

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

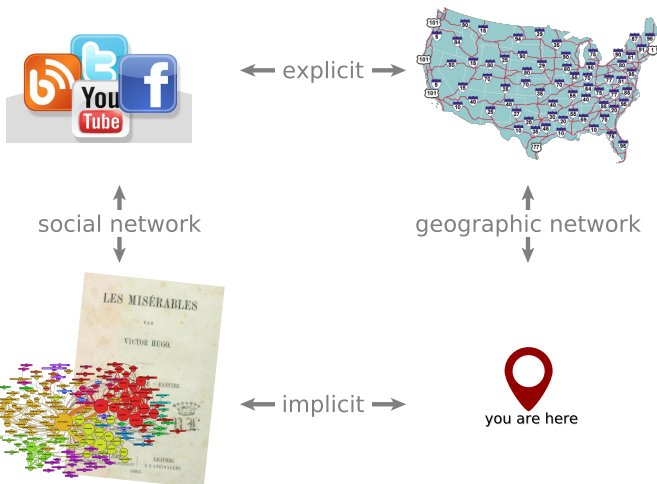
**Andreas Spitz, Johanna Geiß and Michael Gertz**

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



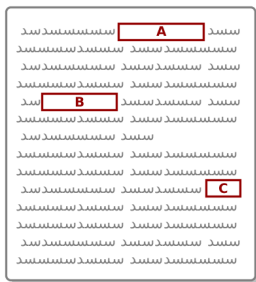
<http://web.mmlc.northwestern.edu/re-visualizing-the-novel/>

# Implicit Text-Based Networks

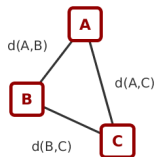
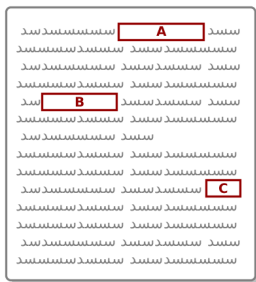
*“Most of the circuits currently in use are specially constructed for competition. The current street circuits are **Monaco, Melbourne, Montreal, Singapore and Sochi**, although races in other urban locations come and go (**Las Vegas and Detroit**, for example) and proposals for such races are often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

[en.wikipedia.org/wiki/Formula\\_One](https://en.wikipedia.org/wiki/Formula_One)

# Graph Extraction from Text



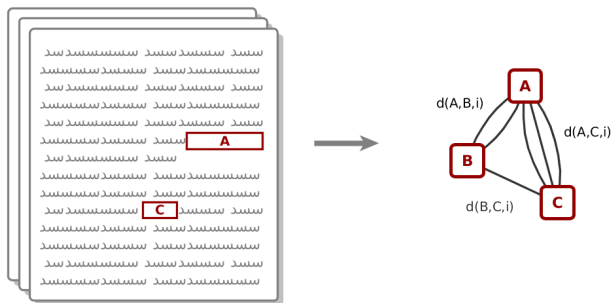
# Graph Extraction from Text



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

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

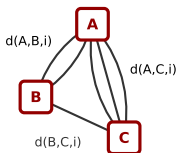
# Graph Extraction from Text



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

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

# Edge Aggregation

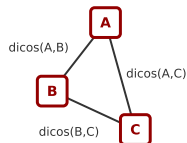


	instances $i$					
	1	2	3	4	5	6
A	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	0
B	$d_1$	$d_2$	0	0	0	$d_6$
C	0	0	$d_3$	$d_4$	$d_5$	$d_6$



Distance-based cosine for nodes  $v$  and  $w$ :

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



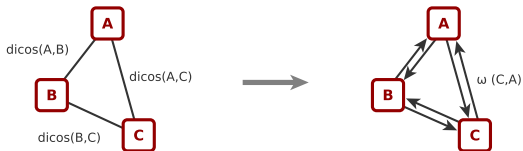
# Nonreciprocal Relationships



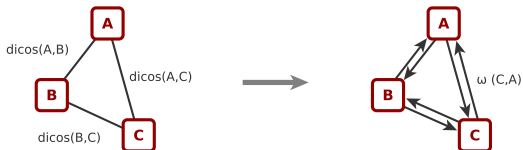
Dirk Beyer, Wikimedia Commons



# Inducing Edge Directions



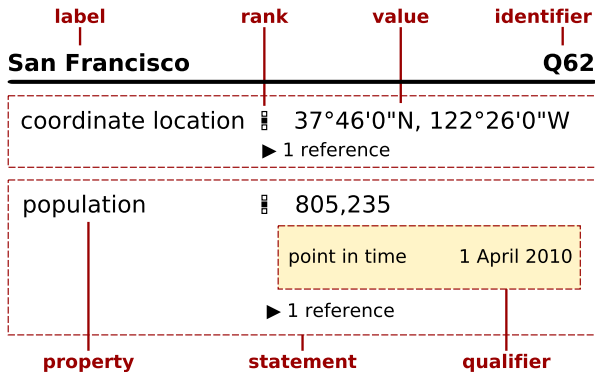
# Inducing Edge Directions



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



**WIKIPEDIA**  
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikipedia store

Interaction

- Help

Article Talk

## 1984 Summer Olympics

From Wikipedia, the free encyclopedia

The **1984 Summer Olympics**, officially known as the **Games of the XXIII Olympiad**, was an international multi-sport event held in **Los Angeles, California, United States** in 1984. When **Tehran**, the only other interested city on the international level, declined to bid due to the concurrent Iranian political and social changes, the IOC awarded Los Angeles the Games for the second occasion Los Angeles hosted the Olympics, having first hosted in 1932.

### Los Angeles

From Wikipedia, the free encyclopedia

Coordinates: 34°03′N 118°15′W

This article is about the U.S. city. For the county in California, see *Los Angeles County*. For other uses, see *Los Angeles (disambiguation)*.  
 "L.A." and "City of Los Angeles" redirect here. For other uses, see *L.A. (disambiguation)* and *City of Los Angeles (disambiguation)*.

**Los Angeles** (/loʊˈlɑːndʒələs/ *and* /lɑːndʒələs/ or /loʊˈʒɑːləs/ *lōʊ-ˈʒɑːləs*)<sup>[k]</sup> officially the **City of Los Angeles** and often known by its initials **L.A.**, is the second-largest city in the United States, the most populous city in the U.S. state of California, and the county seat of Los Angeles County. Situated in Southern California, Los Angeles is known for its mediterranean climate, ethnic diversity, sprawling metropolis, and as a major center of the American entertainment industry. Los Angeles lies in a large coastal basin surrounded on three sides by mountains reaching up to and over 10,000 feet (3,000 m).

#### Los Angeles, California

Charter city

City of Los Angeles



### Tehran

From Wikipedia, the free encyclopedia

Coordinates: 35°41′46″N 51°29′23″E

This article is about the Iranian capital city. For other uses, see *Tehran (disambiguation)*.

**Tehran** (/tɛhˈrɑːn/ pronunciation تهران; Persian: تهران - Tehrān) is the capital of Iran and Tehran Province. With a population of around 9 million in the city and 16 million in the wider metropolitan area,<sup>[k]</sup> Tehran is the largest city and urban area of Iran, the second largest city in Western Asia, and the 3rd largest in the Middle East. It is ranked 29th in the world by the population of its metropolitan area.<sup>[k]</sup>

In the Classical era, part of the present-day city of Tehran was occupied by a Median city which in the *Avesta* occurs as *Rhaga*.<sup>[j]</sup> It was destroyed by the Mongols in the early 13th century, and remains now as a city in Tehran Province, located towards the south end of the modern-day city of Tehran.

#### Tehran

تهران

Metropolis

کلانشهر تهران · Tehran Metropolis













# 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:**

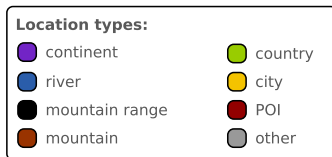
 continent	 country
 river	 city
 mountain range	 POI
 mountain	 other

# Network Overview

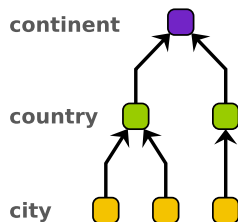
Network statistics:

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

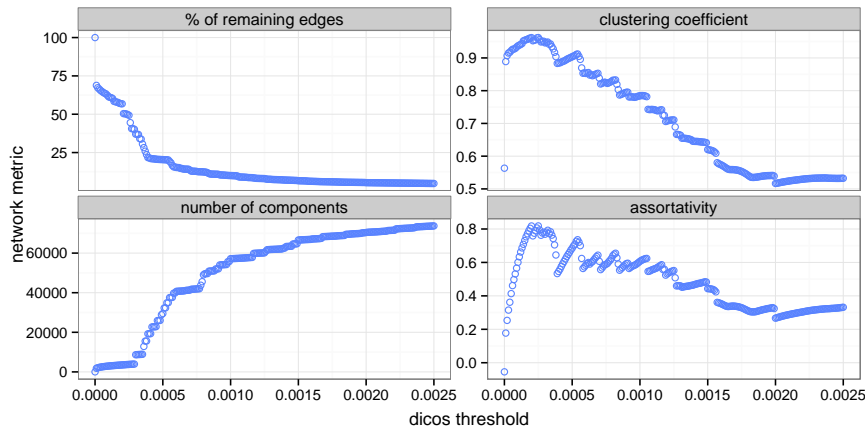
Node types:



Wikidata location hierarchy:



# Network Properties



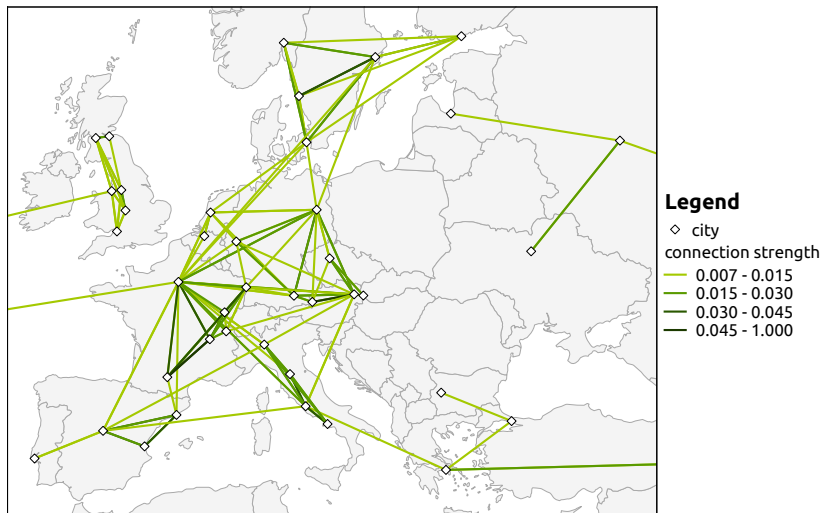
# Network Centrality

city	$C_{deg}$	$C_{indeg}$	$C_{deg}^H$	$C_{indeg}^H$
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

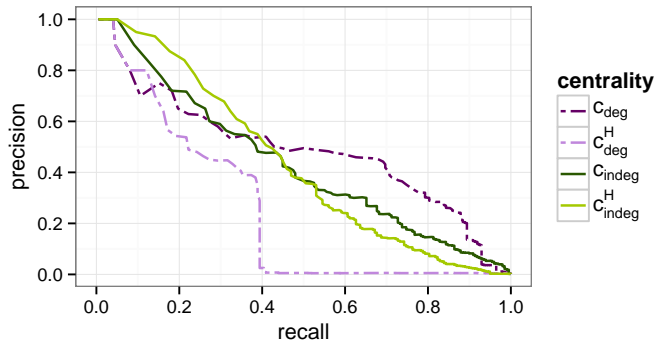
Network between the top 10 European cities by in-degree centrality.



# Geographically Embedded Network



# Centrality-Based Hierarchy Classification



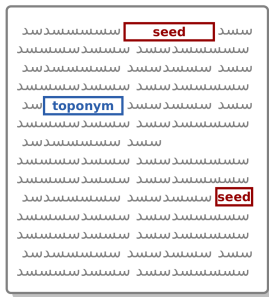
Classification into classes *country* and *city* based on centrality.

# Disambiguation Problem



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

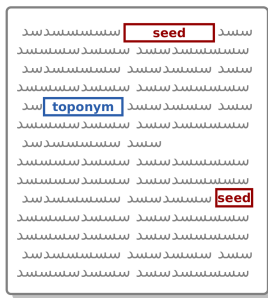
# Network-based Toponym Disambiguation



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	<b>85.7</b>	86.0	<b>85.6</b>	<b>327.5</b>	522.9	<b>179.1</b>
AIDA	84.9	86.0	83.2	120.4	87.7	142.3
B <sub>DIST</sub>	81.6	86.0	78.5	683.1	522.9	800.8
B <sub>MIN</sub>	81.4	86.0	78.8	650.9	522.9	745.0

WLDN Wikipedia Location Network disambiguation

AIDA AIDA named entity disambiguation

B<sub>DIST</sub> Baseline using minimum geographic distance

B<sub>MIN</sub> 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.