Beyond Friendships and Followers: The Wikipedia Social Network

Johanna Geiß, Andreas Spitz, Michael Gertz
Database Systems Research Group
Institute of Computer Science,
Heidelberg University
{johanna.geiss, spitz, gertz}@informatik.uni-heidelberg.de

ASONAM 28.08.2016
Explicit networks

Latently embedded
Explicit networks

Latently embedded

Facebook

Twitter

GitHub
Explicit networks

- persons/relationships are explicitly given
  - by user
  - by user interactions

Latently embedded

GitHub
Explicit networks

- persons/relationships are explicitly given
  - by user
  - by user interactions

Latently embedded

- semi-structured data repositories
  - email archives
  - blogs
  - movie databases ...

- networks are latently embedded
- domain specific extraction methods
Implicit person-centric networks

• New approach for person-centric network extraction from unstructured data (text)
• Text from English Wikipedia (not editor information)
  • large network
  • deals with mostly well-known persons
  • can be combined with other sources
• approach is based on co-occurrences of persons
Outline

• Preliminary considerations
• Extracting person mentions
• The Wikipedia Social Network
• Community evaluation
• Conclusion and ongoing work
Persons co-occur in the same article → persons belong to same context/theme

*Star Wars: The Force Awakens*

From Wikipedia, the free encyclopedia

*Star Wars: The Force Awakens* (also known as *Star Wars Episode VII: The Force Awakens*) is an upcoming American epic space opera film directed by J. J. Abrams. The seventh installment in the episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driver, Oscar Isaac, Andy Serkis, Domhnall Gleeson, and Max von Sydow, with Harrison Ford, Carrie Fisher, Mark Hamill, Anthony Daniels, Peter Mayhew, and Kenny Baker reprising their roles from previous *Star Wars* films. The story is set approximately 30 years after the events of *Return of the Jedi* (1983).

*The Force Awakens* will be the first film in the planned third *Star Wars* trilogy announced after The Walt Disney Company's acquisition of Lucasfilm in October 2012. It is produced by Abrams, his long-time collaborator Bryan Burk, and Lucasfilm president Kathleen Kennedy. Along with directing and producing, Abrams also co-wrote with Lawrence Kasdan, who co-wrote the original trilogy films *The Empire Strikes Back* and *Return of the Jedi*. Abrams and Kasdan
Person co-occurrence

Persons co-occur in the same article → persons belong to same context/theme

Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia

*Star Wars: The Force Awakens* (also known as *Star Wars: Episode VII: The Force Awakens*) is an upcoming American epic space opera film directed by J. J. Abrams. It is the eleventh film in the episodic *Star Wars* film series, and the first film in the planned *Star Wars* sequel trilogy, which will serve as the third *Star Wars* trilogy announced after *The Walt Disney Company*’s acquisition of *Lucasfilm* in October 2012. It is produced by Abrams, his long-time collaborator Bryan Burk, and *Lucasfilm* president Kathleen Kennedy. Along with directing and producing, Abrams also co-wrote with Lawrence Kasdan, who co-wrote the original trilogy films *The Empire Strikes Back* and *Return of the Jedi*. Abrams and Kasdan

*cast members*

Person co-occurrence

Persons co-occur in the same article → persons belong to same context/theme

Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia

*Star Wars: The Force Awakens* (also known as *Star Wars: Episode VII: The Force Awakens*) is an upcoming American epic space opera film directed by J. J. Abrams. It is the episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driver, Andy Serkis, Gwendoline Christie, Christopher Walken, Oscar Isaac, Lupita Nyong'o, Domhnall Gleeson, and Max von Sydow, with Harrison Ford, Carrie Fisher, Mark Hamill, Anthony Daniels, Peter Mayhew, and Kenny Baker reprising their roles from previous *Star Wars* films. The story is set approximately 30 years after the events of *Return of the Jedi* (1983).

The Force Awakens will be the first film of the *Star Wars* sequel trilogy announced after The Walt Disney Company's acquisition of Lucasfilm in October 2012. It is the first film directed by Abrams, his long-time collaborator Bryan Burk, and Lucasfilm president Kathleen Kennedy. Along with directing and producing, Abrams also co-wrote with Lawrence Kasdan, who co-wrote the original trilogy films *The Empire Strikes Back* and *Return of the Jedi*. Abrams and Kasdan

ASONAM 2015  Johanna Geiß - Beyond Friendships and Followers  5
Person co-occurrence

Persons co-occur in the same article → persons belong to same context/theme

Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia

*Star Wars: The Force Awakens* (also known as *Star Wars: Episode VII: The Force Awakens*) is an upcoming American epic space opera film directed by J. J. Abrams. The episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driver, with Harrison Ford, Carrie Fisher, Mark Hamill, Anthony Daniels, Peter Mayhew, and Kenny Baker reprising their roles from previous *Star Wars* films. The story is set approximately 30 years after the events of *Return of the Jedi* (1983).

*The Force Awakens* will be the first film of the *Star Wars* film franchise's *Star Wars: The Force Awakens* trilogy announced after *The Walt Disney Company*’s acquisition of Lucasfilm in 2012. Abrams, his long-time collaborator Bryan Burk, and Lucasfilm president Kathleen Kennedy. Along with directing and producing, Abrams also co-wrote with Lawrence Kasdan, who co-wrote the original *Star Wars* trilogy films *The Empire Strikes Back* and *Return of the Jedi*.
Distance of person mentions

Distance between person mentions in article → measure for the strength of the relationship

**Star Wars: The Force Awakens**

From Wikipedia, the free encyclopedia

*Star Wars: The Force Awakens* (also known as *Star Wars: Episode VII: The Force Awakens*) is an upcoming American epic space opera film directed by J. J. Abrams and released on December 18, 2015 as the episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driver, and Gwendoline Christie, with Harrison Ford, Carrie Fisher, Mark Hamill, Anthony Daniels, Peter Mayhew, and Kenny Baker reprising their roles from previous *Star Wars* films. The story is set approximately 30 years after the events of *Return of the Jedi* (1983).

The Force Awakens will be the first film in the new *Star Wars* trilogy announced after *The Walt Disney Company*’s acquisition of *Lucasfilm* in October 2012. Along with directing and producing, Abrams also co-wrote with Lawrence Kasdan, who co-wrote the original trilogy films *The Empire Strikes Back* and *Return of the Jedi*, and with Bryan Burk, his long-time collaborator.

**cast members**
- John Boyega
- Daisy Ridley
- Adam Driver
- Harrison Ford
- Carrie Fisher
- Mark Hamill
- Anthony Daniels
- Peter Mayhew
- Kenny Baker

**producers**
- J. J. Abrams
- Kathleen Kennedy

**writers**
- Lawrence Kasdan
- Bryan Burk
Frequency of co-occurrence

Frequency of co-occurrences throughout all Wikipedia articles

→ measure for the strength of the relationship

• persons that co-occur frequently
  → stronger relationship

• persons with only a few co-occurrences
  → weaker relationship
Outline

• Introduction

• Extracting person mentions

• The Wikipedia Social Network

• Community evaluation

• Conclusion and ongoing work
Extracting person mentions

1. Extracting person information from Wikidata and Wikipedia

2. Identifying person mentions in a Wikipedia article
   - Interwiki Links (IWLs)
   - String search
Extracting person information - Wikidata

- **Wikidata**: free, collaboratively edited, multilingual database
- 16.8 M items representing real life concepts

<table>
<thead>
<tr>
<th>identifier</th>
<th>Q81328</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>Harrison Ford (en), Харрисон Форд (ru)</td>
</tr>
<tr>
<td>description</td>
<td>American film actor and producer</td>
</tr>
<tr>
<td>alias</td>
<td>Harisson Ford (fr)</td>
</tr>
<tr>
<td>statement</td>
<td><strong>Instance of: human</strong>, date of birth: 13 July 1942</td>
</tr>
<tr>
<td>link</td>
<td>Harrison Ford</td>
</tr>
</tbody>
</table>

- 1.2 M persons with link to EN Wikipedia
Extracting person information - Wikipedia

• 5.29 M articles (as of January 12, 2015)

• Wikipedia categories
  • categories for Harrison Ford:
    • 1942 births  • 20th-century American male actors
    • Living people  • Survivors of aviation accidents or incidents

• no single category for humans

• <year> births and <year> deaths

• 1.05 M articles about persons in EN Wikipedia
Identifying person mentions – IWL (1)

Interwiki-Links (IWL)

*Star Wars: The Force Awakens*

From Wikipedia, the free encyclopedia

*Star Wars: The Force Awakens* (also known as *Star Wars Episode VII: The Force Awakens*) is an upcoming American epic space opera film directed by J. J. Abrams. The seventh installment in the episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driver, Oscar Isaac, Andy Serkis, Domhnall Gleeson, and Max von Sydow, with Harrison Ford, Carrie Fisher, Mark Hamill, Anthony Daniels, Peter Mayhew, and Kenny Baker reprising their roles from previous *Star Wars*
Identifying person mentions – IWL (1)

Interwiki-Links (IWL)

**Star Wars: The Force Awakens**

*From Wikipedia, the free encyclopedia*

*Star Wars: The Force Awakens* (also known as *Star Wars Episode VII: The Force Awakens*) is an upcoming American epic space opera film directed by J. J. Abrams. The seventh installment in the episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driver, Oscar Isaac, Andy Serkis, Domhnall Gleeson, and Max von Sydow, with Harrison Ford, Carrie Fisher, Mark Hamill, Anthony Daniels, Peter Mayhew, and Kenny Baker reprising their roles from previous *Star Wars*. 
Identifying person mentions – IWL (1)

Interwiki-Links (IWL)

[[Harrison Ford]]
Identifying person mentions – IWL (1)

Interwiki-Links (IWL)

Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia

Star Wars: The Force Awakens (also known as Star Wars Episode VII: The Force Awakens) is a 2015 American epic space opera film

[[Harrison Ford]]

[[link target | surface]]

e.g. [[Harrison Ford | Ford]]
Identifying person mentions – IWL (1)

Interwiki-Links (IWL)

### Introduction

Extraction

WSN

Evaluation

Conclusion

---

**Star Wars: The Force Awakens** (also known as *Star Wars: Episode VII: The Force Awakens*) is an upcoming American epic science fiction film directed by J. J. Abrams. The seventh installment in the *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driver, Harrison Ford, Carrie Fisher, Mark Hamill, Anthony Daniels, Peter Mayhew, and Kenny Baker reprising their roles from previous films.

For the unrelated silent film actor, see Harrison Ford (silent film actor).

Harrison Ford (born July 13, 1942) is an American actor and film producer. He gained worldwide fame for his starring roles as Han Solo in the original *Star Wars* epic space opera trilogy and the title character of the *Indiana Jones* film series. Ford is also known for his roles as Rick Deckard in the 1982 neo-noir dystopian science fiction film *Blade Runner*, John Book in the 1985 thriller *Witness*, and Jack Ryan in the 1992 action-suspense film *Patriot Games* and the 1994 spy action thriller film *Clear and Present Danger*.

His career has spanned six decades and includes roles in several Hollywood blockbusters; including the epic war film *Apocalypse Now* (1979), the legal drama *Presumed Innocent* (1990), the action film *The Fugitive* (1993), the science fiction film *Air Force One* (1997), and the biographical drama *Chasing Amy* (1997). Ford was a star of the 1970s film ensemble star known as the Brat Pack. He has received Academy Award nominations for his performances in *American Graffiti* (1973), *The Boys in the Band* (1970), *Hoosiers* (1986), and *The kärtan* (1985), and has won receive a nomination for the Academy Award for Best Supporting Actress for his role in *An Officer and a Gentleman* (1982).

---

e.g. [[Harrison Ford | Ford]]
IWL refers to person if

1. Link target = link entry in Wikidata

[[Harrison Ford |Ford]]

- identifier: Q81328
- label: Harrison Ford (en)
- description: American film actor and producer
- statements: Instance of: human
- link to EN Wikipedia: Harrison Ford

2. Page referred to has category `<year> births` or `<year> deaths` (only 0.01%)
Identifying person mentions - string match

- many person mentions outside of IWLs

*Star Wars: The Force Awakens*

From Wikipedia, the free encyclopedia

[...]

*The Force Awakens* will be the first film in the planned third *Star Wars* trilogy announced after *The Walt Disney Company*’s acquisition of *Lucasfilm* in October 2012. It is produced by Abrams, his long-time collaborator Bryan Burk, and Lucasfilm president Kathleen Kennedy. Along with directing and producing, Abrams also co-wrote with Lawrence Kasdan, who co-wrote the original trilogy films *The Empire Strikes Back* and *Return of the Jedi*. Abrams and Kasdan

- search for surface forms and link target of IWLs on that page
Identifying person mentions - string match

• many person mentions outside of IWLs

Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia

[...]
The Force Awakens will be the first film in the planned third Star Wars trilogy announced after The Walt Disney Company’s acquisition of Lucasfilm in October 2012. It is produced by Abrams, his long-time collaborator Bryan Burk, and Lucasfilm president Kathleen Kennedy. Along with directing and producing, Abrams also co-wrote with Lawrence Kasdan, who co-wrote the original trilogy films The Empire Strikes Back and Return of the Jedi. Abrams and Kasdan

• search for surface forms and link target of IWLs on that page
Identifying person mentions - string match

- many person mentions outside of IWLs

![Star Wars: The Force Awakens Wikipedia Link](en.wikipedia.org/wiki/Star_Wars:_The_Force_Awakens)

- search for surface forms and link target of IWLs on that page
  - [[J. J. Abrams| Abrams]] -> search for “J. J. Abrams” and “Abrams”
Identifying person mentions - statistics

- 13 M person mentions (10.4 M via IWLs)
- of $\approx$ 840,000 different persons
- in 1.8 M Wikipedia articles

TOP 5:
1. Jesus (14) 4. Muhammad (27)
2. Napoleon (3) 5. William Shakespeare (4)
3. Barack Obama (1)

- majority of articles contain up to three person mentions
Outline

• Introduction

• Extracting person mentions

• The Wikipedia Social Network

• Community evaluation

• Conclusion and ongoing work
### Network construction - bipartite graph

- List of persons per sentence in document

<table>
<thead>
<tr>
<th><em>Document</em></th>
<th><em>Sentence</em></th>
<th><em>Persons</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>d1</em></td>
<td>s0</td>
<td>p1, p2</td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td>p1</td>
</tr>
<tr>
<td><em>d2</em></td>
<td>s1</td>
<td>p2</td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td>p3</td>
</tr>
<tr>
<td><em>d3</em></td>
<td>s2</td>
<td>p2</td>
</tr>
<tr>
<td></td>
<td>s12</td>
<td>p3</td>
</tr>
</tbody>
</table>
Network construction - bipartite graph

- List of persons per sentence in document
- Bipartite graph of documents and persons

<table>
<thead>
<tr>
<th>Document</th>
<th>Sentence</th>
<th>Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>s0</td>
<td>p1, p2</td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td>p1</td>
</tr>
<tr>
<td>d2</td>
<td>s1</td>
<td>p2</td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td>p3</td>
</tr>
<tr>
<td>d3</td>
<td>s2</td>
<td>p2</td>
</tr>
<tr>
<td></td>
<td>s12</td>
<td>p3</td>
</tr>
</tbody>
</table>
Network construction - multi graph

- project bipartite graph onto set of persons
- each co-occurrence induces one edge
  - 309 M edges between 799,181 persons
Network construction – edge weights

- \( d(v, w, i) = \) number of sentences between mentions (0 for same sentence)
- Weight function (decaying distance):

\[
\varphi(v, w, i) = \exp\left(\frac{-d(v, w, i)}{2}\right)
\]
Network construction – edge weights

• $d(v,w,i) = \text{number of sentences between mentions (0 for same sentence)}$

• Weight function (decaying distance):

$$\varphi(v, w, i) = \exp\left(\frac{-d(v, w, i)}{2}\right)$$

<table>
<thead>
<tr>
<th>v,w,i</th>
<th>d</th>
<th>$\varphi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1, p2, d1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>p1, p2, d1</td>
<td>3</td>
<td>0.22</td>
</tr>
<tr>
<td>p2, p3, d2</td>
<td>1</td>
<td>0.61</td>
</tr>
<tr>
<td>p2, p3, d3</td>
<td>10</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Network construction – edge weights

- $d(v,w,i) =$ number of sentences between mentions (0 for same sentence)
- Weight function (decaying distance):
  \[
  \varphi(v, w, i) = \exp\left(\frac{-d(v, w, i)}{2}\right)
  \]

<table>
<thead>
<tr>
<th>$v,w,i$</th>
<th>$d$</th>
<th>$\varphi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1, p2, d1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>p1, p2, d1</td>
<td>3</td>
<td>0.22</td>
</tr>
<tr>
<td>p2, p3, d2</td>
<td>1</td>
<td>0.61</td>
</tr>
<tr>
<td>p2, p3, d3</td>
<td>10</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Network construction – edge aggregation

• Node-edge incidence matrix $M$

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>e2</td>
<td>0.22</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>e3</td>
<td>0</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>e4</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Network construction – edge aggregation

- Node-edge incidence matrix $\mathbf{M}$
- Node-edge incidence vector

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>e2</td>
<td>0.22</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>e3</td>
<td>0</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>e4</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Network construction – edge aggregation

- Node-edge incidence matrix $M$
- Node-edge incidence vector
- dicos $\rightarrow$ cosine similarity of incidence vectors

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>e2</td>
<td>0.22</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>e3</td>
<td>0</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>e4</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Network construction – edge aggregation

- Node-edge incidence matrix $M$
- Node-edge incidence vector
- $\text{dicos} \rightarrow \text{cosine similarity of incidence vectors}$

<table>
<thead>
<tr>
<th>$e_1$</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$e_2$</td>
<td>0.22</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>$e_3$</td>
<td>0</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>$e_4$</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Graph representation:

- $p_1$ connected to $p_2$ with weight 1
- $p_2$ connected to $p_3$ with weight 0.22
- $p_2$ connected to $p_3$ with weight 0.61
- $p_1$ connected to $p_3$ with weight 0.01
### Network construction – edge aggregation

- Node-edge incidence matrix $\mathbf{M}$
- Node-edge incidence vector
- $\text{dicos} \rightarrow$ cosine similarity of incidence vectors

<table>
<thead>
<tr>
<th>$\mathbf{M}$</th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>e2</td>
<td>0.22</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>e3</td>
<td>0</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>e4</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

```
M                p1   p2   p3
e1           1    1    0
e2           0.22 0.22 0
e3           0    0.61 0.61
e4           0    0.01 0.01
```

Graph representation:
- Node $p1$ connects with $p2$ and $p3$ with edge weights 0.86 and 0.51, respectively.
Network properties - threshold

- **67 M edges** and roughly **800,000 nodes**
- 99.8% of the nodes are connected in one giant component
- threshold for edges
  - $\text{dicos}(v,w) = 0.0019$
  - high clustering coefficient
  - high assortativity by degree
  - number of edges and component size
Outline

• Introduction
• Extracting Person Mentions
• The Wikipedia Social Network
• Community Evaluation
• Conclusion and Ongoing Work
Evaluation goal

- Did we choose the right threshold?
- Does the network represent meaningful relationships between persons?
Evaluation strategy (1)

• Extract communities from the network
  • Stabilized Label Propagation Algorithm (SLPA)
  • Assign each node to the most probable community
• Compare communities against Wikipedia categories
Evaluation strategy (2)

• A community is compared to each category of its members
  • category with highest F-score is included in the evaluation

• Comparison of the **dicos** threshold to
  • no threshold (none)
  • co-occurrence threshold of 2 (cooc)
Community evaluation - results

- Evaluation on two different sets of communities

<table>
<thead>
<tr>
<th>communities</th>
<th>t</th>
<th>#comms</th>
<th>#persons</th>
<th>F</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>none</td>
<td>4,292</td>
<td>798,777</td>
<td>0.2883</td>
<td>0.6223</td>
<td>0.2316</td>
</tr>
<tr>
<td></td>
<td>cooc=2</td>
<td>3,584</td>
<td>677,880</td>
<td>0.2923</td>
<td>0.6002</td>
<td>0.2467</td>
</tr>
<tr>
<td></td>
<td>dicos</td>
<td>8,193</td>
<td>788,279</td>
<td><strong>0.2954</strong></td>
<td>0.5811</td>
<td><strong>0.2535</strong></td>
</tr>
<tr>
<td>subset</td>
<td>none</td>
<td>90</td>
<td>1,562</td>
<td><strong>0.4612</strong></td>
<td><strong>0.6138</strong></td>
<td>0.4341</td>
</tr>
<tr>
<td>10&lt;n&lt;500</td>
<td>cooc=2</td>
<td>301</td>
<td>10,683</td>
<td>0.4105</td>
<td>0.5583</td>
<td>0.4078</td>
</tr>
<tr>
<td></td>
<td>dicos</td>
<td>713</td>
<td>24,315</td>
<td>0.3889</td>
<td>0.4785</td>
<td>0.4238</td>
</tr>
</tbody>
</table>

WSN Evaluation
# Community evaluation - results

• Evaluation on two different sets of communities

<table>
<thead>
<tr>
<th>communities</th>
<th>t</th>
<th>#comms</th>
<th>#persons</th>
<th>F</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>none</td>
<td>4,292</td>
<td>798,777</td>
<td>0.2883</td>
<td>0.6223</td>
<td>0.2316</td>
</tr>
<tr>
<td></td>
<td>cooc=2</td>
<td>3,584</td>
<td>677,880</td>
<td>0.2923</td>
<td>0.6002</td>
<td>0.2467</td>
</tr>
<tr>
<td></td>
<td>dicos</td>
<td>8,193</td>
<td>788,279</td>
<td>0.2954</td>
<td>0.5811</td>
<td>0.2535</td>
</tr>
<tr>
<td>subset</td>
<td>none</td>
<td>90</td>
<td>1,562</td>
<td>0.4612</td>
<td>0.6138</td>
<td>0.4341</td>
</tr>
<tr>
<td>10&lt;n&lt;500</td>
<td>cooc=2</td>
<td>301</td>
<td>10,683</td>
<td>0.4105</td>
<td>0.5583</td>
<td>0.4078</td>
</tr>
<tr>
<td></td>
<td>dicos</td>
<td>713</td>
<td>24,315</td>
<td>0.3889</td>
<td>0.4785</td>
<td>0.4238</td>
</tr>
</tbody>
</table>
Community evaluation – discussion (1)

Problems with ground truth

• No clear rules for category assignment
• Inconsistency due to large number of editors
• Lack of semantic meaning
  ▪ e.g. categories based on same place of birth
• Hierarchical organization of categories
## Community evaluation – discussion (2)

### Problems with evaluation

- Person was assigned to most probable community -> very large communities
- Comparison to large categories like “Living People” (containing > 400,000 persons)
  - never receives high Precision Recall

**BUT** the social network was not designed to resemble Wikipedia categories
Interesting groups within the network are found, they are:

- semantically very well related
- not covered by any Wikipedia category
  - crew of men’s eight (1962 Commonwealth Games) (9 members)
  - members of Hungarian Heavy Metal band
  - "..."
Finish Archers

Jari Lipponen

Ismo Falck
1988, 1992 TEAM

Tomi Poikolainen

Päivi Meriluoto
1980, 1984

Kyösti Laasonen

Tapio Rautavaara
WC 1958

Sirkka Sokka-Matikainen
WC 2001

Mari Piuva
2004

Matti Hatava
2008

Mirjam Tuokkola
2014 youth Olympics
**Evaluation - Conclusion**

- WSN represents meaningful relationships
  - we find semantically very well related communities, but difficult to evaluate
- community evaluation based on Wikipedia category not optimal
- threshold with dicos weight works well
  - minimized loss of information
  - more communities are detected
  - F-measure nor considerably lower
Outline

• Introduction

• Extracting Person Mentions

• The Wikipedia Social Network

• Community Evaluation

• Conclusion and Ongoing Work
Conclusion

• First framework for extracting person-centric network from unstructured data (Wikipedia)
• Heuristics for determining person mentions
• $dicos$ weight for relationships
• Edge threshold based on $dicos$ weight
• WSN represent valid and meaningful relationships and shows typical properties of social networks
Ongoing work

• Find more person mentions using Stanford NER
• Add more person related information to the network (occupation, citizenship etc.)
• Apply the WSN to other tasks
  ▪ disambiguation person names in text
  Precision of 97%
Contact and resources

johanna.geiss@informatik.uni-heidelberg.de
http://dbs.ifi.uni-heidelberg.de/

Wikipedia Social Network is available at:
http://dbs.ifi.uni-heidelberg.de/index.php?id=data
Questions
Statistics for Wikidata

• **2.6 M** person entries in total
• **1.2 M** entries with [link to EN Wikipedia](https://en.wikipedia.org)
• Top 3 occupations:
  - politician (15.5%)
  - association football player (14.5%)
  - actor (7.4%)
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Extraction</th>
<th>WSN</th>
<th>Evaluation</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

**Identifying person mentions – IWL (4)**

- 76.8 M IWLs in EN Wikipedia
- 10.4 M IWLs refer to persons (13.6%)
- $\approx 840,000$ different persons

**TOP 5:**
- Barack Obama (11,954 IWLs)
- George W. Bush (10,650 IWLs)
- Napoleon (8,918 IWLs)
- William Shakespeare (8,770 IWLs)
- Adolf Hitler (8,229 IWLs)
# Network properties - Pagerank

## TOP 10 ranked persons according to pagerank

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Birth</th>
<th>Death</th>
<th>degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Barack Obama</td>
<td>1961</td>
<td></td>
<td>1561</td>
</tr>
<tr>
<td>2</td>
<td>John Paul II</td>
<td>1920</td>
<td>2005</td>
<td>1449</td>
</tr>
<tr>
<td>3</td>
<td>George W. Bush</td>
<td>1946</td>
<td></td>
<td>1419</td>
</tr>
<tr>
<td>4</td>
<td>Adolf Hitler</td>
<td>1889</td>
<td>1945</td>
<td>1508</td>
</tr>
<tr>
<td>5</td>
<td>Bill Clinton</td>
<td>1946</td>
<td></td>
<td>1249</td>
</tr>
<tr>
<td>6</td>
<td>Franklin D. Roosevelt</td>
<td>1882</td>
<td>1945</td>
<td>1232</td>
</tr>
<tr>
<td>7</td>
<td>Napoleon</td>
<td>1769</td>
<td>1821</td>
<td>1572</td>
</tr>
<tr>
<td>8</td>
<td>Benedict XVI</td>
<td>1927</td>
<td></td>
<td>1142</td>
</tr>
<tr>
<td>9</td>
<td>Elizabeth II</td>
<td>1926</td>
<td></td>
<td>1217</td>
</tr>
<tr>
<td>10</td>
<td>Ronald Reagan</td>
<td>1911</td>
<td>2004</td>
<td>1130</td>
</tr>
<tr>
<td>Introduction</td>
<td>Extraction</td>
<td>WSN</td>
<td>Evaluation</td>
<td>Conclusion</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td>-----</td>
<td>------------</td>
<td>------------</td>
</tr>
</tbody>
</table>

**Community evaluation - ground truth**

- categories from Wikipedia
- more than 840,000 categories for the persons in our network
- most persons > 1 category
- Winston Churchill in 97 categories
- largest category „Living people“ 438,500 members
Community detection

Stabilized Label Propagation Algorithm (SLPA)

• edge weights are not considered

• soft clustering: multiple labels possible
  ▪ overlapping communities
  ▪ probability of community membership depends on label distribution

• threshold for probabilities to reduce number of communities