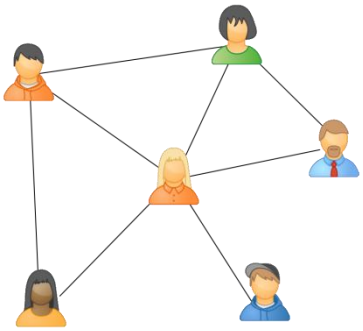
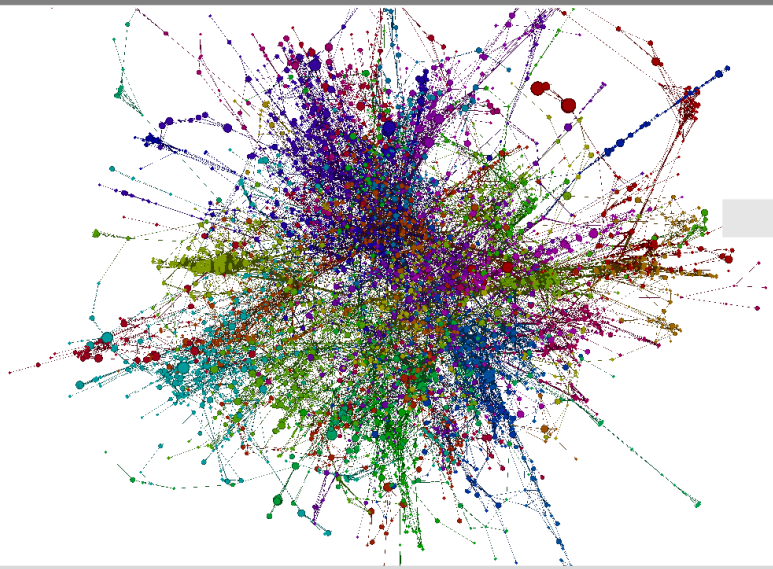
Karlsruher Institut für Technologie

Synergien aus Graph-Theorie und Data-Mining für die Analyse von Netzwerkdaten

Tanja Hartmann, Patricia Iglesias Sánchez, Andrea Kappes, Emmanuel Müller und Christopher Oßner

IPD – Institut für Programmstrukturen und Datenorganisation
ITI – Institut für Theoretische Informatik





Age	Profession	Gender	Publications
21	Student	Female	1
23	Student	Male	0
30	Lecturer	Female	25
29	Student	Male	2
27	Student	Female	3
25	Student	Male	0

Beispiel: Soziale Netzwerke

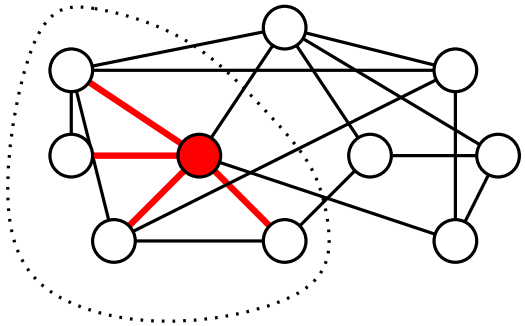
- Modell für Beziehungen zwischen sozialen Entitäten
  - Vernetzung von Freunden, Zusammenarbeit von Firmen, …
- Analyse realer Netzwerke
  - Nicht Plattformen, sondern die zu Grunde liegenden Netzwerke
- Soziale Netzwerke als Graph
  - Knoten repräsentieren Personen, Kanten die Beziehungen
  - Charakteristika: Knotengrad, Durchmesser, durchschnittliche Distanz, Clusterkoeffizient

Karlsruher Institut für Technologie

SEMINARTHEMEN

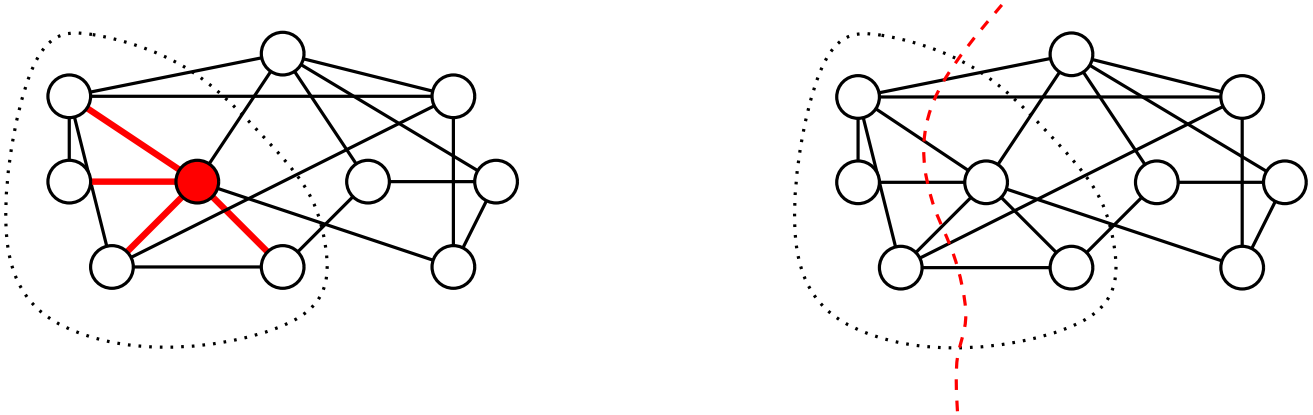
Thema 1: Motivation - Anwendungsbeispiel

- Suche nach Communities mit garantierten Eigenschaften
  - Hier: Pro Konten mehr Links zu Community-Mitgliedern als nach außen



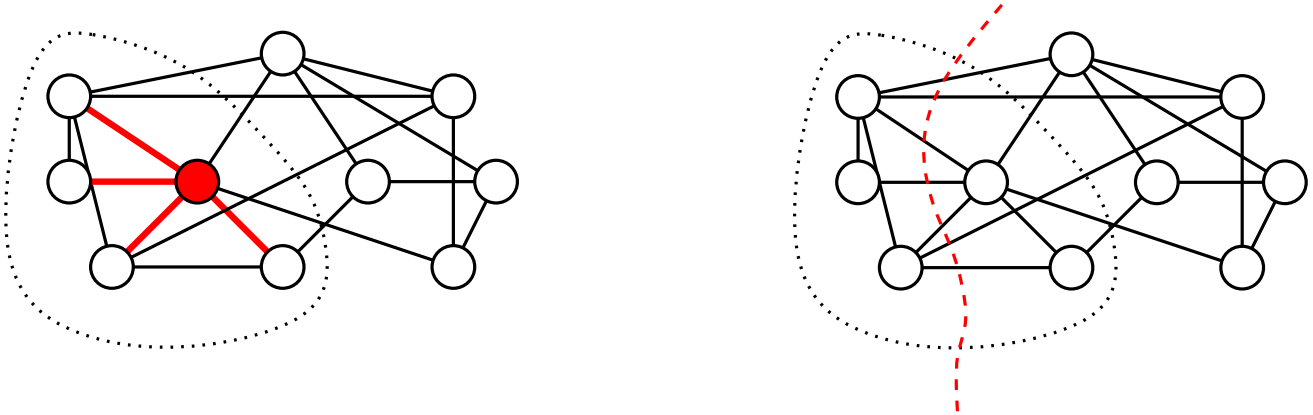
Thema 1: Motivation - Anwendungsbeispiel

- Suche nach Communities mit garantierten Eigenschaften
  - Hier: Pro Konten mehr Links zu Community-Mitgliedern als nach außen
  - Und: Zerschneiden der Community hat ”Mindestpreis”



Thema 1: Motivation - Anwendungsbeispiel

- Suche nach Communities mit garantierten Eigenschaften
  - Hier: Pro Konten mehr Links zu Community-Mitgliedern als nach außen
  - Und: Zerschneiden der Community hat ”Mindestpreis”

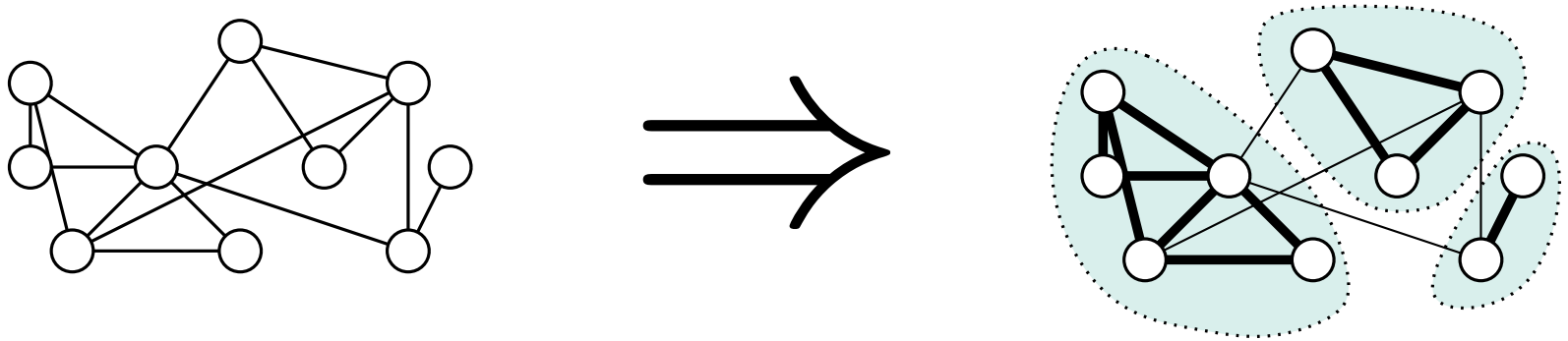


Thema

- Graph-Clustern mit Hilfe von Schnitt-Bäumen
  - Erweitere Graphen um Hilfsknoten und Hilfskanten
  - Berechne minimale *s*-*t*-Schnitte

Thema 2/3: Motivation - Modularity

- Zerlegung in dichte Teilgraphen (Cluster), die nur schwach untereinander verbunden sind
  - Keine formale Problemstellung
  - Wie misst man Qualität einer gegebenen Zerlegung?
  - Wie findet man gute Zerlegung/Clusterung?

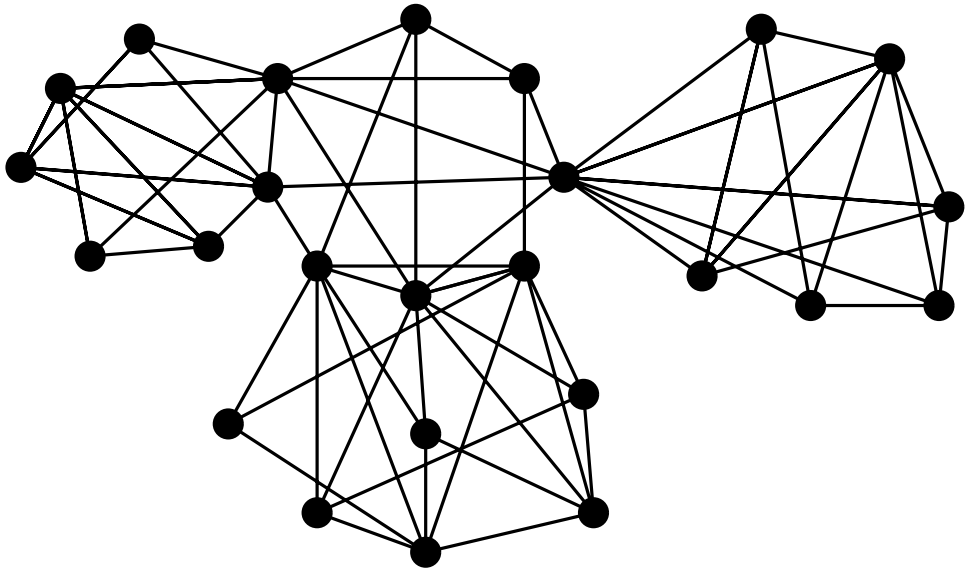


- Ein mögliches Qualitätsmaß: Modularity
  - Wird auch zum Design von Algorithmen verwendet

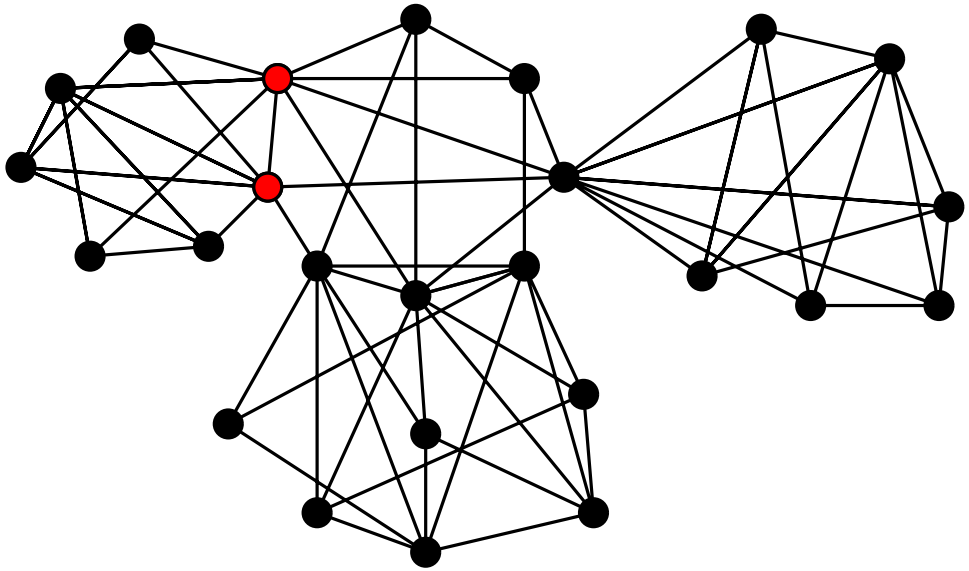
Themen

- Modularity I (Thema 2)
  - Einführung des Maßes Modularity
  - Komplexität der Optimierung von Modularity
- Modularity II (Thema 3)
  - Greedy-Algorithmen zur Berechnung guter Modularity-Clusterungen
  - Theoretische und praktische Eigenschaften
- Clustern oder nicht Clustern (Thema X)
  - Wenn keine gute Clusterung gefunden wird, was heißt das?
  - Wann besitzt ein Graph überhaupt eine sinnvolle Clusterung?

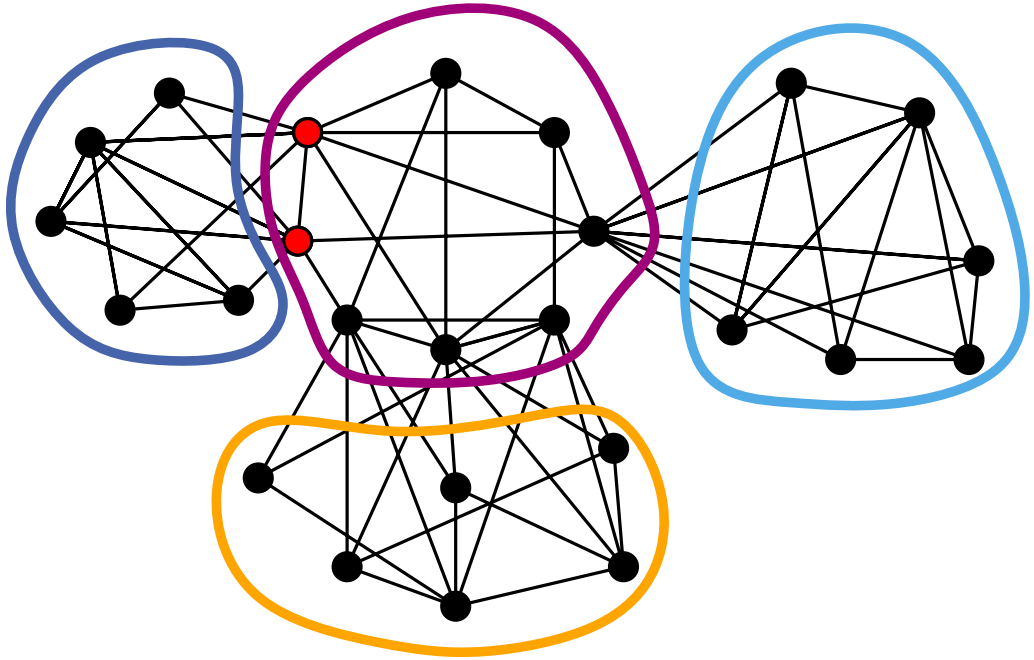
Thema 4/5 : Motivation - Überlappende Clusterungen



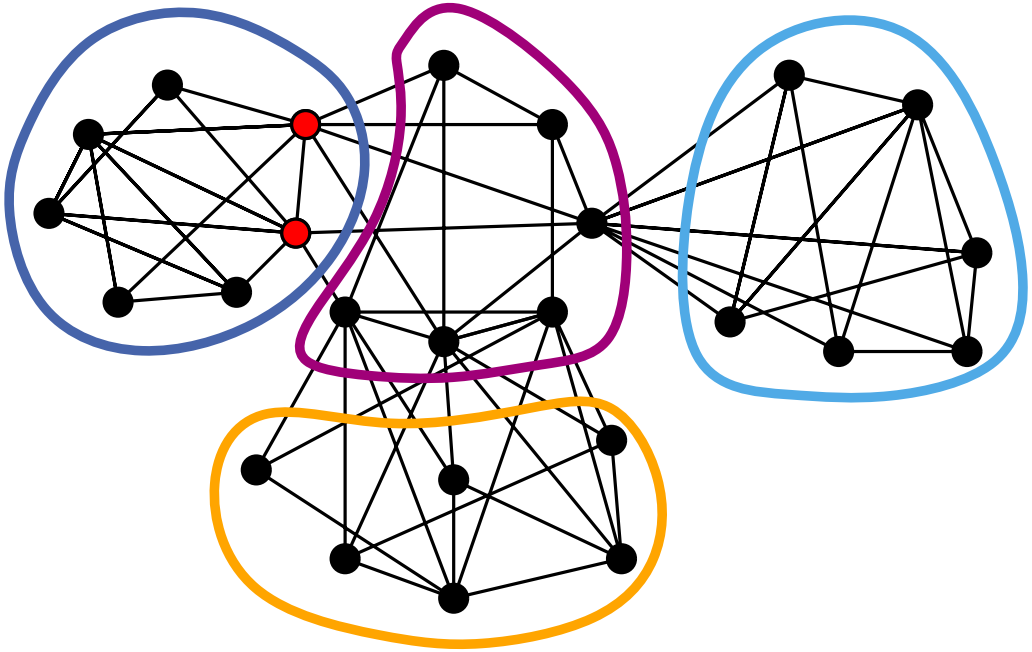
Thema 4/5 : Motivation - Überlappende Clusterungen



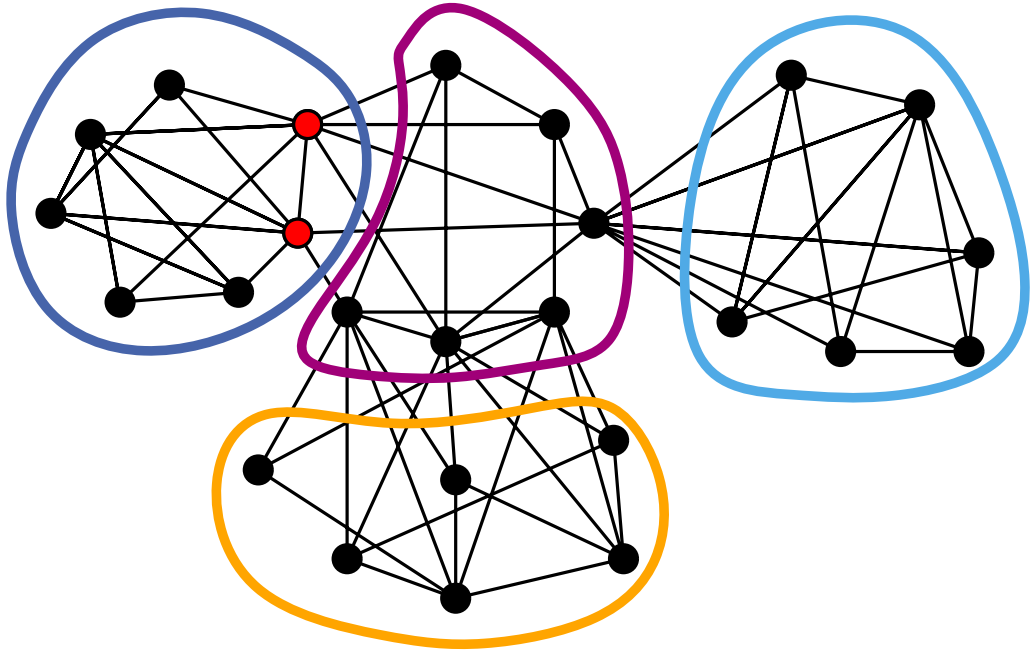
Thema 4/5 : Motivation - Überlappende Clusterungen



Thema 4/5 : Motivation - Überlappende Clusterungen

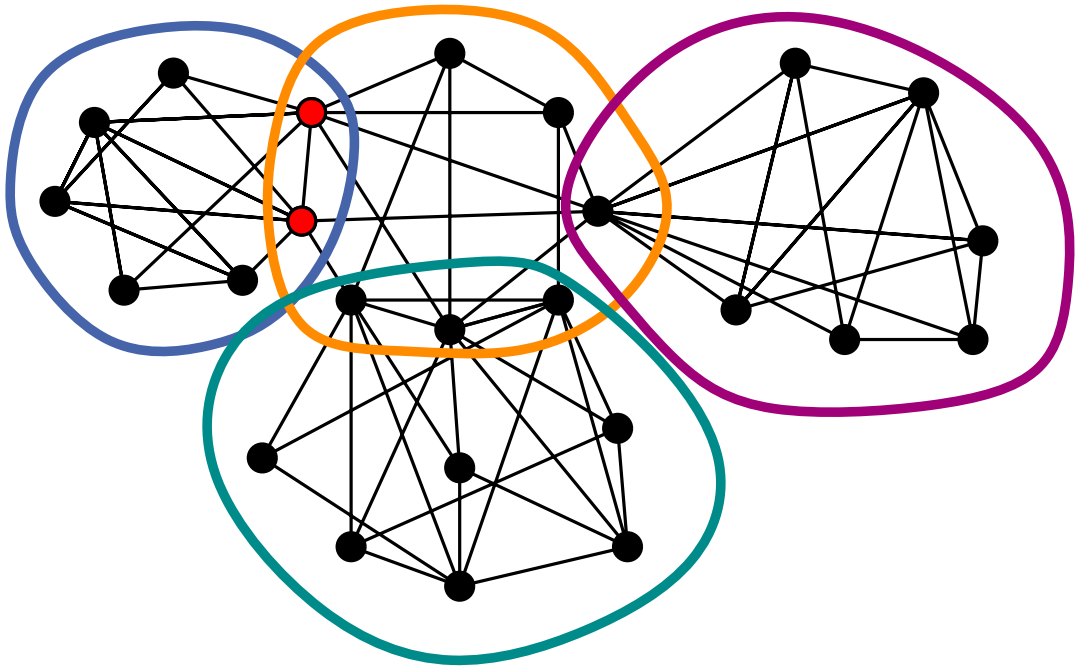


Thema 4/5 : Motivation - Überlappende Clusterungen



Welche Einteilung ist besser?

Thema 4/5 : Motivation - Überlappende Clusterungen



Beispiel soziale Netze: Aktoren gehören zu mehreren Gruppen:

Schulfreunde, Kommilitonen, Sportverein, ...

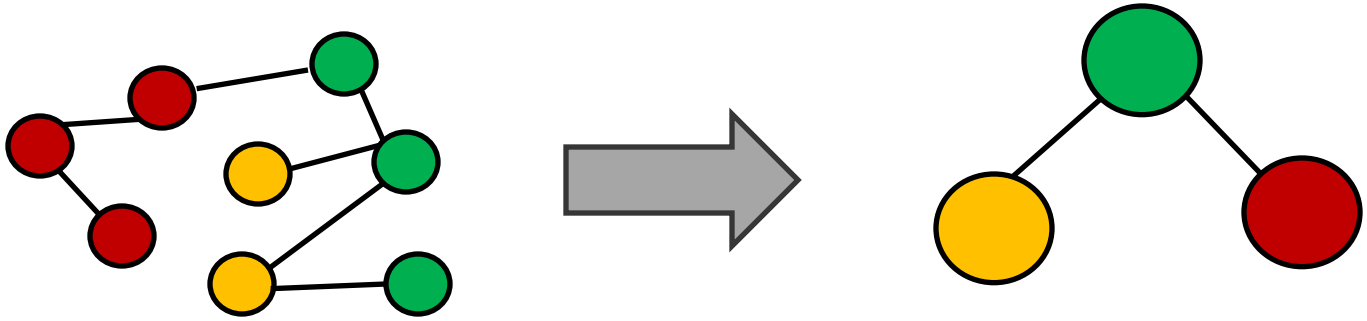
Themen

- Clique Percolation (Thema 4)
  - Adjazente k-Cliquen: k-Cliquen, die sich nur in einem Knoten unterscheiden
  - Cluster = Menge von Knoten, die in einer Folge von paarweise adjazenten Cliquen enthalten ist

- Strongly Knit Clusters (Thema 5)
  - Seminarthema (Master)!
  - Jeder Knoten innerhalb eines Clusters sollte mit mindestens β Prozent der Knoten dort verbunden sein
  - Jeder Knoten außerhalb eines Clusters sollte mit mindestens α Prozent der Knoten dort verbunden sein

Thema 6/7: Motivation - Anwendungsbeispiel

- Große Graphen sind schwer zu verarbeiten
  - Idee: Aggregation
  - Hier: Anhand der Farben



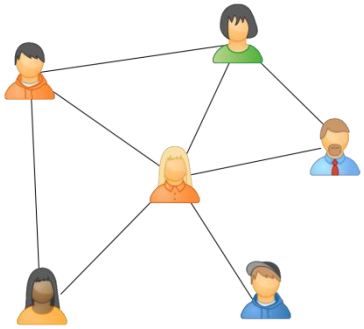
- Was wenn keine Attribute (Farben) vorhanden?
  - Lösung: Graph-Clustering /-Partitionierung

Themen

- METIS (Thema 6)
  - 1. *Coarsening Phase*: Schrittweises zusammenfassen des Graphen
  - 2. Initiale Partitionierung
  - 3. *Uncoarsening Pahse*: Rekursive Partitionierung des Gaphen
- SCAN (Thema 7)
  - *Structural Similarity*: Fasse Knoten zusammen, die viele gemeinsame Nachbarn haben
  - Hubs und Outlier werden erkannt

Thema 8/9: Motivation – Attributierte Graphen

- Kombination aus Graphen + Relationalen Daten
- Analyse der Struktur des Graphen zusammen mit den Attributen
- Cluster/Outlier Erkennung:
  Bezüglich der Struktur aber auch bezüglich der Attribute



Age	Profession	Gender	Publications
21	Student	Female	1
23	Student	Male	0
30	Lecturer	Female	25
29	Student	Male	2
27	Student	Female	3
25	Student	Male	0

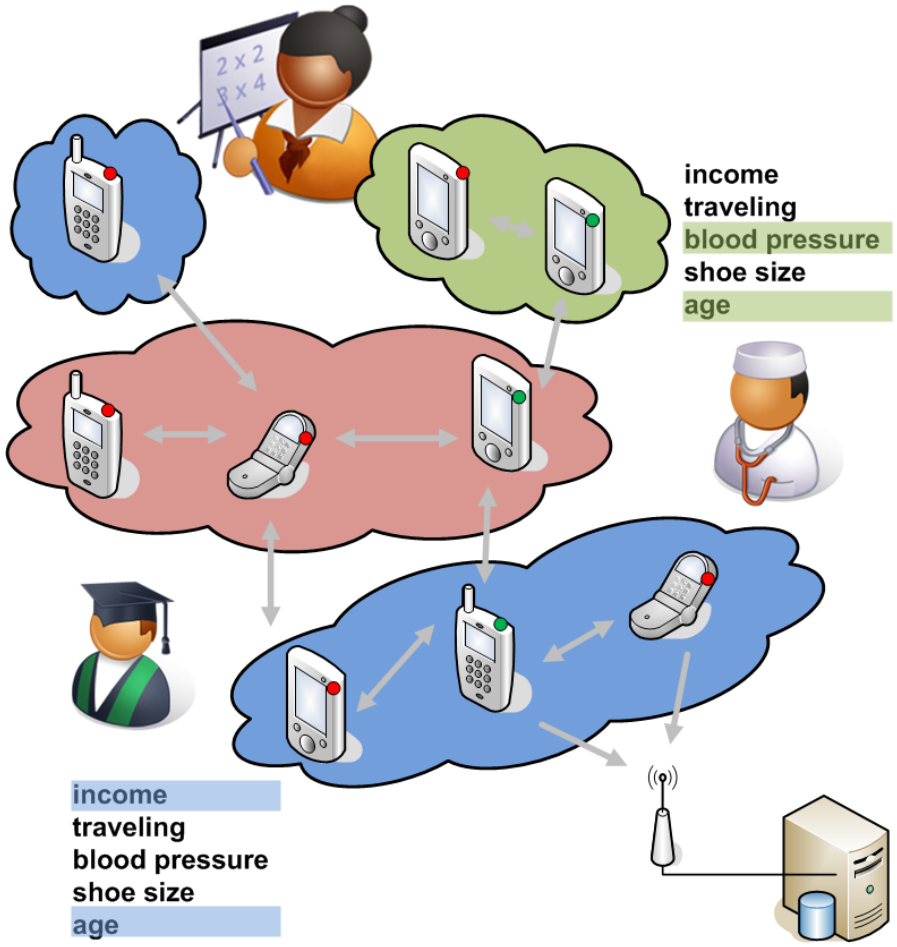
Themen

- Outlier Erkennung mit Zentralitätsmaßen (Thema 8)
  - Outlier im Bezug nur auf die Struktur
  - Outlier: **Einbruch im Netzwerkverkehr**
  - **Empirische Evaluation** on Europäische Internetdienstanbieter

- Cluster und Outlier Erkennung in attributierten Graphen (Thema 9)
  - **Attributen und Graph Struktur** betrachtet
  - Kein Parameter
  - starke verbundene Knoten, die mit Attributen korrelieren (Bsp: Gemeinsame Youtube Gruppen)
  - Outlier als **Nebenprodukt** des Clustering

Thema 10/11: Motivation – Subspace Selektion

- Nicht alle Eigenschaften sind für das Data Mining relevant
- Pro Cluster / Community können unterschiedliche Attribute notwendig sein
- Ziele:
  - Erkennung der Cluster
  - Auswahl der Projektionen
  - Evaluation der Korrelation



Themen

- Feature Selektion in attributierten Graphen (Thema 10)
  - Angenommen man hat Cluster / Communities in einem Graphen
  - Welches sind dann die relevanten Attribute?
  - Beschränkung auf die Auswahl der Dimensionen (Features)

- Subspace clustering in attributierten Graphen (Thema 11)
  - Integration von Cluster und Subspace Erkennung
  - Dichtebasiertes Clustering
  - Evaluation der Korrelation von Graph-Cluster und Attribut-Cluster
  - Wie stark ist die Graph/Attribut Struktur am Ergebnis beteiligt?

Übersicht der Themen

(Pro-)Seminar

1. Graph-Clustern mit Hilfe von Schnitt-Bäumen (TH)
2. Modularity I (Komplexität der Optimierung von Modularity) (TH)
3. Modularity II (Greedy-Algorithmen) (AK)
4. Clique Percolation (AK)
5. Strongly Knit Clusters (AK) (Master!)
6. Graph Partitionierung mit METIS (CO)
7. Cluster, Hub, und Outlier Erkennung mit SCAN (CO)
8. Outlier Erkennung mit Zentralitätsmaßen (PI)
9. Cluster und Outlier Erkennung in attributierten Graphen (PI)
10. Feature Selection in attributierten Graphen (EM)
11. Subspace clustering in attributierten Graphen (EM)

Betreuer

Tanja Hartmann (TH), Patricia Iglesias Sánchez (PI), Andrea Kappes (AK), Emmanuel Müller (EM) und Christopher Oßner (CO)

THEMENVERGABE

Themenvergabe

Thema	Bearbeiter
Graph-Clustern mit Hilfe von Schnitt-Bäumen	
Modularity I (Komplexität der Optimierung von Modularity)	
Modularity II (Greedy-Algorithmen)	
Clique Percolation	
Strongly Knit Clusters	(Master Thema!)
Graph Partitionierung mit METIS	
Cluster, Hub, und Outlier Erkennung mit SCAN	
Outlier Erkennung mit Zentralitätsmaßen	
Cluster und Outlier Erkennung in attributierten Graphen	
Feature Selection in attributierten Graphen	
Subspace clustering in attributierten Graphen	

ORGANISATORISCHES

Anforderungen für den Schein

- Erstellung einer
  - Gliederung und Literaturübersicht
  - Präsentation: 20 Minuten (Proseminar) bzw. 30 Minuten (Seminar) + 15 Minuten Diskussion
  - Ausarbeitung: 8-10 Seiten (Proseminar) bzw. 12-15 Seiten (Seminar)

- Teilnahme an allen Vortragsterminen
  - Das Seminar ist eine Prüfungsleistung (50% Ausarbeitung, 50% Vortrag)
  - Gleichberechtigt mit anderen Prüfungsleistungen
  - Unentschuldigtes Fehlen bedeutet Ausschluss aus dem Seminar

- **Achtung bei Scheinen und Anrechnung in Module**
  - Bitte informiert euch vorher, falls keine „Standardkombination“
  - Proseminar und Seminarbezeichnungen bleiben (auch wenn ECTS von Proseminar für manche ausreichend wären)

Termine

- Abgabe Gliederung und Literaturübersicht: 26.11.2012
- Blocktermine mit 4 Vorträgen pro Termin
  - werden noch bekanntgegeben (doodle Abstimmung)
  - zwischen 07.01.2012 und 08.02.2013
- Bis spätestens 14.12.2012:
  - Abgabe Folien
  - (optional) Probevortrag mit Betreuer
- Erste Version der Ausarbeitung: 07.01.2013
- Abgabe Ausarbeitung: 28.02.2013
- Abgabefristen müssen eingehalten werden!

TECHNISCHES & LITERATURRECHERCHE

Technisches

- Ausarbeitung in Word / Writer oder LaTeX
  - LNCS Stylesheet
  - Vorlage wird auf Web-Seite zur Verfügung gestellt
- Präsentation mit Powerpoint, Impress oder PDF
- Abgabe der Materialien per E-Mail als PDF

Literaturrecherche

- Portal.ACM.org

- IEEE Explore

- citeseer.ist.psu.edu

- Hinweis: Zugang frei, wenn Ihr wie folgt vorgeht:
  - http://www.ubka.uni-karlsruhe.de/
  - „Digitale Biliothek“ anklicken
  - „Elektronische Zeitschriften“ auswählen

Literaturrecherche (Suchen)

Suche Erweiterte Scholar-Suche Scholar-Einstellungen

[BUCH] Biostatistical Analysis - Gruppe von 3 »
JH Zar - 2007 - Prentice-Hall, Inc. Upper Saddle River, NJ, USA
... The **ACM Portal** is published by the Association for Computing Machinery. Copyright © 2007 **ACM**, Inc. Terms of Usage Privacy Policy ...
Zitiert durch: 25214 - Ähnliche Artikel - Websuche - Bibliothekssuche

Richtlinien für den Vortrag

Struktur

- Einleitung:
  - Motivation
  - grundlegende Begriffe, Definitionen
  - Plan, Vorgehensweise aufzeigen
- Hauptteil:
  - Top-Down Ansatz
  - Schlüsselideen
  - Beweisskizze
  - Technische Details
  - Beispiele
- Schluss
  - Zusammenfassung
  - Offene Fragen

Wie erreiche ich mein Publikum?

- Wichtig: Ihr seid die Experten!
- Erfolg = Zuhörer verstehen das Wesentliche
- Wiederholungen verwenden
- Bezug herstellen:
  „Wie wir aus den Theoretischen Grundlagen wissen. . . “
- Erinnern, nicht einfach voraussetzen
- Kontakt durch klare Signale herstellen:
  - „Fragen?“
  - „Danke!“
- Darauf achten, nicht zu schnell zu sprechen!

Gestaltung der Folien

- Soweit möglich, Beispiele und Bilder benutzen!
- Wenig Text
- Ganze Sätze nur in Ausnahmefällen (ausgenommen: Definitionen,Sätze, . . .)!
- Höchstens 10 Zeilen pro Folie
- Zeit: 1,5 bis 2 Minuten pro Folien
- Klare Motivation für den Einsatz von Farben

VIEL ERFOLG BEIM LITERATURSTUDIUM