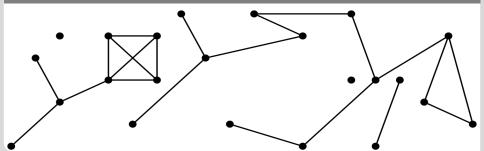


# On Node Classification in Dynamic Content-based Networks

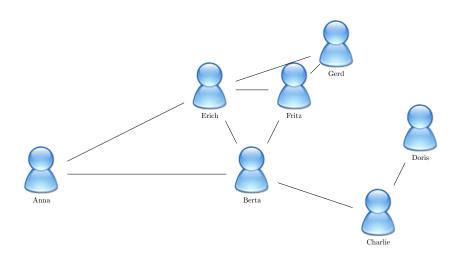
Martin Thoma | 28. Februar 2014

#### INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



## **Social Network**





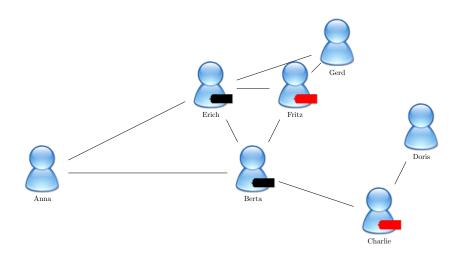
 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Zusammenfassung
 Ende

 ●000
 000
 000
 00
 0
 0

 Martin Thoma – On Node Classification in Dynamic Content-based Networks
 28. Februar 2014
 2/19

# Partially labeled network





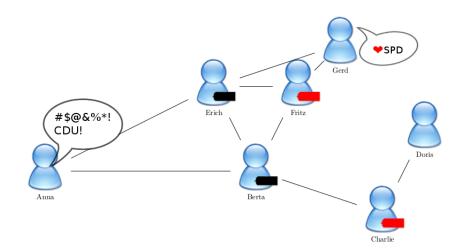
Sprungtypen

Martin Thoma - On	Node Classification in	Dynamic Content-base	d Networks
0000	0000	000	00
Szenario	Uberblick	Vokabular	Sprungtype

Ende

# Partially labeled network with content





Szenario 0000

Überblick

Vokabular

Sprungtypen

Zusammenfassung

Ende 28. Februar 2014

# 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



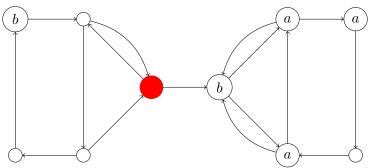
Computer Science

The Identity
Thesis for
Language and
Music



Linguistics

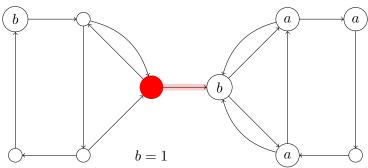




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	ıarı	0
000	00	

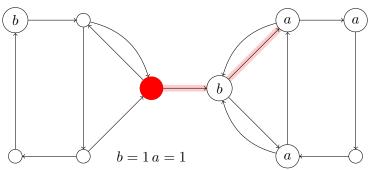




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

SZEHAH	,
0000	
Manata	Th

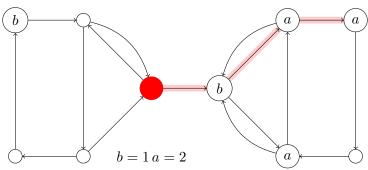




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	nari	0
000	00	

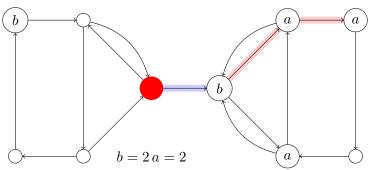




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

10	nari	Sze
	00	000

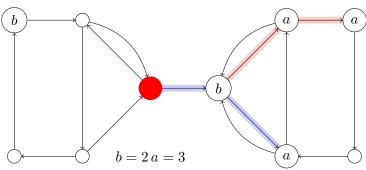




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szer	ıarı	0
000	0	
		_

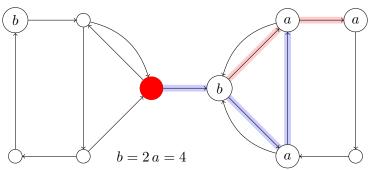




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenari	0
0000	

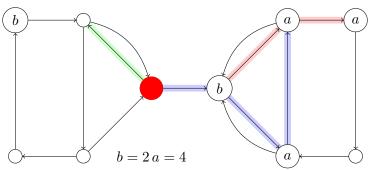




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	narı	0
000	00	

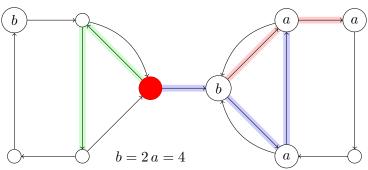




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	nari	0
000	00	

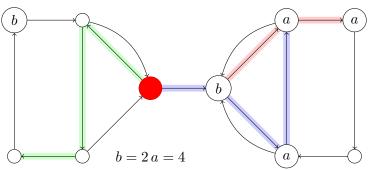




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	narı	0
000	00	





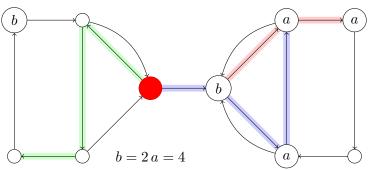
#### Klassifizieren des roten Knotens:

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

Szei	narı	0
000	00	

28. Februar 2014





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

Szen	ario	
000	0	



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



- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweiterr
  - 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 Wortknoter
  - vice versa

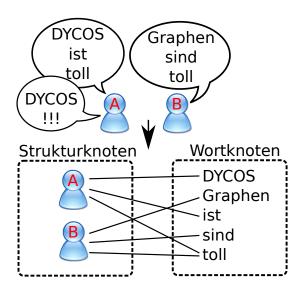


- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoten
  - vice versa



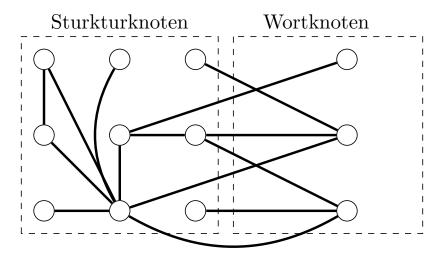
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoten
  - vice versa

Sprungtypen



# Erweiterter, semi-bipartiter Graph







Überblick 000● Vokabular

Sprungtypen

Zusammenfassung

Ende 0000



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

#### Idee:

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

28. Februar 2014



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

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



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

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



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

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



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

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



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

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



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



- 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



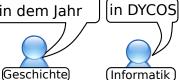
- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$  mit  $p_i$  als relative Häufigkeit
- $g \in (0,1]$
- g nahe bei  $1 \Rightarrow Wort$  ist stark ungleich verteilt
- $\Rightarrow$  Nehme Top-m Wörter mit höchstem Gini-Koeffizient







in der Schule in dem Jahr



Mathematik

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

12/19







in der Schule in dem Jahr



Mathematik

Geschichte

- $lue{}$  Vorkommen insgesamt: 5 imes
- 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"  $2 \times \Rightarrow p_2 = \frac{2}{5}$
- Gini-Koeffizient:  $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$







in der Schule in dem Jahr



Mathematik

(Geschichte)

#### Beispiel: "in"

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

12/19







in der Schule in dem Jahr



(Mathematik)

Geschichte

- Vorkommen insgesamt: 5×
  - 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"  $2 \times \Rightarrow p_2 = \frac{2}{5}$
  - Gini-Koeffizient:  $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$







in der Schule in dem Jahr



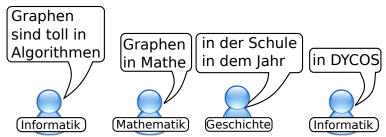
Mathematik

Geschichte

- 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"  $2 \times \Rightarrow p_2 = \frac{2}{5}$
  - Gini-Koeffizient:  $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$







- 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"  $2 \times \Rightarrow p_2 = \frac{2}{5}$

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







in der Schule in dem Jahr



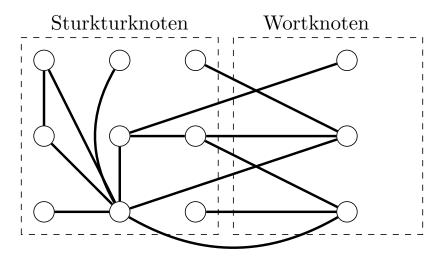
Mathematik

Geschichte

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

## Sprungtypen







Überblick

Vokabular

Sprungtypen

 ${\sf Zusammenfassung}_{\sf O}$ 

Ende 0000 13/19



- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Zweifachsprung: von Strukturknoten v über Wortknoten zu Strukturknoten  $v^\prime$ 
  - 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'

  - Nehme Top-*q*-Knoten (Anzahl der Pfade)



- ullet 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



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

28. Februar 2014



- ullet 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
- Inhaltlicher Zweifachsprung

28. Februar 2014

Vokabular

## Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

28. Februar 2014

## Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

## Danke!



# Gibt es Fragen?

Ende

#### Bildquellen



Crystal\_Clear\_app\_personal.png von Wikipedia Commons

Ende

#### 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, LeTeXund Material



Der Foliensatz und die  $\protect\operatorname{MTE}X$ und  $\protect\operatorname{Ti}kZ$ -Quellen sind unter github.com/MartinThoma/LaTeX-examples/tree/master/presentations/Datamining-Proseminar Kurz-URL: tinyurl.com/Info-Proseminar