UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

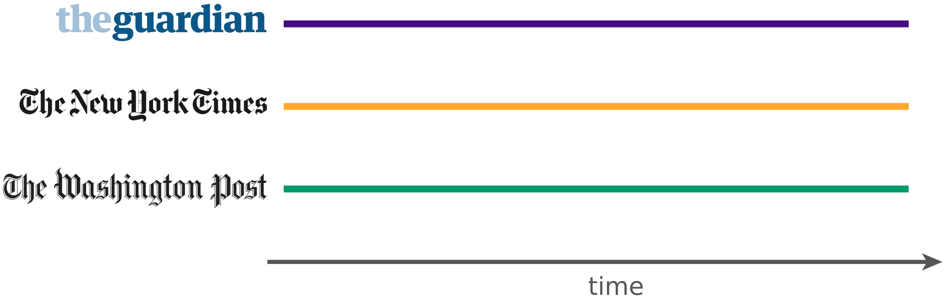
Exploring Entity-centric Networks in Entangled News Streams

Andreas Spitz and Michael Gertz

April 25, 2018 — WWW 2018, Lyon

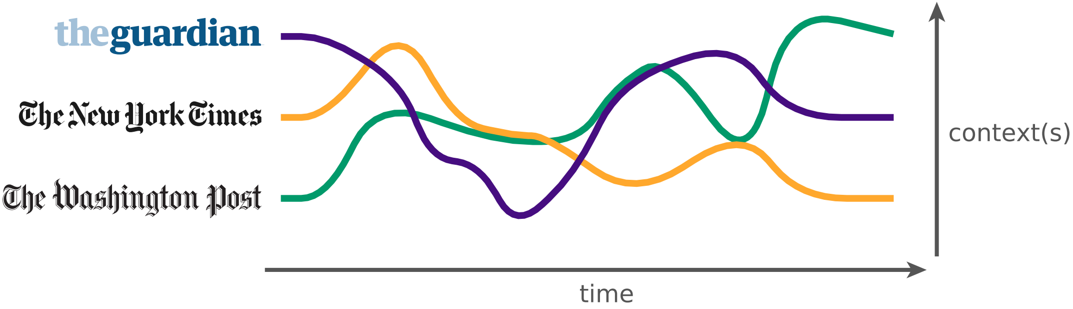
Heidelberg University, Germany
Database Systems Research Group

Parallel News Streams

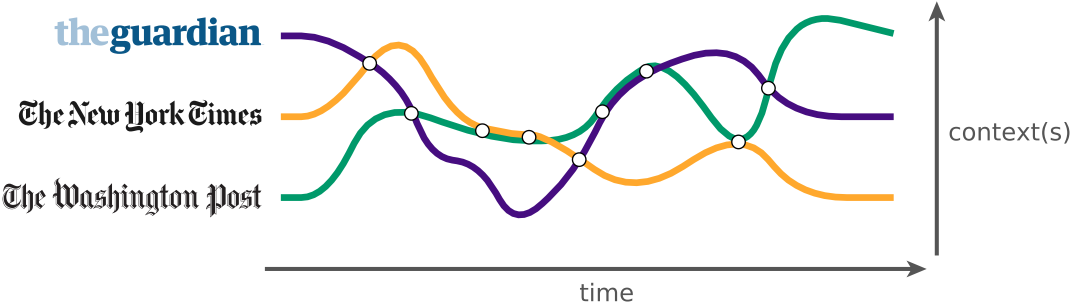


Crossing Streams

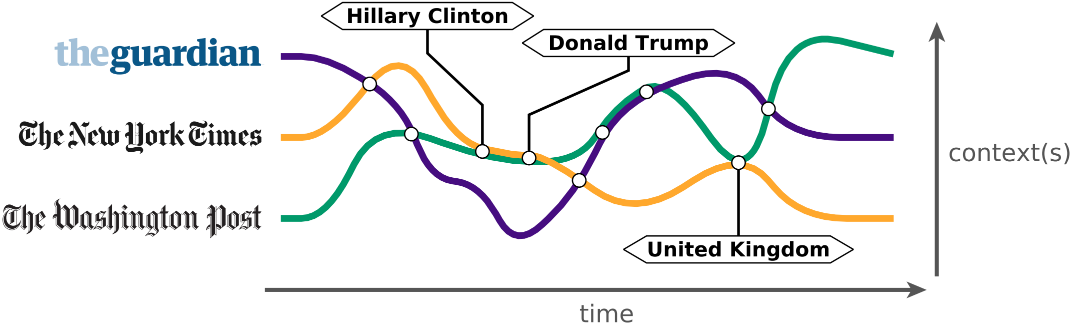
Entangled News Streams



Entangled News Streams



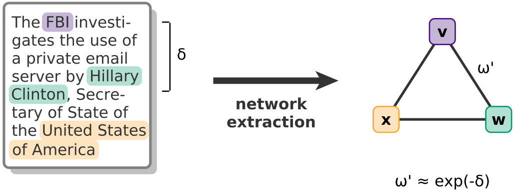
Entangled News Streams



Core idea: entity cooccurrences characterize stitching points between news streams

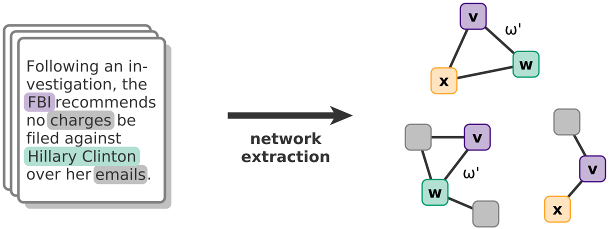
Implicit Entity Networks

Implicit Network Extraction


The FBI investigates the use of a private email server by Hillary Clinton, Secretary of State of the United States of America
δ
network extraction
v
x
w
ω'
ω' ≈ exp(-δ)

Andreas Spitz and Michael Gertz. "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events". In: *SIGIR*. 2016

Implicit Network Aggregation

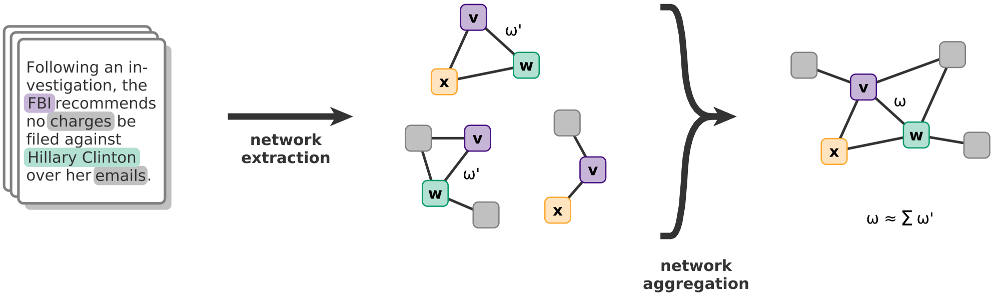


Following an investigation, the FBI recommends no charges be filed against Hillary Clinton over her emails.

network extraction

Andreas Spitz and Michael Gertz. “Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events”. In: *SIGIR*. 2016

Implicit Network Aggregation

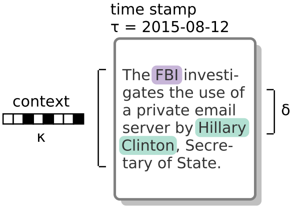


Following an investigation, the FBI recommends no charges be filed against Hillary Clinton over her emails.

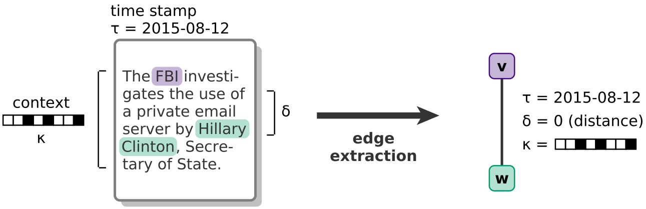
Andreas Spitz and Michael Gertz. "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events". In: *SIGIR*. 2016

Implicit Networks of Text Streams

Edge Context Extraction

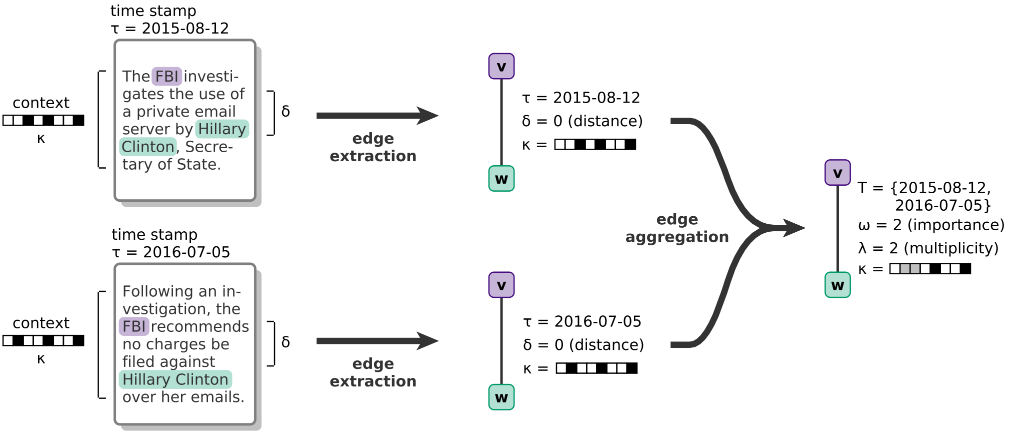

time stamp
τ = 2015-08-12
context
κ
δ
The FBI investigates the use of a private email server by Hillary Clinton, Secretary of State.

Edge Context Extraction



time stamp
τ = 2015-08-12
context
κ
The FBI investigates the use of a private email server by Hillary Clinton, Secretary of State.
δ
edge extraction
v
w
τ = 2015-08-12
δ = 0 (distance)
κ =

Context-based Aggregation of Edges



The FBI investigates the use of a private email server by Hillary Clinton, Secretary of State.

Following an investigation, the FBI recommends no charges be filed against Hillary Clinton over her emails.

Edge Aggregation Approaches

Streaming aggregation:

Static aggregation / clustering:

Edge Aggregation Approaches

Streaming aggregation:

- Compare similarity of new edge $(v, w, \cdot)$ to existing edges $(v, w, \cdot)$
- If similarity threshold is exceeded: merge with existing edge
- Otherwise, insert as new parallel edge

Static aggregation / clustering:

Edge Aggregation Approaches

Streaming aggregation:

- Compare similarity of new edge $(v, w, \cdot)$ to existing edges $(v, w, \cdot)$
- If similarity threshold is exceeded: merge with existing edge
- Otherwise, insert as new parallel edge

Static aggregation / clustering:

- Collect all parallel edges
- Cluster parallel edges (density-based)
- Discard “noisy” edges
- aggregate edges within clusters

Application Examples

News Article Data

English news articles from RSS feeds:

- 14 news outlets (from US, UK, and AU)
- 6 months (Jun 1 - Nov 30, 2016)
- 127.5 thousand articles
- 5.4 million sentences

News Article Data

English news articles from RSS feeds:

- 14 news outlets (from US, UK, and AU)
- 6 months (Jun 1 - Nov 30, 2016)
- 127.5 thousand articles
- 5.4 million sentences

NLP processing pipeline:

- Part-of-speech and sentence tagging:
  Stanford POS tagger
- Temporal tagging: HeidelTime
- Entity classification:
  YAGO classes (LOC, ORG, PER)
- Named entity recognition and linking:

AMBIVERSE
Text to Knowledge

News Article Data

English news articles from RSS feeds:

- 14 news outlets (from US, UK, and AU)
- 6 months (Jun 1 - Nov 30, 2016)
- 127.5 thousand articles
- 5.4 million sentences

The resulting implicit network has

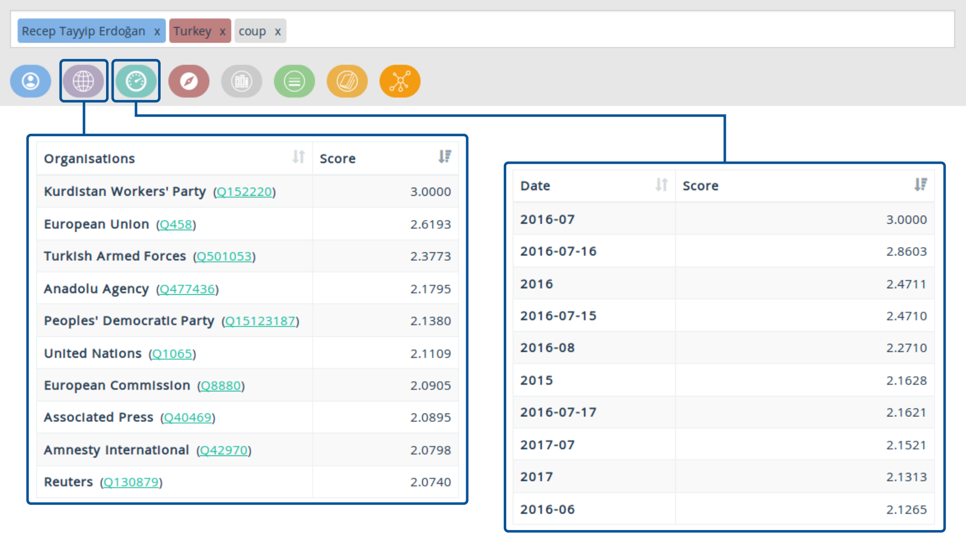
- 125 thousand entities
- 351 thousand terms
- 83.4 million edges

NLP processing pipeline:

- Part-of-speech and sentence tagging:
  Stanford POS tagger
- Temporal tagging: HeidelTime
- Entity classification:
  YAGO classes (LOC, ORG, PER)
- Named entity recognition and linking:

AMBIVERSE
Text to Knowledge

Context Sensitive Entity Search



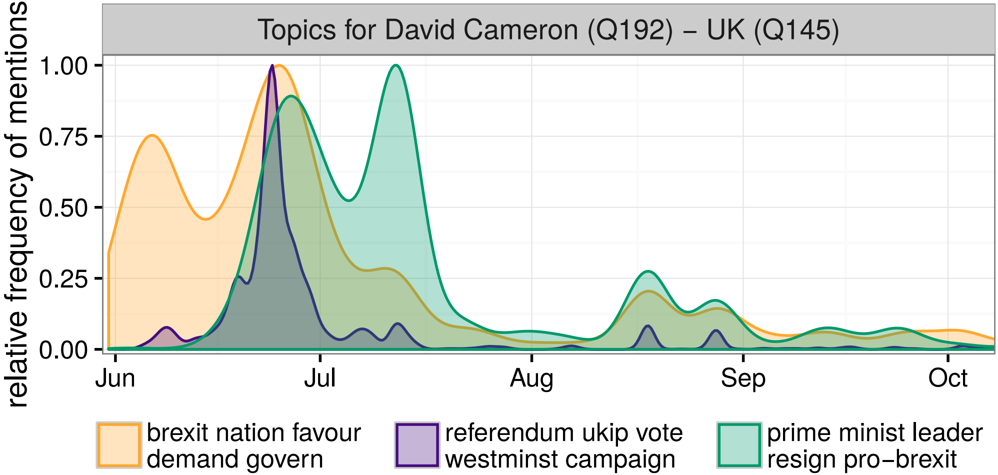
Recep Tayyip Erdoğan x Turkey x coup x

Organisations	Score
Kurdistan Workers' Party (Q152220)	3.0000
European Union (Q458)	2.6193
Turkish Armed Forces (Q501053)	2.3773
Anadolu Agency (Q477436)	2.1795
Peoples' Democratic Party (Q15123187)	2.1380
United Nations (Q1065)	2.1109
European Commission (Q8880)	2.0905
Associated Press (Q40469)	2.0895
Amnesty International (Q42970)	2.0798
Reuters (Q130879)	2.0740

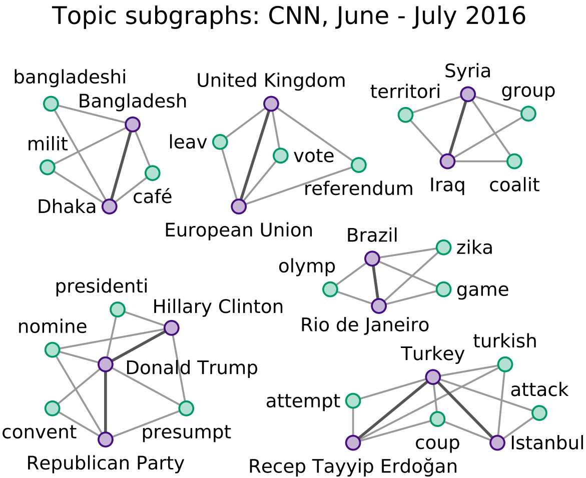
Date	Score
2016-07	3.0000
2016-07-16	2.8603
2016	2.4711
2016-07-15	2.4710
2016-08	2.2710
2015	2.1628
2016-07-17	2.1621
2017-07	2.1521
2017	2.1313
2016-06	2.1265

A. Spitz, S. Almasian, and M. Gertz. "EVELIN: Exploration of Event and Entity Links in Implicit Networks". In: *WWW Companion.* 2017. URL: http://evelin.ifi.uni-heidelberg.de

Evolution of Entity Contexts



Topic Subgraph Exploration



Andreas Spitz and Michael Gertz. "Entity-Centric Topic Extraction and Exploration: A Network-Based Approach". In: *ECIR*. 2018

Further Applications

News analysis and exploration:

- Contrastive source comparison
- Coverage bias
- Evolution of news stories
- Event description
- ...

Further Applications

News analysis and exploration:

- Contrastive source comparison
- Coverage bias
- Evolution of news stories
- Event description
- ...

NLP and IR applications:

- Entity disambiguation
- (Extractive) summarization
- Relationship extraction
- ...

Resources

Resources

Data and implementation are available online:

- [data] Implicit news stream network
- [code] Implicit network extraction
- [code] Entity query and topic extraction

`https://dbs.ifi.uni-heidelberg.de/resources/newsstream/`

Resources

Data and implementation are available online:

- [data] Implicit news stream network
- [code] Implicit network extraction
- [code] Entity query and topic extraction

`https://dbs.ifi.uni-heidelberg.de/resources/newsstream/`