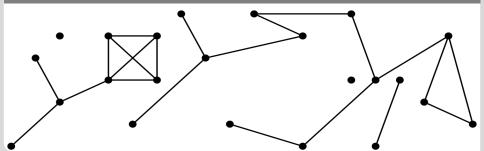


### Knotenklassifikation in dynamischen Graphen mit Texten

Martin Thoma | 25. Februar 2014

#### INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



### Zitationsdatenbanken



- Publikationen oder Autoren können Knoten sein,
- Zitate oder Mehrautorenschaftkönnen Kanten sein und
- Kategorien können Beschriftungen sein

**Problem**: Nicht alle Knoten sind beschriftet **Anwendungsideen**:

- Kategorievorschläge bei neuen Einträgen
- Korrekturvorschläge für alte Einträge

### Herausforderungen

- Große Graphen,
- Dynamische Graphen,
- Texte sollen verwendet werden

| Name | Knoten  | davon beschriftet | Kanten    | Beschriftungen |
|------|---------|-------------------|-----------|----------------|
| CORA | 19 396  | 14814             | 75 021    | 5              |
| DBLP | 806 635 | 18 999            | 4 414 135 | 5              |

#### DYCOS is

- effizient,
- einfach,
- und nutzt Struktur und Texte

| Szenano  |
|----------|
| 00000    |
| Martin T |

### Herausforderungen

- Große Graphen,
- Dynamische Graphen,
- Texte sollen verwendet werden

| Name | Knoten  | davon beschriftet | Kanten    | Beschriftungen |
|------|---------|-------------------|-----------|----------------|
| CORA | 19 396  | 14814             | 75 021    | 5              |
| DBLP | 806 635 | 18 999            | 4 414 135 | 5              |

#### **DYCOS** ist

- effizient,
- einfach,
- und nutzt Struktur und Texte

### Zitationsdatenbanken



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

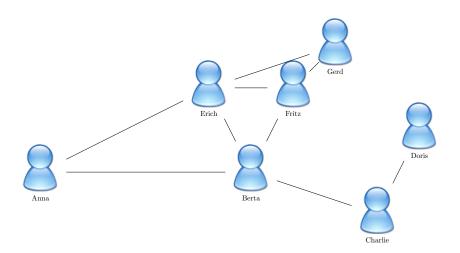
The Identity
Thesis for
Language and
Music



Linguistics

### **Social Network**





Szenario 000000 Vokabular

Sprungtypen

Evaluation

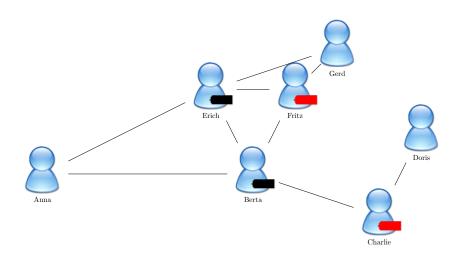
Zusammenfassung

Ende

Überblick

### Partially labeled network





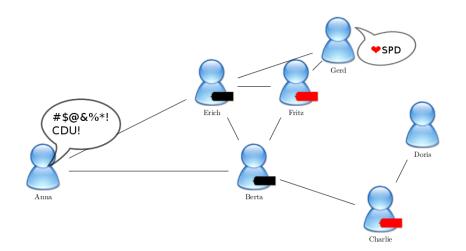
Szenario 0000•0 Vokabular 000 Sprungtypen 00 Evaluation 00 Zusammenfassung

Ende 0000 6/25

Überblick

### Partially labeled network with content





Szenario 00000●

Überblick

Vokabular 000 Sprungtypen 00 Evaluation 00

Zusammenfassung

ssung Ende

### Überblick



- Graph ist gegeben
- Knoten sind teilweise beschriftet
- Fehlende Beschriftungen sollen berechnet werden

**Idee**: Homophilie nutzen Nahe Knoten sind ähnlich

⇒ Random Walks zur Klassifizierung nutzen

### Überblick



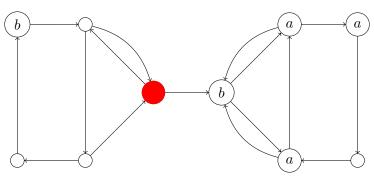
- Graph ist gegeben
- Knoten sind teilweise beschriftet
- Fehlende Beschriftungen sollen berechnet werden

Idee: Homophilie nutzen

Nahe Knoten sind ähnlich

⇒ Random Walks zur Klassifizierung nutzen

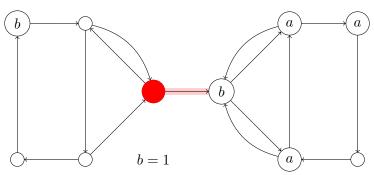




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

| Szenano        | Operblick            | VOKADUIAI         | Sprungtypen        |
|----------------|----------------------|-------------------|--------------------|
| 000000         | 00000                | 000               | 00                 |
| Martin Thoma - | - Knotenklassifikati | on in dynamischen | Graphen mit Texten |

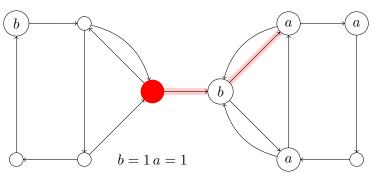




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

| Szeriario |
|-----------|
| 000000    |
| Mantin T  |

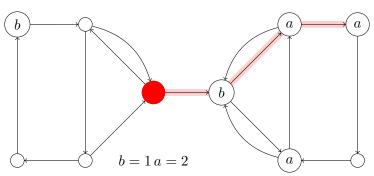




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

| N / :   | т. |
|---------|----|
| 00000   | 0  |
| Szenanc | ,  |

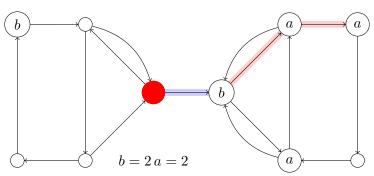




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

| Szenano   |
|-----------|
| 000000    |
| Manager T |

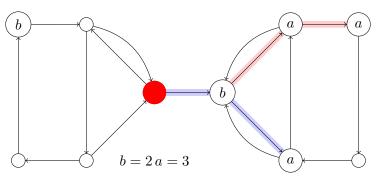




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

|    | 4    | - |
|----|------|---|
| 0  | 0000 | 0 |
| 32 | enan | J |

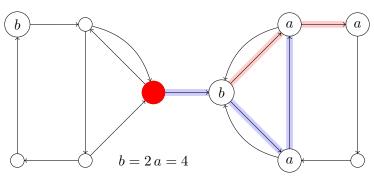




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

| 000000 |
|--------|
|        |

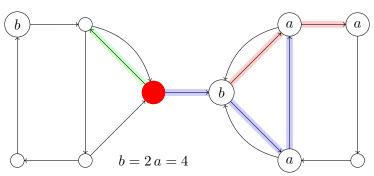




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

|    | 4    | - |
|----|------|---|
| 0  | 0000 | 0 |
| 32 | enan | J |

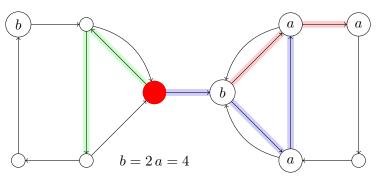




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

| Szenano |   |
|---------|---|
| 000000  |   |
| Manager | - |

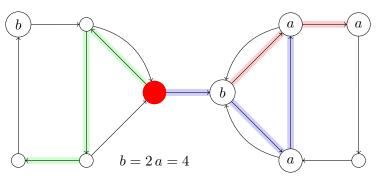




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

|   | NA-mti-  |   |
|---|----------|---|
|   | 00000    | 0 |
| • | Szemanic | , |

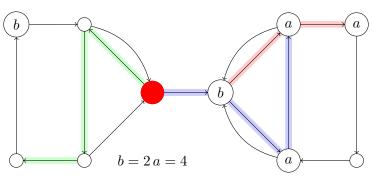




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$ ,  $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

| Szenano   |
|-----------|
| 000000    |
| Mantin Ti |





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



- 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



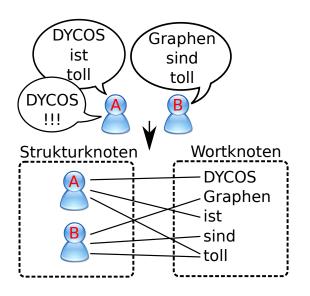
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoter
  - vice versa



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

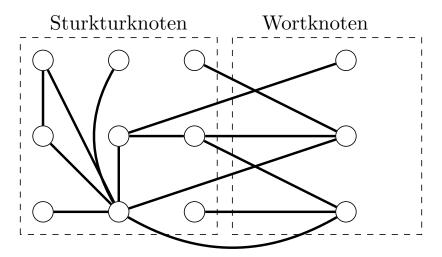


- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
  - Texte als Wortmengen
  - Strukturknoten verweisen auf Wortknoten
  - vice versa



### Erweiterter, semi-bipartiter Graph





25. Februar 2014

Ende



- 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



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

#### Idee:

- Zufällige Beispielmenge von Texten für Vokabularbildung

Ende



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

#### Idee:

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten



- 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
- Hier:  $g \in (0,1]$
- $lue{g}$  nahe bei  $1 \Rightarrow \text{Wort}$  ist stark ungleich verteilt
- $\Rightarrow$  Nehme Top-m Wörter mit höchstem Gini-Koeffizien

### Gini-Koeffizient



- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$  mit  $p_i$  als relative Häufigkeit
- Hier:  $g \in (0,1]$
- $lue{g}$  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
- Hier:  $g \in (0,1]$
- $lue{g}$  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
- Hier:  $g \in (0,1]$
- g nahe bei  $1 \Rightarrow \mathsf{Wort}$  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
- Hier:  $g \in (0,1]$
- 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





Beispiel: "in"

■ 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



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







in der Schule in dem Jahr





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









in der Schule in dem Jahr



(Mathematik)

Geschichte

Beispiel: "in"

■ Vorkommen insgesamt: 5×

• Vorkommen in "Informatik"  $2 imes \Rightarrow p_1=rac{2}{5}$ 

• 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:  $5 \times$
- 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

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



Sprungtypen







in der Schule in dem Jahr



hematik) (

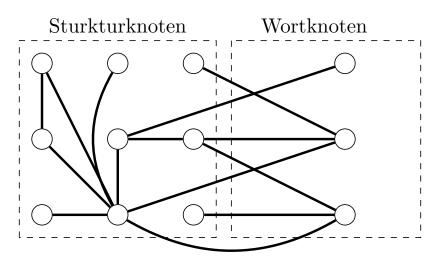
Geschichte

Beispiel: "in"

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

## Sprungtypen







Überblick 00000 Vokabular 000 Sprungtypen ●○ Evaluation 00 Zusammenfassung

ung Ende



- 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



- 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'
  - 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'
  - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
  - Nehme Top-q-Knoten (Anzahl der Pfade)
  - Wähle zufällig einen davor



- 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



| Name | Knoten  | davon beschriftet | Kanten    | Beschriftungen |
|------|---------|-------------------|-----------|----------------|
| CORA | 19 396  | 14814             | 75 021    | 5              |
| DBLP | 806 635 | 18 999            | 4 414 135 | 5              |

Martin Thoma - Knotenklassifikation in dynamischen Graphen mit Texten



#### Performance:

- Klassifizierung aller Knoten
- Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
- DBLP: < 25 s
- $\bullet$  CORA: < 5 s

### Klassifikationsgüte:

- CORA: 82% 84%
- DBLP: 61% 66%

Martin Thoma - Knotenklassifikation in dynamischen Graphen mit Texten



#### Performance:

- Klassifizierung aller Knoten
- Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
- DBLP: < 25 s
- CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% 84%
  - DBLP: 61% 66%



- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - $\bullet$  CORA: < 5 s
- Klassifikationsgüte:
  - CORA: 82% 84%
  - DBLP: 61% 66%



- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s</p>
- Klassifikationsgüte:
  - CORA: 82% 84%
  - DBLP: 61% 66%



#### Performance:

- Klassifizierung aller Knoten
- Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
- DBLP: < 25 s</li>
- CORA: < 5s</p>
- Klassifikationsgüte:



- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s</li>
  - CORA: < 5s</p>
- Klassifikationsgüte:



- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s</p>
- Klassifikationsgüte:
  - CORA: 82% 84%
  - DBLP: 61% 66%



- Performance:
  - Klassifizierung aller Knoten
  - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
  - DBLP: < 25 s
  - CORA: < 5 s</p>
- Klassifikationsgüte:
  - CORA: 82% 84%
  - DBLP: 61% 66%

## Wichtige Ideen



- Random Walk
- Inhaltlicher Zweifachsprung

## Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

## Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Martin Thoma - Knotenklassifikation in dynamischen Graphen mit Texten



- DYCOS ist nur von der lokalen Situation abhängig

Überblick



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



- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach
- ⇒ Der Graph darf dynamisch sein; DYCOS funktioniert dennoch

Martin Thoma - Knotenklassifikation in dynamischen Graphen mit Texten



- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach
- ⇒ Der Graph darf dynamisch sein; DYCOS funktioniert dennoch

Martin Thoma - Knotenklassifikation in dynamischen Graphen mit Texten

## 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.
- 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 LATEX und TikZ-Quellen sind unter github.com/MartinThoma/LaTeX-examples/tree/master/presentations/Datamining-Proseminar Kurz-URL: tinyurl.com/Info-Proseminar