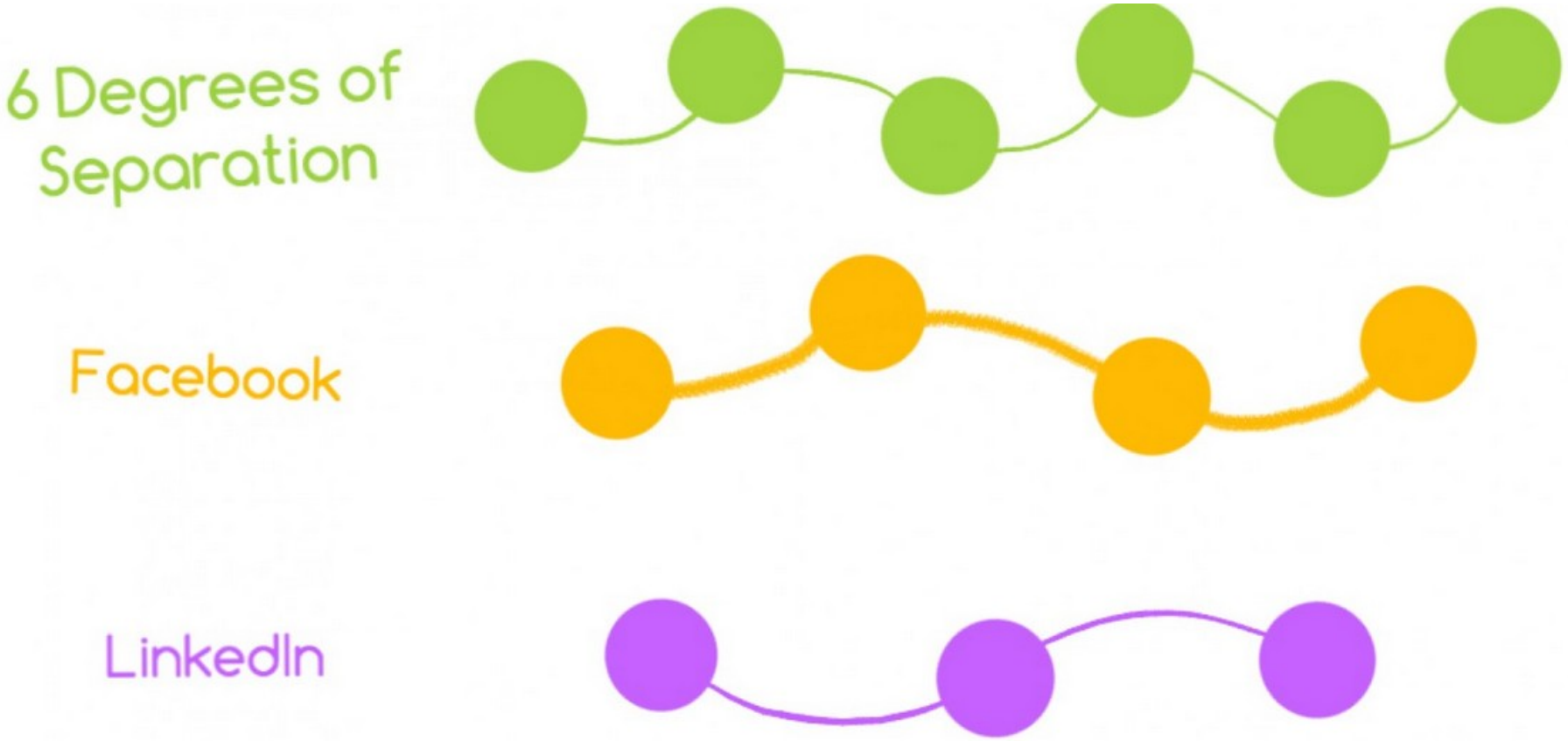
Skalenfreie Netzwerke





Power-Law Distribution

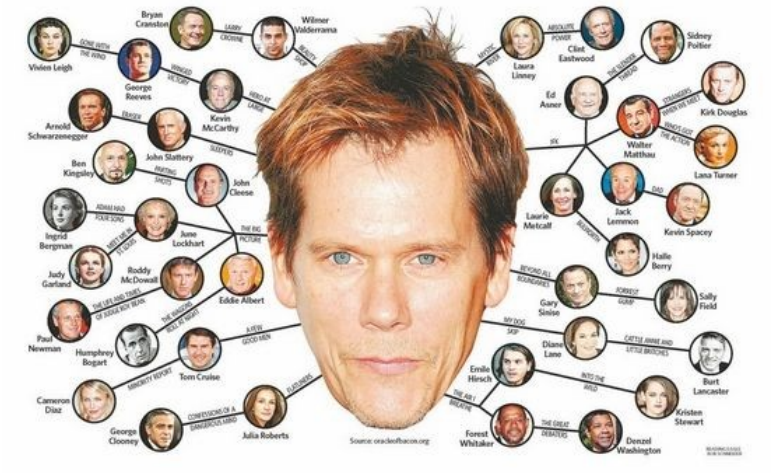
Beispiel: “six degrees of separation”



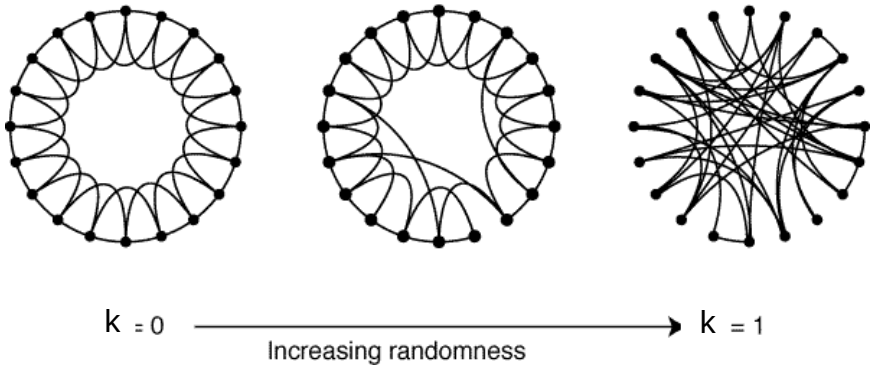
Beispiel: “six degrees of separation”

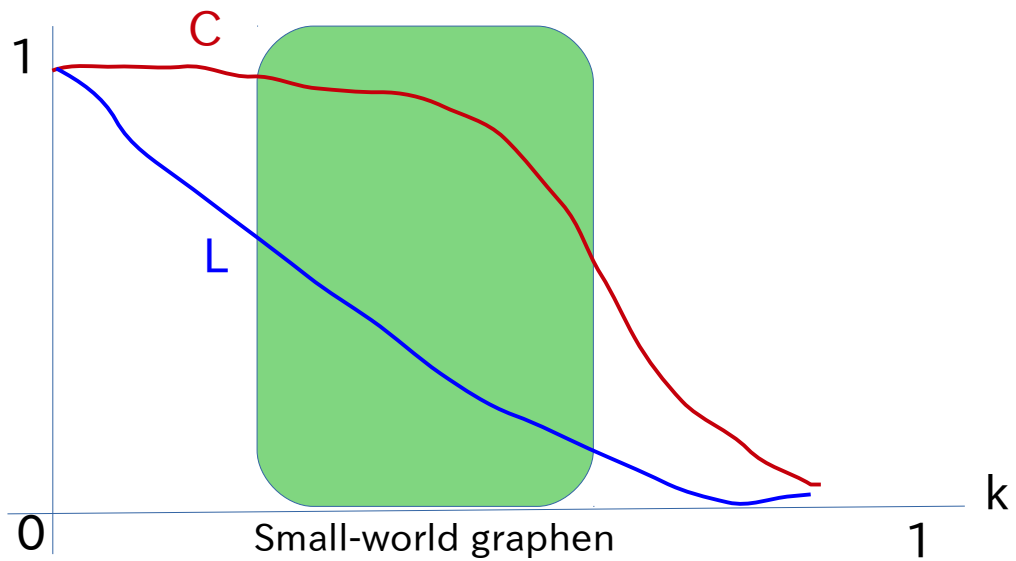
- Stanley Milgram: Paket-Experiment (1967):
- Frage: Wie weit sind Menschen voneinander entfernt?
- Aufbau: Zufällig ausgewählte Menschen geben einen Brief nur über Bekannte an eine andere Person weiter
- Ergebnis: Alle Menschen sind rund 6 Schritte voneinander entfernt

Soziale Netzwerke sind enger, als man vermuten würde.
Warum?



Small-World Modell





- Übergang von einer regulären Struktur zu einem Zufallsgraph
- Dabei werden Kanten zufällig neu verknüpft
- Der Parameter k differenziert zwischen geordnteten und zufälligen Graphen

VIDEO

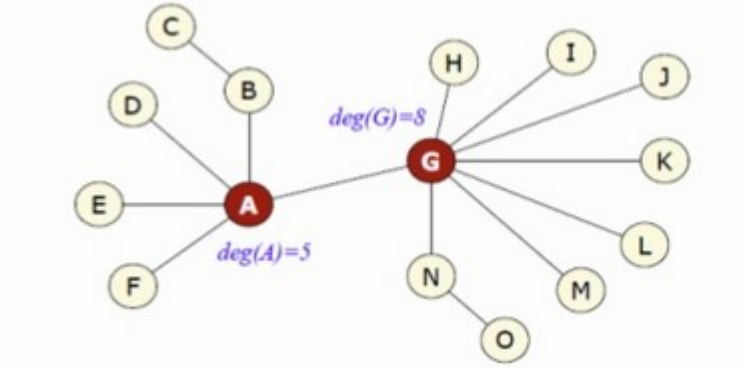
Zentralitätsmassen

- Grad-basierte Zentralität (degree centrality)
- Nähe-basierte Zentralität (closeness centrality)
- Zwischenzentralität (betweeness centrality)

VIDEO

Degree Centrality: “Wer hat viele Freunde?”

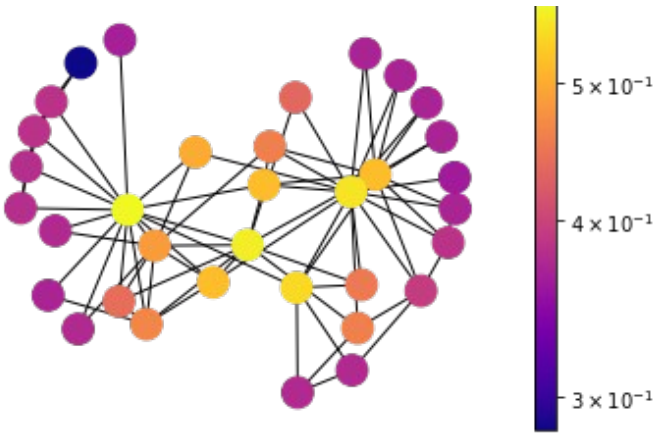
- Die Anzahl der direkten Beziehungen, die ein Knoten hat
- Ein Knoten mit hoher Degree Centrality:
  - Ist ein aktiver Spieler
  - Ist oft ein Anschluss oder Hub

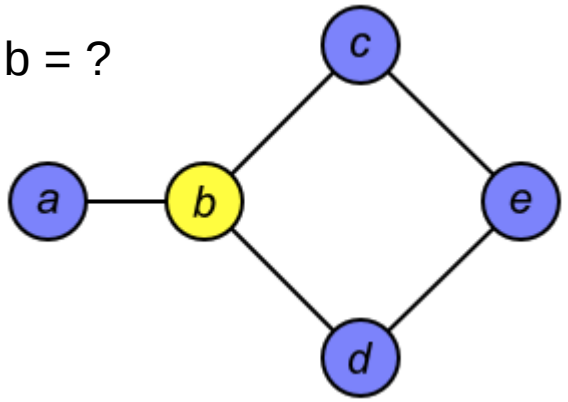


Closeness Centrality: ”Wo sind die Hubs?”

- Misst, wie schnell ein Knoten auf andere Knoten zugreifen kann
- Ein Knoten mit hoher Closeness Centrality:
  - Hat einen schnellen Zugang zu anderen Knoten (kurzen Weg)
- Die Closeness Centrality eines Knotens n ist definiert als der Kehrwert der durchschnittlichen kürzesten Weglänge.

$$Cc(n) = 1 / avg(L(n,m))$$



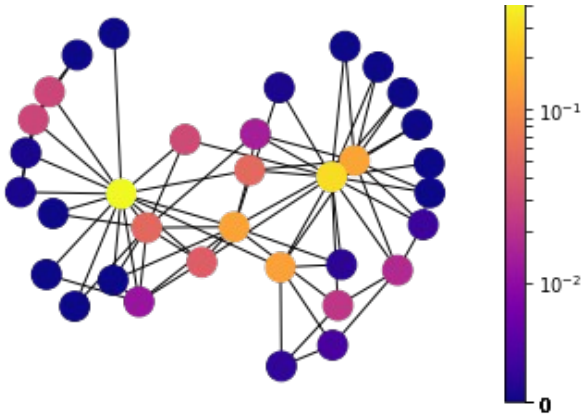


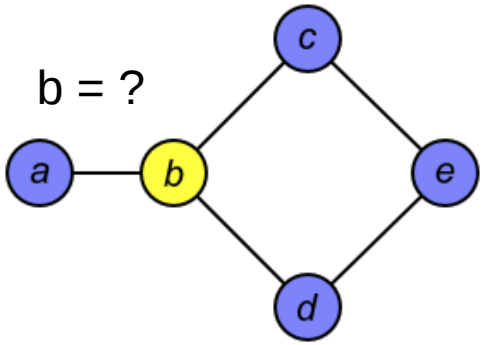
Betweeness Centrality: “Wo sind die Brücke?”

- Identifiziert die Position eines Knotens in Bezug auf sein Fähigkeit, Verbindungen zu anderen Gruppen herzustellen
- Ein Knoten mit hoher Betweeness:
  - Hält eine beliebte oder starke Position
  - Grossen Einfluss darauf, was im Netzwerk passiert
- Wie viele kürzeste Pfade, die nicht an einem bestimmten Knoten beginnen oder enden, durch ihn durchgehen.

$$Cb(n) = \sum s \neq n \neq t \; (\sigma st \; (n) \; / \; \sigma st)$$

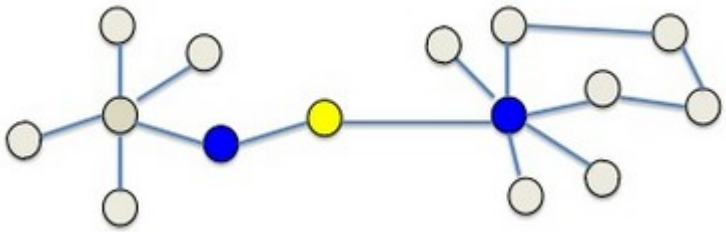
$$\text{Betweenness Centrality of } \textit{node X} = \sum_{\text{for all pairs of nodes}} \frac{\text{fraction of shortest paths that go through } \textit{node X}}{\text{all shortest paths between every node}}$$

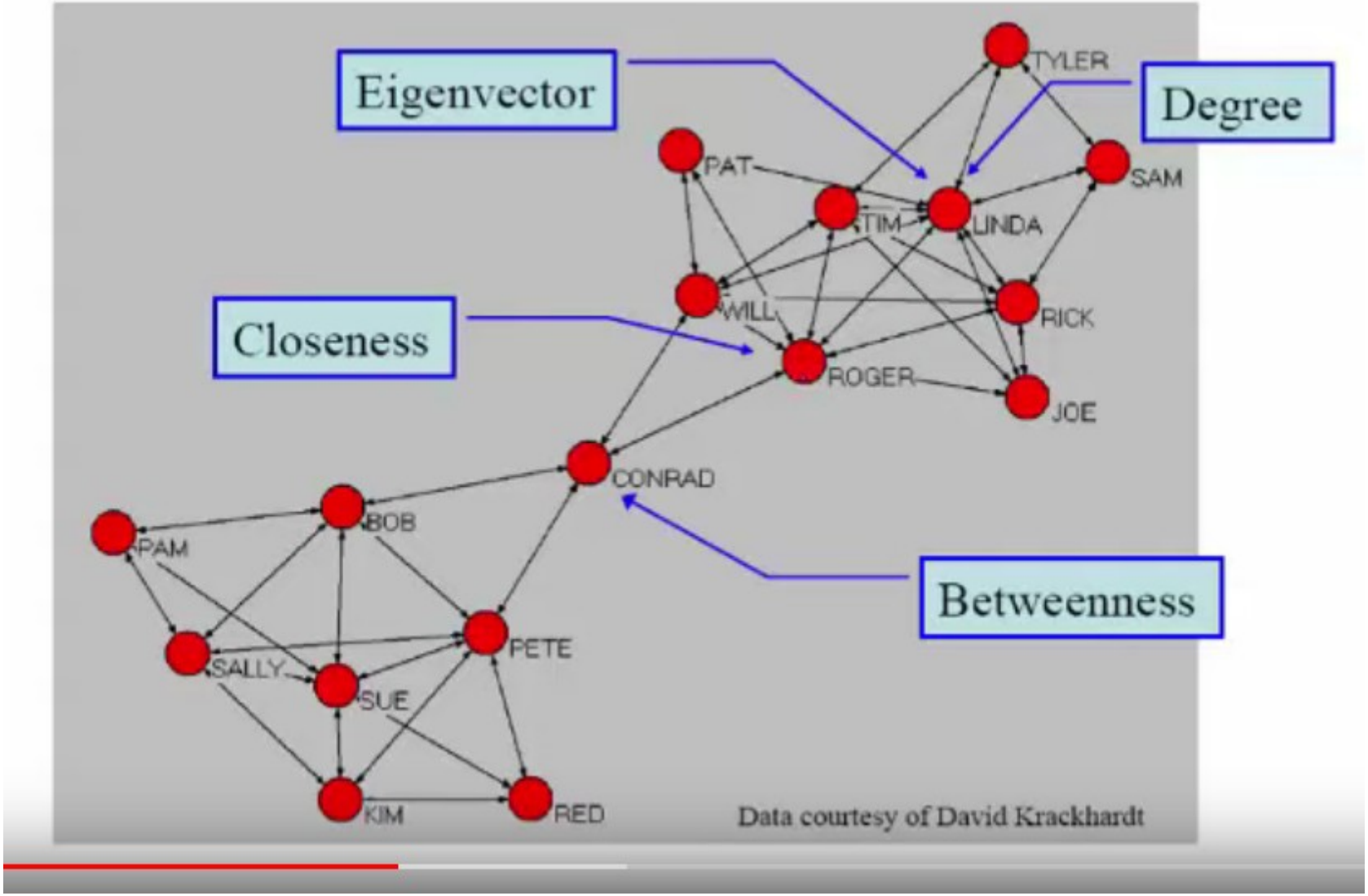




Eigenvalue Centrality: “Wer hat am meinsten Einfluss?”

- Eigenvalue Centrality misst, wie viel Einfluss ein Knoten im ganzen Netzwerk hat, unter Berücksichtigung seiner Verbindungen zu anderen hochverbundenen Knoten.
- Eigenvektor-Zentralität berechnet wie gut ein Knoten mit den Teilen des Netzwerks mit der besten Konnektivität verbunden ist
- “Könige des Netzes”: sia haben vielleicht kein grosse Closeness oder Betweeness, aber sie haben viel Einfluss
- PageRank ist auf Eigenvalue Centrality basiert




Eigenvector
Degree
Closeness
Betweenness
TYLER
PAT
SAM
TIM
LINDA
WILL
RICK
ROGER
JOE
CONRAD
BOB
PAM
PETE
SALLY
SUE
KIM
RED
Data courtesy of David Krackhardt

PageRank

- PageRank verwendet, um **Websites** in den Suchergebnissen von Google **zu bewerten**.
- Wichtige Seiten mit grösserer Wahrscheinlichkeit ein höheres Volumen an Links von anderen Seiten erhalten.
- **Use-Cases**
  - Personalisierter PageRank ist von Twitter verwendet, um den Usern Empfehlungen für andere Accounts zu geben.
  - PageRank wurde verwendet, um Strassen zu bewerten und den Verkehrsfluss und die Bewegung von Menschen in diesen Bereichen vorherzusagen.
  - PageRank kann als Teil eines Anomalie- oder Betrugserkennungssystems in der Gesundheits- und Versicherungsbranche eingesetzt werden.

PageRank – Beispiel mit 5 Seiten

Initialisierung

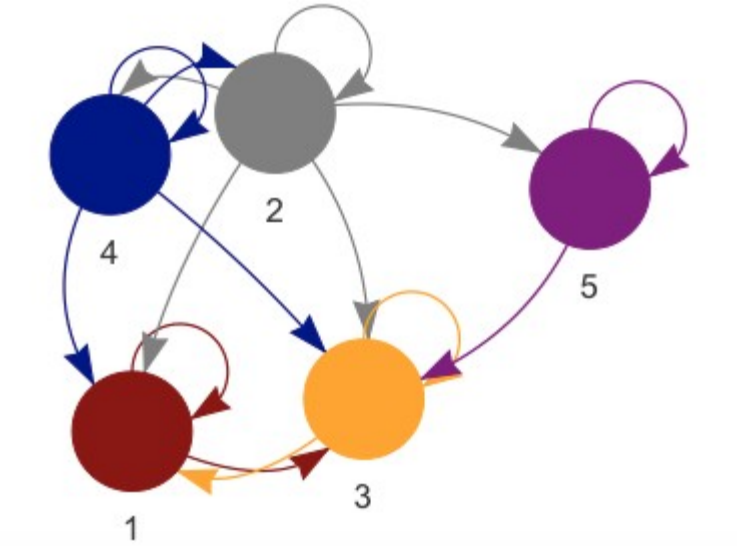
$$\rho(P_1, P_2, P_3, P_4, P_5) = \frac{1}{5}(1, 1, 1, 1, 1)$$

Update

PR(i)

$$\rho(X) = \frac{1-\alpha}{N} + \alpha \sum_{i=1}^{n} \frac{\rho(P_i)}{|P_i|}$$

Outlinks



- Jeder Knoten (Webseite) überträgt die aktuelle PR (Relevanz) gleichmässig auf die Knoten (Seiten), auf die er verlinkt
- Alpha = 0.85
- % der Zeit, in der der Benutzer sich im Web befindet und Links zwischen den verschiedenen Seiten folgt

PageRank – Beispiel mit 5 Seiten

Initialisierung

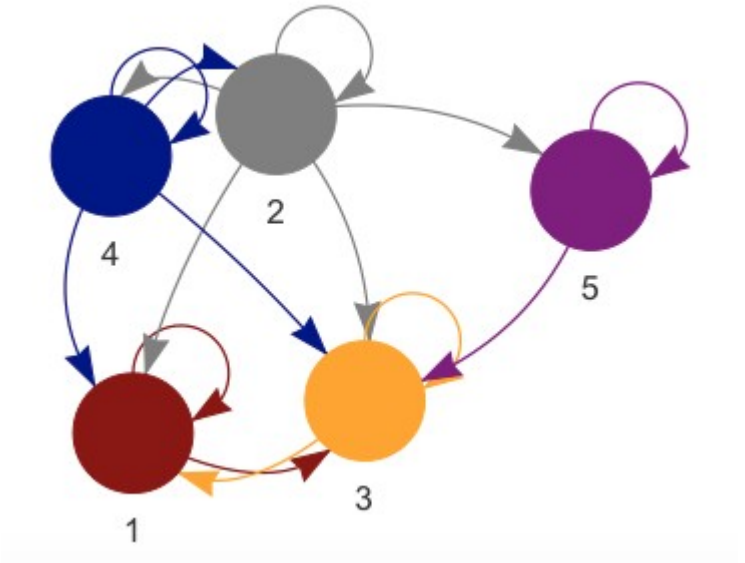
$$\rho(P_1, P_2, P_3, P_4, P_5) = \frac{1}{5}(1, 1, 1, 1, 1)$$

Update

PR(i)

$$\rho(X) = \frac{1-\alpha}{N} + \alpha \sum_{i=1}^{n} \frac{\rho(P_i)}{|P_i|}$$

Outlinks



P1:
#outlinks = 2
P1=[P1,P2,P3,P4]
PR1=0.03+0.85*((1/5)/2+(1/5)/5+(1/5)/2+(1/5)/5)

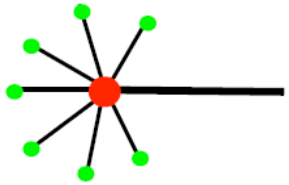
PR = (0.28,0.11,0.39,0.11,0.15)

Knotenausfall

- Wie robust ist ein Netzwerk beim Ausfall von Knoten?
- Können noch Informationen ausgetauscht werden?
- Zwei Arten von Störungen:
  - Zufälliger Ausfall von Knoten
  - Gezielte Attacke
- Die Störungen haben unterschiedliche Folgen auf Zufallsgraphen und Skalenfreie Netzwerke

Knotenausfall

- Ein Zufallsgraphs aus dem Knoten zufällig entfernt werden, ist immer noch ein Zufallsgraphs
- Skalenfreie Netzwerke sind robuster bei zufälligem Ausfall von Knoten, aber Schwächer bei gezielten Attacken auf Knoten mit hohen Grad



P(Hub) : P(Peripherie) = 1 : 7

- Die Wahrscheinlichkeit einen unwichtigen peripheren Knoten zu erwischen, ist extrem viel höher, als einen Hub zu treffen
- Bei Ausfall von nur 1% der Knoten mit höchtem Grad zerfällt das Netzwerk in kleine isoliterte Teile

Impfen in Skalenfreie Netzwerke

- Impfen/Immunisieren von Menschen/Tieren bei Epidemien
- Immunisierung von Knoten in Mailnetzwerken, Routern, Servern etc. um Viren oder Fehlrepropagierung zu Stoppen
- Impfen der minimale Mange von Knoten, um die Ausbreitung des Virus zu stoppen

Impfen in Skalenfreie Netzwerke

- Impfen/Immunisieren von Menschen/Tieren bei Epidemien
- Immunisierung von Knoten in Mailnetzwerken, Routern, Servern etc. um Viren oder Fehlrepropagierung zu Stoppen
- Impfen der minimale Mange von Knoten, um die Ausbreitung des Virus zu stoppen

Die Hubs impfen

Workshop
Les Miserables & Karate

- Files
- LesMiserables
- Karate

Workshop
Social Network Analysis of 2016 US Presidential Election Candidates

- https://github.com/Rameshb-umd/Social-Network-Analysis
- Files
- realdonaldtrump_no_tweet_10k.gephi: Fig.6, Fig.15 (Ego Network Hillary)
- HillaryClinton_Notweets.gephi: Fig.7, Fig.17 (Ego Network Trump)
- Trump2016.gephi: Fig.8 (nicht sicher)
- RealDonaldTrumpFriends.gml: Fig.10
- Bearbeiten bis Fig. 7

Workshop
Game of Thrones Season 1

- https://gameofnodes.wordpress.com/2015/05/06/game-of-nodes-a-social-network-analysis-of-game-of-thrones/
- Files
  - Game of Thrones nodes
  - Game of Thrones edges
- Gephi → File → Import Spreadsheet → gotnodesseason1.csv (Comma separated)
- Gephi → File → Import Spreadsheet → gotedgeseason1.csv (Comma separated)

Workshop
Facebook Network

- Visualize Facebook network with Chrome extension “Lost Circles”:
- http://2centsapiece.blogspot.ch/2016/08/visualize-your-facebook-network.html