So Far Away and Yet so Close: Augmenting Toponym Disambiguation and Similarity with Text-Based Networks

Andreas Spitz, Johanna Geiß and Michael Gertz

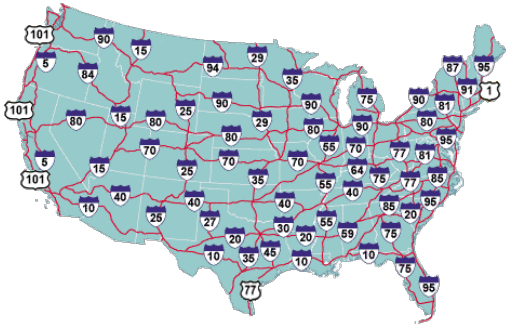
Heidelberg University, Institute of Computer Science
Database Systems Research Group, Heidelberg

{spitz, geiss, gertz}@informatik.uni-heidelberg.de

3rd GeoRich Workshop
San Francisco, June 26, 2016

Implicit Networks

← explicit →



social network

geographic network

http://web.mmlc.northwestern.edu/re-visualizing-the-novel/

← implicit →

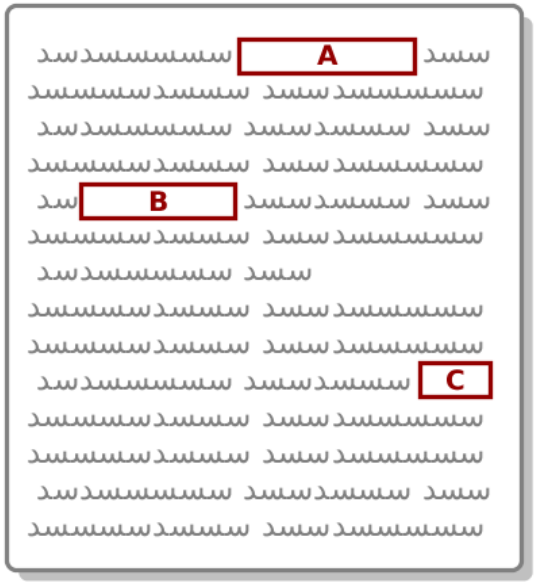
you are here

Implicit Text-Based Networks

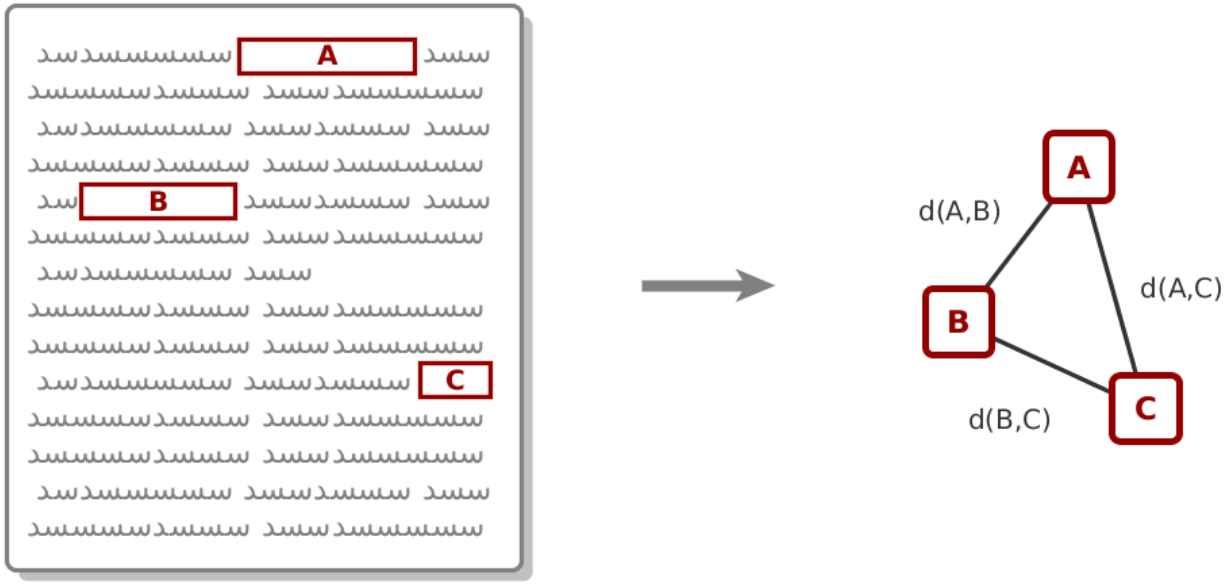
"Most of the circuits currently in use are specially constructed for competition. The current street circuits are **Monaco**, **Melbourne**, **Montreal**, **Singapore** *and* **Sochi**, *although races in other urban locations come and go (***Las Vegas** *and* **Detroit**, *for example) and proposals for such races are often discussed – most recently* **New Jersey**."

en.wikipedia.org/wiki/Formula_One

Graph Extraction from Text



Graph Extraction from Text



$s(v, w) :=$ distance in sentences between toponyms v and w

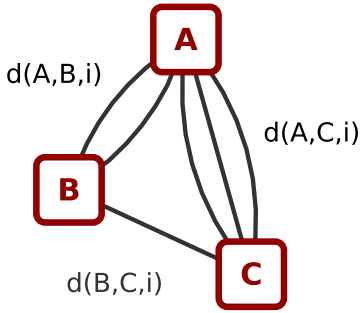
$$d(v, w) := \exp\left(-\frac{s(v, w)}{2}\right)$$

Graph Extraction from Text

$s(v, w) :=$ distance in sentences between toponyms v and w

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

Edge Aggregation



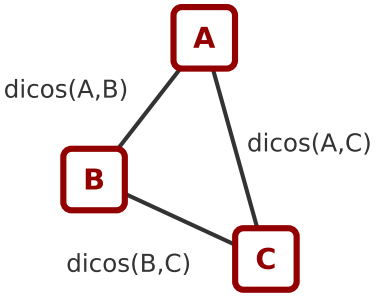
instances i

	1	2	3	4	5	6
A	d_1	d_2	d_3	d_4	d_5	0
B	d_1	d_2	0	0	0	d_6
C	0	0	d_3	d_4	d_5	d_6

dicos

Distance-based cosine for nodes v and w:

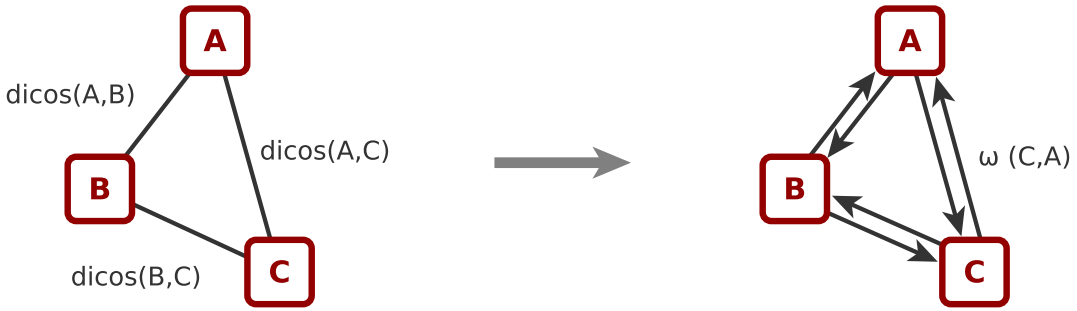
$$dicos(v,w) := \frac{\sum_i d_i(v)\ d_i(w)}{\sqrt{\sum_i d_i(v)^2}\sqrt{\sum_i d_i(w)^2}}$$



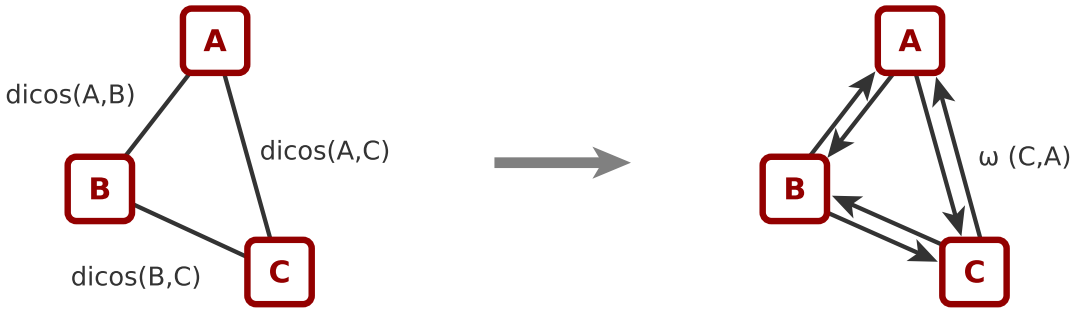
Nonreciprocal Relationships

Dirk Beyer, Wikimedia Commons

Inducing Edge Directions



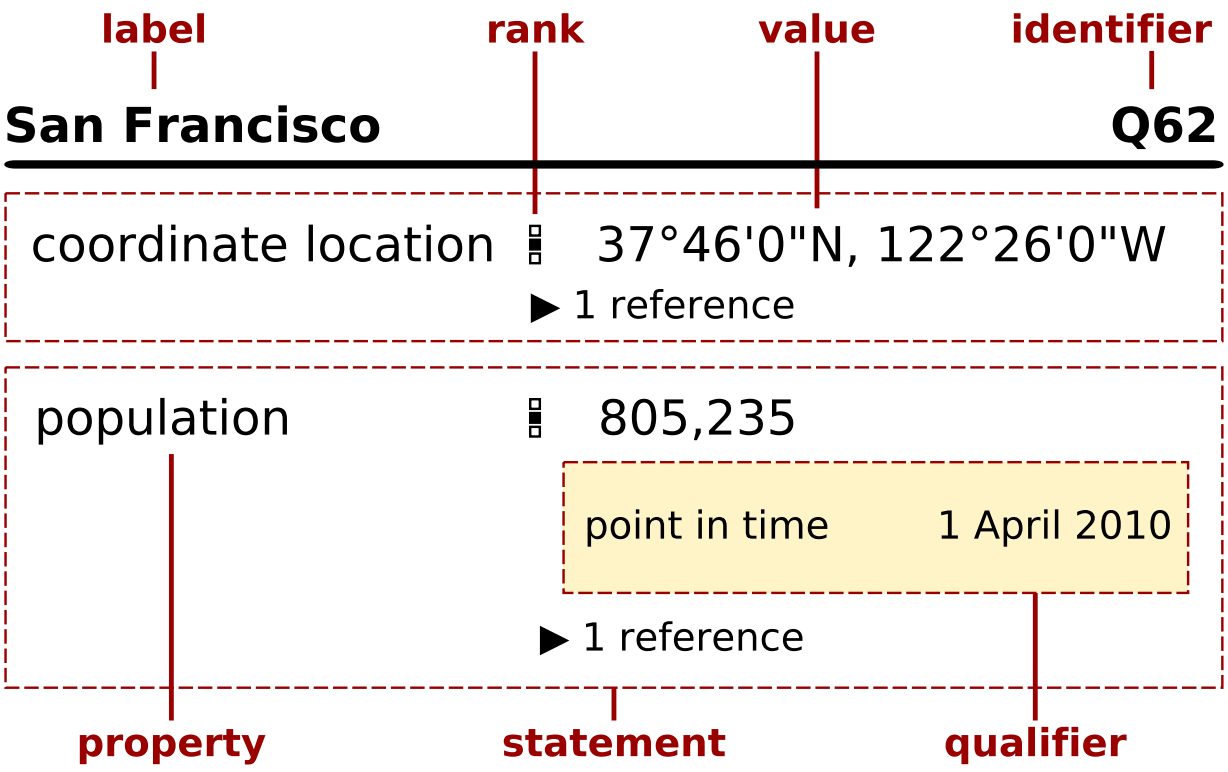
Inducing Edge Directions



Normalize weights of outgoing edges:

$$\omega(v \rightarrow w) := \frac{dicos(v,w)}{\sum_{x \in V} dicos(v,x)}$$

Adding Knowledge Base Support: Wikidata



Toponym Extraction in Wikipedia & Wikidata

Network Overview

Network statistics:

$\|V\|$	$\|E\|$	density	clustering coefficient
$723,779$	$178,890,238$	$6.8 \cdot 10^{-4}$	0.56

Node types:

Location types:

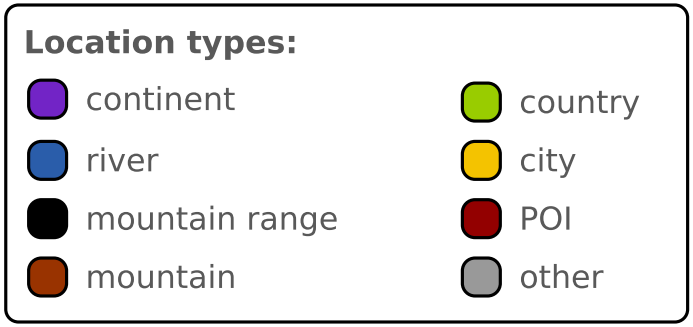
- continent
- river
- mountain range
- mountain
- country
- city
- POI
- other

Network Overview

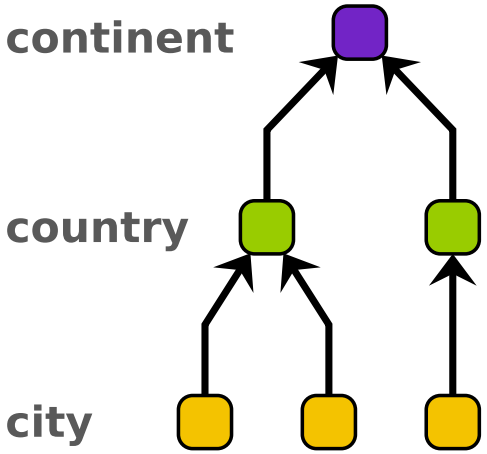
Network statistics:

$\|V\|$	$\|E\|$	density	clustering coefficient
$723,779$	$178,890,238$	$6.8 \cdot 10^{-4}$	0.56

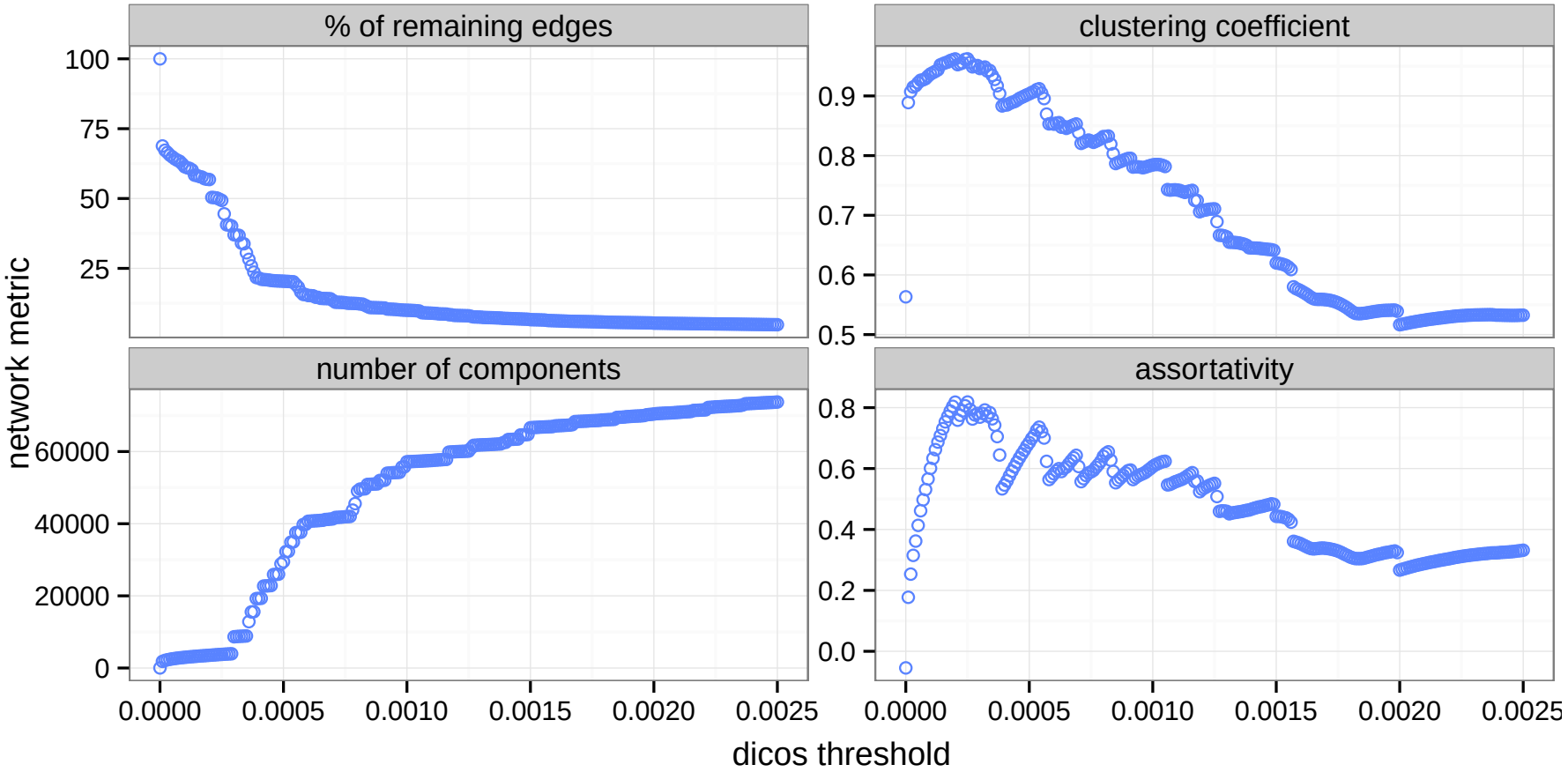
Node types:



Wikidata location hierarchy:



Network Properties

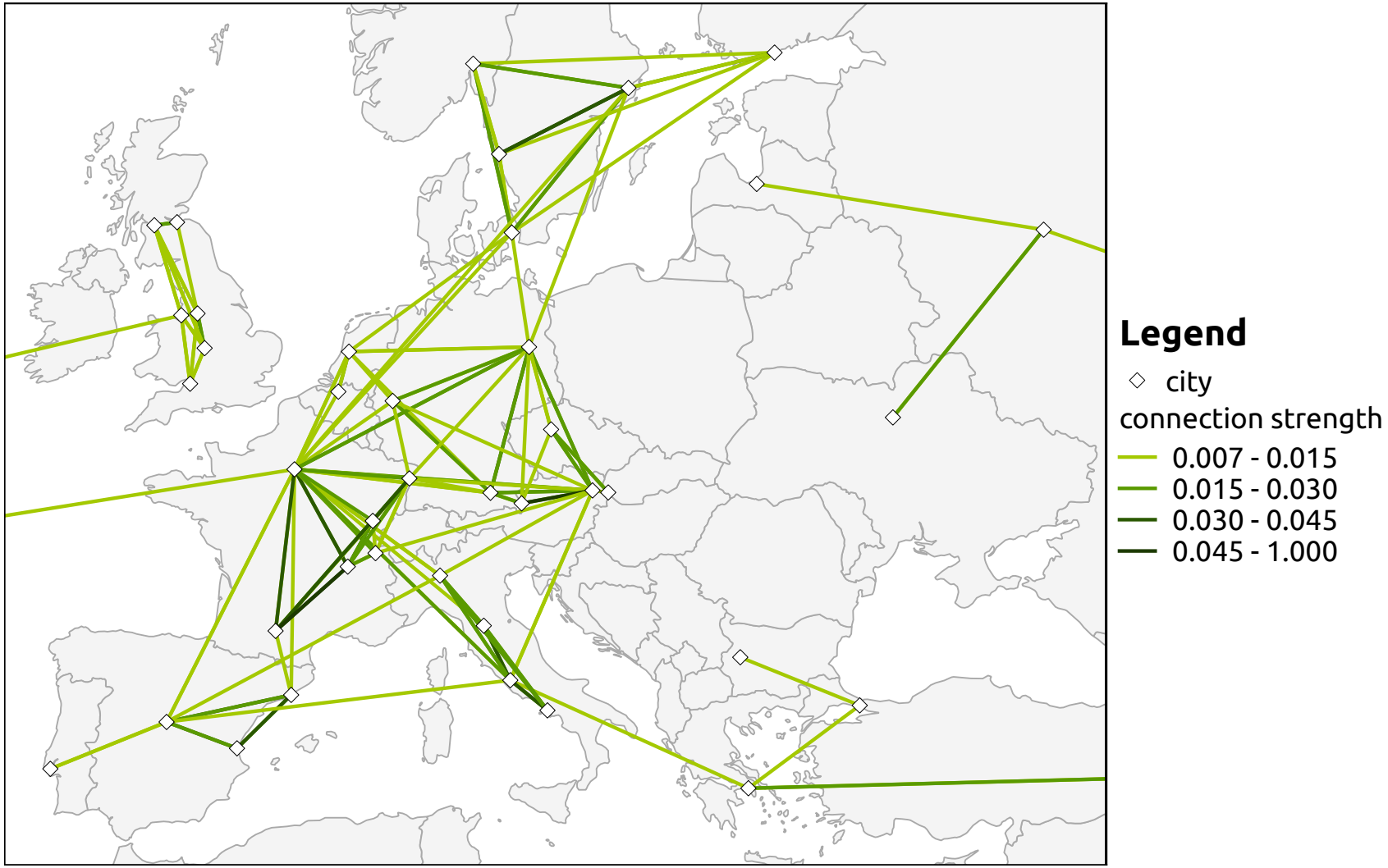


Network Centrality

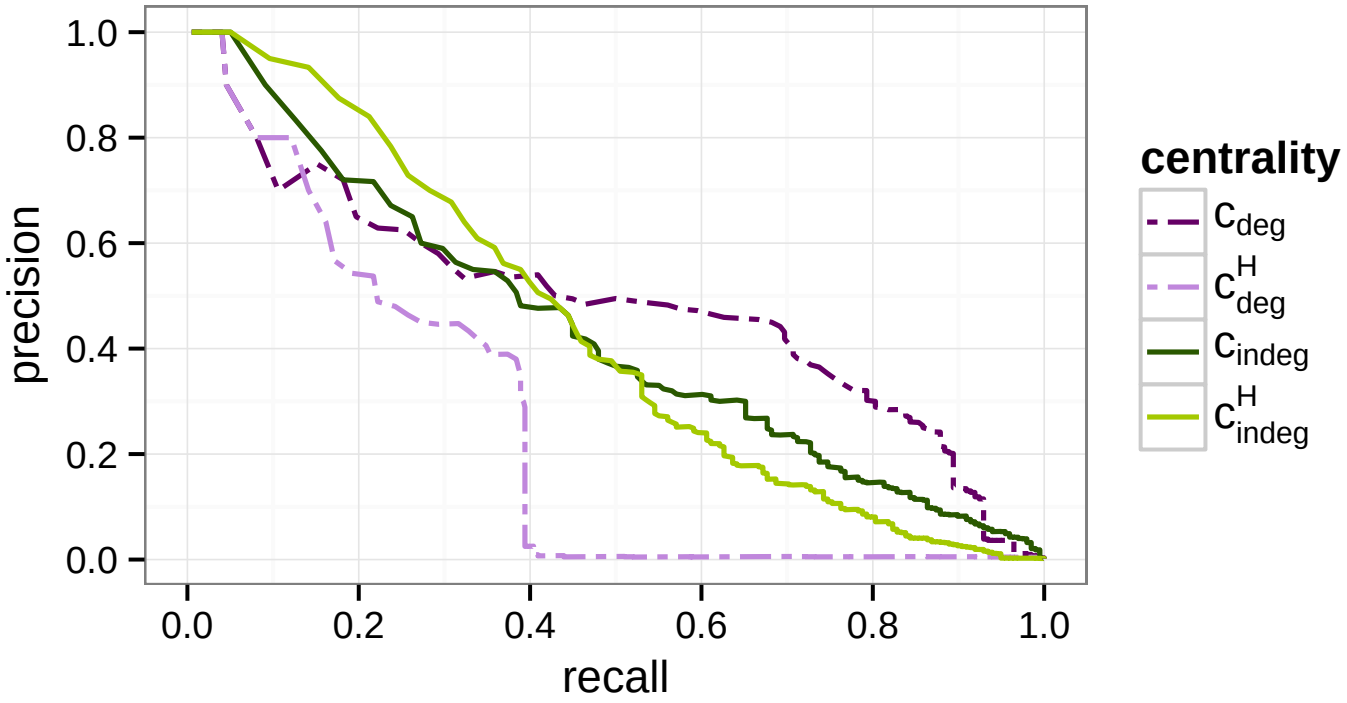
city	c_{deg}	c_{indeg}	c^H_{deg}	c^H_{indeg}
Paris	63,150	89.87	8,064	7.56
New York City	79,398	71.74	9,294	12.12
Chicago	54,217	51.84	8,074	7.70
Los Angeles	49,961	51.47	7,276	7.76
Washington, D.C.	62,858	51.05	8,138	8.65
Boston	45,895	50.43	6,121	6.08
Philadelphia	51,237	45.19	6,372	5.03
Vienna	35,724	44.55	4,827	7.44
Moscow	29,026	43.77	4,644	19.47
San Francisco	43,759	40.87	6,029	4.76

Network between the top 10 European cities by in-degree centrality.

Geographically Embedded Network



Centrality-Based Hierarchy Classification

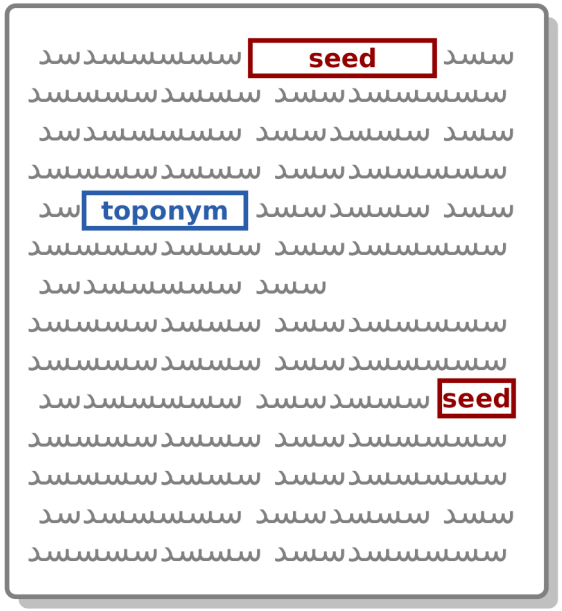


Classification into classes *country* and *city* based on centrality.

Disambiguation Problem

Locations of towns and cities with the name *Heidelberg*.

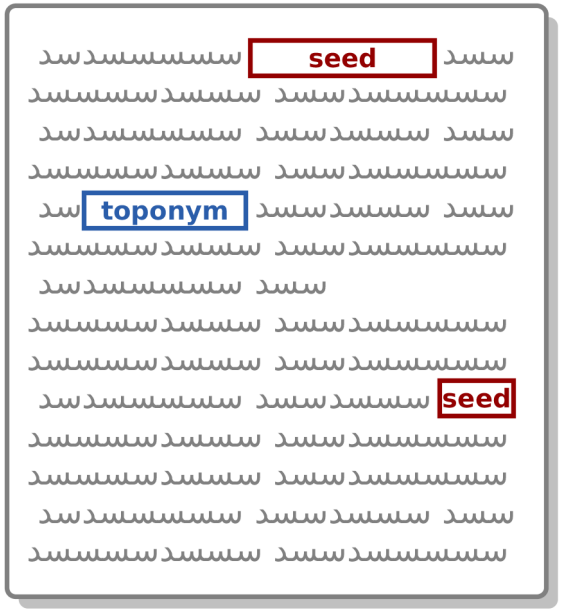
Network-based Toponym Disambiguation



Given a document with toponyms, the following information is available:

- a set of locations L in the network
- a set of seeds $S \subseteq L$ in the document (unambiguous toponyms)
- an ambiguous toponym t in the document with candidates $l \in L$

Network-based Toponym Disambiguation



Given a document with toponyms, the following information is available:

- a set of locations L in the network
- a set of seeds $S \subseteq L$ in the document (unambiguous toponyms)
- an ambiguous toponym t in the document with candidates $l \in L$

Resolve toponyms by their neighbourhood in the network:

$$\text{resolve}(t) := \arg\max_{l \in L} \sum_{s \in S} \omega(l, s)$$

Evaluation on AIDA CoNLL-YAGO data set

	Precision in %			mean distance in km		
	all	seeds	ambig.	all	seeds	ambig.
WLND	**85.7**	86.0	**85.6**	**327.5**	522.9	**179.1**
AIDA	84.9	86.0	83.2	*120.4*	*87.7*	*142.3*
B_{DIST}	81.6	86.0	78.5	683.1	522.9	800.8
B_{MIN}	81.4	86.0	78.8	650.9	522.9	745.0

WLDN Wikipedia Location Network disambiguation
AIDA AIDA named entity disambiguation
B_{DIST} Baseline using minimum geographic distance
B_{MIN} Baseline using lowest Wikidata ID

Summary

New method for implicit network extraction that

- is based on text distances of toponyms,
- works across documents,
- can be applied to any geo-tagged corpus.

Application to Wikipedia & Wikidata

- creates an accurate and reliable network,
- supports disambiguation and entity linking,
- provides a *language-agnostic* tool for NLP tasks

The Wikipedia Location Network
is available for download.

http://dbs.ifi.uni-heidelberg.de/index.php?id=data

The Wikipedia Location Network
is available for download.

http://dbs.ifi.uni-heidelberg.de/index.php?id=data

Thank you!
Questions?

Bibliography

Johanna Geiß and Michael Gertz.
With a Little Help from my Neighbors: Person Name Linking Using the Wikipedia Social Network.
In *WWW Companion*, 2016.

Johanna Geiß, Andreas Spitz, Jannik Strötgen, and Michael Gertz.
The Wikipedia Location Network - Overcoming Borders and Oceans.
In *GIR*, 2015.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum.
Robust Disambiguation of Named Entities in Text.
In *EMNLP*, 2011.

Michael Speriosu and Jason Baldridge.
Text-Driven Toponym Resolution using Indirect Supervision.
In *ACL*, 2013.