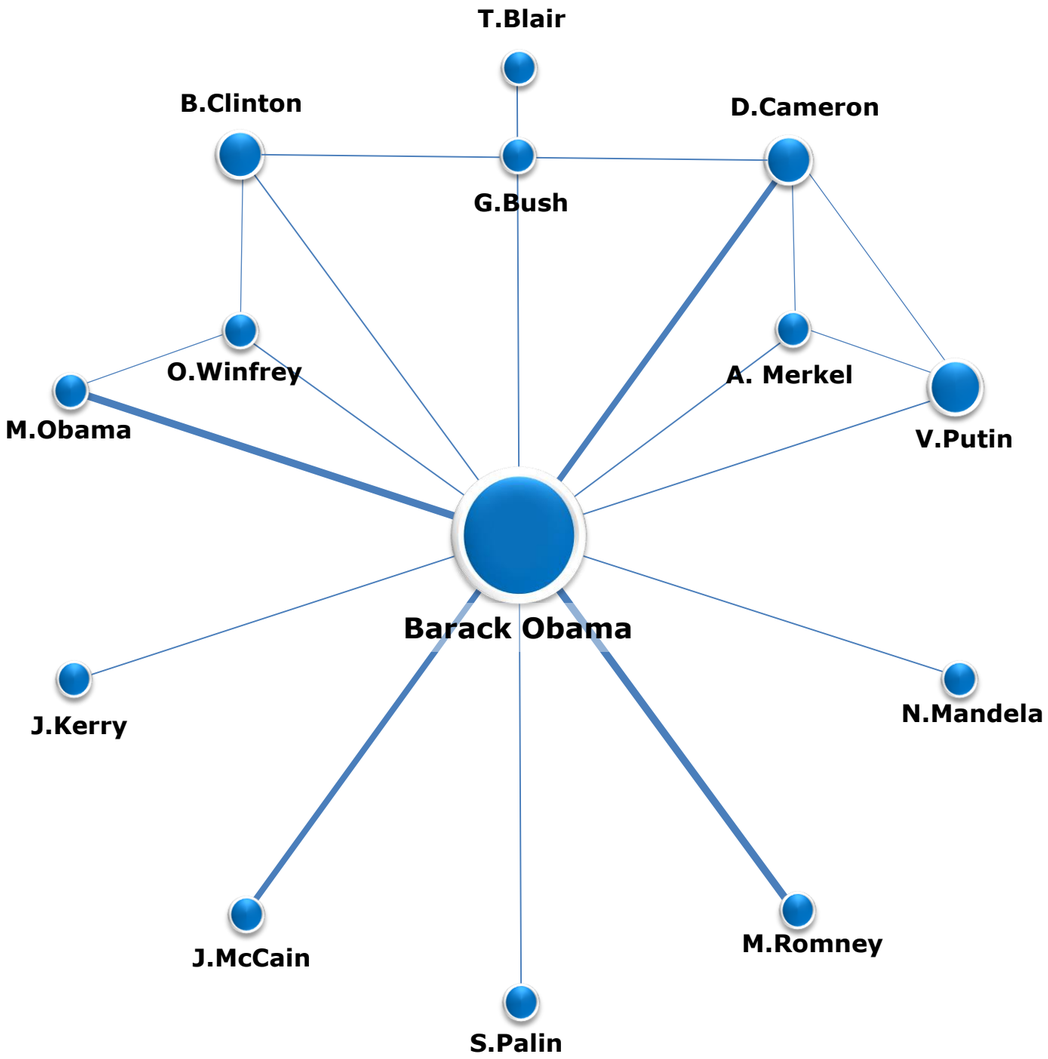
Temporal Analysis of Social Networks
(on the Web and in Archives)

Stefan Siersdorfer, Philipp Kemkes,
Hanno Ackermann, Miroslav Shaltev, **Sergej Zerr**

Social Network Applications



Social Network sample extracted using our methods

Social networks can be leveraged to:

- Identify influential entities
- Detect communities with special interests
- Spread of ideas / diseases
- Entity disambiguation
- etc.

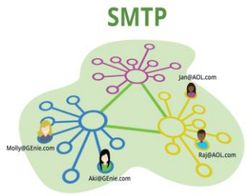
Social Network Sources

Structured

- Social applications
- Knowledge bases

Semi-structured

- Email traffic
- DBLP



Unstructured

- Photos
- Characters from books
- Web page content

Overview: Related work

Relation mining in Web documents (through crawling):

- **Textrunner**
  (Yates et all, 2007)
- **KnowItAll**
  (Etzioni et all, 2004)

Relation mining through search engines:

- **Polyphonet**
  (Matsuo et all, 2006)

Overview: Related work

Relation mining in Web documents (through crawling):

- **Textrunner**
  (Yates et all, 2007)
- **KnowItAll**
  (Etzioni et all, 2004)

Fixed to a dataset

Relation mining through search engines:

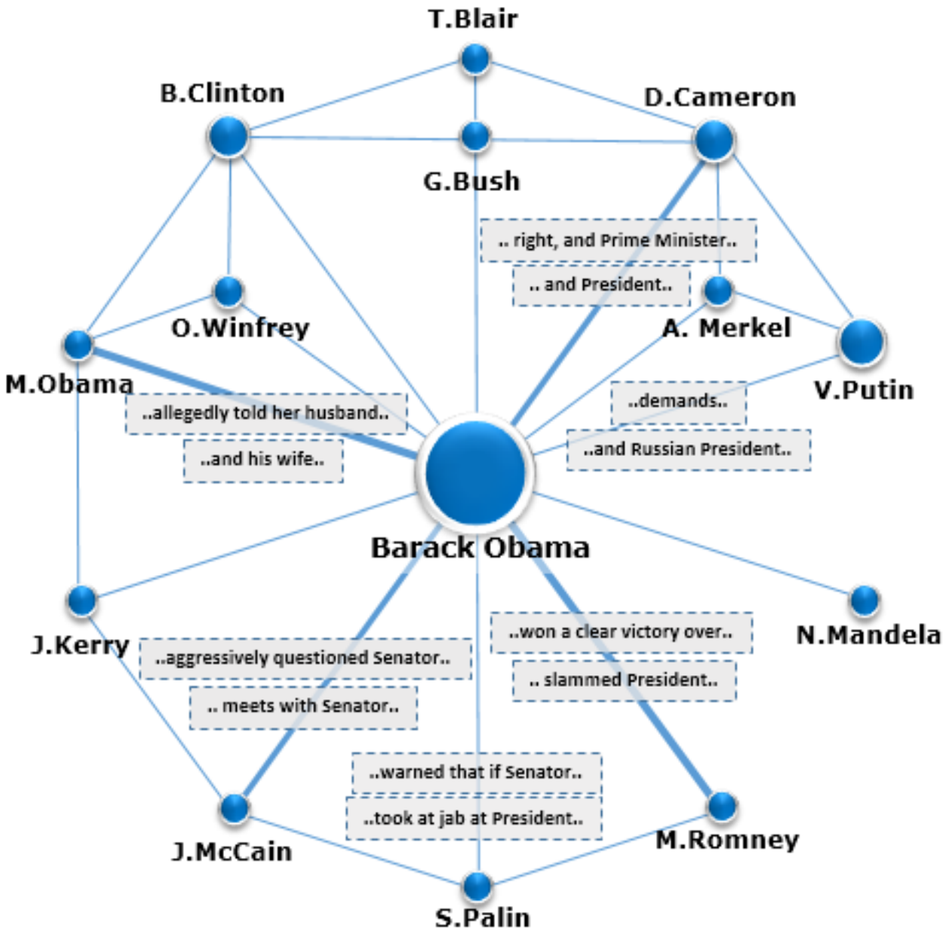
- **Polyphonet**
  (Matsuo et all, 2006)

Fixed to a set of entities

Who with Whom and How? - Guided Pattern Mining for Extracting Large Social Networks using Search Engines

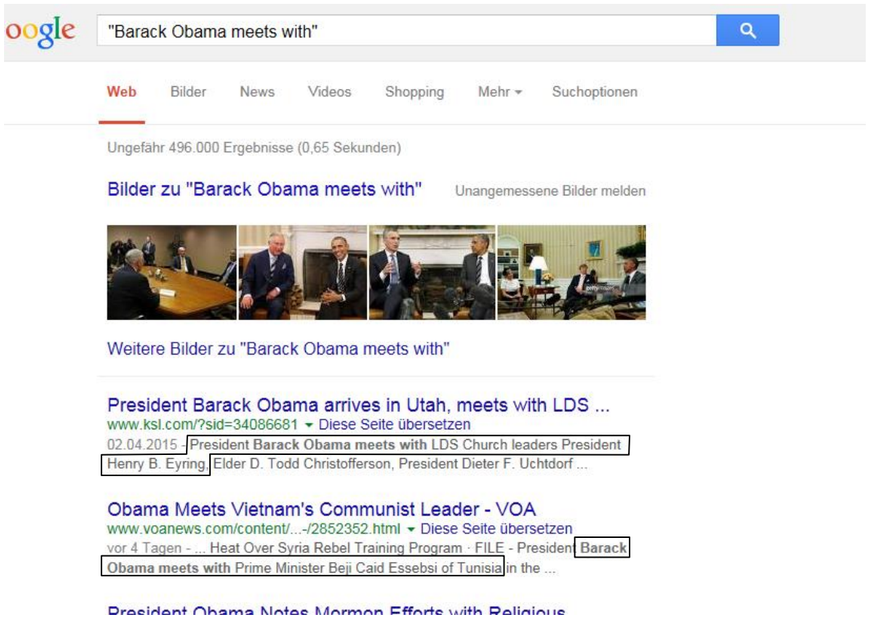
Use intelligently formulated Search Engine requests to discover relations of a person from the seed set

Who with Whom and How? - Guided Pattern Mining for Extracting Large Social Networks using Search Engines



Network beeing extracted with a seed set of politician names

Use intelligently formulated Search Engine requests to discover relations of a person from the seed set



Our Approach for Efficient Graph Mining

- Network expansion using pattern based search engine queries

  Getting Children Nodes

  - *<Barack Obama> <with> ?*
  - *<Angela Merkel> <and her colleague> ?*

Our Approach for Efficient Graph Mining

- Network expansion using pattern based search engine queries

  Getting Children Nodes

  - *<Barack Obama> <with>* ?
  - *<Angela Merkel> <and her colleague>* ?

- Extraction of pairs from snippets

  President *Barack Obama took a jab* at Mitt Romney on Thursday evening

Our Approach for Efficient Graph Mining

- Network expansion using pattern based search engine queries

  Getting Children Nodes

  - *<Barack Obama> <with> ?*
  - *<Angela Merkel> <and her colleague> ?*

- Extraction of pairs from snippets

  President *Barack Obama took a jab* at Mitt Romney on Thursday evening

- Intelligent prioritization approaches for discovered nodes

Our Approach for Efficient Graph Mining

- Network expansion using pattern based search engine queries

  Getting Children Nodes

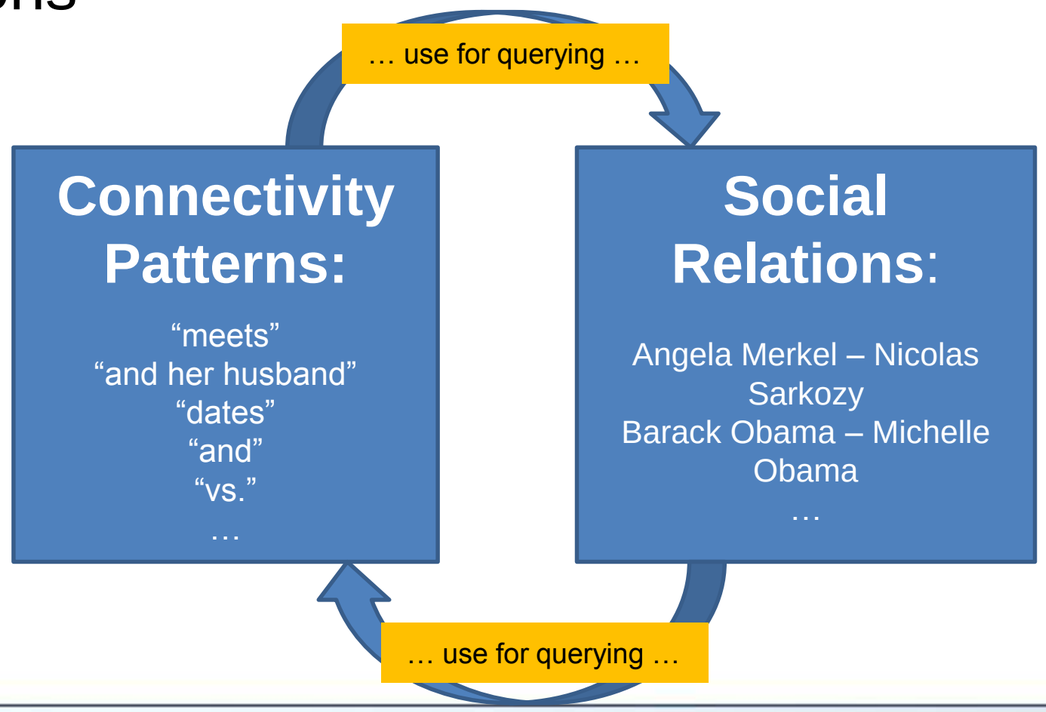
  - *<Barack Obama> <with>* ?
  - *<Angela Merkel> <and her colleague>* ?

- Extraction of pairs from snippets

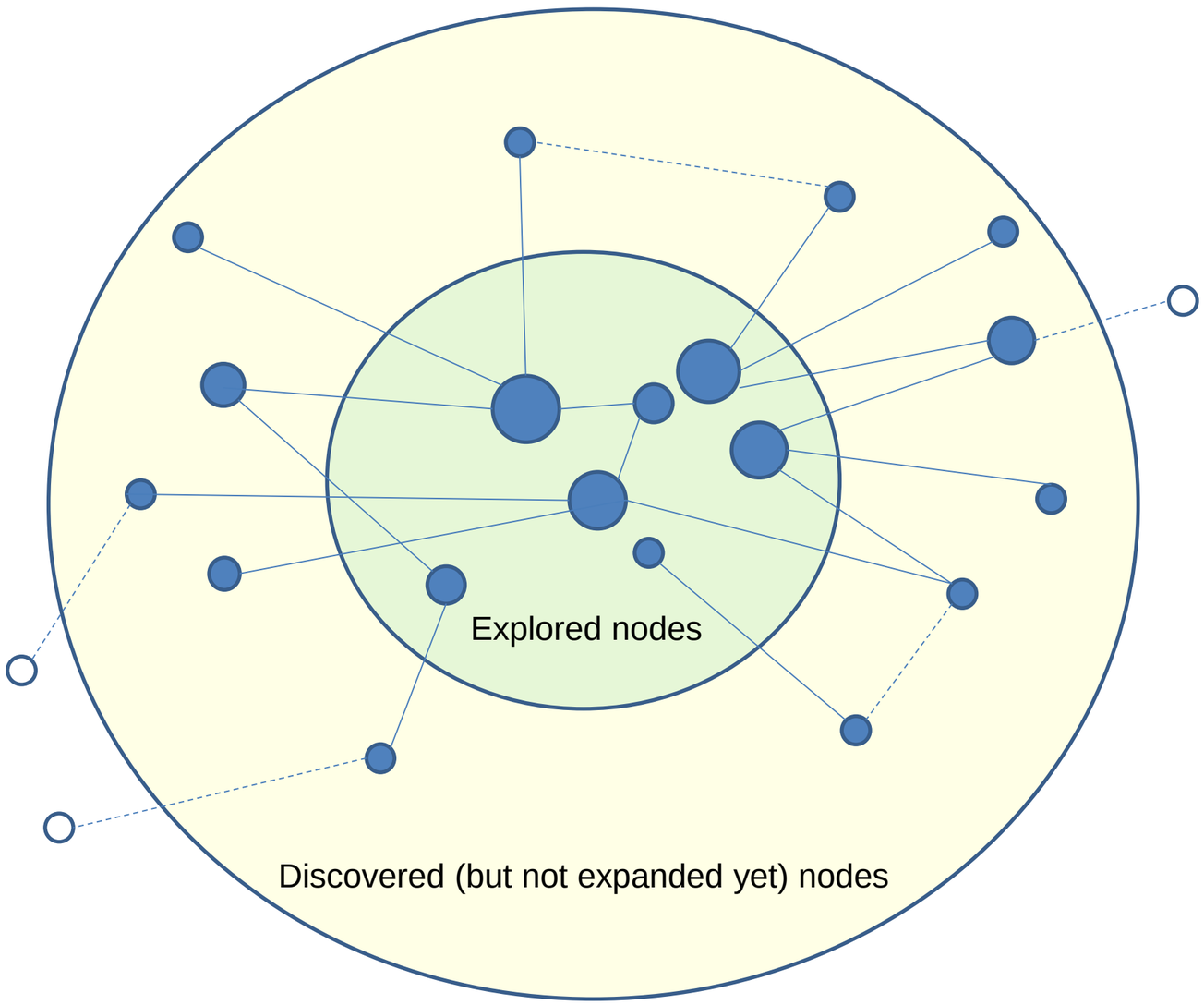
  President *Barack Obama took a jab* at Mitt Romney on Thursday evening

- Intelligent prioritization approaches for discovered nodes

- Bootstrapping approach for covering multiple aspects of social relations

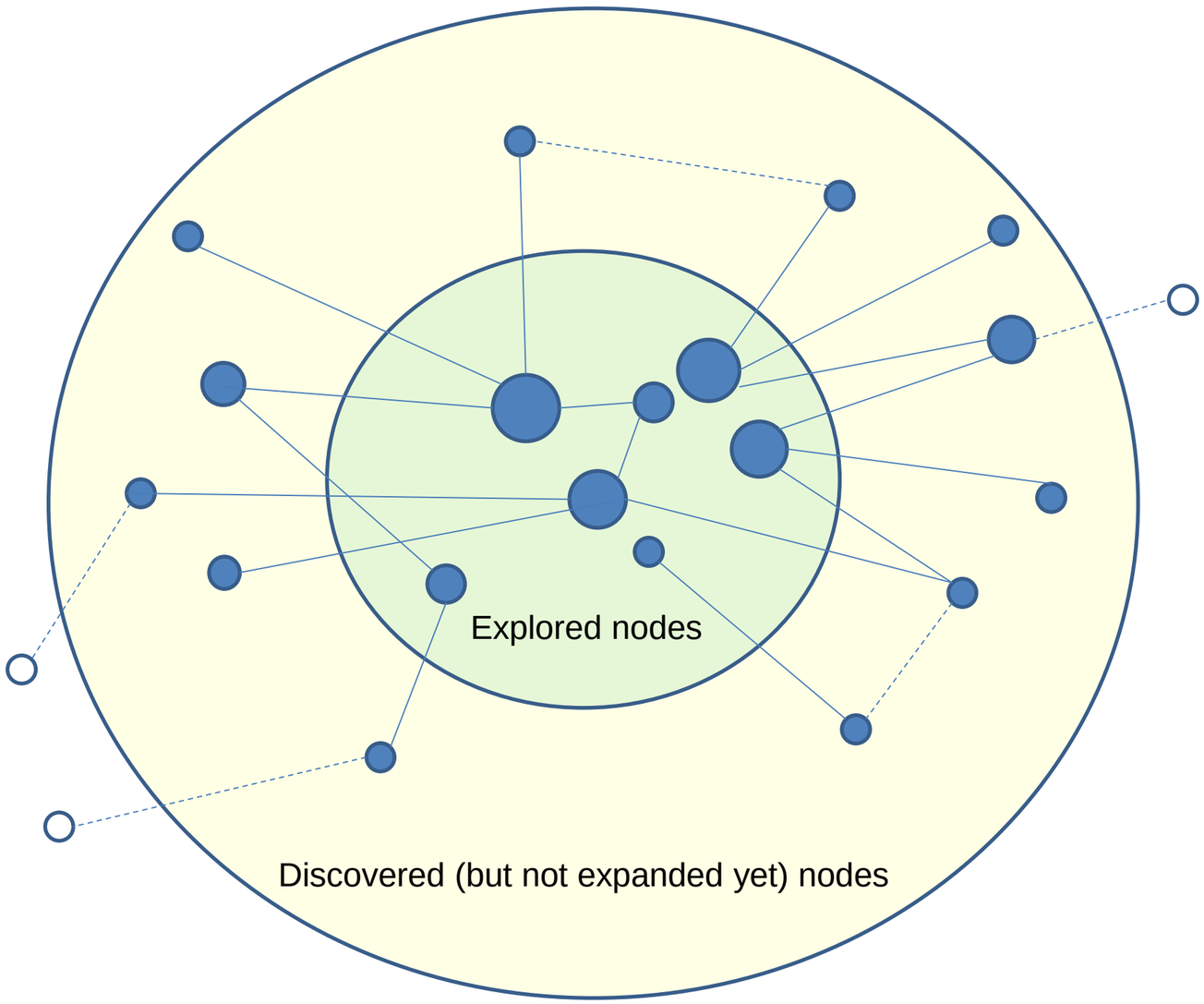


Traversal Strategies



Minimize the number of search engine API requests

Traversal Strategies



Breath first

- good coverage

Decay prioritization

- Prefere „essential“ nodes
- reduce search requests

$$\varphi(v) = \rho * e^{\alpha * t}$$

ρ Popularity

t Expansion steps (novelty)

α Balance parameter

Minimize the number of search engine API requests

Evaluation: Data

Two evaluation sets from different domains

- Wikipedia (1 million persons)
  - Seed set: 342 current heads of state and government
- IMDb (2 million persons)
  - Seed set: 100 current and former leading actors

Evaluation: Efficiency

- We varied the alpha parameter and compared to naive BF approach and the baseline from the literature (Polyphonet)
- We tested each approach with 200,000 Search Engine requests

Algorithm	Nodes	Edges
WikiNet	baseline t=0.4	2,925	3,838
IMDbNet	baseline t=0.4	2,192	6,666

Evaluation: Efficiency

- We varied the alpha parameter and compared to naive BF approach and the baseline from the literature (Polyphonet)
- We tested each approach with 200,000 Search Engine requests

Algorithm	Nodes	Edges
WikiNet	baseline t=0.4	2,925	3,838
	BF	113,988	368,806
IMDbNet	baseline t=0.4	2,192	6,666
	BF	109,453	376,429

Evaluation: Efficiency

- We varied the alpha parameter and compared to naive BF approach and the baseline from the literature (Polyphonet)
- We tested each approach with 200,000 Search Engine requests

	Algorithm	Nodes	Edges
WikiNet	baseline t=0.4	2,925	3,838
	BF	113,988	368,806
	decay α=0	98,479	456,223
	decay α=0.005	110,260	346,663
	decay α=0.01	116,081	323,372
IMDbNet	baseline t=0.4	2,192	6,666
	BF	109,453	376,429
	decay α=0	107,085	425,830
	decay α=0.005	112,736	264,679
	decay α=0.01	112,782	242,184

Popularity

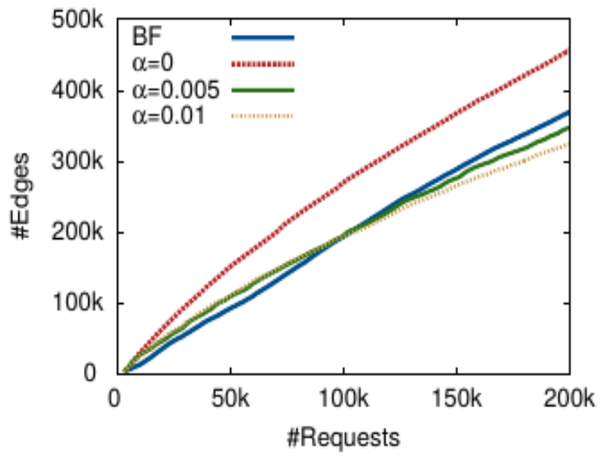
Novelty

$$\boldsymbol{\varphi}(v) = \boldsymbol{\rho} * \ \boldsymbol{e}^{\boldsymbol{\alpha} * \boldsymbol{t}}$$

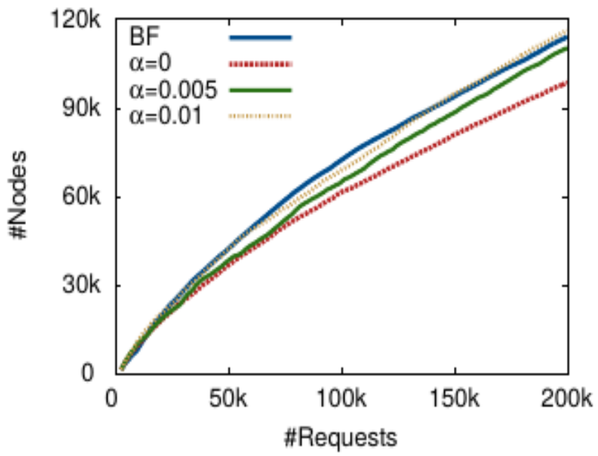
Evaluation: Efficiency

- We varied the alpha parameter and compared to naive BF approach and the baseline from the literature (Polyphonet)
- We tested each approach with 200,000 Search Engine requests

Algorithm	Nodes	Edges
WikiNet	baseline t=0.4	2,925	3,838
	BF	113,988	368,806
	decay α=0	98,479	456,223
	decay α=0.005	110,260	346,663
	decay α=0.01	116,081	323,372
IMDbNet	baseline t=0.4	2,192	6,666
	BF	109,453	376,429
	decay α=0	107,085	425,830
	decay α=0.005	112,736	264,679
	decay α=0.01	112,782	242,184

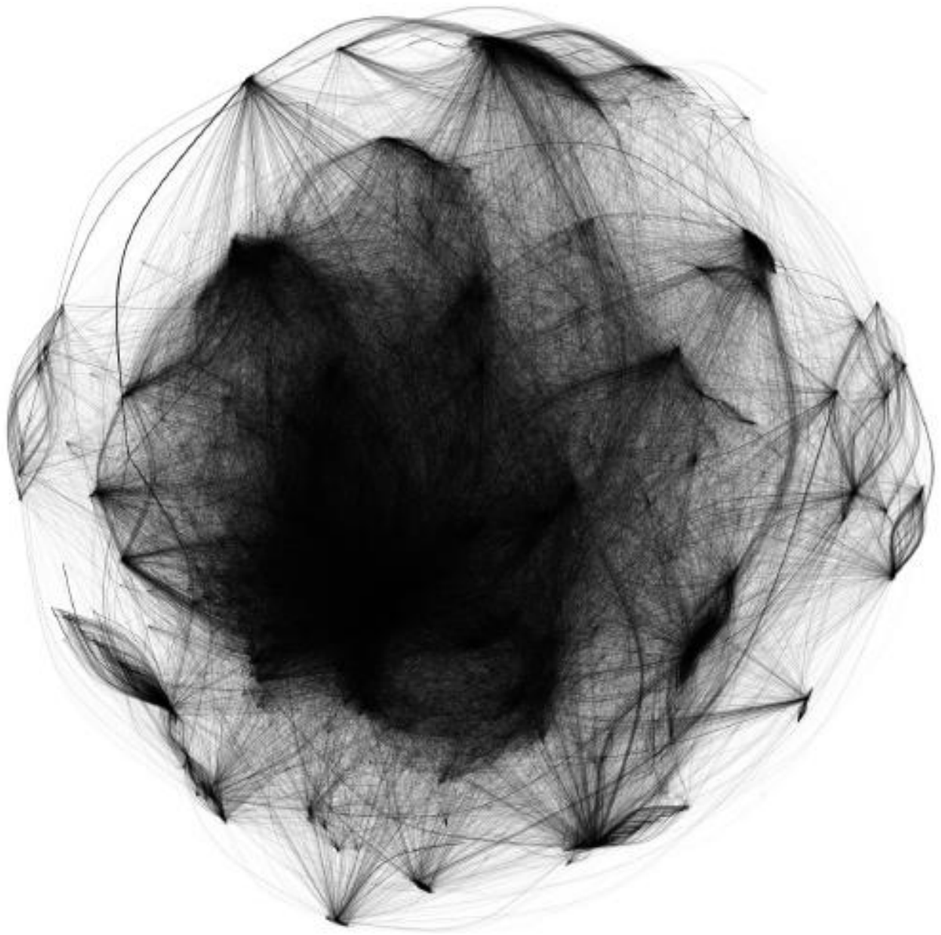


(a) WikiNet edges

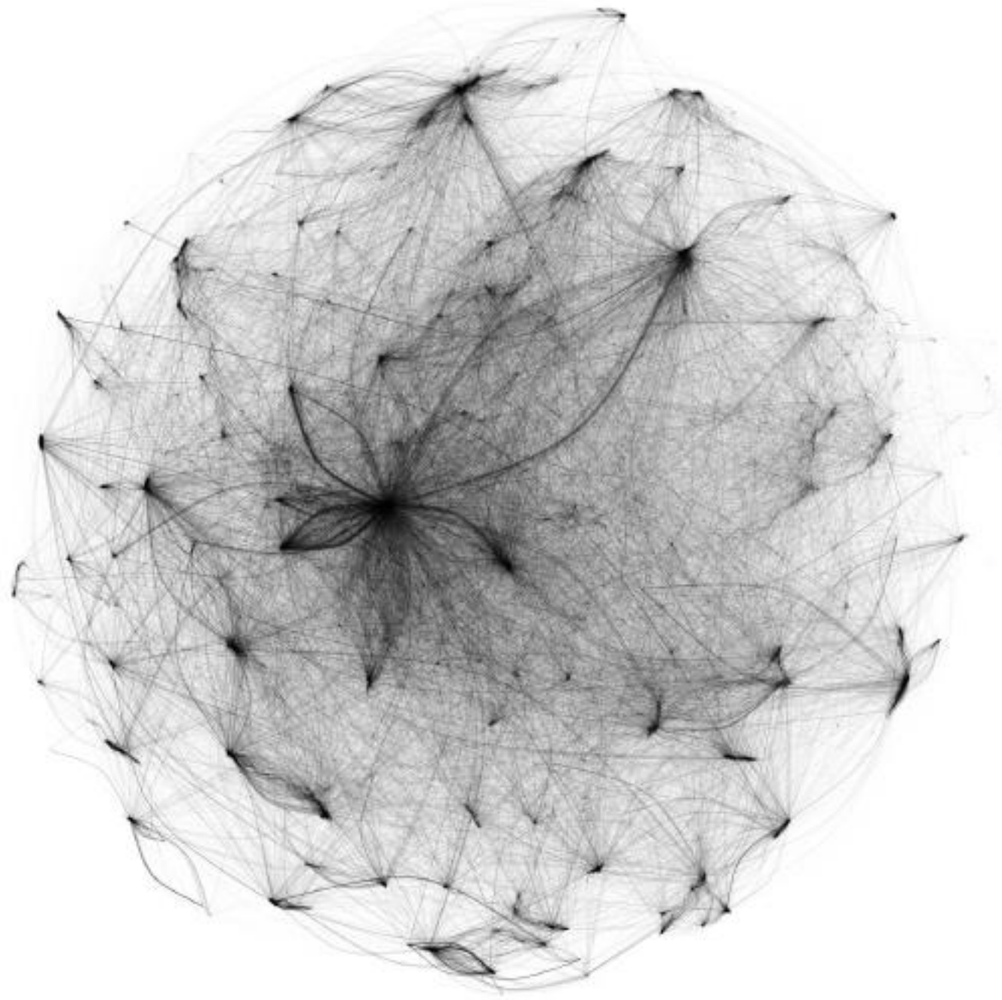


(b) WikiNet nodes

Example Networks



<u>More popularity focused:</u>
98,000 Persons
456,000 Social Relations



<u>More novelty focused:</u>
113,000 Persons
323,000 Social Relations

Extracted relations

PERSON 1		PERSON 2
Brad Pitt	–	Angelina Jolie
Trey Parker	–	Matt Stone
Spencer Pratt	–	Heidi Montag
Stephen Moyer	–	Anna Paquin
Bud Abbott	–	Lou Costello
Robert Pattinson	–	Kristen Stewart
Bruce Robison	–	Kelly Willis
Ryan Gosling	–	Eva Mendes
Dev Patel	–	Freida Pinto
Terence Hill	–	Bud Spencer
Sergey Brin	–	Larry Page
Nick Cannon	–	Mariah Carey
Tod Williams	–	Billie Tsien
Josh Dallas	–	Ginnifer Goodwin
Eric Dane	–	Rebecca Gayheart

Top-15 connections in the IMDbNet

- famous co-actors
- romantic couples
- other domains

Quality Evaluations

	Algorithm	Votes	Accuracy
WikiNet	BF	800	0.969 ± 0.008
	decay a=0.0	800	0.990 ± 0.007
	decay a=0.01	800	0.983 ± 0.009
IMDbNet	BF	800	0.958 ± 0.014
	decay a=0.0	800	0.983 ± 0.009
	decay a=0.01	800	0.986 ± 0.008
Overall		4800	0.978 ± 0.004

Pairwise edge assessment
20 participants with over 5000 votes
Accuracy between 95% and 99%

Inter-rater agreement with 5 users for a subset of 200 edges:
- pairwise percent agreement: 93%-97%
- Fleiss' Kappa 0.83 – 0.97

Individual edge assessment
5 participants with over 600 votes

Average Rating on 5pt Likert scale:
- Connected nodes: 4.55
- Disconnected nodes 1.81

Pattern Analysis

- Iterative pattern extraction:
  - Language agnostic
  - Spam resistent
- Frequent patterns lead to promising queries
- Unfrequent patterns tend to be domain specific
  - Can be leveraged to identify communitiies/subgraphs

Pattern	Edges	Pairs	Domains
and	4,230	94	91
,	1,656	93	87
&	828	90	95
␣	806	91	89
und	1,124	79	71
;	54	33	25
:	36	27	24
-	36	23	20
or	33	19	14
/	28	22	11
with	20	15	15
on	23	15	14
y	16	14	13
and his wife	32	11	10
et	16	11	9
|	18	13	7
mit	14	12	9
e	14	10	9
vs.	13	10	9
and wife	25	7	7
+	9	8	7
left and	13	7	6
and actor	12	7	7
hat	8	7	8
Pictures Photo of	31	28	1

Social Network from Archives

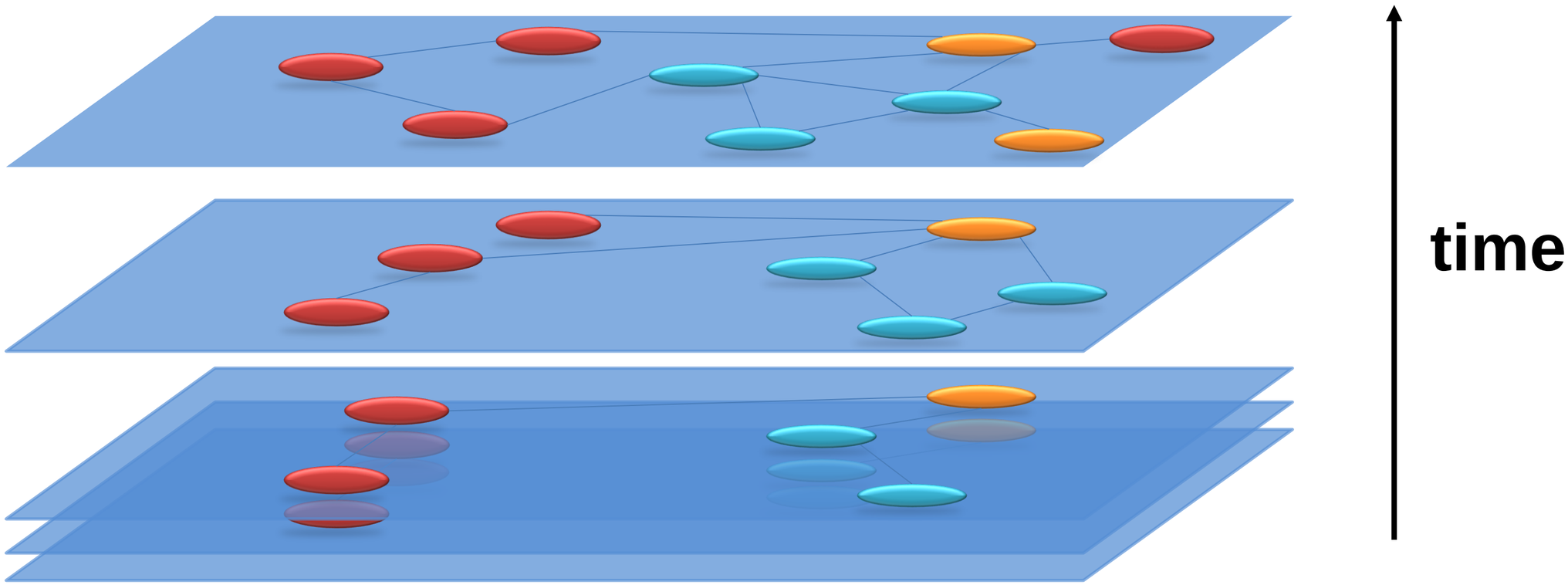
So far:

- Efficient and accurate methods for extracting social networks from unstructured Web data
  - Iterative approach for finding new query phrase patterns
  - Intelligent prioritization approaches for network expansion

Next Steps

- Understand relation development over time
  - Relationship type?
  - Bidirectional or unidirectional relationship?
  - etc.

Why study networks in Archives?



Temporal dynamics:

- new relationships
- new players
- evolution of communities
- evolution of relationship types

First Experiments and Statistics

- Data set:
  - About half of the internet archive text warc data (“/data/ia/w/de/TB/*.warc.gz”)
  - 52,695 WARC files, 360 Mio pages
  - Size: 5,6 TB
- Data Collections and Efficiency:
  - Sliding Window scan for identifying entities with high proximity
  - Time for parsing files: ~12h
  - Storing social graph information in DB and indexing: ~2h
- Graph size:
  - 587,235 nodes
  - 3,389,298 edges

Identifying pattern types

	"and her husband"	"kissed"	"dates"	"confers with"	"works with"	"has a meeting with"
Romeo --- Juliet		20	10	1		
Bonnie --- Clyde	20	56	5		5	
Jolie --- Pitt	10	10	10	10	10	10
Holmes --- Watson				45	78	67
Batman --- Robin				6	89	

Identifying pattern types

	"and her husband"	"kissed"	"dates"	"confers with"	"works with"	"has a meeting with"
Romeo --- Juliet		20	10	1		
Bonnie --- Clyde	20	56	5		5	
Jolie --- Pitt	10	10	10	10	10	10
Holmes --- Watson				45	78	67
Batman --- Robin				6	89	

LDA

Characterize relations by latent topics

Examples of Latent Topics

Topic 32	Topic 47	Topic 53	Topic 39	Topic 45	Topic 82
written_by	date	was_marri_to	beat	and_husband	and_presid
direct_by	girlfriend	divorc	replac	and_her_husband	and_prime_minist
writer	and_girlfriend	s_ex_wife	defeat	marri	presid
produc	and_boyfriend	affair	versus	husband	usa
and_written_by	boyfriend	marri	take_on	wed	prime_minist
director	is_date	kid	over	and_wife	and_former_presid
and_direct_by	and_his_girlfriend	and_his_ex_wife	to_replac	is_marri_to	and_us_presid
and_produc	and_her_boyfriend	latest	against	and_his_wife	season
and_director	s_girlfriend	s_ex	gegen	with_husband	and_opposit_leader
produc_by	dump	pic	fight	and_hubbi	and_foreign_minist

Examples of Latent Topics

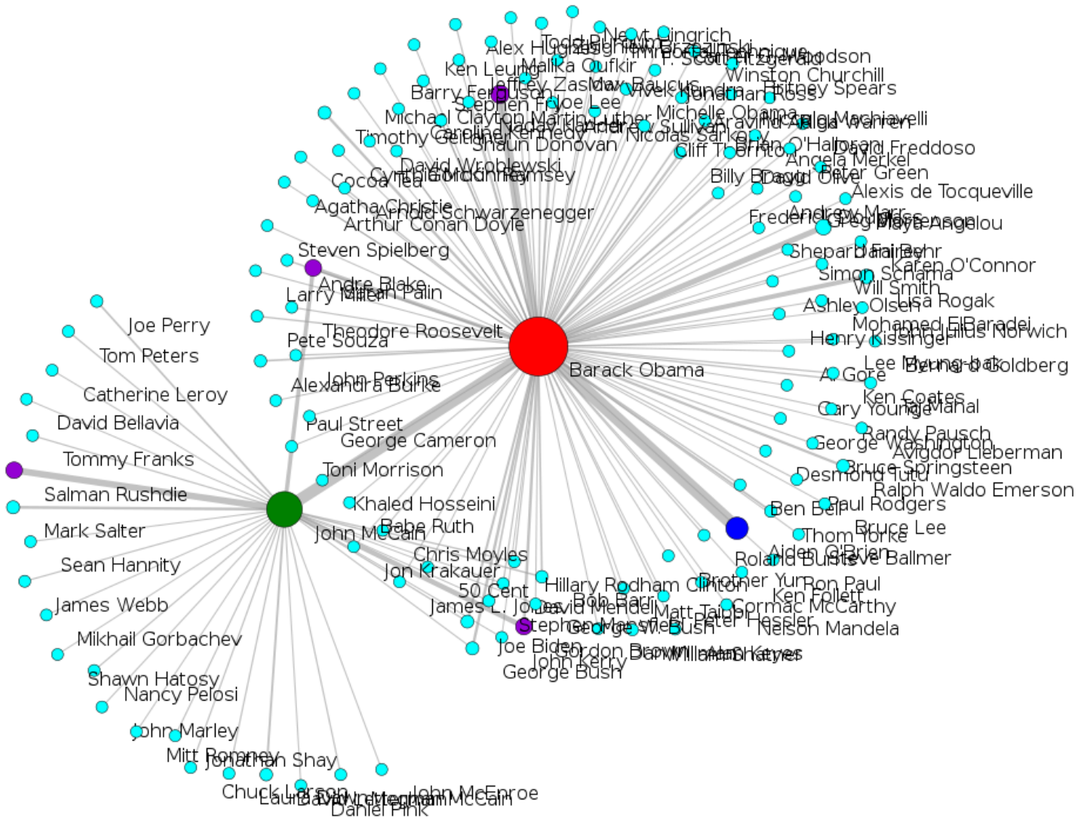
Movie making	Dating	Divorce	Competition	Marriage	Politics
Topic 32	**Topic 47**	**Topic 53**	**Topic 39**	**Topic 45**	**Topic 82**
written_by	date	was_marri_to	beat	and_husband	and_presid
direct_by	girlfriend	divorc	replac	and_her_husband	and_prime_minist
writer	and_girlfriend	s_ex_wife	defeat	marri	presid
produc	and_boyfriend	affair	versus	husband	usa
and_written_by	boyfriend	marri	take_on	wed	prime_minist
director	is_date	kid	over	and_wife	and_former_presid
and_direct_by	and_his_girlfriend	and_his_ex_wife	to_replac	is_marri_to	and_us_presid
and_produc	and_her_boyfriend	latest	against	and_his_wife	season
and_director	s_girlfriend	s_ex	gegen	with_husband	and_opposit_leader
produc_by	dump	pic	fight	and_hubbi	and_foreign_minist

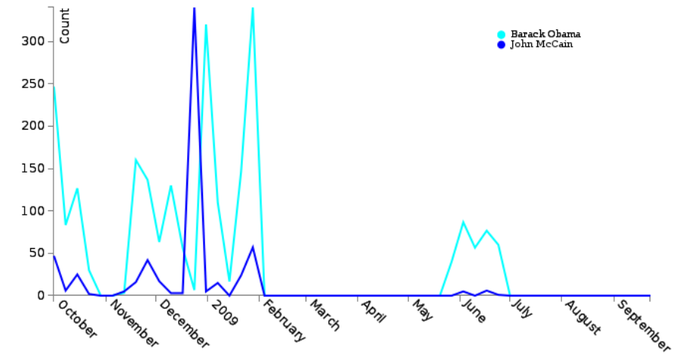
Example Connections and Relationship Types

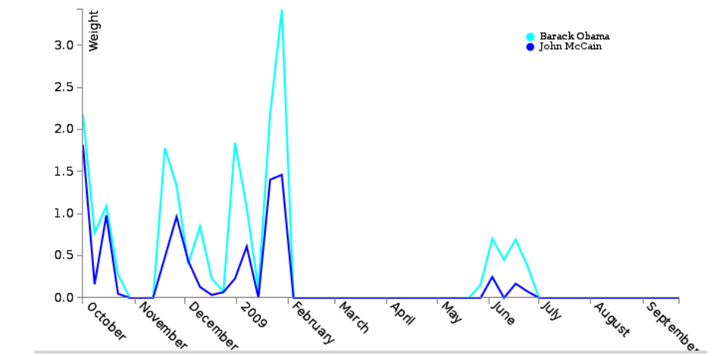
David Garrett	Andrea Berg	5922
Robert Pattinson	Kristen Stewart	5683
Katie Melua	Xavier Naidoo	4852
Brad Pitt	Angelina Jolie	4246
Andrea Berg	Helene Fischer	3351
Pietro Lombardi	Sarah Engels	2622
Angela Merkel	Nicolas Sarkozy	2454
Bud Spencer	Terence Hill	2366
Freddie Mercury	Helene Fischer	2295
Erwin Sellering	Lorenz Caffier	1886
Michael Schumacher	Sebastian Vettel	1834
Sebastian Vettel	Fernando Alonso	1750
Tom Cruise	Katie Holmes	1738
James Bond	Daniel Craig	1715
Angela Merkel	Guido Westerwelle	1675
Barack Obama	Mitt Romney	1667
Mark Webber	Sebastian Vettel	1657
Tommy Hilfiger	Ralph Lauren	1657

Pattern	Distinct pairs	Distinct domains
_	1332324	108489
und	193575	54480
oder	32753	15830
and	38077	11319
&	27557	10287
_	47963	1687
-	23692	4657
/	24366	3802
für	43863	804
mit	14498	4685
Jersey	27615	697
Darsteller:	16619	967
vs.	7672	1804
Regie:	12416	779
von	5357	2808
|	9298	1047
sowie	5033	2940

Demonstration: Temporal Analysis of Social Networks in Archives







240K distinct person names, 10Mio connections

Conclusion and Future Work

Conclusion

- Efficient and accurate methods for extracting social networks from unstructured Web data
  - Iterative approach for finding new query phrase patterns
  - Intelligent prioritization approaches for network expansion

Future work

- Efficient named entity recognition
- Temporal analysis of relationship development
- Entity disambiguation using social networks
- Involving of larger pair context
- Efficient community detection using domain specific pattern

Conclusion and Future Work

Conclusion

- Efficient and accurate methods for extracting social networks from unstructured Web data
  - Iterative approach for finding new query phrase patterns
  - Intelligent prioritization approaches for network expansion

Future work

- Efficient named entity recognition
- Temporal analysis of relationship development
- Entity disambiguation using social networks
- Involving of larger pair context
- Efficient community detection using domain specific pattern

Thank you!