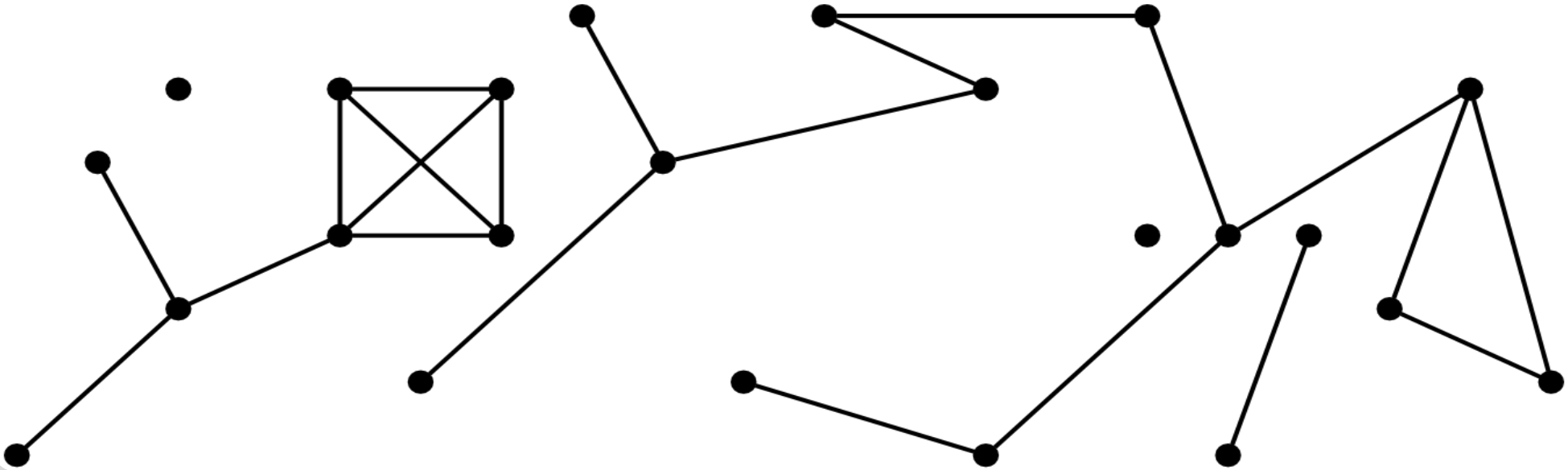
Karlsruher Institut für Technologie

Knotenklassifikation in dynamischen Graphen mit Texten

Martin Thoma | 25. Februar 2014

INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



Zitationsdatenbanken

- **Publikationen** oder **Autoren** können **Knoten** sein,
- **Zitate** oder **Mehrautorenschaft**können **Kanten** sein und
- **Kategorien** können **Beschriftungen** sein

Problem: Nicht alle Knoten sind beschriftet
Anwendungsideen:

- Kategorievorschläge bei neuen Einträgen
- Korrekturvorschläge für alte Einträge

Herausforderungen

- Große Graphen,
- Dynamische Graphen,
- Texte sollen verwendet werden

Name	Knoten	davon beschriftet	Kanten	Beschriftungen
CORA	19 396	14 814	75 021	5
DBLP	806 635	18 999	4 414 135	5

DYCOS ist

- effizient,
- einfach,
- und nutzt Struktur und Texte

Herausforderungen

- Große Graphen,
- Dynamische Graphen,
- Texte sollen verwendet werden

Name	Knoten	davon beschriftet	Kanten	Beschriftungen
CORA	19 396	14 814	75 021	5
DBLP	806 635	18 999	4 414 135	5

DYCOS ist

- effizient,
- einfach,
- und nutzt Struktur und Texte

Zitationsdatenbanken

The Development of the C Language Interprocess Communication in the Ninth Edition Unix System

Computer Science

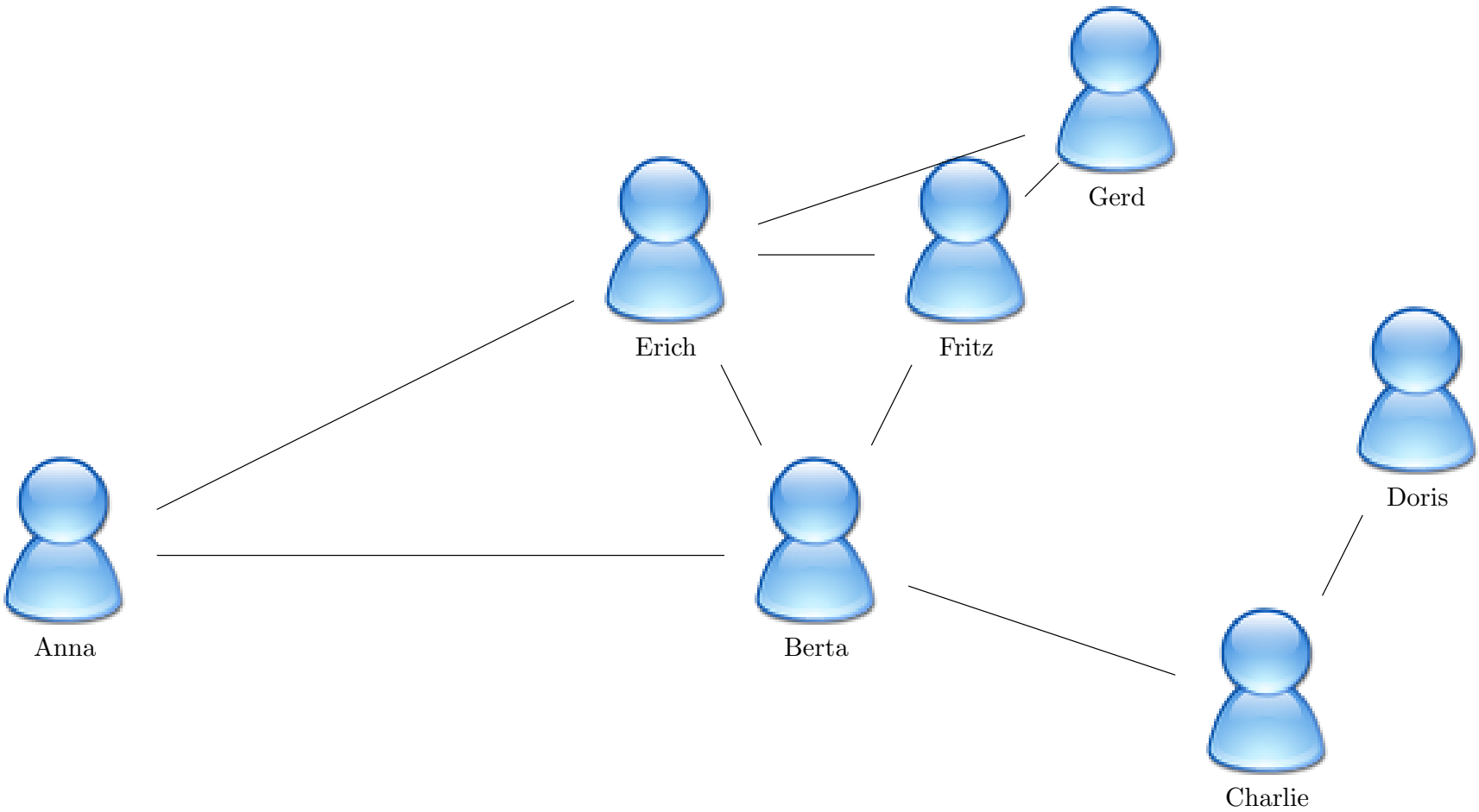
The C Programming Language digital restoration and typesetter

Computer Science

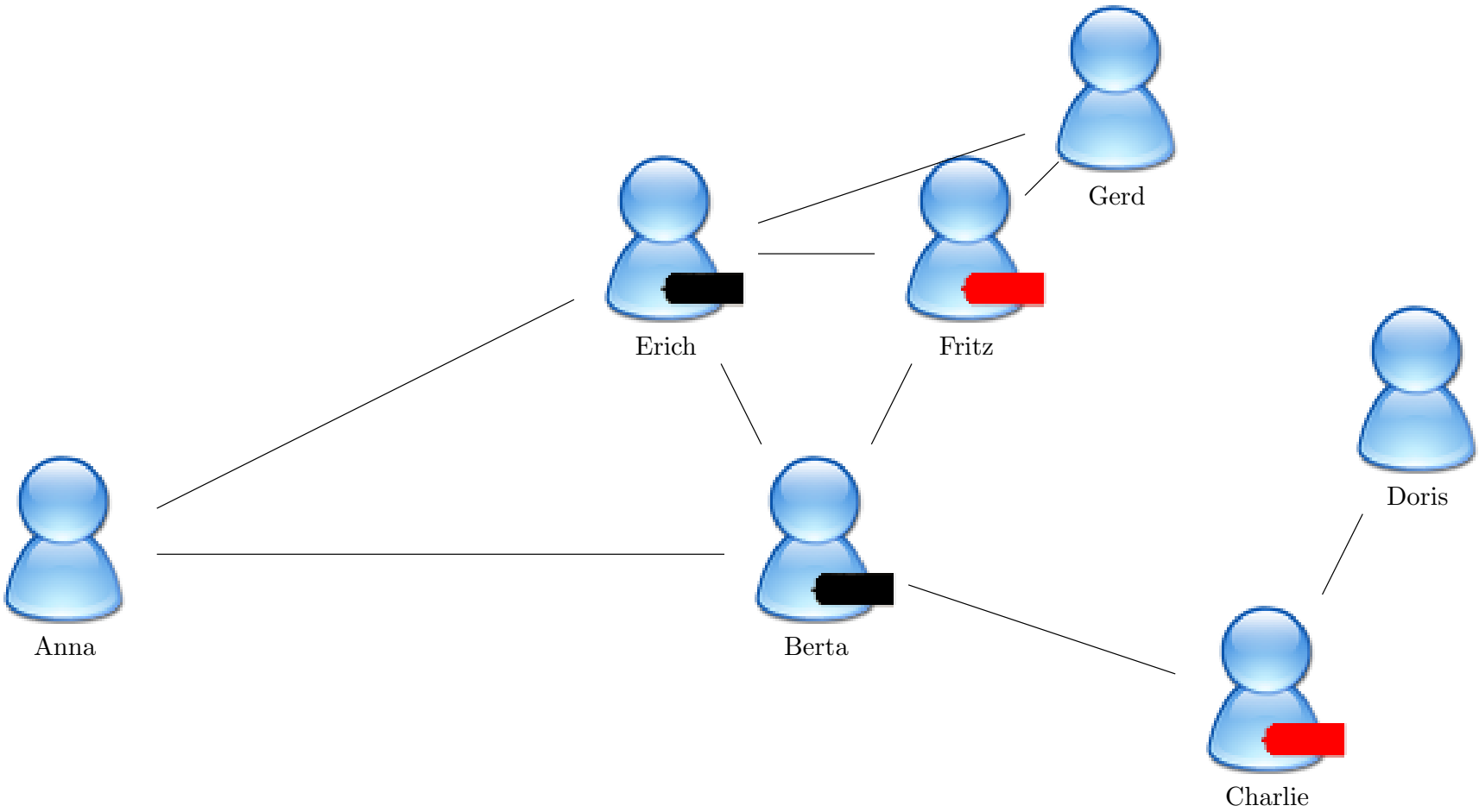
The Identity Thesis for Language and Music

Linguistics

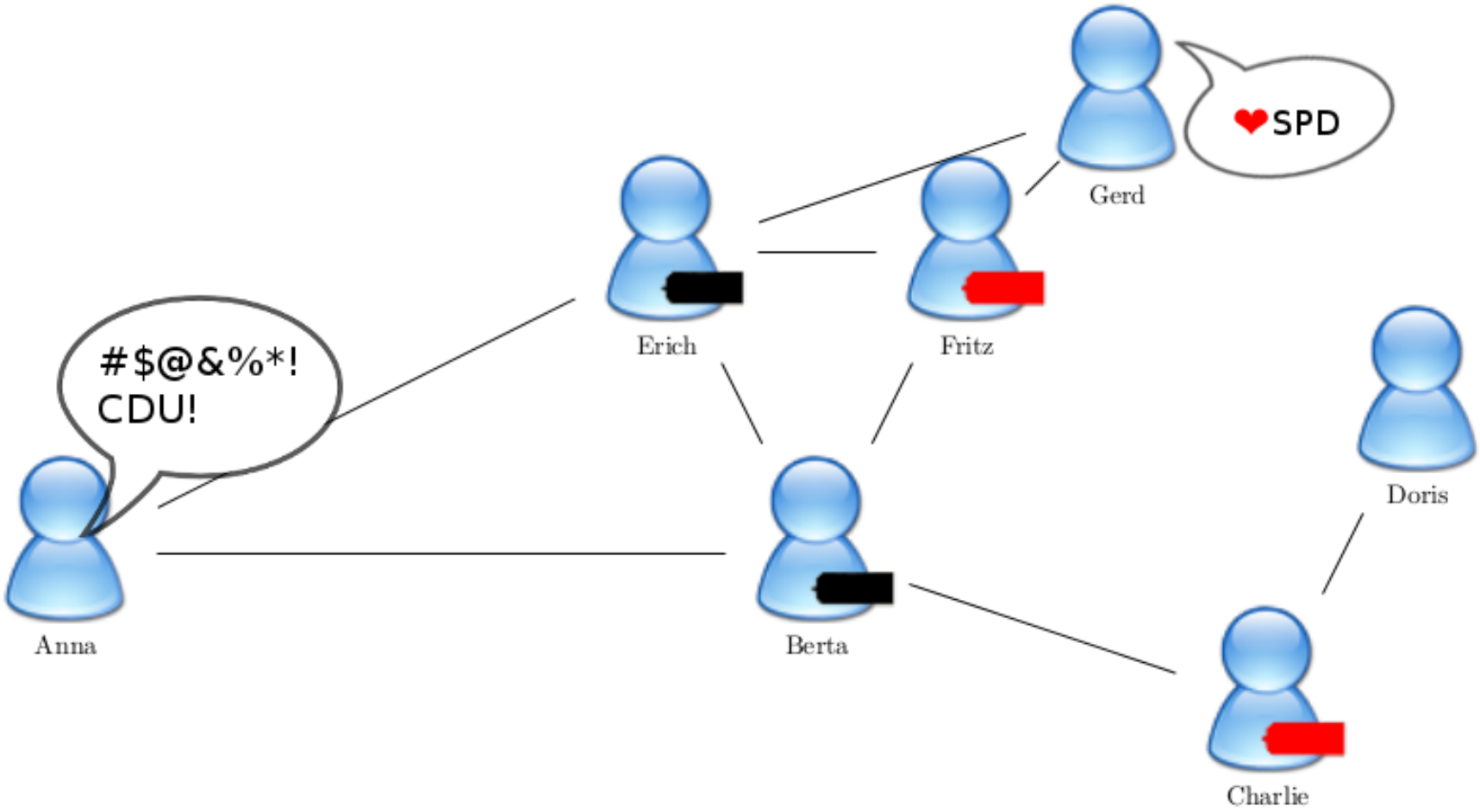
Social Network



Partially labeled network



Partially labeled network with content



Überblick

- Graph ist gegeben
- Knoten sind teilweise beschriftet
- Fehlende Beschriftungen sollen berechnet werden

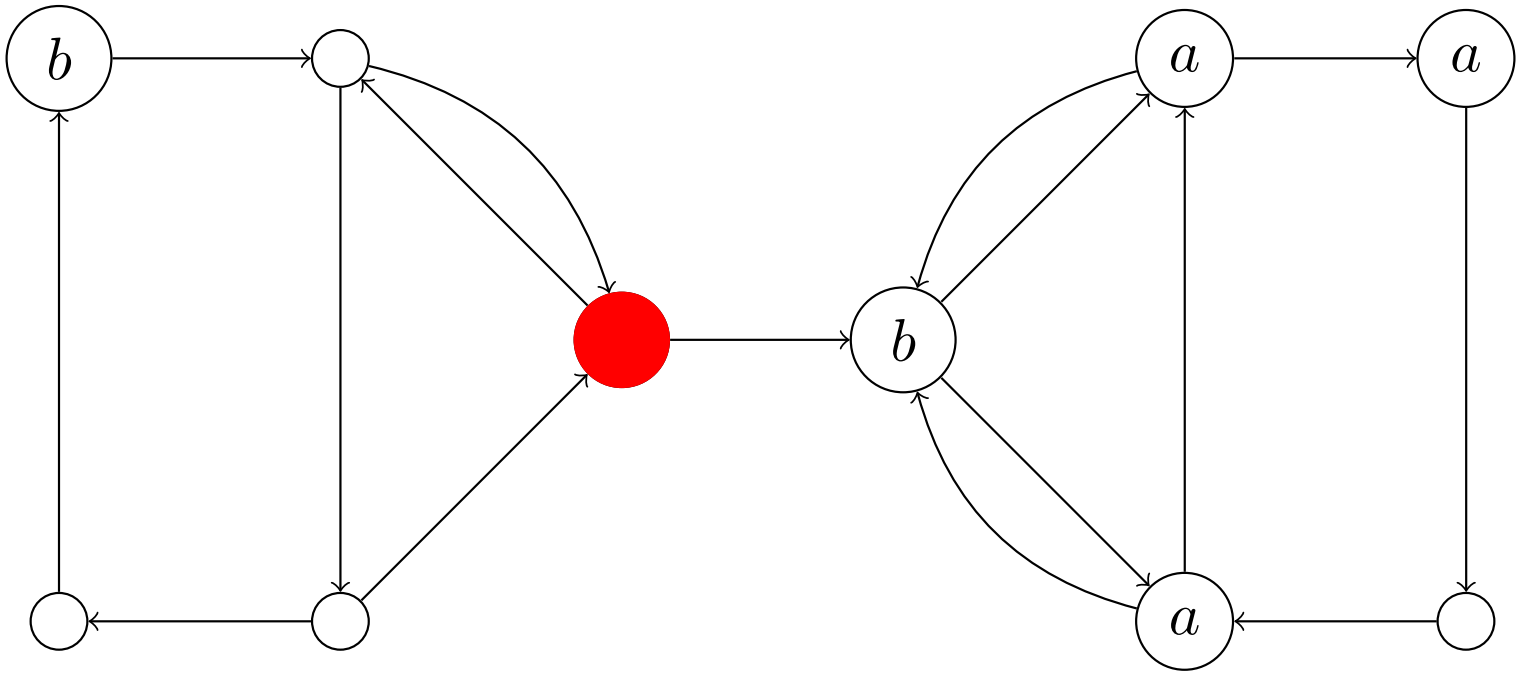
Idee: Homophilie nutzen
Nahe Knoten sind ähnlich
$\Rightarrow$ Random Walks zur Klassifizierung nutzen

Überblick

- Graph ist gegeben
- Knoten sind teilweise beschriftet
- Fehlende Beschriftungen sollen berechnet werden

Idee: Homophilie nutzen
Nahe Knoten sind ähnlich
$\Rightarrow$ Random Walks zur Klassifizierung nutzen

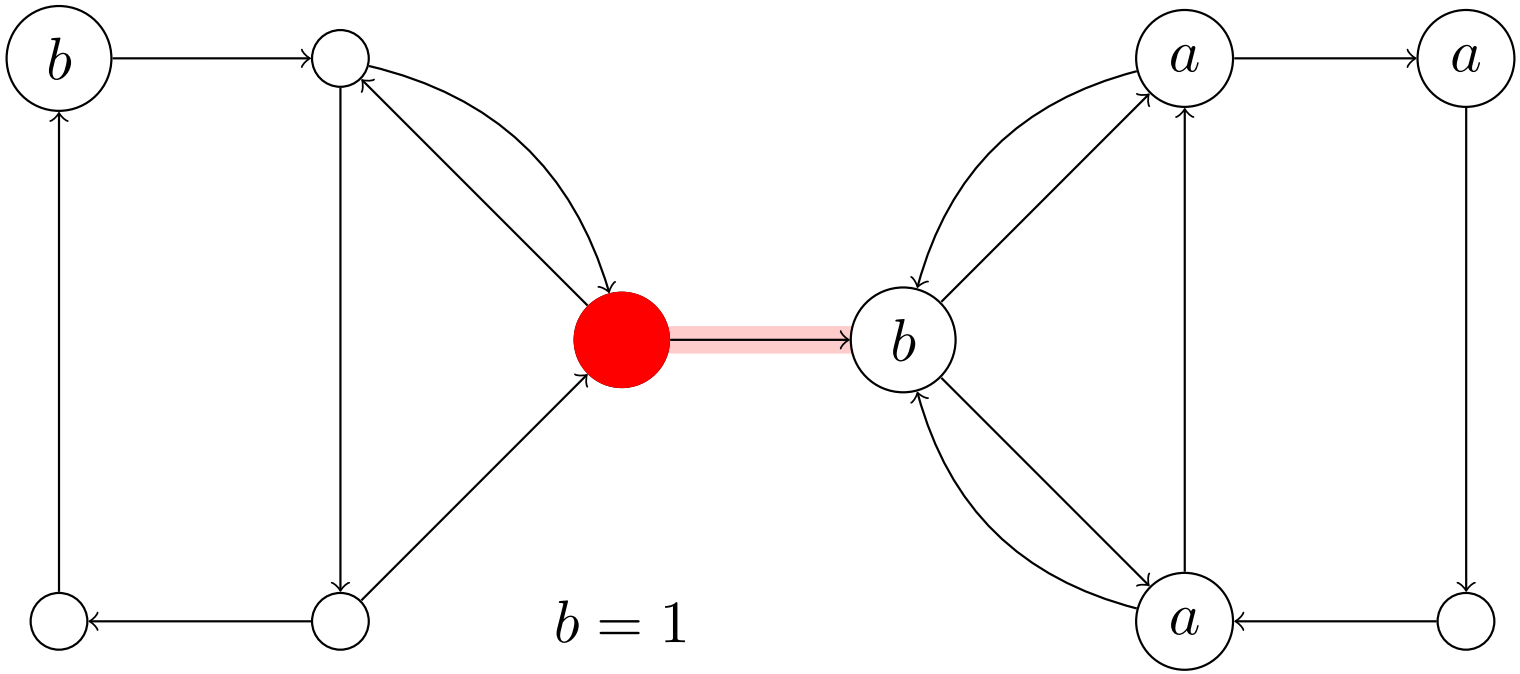
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

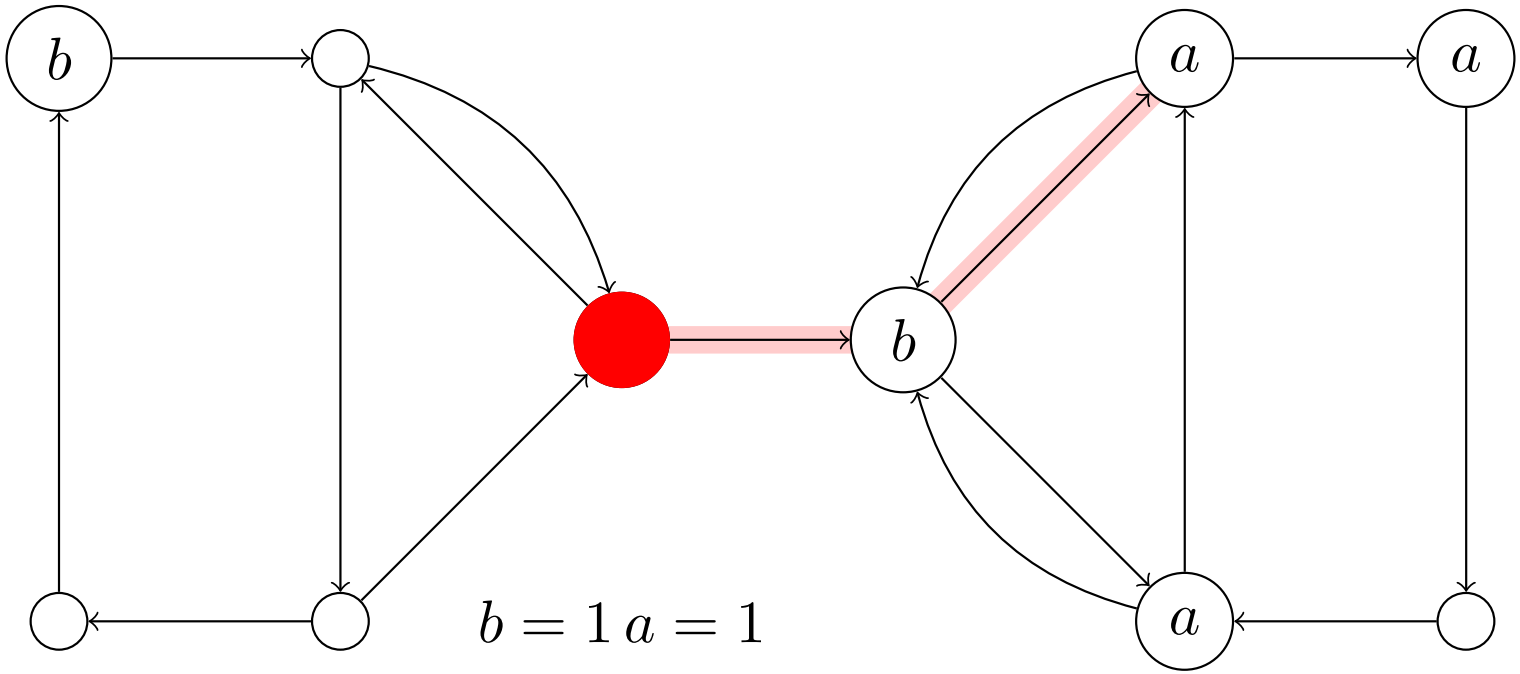
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

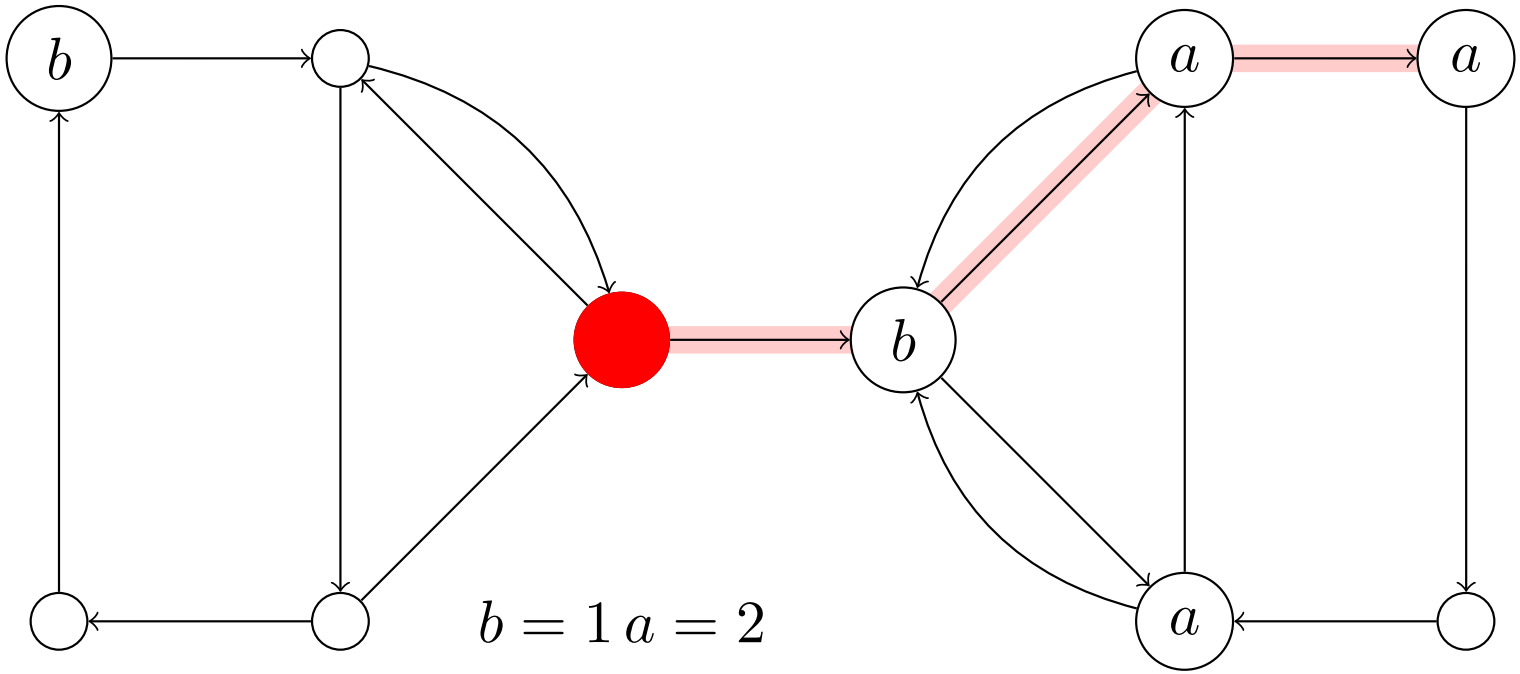
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

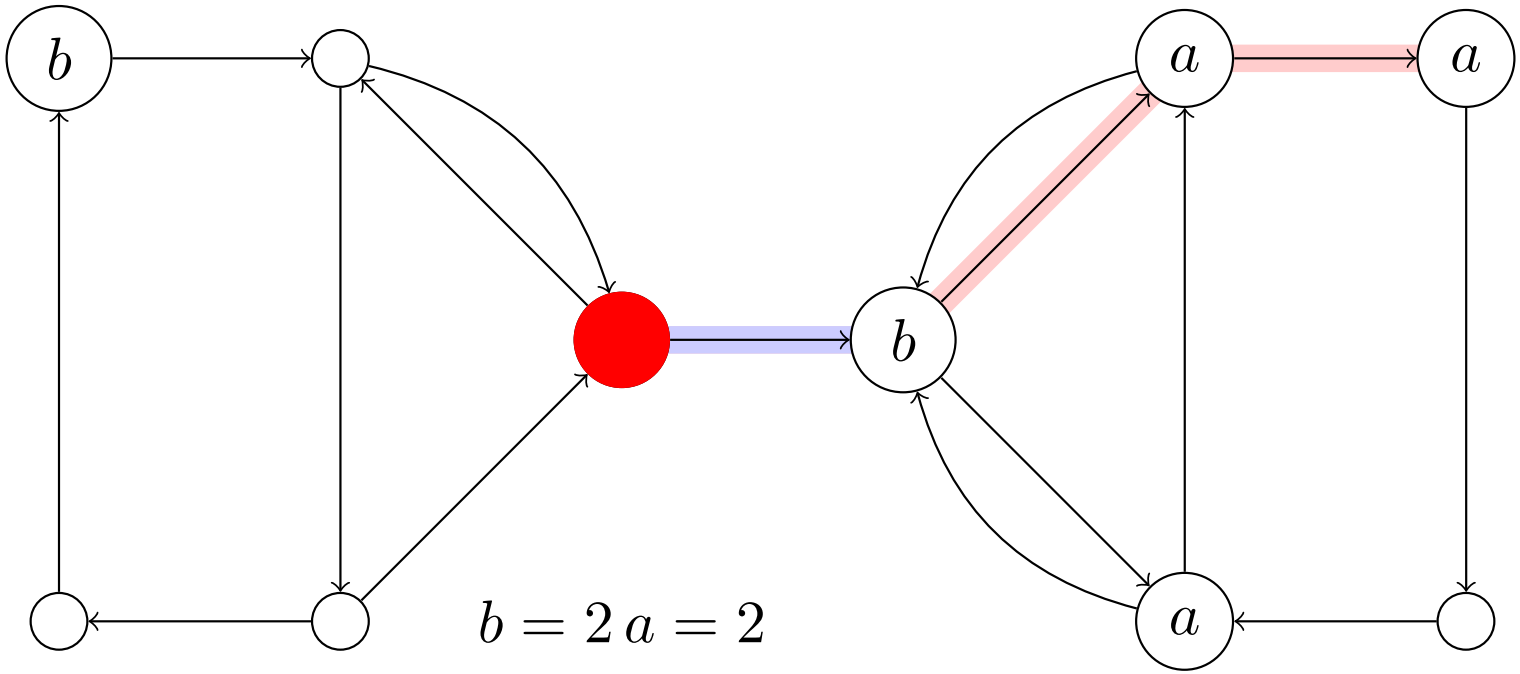
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

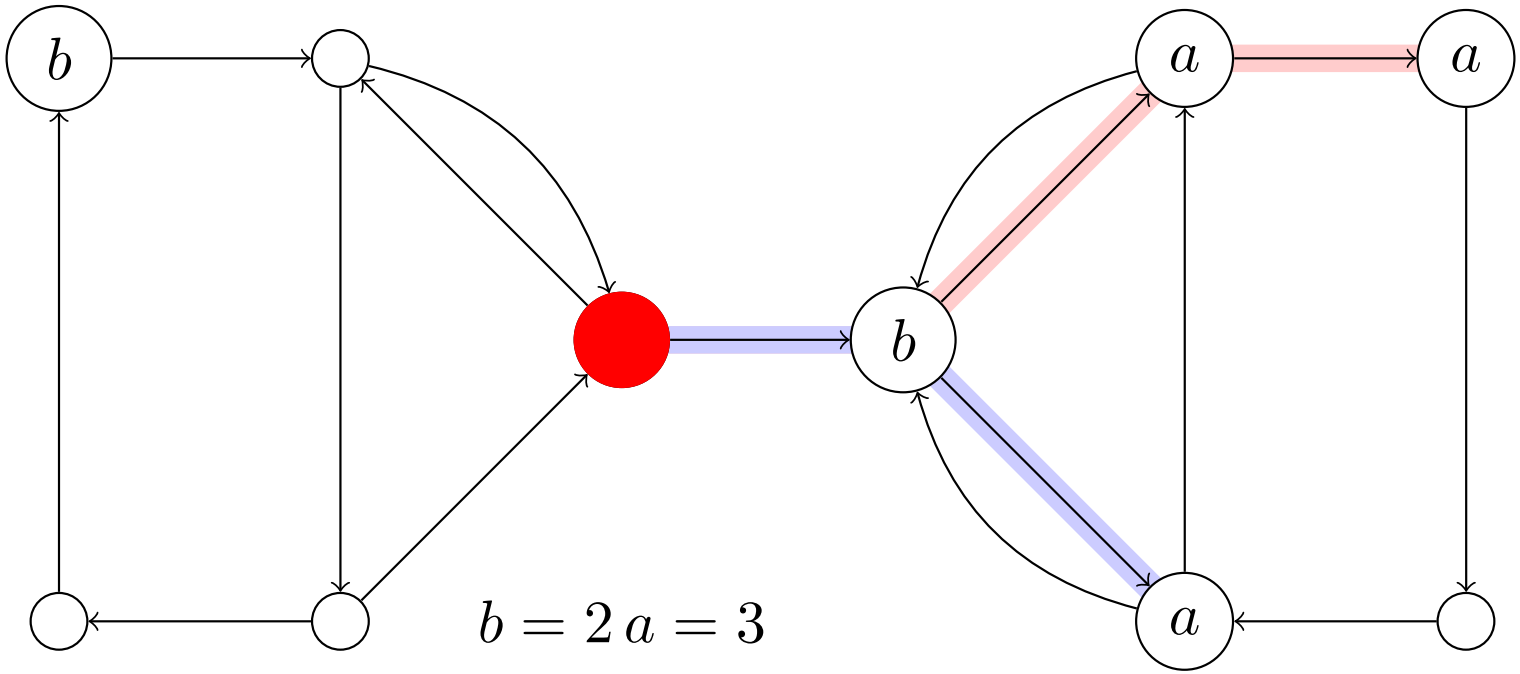
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

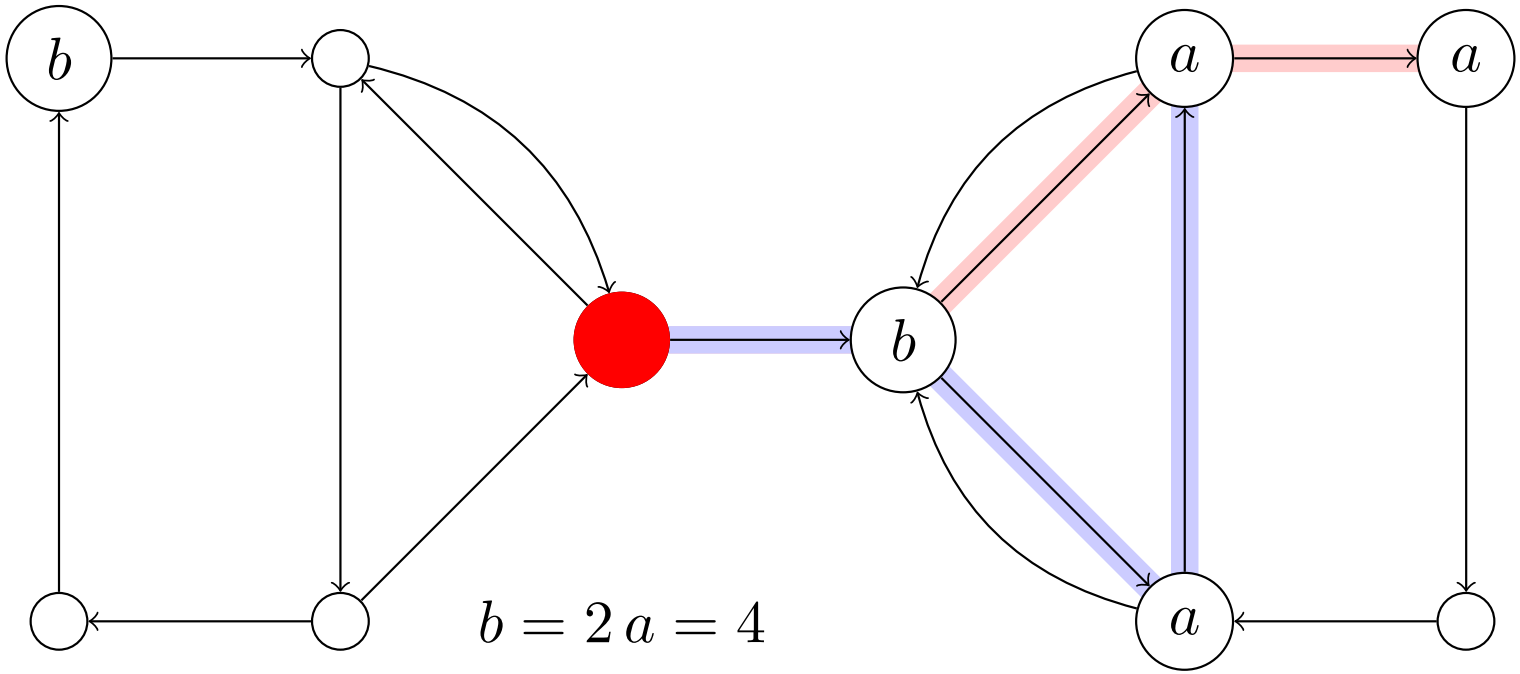
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

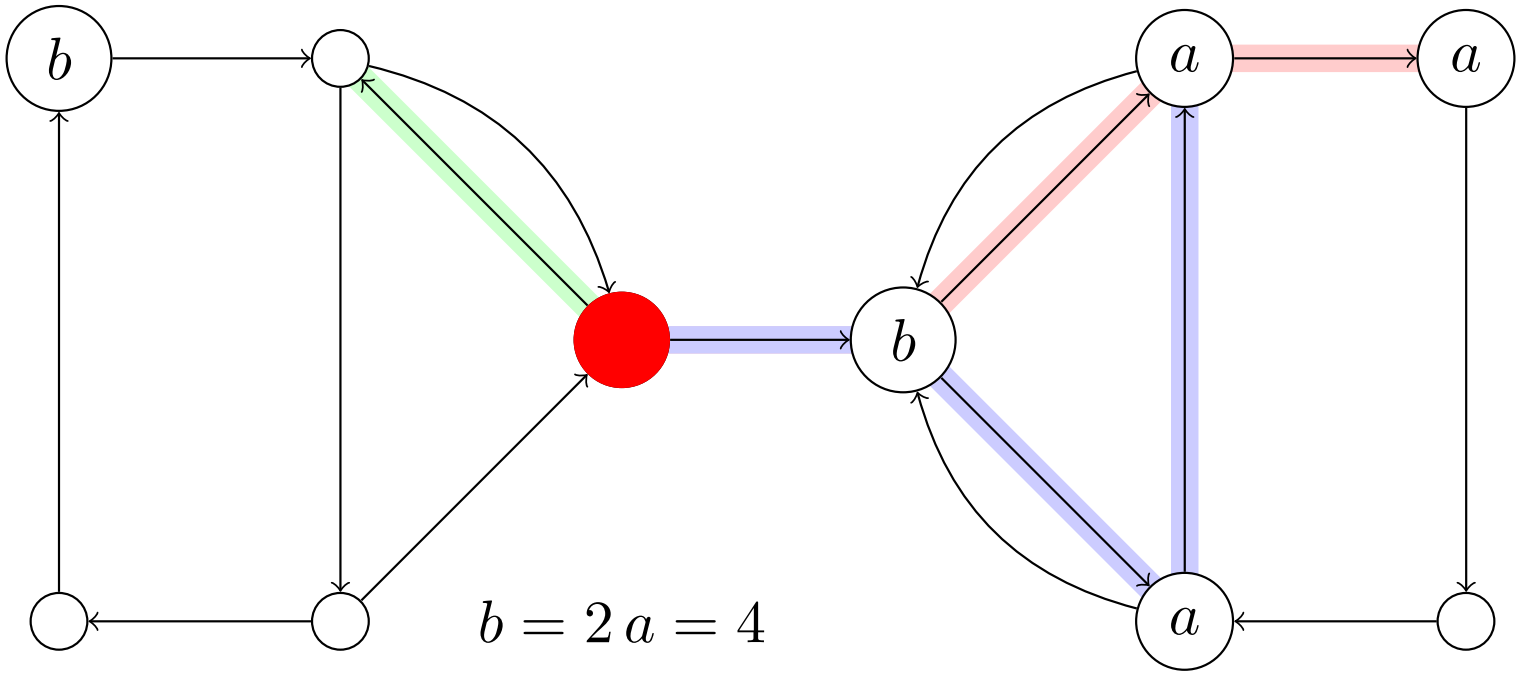
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

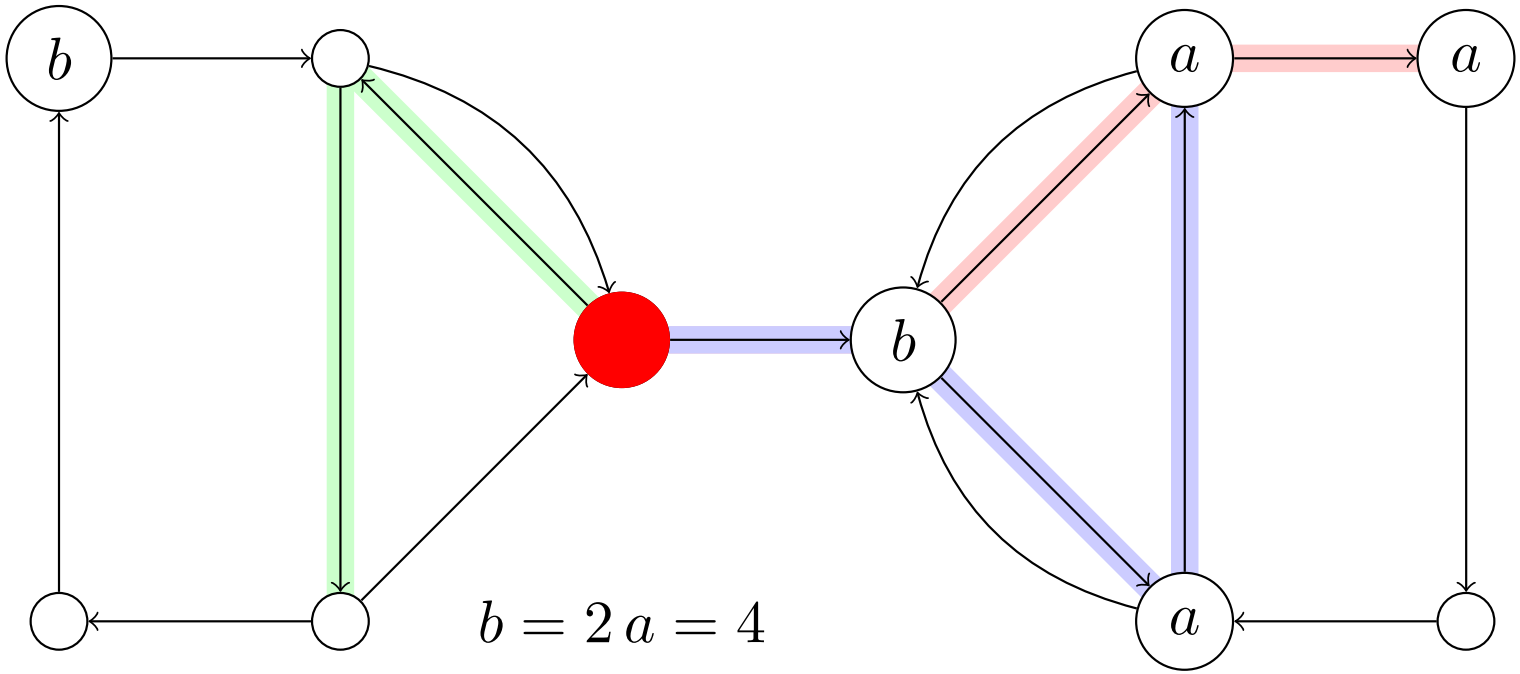
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

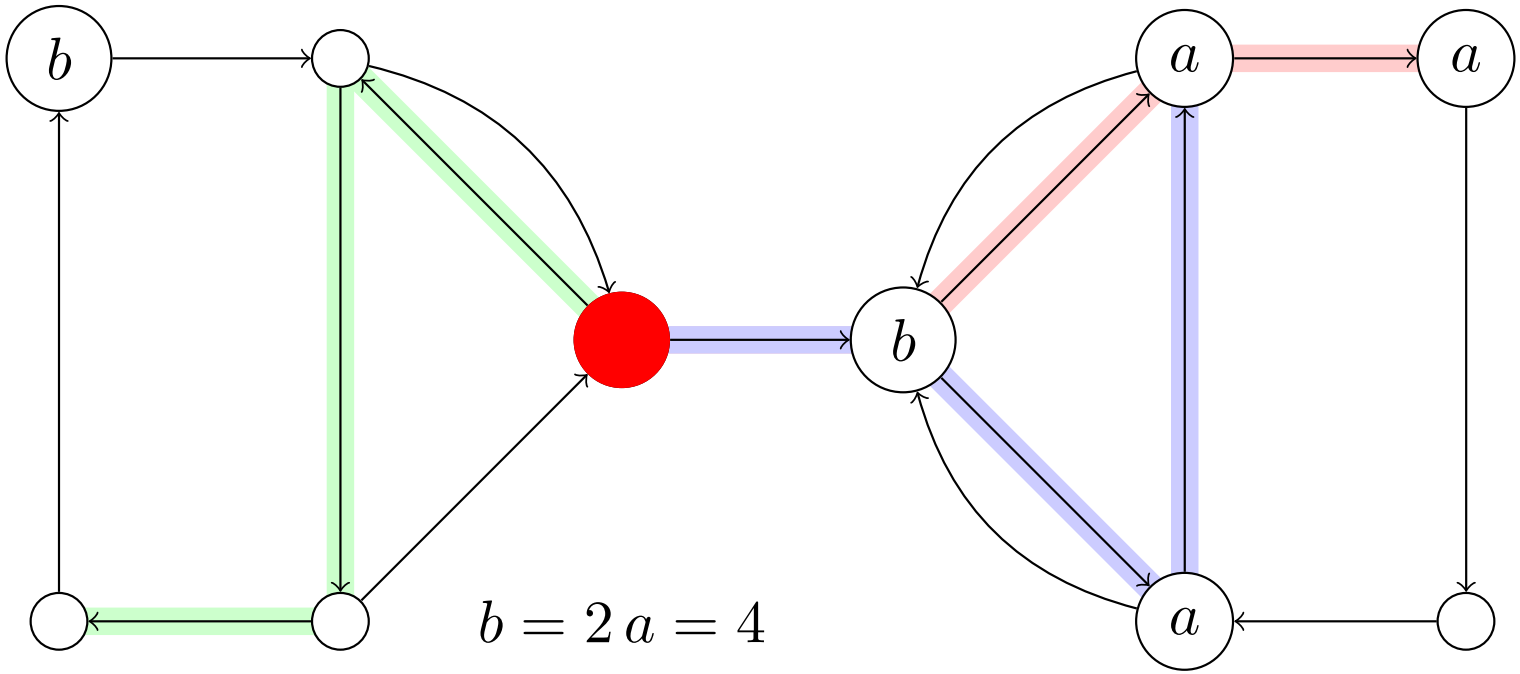
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

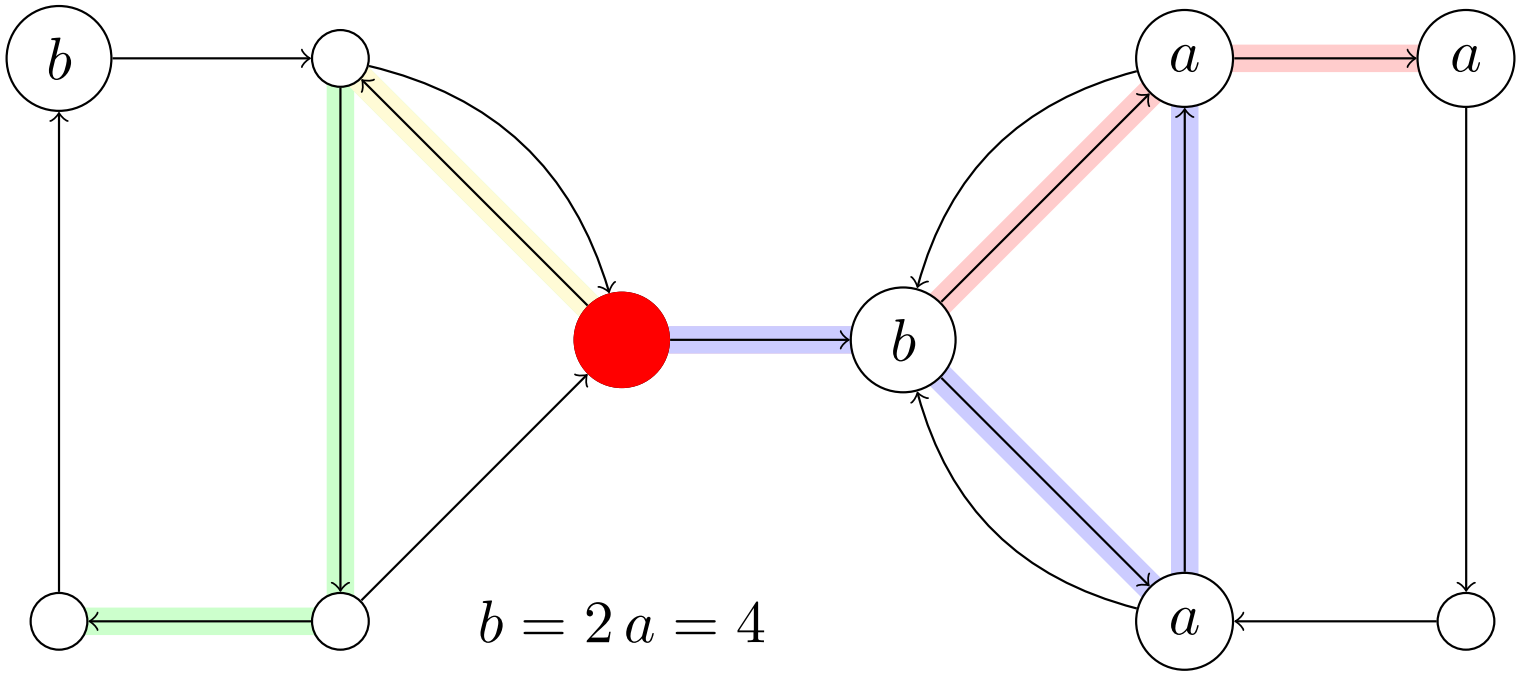
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

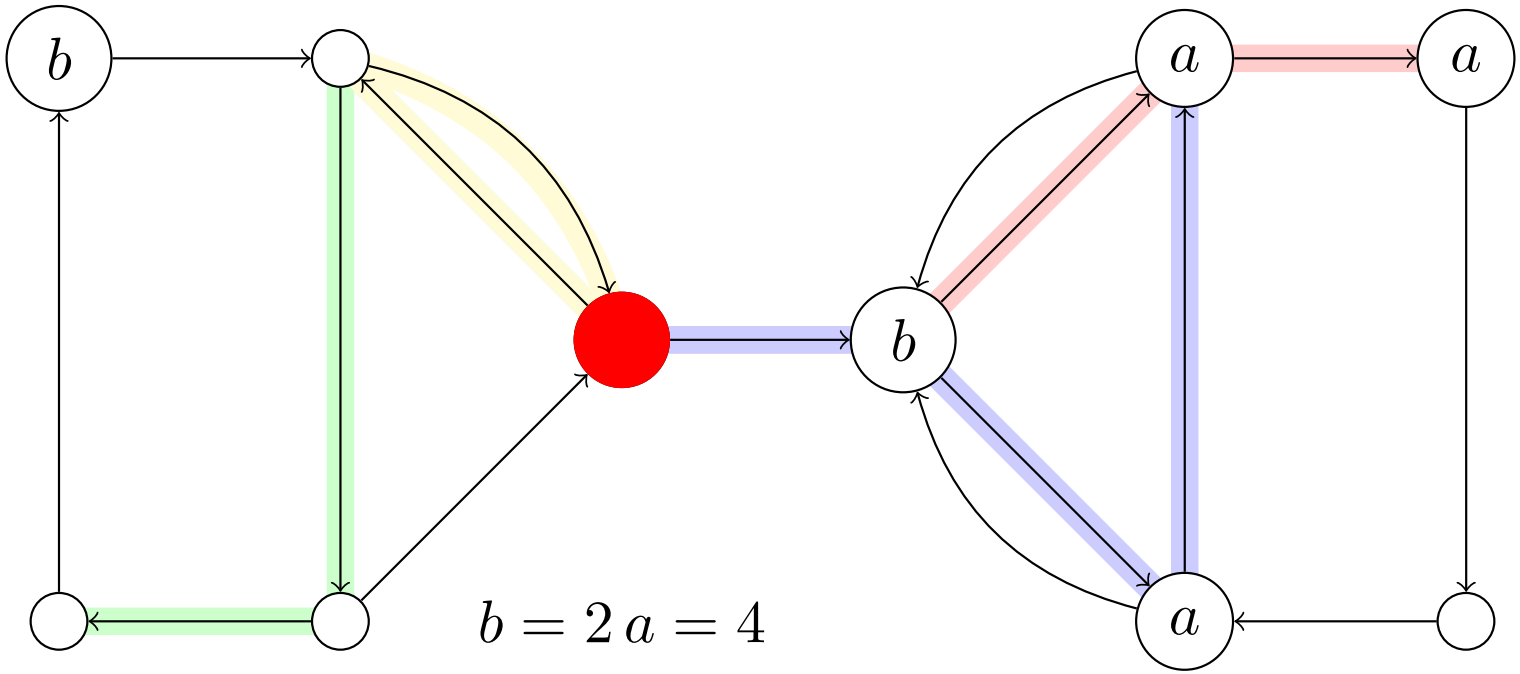
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

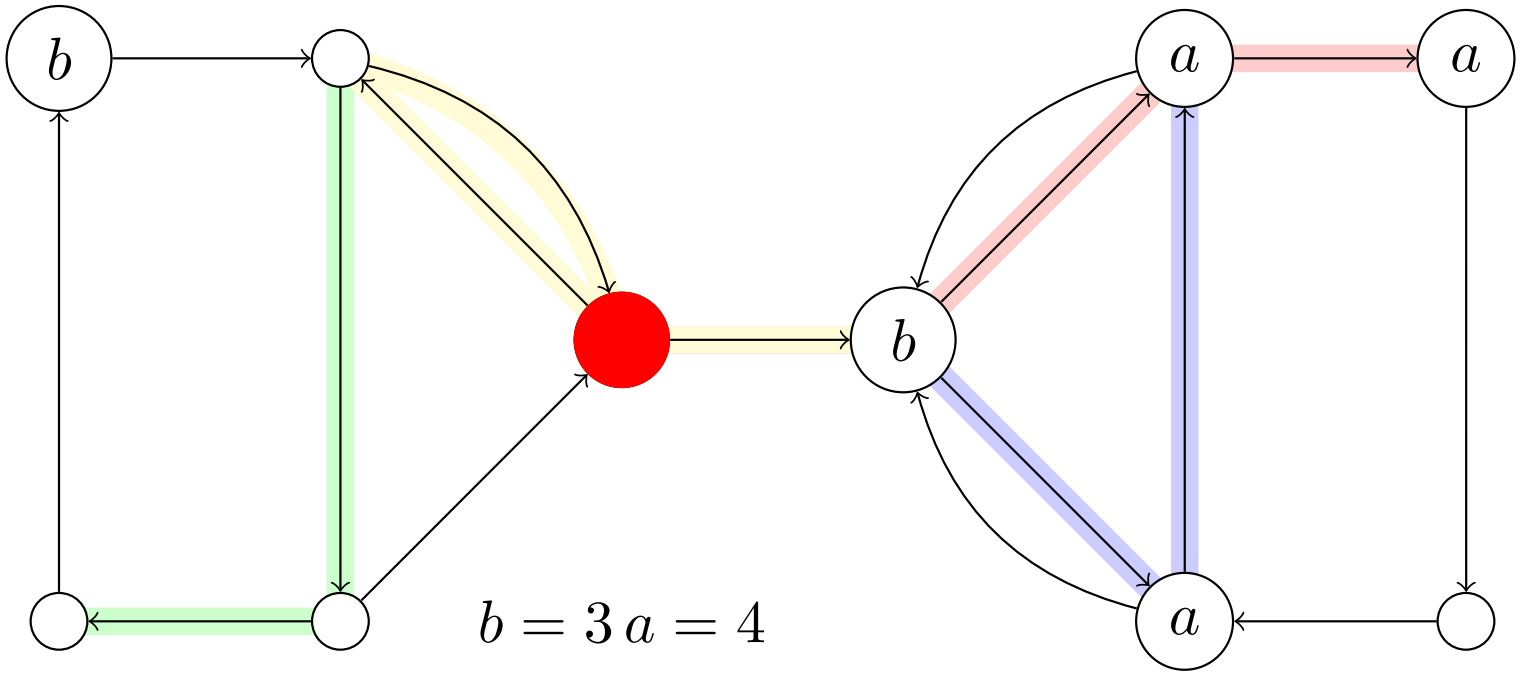
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

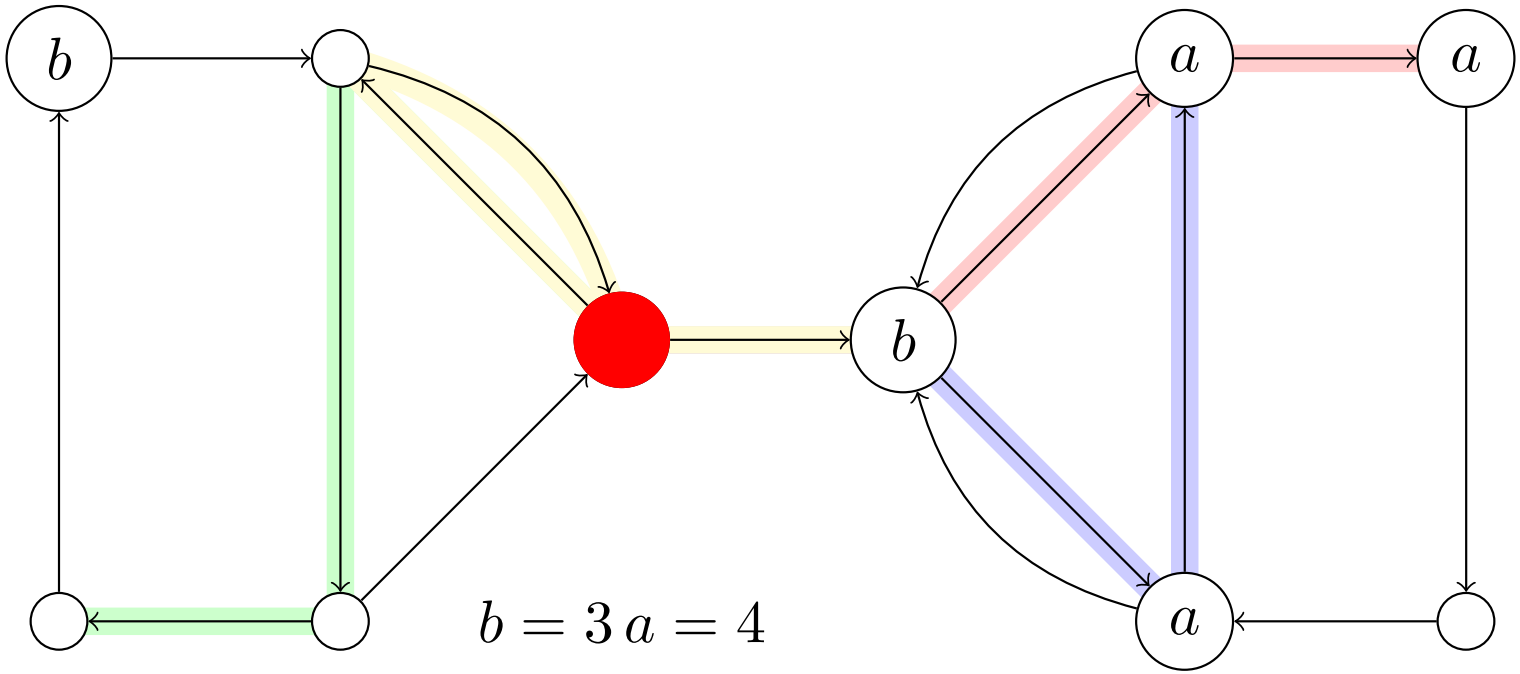
Knotenklassifizierung mit Random Walks



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

Knotenklassifizierung mit Random Walks

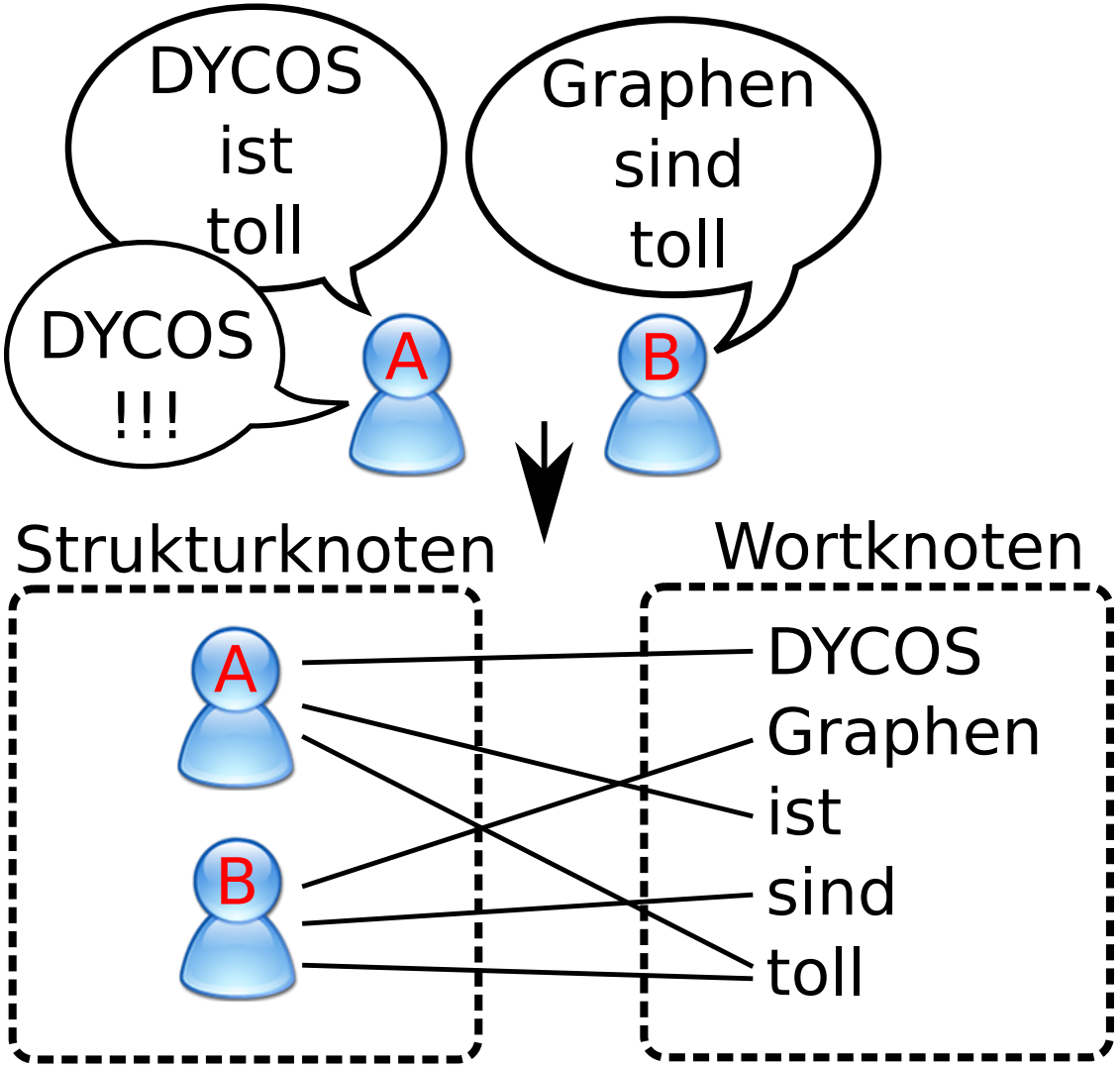


Klassifizieren des roten Knotens:

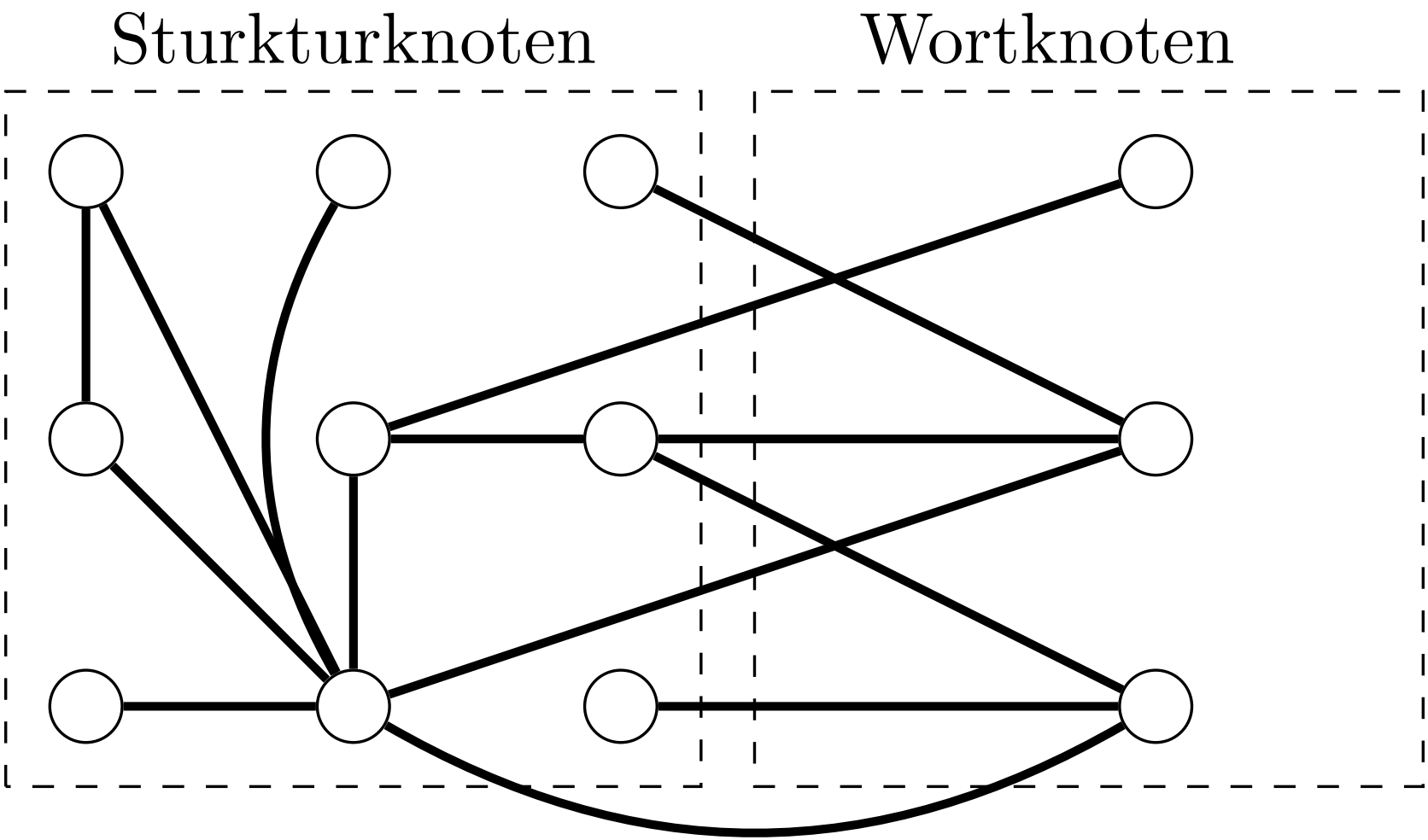
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $3 \cdot b \Rightarrow$ Rot mit a klassifizieren

Wortknoten

- Bisher wurden keine Texte genutzt
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoten
  - vice versa


DYCOS ist toll
Graphen sind toll
DYCOS !!!
A
B
Strukturknoten
Wortknoten
A
B
DYCOS
Graphen
ist
sind
toll

Erweiterter, semi-bipartiter Graph



Vokabular

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$ Beschränkung des Vokabulars sinnvoll

Idee:

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen

Vokabular

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$ Beschränkung des Vokabulars sinnvoll

Idee:

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen

Vokabular

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$ Beschränkung des Vokabulars sinnvoll

Idee:

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen

Vokabular

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$ Beschränkung des Vokabulars sinnvoll

Idee:

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen

Vokabular

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$ Beschränkung des Vokabulars sinnvoll

Idee:

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen

Gini-Koeffizient

- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- Hier: $g \in (0, 1]$
- g nahe bei $1 \Rightarrow$ Wort ist stark ungleich verteilt

$\Rightarrow$ Nehme Top-m Wörter mit höchstem Gini-Koeffizient

Gini-Koeffizient

- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- Hier: $g \in (0, 1]$
- g nahe bei $1 \Rightarrow$ Wort ist stark ungleich verteilt

$\Rightarrow$ Nehme Top-m Wörter mit höchstem Gini-Koeffizient

Gini-Koeffizient

- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- Hier: $g \in (0, 1]$
- g nahe bei $1 \Rightarrow$ Wort ist stark ungleich verteilt

$\Rightarrow$ Nehme Top-m Wörter mit höchstem Gini-Koeffizient

Gini-Koeffizient

- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- Hier: $g \in (0, 1]$
- g nahe bei $1 \Rightarrow$ Wort ist stark ungleich verteilt

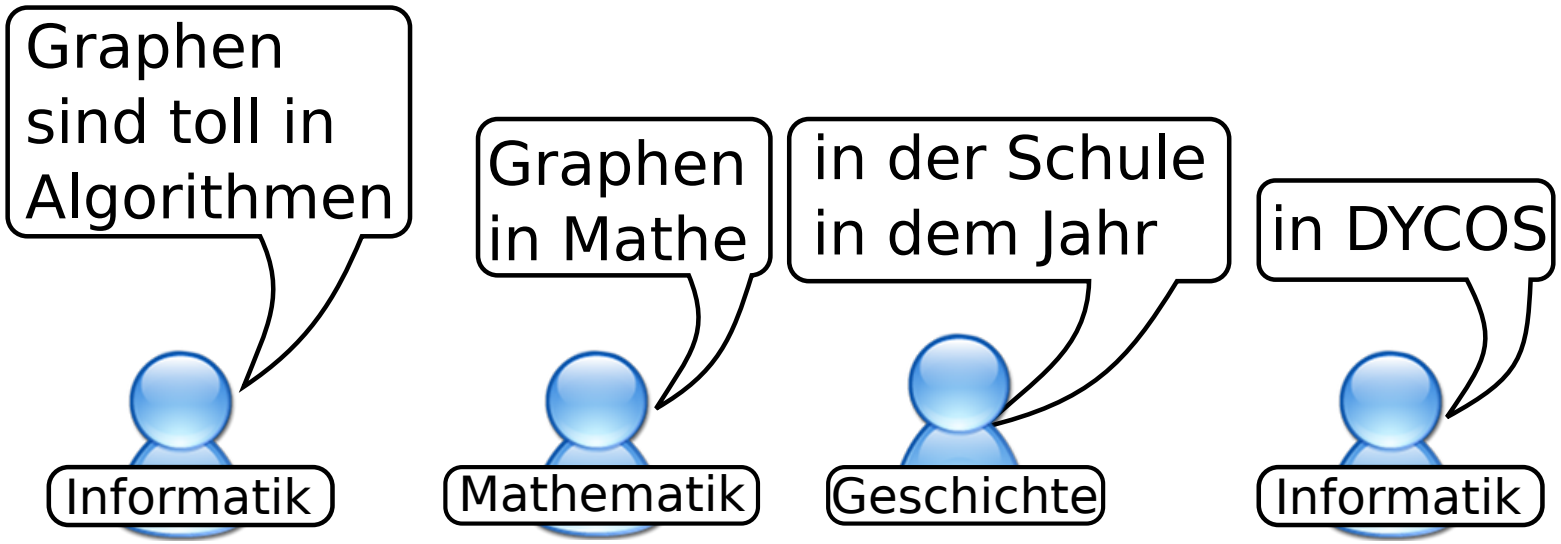
$\Rightarrow$ Nehme Top-m Wörter mit höchstem Gini-Koeffizient

Gini-Koeffizient

- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- Hier: $g \in (0, 1]$
- g nahe bei $1 \Rightarrow$ Wort ist stark ungleich verteilt

$\Rightarrow$ Nehme Top-m Wörter mit höchstem Gini-Koeffizient

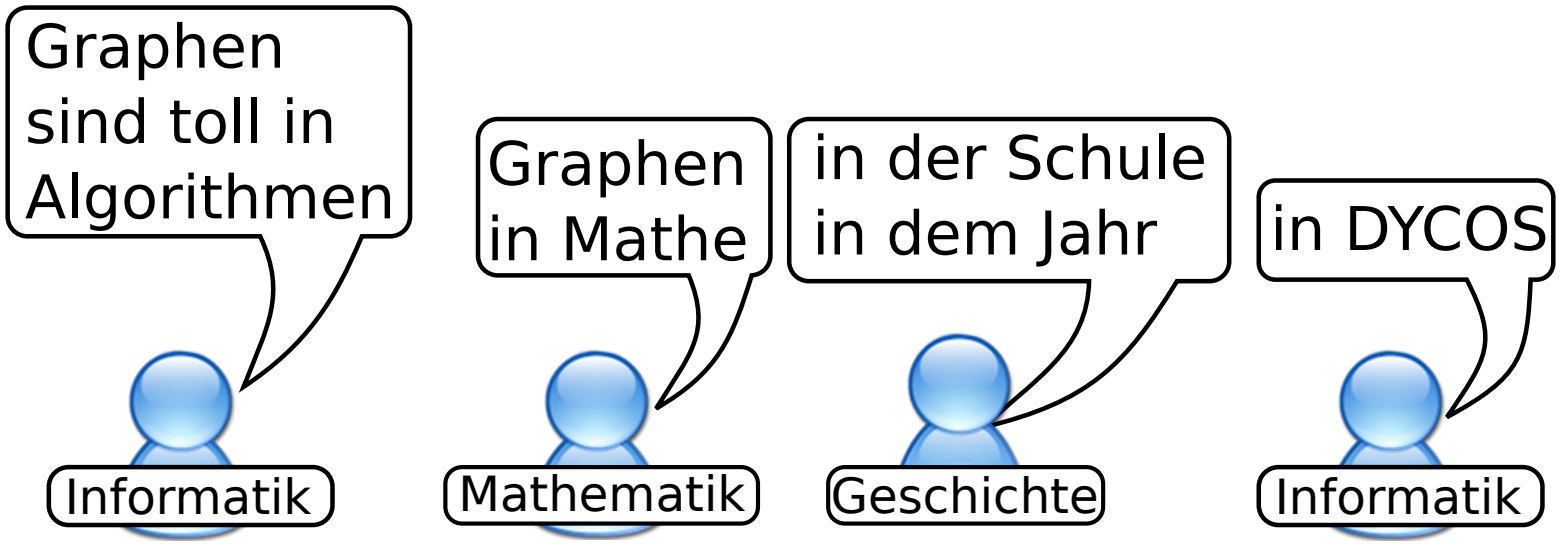
Gini-Koeffizient



Beispiel: „in“

- Vorkommen insgesamt: $5\times$
- Vorkommen in „Informatik“ $2\times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in „Mathematik“ $1\times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in „Geschichte“ $2\times \Rightarrow p_3 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

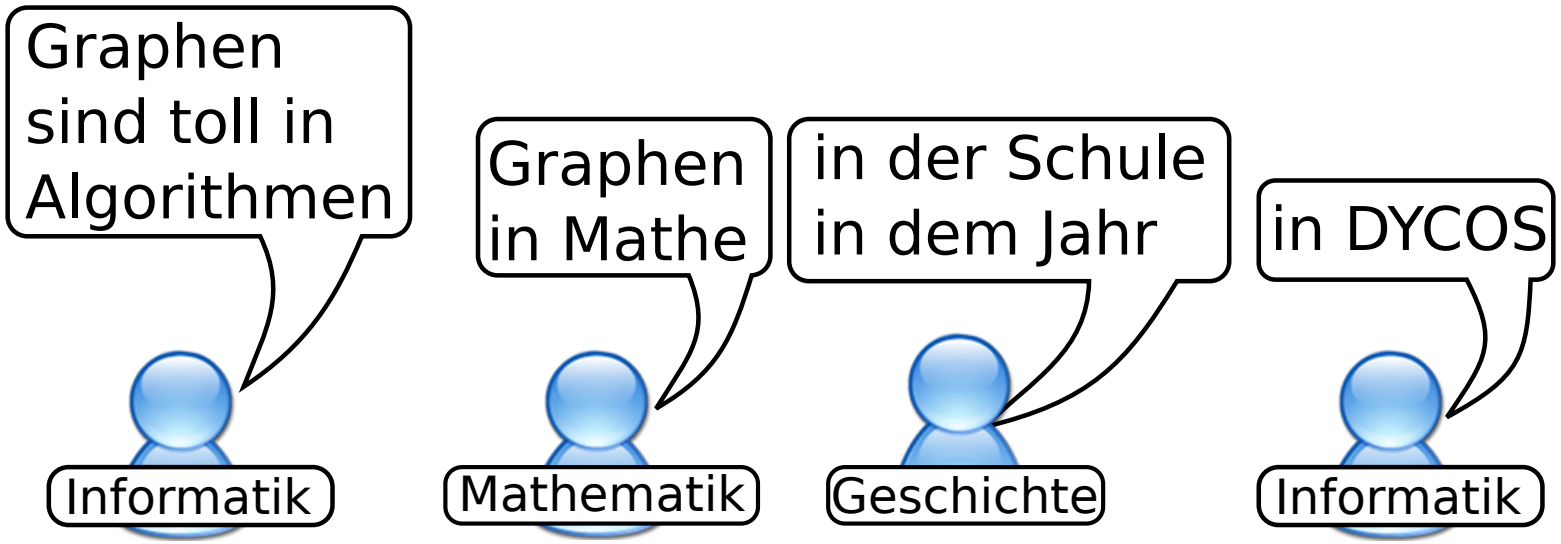
Gini-Koeffizient



Beispiel: „in“

- Vorkommen insgesamt: $5\times$
- Vorkommen in „Informatik“ $2\times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in „Mathematik“ $1\times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in „Geschichte“ $2\times \Rightarrow p_3 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

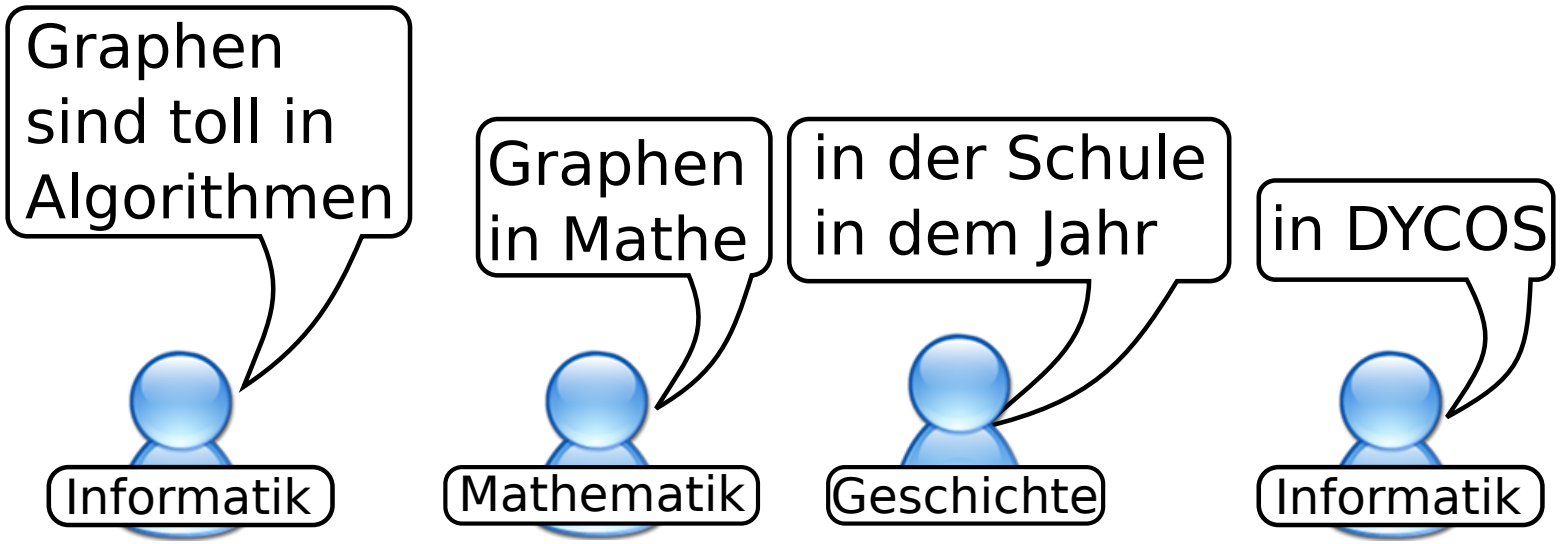
Gini-Koeffizient



Beispiel: „in“

- Vorkommen insgesamt: $5\times$
- Vorkommen in „Informatik“ $2\times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in „Mathematik“ $1\times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in „Geschichte“ $2\times \Rightarrow p_3 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

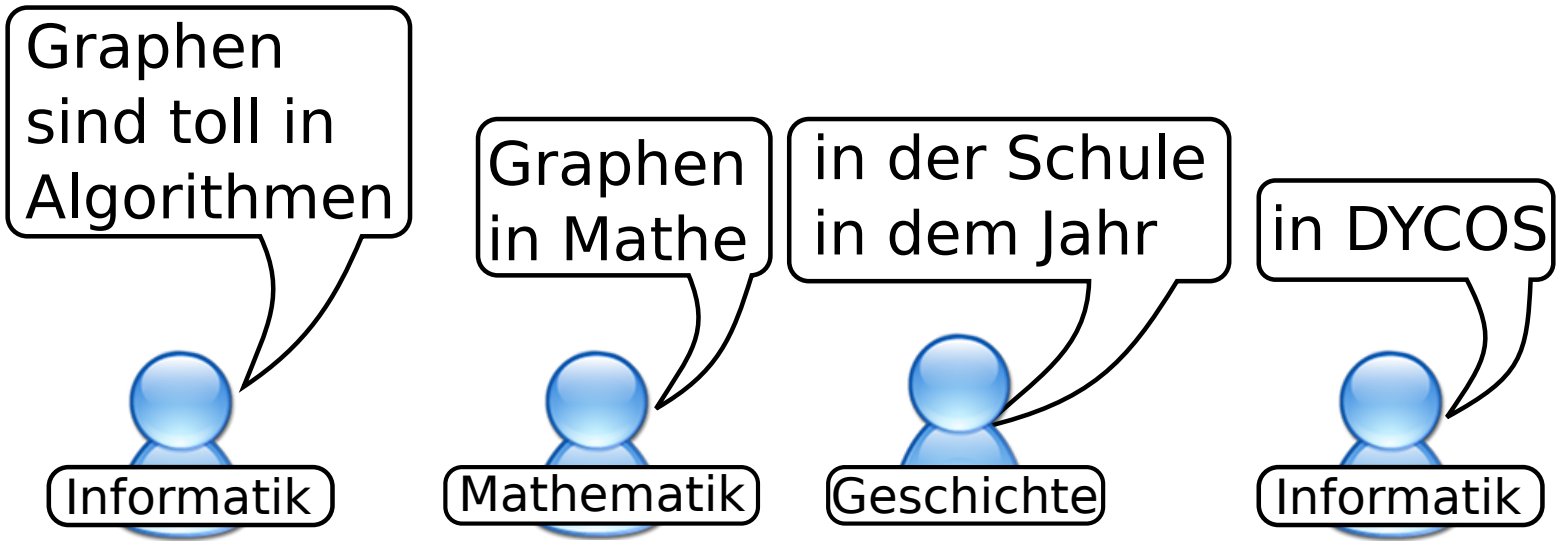
Gini-Koeffizient



Beispiel: „in“

- Vorkommen insgesamt: $5\times$
- Vorkommen in „Informatik“ $2\times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in „Mathematik“ $1\times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in „Geschichte“ $2\times \Rightarrow p_3 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

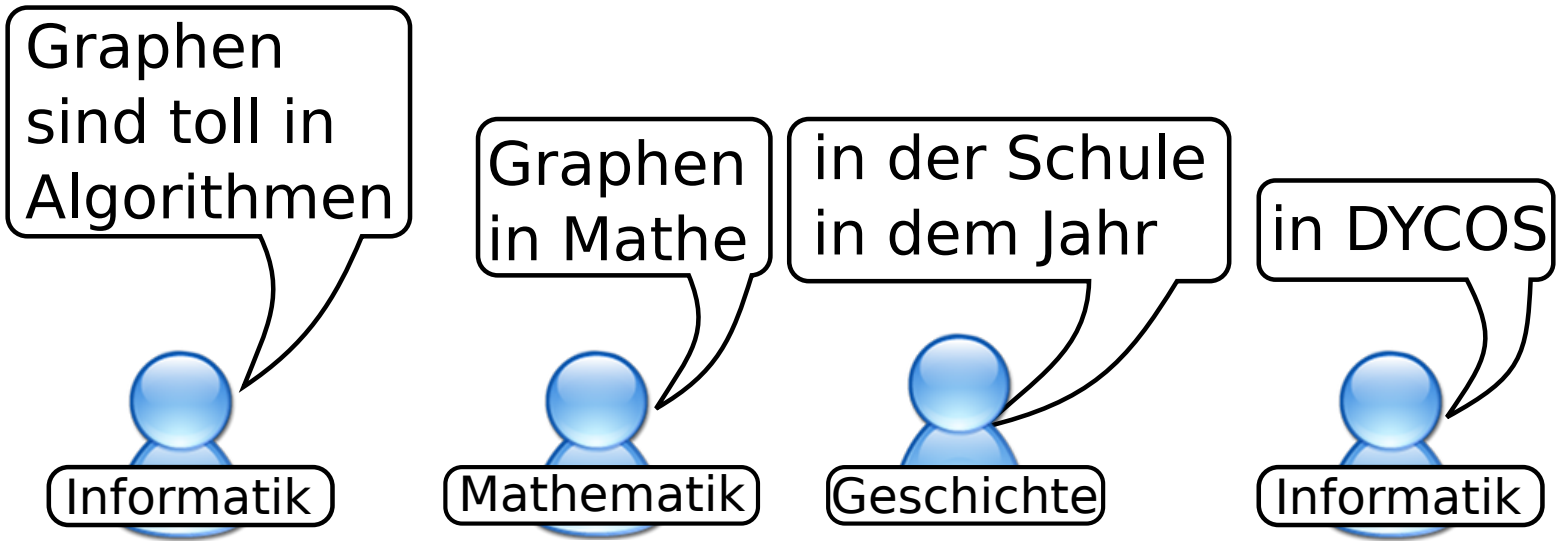
Gini-Koeffizient



Beispiel: „in“

- Vorkommen insgesamt: $5\times$
- Vorkommen in „Informatik“ $2\times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in „Mathematik“ $1\times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in „Geschichte“ $2\times \Rightarrow p_3 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

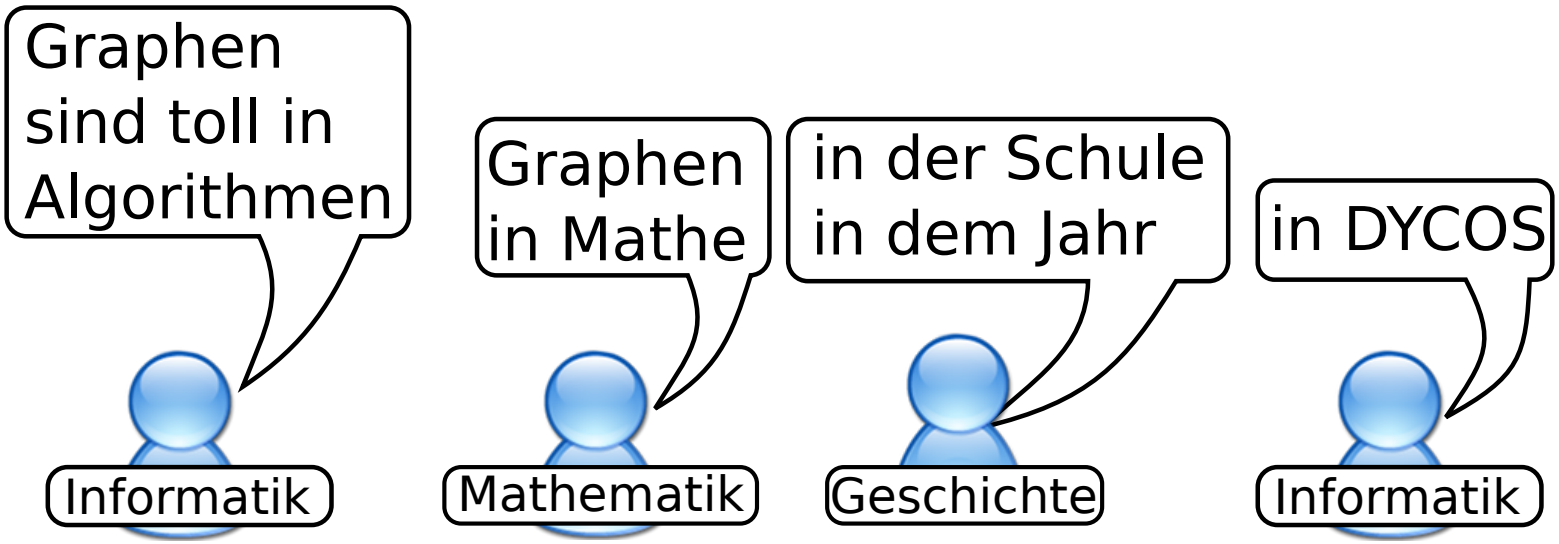
Gini-Koeffizient



Beispiel: „in“

- Vorkommen insgesamt: $5\times$
- Vorkommen in „Informatik“ $2\times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in „Mathematik“ $1\times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in „Geschichte“ $2\times \Rightarrow p_3 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

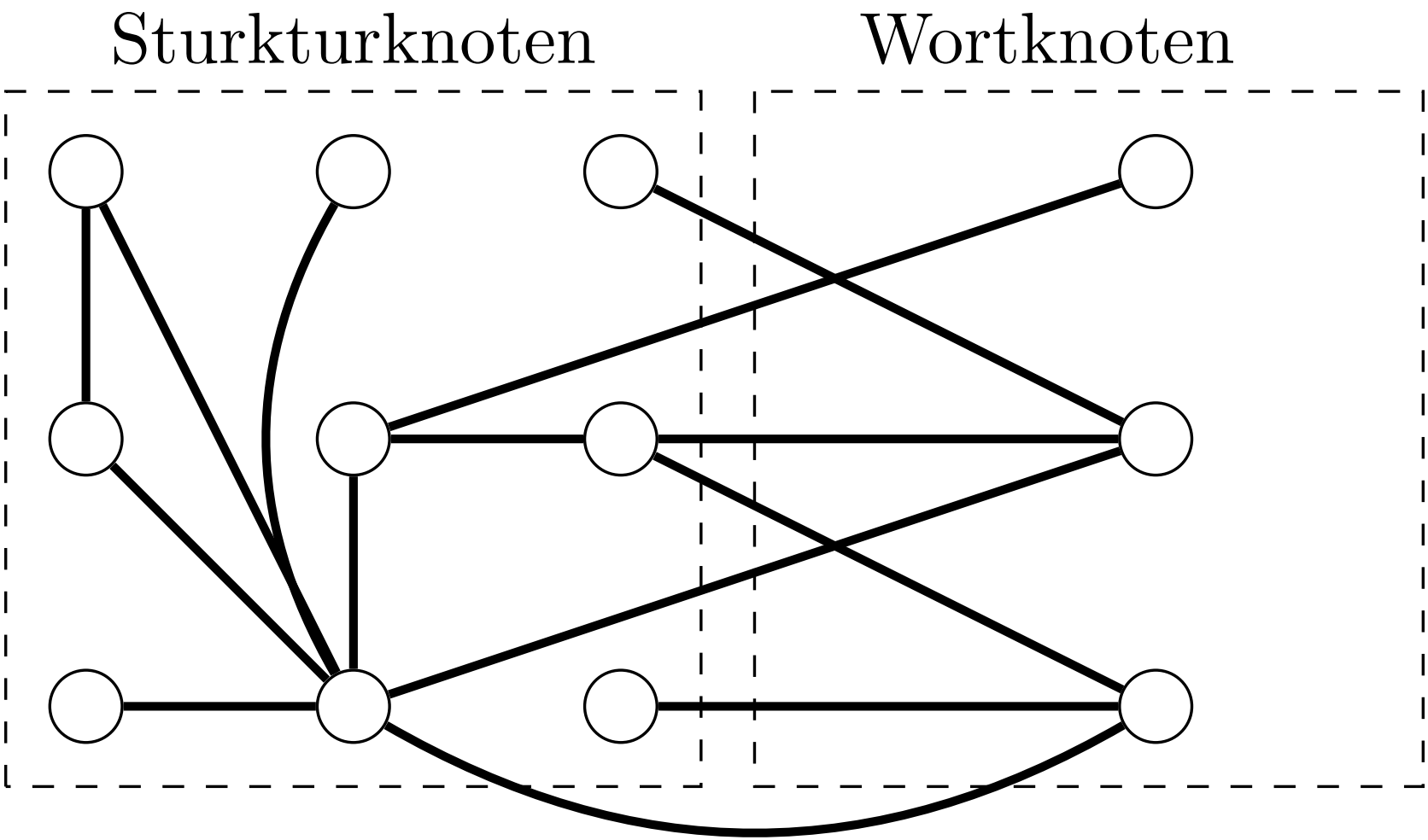
Gini-Koeffizient



Beispiel: „in“

- Vorkommen insgesamt: $5\times$
- Vorkommen in „Informatik“ $2\times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in „Mathematik“ $1\times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in „Geschichte“ $2\times \Rightarrow p_3 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

Sprungtypen



Inhaltlicher Zweifachsprung

- **Strukturspung**: von Strukturknoten v zu Strukturknoten v'
- **Inhaltlicher Zweifachsprung**: von Strukturknoten v über Wortknoten zu Strukturknoten v'
  - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
  - Nehme Top-q-Knoten (Anzahl der Pfade)
  - Wähle zufällig einen davon

Inhaltlicher Zweifachsprung

- **Strukturspung**: von Strukturknoten v zu Strukturknoten v'
- **Inhaltlicher Zweifachsprung**: von Strukturknoten v über Wortknoten zu Strukturknoten v'
  - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
  - Nehme Top-q-Knoten (Anzahl der Pfade)
  - Wähle zufällig einen davon

Inhaltlicher Zweifachsprung

- **Strukturspung**: von Strukturknoten v zu Strukturknoten v'
- **Inhaltlicher Zweifachsprung**: von Strukturknoten v über Wortknoten zu Strukturknoten v'
  - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
  - Nehme Top-q-Knoten (Anzahl der Pfade)
  - Wähle zufällig einen davon

Inhaltlicher Zweifachsprung

- **Strukturspung**: von Strukturknoten v zu Strukturknoten v'
- **Inhaltlicher Zweifachsprung**: von Strukturknoten v über Wortknoten zu Strukturknoten v'
  - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
  - Nehme Top-q-Knoten (Anzahl der Pfade)
  - Wähle zufällig einen davon

Inhaltlicher Zweifachsprung

- **Strukturspung**: von Strukturknoten v zu Strukturknoten v'
- **Inhaltlicher Zweifachsprung**: von Strukturknoten v über Wortknoten zu Strukturknoten v'
  - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
  - Nehme Top-q-Knoten (Anzahl der Pfade)
  - Wähle zufällig einen davon

Evaluation

Name	Knoten	davon beschriftet	Kanten	Beschriftungen
CORA	19 396	14 814	75 021	5
DBLP	806 635	18 999	4 414 135	5

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Evaluation

- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% - 84%
  - DBLP: 61% - 66%

Wichtige Ideen

- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Wichtige Ideen

- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Wichtige Ideen

- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Was ist an DYCOS dynamisch?

- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach

$\Rightarrow$ Der Graph darf dynamisch sein; DYCOS funktioniert dennoch

Was ist an DYCOS dynamisch?

- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach

$\Rightarrow$ Der Graph darf dynamisch sein; DYCOS funktioniert dennoch

Was ist an DYCOS dynamisch?

- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach

$\Rightarrow$ Der Graph darf dynamisch sein; DYCOS funktioniert dennoch

Was ist an DYCOS dynamisch?

- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach

$\Rightarrow$ Der Graph darf dynamisch sein; DYCOS funktioniert dennoch

Danke!

Gibt es Fragen?

Bildquellen

- Crystal_Clear_app_personal.png von Wikipedia Commons

Literatur

- Charu C. Aggarwal, Nan Li: *On Node Classification in Dynamic Content-based Networks*.
- Smriti Bhagat, Graham Cormode und S. Muthukrishnan. *Node Classification in Social Networks*.
- M. F. Porter. Readings in Information Retrieval. Kapitel *An Algorithm for Suffix Stripping*.
- Jeffrey S. Vitter. *Random Sampling with a Reservoir*.

Folien, LaTeXund Material

Der Foliensatz sowie die LaTeXund TikZ-Quellen sind unter
github.com/MartinThoma/LaTeX-examples/tree/master/presentations/Datamining-Proseminar
Kurz-URL: tinyurl.com/thoma-ps