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

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

ASONAM 28.08.2016

Explicit networks

Latently embedded

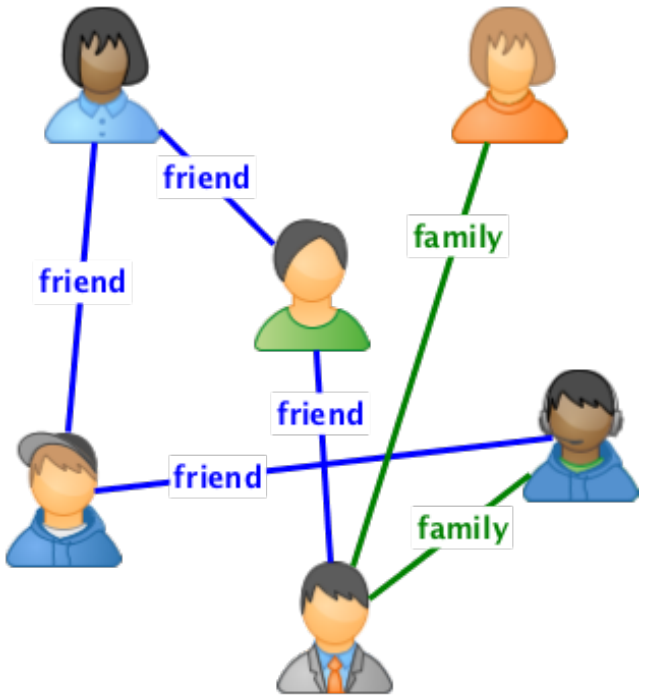
Explicit networks

GitHub

Latently embedded

Explicit networks

GitHub



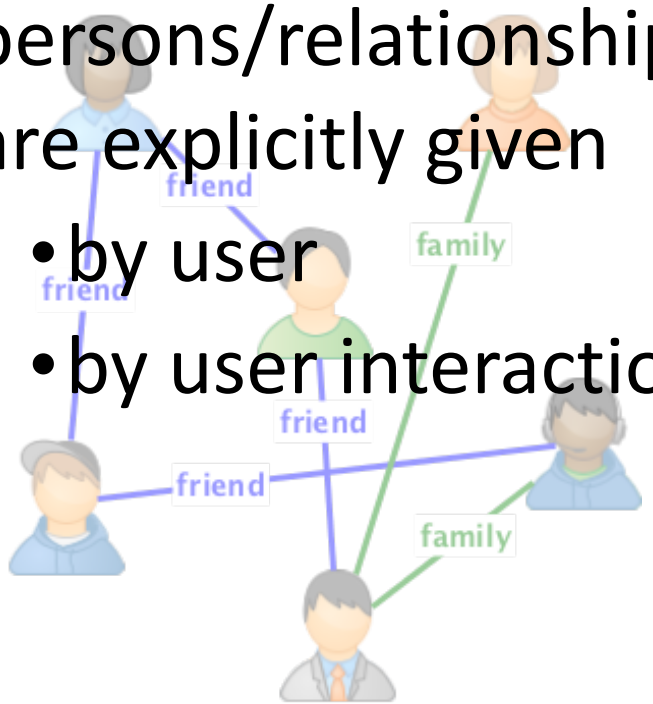
Latently embedded

Explicit networks

Latently embedded

GitHub

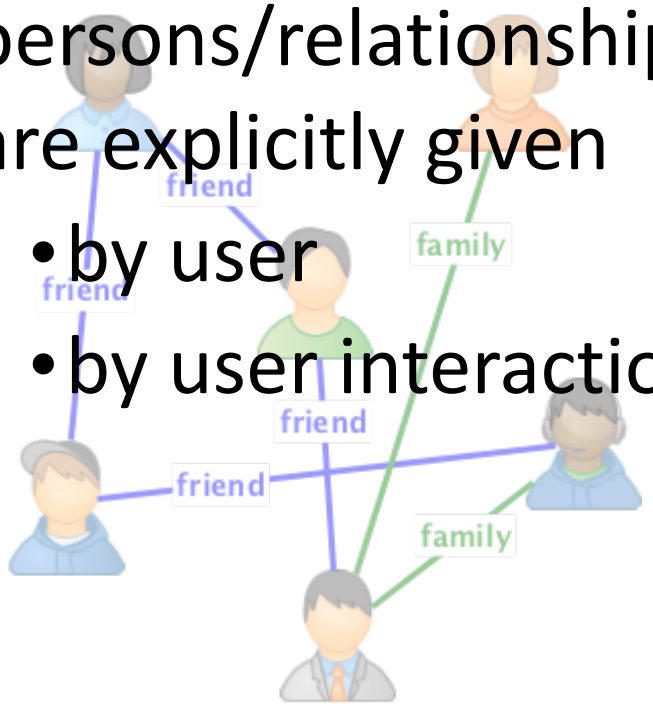
- persons/relationships are explicitly given
  - by user
  - by user interactions



Explicit networks

GitHub

- 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

Person co-occurence

Persons co-occur in the **same** article

➔ persons belong to same context/theme

Star Wars: The Force Awakens

en.wikipedia.org/wiki/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-occurence

Persons co-occur in the **same** article

➔ persons belong to same context/theme

Star Wars: The Force Awakens

en.wikipedia.org/wiki/Star_Wars:_The_Force_Awakens

From Wikipedia, the free encyclopedia

Star Wars: The Force Awakens (also known ***wakens***) is an upcoming American epic space opera film directed by J. episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driv 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).

cast members

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-occurence

Persons co-occur in the **same** article

➔ persons belong to same context/theme

Star Wars: The Force Awakens

en.wikipedia.org/wiki/Star_Wars:_The_Force_Awakens

From Wikipedia, the free encyclopedia

Star Wars: The Force Awakens (also known ***wakens***) is an upcoming American epic space opera film directed by J. episodic *Star Wars* film series, it stars John Boyega, Daisy Ridley, Adam Driv 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 trilogy announced after The Walt Disney Company's acquisition of Lucasfilm in 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

producers

Person co-occurence

Persons co-occur in the **same** article

➔ persons belong to same context/theme

Distance of person mentions

Distance between person mentions in article

➔ measure for the strength of the relationship

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

identifier	Q81328
label	Harrison Ford (en), Харрисон Форд (ru)
description	American film actor and producer
alias	Harisson Ford (fr)
statement	**Instance of: human**, date of birth: 13 July 1942
link	Harrison Ford

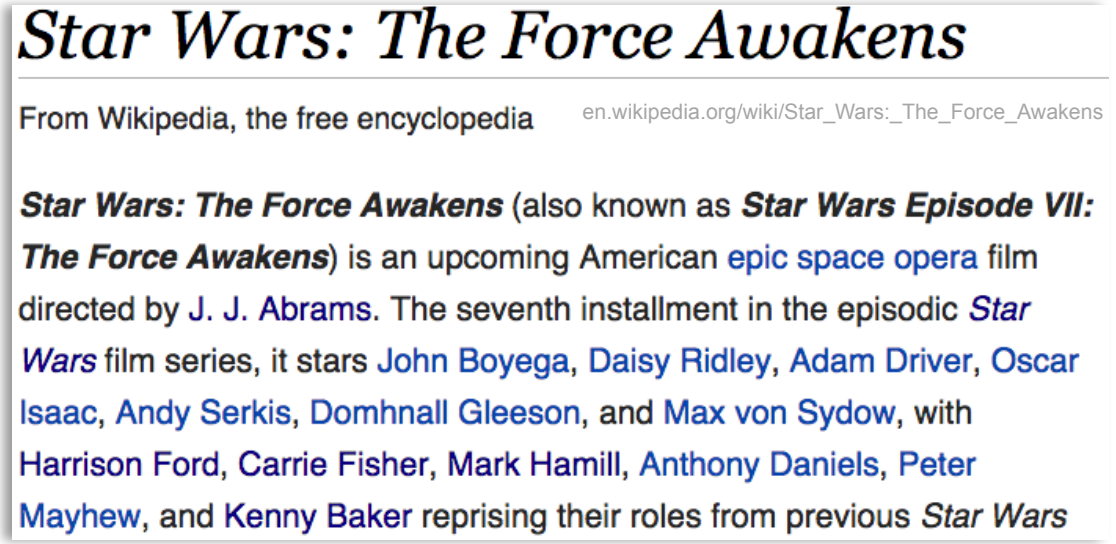
- **1.2 M** persons with **link to EN Wikipedia**

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)



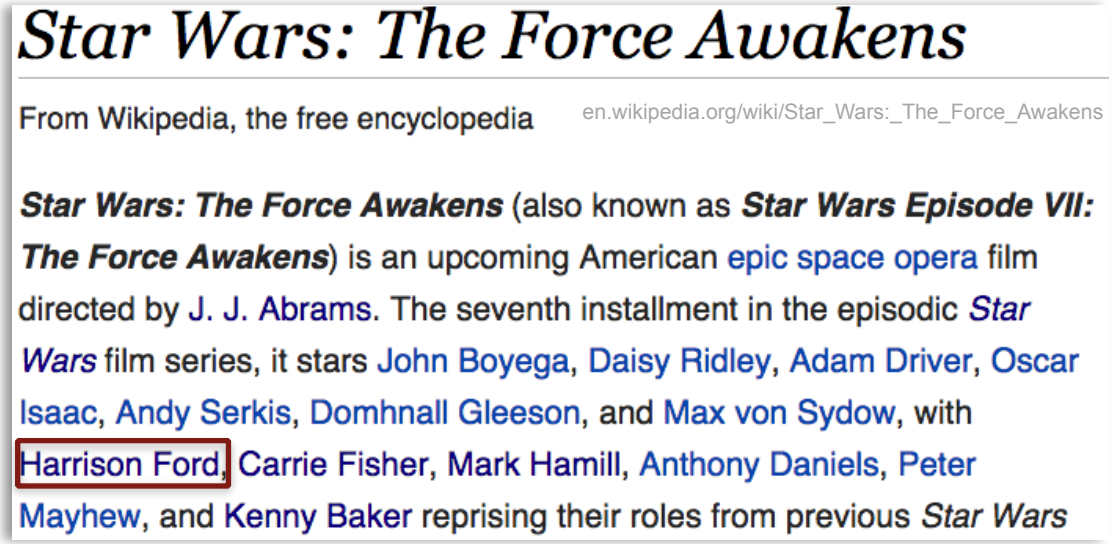
Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Star_Wars:_The_Force_Awakens

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)



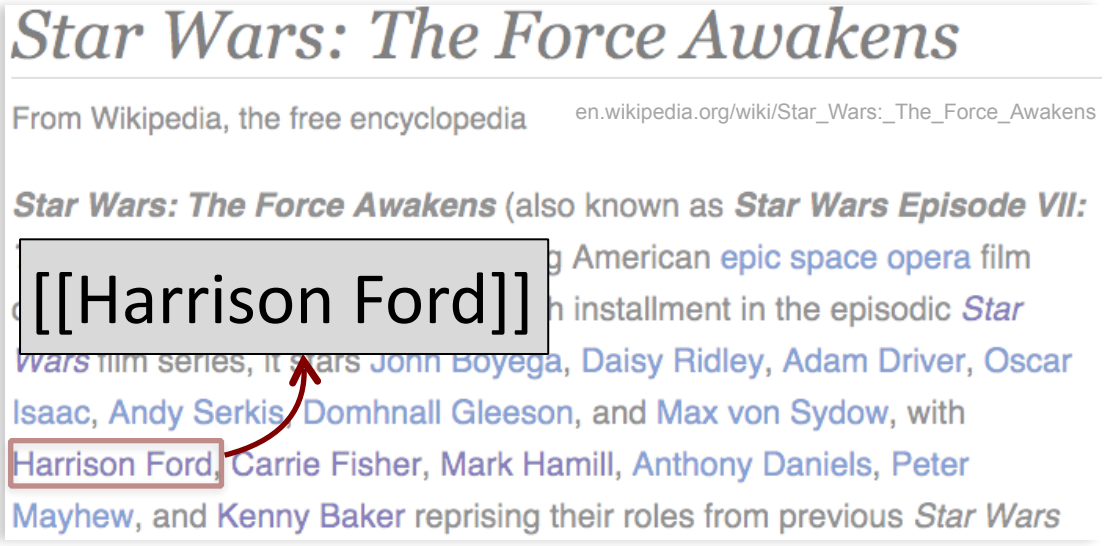
Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Star_Wars:_The_Force_Awakens

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*

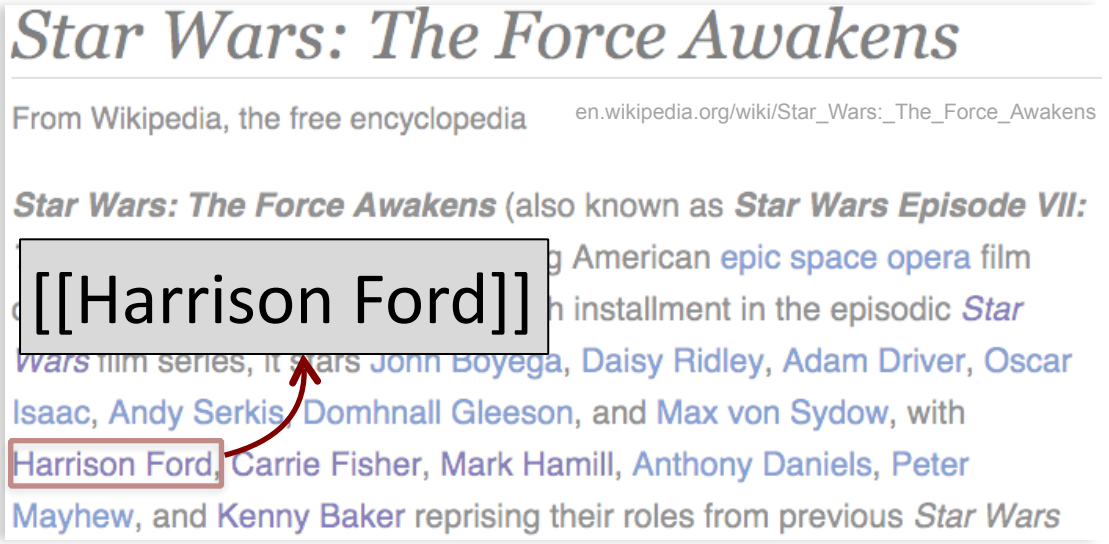
Identifying person mentions – IWL (1)

Interwiki-Links (IWL)



Identifying person mentions – IWL (1)

Interwiki-Links (IWL)



[[link target | surface]]

e.g. [[Harrison Ford | Ford]]

Identifying person mentions – IWL (1)

Interwiki-Links (IWL)

Star Wars: The Force Aw

From Wikipedia, the free encyclopedia en.wikipedia.org/wiki/

Star Wars: The Force Awakens (also known as ***Sta***

The Force Awakens) is an upcoming American epic

directed by J.J. Abrams. The seventh installment in

Wars film series, it stars John Boyega, Daisy Ridley,

Isaac, Andy Serkis, Domhnall Gleeson, and Max von

Harrison Ford, Carrie Fisher, Mark Hamill, Anthony D

Mayhew, and Kenny Baker reprising their roles from

[[Harrison Ford]]

Harrison Ford en.wikipedia.org/wiki/Harrison_Ford

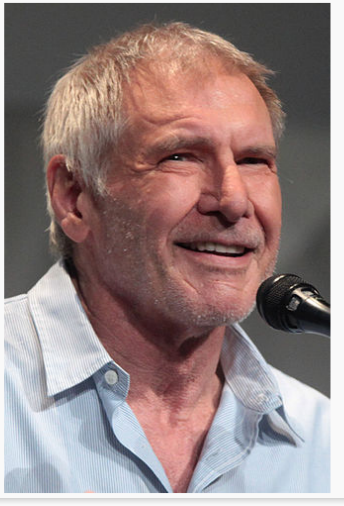
From Wikipedia, the free encyclopedia

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

Harrison Ford



[[link target | surface]]

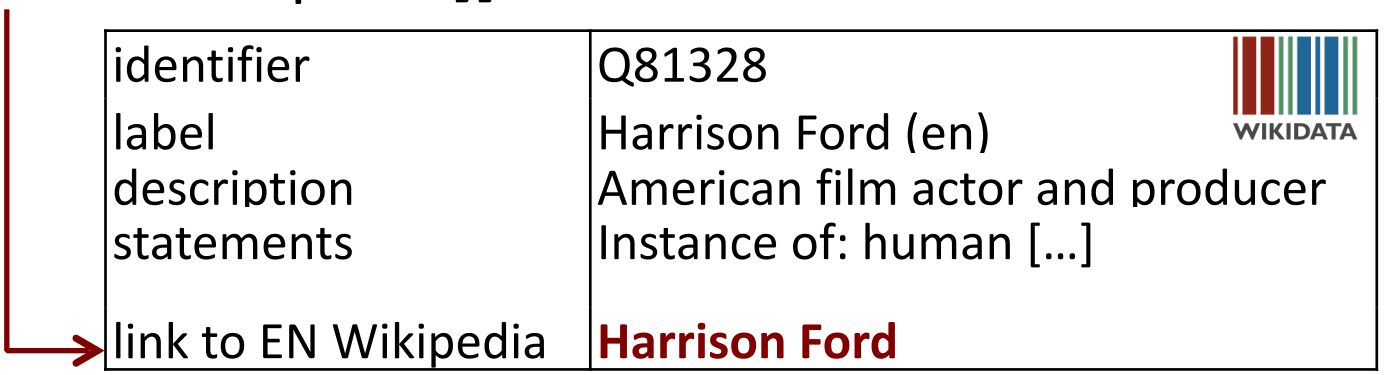
e.g. [[Harrison Ford | Ford]]

Identifying person mentions – IWL (2)

IWL refers to person if

1. Link target = link entry in Wikidata

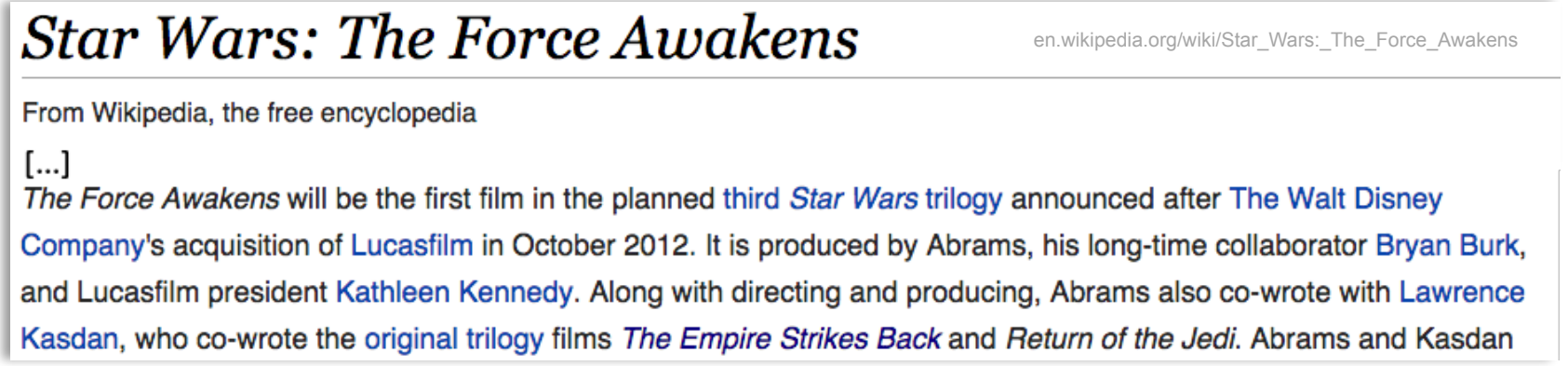
   [[Harrison Ford |Ford]]



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

en.wikipedia.org/wiki/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

en.wikipedia.org/wiki/Star_Wars:_The_Force_Awakens

From Wikipedia, the free encyclopedia

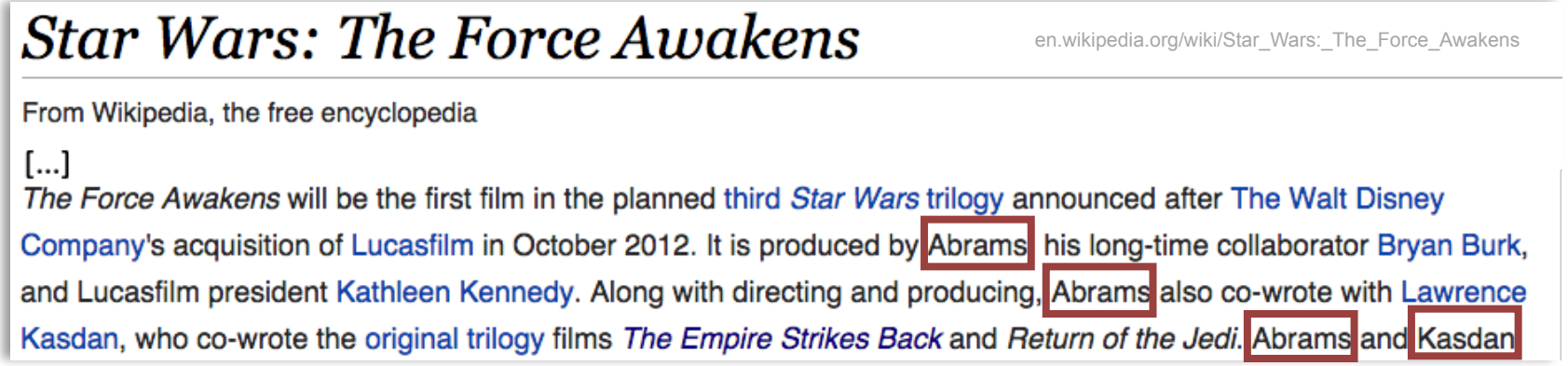
[...]

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

en.wikipedia.org/wiki/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
  - [[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 ≈ 840,000 different persons
- in 1.8 M Wikipedia articles
- TOP 5:
  1. Jesus (14)
  2. Napoleon (3)
  3. Barack Obama (1)
  4. Muhammad (27)
  5. William Shakespeare (4)
- 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

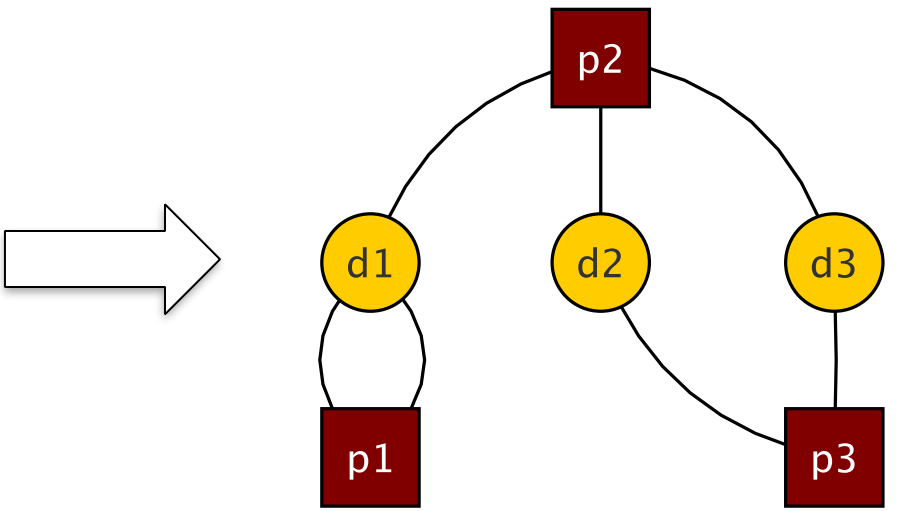
- List of persons per sentence in document

Document	***Sentence***	***Persons***
d1	s0	p1, p2
	s3	p1
d2	s1	p2
	s2	p3
d3	s2	p2
	s12	p3

Network construction - bipartite graph

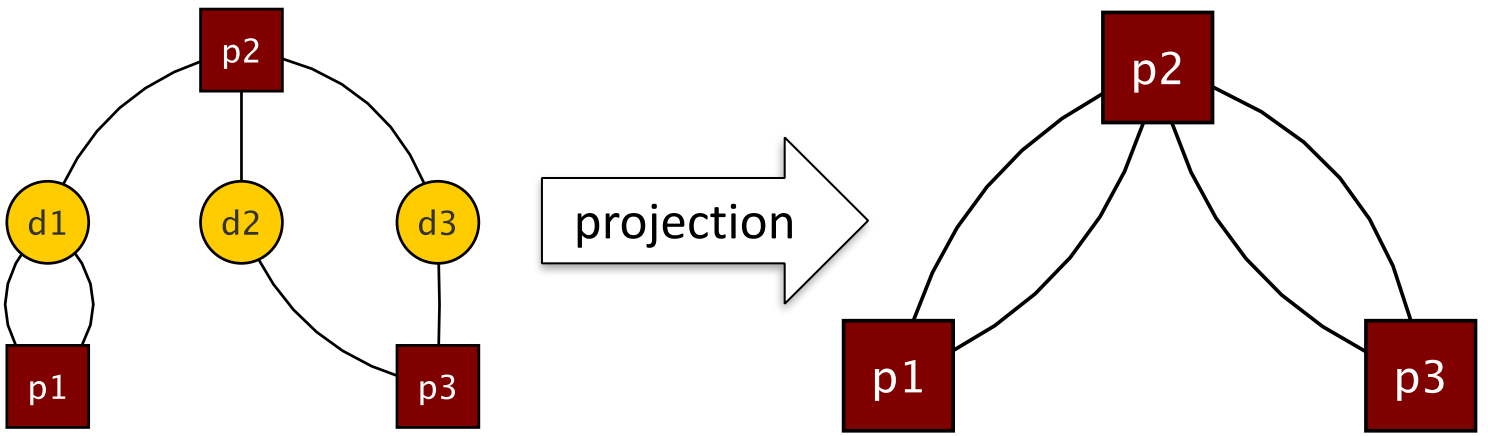
- List of persons per sentence in document
- Bipartite graph of documents and persons

Document	*Sentence*	*Persons*
d1	s0	p1, p2
	s3	p1
d2	s1	p2
	s2	p3
d3	s2	p2
	s12	p3



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):

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

Network construction – edge weights

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

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

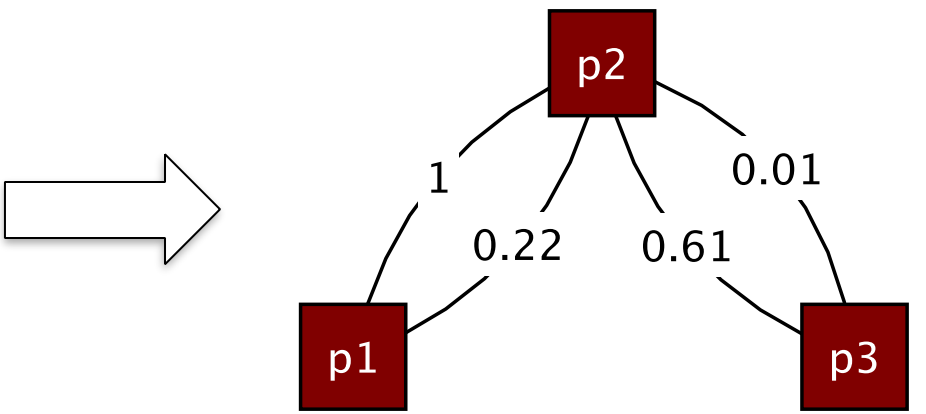
v,w,i	d	φ
p1, p2, d1	0	1
p1, p2, d1	3	0.22
p2, p3, d2	1	0.61
p2, p3, d3	10	0.01

Network construction – edge weights

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

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

v,w,i	d	φ
p1, p2, d1	0	1
p1, p2, d1	3	0.22
p2, p3, d2	1	0.61
p2, p3, d3	10	0.01



Network construction – edge aggregation

- Node-edge incidence matrix **M**

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

Network construction – edge aggregation

- Node-edge incidence matrix **M**
- Node-edge incidence vector

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

Network construction – edge aggregation

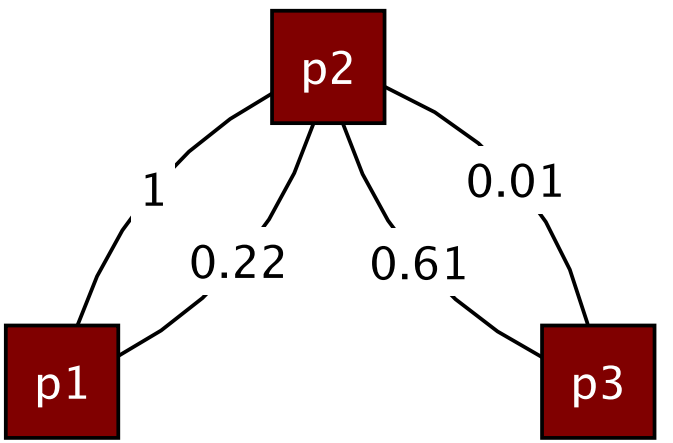
- Node-edge incidence matrix **M**
- Node-edge incidence vector
- **dicos** → cosine similarity of incidence vectors

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

Network construction – edge aggregation

- Node-edge incidence matrix **M**
- Node-edge incidence vector
- **dicos** → cosine similarity of incidence vectors

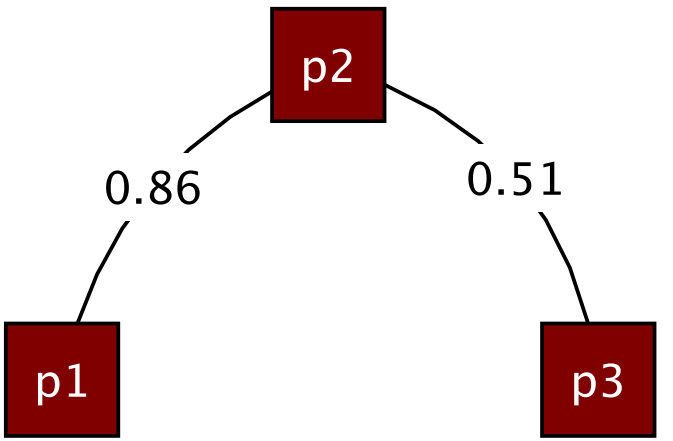
M	p1	p2	p3
e1	1	1	0
e2	0.22	0.22	0
e3	0	0.61	0.61
e4	0	0.01	0.01



Network construction – edge aggregation

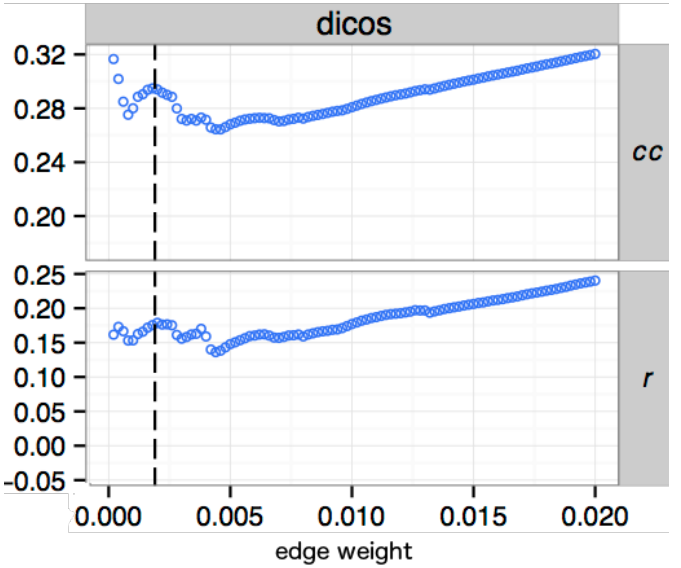
- Node-edge incidence matrix **M**
- Node-edge incidence vector
- **dicos** → cosine similarity of incidence vectors

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



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
  - 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

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

Community evaluation - results

- Evaluation on two different sets of communities

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

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

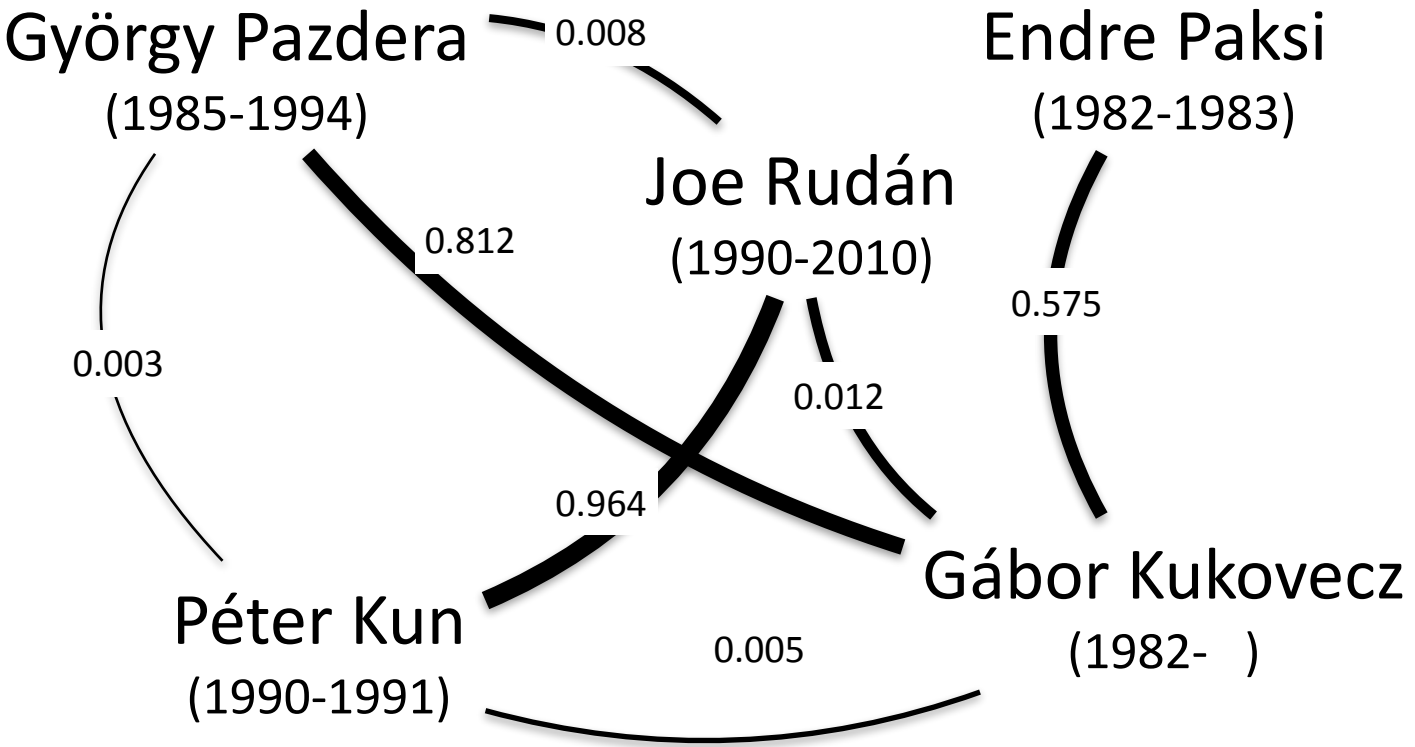
BUT the social network was not designed to resemble Wikipedia categories

Community evaluation – discussion (3)

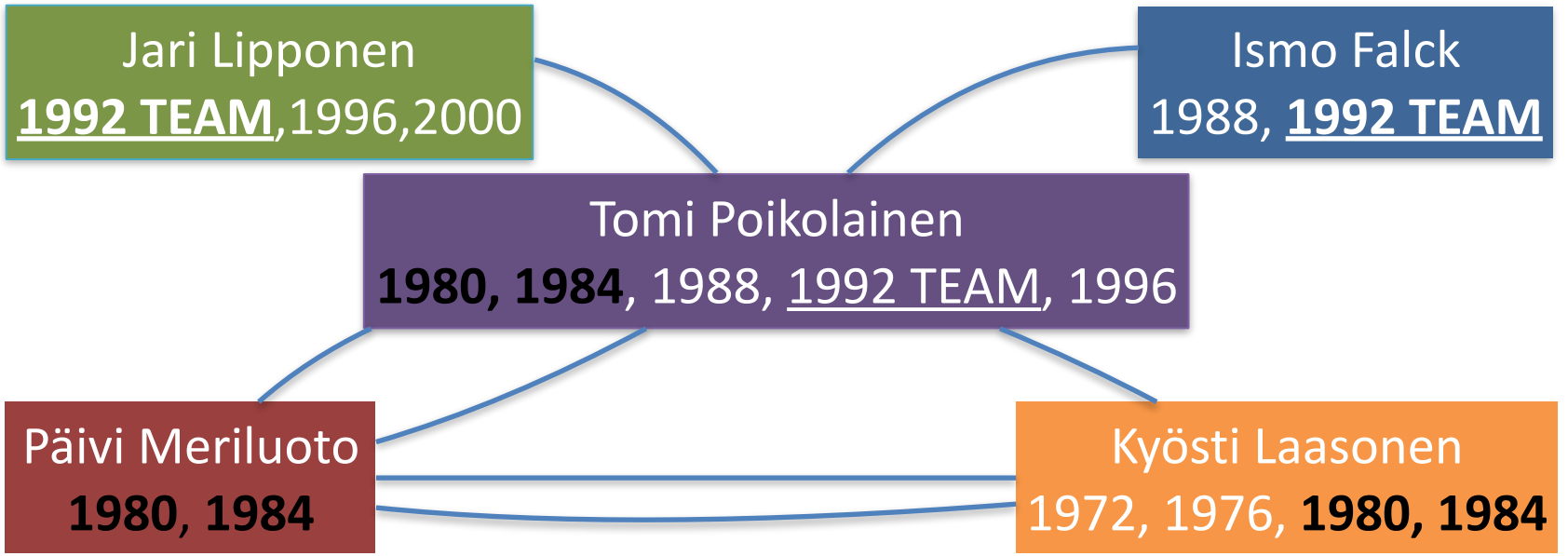
Interesting groups within the network are found, they are:

- 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
  - "…"

Pokolgép- Hungarian Heavy Metal Band



Finish Archers



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

Additional material

Statistics for Wikidata

- 2.6 M person entries in total
- **1.2 M** entries with **link to EN Wikipedia**
- Top 3 occupations:
  - politician (15.5%)
  - association football player (14.5%)
  - actor (7.4 %)

Identifying person mentions – IWL (4)

- 76.8 M IWLs in EN Wikipedia
- 10.4 M IWLs refer to persons (13.6%)
- ≈ 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

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

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