UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Predicting Document Creation Times in News Citation Networks

Andreas Spitz[1], Jannik Strötgen[2], and Michael Gertz[1]

April 23, 2018 — TempWeb 2018, Lyon

[1] Database Systems Research Group
Heidelberg University, Germany

[2] Bosch Center for Artificial Intelligence
Germany

Hm, when did this happen again?

Support The Guardian | Subscribe | Find a job | Sign in | Search

International edition

The Guardian

News | Opinion | Sport | Culture | Lifestyle | More

World ▸ Europe US Americas Asia Australia Middle East Africa Inequality Cities Global development

David Cameron

David Cameron resigns after UK votes to leave European Union

PM announces resignation following victory for leave supporters after divisive referendum campaign

EU referendum fallout - live

Heather Stewart, Rowena Mason and Rajeev Syal

Fri 24 Jun 2016 10.44 BST

This article is over 1 year old

<24,779

▲ David Cameron: a political obituary

David Cameron has resigned, bringing an abrupt end to his six-year premiership, after the British public took the momentous decision to reject his entreaties and turn their back on the European Union.

Just a year after he clinched a surprise majority in the general election, a visibly emotional Cameron, standing outside Number 10 on Friday morning alongside his wife, Samantha, said: "The will of the British people is an instruction that must be delivered."

The prime minister campaigned hard in the divisive referendum on Britain's relationship with the EU, appearing at hundreds of public events up and down the country to argue that Brexit would be an act of "economic self-harm".

How do you choose a new prime minister?

The prime minister is notionally picked by the Crown, but in practice the monarch has – for centuries – been obliged to pick the party leader who can command the support of most MPs in the House of Commons.

In the distant past this was established by taking informal soundings, but the Labour party from its inception elected its leader by a ballot of its ... Show more

most viewed


Arsène Wenger to leave Arsenal at end of the season


Jacinda Ardern wears Māori cloak to Buckingham Palace


Yanis Varoufakis: Marx predicted our present crisis - and points the way out


Brexit divorce bill will surpass £39bn, warns Whitehall watchdog


Mystery of sea nomads' amazing ability to freedive is solved

News Citation Networks

News Citation Network Extraction

navigational links

Sections

The Washington Post
Democracy Dies in Darkness

Sign In | Subscribe

Politics • Analysis

If Rosenstein is fired, this may be the timeline used to rationalize it

By Philip Bump April 19 at 10:52 AM | Email the author

Deputy Attorney General Rod J. Rosenstein. (Brynn Anderson/AP)

anchored references

There are three possible explanations for the ongoing tension between congressional investigators and the Department of Justice.

The first is that the investigators, led by Rep. Devin Nunes (R-Calif.), are using Justice's reluctance to share classified documents as a means of undercutting the investigation into Russian interference in the 2016 election and, more specifically, to cast Deputy Attorney General Rod J. Rosenstein as a bad actor to facilitate his firing. Rosenstein, as you may be aware, both appointed special counsel Robert S. Mueller III and has sole authority over Mueller and his investigation; ousting Rosenstein could severely hamper Mueller's probe.

advertisements

internal links

Most Read Politics

1 Trump hires Giuliani, two other attorneys amid mounting legal turmoil over Russia

2 Analysis
The first domino just fell after the Michael Cohen raid

3 Giuliani says he is joining Trump's legal team to 'negotiate an end' to Mueller probe

4 Analysis
What the Comey memos say

5 'He knows how to read a room really, really well': How White House physician Rony L. Jackson became Trump's nominee to lead VA

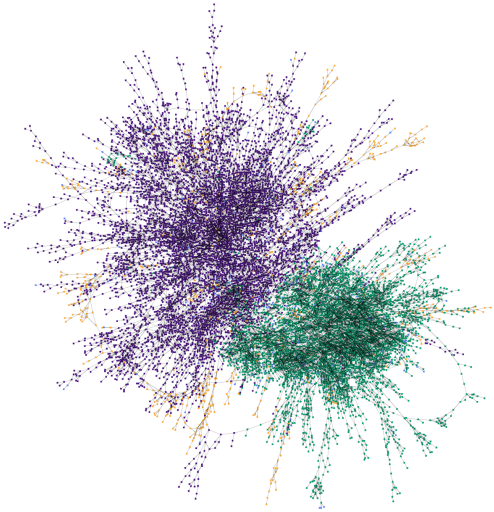
News Citation Network Overview

News articles from RSS feeds:

- Politics and business feeds
- 34 English news outlets
  (USA, UK, AUS, CAN, GER, CHN, QAT)
- 2 years (Nov 2015 - Oct 2017)
- 244.6 thousand articles
- 367.2 thousand edges

Used data:

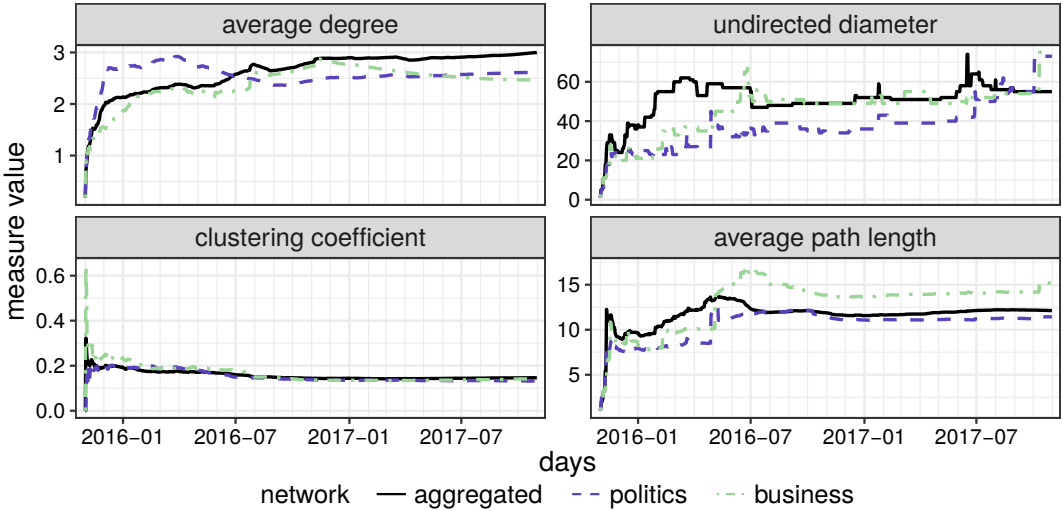
- Hyperlinks in the article body
- Publication dates
- Temporal expressions



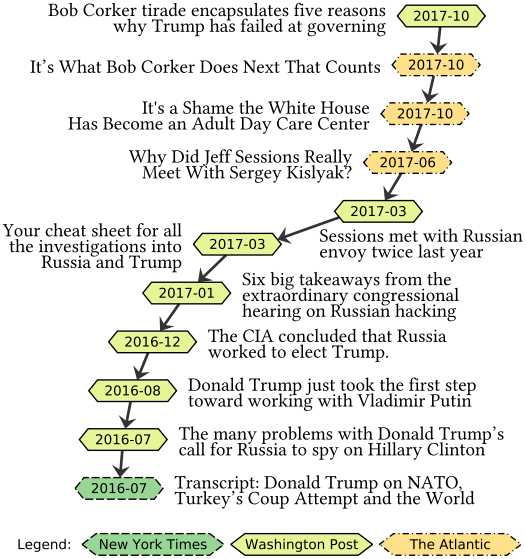
News Outlet Statistics (sample)

short	news outlet	*days*	$\langle articles \rangle$	$\langle temp\ exp \rangle$	$other_{in}$	$other_{out}$
AT	The Atlantic	334	7.2	10.5	16.7	50.6
BBC	British Bc. Corp.	730	8.1	6.5	19.1	8.0
DW	Deutsche Welle	334	1.2	6.1	48.1	5.9
FOX	Fox News	548	2.7	9.8	0.0	0.0
NPR	National Public Radio	334	0.4	8.4	63.6	58.5
NY	The New Yorker	548	3.0	13.2	33.5	30.6
NYT	New York Times	669	23.8	10.7	26.8	4.7
SMH	Sydney Morn. Herald	548	2.3	7.0	3.0	51.9
WP	Washington Post	548	62.7	9.4	13.7	5.1

Evolution of Network Metrics

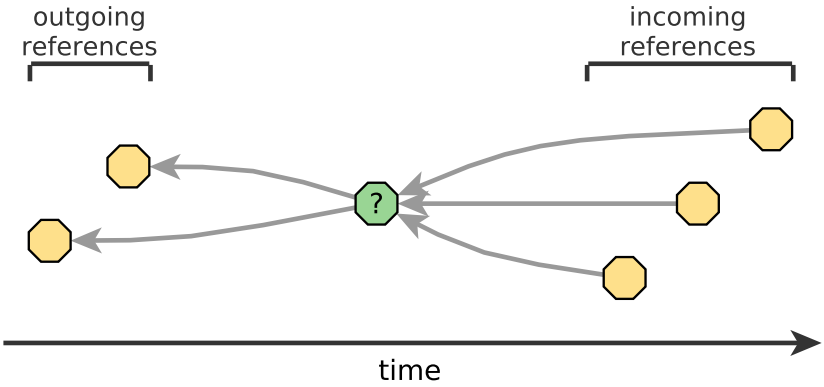


Exploring Citation Chains



Article Publication Time Prediction

Task Definition: Publication Time Prediction



Available News Citation Network Data

Predict article publication times from:

- Citation network topology
- Publication dates of adjacent articles
- Temporal expressions in adjacent articles

Available News Citation Network Data

Predict article publication times from:

- Citation network topology
- Publication dates of adjacent articles
- Temporal expressions in adjacent articles
- **Not** the metadata of the article itself
- **Not** the article content

Feature Extraction

Network Topology Features

Node degree-based features:

- Incoming degree
- Outgoing degree
- Undirected degree

Network Topology Features

Node degree-based features:

- Incoming degree
- Outgoing degree
- Undirected degree

Centrality-based features:

- Betweenness centrality
- Incoming closeness centrality
- Outgoing closeness centrality
- Page Rank centrality

Network Topology Features

Node degree-based features:

- Incoming degree
- Outgoing degree
- Undirected degree

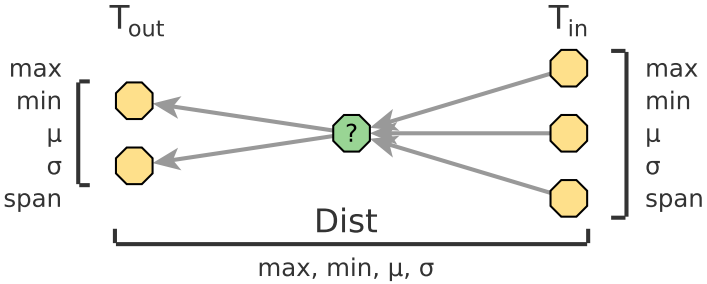
Density-based features:

- Undirected local clustering coefficient

Centrality-based features:

- Betweenness centrality
- Incoming closeness centrality
- Outgoing closeness centrality
- Page Rank centrality

Temporal Network Features



Temporal Expression Features

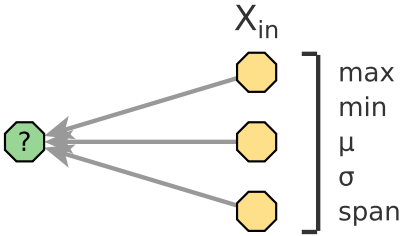
Correlation of temporal expressions:

- **good** with publication dates of referencing articles (incoming edges)
- **bad** with publication dates of referenced articles (outgoing edges)

Temporal Expression Features

Correlation of temporal expressions:

- **good** with publication dates of referencing articles (incoming edges)
- **bad** with publication dates of referenced articles (outgoing edges)



Missing Features and Imputation

Missing features

- 30.8% of feature values are missing
- 89.6% of articles are missing at least one feature

Missing Features and Imputation

Missing features

- 30.8% of feature values are missing
- 89.6% of articles are missing at least one feature

Imputation of missing values

- Column mean of the feature

Evaluation

Regression Methods

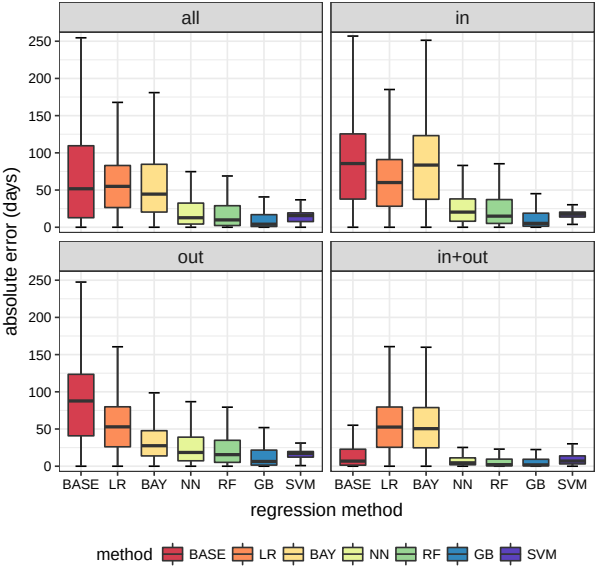
Used regression methods:

- **BASE**: Baseline (average publication date of adjacent articles)
- **LR**: Linear regression
- **BAY**: Bayesian ridge regression (Laplace model)
- **RF**: Random forest
- **GB**: Gradient boosting (Laplace distribution, decision trees)
- **SVM**: Support vector machine (radial kernel)
- **NN**: Neural network (feedforward, one hidden layer)

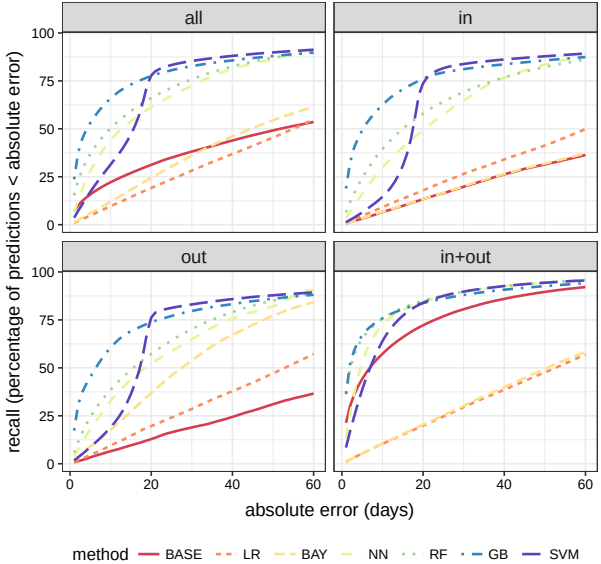
Evaluation Results: Mean Absolute Error (days)

	BASE	LR	BAY	NN	RF	GB	SVM
all	66.72	60.46	59.61	26.88	24.98	**22.66**	26.19
in	88.88	66.48	87.55	34.03	32.25	**27.49**	32.29
out	87.32	59.54	40.24	32.52	30.10	**26.68**	30.77
in+out	18.68	55.45	54.95	12.62	**11.23**	12.76	14.31

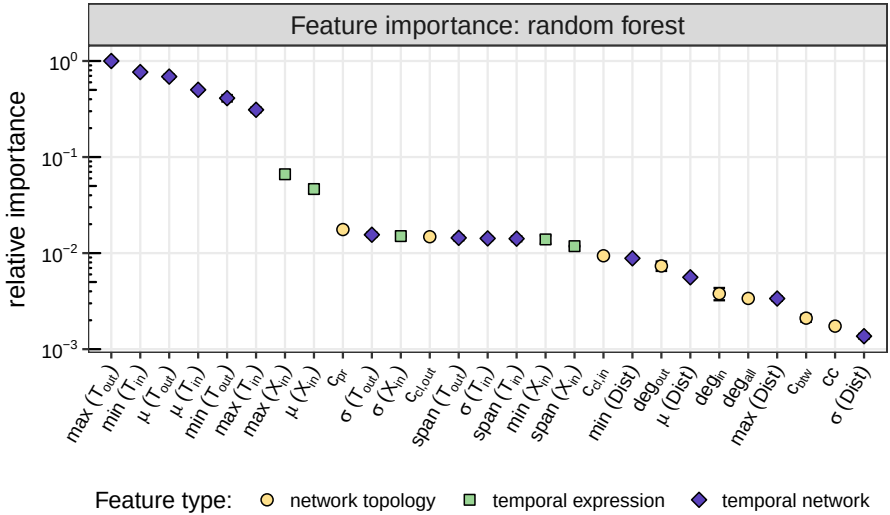
Distribution of Absolute Errors



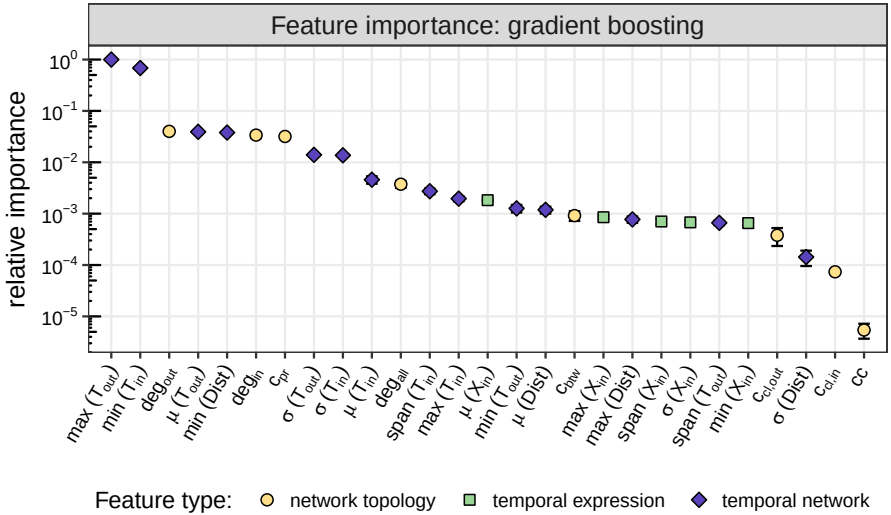
Recall by Varying Absolute Error



Feature Importance: Random Forest



Feature Importance: Gradient Boosting



Summary & Resources

Summary

News citation networks:

- Focus on anchored links inside the article body
- Constructed like a citation network between articles

Publication date prediction:

- Can be framed as a regression problem
- Average prediction error of 3 weeks
- Temporal network features are most discriminative

Resources

Data and implementation are available online:

- [data] News citation network (including URLs)
- [data] Temporal annotations
- [code] Publication date prediction

https://dbs.ifi.uni-heidelberg.de/resources/data/

Resources

Data and implementation are available online:

- [data] News citation network (including URLs)
- [data] Temporal annotations
- [code] Publication date prediction

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

Thank You!
Questions?

actual content

Interested in more network-based news analysis? Click here:

Exploring Entity-centric Networks in Entangled News Streams
Track: Journalism, Misinformation and Fact Checking III
Wednesday, 15:40 - 17:00, Salle Rhône 2

shameless advertisement