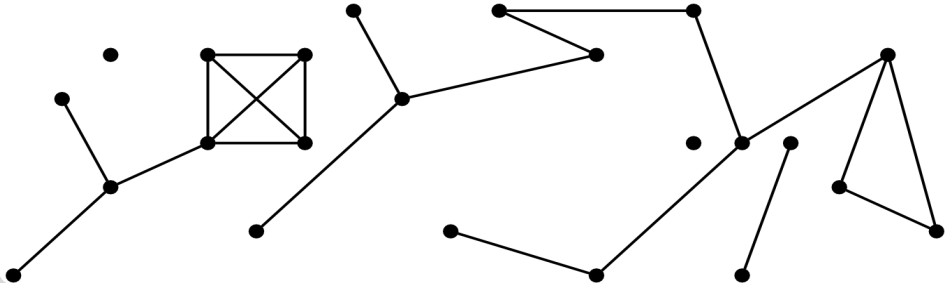
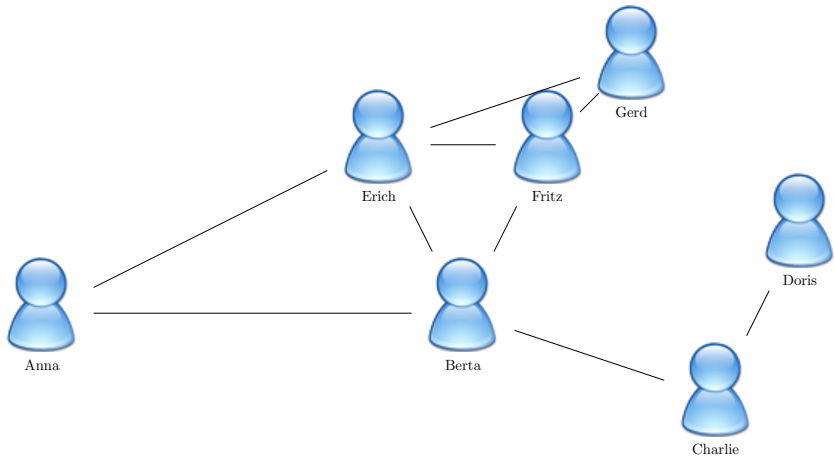


On Node Classification in Dynamic Content-based Networks

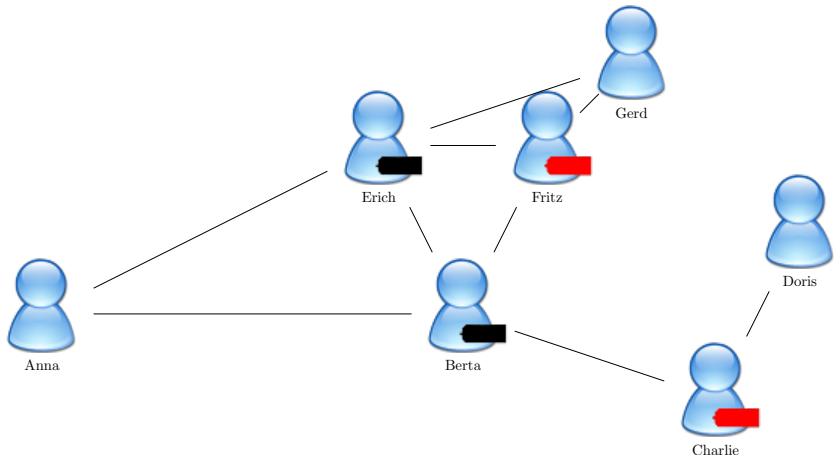
Martin Thoma | 28. Februar 2014

INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION

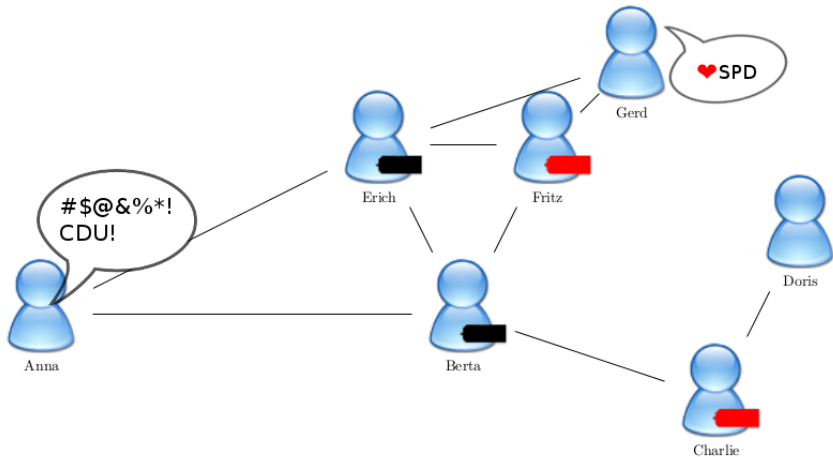


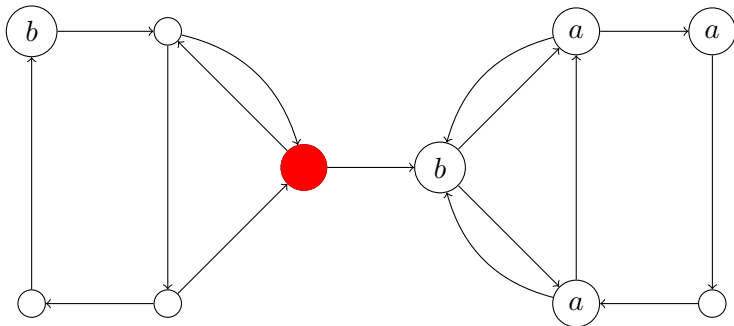


Partially labeled network



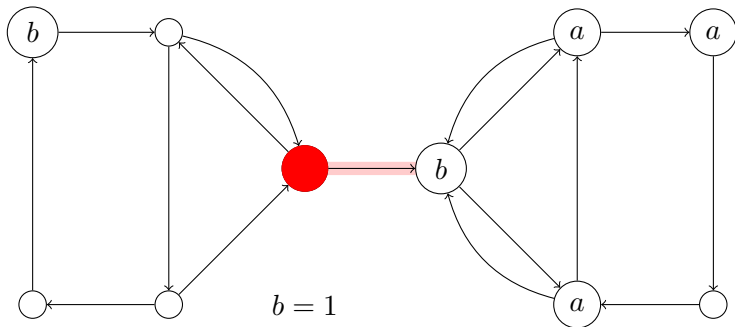
Partially labeled network with content





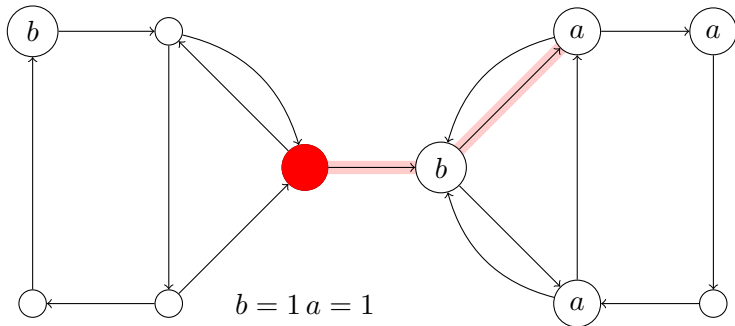
Klassifizieren des roten Knotens:

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



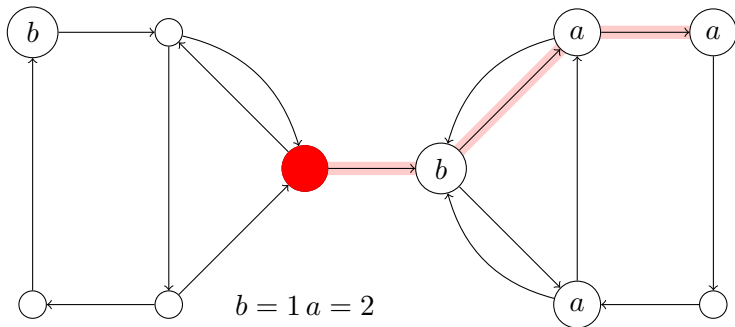
Klassifizieren des roten Knotens:

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



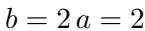
Klassifizieren des roten Knotens:

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

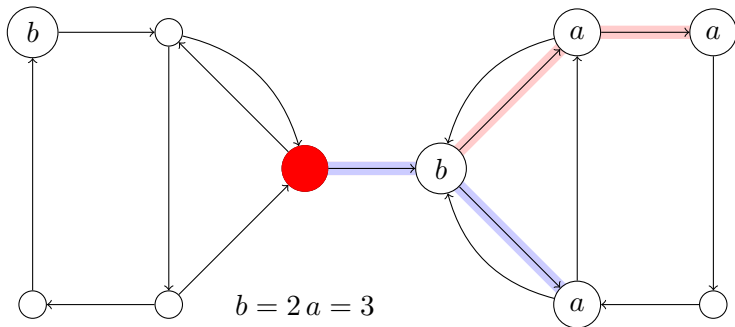


Klassifizieren des roten Knotens:

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

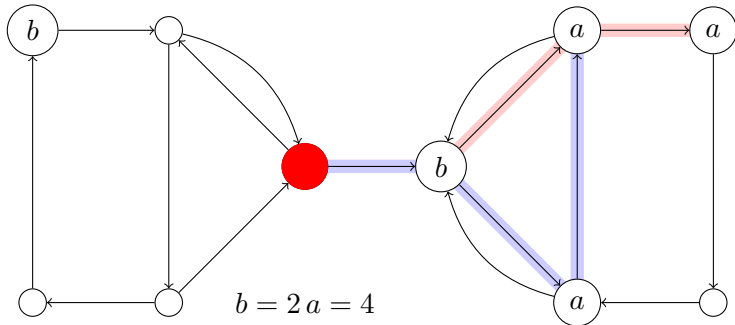


- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks
- 3 Sprünge pro Random Walk
- $4 \cdot a, 2 \cdot b \Rightarrow$ Rot mit a klassifizieren



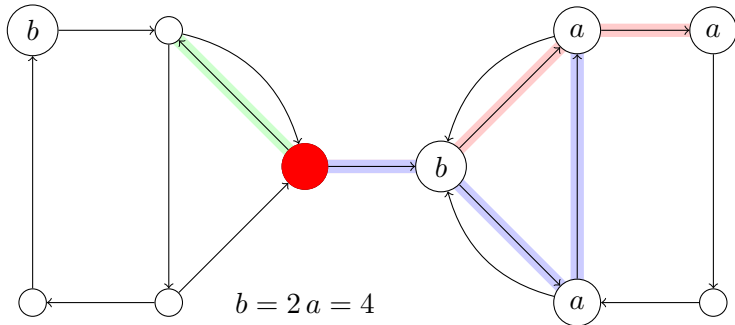
Klassifizieren des roten Knotens:

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



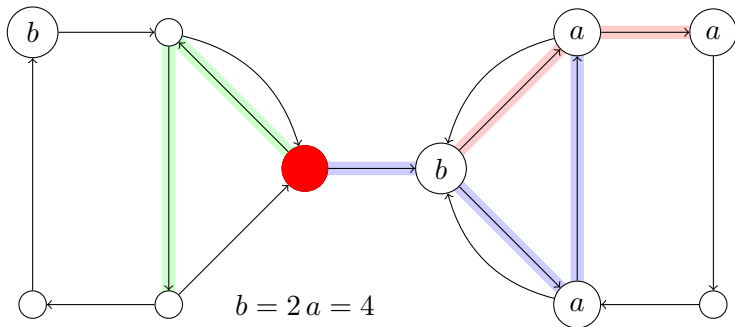
Klassifizieren des roten Knotens:

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



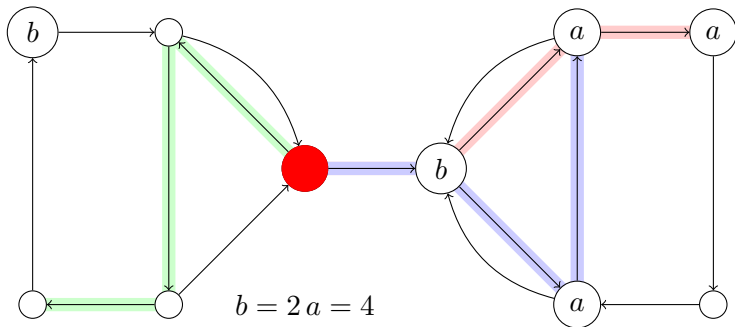
Klassifizieren des roten Knotens:

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

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

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

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

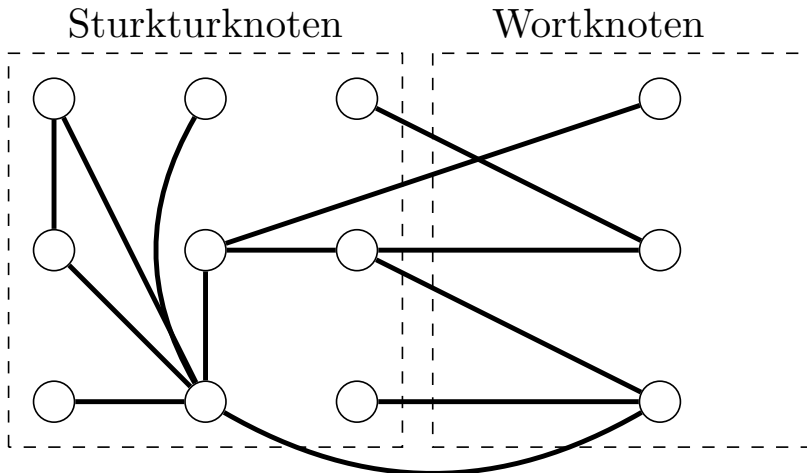
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Wünschenswert: Wenig weiterer Programmieraufwand
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa

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

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

- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Wünschenswert: Wenig weiterer Programmieraufwand
- 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

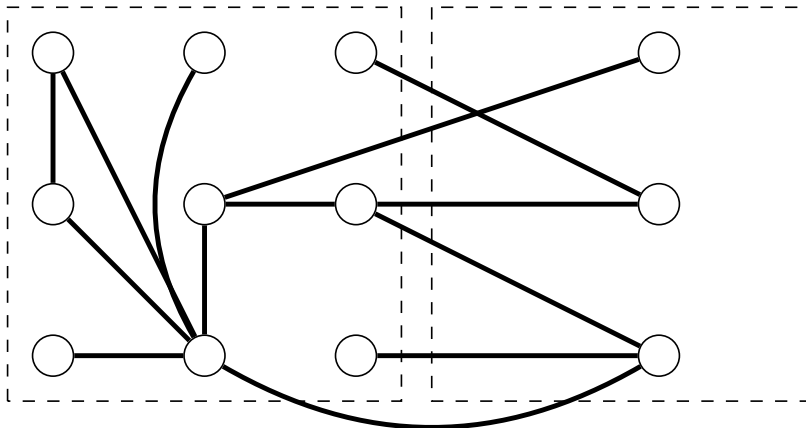
- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0, 1]$
- g nahe bei 1 \Rightarrow Wort ist stark ungleich verteilt

\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

Strukturknoten

Wortknoten



- **Struktursprung:** von Strukturnoten v zu Strukturnoten v'
- **Inhaltlicher Mehrfachsprung:** von Strukturnoten v über Wortknoten zu Strukturnoten v'

- **Struktursprung:** von Strukturnoten v zu Strukturnoten v'
- **Inhaltlicher Mehrfachsprung:** von Strukturnoten v über Wortknoten zu Strukturnoten v'

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*

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