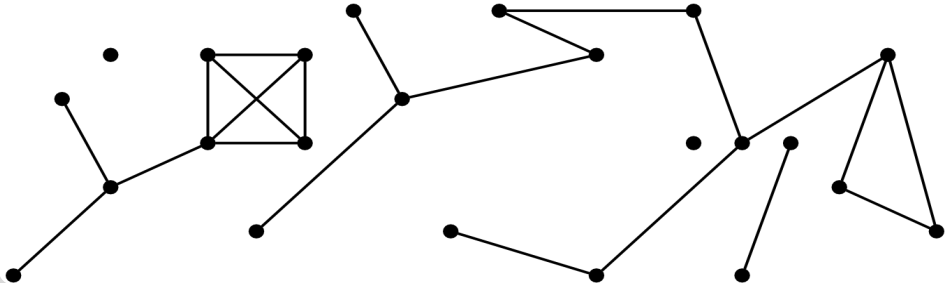
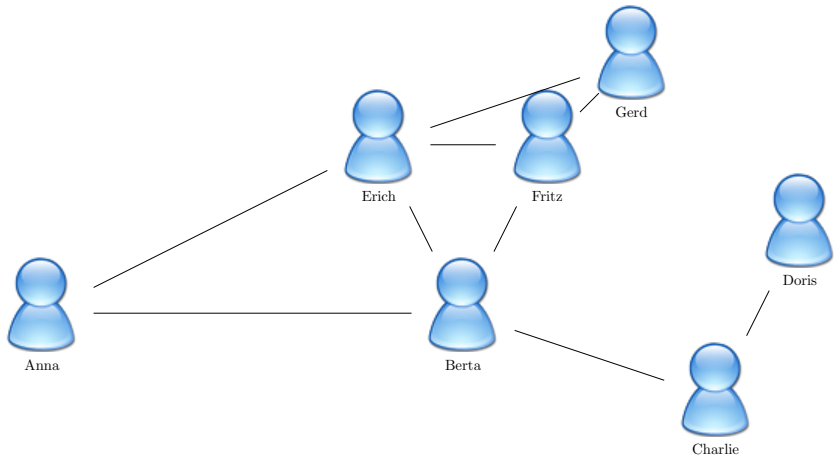


# On Node Classification in Dynamic Content-based Networks

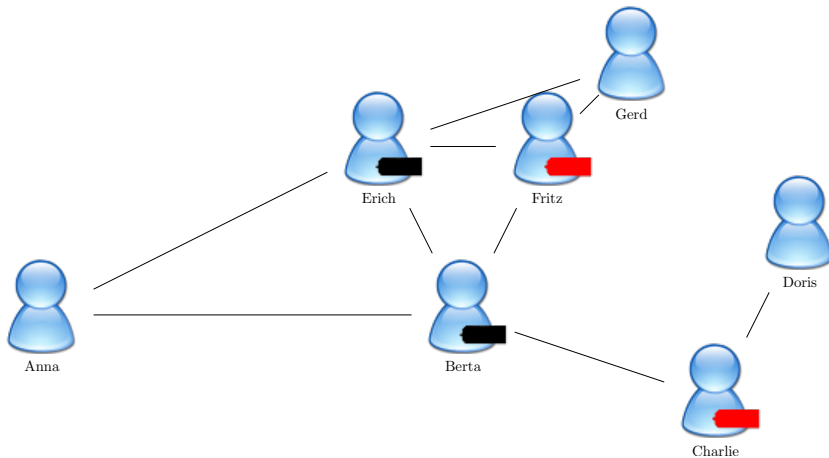
Martin Thoma | 28. Februar 2014

INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION

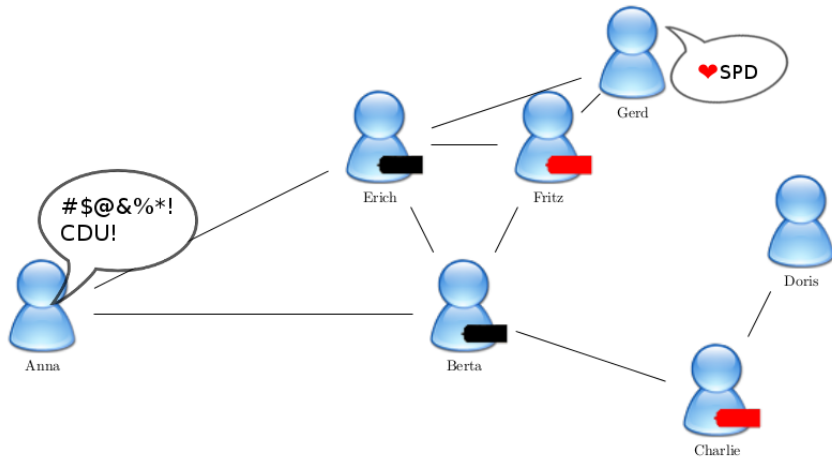




# Partially labeled network



# Partially labeled network with content



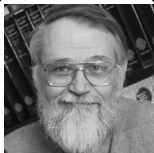
# Beispiel 2: Literaturdatenbanken

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



Computer Science

The C Programming  
Language  
digital restoration  
and typesetter

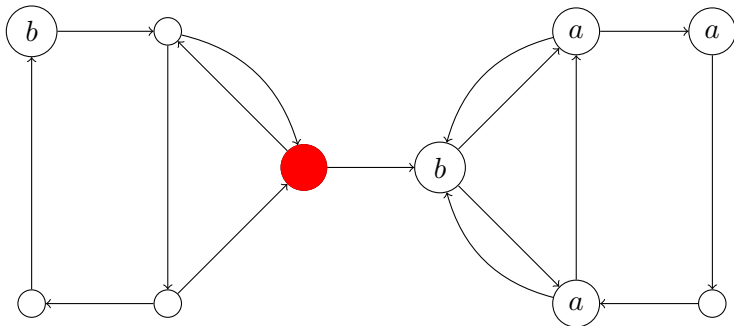


Computer Science

The Identity  
Thesis for  
Language and  
Music

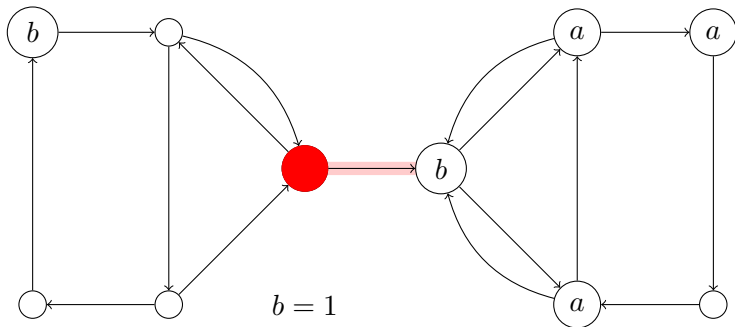


Linguistics



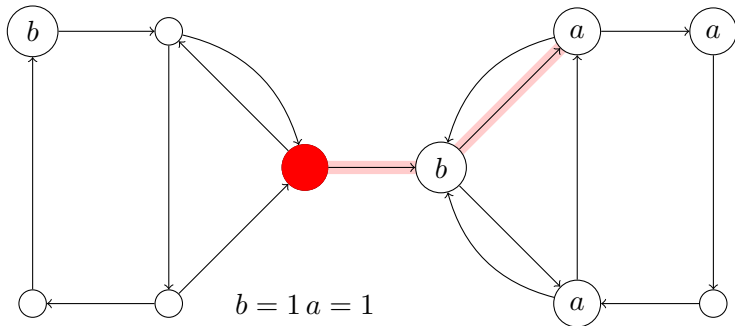
Klassifizieren des roten Knotens:

- 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



Klassifizieren des roten Knotens:

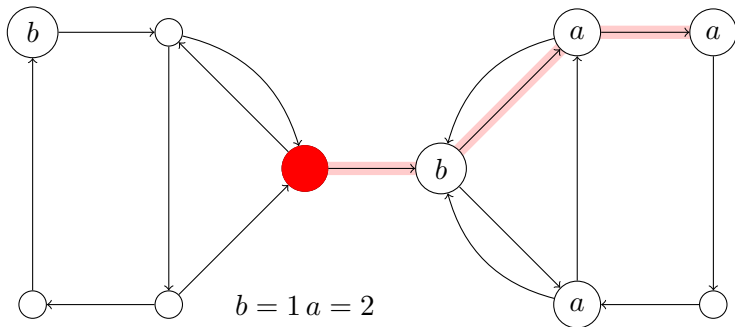
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$  Rot mit  $a$  klassifizieren



Klassifizieren des roten Knotens:

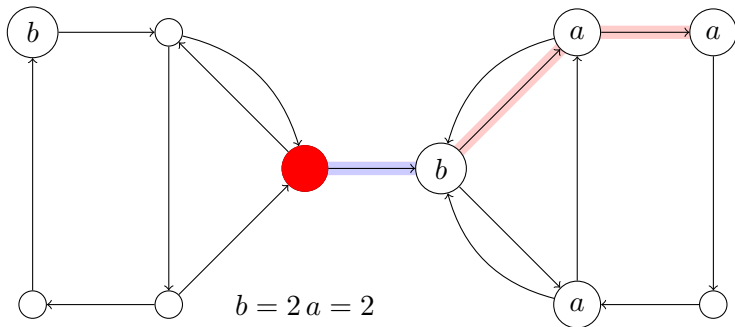
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$  Rot mit  $a$  klassifizieren





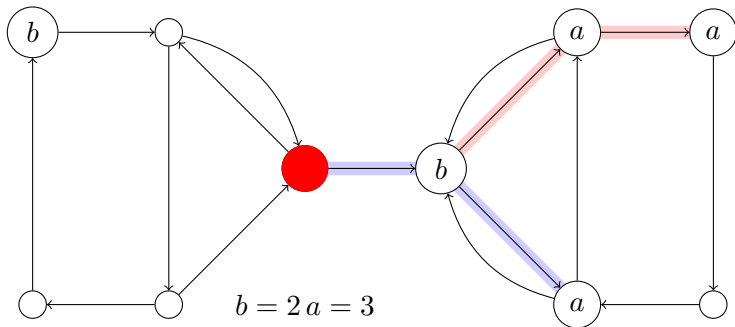
Klassifizieren des roten Knotens:

- 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



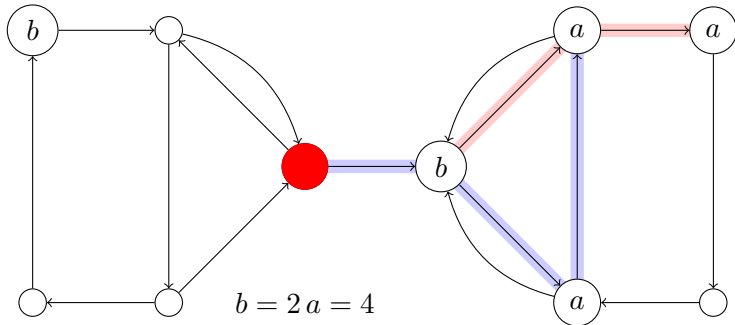
Klassifizieren des roten Knotens:

- 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



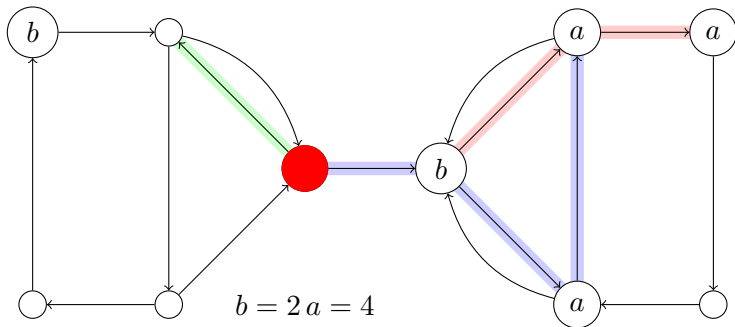
Klassifizieren des roten Knotens:

- 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



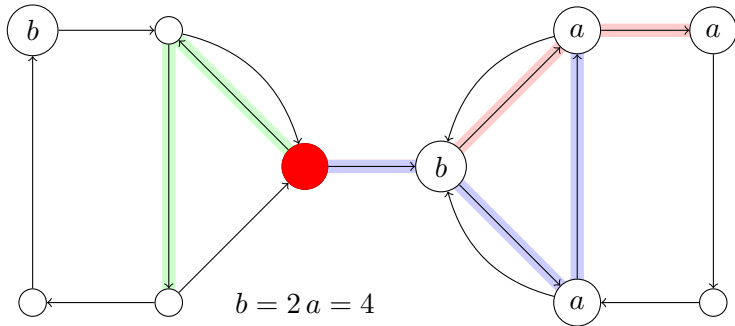
Klassifizieren des roten Knotens:

- 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



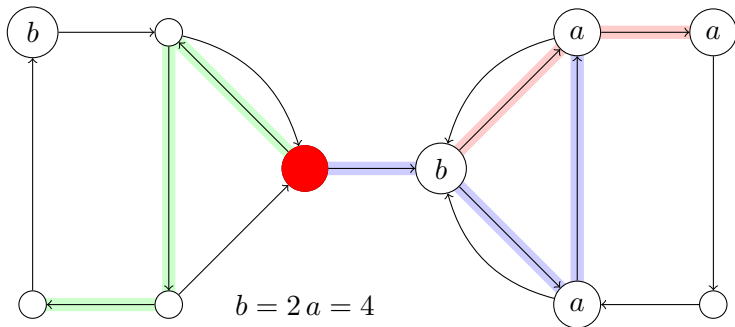
Klassifizieren des roten Knotens:

- 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



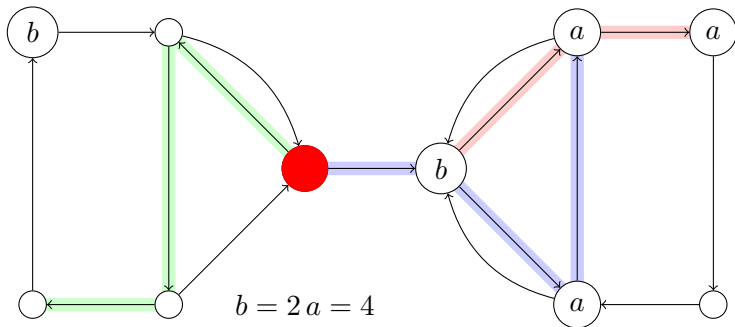
Klassifizieren des roten Knotens:

- 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



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



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

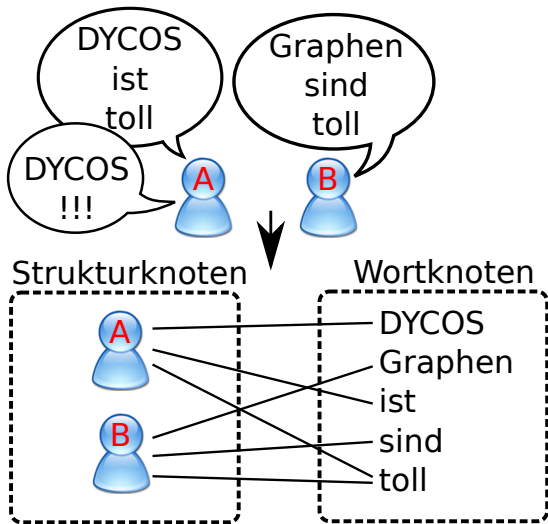
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoten
  - vice versa

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

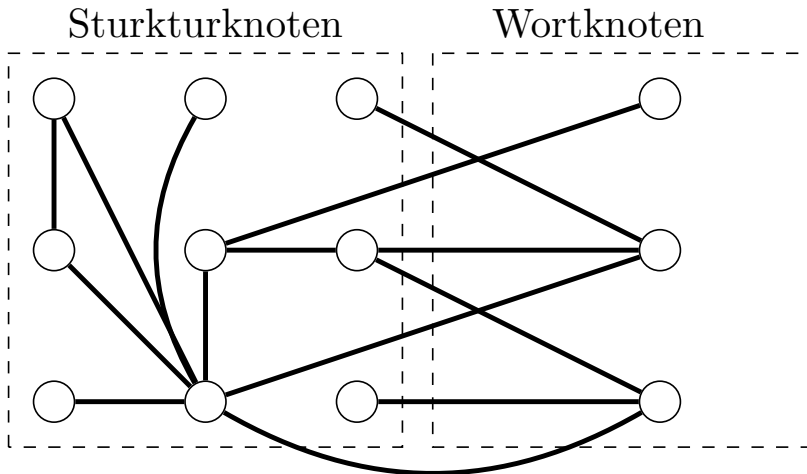
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoten
  - vice versa

- 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





- Viele Texte  $\Rightarrow$  Komplette Textanalyse nicht möglich

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$  Beschränkung des Vokabulars sinnvoll

## Idee:

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

- Viele Texte  $\Rightarrow$  Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...

$\Rightarrow$  Beschränkung des Vokabulars sinnvoll

## Idee:

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

- Viele Texte  $\Rightarrow$  Komplette Textanalyse nicht möglich

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$  Beschränkung des Vokabulars sinnvoll

## Idee:

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

- Viele Texte  $\Rightarrow$  Komplette Textanalyse nicht möglich

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$  Beschränkung des Vokabulars sinnvoll

## Idee:

- Zufällige Beispielmengende von Texten für Vokabularbildung betrachten

- Gini-Koeffizient nutzen

- Viele Texte  $\Rightarrow$  Komplette Textanalyse nicht möglich

- Füllwörter: und, oder, im, in, ...

$\Rightarrow$  Beschränkung des Vokabulars sinnvoll

## Idee:

- Zufällige Beispielmengende von Texten für Vokabularbildung betrachten

- Gini-Koeffizient nutzen

- Viele Texte  $\Rightarrow$  Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...

$\Rightarrow$  Beschränkung des Vokabulars sinnvoll

## Idee:

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

- 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

- 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



- statistisches Maß für Ungleichverteilung

- $g = \sum_i p_i^2$  mit  $p_i$  als relative Häufigkeit

- $g \in (0, 1]$

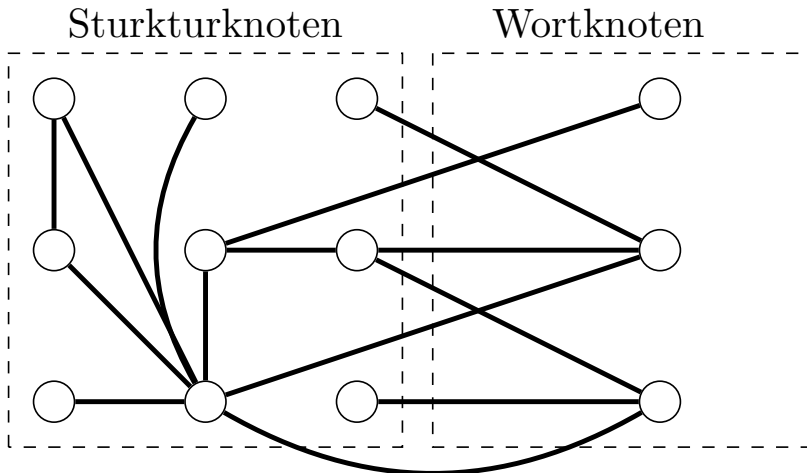
- $g$  nahe bei 1  $\Rightarrow$  Wort ist stark ungleich verteilt

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

- 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

- 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



- **Struktursprung:** von Strukturknoten  $v$  zu Strukturknoten  $v'$
- **Inhaltlicher Mehrfachsprung:** 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 Mehrfachsprung:** 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 Mehrfachsprung:** 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 Mehrfachsprung:** 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 Mehrfachsprung:** 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

# 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  $\LaTeX$  und TikZ-Quellen sind unter  
[github.com/MartinThoma/LaTeX-examples/tree/master/presentations/Datamining-Proseminar](https://github.com/MartinThoma/LaTeX-examples/tree/master/presentations/Datamining-Proseminar)  
Kurz-URL: [tinyurl.com/Info-Proseminar](https://tinyurl.com/Info-Proseminar)