Big Data Anwendungen

Social Network Analysis

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining

- Social Network Analysis
  - Deskriptive Eigenschaften Sozialer Netzwerke
  - Community Detection
  - Collective Classification
  - Link Prediction

- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Wandel der Kommunikation

- Früher (also vor ca. 20 Jahren)
  - Face-to-Face Kommunikation
  - Analoge Post
  - Telephonie

- Seitdem
  - E-Mail
  - Kurznachrichtendienste (SMS, Whatsapp)
  - Blogs (erst persönliche Webseiten wie Geocities, MySpace)
  - Bild-Telefonie (erste Ansätze in den 90ern, Skype, Facetime)
  - Social Network Sites (Twitter, Xing, Facebook)

- Wesentliche Änderung
  - Verstärktes Gefühl der Nähe (über visuelle Übertragung)
  - Möglichkeit auch entfernte Bekannte kontinuierlich zu informieren
  - Neues Marketingpotential (über Konsumverhalten aller Bekannten)
  - Perfekte Überwachungsmöglichkeit (Freunde finden, aber auch mit Rest)

Implikationen des Wandels

- Social Network Forschung
  - Hundert Jahre alte Disziplin (insb. Soziologie)
  - Früher Mangel an technischen Untersuchungsmöglichkeiten
  - Studien sehr aufwendig (z.B. Milgram‘s „Six degrees of seperation“, 60er)

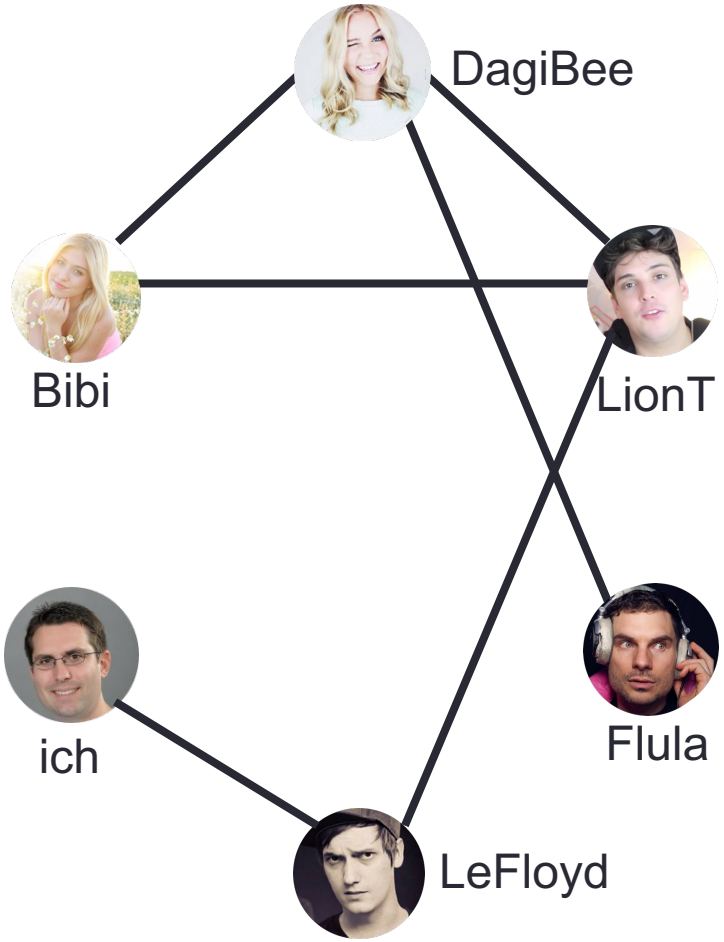
- Indirekte Kommunikation (E-Mail, Messanger)
  - Früher primär direkte Kommunikation
  - Heute Entwicklung Richtung indirekter Kommunikation

- Content Sharing
  - Austausch von bestehendem Content möglich
  - Erstellung und Bereitstellung eigenen Contents
  - Großes Analyse Potential (unstrukturierte Daten!)

- Möglichkeit Soziale Netzwerke indirekt abzuleiten
  - Zitationsnetzwerke (wer kennt wen, über wer zitiert wen)
  - Bewegungsprofile (Erkennung von Änderungen im Beziehungsstatus)

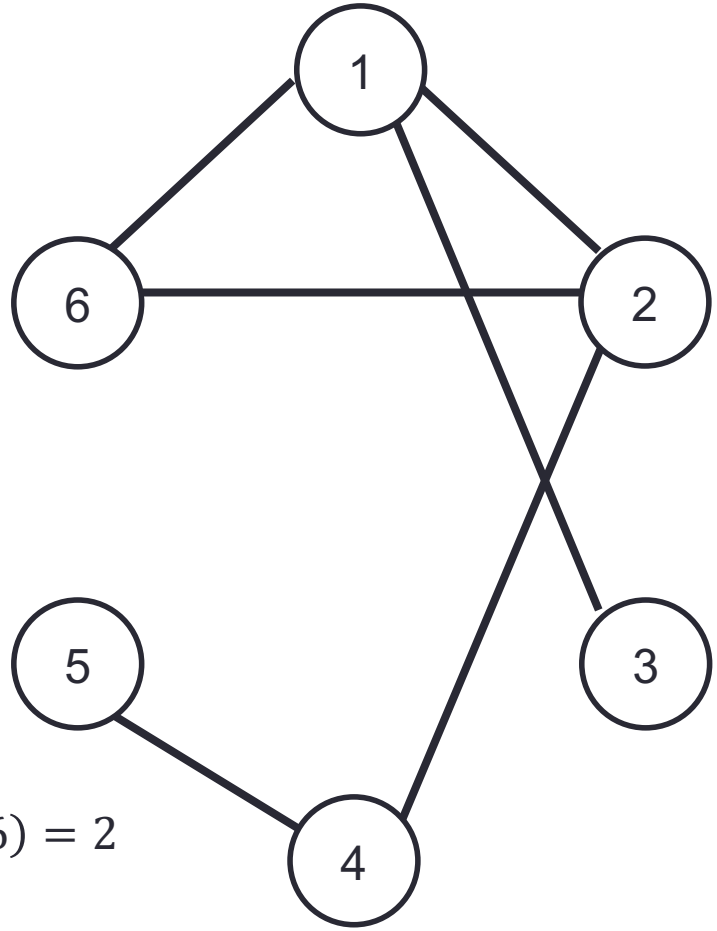
Abbildung Sozialer Netzwerke auf Graphen

- Zwei Elemente
  - Knoten: Teilnehmer
  - Kanten: Verbindungen
- Optional
  - Kantengewichte: Stärke der Verbindung
  - Richtung der Kanten: Ursache der Kante
  - Strategie: Verhalten der Knoten
- Beobachtungen
  - Isolierte Knoten
  - Untereinander stark vernetzte Knoten
  - Knoten als Verbinder unterschiedlicher... ... Teilnetze
- Zentrale Punkte
  - Identifikation von stark vernetzten Knoten
  - Analyse der Struktur



Formalisierung von Graphen

- Graph $G = \{V, E\}$ besteht aus
  - Einer Menge Knoten V
  - Einer Menge Kanten E
- Knoten werden oft mit Ganzzahlen...
  ... repräsentiert $V = \{1, 2, 3, 4, 5, 6\}$
- Kanten werden als Vektoren der...
  ... verbundenen Knoten dargestellt...
  ... $E = \{(1,2), (1,3), (1,6), (2,4), (2,6), (4,5)\}$
- Jeder Knoten hat einen Grad...
  ... $d(e_i) = |\{(j,k): j = i \vee k = i\}|$...
  ... (Anzahl der Kanten am Knoten)...
  ... $d(1) = |(1,2), (1,3), (1,6)| = 3$
  $d(2) = 3;\ d(3) = 1; d(4) = 2; d(5) = 1; d(6) = 2$
- Zwei Knoten haben Abstand $dist\ (e_i, e_j)$...
  ... d.h. minimale Pfadlänge zwischen beiden Knoten...
  ... $dist\ (1,2) = 1;\ dist\ (1,3) = 1;\ dist\ (1,4) = 2;\ dist\ (1,5) = 3;\ dist\ (1,6) = 1$



Soziale Homophilie

- Beobachtung
  „Gleich und Gleich gesellt sich gern!“

- Implikationen für Netzwerke
  - Vernetzte Knoten ähneln einander oft bzgl.
    Geschlecht, ethnischer Herkunft, sozioökonomischen Status, Bildung, ...
  - Vernetzte Knoten haben oft ähnliche Interessen, Hobbies, ..

- Implikationen für die Knoten
  - Erleichterte Kommunikation und Koordination von Handlungen/Aktivitäten
    (Relevanz für Marketing – Facebook verkauft entsprechende Daten!)
  - Selektiver Informationsgewinn (wg. Gruppendenken, ...)
    (Donald Trump: Um Ausschluss vorzukommen, keine Widerrede, ...)

- Additivität
  - Sind Knoten in verschiedenen Netzwerken verbunden...
    ... impliziert das oft stärkere Bindung
  - Ableitung oft schwierig (Untersuchung verschiedener Netzwerke, ...)

Triadic Closure

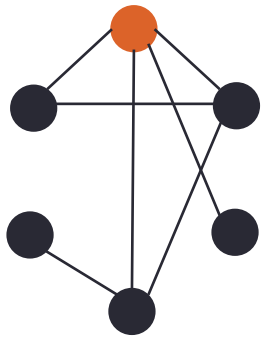
- Idee
  - Entsprechung der sozialen Homophilie bzgl. Netzwerk
  - Knoten mit ähnlichen Kontakten sind auch mit ähnlichen Knoten verknüpft (Xing: „Erweitern Sie Ihr Kontaktnetzwerk“, Facebook: „Freunde finden“)

- 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: $\eta(0) = \frac{2}{6} = \frac{1}{3}$

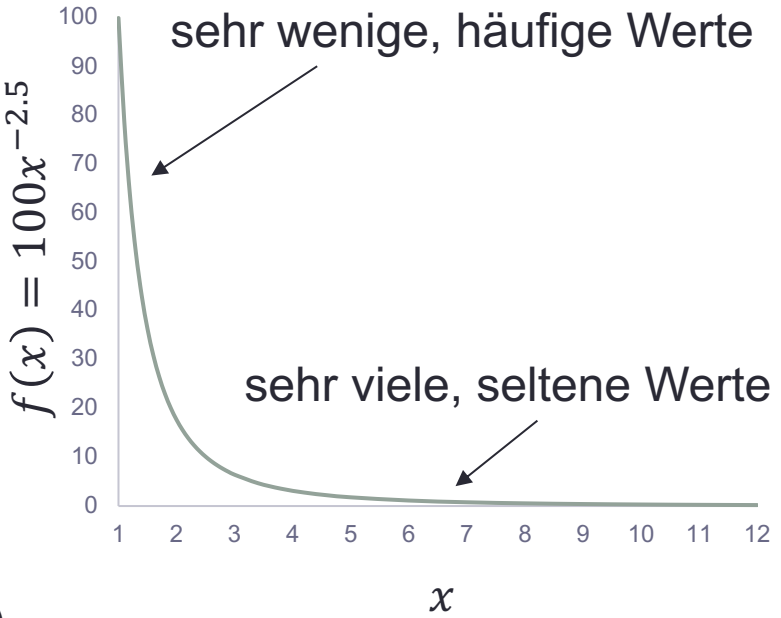


Power Law Verteilung

- 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

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 α 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, ...)

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

Verdichtung

- Beobachtung
  - Mehr Knoten und Kanten kommen hinzu als gelöscht werden
  - Anzahl neuer Kanten übersteigt meist Anzahl neuer Knoten

- Annahme (formal)

$$|V| \sim |E|^{\beta}$$

  wobei $1 \leq \beta \leq 2$

- Besondere Werte für β
  - $\beta = 1$: Grad der Knoten nicht durch die Größe des Netzwerks beeinflusst
  - $\beta = 2$: $|V|/|E|$ wächst gleich schnell wie $|E|$

- Anmerkungen
  - Giant connected component entsteht
    Bereich im Graph der sehr stark verbunden ist
  - Hubs entstehen
    Einzelne Knoten verbinden – eigentlich unzusammenhängende Knoten

Zentralität und Prestige

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

- Definitionen

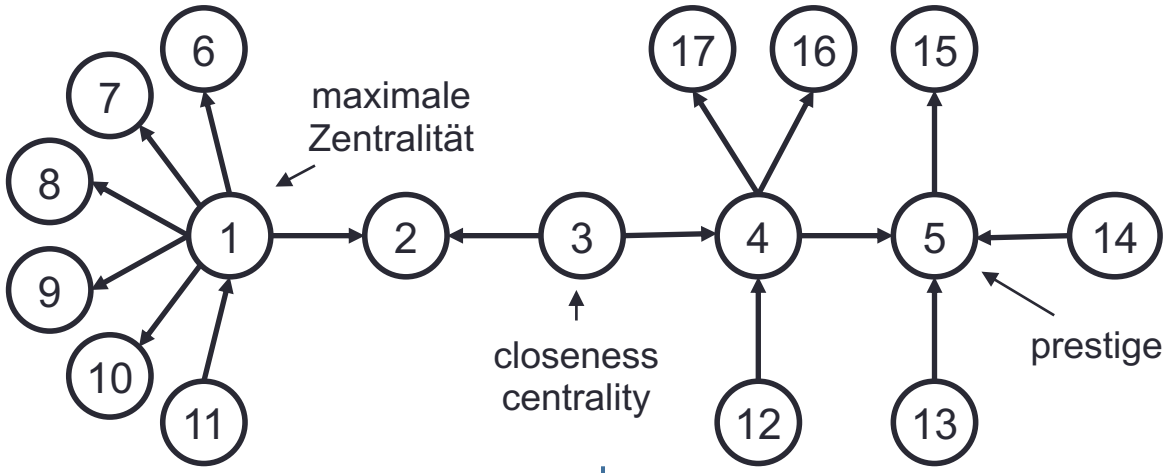
$$\text{Zentralität: } c(e_i) = \frac{d(e_i)}{|V|-1}$$

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

- Anmerkungen
  - Oft „Closeness Centrality“: $c_c(e_i) = \frac{|V|-1}{\sum_{j=1}^{|V|} dist(e_i,e_j)}$

    (Zentralität bevorzugt oft zentrale Knoten am Rand)
  - Prestige nur für „gerichtete“ Graphen möglich
    (Anmerkung: Pagerank [Algorithmus hinter Google] ähnlich Prestige)

Zentralität und Prestige – Beispiel



Zentralität

- $c(e_i) = \frac{d(e_i)}{|V|-1}; c_c(e_i) = \frac{|V|-1}{\sum_{j=1}^{|V|} dist(e_i,e_j)}$
- $c(1) = 7/16; c_c(1) = 16/43$
- $c(2) = 2/16; c_c(2) = 16/40$
- $c(3) = 2/16; c_c(3) = 16/39$
- $c(4) = 5/16; c_c(4) = 16/40$
- $c(5) = 4/16; c_c(5) = 16/48$

Prestige

- $p(e_i) = \frac{\text{Indegree}(e_i)}{|V|-1}$
- $p(1) = 1/16$
- $p(2) = 2/16$
- $p(3) = 0/16$
- $p(4) = 2/16$
- $p(5) = 3/16$

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining

- Social Network Analysis
  - Deskriptive Eigenschaften Sozialer Netzwerke
  - Community Detection
  - Collective Classification
  - Link Prediction

- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Community Detection

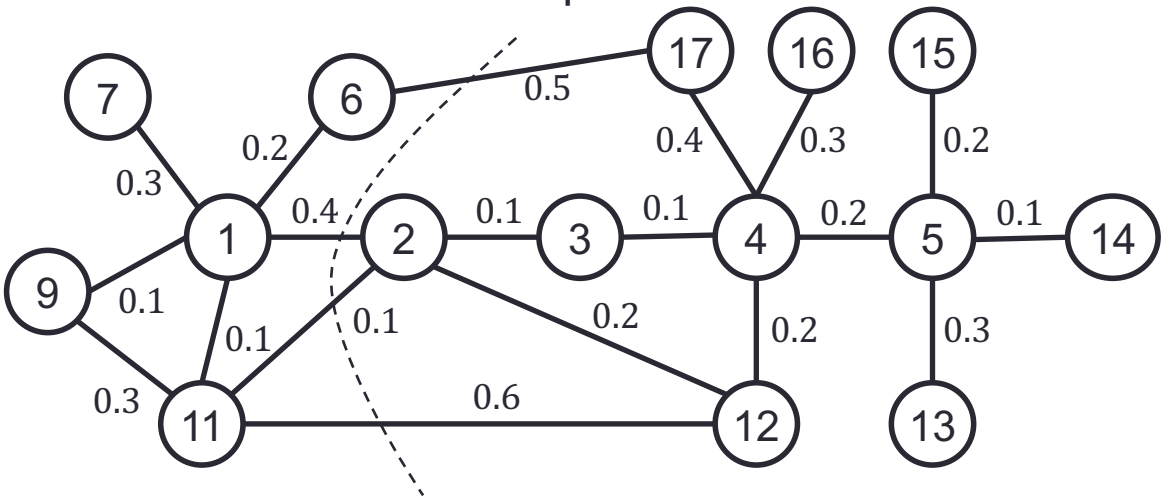
- Idee (analog k-Means Clustering)
  Finden von Clustern im sozialen Netzwerk in dem...
  ... die Knoten des Clusters eine geringe Distanz besitzen, ...
  ... während alle anderen Knoten weit entfernt sind

- Typische Eigenschaften bei sozialen Netzwerken
  - Distanz zweier Knoten als Distanzmaß nicht feingranular genug
    (insbesondere bei kleinen Netzwerken)
  - Hubs verbinden Cluster
    (... und Verringern damit den Abstand zwischen Clustern)
  - Unterschiedlich dichte Bereich im Netzwerk
    (Kennzahlen lokal stark unterschiedlich ⇒ globale Bewertung schwierig)
  - Giant Components führen zu unbalancierten Clustern
    (Dicht verknüpfte Teilgraphen beeinflussen Gesamtgraph)

- Lösung
  - Spezielle „Community Detection“ Algorithmen
    (insbesondere auf gewichteten Graphen, d.h. Kante (i, j) hat Gewicht w_{ij}

Kernigham-Lin Algorithmus I

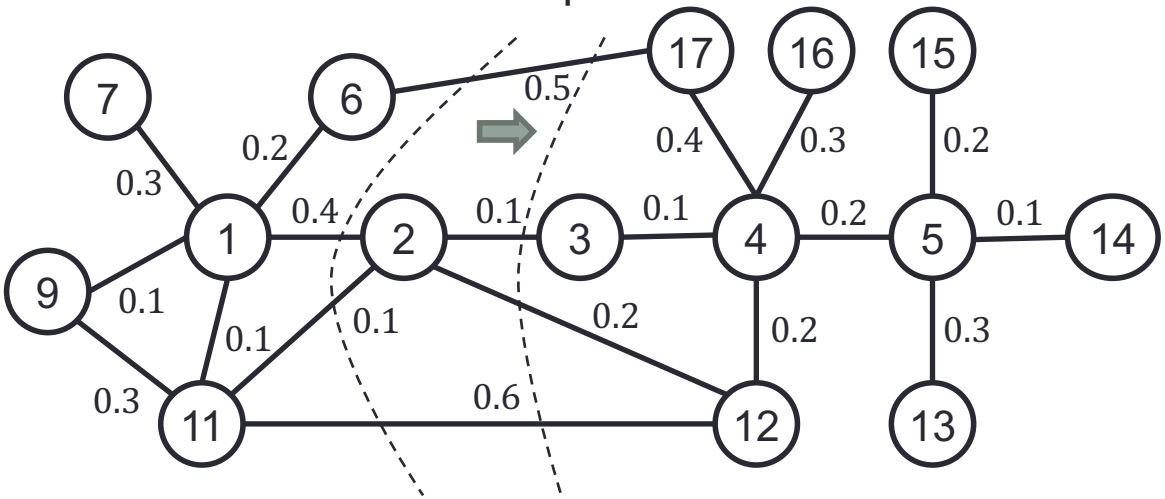
- Ausgangssituation: Gewichteter Graph



- Einfügen einer zufälligen Grenze (damit zwei Cluster γ_1 und γ_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 γ_2

Kernigham-Lin Algorithmus II

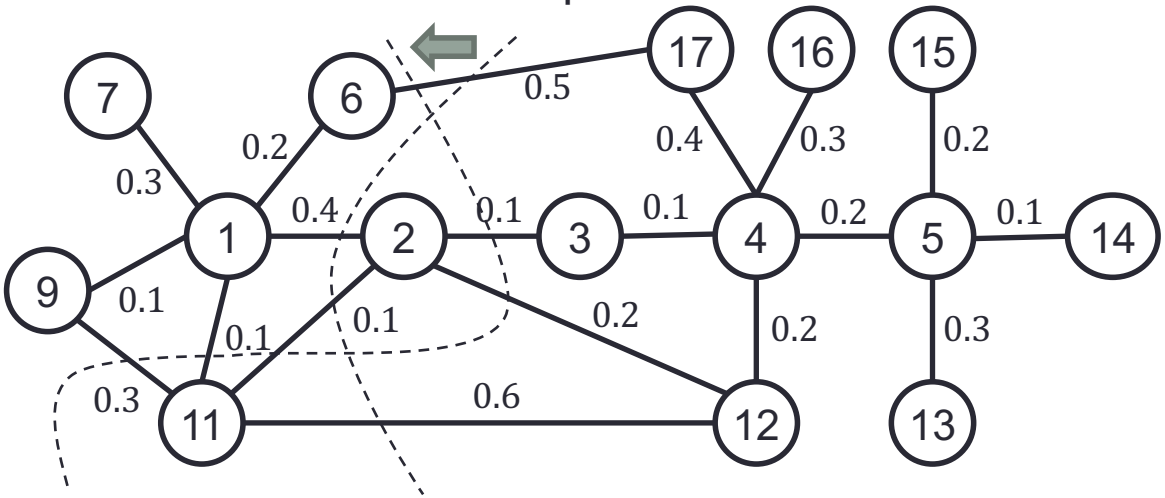
- Ausgangssituation: Gewichteter Graph



- Einfügen einer zufälligen Grenze (damit zwei Cluster γ_1 und γ_2)
- Für jeden Knoten (hier: $2 \in \gamma_1$)
  - Berechnen der internen Kosten: $c_2^{int} = \sum_{j \in \gamma_1} w_{2j} = 0.1 + 0.2 = 0.3$
  - Berechnen der externen Kosten: $c_2^{ext} = \sum_{j \in \gamma_2} w_{2j} = 0.4 + 0.1 = 0.5$
  - Berechnen des Gewinns: $g_2 = c_2^{ext} - c_2^{int} = 0.5 - 0.3 = 0.2$
  - Gewinn g_2 entspricht Auszahlungsänderung bei Tausch von 2 nach γ_2

Kernigham-Lin Algorithmus III

- Ausgangssituation: Gewichteter Graph

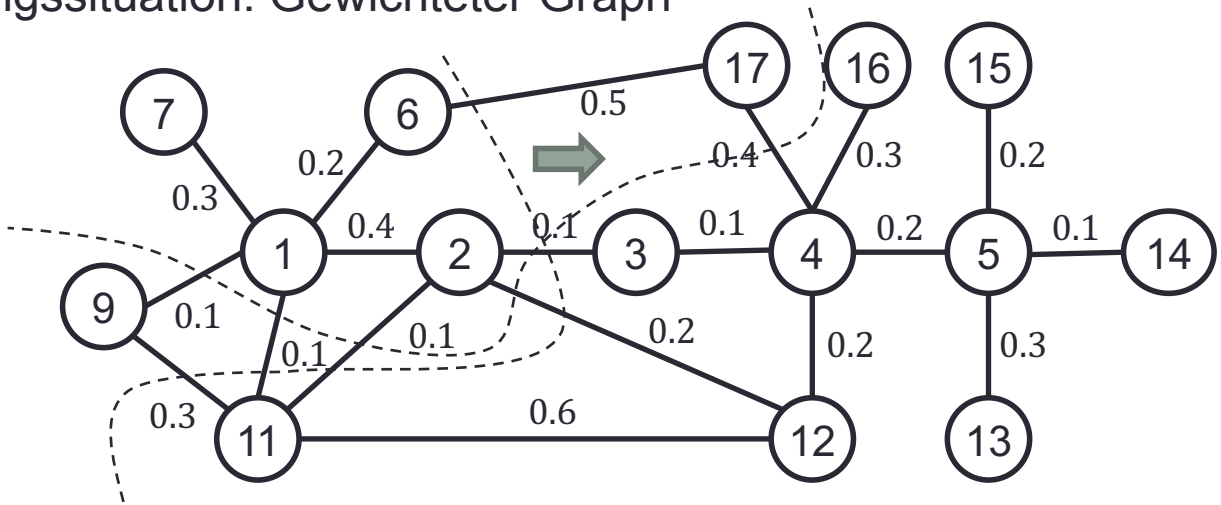


- Typischerweise: Austausch von Knoten zwischen den Clustern i, j
  - Kosten abzüglich der Gewichte, die künftig in gleichem Cluster bleiben
  - Berechnen des Gewinns: $G_{ij} = g_i + g_j - 2w_{ij}$

- Beispiel
  - $g_2 = 0.5 - 0.3 = 0.2$
  - $g_{11} = 0.6 - (0.3 + 0.1 + 0.1) = 0.1$
  - $G_{2;11} = g_2 + g_{11} - w_{2;11} = 0.2 + 0.1 - 2 \cdot 0.1 = 0.1$ ⇨ Tausch!

Kernigham-Lin Algorithmus IV

- Ausgangssituation: Gewichteter Graph

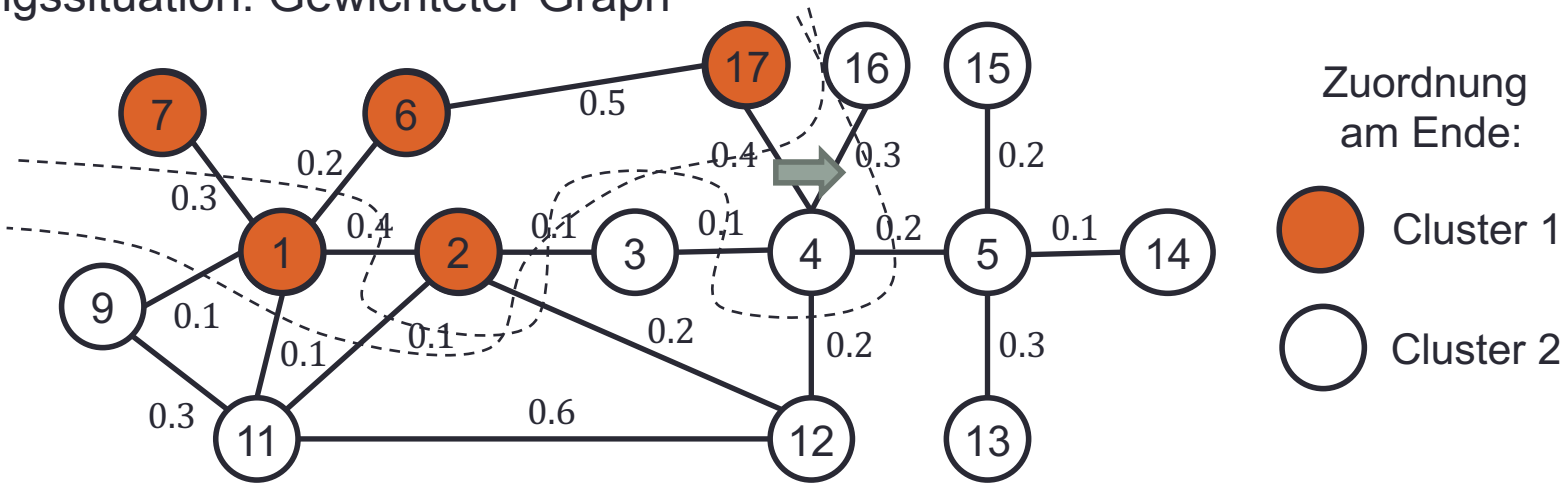


- Tausch wird nicht direkt ausgeführt, sondern so lange gesucht, ...
  ... bis keine weitere Verbesserung möglich

- Beispiel
  - $G_{2;11} = 0.1$
  - $g_{17} = 0.5 - 0.4 = 0.1$
  - $g_9 = 0.3 - 0.1 = 0.2$
  - $G_{9;17} = g_{17} + g_9 - w_{9;17} = 0.1 + 0.2 - 2 \cdot 0.0 = 0.3$ ⇨ Tausch!

Kernigham-Lin Algorithmus V

- Ausgangssituation: Gewichteter Graph



- Tausch wird nicht direkt ausgeführt, sondern so lange gesucht, ...
  ... bis keine weitere Verbesserung möglich

- Beispiel
  - $G_{2;11} = 0.1; G_{9;17} = 0.3$
  - $g_1 = (0.1 + 0.1) - (0.3 + 0.2 + 0.4) = -0.7$
  - $g_4 = 0.4 - (0.3 + 0.2 + 0.1) = -0.2$
  - $G_{1;4} = g_1 + g_4 - w_{1;4} = -0.7 - 0.2 - 2 \cdot 0.0 = -0.9$ ⇨ Kein Tausch!

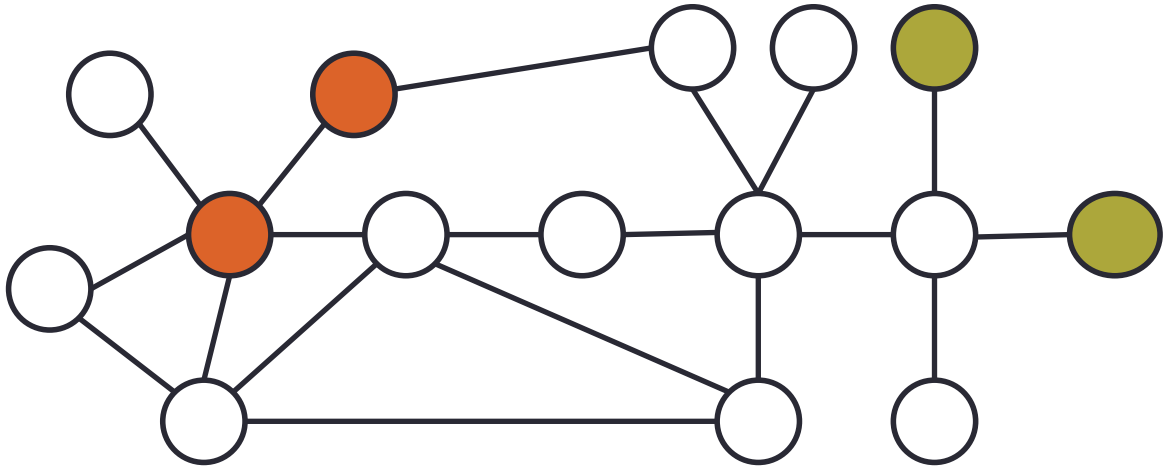
Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining

- Social Network Analysis
  - Deskriptive Eigenschaften Sozialer Netzwerke
  - Community Detection
  - Collective Classification
  - Link Prediction

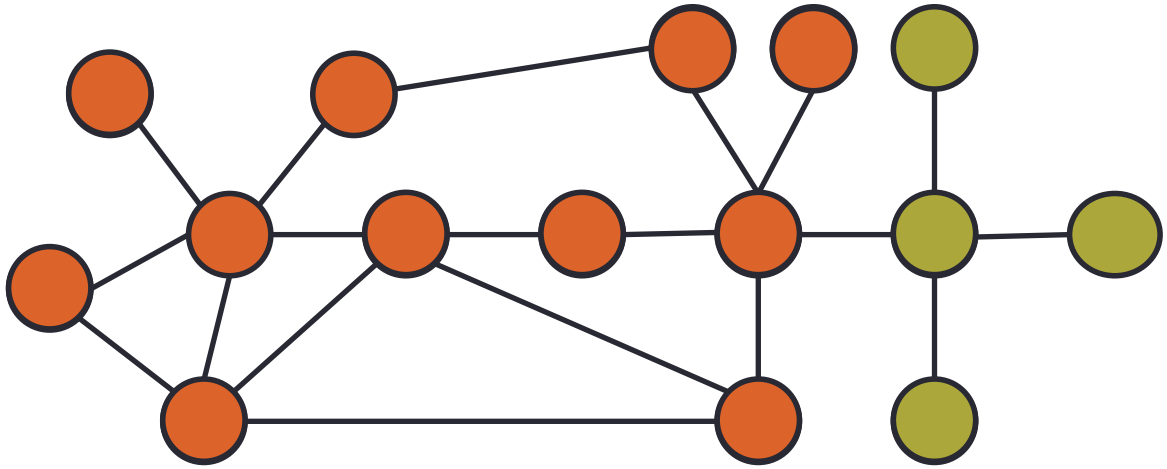
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Collective Classification



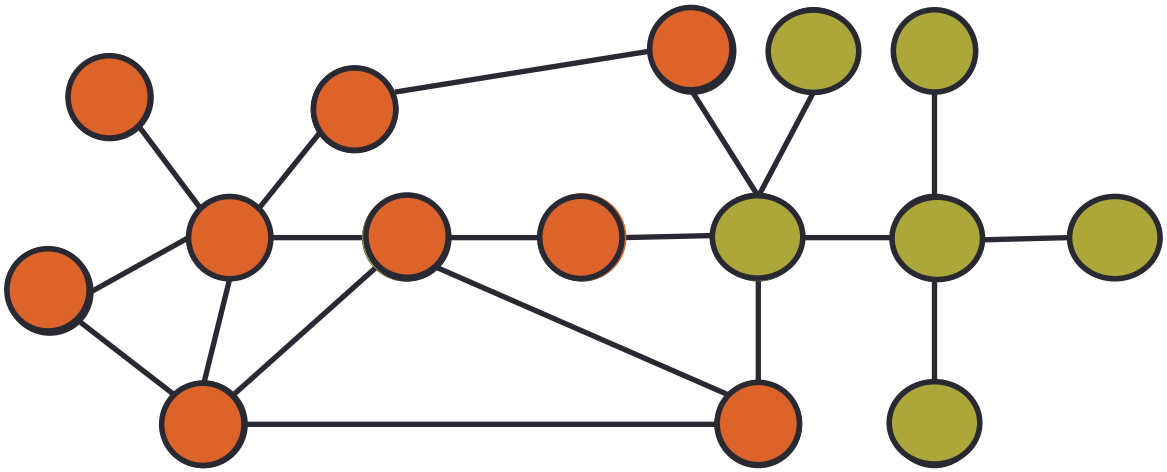
- Idee
  - Ungerichteter Graph bildet soziales Netzwerk ab
  - Für Teil der Knoten ist eine Eigenschaft bekannt (hier: ist orange, ist grün)
  - Ableiten der Eigenschaft (orange, grün) für alle anderen Knoten
- Voraussetzung
  - Homophilie Eigenschaft ist erfüllt (verbundene Knoten sind ähnlich) oder...
  - ... gegenseitige Beeinflussung der Knoten
- Hier: Fokus auf binäre Attribute
  (analog auch für kategorische, numerische Attribute)

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

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

Agenda

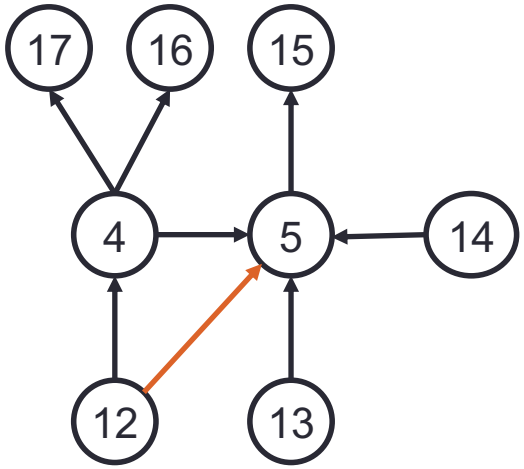
- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining

- Social Network Analysis
  - Deskriptive Eigenschaften Sozialer Netzwerke
  - Community Detection
  - Collective Classification
  - Link Prediction

- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

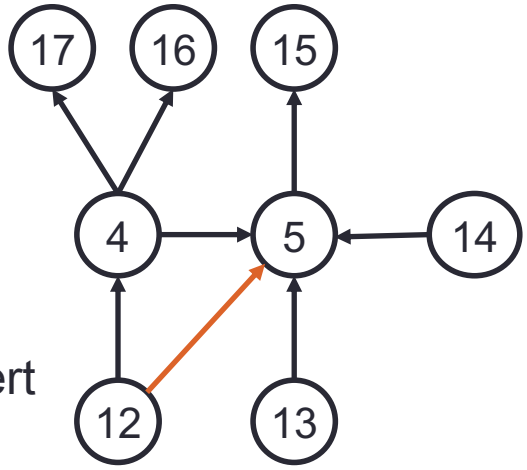
Link Prediction

- Ursachen für Vorhersage von Verbindungen
  - Netzwerk ändert sich über die Zeit
  - Verbindungen (bspw. aus der offline Welt) sind unbekannt
  - Verbindungen sind unzuverlässig
- Mögliche Vorhersagen
  - Existenz von Verbindungen
  - Gewicht einer Verbindung
  - Art der Verbindung (Richtung, ...)
- Grundlegende Idee
  - Für zwei Knoten wird Überschneidung in... ... Nachbarschaft bewertet
  - Ist Überschneidung hoch, ... wird mit neuen Verbindungen gerechnet
  - Verfahren unterscheiden sich im Maß für die Überschneidung



Maße zur Ermittlung der Überschneidung

- Common Neighbor Maß
  - Idee: Anzahl der gemeinsamen Nachbarn
  - Formal: $CommonNeighbors(i,j) = |E_i \cap E_j|$
  - Beispiel: $CommonNeighbors(5,12) = |\{4\}| = 1$
  - Nachteil: Anzahl der Kontakte unberücksichtigt



- 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

- Adamic-Adar Maß
  - Idee: Anzahl der gemeinsamen Nachbarn normalisiert
  - Formal: $AdamicAdar(i,j) = \sum_{k \in E_i \cap E_j} \frac{1}{\log(|E_k|)}$
  - Beispiel: $AdamicAdar(5,12) = \frac{1}{\log(4)} = 1.6610$