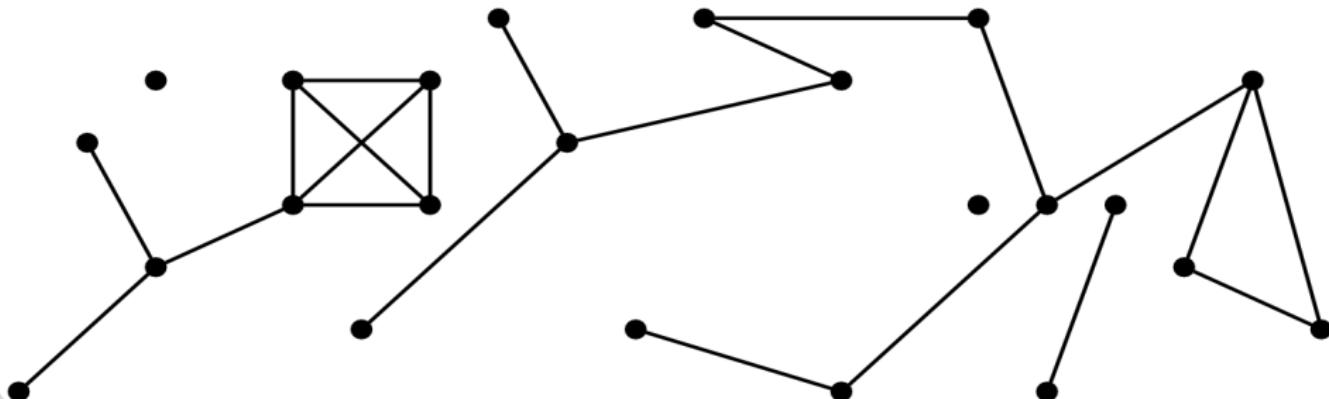


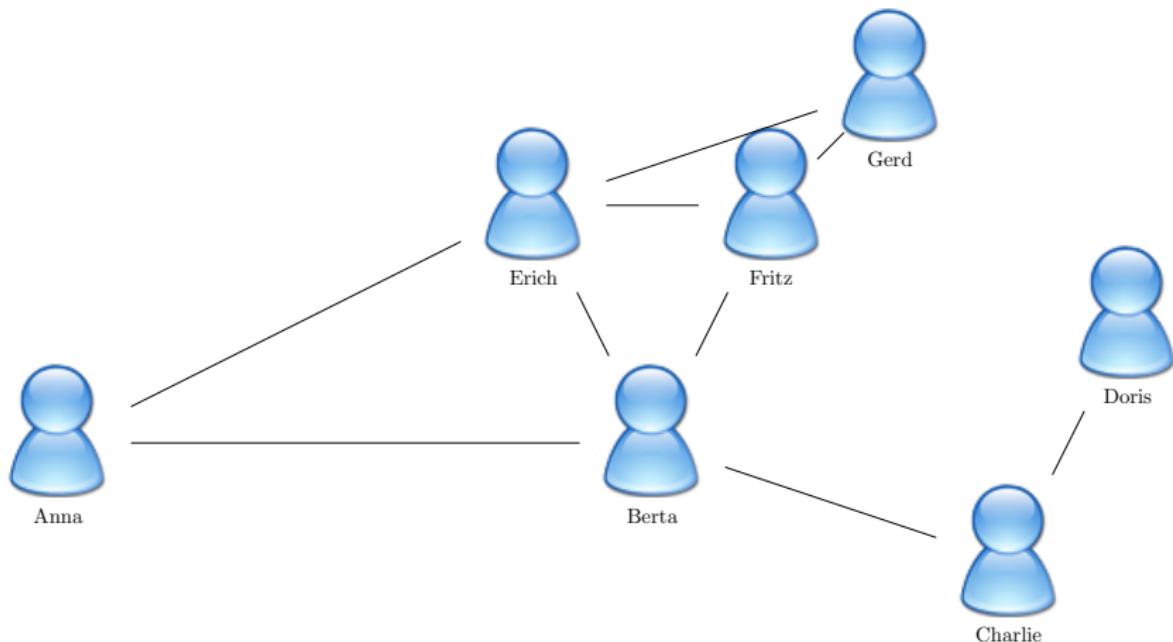
On Node Classification in Dynamic Content-based Networks

Martin Thoma | 28. Februar 2014

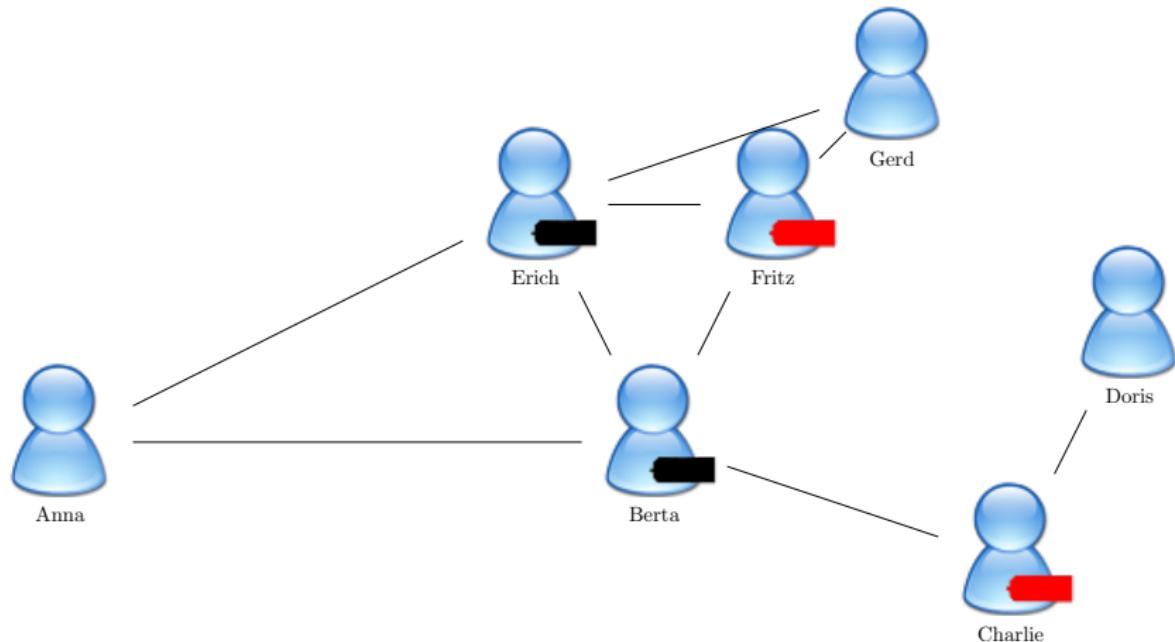
INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



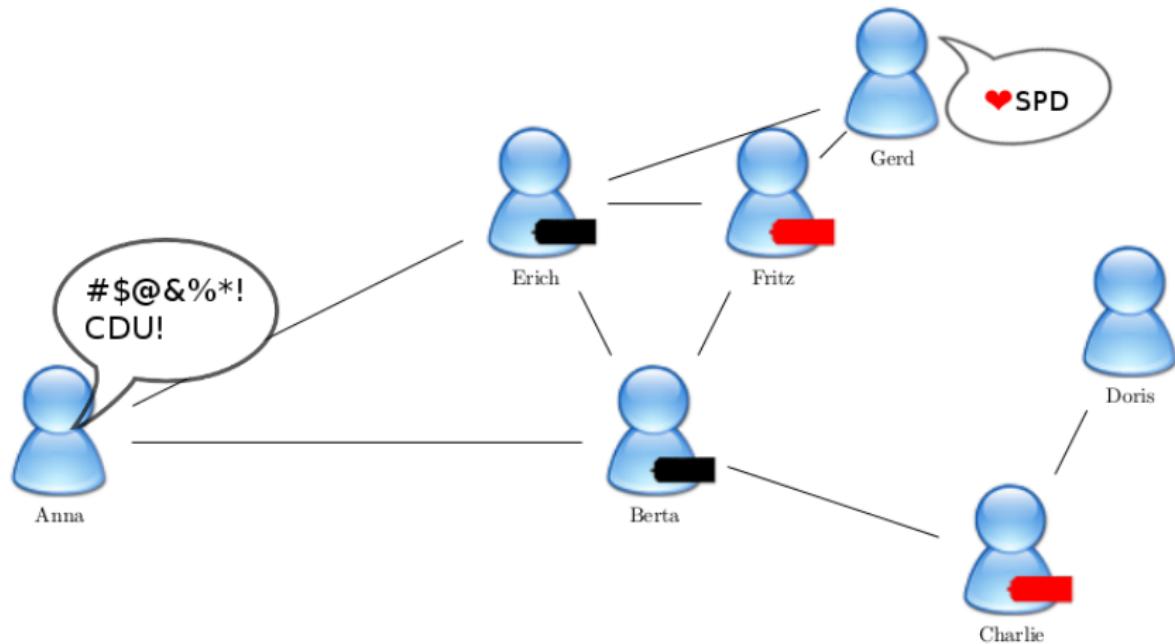
Social Network



Partially labeled network



Partially labeled network with content



Beispiel 2: Literaturdatenbanken

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

The C Programming
Language
digital restoration
and typesetter

The Identity
Thesis for
Language and
Music



Computer Science

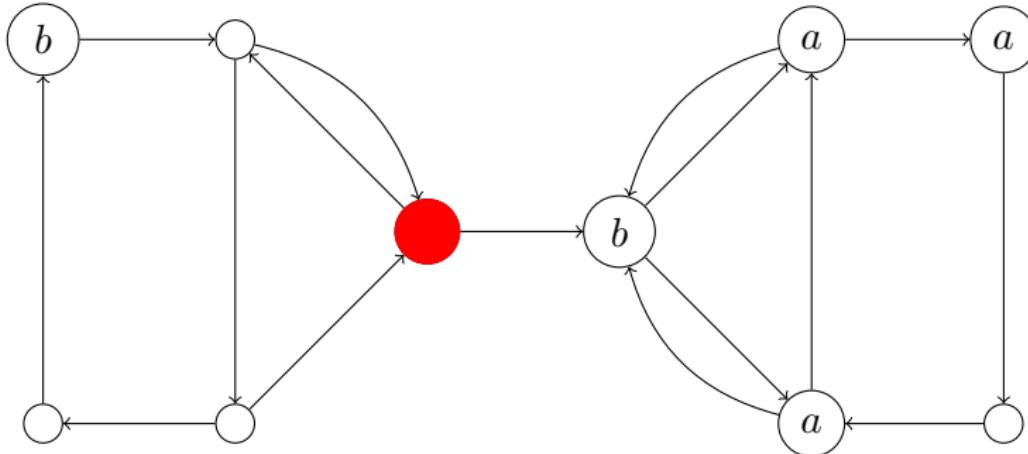


Computer Science



Linguistics

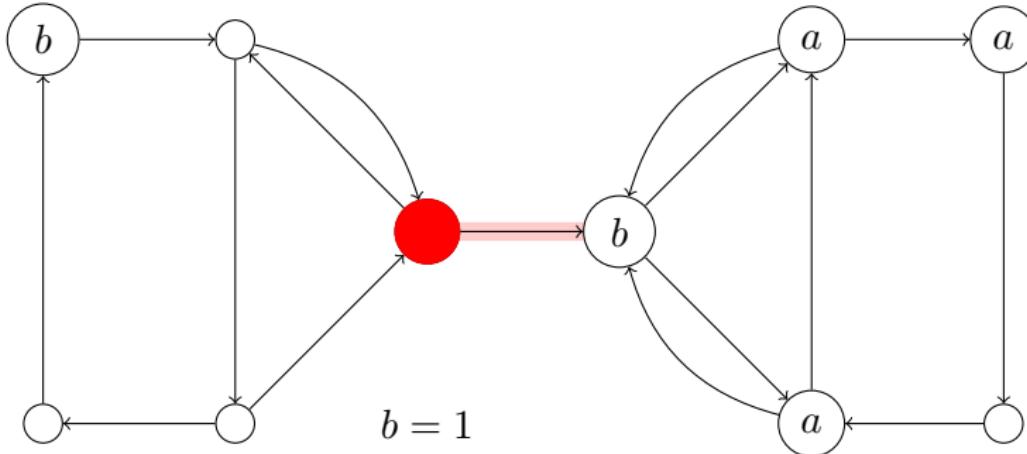
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

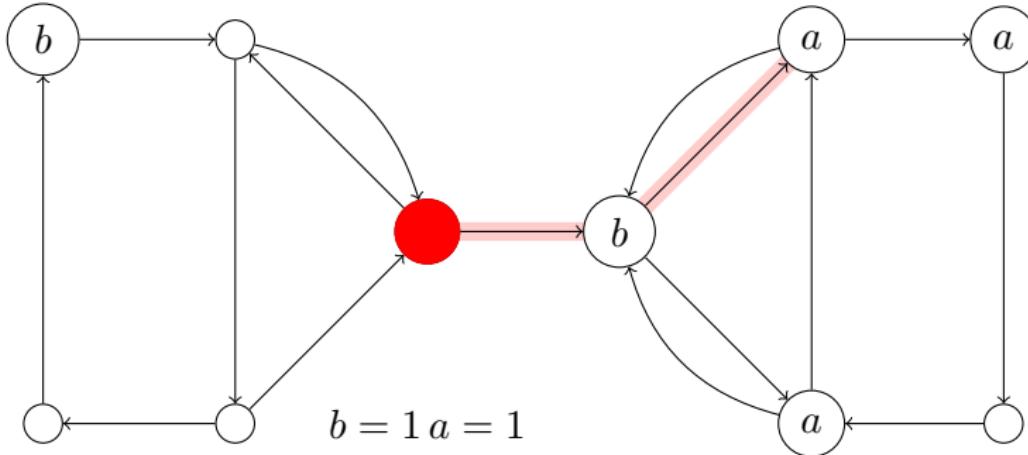
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

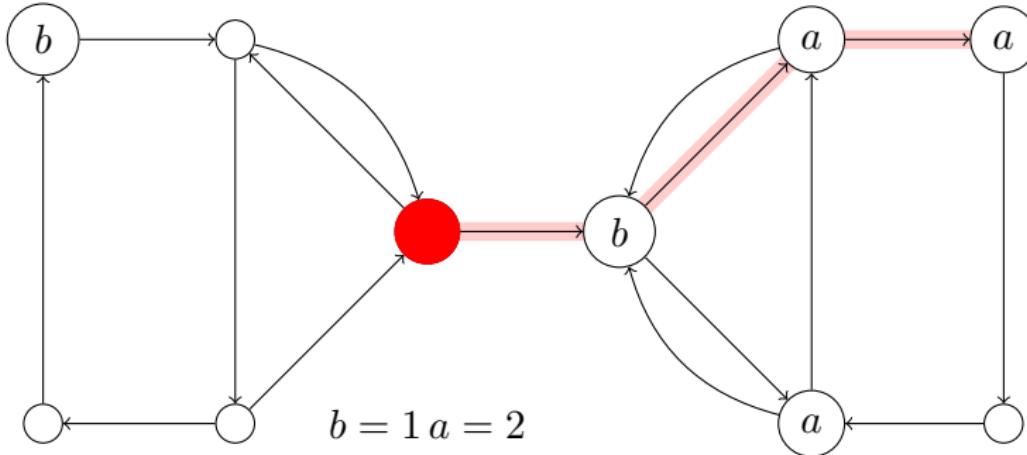
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

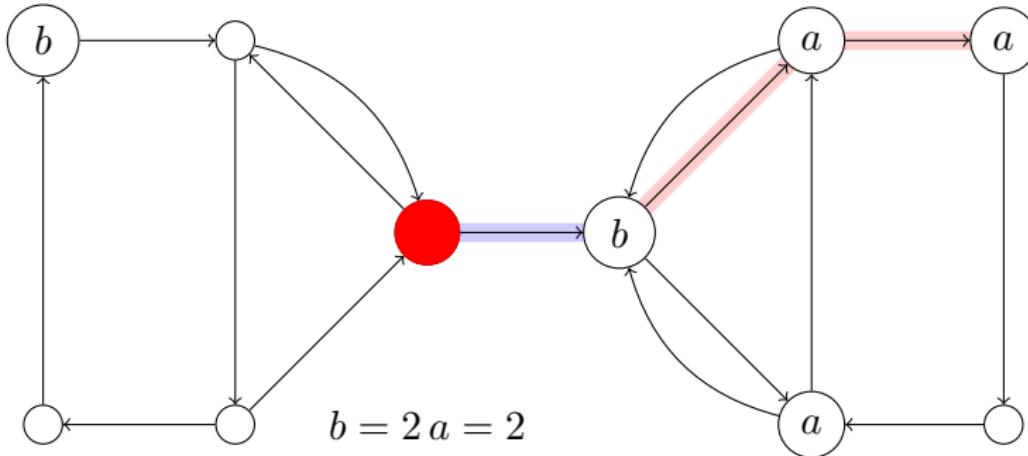
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

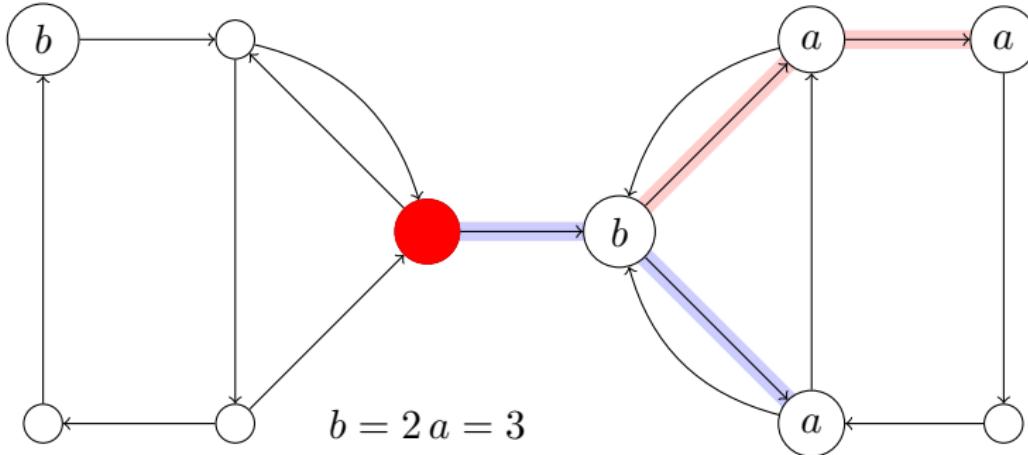
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

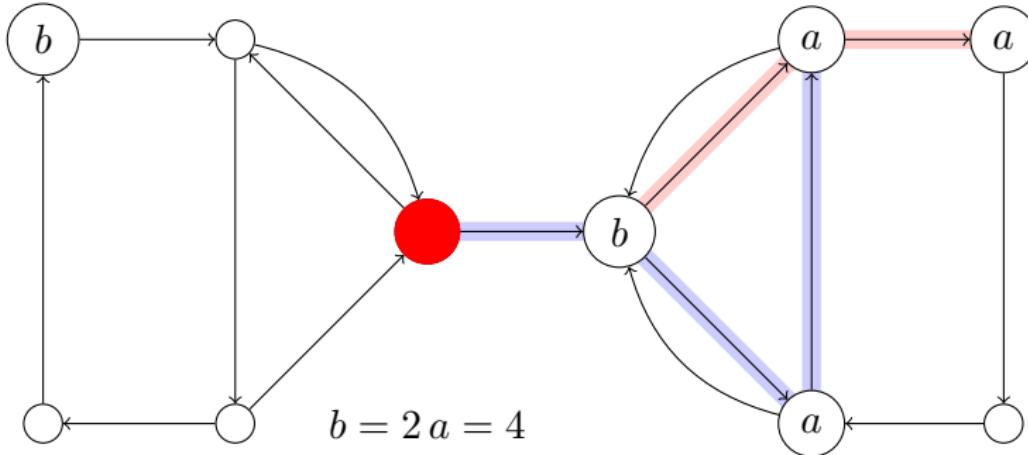
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

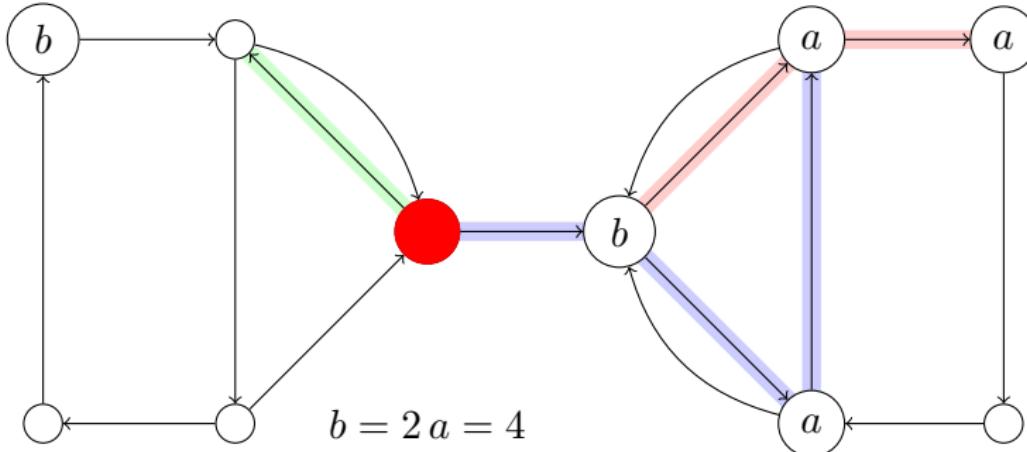
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

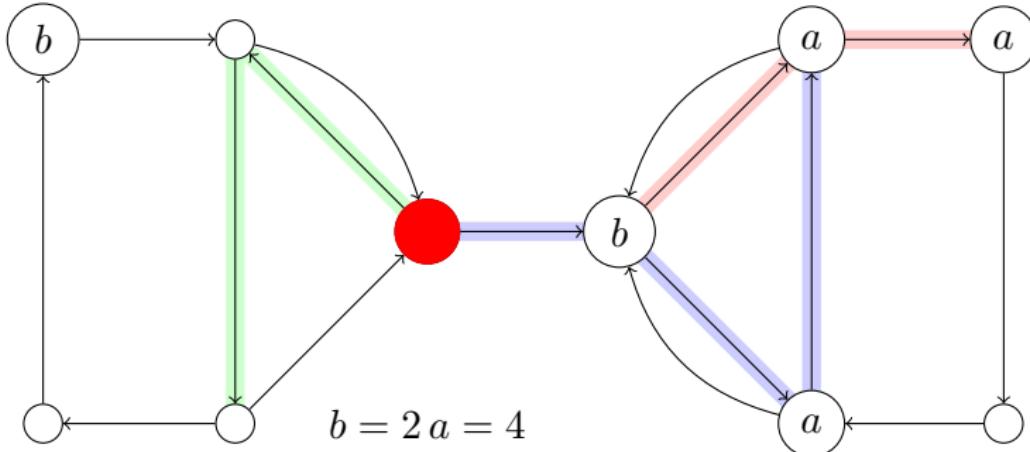
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

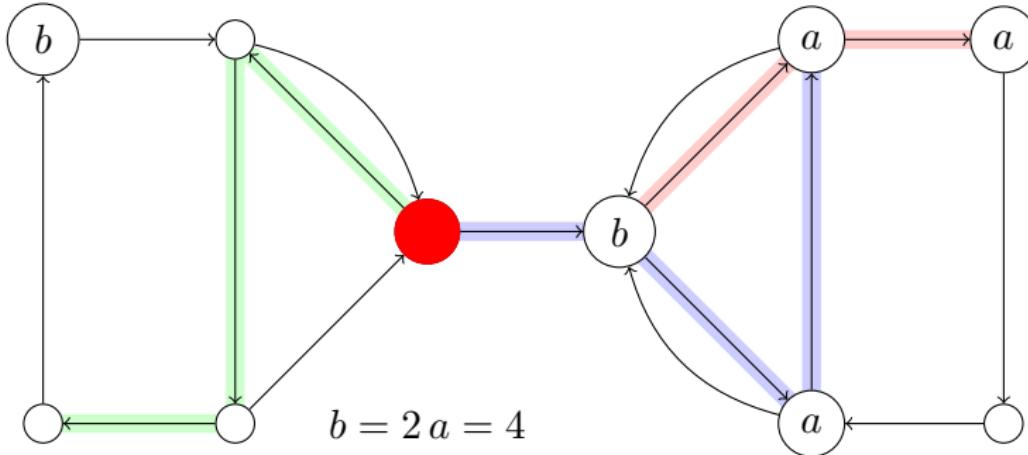
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

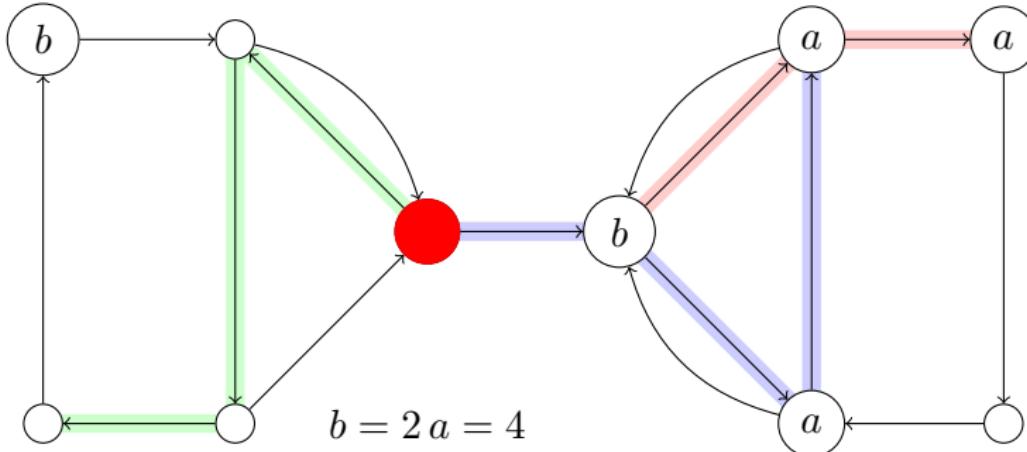
Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

Random Walk



Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren

- Neben Struktur können Texte genutzt werden
 - Einschränkung: Effizienz!
 - Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa

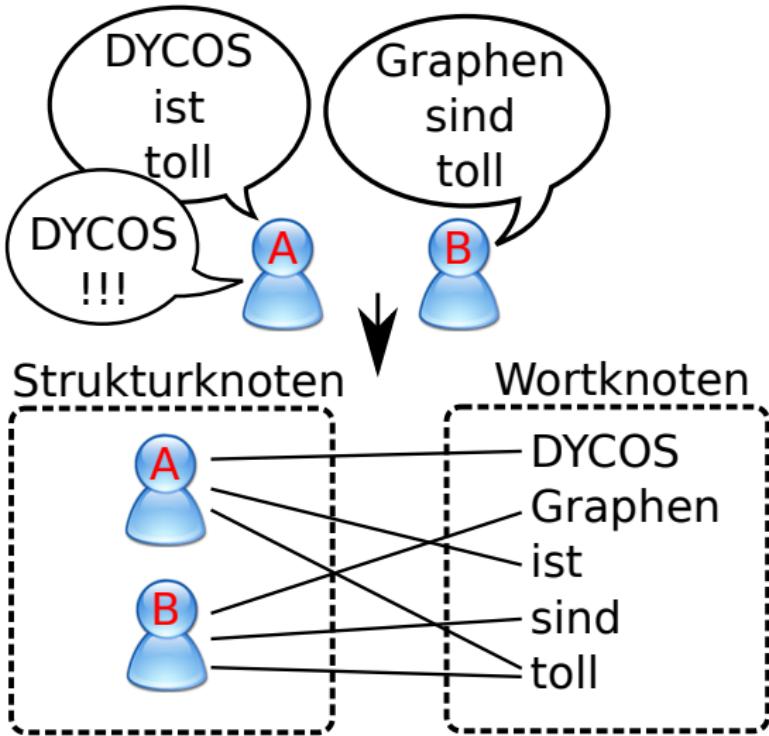
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa

- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa

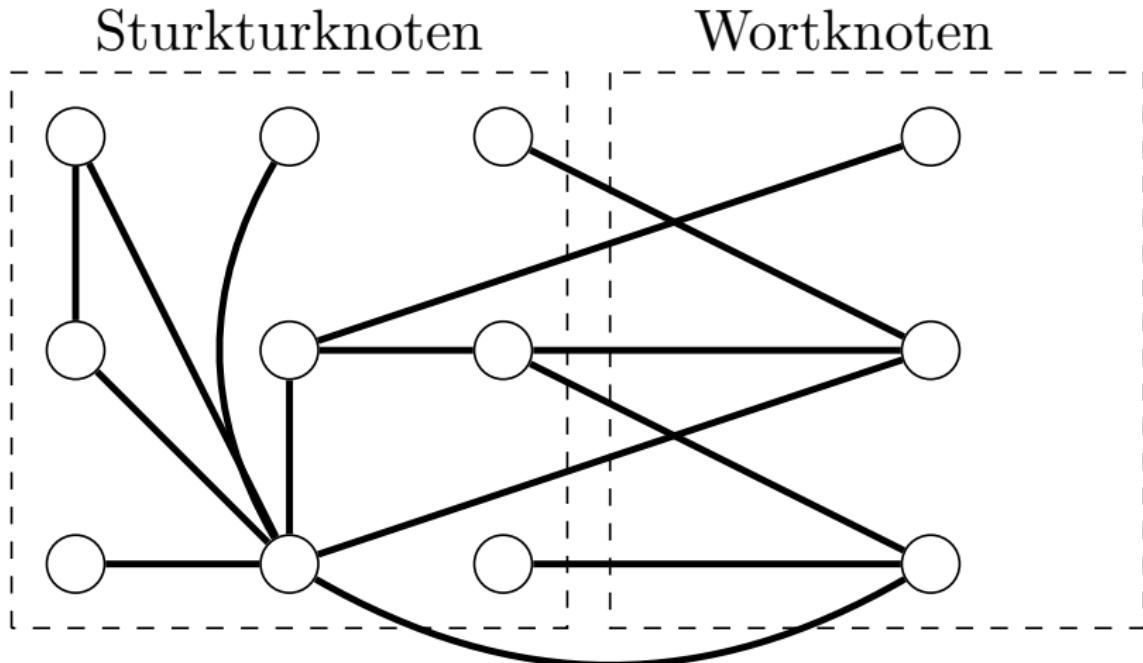
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa

- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa

- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa



Erweiterter, semi-bipartiter Graph



- Füllwörter: und, oder, im, in, ...
⇒ Beschränkung des Vokabulars sinnvoll

Idee:

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

- Füllwörter: und, oder, im, in, ...
⇒ Beschränkung des Vokabulars sinnvoll

Idee:

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

- Füllwörter: und, oder, im, in, ...
⇒ Beschränkung des Vokabulars sinnvoll

Idee:

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

- Füllwörter: und, oder, im, in, ...
⇒ Beschränkung des Vokabulars sinnvoll

Idee:

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

- Füllwörter: und, oder, im, in, ...
⇒ Beschränkung des Vokabulars sinnvoll

Idee:

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

- Füllwörter: und, oder, im, in, ...
⇒ 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
 - $g \in (0, 1]$
 - 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
 - $g \in (0, 1]$
 - 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
 - $g \in (0, 1]$
 - g nahe bei 1 \Rightarrow Wort ist stark ungleich verteilt
- \Rightarrow Nehme Top- m Wörter mit höchstem Gini-Koeffizient

Gini-Koeffizient

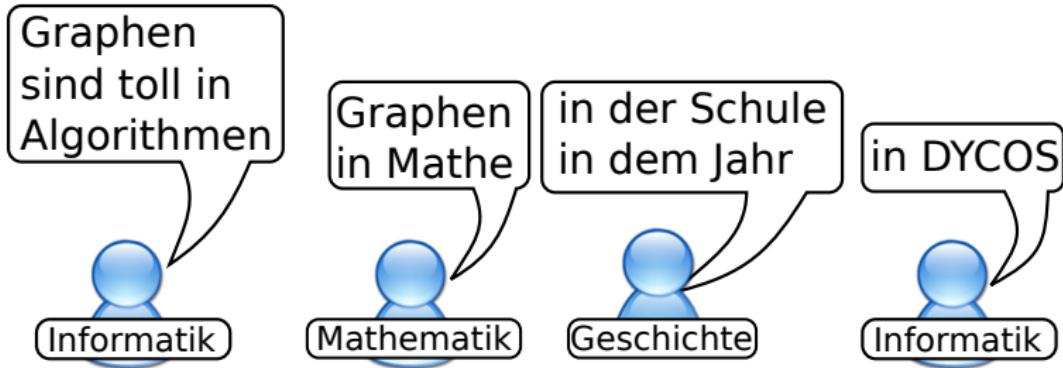
- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0, 1]$
- g nahe bei 1 \Rightarrow Wort ist stark ungleich verteilt

\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
 - $g \in (0, 1]$
 - g nahe bei 1 \Rightarrow Wort ist stark ungleich verteilt
- \Rightarrow Nehme Top- m Wörter mit höchstem Gini-Koeffizient

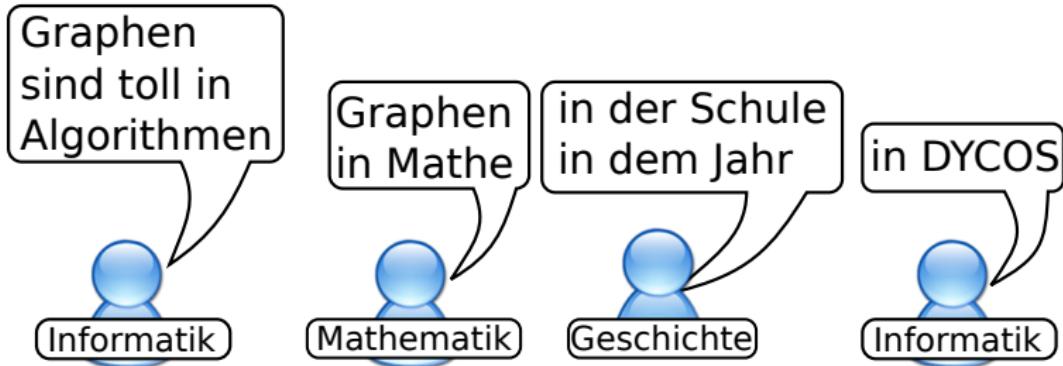
Gini-Koeffizient



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: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

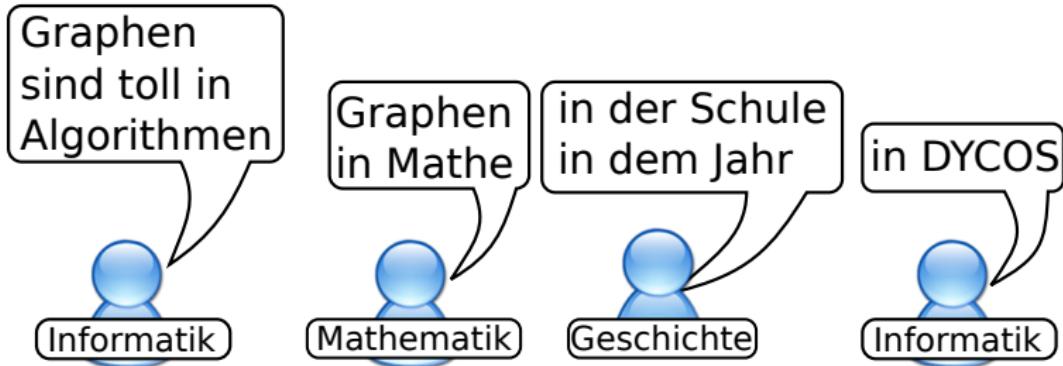
Gini-Koeffizient



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: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

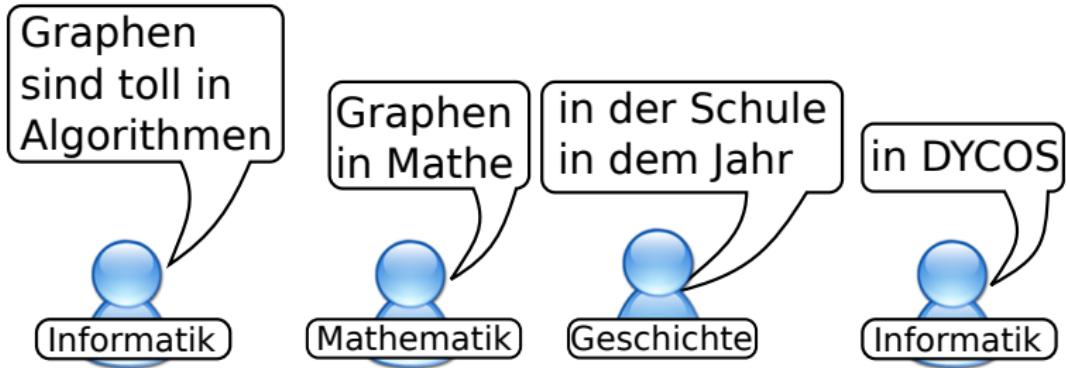
Gini-Koeffizient



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: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

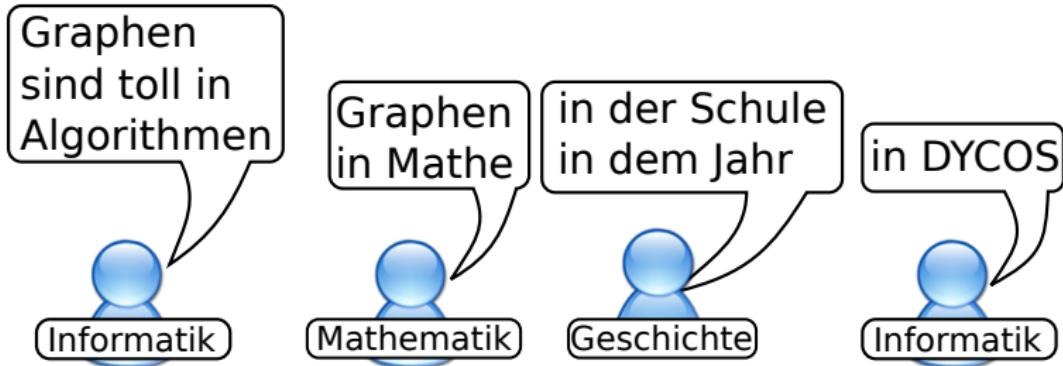
Gini-Koeffizient



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: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

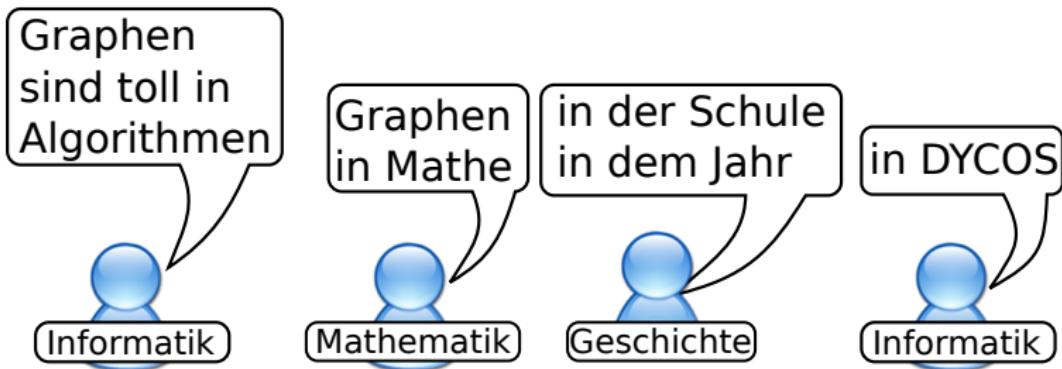
Gini-Koeffizient



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: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

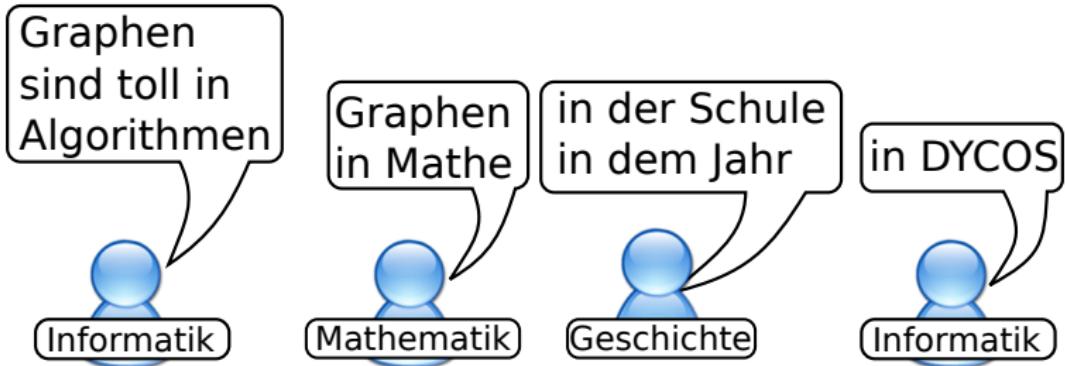
Gini-Koeffizient



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: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

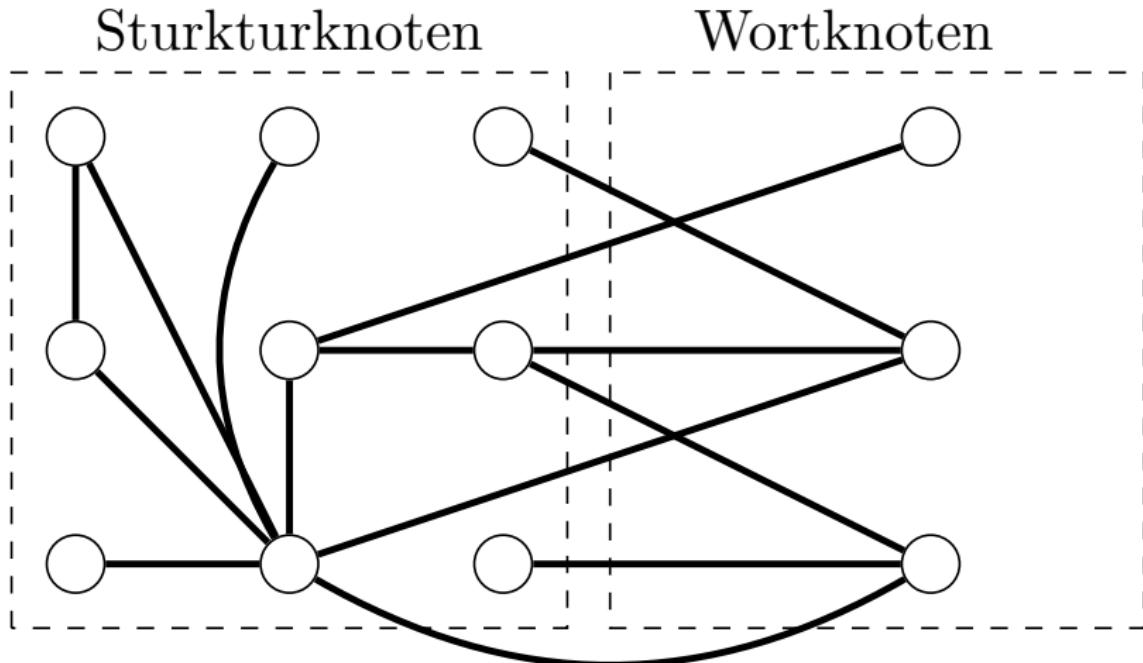
Gini-Koeffizient



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

- **Struktursprung:** 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- g -Knoten (Anzahl der Pfade)
 - Wähle zufällig einen davon

- **Struktursprung:** 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

- **Struktursprung:** 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

- **Struktursprung:** 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

- **Struktursprung:** 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

Wichtige Ideen

- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Wichtige Ideen

- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Wichtige Ideen

- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Dynamisch?

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

Dynamisch?

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

Dynamisch?

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

Danke!

Gibt es Fragen?

- Crystal_Clear_app_personal.png von [Wikipedia Commons](#)

- 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*.

Der Foliensatz und die L^AT_EX und TikZ-Quellen sind unter
github.com/MartinThoma/LaTeX-examples/tree/master/presentations/Datamining-Proseminar
Kurz-URL: tinyurl.com/Info-Proseminar