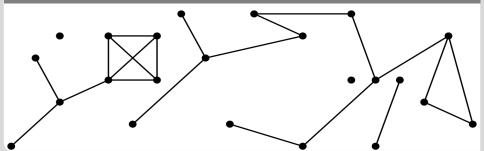


# On Node Classification in Dynamic Content-based Networks

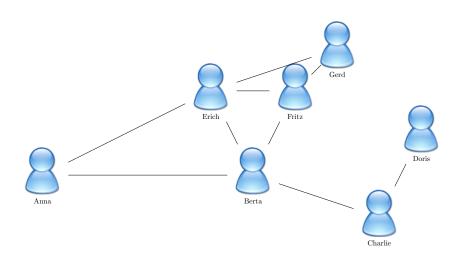
Martin Thoma | 28. Februar 2014

#### INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



## **Social Network**



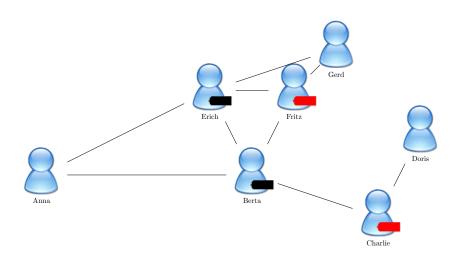


 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Ende

 ●00
 000
 00
 000
 000

# Partially labeled network

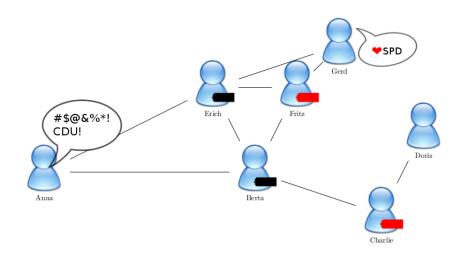




Szenario	Überblick	Vokabular	Sprungtypen	Ende
000	000	00	00	0000

# Partially labeled network with content

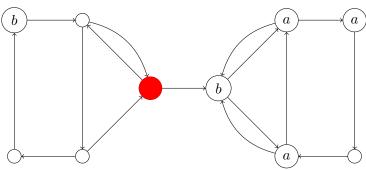




 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Ende

 00●
 000
 00
 00
 000





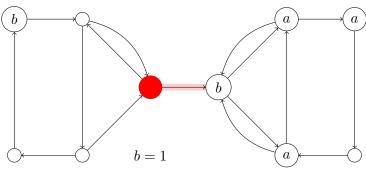
#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma - On Node	Classification in	Dynamic Content based Nature





Vakabular

#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

Sanaria

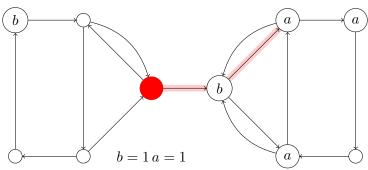
3 Sprünge pro Random Walk

Überblick

•  $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$ 

OLCHUIO	O D C I D II C II	v Ortabarar
000	●00	00
Martin Thoma - On Node C	Tassification in Dynamic Conte	nt based Network





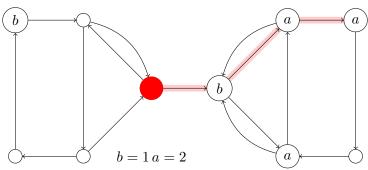
#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma -	On Node Classification in Dyna	amic Content-based Networks





#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

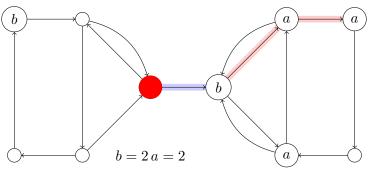
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma -	On Node Classification in Dyna	amic Content-based Networks

28. Februar 2014

Ende





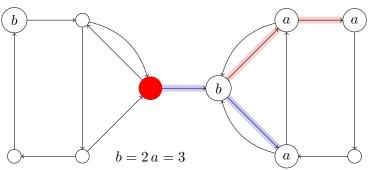
#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma -	On Node Classification in Dyna	amic Content-based Networks





#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

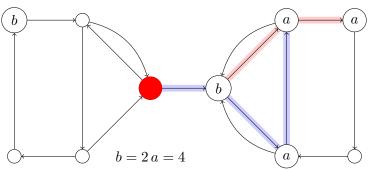
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma -	- On Node Classification in Dyna	mic Content-based Networks

28. Februar 2014

Ende





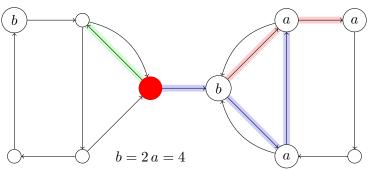
#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma -	On Node Classification in Dyna	amic Content-based Networks





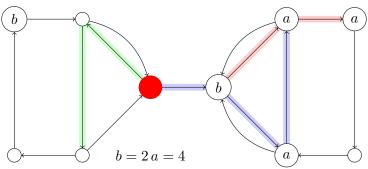
#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma -	On Node Classification in Dyna	amic Content-based Networks





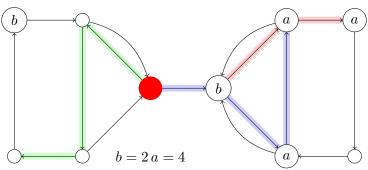
#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma -	On Node Classification in Dyna	amic Content-based Networks





#### Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks

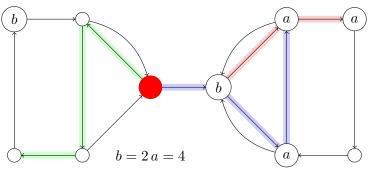
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenario	Überblick	Vokabular
000	●00	00
Martin Thoma - On Node (	Classification in Dynamic Contr	ent-based Network

28. Februar 2014

Ende





#### Klassifizieren des roten Knotens:

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

Szenano	
000	

5/15



- 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
  - lexte 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

6/15



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



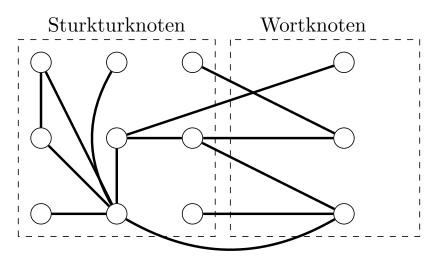
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Wünschenswert: Wenig weiterer Programmieraufwand
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoten
  - vice versa



- 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 ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

#### ldee

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

#### ldee

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

#### ldee

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

#### Idee:

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

#### Idee:

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

#### Idee:

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



- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$  mit  $p_i$  als relative Häufigkeit
- $g \in (0,1]$
- lack q nahe bei  $1 \Rightarrow \text{Wort}$  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
- $g \in (0,1]$
- lack q nahe bei  $1 \Rightarrow \text{Wort}$  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
- $g \in (0,1]$
- $\blacksquare$  g nahe bei 1  $\Rightarrow$  Wort ist stark ungleich verteilt
- $\Rightarrow$  Nehme Top-m Wörter mit höchstem Gini-Koeffizient



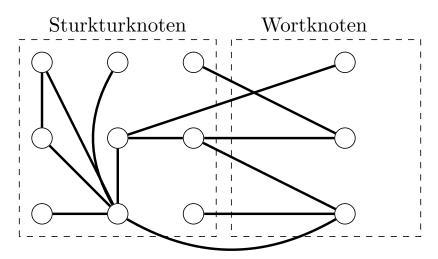
- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$  mit  $p_i$  als relative Häufigkeit
- $g \in (0,1]$
- g nahe bei  $1 \Rightarrow Wort$  ist stark ungleich verteilt
- $\Rightarrow$  Nehme Top-m Wörter mit höchstem Gini-Koeffizient



- 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

# Sprungtypen







Überblick

Vokabular

Sprungtypen

Ende 28. Februar 2014

# Inhaltlicher Mehrfachsprung



- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Mehrfachsprung: von Strukturknoten v über Wortknoten zu Strukturknoten v'

# Inhaltlicher Mehrfachsprung



- $lacktriang{lacktriang}{lacktriang}$  von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Mehrfachsprung: von Strukturknoten v über Wortknoten zu Strukturknoten v'

# 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

# Folien, LaTeXund Material



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

Szenario

Überblick

Vokabular

Sprungtypen

28. Februar 2014

Ende