

# **Social Network Analysis**

Roland Hediger

22. Oktober 2014

# Inhaltsverzeichnis

## I. Grundkonzepte

### 1. Grundkonzepte 1

1.1. Andere Begriffe . . . . .	4
1.2. SNA Allgemein . . . . .	4

### 2. Grundkonzepte 2

2.1. Eigenschaften von Dyaden . . . . .	6
2.2. Datenformate . . . . .	6
2.3. Datenerhebung . . . . .	7
2.4. Social Media Monitoring . . . . .	8
2.5. Definitions Extra . . . . .	9

### 3. Grundkonzepte 3

3.1. Arten von Fragestellungen . . . . .	10
3.2. Zusammenhänge . . . . .	11
3.3. Kriterien einer guten Visualisierung . . . . .	12

## II. Mathematischer Hintergrund

### 4. Grundlagen zur Graphen

13

14

### 5. Interpretation/Metriken

16

5.1. Zentralisierung . . . . .	18
5.2. Metriken . . . . .	19
5.3. Communities . . . . .	20
5.4. Clustering und Communities . . . . .	21

### 6. Small World

25

6.1. Experimente . . . . .	25
6.2. Eigenschaften . . . . .	26

# **Teil I.**

## **Grundkonzepte**

# 1. Grundkonzepte 1

## Was ist ein Knoten?

- Personen
- Gruppen
- Events
- Proteine

## Knotenbezeichnungen

Actor,node,site(Physik, selten gebraucht),vertex.

## Was ist eine Verbindung?

Beziehungen - :

- Freundschaft(frei gewählt)
- Verwandschaft in einer Organisation (vorgegeben)

## Verbindungsbezeichnungen

tie,link,bond,relationship,connection.

## 1.1. Andere Begriffe

**Edge** Ungerichtete Verbindung (Kante)

**Dyade** Knotenpaar mit möglicher Verbindung

**Triade** drei Knoten mit möglichen Verbindungen

**Pfad** direkt verbundene Knotenkette

**Geodesic** Kurzester Pfad

**Clique** Komplette verbundener Subgraph

**Community** Gemeinschaft - keine eindeutige Struktur

**Komponent** allein stehender Teil eines Netzwerks

**Graph** Mathematische Bezeichnung für Netzwerk

## 1.2. SNA Allgemein

### Herausforderungen

Begrenzung der Population,Fehlende Datepunkte,Rauschen in den Verhaltensdaten, Semantik von Verhaltensdaten,Komplexität.

### Analyse Ebenen

**Mikro** Knoten und ihre Umgebung

**Meso** Elemente innerhalb eines Netzwerks

**Makro** Gesamte Netzwerk Struktur

### Metriken

- Dichte
- # bestehende / # mögliche Verbindungen
- $\Delta = \frac{\sum_{i,j} x_{ij}}{n(n-1)}$
- Zentralitätsmasse Position v Aktoren
- Graph Korrelation und Regression
- Random und bias nets, QAP EGRM
- Simulationen

### Hypothesen, Selektion vs Einfluss

Netzwerkstruktur  $\Rightarrow$  Verhalten, Einstellung

Machen mehr freunden glücklicher?

## 2. Grundkonzepte 2

### 2.1. Eigenschaften von Dyaden

#### Dyadischer Zensus

- Mutual/ Reciprocal



- Asynchron - eingehend oder ausgehend:



- No tie
- Multiplexität - Überlappung Freund & Hilfe



- Austausch:



### 2.2. Datenformate

- Einträge: Sender, Empfänger, Gewichtung, Attribute
- Nachteil: Schwierige Übersicht
- Vorteil: Skalierbar (effizient bei vielen Daten)

Beispiel:

Sender	Receiver	Weight	Timestamp
Laura	Kari	4	Sept
Kari	Sepp	3	Sept
Fritz	Laura	10	Nov

Beim Import werden oft Angaben über Netzwerk gefragt

Abbildung 2.1.: Datenformate: Linked List

- Tabelle mit Sendern (Col) und Empfaenger (Row)
- Nachteil: schlecht skalierbar
- Vorteil: Dichte, uebersichtliche Darstellung der Daten

Beispiel:

	Laura	Kari	Fritz
Laura		2	
Kari	1		3
Fritz	2		

Beim import muessen oft  
Fragen zur Symmetrie,  
Diagonale, Missing  
Werte etc. definiert  
werden!

Viele weitere, neuere Formate: GraphML (XML basiert)...

Abbildung 2.2.: Datenformate:Matrix

## Matrix Sorting

		1		1 1 1	1 1 1 1	1 2 2	2 3	2 2	2 1 2 2 2 2
	1 2 3 3 7 6	4 2 6 8 8	6 3 7 9 9 1 6 2	0 5	4 7	5 4 1 8 9 3			
	B T S H J L S A H H	Z F H G S Z S S G	X N Y G	X J H Q N G					
1 Beijing	1 1 1 1 1 1 1	1 1	1 1 1 1 1	1 1 1 1					
2 Tianjin	1 1 1	1	1		1	1			
15 Shandong	1 1 1	1							
3 Hebei	1	1							
7 Jin	1	1 1			1 1				
6 Liaoning	1	1 1 1 1	1	1	1 1				
4 Shanxi			1		1	1			
12 Anhui			1 1	1 1	1				
16 Hunan	1		1	1		1			
18 Hunan	1	1	1 1	1 1 1 1	1				
8 Heilongjiang			1	1	1				
10 Jiangsu	1	1	1	1 1 1 1 1 1	1				
13 Fujian	1			1 1 1 1	1				
17 Hubei	1 1	1	1	1 1 1 1 1	1 1				
19 Guangdong	1		1	1 1					
9 Shanghai	1 1	1	1 1 1	1 1 1 1 1 1	1 1	1 1	1		
11 Zhejiang	1			1 1 1 1		1			
26 Shanxi	1 1 1			1 1 1 1 1 1			1		
22 Sichuan	1 1		1	1 1 1 1 1 1	1	1		1	
20 Guangxi					1				
30 Xinjiang						1			
5 Henan							1		
24 Yunnan								1 1	
27 Gansu			1					1 1	
25 Inner Mongolia									
14 Jiangxi									
21 Hainan									
28 Qinghai									
29 Ningxia									
23 Guizhou									

Abbildung 2.3.: Matrix Sorting

## 2.3. Datenerhebung

1. Fragebogen
2. Beobachtungen
3. Datenbanken
4. Verhaltensdaten

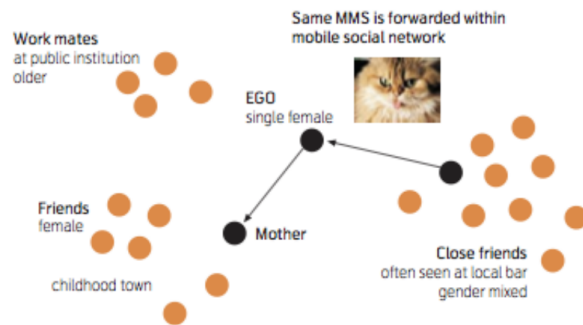


Figure 1 Sample sketch of a personal social network

Abbildung 2.4.: Beispiel Quantatives Mapping

**Location of conversations** self explanatory.

**Volume of Conversations** Conversation count.

**Latency** Conversation speed.

## 2.4. Social Media Monitoring

Continuous analysis of available Social Media Content.


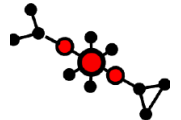
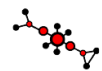
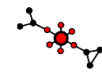
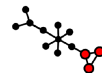
- Customer service issues
- Competitor infos
- PR Disaster Warning System
- Company prepared for social conversations??

**Monitoring Features Coverage** types of media or geographic markets

**Sentiment Analysis** Accuracy at a level varied 59-87%



## 2.5. Definitions Extra

Term	Definition	Interpretation	Example
Degree (Most Reachable)	Number of direct relations to other nodes in the network. (direction of connection is ignored)	An actor with high degree centrality is active and might have immediate power / prominence within a network.	
Closeness(Efficient Reach)	Shortest path distance to all other actors of the network (inverse sum).	An actor with high closeness centrality can access efficiently all other nodes in a network.	
Betweenness (Control flow)	Number of shortest paths a node lies between two other nodes in the network.	An actor with high betweenness centrality can control relations between others in the form of preventing or promoting the flow of information.	
Eigenvector (Influential Friends)	Importance of a node based on the connections of his connections using the eigenvalue of a node.	An actor with high eigenvector centrality might only have a few connections but to strategically important key-players within a network.	
Clustering(Local Embedded)	Likelihood that two nodes, which are connected to a node, are connected between themselves. An actor with a high clustering coefficient has local influence and might be less susceptible for outside information.	An actor with a high clustering coefficient has local influence and might be less susceptible for outside information.	

### 3. Grundkonzepte 3

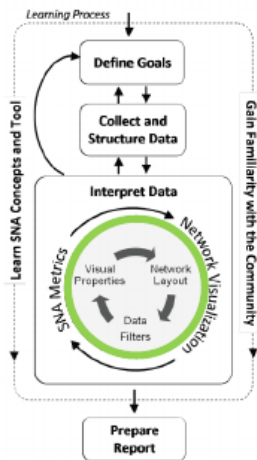


Figure 3. Network Analysis and Visualization Process Model derived from the students' practices.

#### 3.1. Arten von Fragestellungen

##### Knoten Basiert

- Sind Mitarbeiter mit mehr Verbindungen erfolgreicher?

(Social Capital)

- Haben Kunden mit mehr Umsatz auch mehr Kommunikations-Partner?

##### Verbindungsbasiert

- Führen Freundschafts-Verbindungen zu Business-Verbindungen?
- Gibt es einen Zusammenhang zwischen der Migration und dem
- kommunikations-Volumen zwischen Ländern?

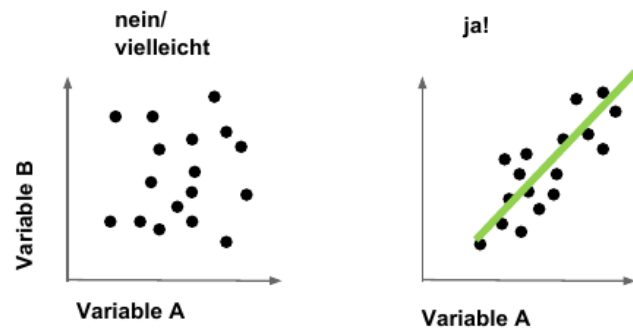
##### Netzwerkbasiert

- Sind Teams mit höherer Kommunikations-Dichte effizienter?
- Je grösser ein Ego-Netzwerk desto weniger Triaden?

##### Gemischt

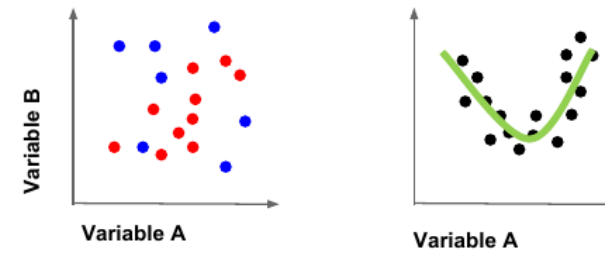
- Beeinflussen sich Freunde gegenseitig in ihrer Meinung zu politischen Themen oder bezüglich Konsumverhalten? (Diffusion, Influence) T
- endieren Mitarbeiter dazu bei Personen Hilfe zu suchen, welche dasselbe Geschlecht haben? (Homophily, Selection)

## 3.2. Zusammenhänge



Illustrative Abbildung,  
Messung fuer den linearen Zusammenhang ueber den Korrelationskoeffizienten  $R \pm 1$

Abbildung 3.1.:



Illustrative Abbildung,  
Messung fuer den linearen Zusammenhang ueber den Korrelationskoeffizienten  $R \pm 1$

Abbildung 3.2.:

## Vorgehen in ihrem Projekt



Abbildung 3.3.:

- Kontinuierliche Variablen (z.B. Anzahl Follower, etc.) sind mit Grösse, Strichdicke, Farbverläufen dargestellt.
- Eine Legende der Symbole die Anzahl Knoten und Verbindungen sind gegebenenfalls ersichtlich.

## 3.3. Kriterien einer guten Visualisierung

- Möglichst wenig Überlappungen von Verbindungs- strukturen (Ausprobieren verschiedener Layoutalgo- rhythmien und Parameter: Atlas, Fruchterman, etc.)
- Netzwerk-Komponenten und isolierte Akteure erkennbar.
- Kategoriale Variablen (z.B. Frau/Mann) werden mit verschiedenen Farben, Symbolen repräsentiert.

**Teil II.**

## **Mathematischer Hintergrund**

## 4. Grundlagen zur Graphen

Ein Graph besteht aus Knoten (auch Akteure oder Nodes genannt) und Kanten (Edges). Im folgenden kleinen Beispiel-Graphen haben wir die Knoten A bis D. Die Kanten zwischen den Knoten sind hier ungerichtet was bedeutet, dass die Interaktion in beide Richtungen möglich ist. Personen C und D zu verschiedenen Teams gehören (z.B Teamleiter sind) und die Person B beide kennt. In der Sozialen Netzwerkanalyse stellen die Kanten eine direkte oder indirekte Kommunikation / Verbindung zwischen verschiedenen Akteuren dar. Einer Kante kann dabei verschiedene Bedeutungen zugetragen werden. Mögliche Bedeutungen:

- Informationsaustausch
- Ressourceaustausch
- Beeinflussung
- Mitgliedschafts-Beziehung
- Verwandschafts-Beziehung
- Persönliche Beziehung usw.

Oftmals erhalten die Kanten auch ein Gewicht (z.B. Anzahl Benachrichtigungen untereinander). Bei solchen Graphen wird auch von **gewichteten Graphen** gesprochen. Die Netzwerk-Struktur sowie die Eigenschaften von Knoten/Kanten lassen interessante Analysen zu. Zum einen bietet die Netzwerk-Struktur bereits viel Potenzial um Schlüsselpersonen durch Berechnung von Zentralitätsmassen zu erkennen. Zum anderen kann das Netzwerk anhand der vorhandenen Eigenschaften gefiltert werden.

### Connected Components

- Menge von Knoten verstanden die über beliebige Pfade miteinander verbunden sind. Ist ein Teilnetzwerk abgeschnitten, bildet es einen zweiten **Component**
- Unterschied Strongly Weakly Connected : jeder Node in **genau einer Component**
- Strongly Connected : dass ich alle Knoten entlang der Kantenrichtung erreichen. Hier ein Beispiel:

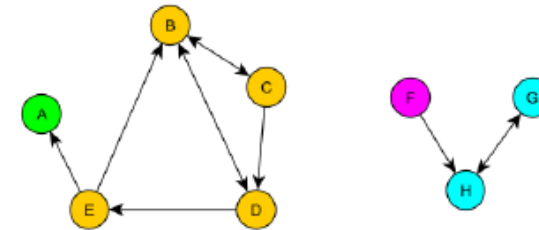


Abbildung 4.1.: Connected Components

- Components im Beispiel:  $\{B, C, D, E\}$ ,  $\{A\}$ ,  $\{G, H\}$ ,  $\{F\}$

- Bei Weakly Connected Components wird untersucht, welche Knoten einander erreichen, wobei die Kantenrichtung ignoriert wird und die Kanten einfach als bidirektional angesehen werden. Im folgenden Beispiel existieren die Weakly Connected Components:

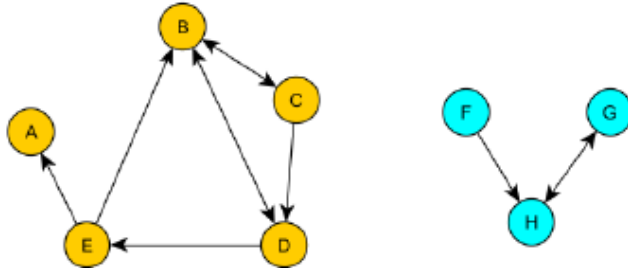


Abbildung 4.2.: Weakly Connected Comonents

## 5. Interpretation/Metriken

**Actor Zentralität** Wichtigkeit einzelner Knoten. „Sichtbarkeit des Aktors“, Kontrollmöglichkeiten auf den Informationsfluss.

**Degree Centrality** Kanten die von Knoten weg gehen oder zu einem Knoten führen - direkten Verbindungen zum Nachbar. **Gerichteten Graphen:** InDegree, OutDegree.

Normalisierung: Degree Centrality kann als **lokales Mass** angesehen werden. Isolierte nicht. Normalisierung berücksichtigt

**Grösse des Netzwerks**

Für die Normalisierung muss der Degree Centrality Wert des Knotens durch die maximale Anzahl möglichen Verbindungen dividiert werden. Dies sind:

Bei ungerichteten Graphen:  $n-1$

Bei gerichteten Graphen:  $2(n-1)$

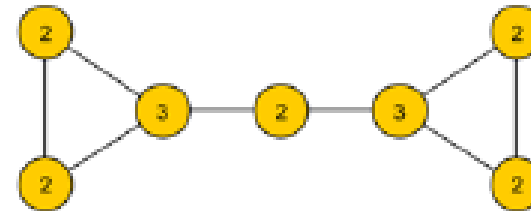


Abbildung 5.1.: Degree Centrality

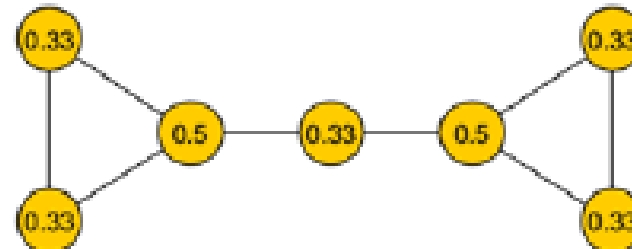


Abbildung 5.2.: Degree Centrality Normalisiert

Interpretation: Degree Wert = Wahrnehmung / Einfluss / Prestige

**Closeness Centrality** Nicht lokales Mass. Berechnet für jenen Knoten, wie effizient von einem Knoten aus alle anderen erreichbar sind. **Berechnung:** Inversen der kürzesten Distanzen zu allen anderen Knoten werden aufsummiert. Zeitintensives verfahren:



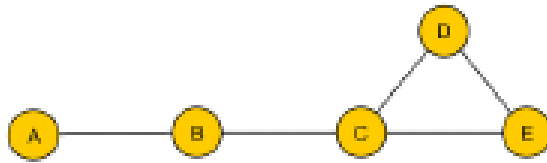


Abbildung 5.3.: Closeness Centrality

$$C_c(A) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{3} = \frac{13}{6}$$

$$\text{Normalisiert} := \frac{1}{4} * \frac{13}{6}$$

Für die Normalisierung wird diese Summe dann noch durch die Anzahl Knoten -1 geteilt. Formel:

$$C_c(i) = \sum_{j=0}^N [d(i, j)]^{-1}$$

**Betweenness Centrality** Position innerhalb des ganzen Netzwerkes für alle Knoten.

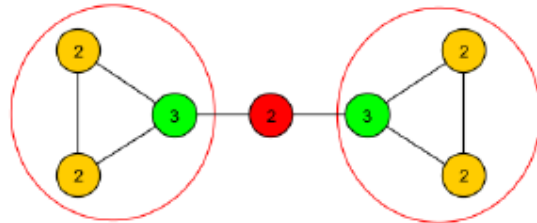


Abbildung 5.4.: Betweenness Centrality

- Brokerage Position oder Gatekeeper.
- Berechnet für jeden Knoten wie stark sich dieser in einer *Brokerage Position* findet

<sup>1</sup>Betweenness Centrality Wert

- kürzeste Pfad zu allen anderen gesucht. Liegt der Knoten auf viele dieser kürzesten Pfade desto höher Betweenness Centrality.

- Algorithmus:

1. Alle Knoten auf BCW<sup>1</sup> 0
2. Der kürzeste Pfad zum nächsten Knoten geht über den vorherigen. Deshalb wird der BCW dieses Knotens um 1 erhöht - rekursiv weiter

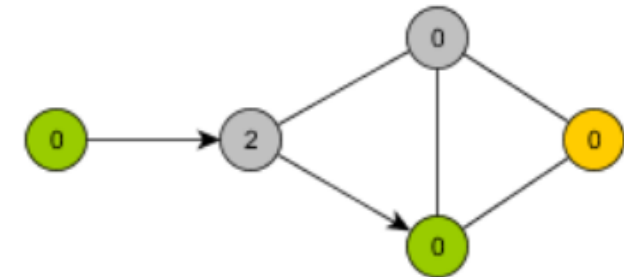


Abbildung 5.5.:

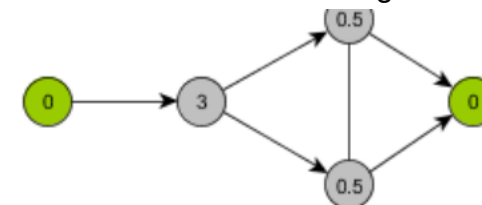


Abbildung 5.6.:

Vom ersten zum letzten Knoten finden wir einen Spezialfall. Es gibt zwei kürzeste Pfade! In diesem Fall

wird jedem Knoten auf dem Pfad jeweils 1 / #kürzeste Pfade addiert. In unserem Fall würden wir also allen Knoten 0.5 addieren auf beiden Wegen. Die restlichen Schritte werden nach dem gleichen Schema ausgeführt. Weil es sich um ein bidirektionales Netzwerk handelt muss ein Knotenpaar lediglich einmal berücksichtigt werden.

- Normalisierung: Gerichtete Graphen:  $(n-1) * (n-2)$   
Ungerichtete Graphen:  $(n-1) * (n-2) / 2$
- Formel:  $C_B = \sum_{s \neq v \neq t \in V} \sigma_{st}(V)$
- Normalisierung Formel:  $C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$

## 5.1. Zentralisierung

**Netzwerk Zentralisierung:** Wie zentral der Akteur ist anhand von Gegebenheiten oder seiner Position innerhalb des Netzwerks. Eigenschaften :

- Masse der zentralste Akteur die Zentralität der anderen Akteure überschreitet.
- Auf Max Wert bezögen (Netz-abhängig).
- Formel:  $\frac{\sum_{i=1}^n |C_x(p^*) - C_x(p_i)|}{\max \sum_{i=1}^n |C_x(p^*) - C_x(p_i)|}$
- In Worte gefasst bedeutet die Formel: Es wird im Zähler die Differenz zwischen dem höchsten Zentralitätswert zu

allen Akteur-Zentralitätswerte aufsummiert. Der Nenner bekommt den Wert der maximal möglichen Differenz. So resultiert immer ein Wert zwischen 0 und 1, wobei 0 bedeutet, dass alle Knoten gleich zentral sind und 1, dass ein Akteur maximale Zentralität besitzt und alle anderen die tiefste Zentralität. Es besteht also die höchst mögliche Ungleichheit.

**Degree Zentralisierung** Ausgehend von den **nicht-normalisierten Degree Centrality Werten**, berechnet sich in einem **ungerichteten Graphen**

**Formel:**  $C_D = \frac{\sum_{i=1}^n |C_D(p^*) - C_D(p_i)|}{(n-1)(n-2)}$

**Nicht normalisiert**

Im fall von einem gerichteten Graph muss der Nenner mit 2 multipliziert w

**Betweenness Zentralisierung** Geht von bereits **normalisierten Akteur zentralitäten** aus<sup>2</sup>

**Formel:**  $C_B = \frac{\sum_{i=1}^n |C'_B(p^*) - C'_B(p_i)|}{n-1}$

**Closeness Zentralisierung**

$$C'_c(n_i) = \frac{n-1}{(\sum_{j=1}^n d(n_i, n_j))}$$

Der grosse Nachteil an dieser Formel ist, dass für Graphen mit verschiedenen Komponenten der Closeness-Centrality Wert für alle Knoten 0 wird, da die Distanz zwischen zwei sich nicht erreichbaren Akteuren per Definition  $\infty$  ist und somit der Nenner  $\infty$ .

Abbildung 5.7.:

**Guter Formel:**  $C_c = \frac{\sum_{i=1}^n |C'_c(p^*) - C'_c(p_i)|}{\frac{n-2}{2}}$

<sup>2</sup>Alle variablen in diesem Abschnitt mit Strich representieren normalisierte Werten

## 5.2. Metriken

- Kennzahlen zur komplette Netzwerkstruktur
- Netzwerkvergleich
- Schlussfolgerungen ziehen
- Kommunikation im Zentrum
- Indirekte Kommunikation proportional zum Manipulationsgefahr.

**Graph Density** Wie gut der Graph insgesamt verbunden ist.

Wert zwischen 0 und 1

**Density für gerichteten Graphen:**  $\frac{|E|}{|V|(|V|-1)}$

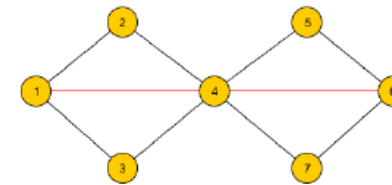
**Density ungerichteten Graphen:**  $\frac{2|E|}{|V|(|V|-1)}$

Dichte gibt Auskunft darüber wie schnell sich Informationen in einem Netzwerk verbreiten da in einem dichten Netzwerk die Kommunikation sehr direkt verläuft und sich somit schnell verbreiten kann. Die Dichte nimmt mit der Grösse des Netzwerks ab da der direkte Kontakt zu einer kleineren Menge von Leuten gepflegt werden kann, aber nicht zu einer grossen Menge.

**Graph Diameter** längste kürzeste Pfad zwischen alle Knotenpaaren in einem Graphen.

Mehrere Komponenten = unendlich

PROZESSUSTRUKTUR FÜR RECHNUNG



Dieser Graph besitzt einen Diameter  $d(1,6)$  resp.  $d(1,7) = 4$ .

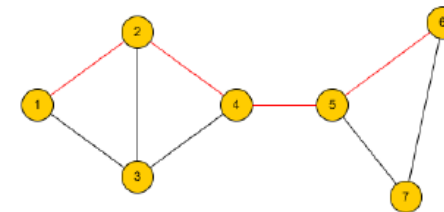


Abbildung 5.8.:

Auskunft über Entfernung der Knoten. Maximale Knoten auf kürzesten Pfad um einen Akteur zu erreichen.

### Cluster Coefficient

- Cliquebildung - 1 ist max Wert - Clique.
- Lokalen vs Globalen Cluster Coefficient. Global = Mittelwert von lokalen Werten.
- **Formel:** Quotienten der Anzahl Kanten zwischen den Nachbarn eines Knoten (**ungerichteten Netzwerken mit 2 multipliziert**). und der maximal Anzahl Kanten möglich.

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

**Globale Clustering:**  $C' = \frac{1}{N} \sum_i^N C_i$

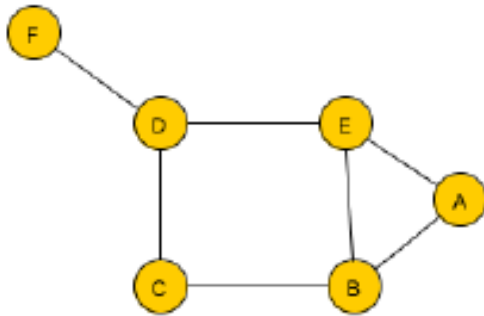


Abbildung 5.9.:

- $C_A = 1$
- $C_B = 1/3$
- $C_C = 0$
- $C_D = 0$
- $C_E = 1/3$
- $C_F = 0$
- $C' = 1/6 * 5/3 = 5/18$

### 5.3. Communities

- Subgraph innerhalb Graphen, stark ziemlich direkt verbunden
- Zeigt vertrauensverhältnis

- Cliques haben redundanten Informations-lieferanten.
- Cliques von innovationen abgekoppelt.
- Identifizierung eine Community :
  - Alle kennen einander (Direkte Beziehung)
  - Alle Knoten haben mindestens k Links zu anderem Knoten.  
(K Core, Grad der Verbundenheit)
  - Individuellen erreichen einander mit max n Hops (n-Clique).

**KCore** k-Core hat eine weniger einschränkende Bedingung als Clique. Hier muss jeder Knoten mindestens k andere Knoten des Clusters verbunden sein.

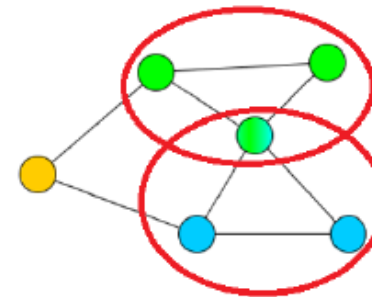


Abbildung 5.10.:

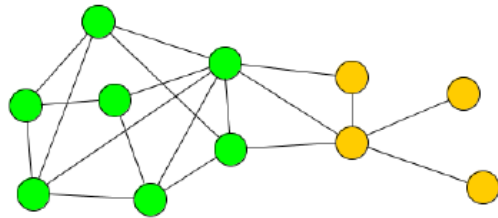
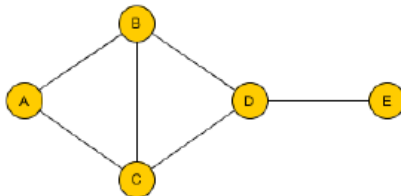


Abbildung 5.11.: 3 core Beispiel

**n-cliques** Alle innerhalb einer n-Clique erreichen einander mit maximal n Hops. Eine 1-clique wäre somit eine Clique wie sie vorher dargestellt wurde, wo jeder jeden direkt erreichen kann. Im folgenden Graphen befinden sich drei 1-Cliques, wobei die Knoten B,C und D zu zwei Cliques gehören:

- A, B, C
- B, C, D
- D, E



Der Graph enthält folgende 2-Cliques:

- A, B, C, D
- B, C, D, E

Abbildung 5.12.:

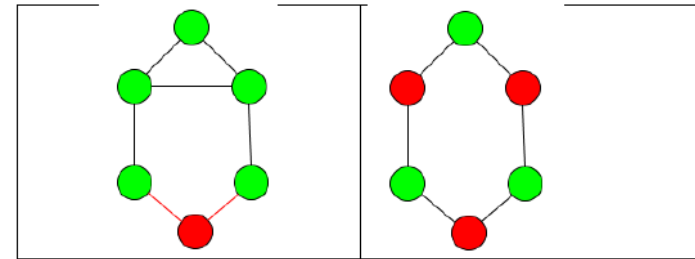


Abbildung 5.13.:

- Durchmesser kann grösser sein als n. Grun ist 2 Cluque.
- Knoten von n Clique können z.B nicht mit einander verbunden sein. Roten und Grünen können als 2 Cliques interpretiert werden (rechts).

**p-Cliques** Bei p-Cliques muss mindestens ein Bruchteil p (Angabe zwischen 0 und 1) aller Kanten eines Knotens zu anderen Knoten führen, welche sich im Cluster befinden. Somit werden viele der oben erwähnten Nachteile beseitigt.

## 5.4. Clustering und Communities

**Hierarchical Clustering** Bottom up Clustering verfahren. Jeden Knoten bildet einen eigenen Cluster am Anfang. Zusammenfassung des Graphens Schritt für Schritt. Bei jeder iteration werden die Ähnlichkeiten zwischen allen Clusters untersucht. Stopp bei gewünschter Anzahl Clusters oder alle Knoten befinden sich im selben Cluster.

Beispiel:

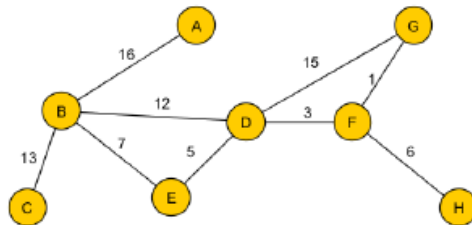


Abbildung 5.14.:

Zu Beginn bildet jeder Knoten seinen eigenen Cluster, es existieren also 8 Clusters. Jetzt werden diejenigen Knoten zusammengeführt, welche am meisten kommunizieren. Das sind die Knoten A und B:

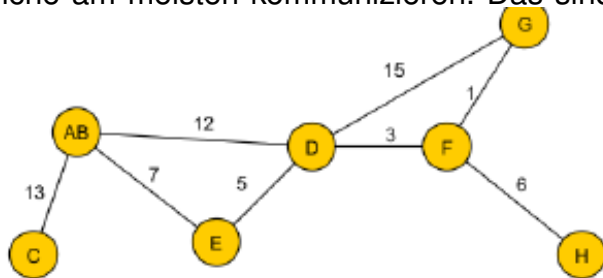
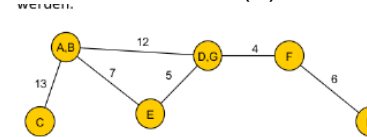


Abbildung 5.15.:

Nun werden die Knoten D und G zusammengefasst. Da diese nun einen neuen Cluster bilden muss die Kommunikation zwischen dem neuen Cluster und aussenstehenden Knoten aktualisiert werden. Beide Knoten kommunizieren mit dem Knoten F. Deshalb muss beim neuen Cluster für die Kommunikation mit dem Knoten F die Summe zwischen den Knoten-

paaren G und F (1) sowie D und F (3) verwendet werden:



Dieses Verfahren wird nun Schritt für Schritt fortgesetzt bis schlussendlich zwei Clusters entstehen:

Abbildung 5.16.:

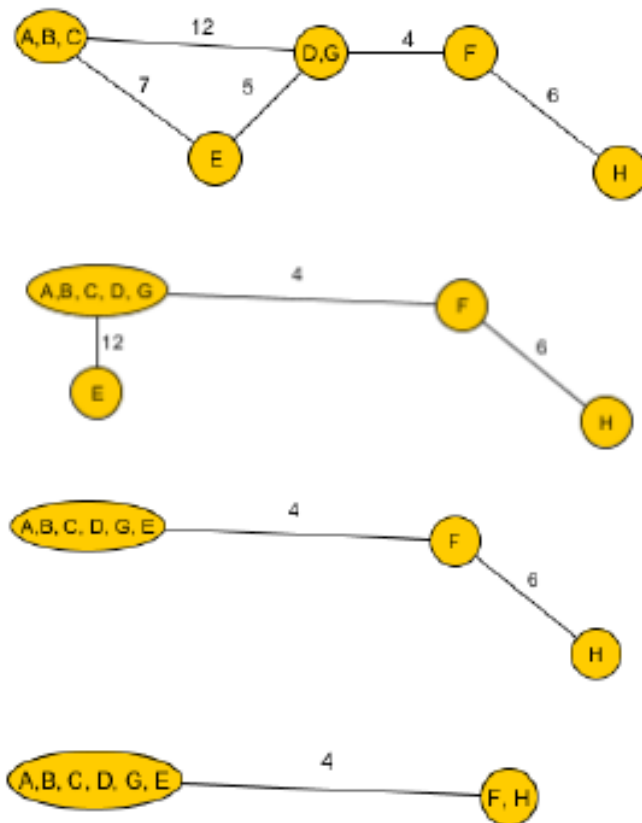


Abbildung 5.17.:

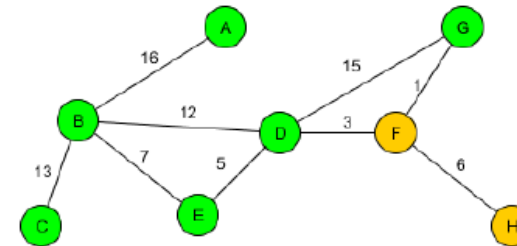
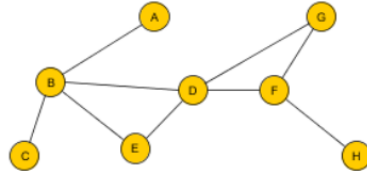


Abbildung 5.18.:

### Edge Betweenness Clustering

Das Edge-Betweenness-Clustering, auch bekannt unter dem Girvan-Neuman Algorithmus, ist ein Top-Down Clustering Verfahren. Dies bedeutet, dass sich zuerst alle Knoten im selben Cluster befinden und dann sukzessive aufgeteilt werden. Wie der Name bereits vermuten lässt wird als Ähnlichkeitsmass der Edge-Betweenness Centrality Wert berechnet. Die Centrality-Berechnung ist genau dasselbe wie bei der Node Betweenness Centrality, nur jetzt halt für Kanten. Ein hoher Edge-Betweenness Wert bedeutet, dass es sich um eine Kante zwischen zwei Knoten-Gruppen handelt. Es gibt wieder mehrere Iterationen, wobei in jeder Iteration diejenige(n) Kante(n) mit dem höchsten Betweenness-Wert entfernt wird, bis schlussendlich die gewünschte Anzahl Clusters erreicht worden ist. Dieses Verfahren ist sehr rechenintensiv, da nach jeder Iteration die Edge-Betweenness Werte neu berechnet werden

müssen, der komplette Algorithmus liegt in  $O(n^3)$ . Deshalb wird es selten eingesetzt. Hier ein Beispiel für das Betweenness Clustering. Die Kante(n) mit dem höchsten Betweenness-Wert sind jeweils rot eingefärbt und werden nacheinander entfernt:



Berechnete Edge-Betweenness Centralities (nicht normalisiert):

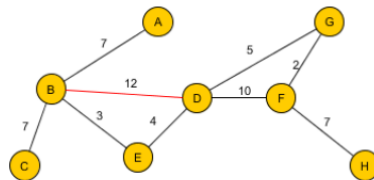
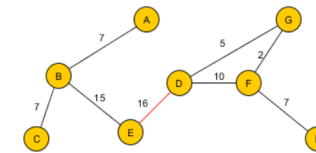


Abbildung 5.19.:

Nachdem die Kante mit dem höchsten Betweenness-Wert entfernt wurde, werden die Kanten- Betweenness Werte neu be-



Nachdem wieder die Kante mit dem höchsten Betweenness-Wert entfernt wurde erhalten wir den in zwei Clusters aufgeteilte Graphen:

rechnet:

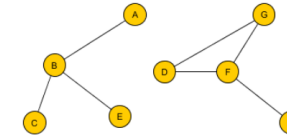


Abbildung 5.20.:



## 6. Small World

### 6.1. Experimente

**Milgram** Ursprüngliches Paket-Experiment von Milgram Im Jahre 1967 hat Stanley Milgram folgendes Experiment durchgeführt: 60 Teilnehmer der USA mussten ein Paket an eine festgelegte Person in Boston, die sozial und geografisch weit von der Ursprungs-Person entfernt war, ein Paket zusenden. Sie durften es aber nur der Person direkt zustellen falls Sie diese auch persönlich kennen und mit Vornamen ansprechen. Ansonsten mussten sie es einer bekannten Person mit gleichem Kriterium weitersenden, bei der die Wahrscheinlichkeit hoch war, dass sie die Zielperson kennt. Der Weg des Paketes wurde protokolliert. So wurde untersucht, wie viele Personen zwischen Absender und Empfänger lagen (Anzahl Hops) Die durchschnittliche Pfadlänge lag bei 5.5. Daraus schliesst sich, dass im Durchschnitt innerhalb der USA jede Person jede andere über 6 Personen erreichen kann. Es entstand auch der Ausdruck Six Degrees of Separation Theorie für dieses Phänomen. Dieser Ausdruck wurde jedoch nie von Milgram verwendet.

**Erdos** Co-Authorship Experiment von Paul Erdős Ein weiteres Small World Experimente wurde vom Mathematiker Paul Erdős

durchgeführt. In einem Graphen stellt er alle Autoren von Publikationen als Knoten dar. Kanten zwischen zwei Knoten wurden erzeugt, wenn diese gemeinsam eine Publikation verfasst haben (Co-Autor). Paul Erdős gab sich selbst die Erdős-Zahl 0. Personen, mit denen er publiziert hatte, erhielten die Zahl 1. Autoren, welche mit Co- Autoren von Paul Erdős eine Publikation verfasst haben die Zahl 2 usw. Autoren, welche nicht erreicht werden konnten, erhielten die Zahl „unendlich“. Es zeigte sich, dass die Zahl entweder unendlich oder sehr klein war. Bei 268'000 Personen konnte ein endlicher Wert ermittelt werden, der Durchschnitt bei 4.65 lag. (Erdős hat in sehr vielen Teilen der Mathematik publiziert).

**MSN** Microsoft analysierte 90 Millionen täglich aktive Messenger-Accounts. Jeder Account wurde als Knoten und die Kommunikation zwischen zwei Personen als Kanten dargestellt. Die Analyse ergab schlussendlich, dass zwei beliebige Personen durchschnittlich 6.6 Schritte voneinander getrennt waren. Es gab auch Pfade bis zu einer Länge von 29 Schritten. Damit wurde die Theorie der Small World anhand eines riesigen, globalen Netzwerks bestätigt, auch wenn die beiden Forscher Eric Horvitz und Jure Leskovec eher „Seven Degrees of Separation“ als Mass vorschlagen würden.

## 6.2. Eigenschaften

- Wenige Hops um Leute zu verbinden
- Neigen nahezu Cliquen zu formen
- Hub Nodes (Indegree/Outdegree viel)
- Die Distanz zwischen zwei zufällig gewählten Knoten entspricht etwa dem Logarithmus der Anzahl Knoten. Duncan Watts and Steven Strogatz haben entdeckt, dass Graphen anhand zwei unabhängigen Metriken klassifiziert werden können: Clustering Coefficient, Average Shortest Path
- **Um zu erkennen, ob es sich um ein Small-World Graphen handelt** werden die beiden Masse mit einem Random-Netzwerk mit ungefähr gleicher Degree-Verteilung verglichen. Für ein Small World Netzwerk sind dann die folgenden beiden Eigenschaften erfüllt:

$$L_{SW} \leq L_{RAND}$$

$$CC_{SW} \leq CC_{RAND}$$