# Big Data Anwendungen

Aufgabenblatt 6

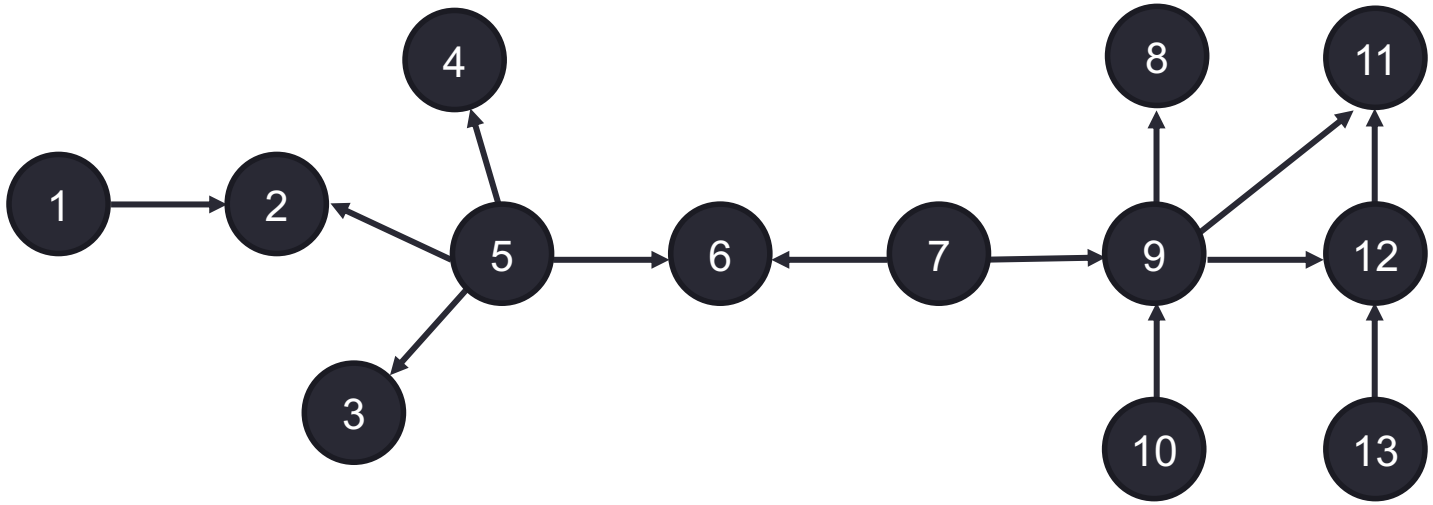
# Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining

- Aufgabenblatt 6 – Social Network Analysis
  - Aufgabe 1 – Kennzahlen
  - Aufgabe 2 – Kerningham-Lin
  - Aufgabe 3 – Verständnisfragen

# Aufgabe 1 (a) - Aufgabenstellung

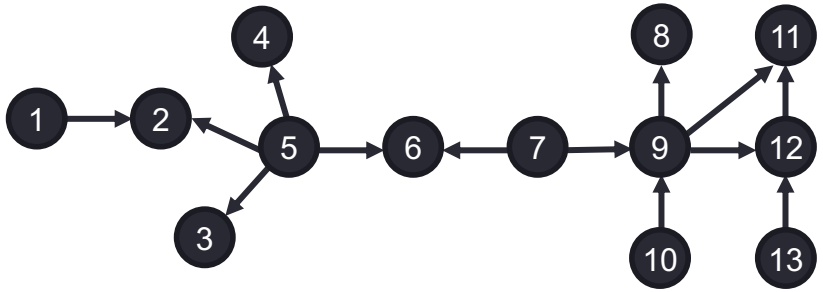
- Gegeben sei die folgende Illustration eines sozialen Netzwerks:




- Geben Sie die zwei Knoten mit der höchsten minimalen Pfadlänge an und ermitteln Sie die minimale Pfadlänge. **(3 Punkte)**

# Aufgabe 1 (a) – Lösung (Minimale Pfadlänge)

- Minimale Pfadlänge
  - Ermittlung der Pfade (Verbindungen) zwischen zwei Knoten
  - Minimale Pfadlänge ist Länge des kürzesten Pfads

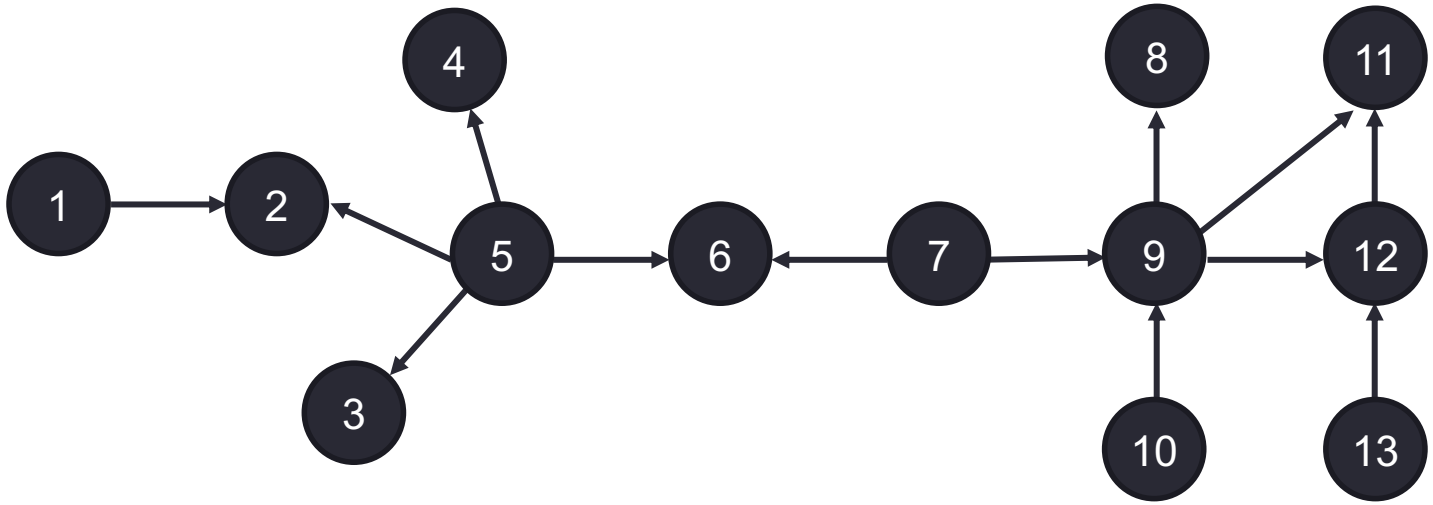



- Hier:
  - Knoten mit höchster minimalen Pfadlänge: Knoten 1 und 13
    Exkurs: 1 und 11 keine Lösung da kürzerer Pfad von 9 nach 11
  - Kürzeste Pfad zwischen 1 und 13
    - (1 → 2 → 5 → 6 → 7 → 9 → 12 → 13)
    - Länge: 7

- Ergebnis: höchste minimale Pfadlänge: 7

# Aufgabe 1 (b) - Aufgabenstellung

- Gegeben sei die folgende Illustration eines sozialen Netzwerks:




- Bestimmen Sie die Triadic Closure für die Knoten 5 und 9. **(3 Punkte)**

# Wiederholung Vorlesung: Triadic Closure

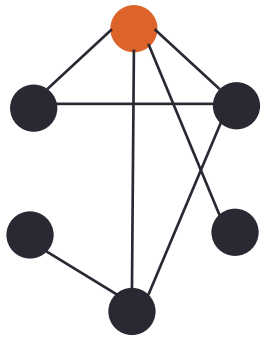
- Idee
  - Entsprechung der sozialen Homophilie bzgl. Netzwerk
  - Knoten mit ähnlichen Kontakten sind auch mit ähnlichen Knoten verknüpft (Xing: „Erweitern Sie Ihr Kontaktnetzwerk“, Facebook: „Freunde finden“)

- Intuition
  Anzahl Verbindungen zwischen Kontakten geteilt durch...
  ... Anzahl möglicher Verbindungen zwischen den Kontakten von $i$

- Formal
  - Clustering Koeffizient des Knoten $i$: $\eta(i) = \frac{|\{(j,k) \in E : j \in V_i, k \in V_i\}|}{\binom{|V_i|}{2}}$

    mit $V_i$ ist die Menge der Kontakte von $i$

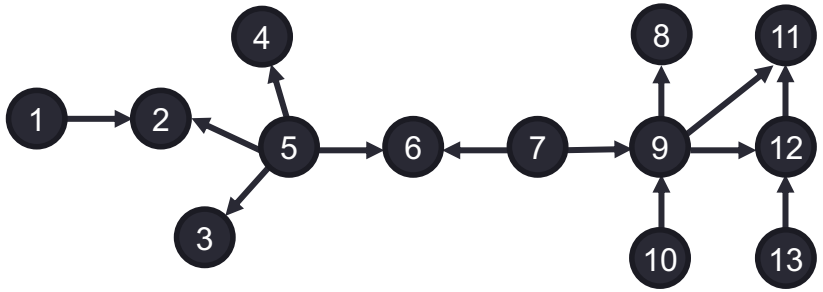
- Beispiel
  - Anzahl Kontakte von 0: $V_0 = 4$
  - Anzahl möglicher Verbindungen: $\binom{|V_0|}{2} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$
  - Anzahl bestehender Verbindungen: $|\{(j,k) \in E : j \in V_0, k \in V_0\}| = 2$
  - Triadic Closure: $\eta(0) = \frac{2}{6} = \frac{1}{3}$



# Aufgabe 1 (b) – Lösung (Triadic Closure)

- Allgemein

$$\eta(i) = \frac{|\{(j,k) \in E : j \in V_i, k \in V_i\}|}{\binom{|V_i|}{2}}$$




- Knoten 5

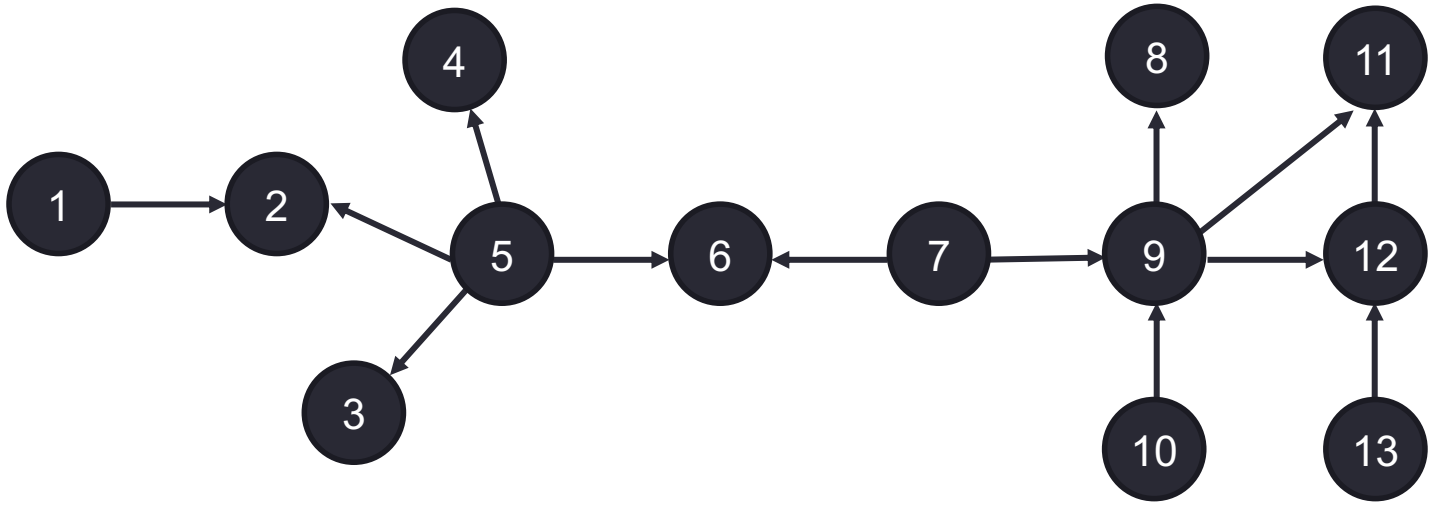
$$\eta(5) = \frac{|\{\}|}{\binom{|\{2,3,4,6\}|}{2}} = \frac{0}{\binom{4}{2}} = \frac{0}{\frac{4!}{2!(4-2)!}} = 0$$

- Knoten 9

$$\eta(9) = \frac{|\{(12,11)\}|}{\binom{|\{7,8,10,11,12\}|}{2}} = \frac{1}{\frac{5!}{2!(5-2)!}} = \frac{2! \cdot 3!}{5!} = \frac{2}{5 \cdot 4} = \frac{1}{10}$$

# Aufgabe 1 (c) - Aufgabenstellung

- Gegeben sei die folgende Illustration eines sozialen Netzwerks:




- Ermitteln Sie die Zentralität für die Knoten 5 und 7. **(3 Punkte)**

# Wiederholung Vorlesung: Zentralität

- Beobachtung
  - Hohe Bedeutung von Knoten mit vielen Verbindungen
  - Eingehende Kanten sind wertvoller als ausgehende
    (auch weil ausgehende Kanten „automatisch“ generiert werden können)

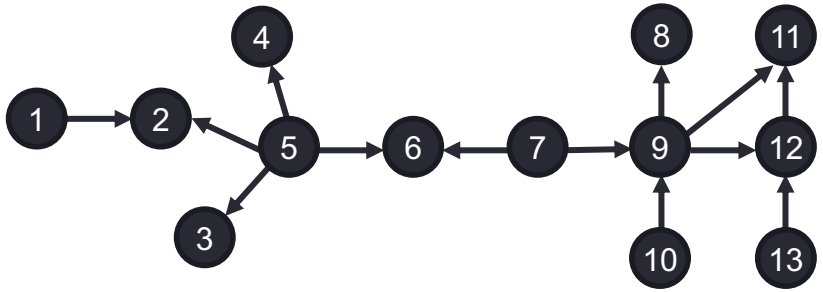
- Definition

$$\text{Zentralität: } c(e_i) = \frac{d(e_i)}{|V|-1}$$

# Aufgabe 1 (c) – Lösung (Zentralität)

- Allgemein

$$c(e_i) = \frac{d(e_i)}{|V|-1}$$




- Knoten 5

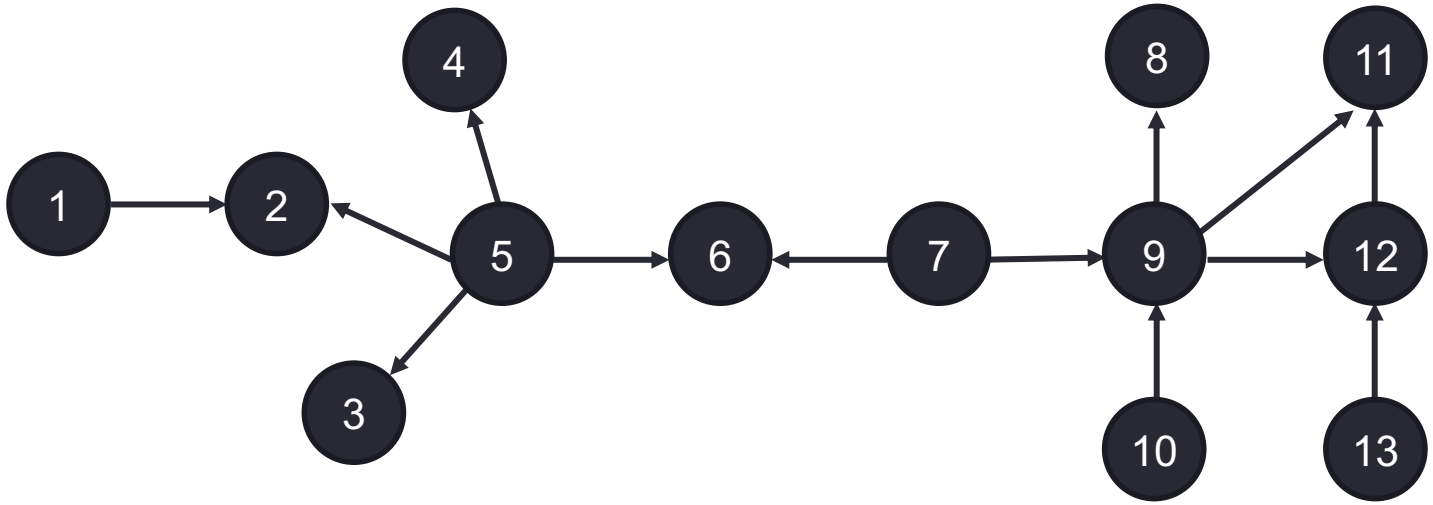
$$c(5) = \frac{|\{(5,2),(5,3),(5,4),(5,6)\}|}{|\{1,2,3,4,5,6,7,8,9,10,11,12,13\}|-1} = \frac{4}{13-1} = \frac{4}{12} = \frac{1}{3}$$

- Knoten 7

$$c(7) = \frac{|\{(7,6),(7,9)\}|}{|\{1,2,3,4,5,6,7,8,9,10,11,12,13\}|-1} = \frac{2}{13-1} = \frac{2}{12} = \frac{1}{6}$$

# Aufgabe 1 (d) - Aufgabenstellung

- Gegeben sei die folgende Illustration eines sozialen Netzwerks:




- Ermitteln Sie das Prestige für die Knoten 2 und 12. **(3 Punkte)**

# Wiederholung Vorlesung: Prestige

- Beobachtung
  - Hohe Bedeutung von Knoten mit vielen Verbindungen
  - Eingehende Kanten sind wertvoller als ausgehende
    (auch weil ausgehende Kanten „automatisch“ generiert werden können)

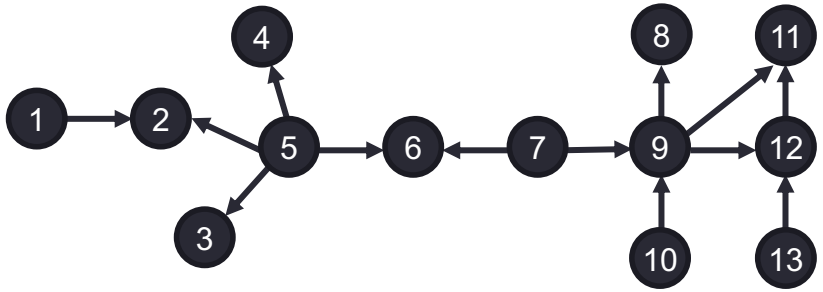
- Definitionen

Prestige: $p(e_i) = \frac{\text{Indegree}(e_i)}{|V|-1}$

# Aufgabe 1 (d) – Lösung (Prestige)

- Allgemein

$$p(e_i) = \frac{\text{Indegree}(e_i)}{|V|-1}$$




- Knoten 2

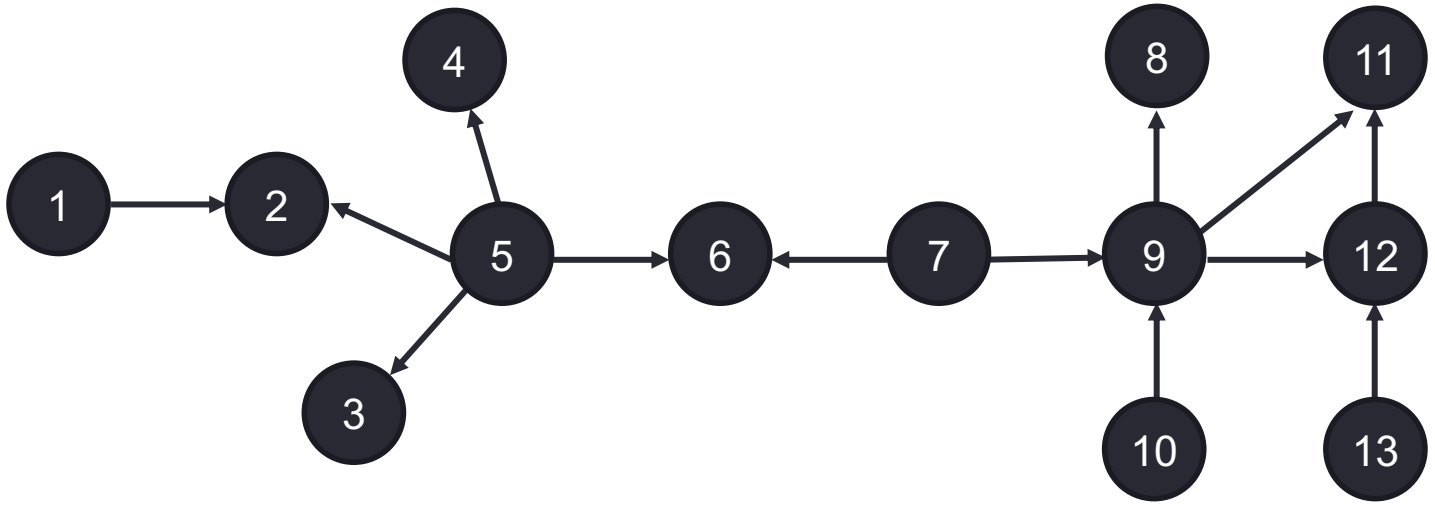
$$p(2) = \frac{|\{(1,2),(5,2)\}|}{|\{1,2,3,4,5,6,7,8,9,10,11,12,13\}|-1} = \frac{2}{13-1} = \frac{2}{12} = \frac{1}{6}$$

- Knoten 12

$$p(2) = \frac{|\{(9,12),(13,12)\}|}{|\{1,2,3,4,5,6,7,8,9,10,11,12,13\}|-1} = \frac{2}{13-1} = \frac{2}{12} = \frac{1}{6}$$

# Aufgabe 1 (e) - Aufgabenstellung

- Gegeben sei die folgende Illustration eines sozialen Netzwerks:



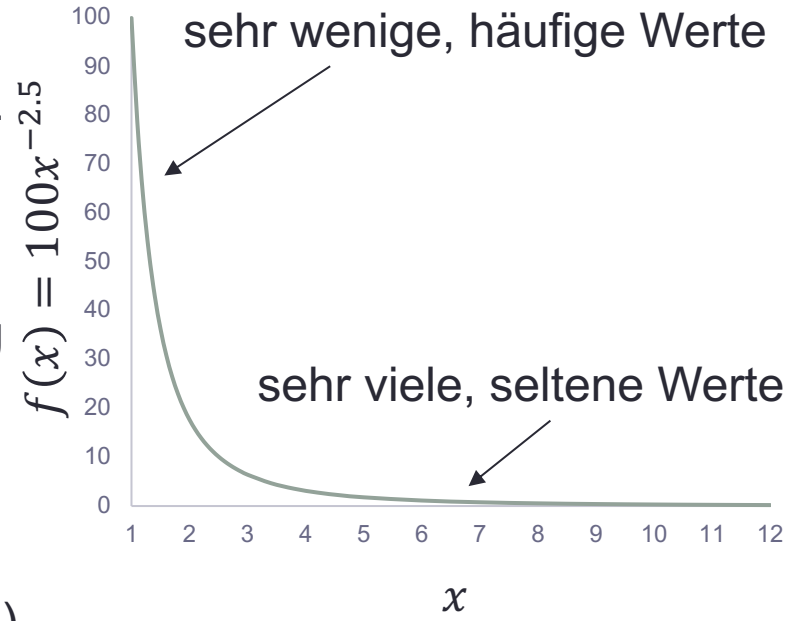

- Erläutern Sie die Bedeutung der Power Law Verteilung für moderne E-Commerce-Szenarien und diskutieren Sie die volkswirtschaftlichen Implikationen. **(5 Punkte)**

# Wiederholung Vorlesung: Power Law Verteilung

- Häufiger funktionaler Zusammenhang in sozialen Netzwerken

$$f(x) = ax^{-k}$$

  mit $2 \leq k \leq 3$ und $a > 0$




- Je größer $k$ umso „extremer“ die Verteilung...
  ... d.h. je größer $k$ umso weniger häufige

- Anzahl der Kanten pro Knoten in sozialen...
  ... Netzwerken folgt oft Power Law Verteilung
  (z.B. Internet, Facebook)

- Weitere Auftreten der Power Law Verteilung
  - Anzahl Besuche pro Webseite
    (z.B. Amazon hat das früh erkannt, GAFA!)
  - Größe von Mondkratern, Wordhäufigkeiten in vielen Sprachen,...

- Ursachen für Power Law Verteilung...
  ... vergleiche folgende Folien

# Aufgabe 1 (e) – Lösung (Power Law Verteilung)

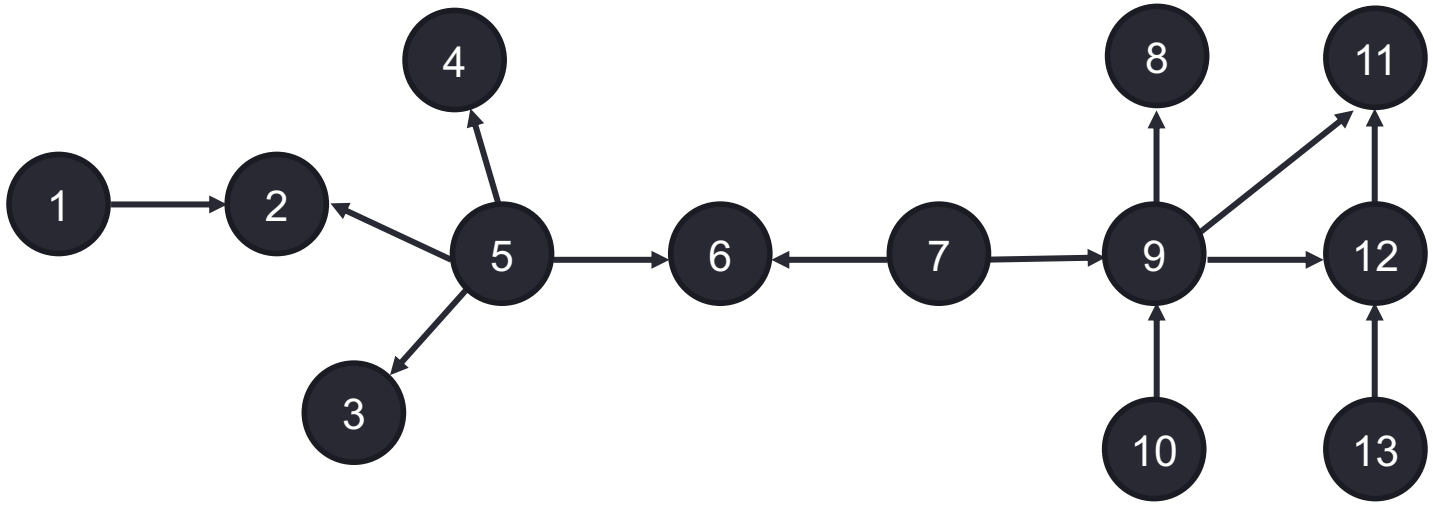
- Power Law Verteilung
  - Wenige Attributwerte sind sehr häufig
  - Meiste Attributwerte sehr selten

- Empirische Bedeutung
  - In Netzwerken oft Power Law Verteilungen
    (z.B. Anzahl von Kontakten, Indegree, Beiträge in Wikipedia)

- Konsequenz
  - Einzelne Mitglieder als Multiplikatoren für Werbung
  - Markteintrittbarieren für Wettbewerber von Amazon, ebay, etc.

# Aufgabe 1 (f) - Aufgabenstellung

- Gegeben sei die folgende Illustration eines sozialen Netzwerks:




- Erläutern Sie die Small World Eigenschaft. **(3 Punkte)**

# Wiederholung Vorlesung: Small World Eigenschaft

- Beobachtung
  - Jeder Mensch ist über 6 Kontakte mit allen anderen Menschen verbunden (nachgeprüft in den 1960er Jahren durch Milgram)
  - Allgemein gilt: Mittlere Pfadlänge $l(i,j)$ zwischen zwei Knoten...
    ... in einem Netzwerk mit $|V|$ Knoten ist gering

- Annahme (formal)

$$l(i,j) \sim \log(|V|)$$

- Anmerkungen
  - Eigenschaft wurde für mehrere Soziale Netzwerke nachgewiesen
  - Logarithmisches Wachstum ist obere Grenze...
    ... oft: (Sogar) sinkende Pfadlänge / sinkender Durchmesser
    - Hinzukommen neuer Kanten übersteigt Hinzukommen von Knoten
    - Maximale Pfadlänge (d.h. Durchmesser) sinkt
    - Mittlere Pfadlänge zwischen zwei Knoten sinkt

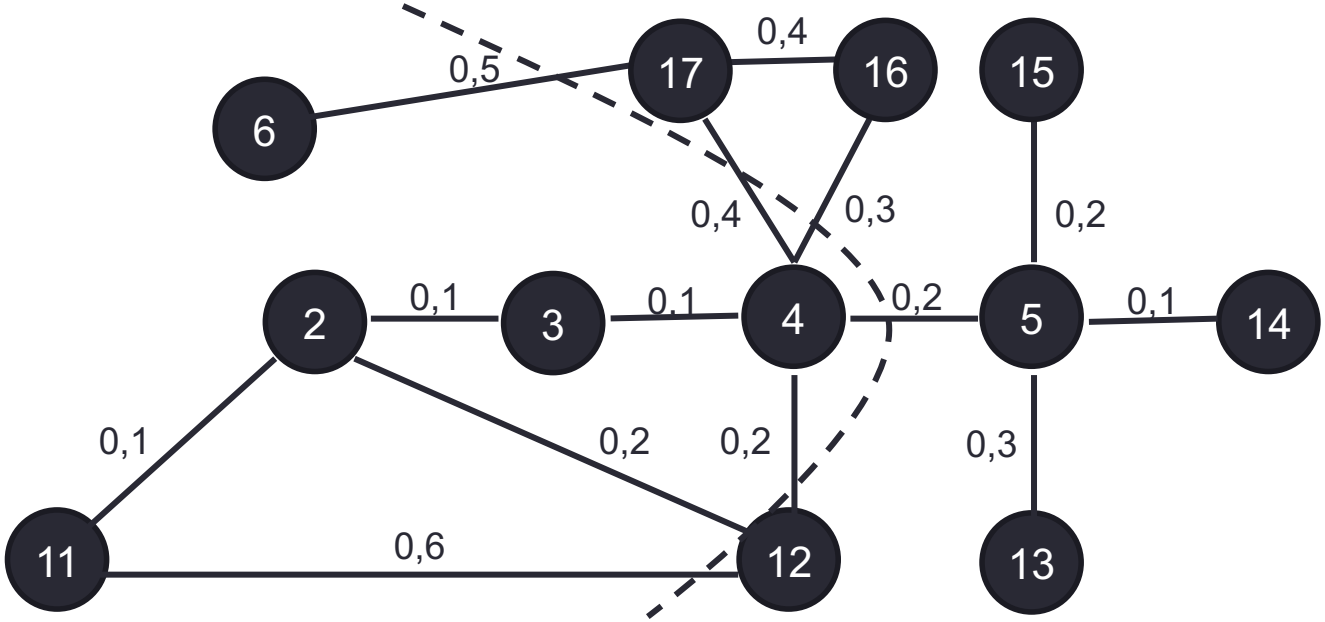
# Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining

- Aufgabenblatt 6 – Social Network Analysis
  - Aufgabe 1 – Kennzahlen
  - Aufgabe 2 – Kerningham-Lin
  - Aufgabe 3 – Verständnisfragen

# Aufgabe 2 - Aufgabenstellung

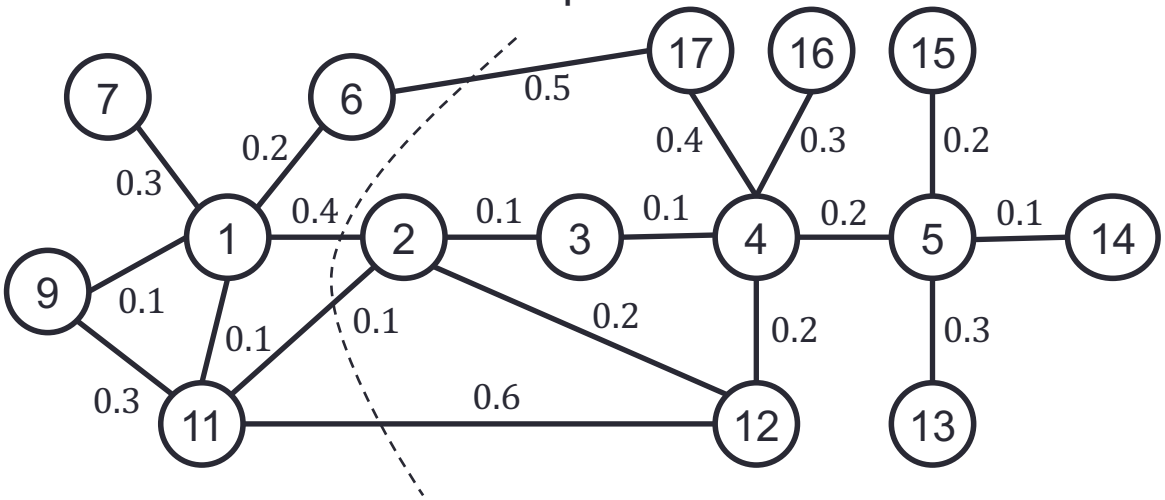
- Gegeben sei die folgende Illustration eines sozialen Netzwerks:




- Die gestrichelte Linie stellt eine initiale Unterteilung des sozialen Netzwerks in zwei Communities dar. Prüfen Sie mit Hilfe des Kerningham-Lin-Algorithmus, wie sich die Unterteilung in den nächsten drei Schritten verschieben würde. Prüfen Sie dabei die Zuordnung von Knoten in der rechten Community zur linken Community und betrachten Sie zuerst immer die in Frage kommenden Knoten mit der geringsten Identifikationsnummer. **(10 Punkte)**

# Wiederholung Vorlesung: Kernigham-Lin

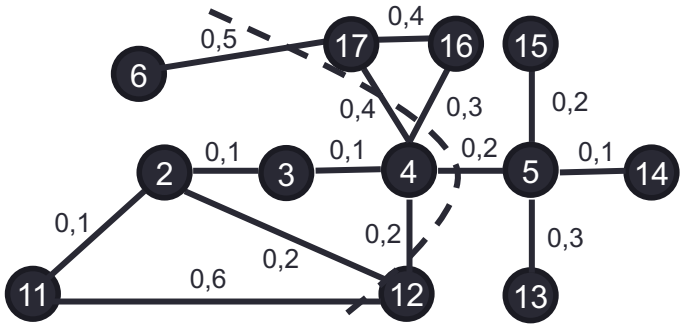
- Ausgangssituation: Gewichteter Graph




- Einfügen einer zufälligen Grenze (damit zwei Cluster $\gamma_1$ und $\gamma_2$)
- Für jeden Knoten (hier: $i \in \gamma_1$)
  - Berechnen der internen Kosten: $c_i^{int} = \sum_{j \in \gamma_1} w_{ij}$
  - Berechnen der externen Kosten: $c_i^{ext} = \sum_{j \in \gamma_2} w_{ij}$
  - Berechnen des Gewinns: $g_i = c_i^{ext} - c_i^{int}$
  - Gewinn $g_i$ entspricht Auszahlungsänderung bei Tausch von $i$ nach $\gamma_2$

# Aufgabe 2 – Lösung (Kerningham-Lin)

- Vorgehen allgemein:
  - Berechnen der internen Kosten
  - Berechnen der externen Kosten
  - Berechnen des Gewinns aus den Kosten
  - Ggfs. Tausch der Knoten

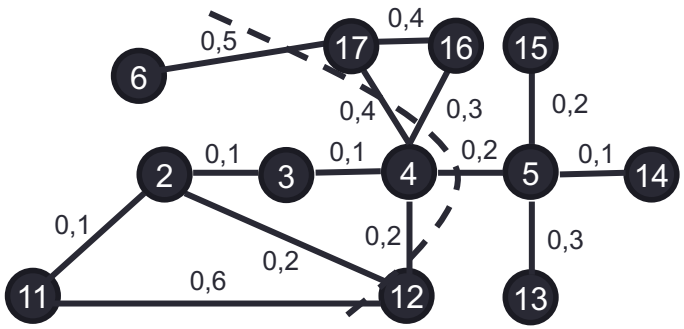



- Knoten 5
  - Berechnen der internen Kosten: $c_i^{int} = 0{,}2 + 0{,}1 + 0{,}3 = 0{,}6$
  - Berechnen der externen Kosten: $c_i^{ext} = 0{,}2$
  - Berechnen des Gewinns: $g_i = 0{,}2 - 0{,}6 = -0{,}4$

  ⇒ Kein Tausch

# Aufgabe 2 – Lösung (Kerningham-Lin)

- Vorgehen allgemein:
  - Berechnen der internen Kosten
  - Berechnen der externen Kosten
  - Berechnen des Gewinns aus den Kosten
  - Ggfs. Tausch der Knoten

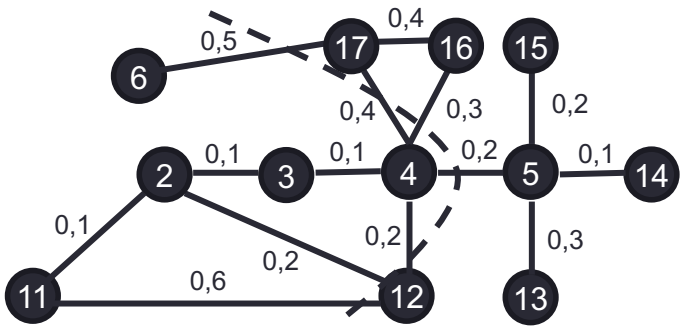



- Knoten 12
  - Berechnen der internen Kosten: $c_i^{int} = 0{,}0$
  - Berechnen der externen Kosten: $c_i^{ext} = 0{,}2 + 0{,}2 + 0{,}6 = 1{,}0$
  - Berechnen des Gewinns: $g_i = 1{,}0 - 0{,}0 = 1{,}0$

  ⇒ Tausch

# Aufgabe 2 – Lösung (Kerningham-Lin)

- Vorgehen allgemein:
  - Berechnen der internen Kosten
  - Berechnen der externen Kosten
  - Berechnen des Gewinns aus den Kosten
  - Ggfs. Tausch der Knoten




- Knoten 16
  - Berechnen der internen Kosten: $c_i^{int} = 0{,}4$
  - Berechnen der externen Kosten: $c_i^{ext} = 0{,}3$
  - Berechnen des Gewinns: $g_i = 0{,}3 - 0{,}4 = -0{,}1$

  $\Rightarrow$ kein Tausch

# Agenda

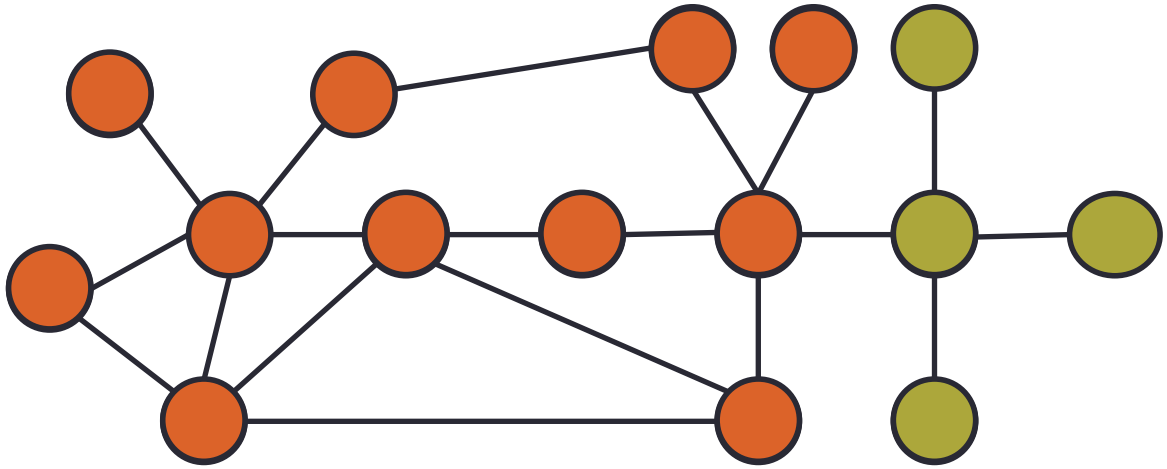
- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
- Aufgabenblatt 5 – Stream Mining

- Aufgabenblatt 6 – Social Network Analysis
  - Aufgabe 1 – Kennzahlen
  - Aufgabe 2 – Kerningham-Lin
  - Aufgabe 3 – Verständnisfragen

# Aufgabe 3 (a) - Aufgabenstellung

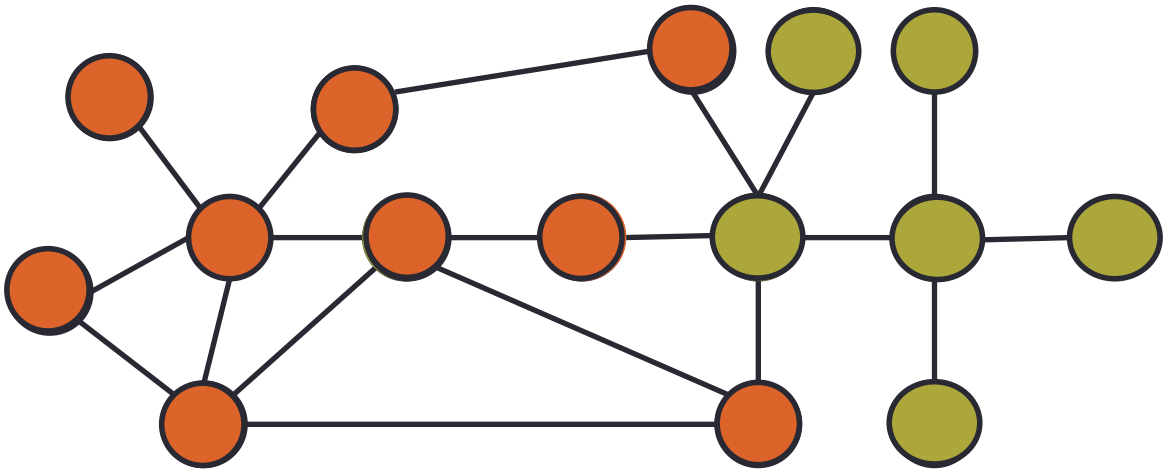
- Nennen Sie die zwei in der Vorlesung vorgestellten Algorithmen zur Collective Classification und beschreiben Sie diese kurz. **(9 Punkte)**

# Wiederholung: Iteratives Klassifikationsverfahren



- Algorithmus (Idee)
  - Bekannte Eigenschaften sind Eigenschaften der Nachbarn
  - Vorhersage Knoten mit bekannten Nachbarn
    (Iteration bis alle Knoten klassifiziert wurden)
  - Einsatz von bekannten Klassifikationsverfahren
    - Gewichtetes Mittel: $p(v_i = c|E_i) = \sum_{j|(i,j)\in E_i} w_{ij} \cdot p(v_j = c)$
    - Bayes Klassifikator: $p(v_i = c|E_i) = \frac{p(E_i|c)\cdot p(c)}{p(E_i)}$
- Herausforderung: Algorithmus muss konvergieren

# Wiederholung: Random Walk



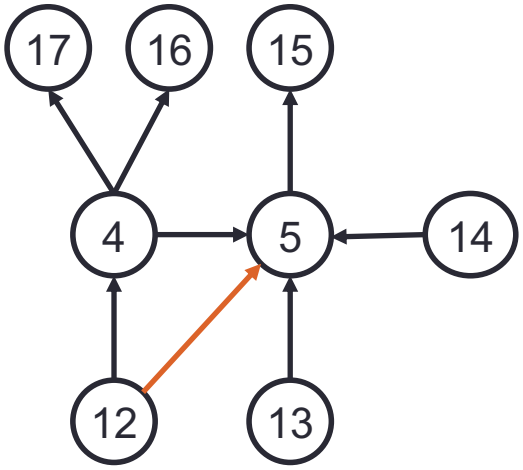
- Algorithmus (Idee)
  - Beginn an einem ungelabelten Knoten
  - Random Walk nach $n$ Schritten
  - Prüfen der Wahrscheinlichkeit für Klasse $p(v_i = c|E_i) = \sum_{j \in V_i} p(v_j = c)$
  - Wahl der Klasse mit höchster Wahrscheinlichkeit
- Hinweise
  - Algorithmus ist über Simulation lösbar (vgl. Beispiel)...
    ... Alternativ formal lösbar (Performanz!)
  - Weitere Verfahren für Collective Classification möglich

# Aufgabe 3 (b) - Aufgabenstellung

- Erläutern Sie, wie das Jaccard-Maß für die Link Prediction genutzt werden kann. **(5 Punkte)**

# Wiederholung Vorlesung: Link Prediction

- Jaccard Maß
  - Idee:
    Anzahl der gemeinsamen Nachbarn normalisiert
  - Formal: $CommonNeighbors(i,j) = \frac{|E_i \cap E_j|}{|E_i \cup E_j|}$
  - Beispiel: $CommonNeighbors(5,12) = \frac{|\{4\}|}{|\{4,13,14,15\}|} = \frac{1}{4}$
  - Nachteil:
    Bedeutung der Kontakte (Prestige!) unberücksichtigt



# Aufgabe 3 (b) – Lösung (Link Prediction)

- Idee
  Wahrscheinlichkeit jemanden kennenzulernen…
  … steigt mit der Anzahl gemeinsamer Kontakte

- Jaccard Maß bildet dies ab
  - Zähler: Anzahl gemeinsamer Kontakte
  - Nenner: Anzahl aller Kontakte der Knoten

- Ausgehend von einem Knoten wird…
  … Jaccard Maß zu allen anderen Knoten im Netzwerk ermittelt

- Wahrscheinlichster neuer Kontakt ist der mit dem höchsten Jaccard Maß

# Aufgabe 3 (c) - Aufgabenstellung

- Erläutern Sie, wie Preferential Attachment dazu führen kann, dass soziale Netzwerke für neue Teilnehmer unattraktiv werden. **(5 Punkte)**

# Wiederholung: Preferential Attachment

- Beobachtung
  - Die Wahrscheinlichkeit von Verknüpfungen zu anderen Knoten $\pi(e_i)$...
    ... steigt mit Grad eines Knotens
  - Analog zu Realwelt: Stark vernetzte Knoten finden leichter neue Kontakte

- Annahme (formal)

$$\pi(e_i) \sim d(e_i)^{\alpha}$$

  wobei $\alpha$ stark von Anwendungsdomäne abhängt

- Annahme im Kontext von Online meist „Skalenfreiheit“, ...
  ... d.h. $\alpha \approx 1$ (Zusammenhang ist linear)

- Anmerkungen
  - Bonus in der realen Welt: Kontakte verblasen!
  - Attraktivität für Beiträge in sozialen Netzwerken sinkt mit der Zeit
    (Wikipedia, MySpace, meine Vermutung: Facebook wird folgen, ...)

# Aufgabe 3 (c) – Lösung (Preferential Attachment)

- Idee
  - Verknüpfungen bevorzugt zu Teilnehmern mit vielen Kontakten
  - Wahrscheinlichkeit für neue Kontakte für neue Teilnehmer gering

- Problem
  Ist soziales Netzwerk älter…
  … ist der Aufstieg zu „populärem“ Knoten schwierig

- Konsequenz
  Soziales Netzwerk für neue Gruppen uninteressant …
  … da diese sich nicht schnell untereinander verbinden

# Aufgabe 3 (d) - Aufgabenstellung

- Gehen Sie davon aus, Sie erhalten die Aufgabe für ein existierendes soziales Netzwerk einen möglichst sinnvollen Werbeträger für ein neues Produkt zu identifizieren. Wie gehen Sie vor und diskutieren Sie welchen Vorteil hierbei die Netzwerkstruktur gegenüber einem unstrukturierten Netzwerk besitzen.

**(6 Punkte)**

# Aufgabe 3 (d) – Lösung (Marketing)

- Idee
  - Wahl von Teilnehmern mit hohem Prestige oder hoher Zentralität
  - Bewerbung von Produkt durch diese Teilnehmer
  - Ist Person glaubwürdiger Nutzer des Produkts (Homophilie!)…
    … ist die Wahrscheinlichkeit groß, dass Peergroup Produkt kauft
  - Aktuell leben mehrere Facebook Sternchen von diesem Konzept

- Attraktivität des Ansatzes
  - Geringe Streuverluste
  - Junges Publikum kaum anders zu erreichen
  - Werbung wird nicht als Werbung wahrgenommen
  - Höhere Glaubwürdigkeit der Werbung

# Aufgabe 3 (e) - Aufgabenstellung

- Diskutieren Sie, welche Gefahren vom Einsatz von Big Data für den Einzelnen ausgehen und welche Vorteile Big Data besitzt. **(5 Punkte)**

# Aufgabe 3 (e) – Lösung (Big Data)

- Risiken
  - Privatwirtschaft als Datensammler
    - Wenige Anbieter haben Monopol auf Daten
      (vgl. Britta Wasserfilter bei Amazon oder Apples Strategie bei Covid19)
    - Primärkreislauf verstärkt Effekt
  - Überforderung der Politik mit Thema
    - Möglichkeiten gegenzusteuern werden nicht erkannt / genutzt
    - Internationaler Unterbietungswettbewerb bei Kontrolle
  - Beobachtung des einzelnen sehr tief
    - Amazon kennt Konsumverhalten, Stimmungen, …
    - Apple kennt Freunde, Bewegungsprofile, …
    - Wenn dieses Wissen an einen Geheimdienst kommt…

- Zentraler Vorteil:
  - Menge konsumierter, irrelevanter Werbung wird reduziert