Breaking the News: Extracting the Sparse Citation Network Backbone of Online News Articles

Andreas Spitz and Michael Gertz

Heidelberg University
Institute of Computer Science
Database Systems Research Group
http://dbs.ifi.uni-heidelberg.de

spitz@informatik.uni-heidelberg.de

statNLP Kolloquium
Heidelberg, June 26, 2015

Handelsblatt

ÜBERNAHME VON SECUSMART

Blackberry kauft „Merkel-Phone“

Autor: dpa
Datum: 29.07.2014 15:20 Uhr • Update: 29.07.2014, 17:15 Uhr

Handelsblatt

ÜBERNAHME VON SECUSMART

Blackberry kauft „Merkel-Phone“

Autor: dpa
Datum: 29.07.2014 15:20 Uhr • Update: 29.07.2014, 17:15 Uhr

theguardian
Winner of the Pulitzer prize 2014

Germany opens inquiry into claims NSA tapped Angela Merkel's phone

Unexpected inquiry, announced by federal prosecutor, will determine if US actively listened in to calls

Networks of news articles are analyzed frequently, e.g. for

- Information diffusion
- Event detection
- Information cascades
- Media dynamics

Networks of news articles are analyzed frequently, e.g. for

- Information diffusion
- Event detection
- Information cascades
- Media dynamics

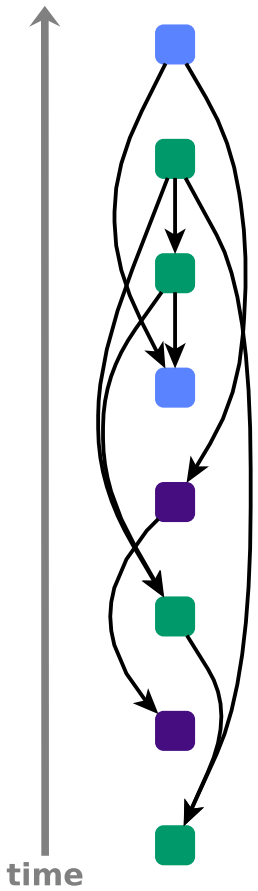
But what about network extraction and emergence?

- Are all networks of news articles born equal?
- Or: when is a link a link?

Overview

1) **Data Extraction** of networks of news articles
2) **Network Structure** of the News Citation Network
3) **Citation Characteristics** of the network
4) **Applications** and Analysis on the network
5) **Traditional Networks** in comparison
6) **Summary**

The Ideal Network of News Articles



SZ [politics] (2015.03.05)
Posse im NSA-Ausschuss - Handy weg, Akten weg

Zeit [politics] (2015.03.05)
NSA-Ausschuss - Abgeordnete halten BND-Akten für manipuliert

Zeit [politics] (2015.03.05)
NSA-Affäre - BND enthielt aus Versehen 130 Akten vor

Spiegel [politics] (2015.03.05)
Aufklärung der Spähaffäre - BND gibt Akten-Schluderei

Welt [politics] (2015.03.03)
Handy ausgespäht - Spionageverdacht im Geheimdienst-Ausschuss

Zeit [politics] (2015.02.05)
NSA-Affäre - NSA-Ausschuss fühlt sich eingeschüchtert

Welt [business] (2014.11.27)
Merkel-Phones - Bund schließt Anti-Spionage-Vertrag mit Blackberry

Zeit [politics] (2014.09.09)
Überwachungsaffäre - NSA-Ausschuss sieht nur schwarz

Directed, acyclic network with time ordering of nodes

Types of Links Between News Articles

Classification of links by location and target:

a) navigational links

b) anchored references

c) internal links

d) advertisement

The Established Approach: Crawling

For very large data sets:

- Select a large number of news outlets
- Crawl the web pages and follow links
- Extract all articles along the way

The Established Approach: Crawling

For very large data sets:

- Select a large number of news outlets
- Crawl the web pages and follow links
- Extract all articles along the way

Problems:

- Determining publication time
- Extracting the article's content
- Recombining multi-page articles
- Distinguishing between link types

The Established Approach: RSS-Feeds

For streams of news articles:

- Select news outlets that publish RSS-Feeds
- Periodically check Feeds
- Download new articles

The Established Approach: RSS-Feeds

For streams of news articles:

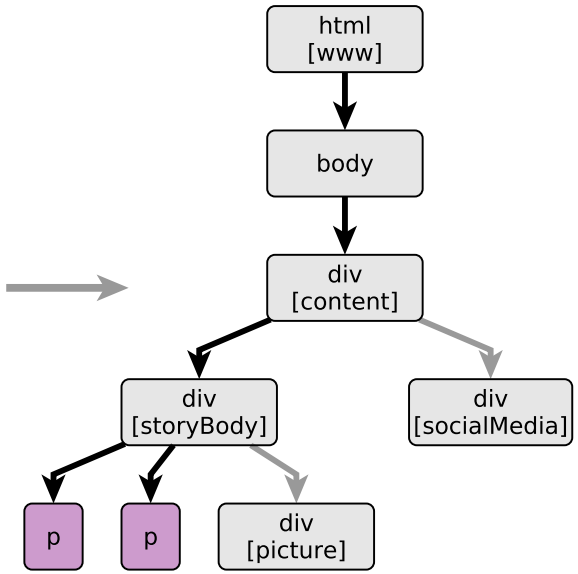
- Select news outlets that publish RSS-Feeds
- Periodically check Feeds
- Download new articles

Problems:

- Determining publication time
- Extracting the article's content
- Recombining multi-page articles
- Distinguishing between link types

Structural Basics of News Articles: HTML DOM-Tree

```
<html class="www" lang="de">
  <body>
    <div class="content">
      <div class="storyBody">
        <p>Die Proteste in Hongkong haben am Mittwoch...</p>
        <p>Mit einem Schweigeprotest begleiteten die Demonstranten...</p>
        <div class="picture">
          <a href="http://www.welt.de/.../Hong-Kong-Occupy.jpg"></a>
        </div>
      </div>
      <div class="socialMedia">
        ...
      </div>
    </div>
  </body>
</html>
```



A Rule-based Approach

Create a network by

- limiting the set of nodes to articles published by news outlets
- downloading all pages of multi-page articles
- using outlet-dependent rules to extract the article text
- extracting anchored references within the texts as edges

A Rule-based Approach

Create a network by

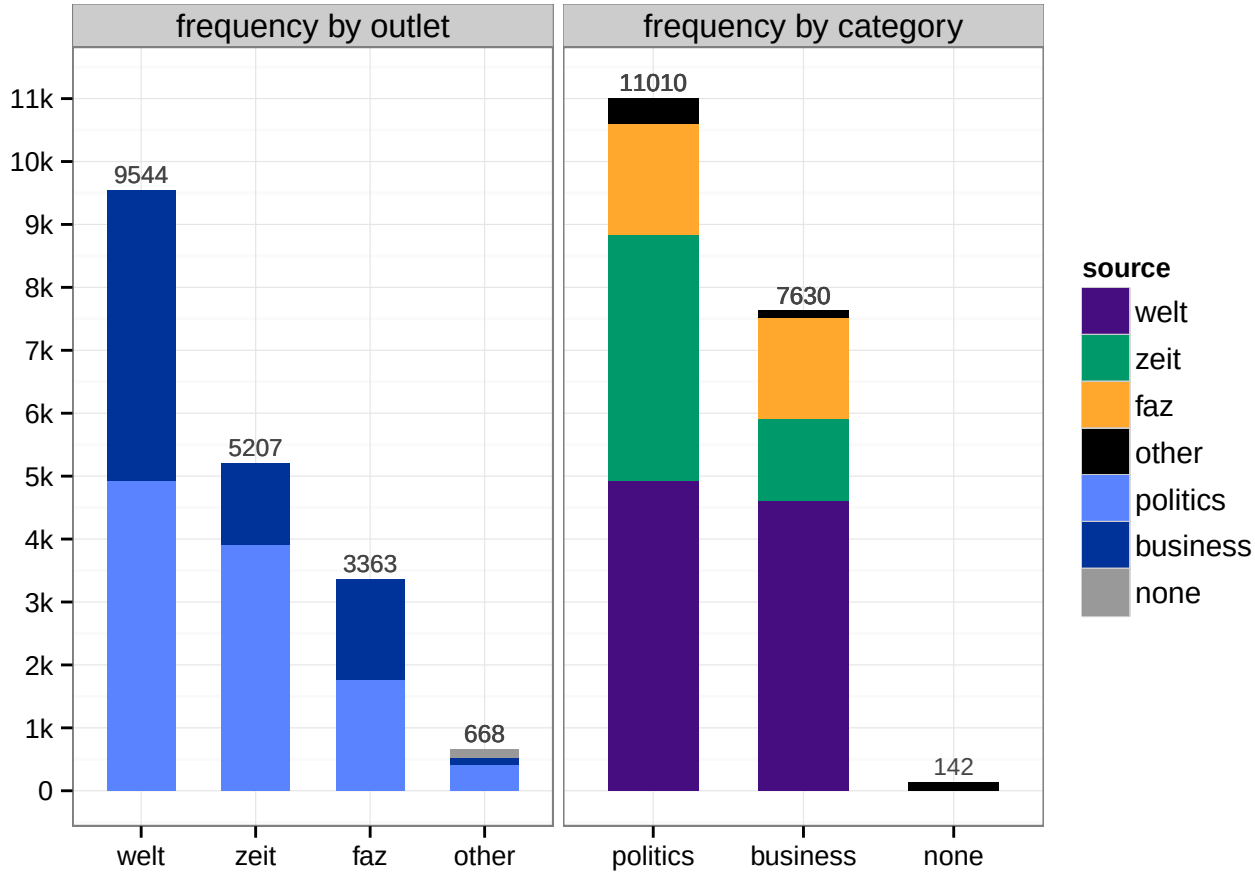
- limiting the set of nodes to articles published by news outlets
- downloading all pages of multi-page articles
- using outlet-dependent rules to extract the article text
- extracting anchored references within the texts as edges

Problems:

- Determining publication time
- Extracting the article's content
- Recombining multi-page articles
- Distinguishing between link types
- Additional effort to find extraction rules

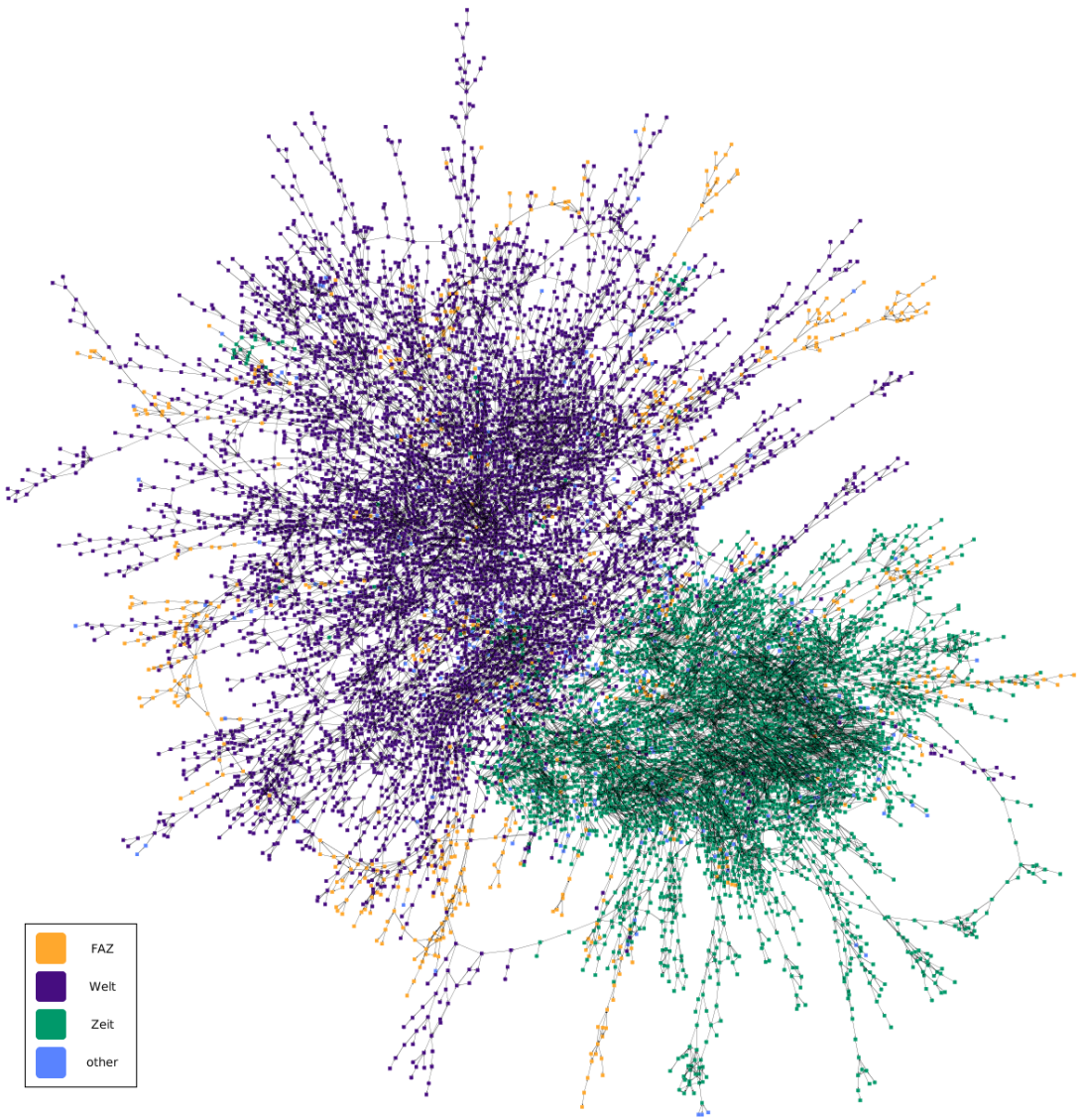
The News Citation Network

Data collected from 6 German news outlets over 10 months



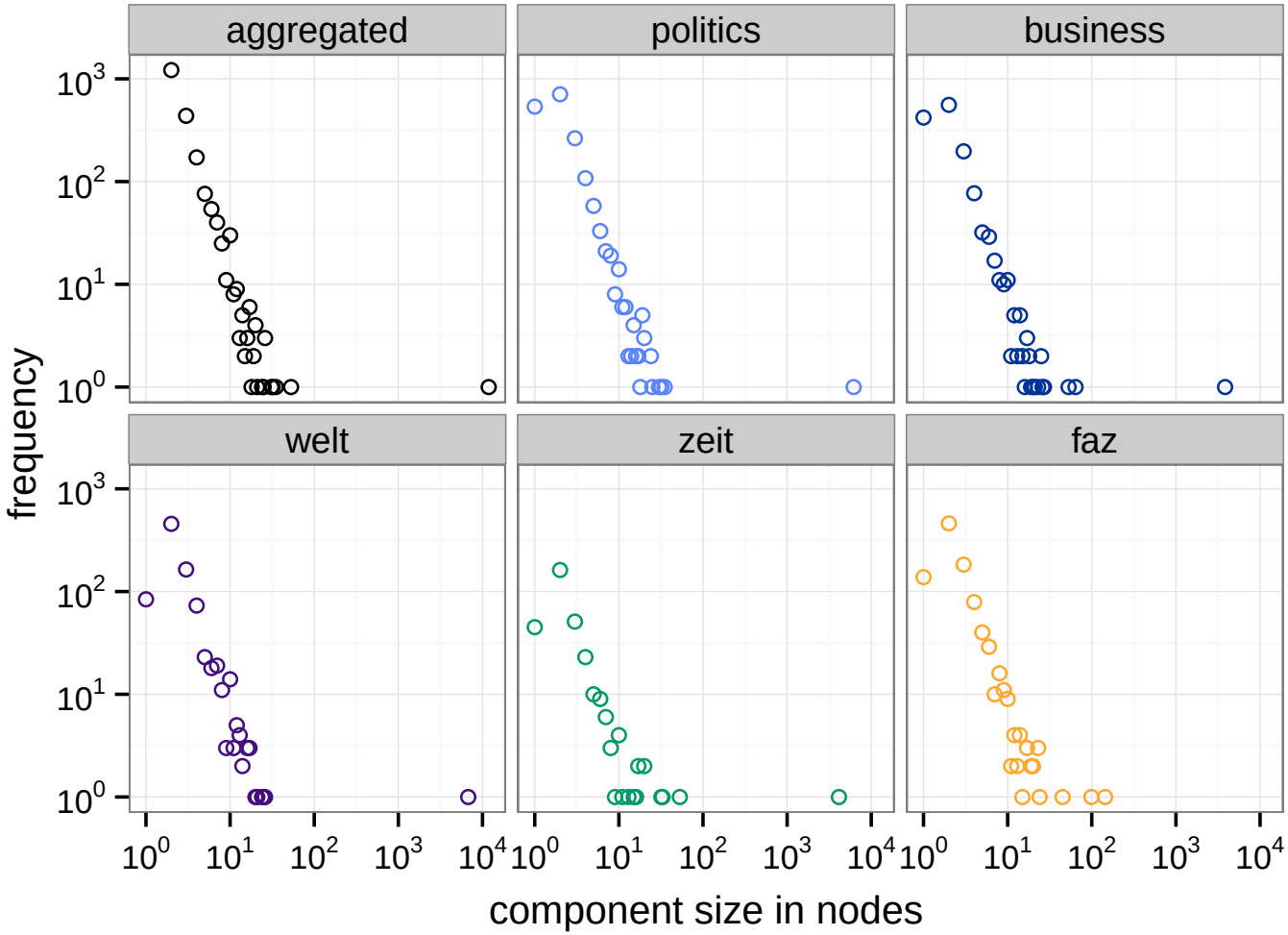
$|V| = 18,782$ articles and $|E| = 21,581$ references between them

Components of the News Network

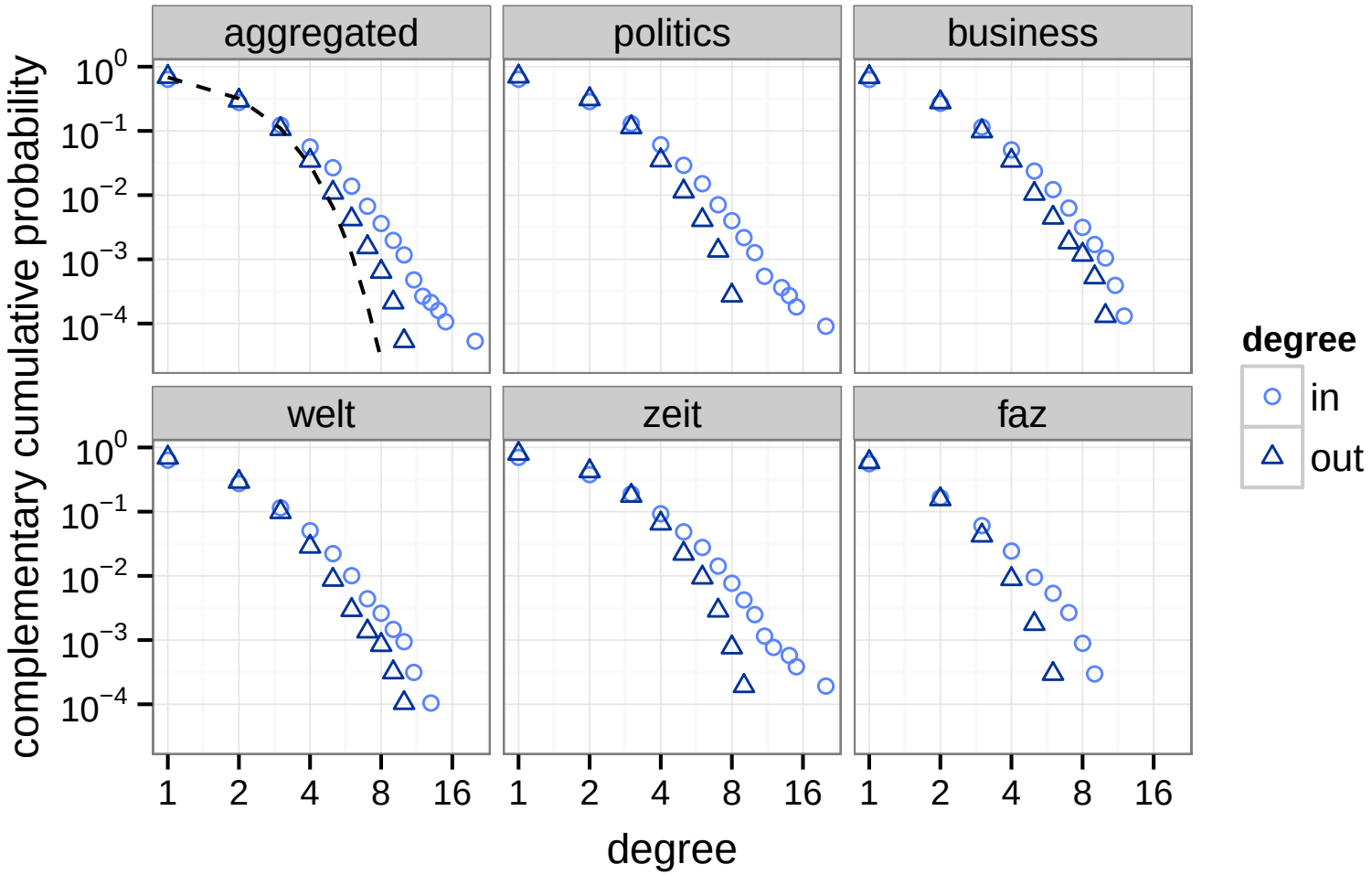


- 63.1% of nodes in one giant connected component
- Component consists of two clusters of articles from Zeit and Welt
- Other articles are mixed in or form small, homogeneous components

Component Size Distribution



Degree Distribution

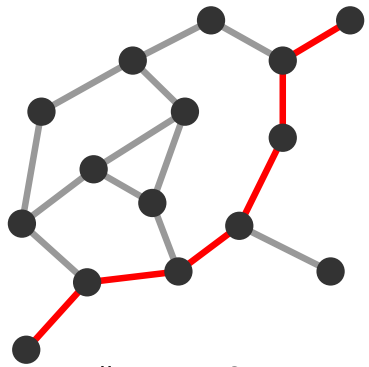


Structural Measures (Definitions)

Structural measures for a network:

- **Average degree**: mean number of neighbours of a node in the network
- **Clustering coefficient**: $cc = \frac{3\Delta}{T}$
  Δ is the number of closed triangles
  T is the number of connected triples.
- **Diameter** $\varnothing$: the longest shortest path between any two nodes
- **Average path length** $\langle l \rangle$: average length of pairwise shortest paths

connected triple closed triangle



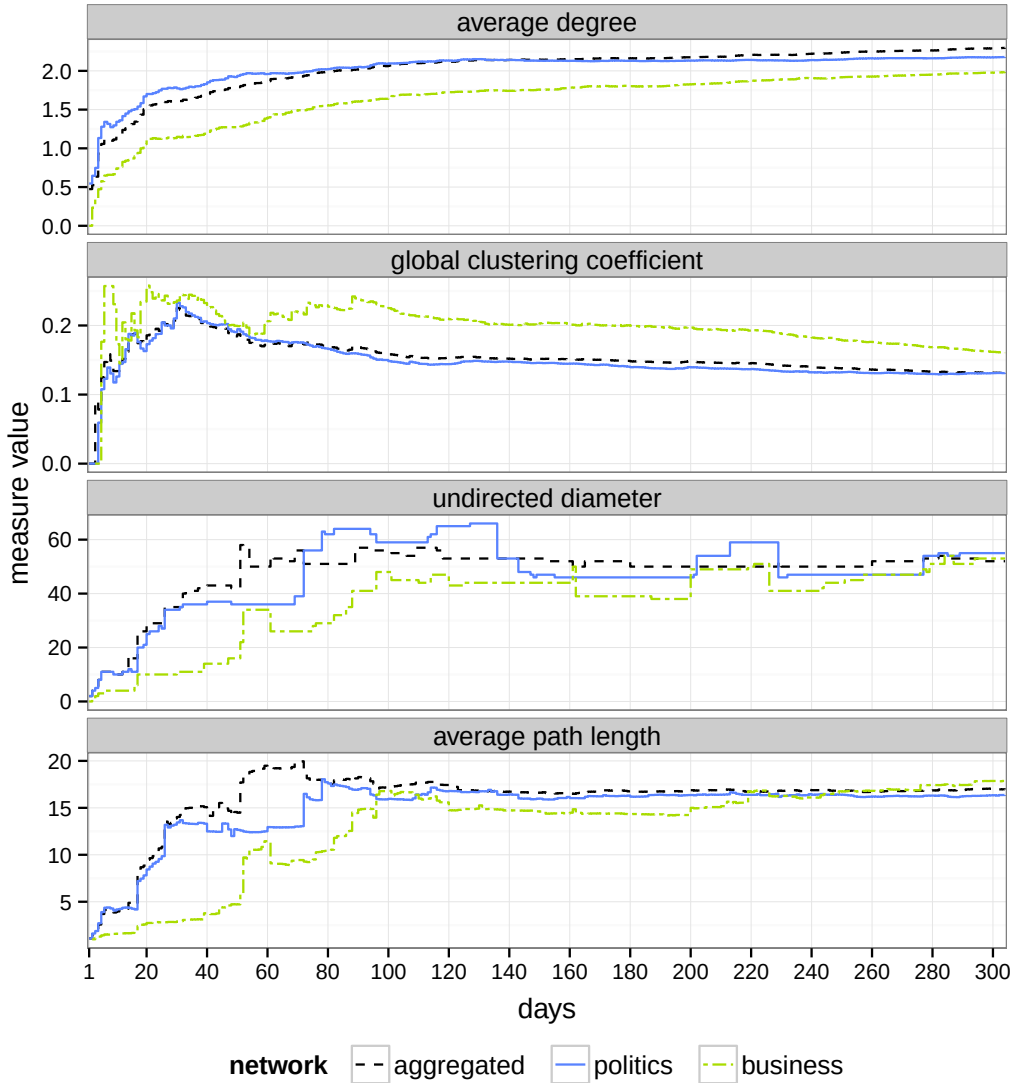
diameter: 6

Structural Measures

network	$\|V\|$	$\|E\|$	cc	$\varnothing_d$	$\varnothing_u$	$\langle l_d \rangle$	$\langle l_u \rangle$
aggregated	18782	21581	0.13	38	52	11.0	16.9
politics	11010	11996	0.13	37	55	11.0	16.4
business	7630	7579	0.16	16	53	3.6	17.8
welt	9544	10536	0.11	24	47	6.2	16.2
zeit	5207	7594	0.16	37	37	11.9	11.6
faz	3363	2603	0.13	12	23	2.4	7.0

Clustering coefficient cc, diameters $\varnothing$ and average path lengths $\langle l \rangle$.

Network Evolution

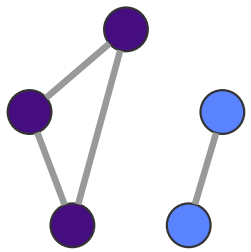


Modularity and Assortativity (I)

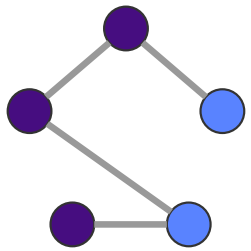
$$Q := \frac{1}{2|E|} \sum_{i,j} \left[A_{ij} - \frac{deg(v_i)deg(v_j)}{2|E|} \right] \delta(v_i, v_j)$$

Where:

- A is the $\{0, 1\}$-valued adjacency matrix
- $deg(v)$ is the number of neighbours of node v
- $\delta(v_i, v_j) := \begin{cases} 1 & \text{if } outlet(v_i) = outlet(v_j) \\ 0 & \text{if } outlet(v_i) \neq outlet(v_j) \end{cases}$



modular



not modular

Newman (2003)

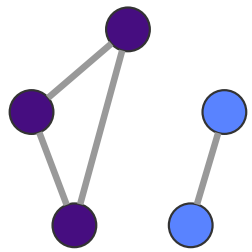
Modularity and Assortativity (I)

$$Q := \frac{1}{2|E|} \sum_{i,j} \left[A_{ij} - \frac{deg(v_i)deg(v_j)}{2|E|} \right] \delta(v_i, v_j)$$

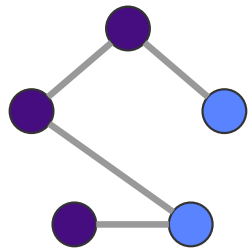
Where:

- A is the $\{0, 1\}$-valued adjacency matrix
- $deg(v)$ is the number of neighbours of node v
- $\delta(v_i, v_j) := \begin{cases} 1 & \text{if } outlet(v_i) = outlet(v_j) \\ 0 & \text{if } outlet(v_i) \neq outlet(v_j) \end{cases}$

The complete news network is highly modular by news outlet with $Q = 0.582$



modular



not modular

Newman (2003)

Modularity and Assortativity (II)

network		Q_{cat}	Q_{ol}	r	r_{ii}	r_{io}	r_{oi}	r_{oo}
	obs	0.39	0.57	0.25	0.13	0.16	0.52	0.19
aggreg.	mod			0.06	0.17	0.00	0.51	0.13
	δ			0.19	-0.03	0.16	0.01	0.06
	obs		0.56	0.23	0.13	0.15	0.51	0.18
politics	mod			0.10	-0.13	0.09	0.43	-0.15
	δ			0.13	**0.26**	0.06	0.08	**0.33**
	obs		0.49	0.31	0.10	0.19	0.53	0.16
business	mod			0.12	-0.27	0.32	0.36	-0.26
	δ			0.20	**0.36**	-0.13	0.16	**0.41**

Modularity by category Q (by category and news outlet), assortativity by degree r and directed assortativity $r_{in,in}$, $r_{in,out}$, $r_{out,in}$ and $r_{out,out}$.

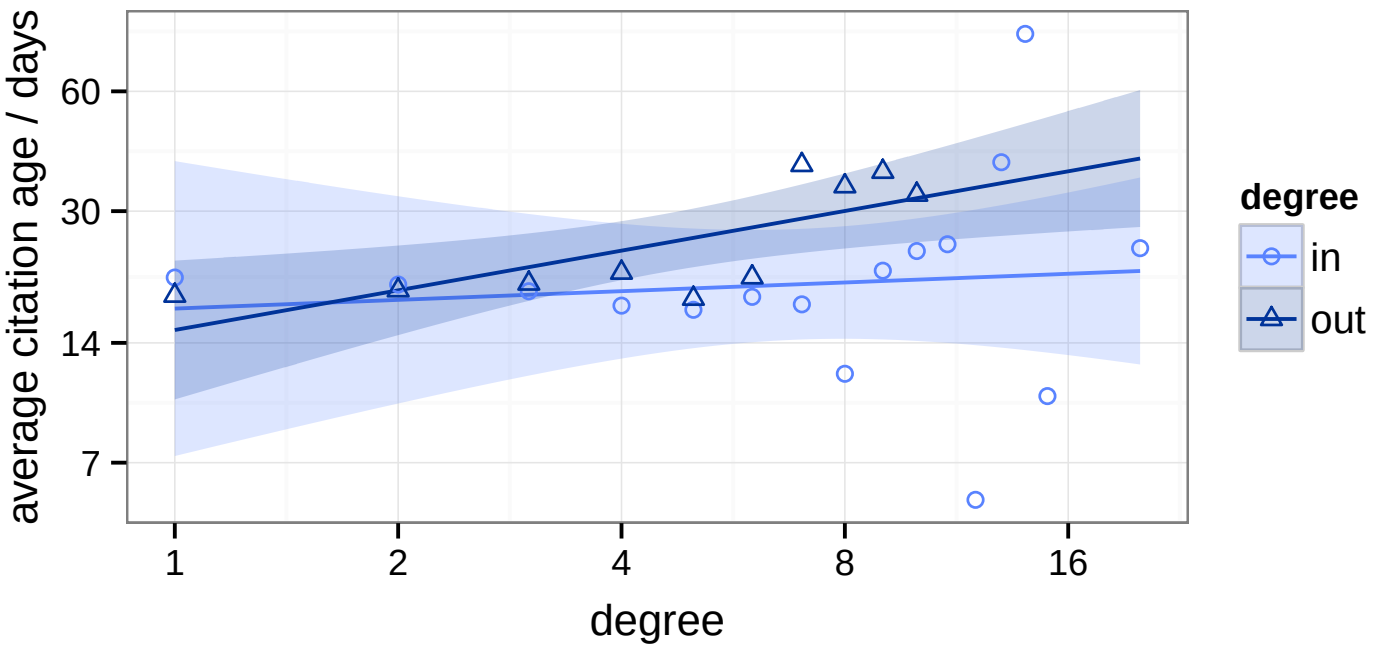
Summary of Network Structure

The News Citation Network

- is very sparse and largely connected
- is highly modular and assortative
- has constant clustering coefficient
- has no shrinking diameter
- has long, constant average path length

$\Rightarrow$ This indicates a hierarchical structure

The Effect of Age on Citations



Models for Citation Networks

Models and applications for citation networks are well established:

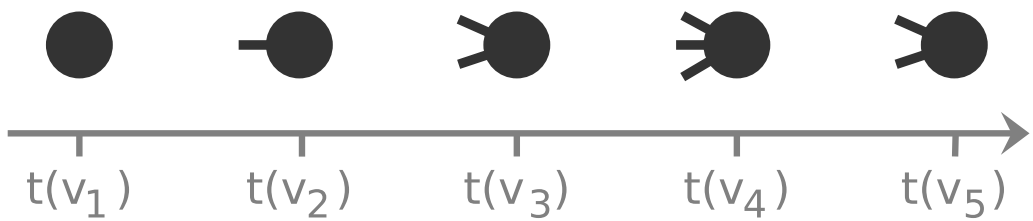
- de Solla Price (1965)
- Garfield (1972) and Hirsch (2005)
- Barabási and Albert (1999)
- Dorogovtsev and Mendez (2000)

Models usually include:

- High clustering coefficient
- Preferential attachment
  - by degree (i.e. popularity)
  - by age (i.e. relevance)
- Long tailed degree distribution

The Triadic Closure Model for DAGs

The nodes are sorted topologically. Outgoing degrees are fixed and parameters $\alpha \in \mathbb{R}$, $\beta \in [0, 1]$ are selected. New edges are then generated for each node v_i, starting with $i = 1$:

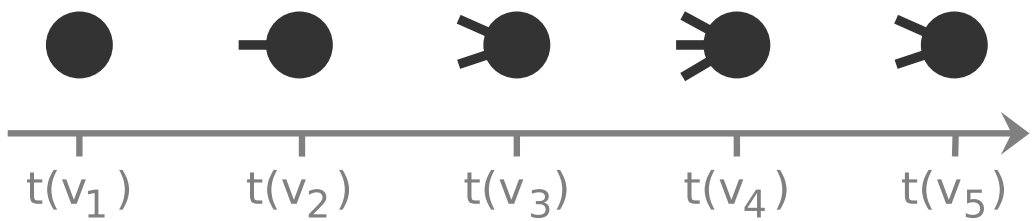


Wu and Holme (2009)

The Triadic Closure Model for DAGs

The nodes are sorted topologically. Outgoing degrees are fixed and parameters $\alpha \in \mathbb{R}$, $\beta \in [0, 1]$ are selected. New edges are then generated for each node v_i, starting with $i = 1$:

- **Decay with age**: The first edge of a node is attached to a random older node v_j with probability $\Pi_{ij} \sim (t(v_i) - t(v_j))^{\alpha}$.

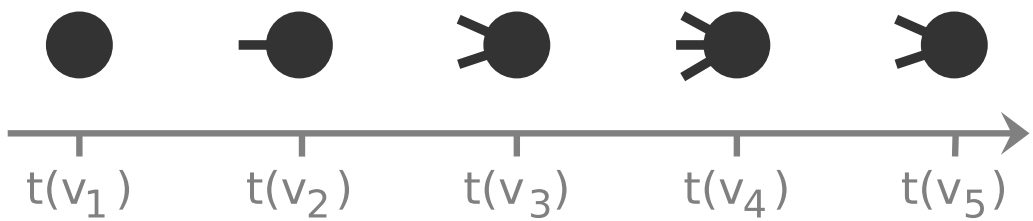


Wu and Holme (2009)

The Triadic Closure Model for DAGs

The nodes are sorted topologically. Outgoing degrees are fixed and parameters $\alpha \in \mathbb{R}$, $\beta \in [0, 1]$ are selected. New edges are then generated for each node v_i, starting with $i = 1$:

- **Decay with age**: The first edge of a node is attached to a random older node v_j with probability $\Pi_{ij} \sim (t(v_i) - t(v_j))^{\alpha}$.
- **Triangle creation**: With probability β, the next edge is attached to a randomly selected neighbour of v_j.

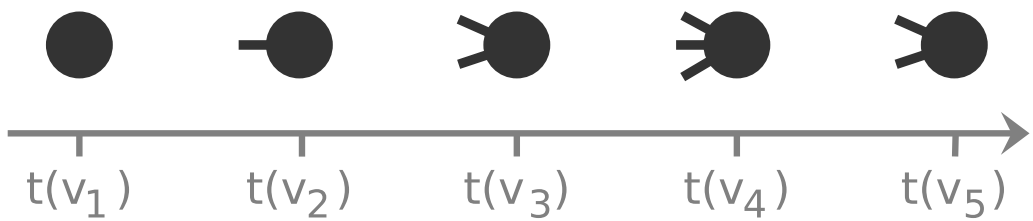


Wu and Holme (2009)

The Triadic Closure Model for DAGs

The nodes are sorted topologically. Outgoing degrees are fixed and parameters $\alpha \in \mathbb{R}$, $\beta \in [0, 1]$ are selected. New edges are then generated for each node v_i, starting with $i = 1$:

- **Decay with age**: The first edge of a node is attached to a random older node v_j with probability $\Pi_{ij} \sim (t(v_i) - t(v_j))^{\alpha}$.
- **Triangle creation**: With probability β, the next edge is attached to a randomly selected neighbour of v_j.
- With probability $1 - \beta$, the edge is instead attached to any older node as in the first step.



Wu and Holme (2009)

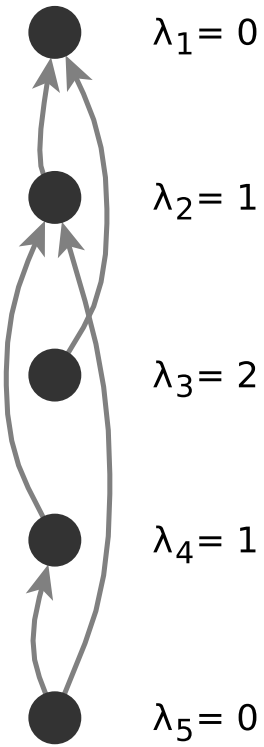
Goodness of Fit

The goodness of fit F depends on:

- The number of transient edges λ_i passing each node v_i:

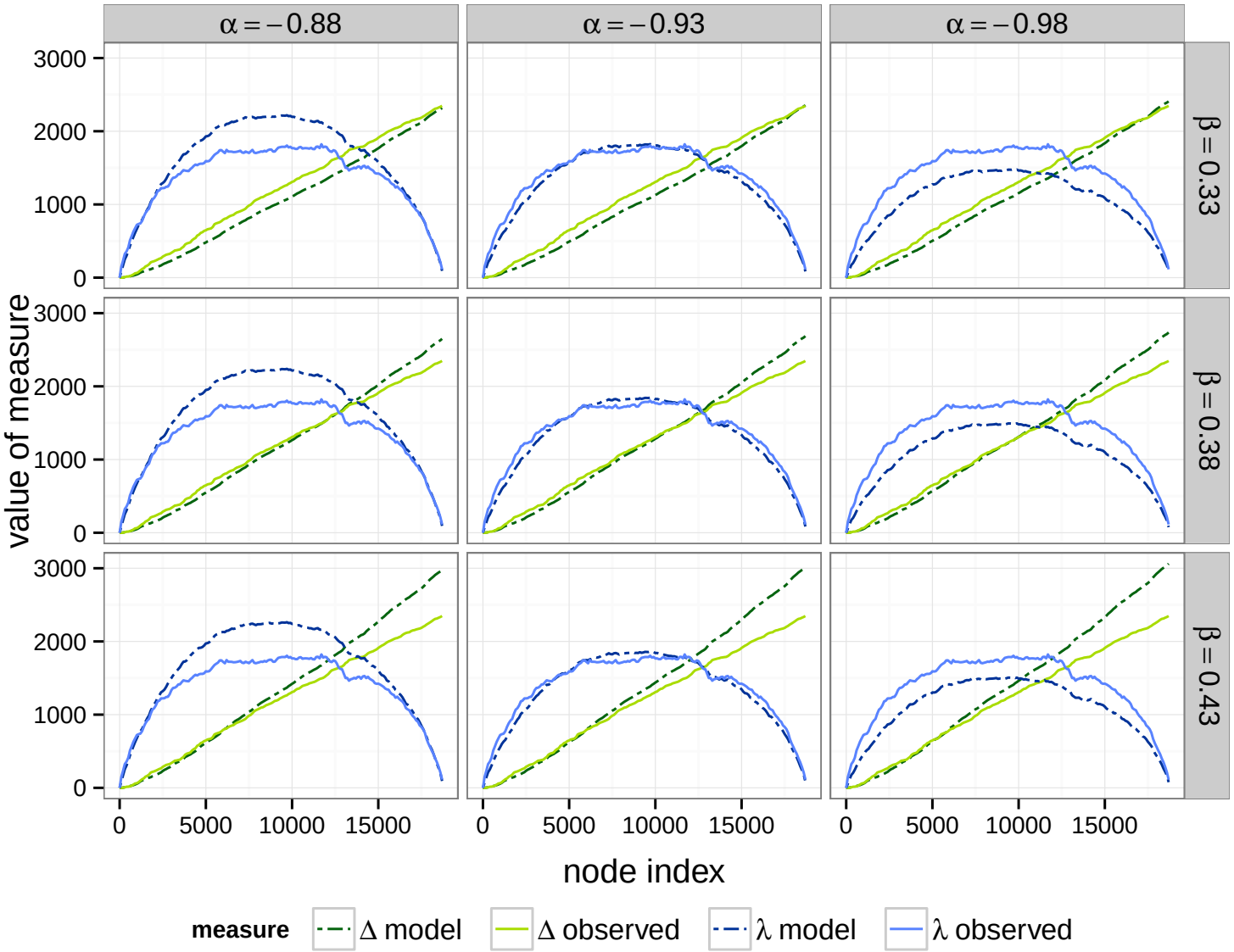
$$\lambda_i := \sum_{j=1}^{i-1} deg_{in}(v_j) - \sum_{j=1}^{i} deg_{out}(v_j)$$

- The number of triangles Δ_i in the graph after node v_i is included.

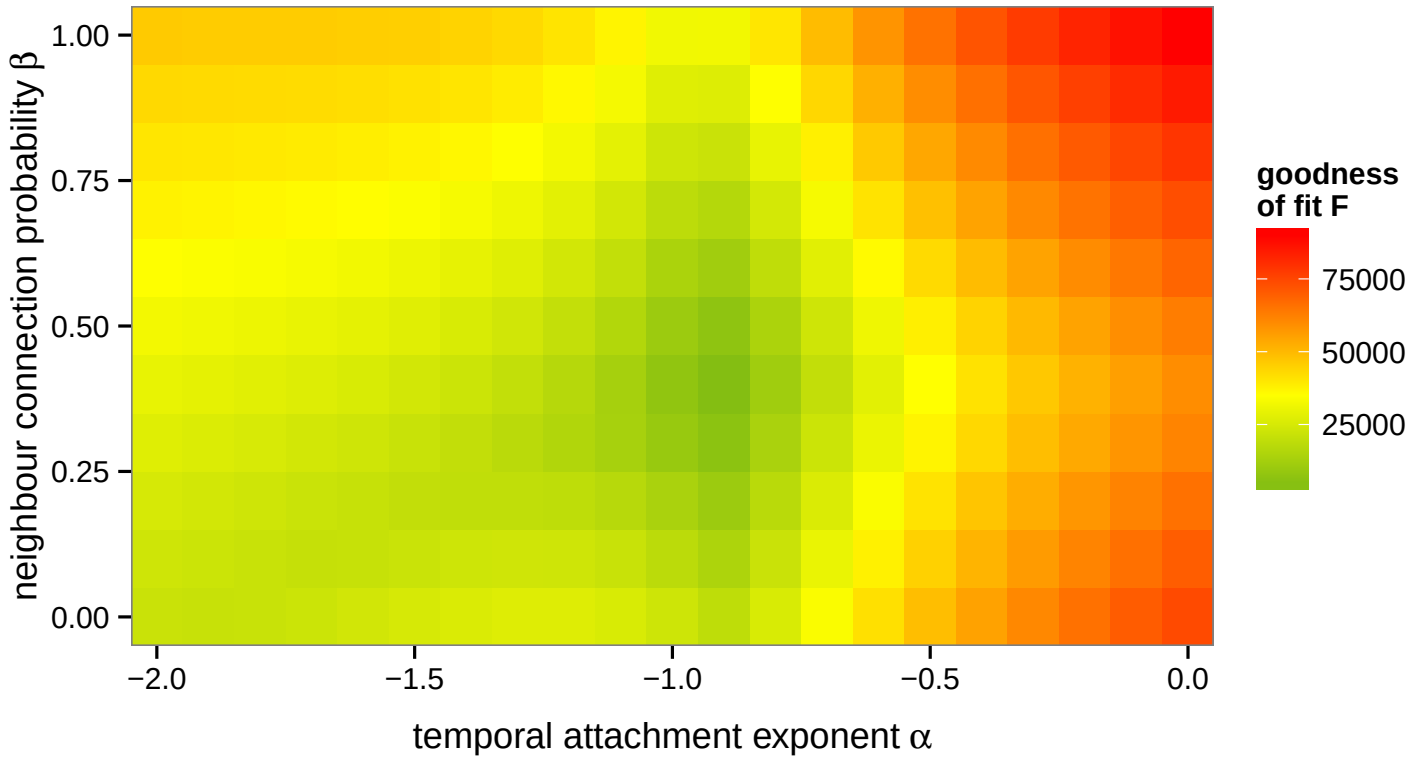


$$F := \sum_{i=1}^{|V|} \frac{|\Delta_i - \Delta_i^{obs}|}{\Delta_i^{obs}} + \sum_{i=1}^{|V|} \frac{|\lambda_i - \lambda_i^{obs}|}{\lambda_i^{obs}}$$

Fitting the Model (I)

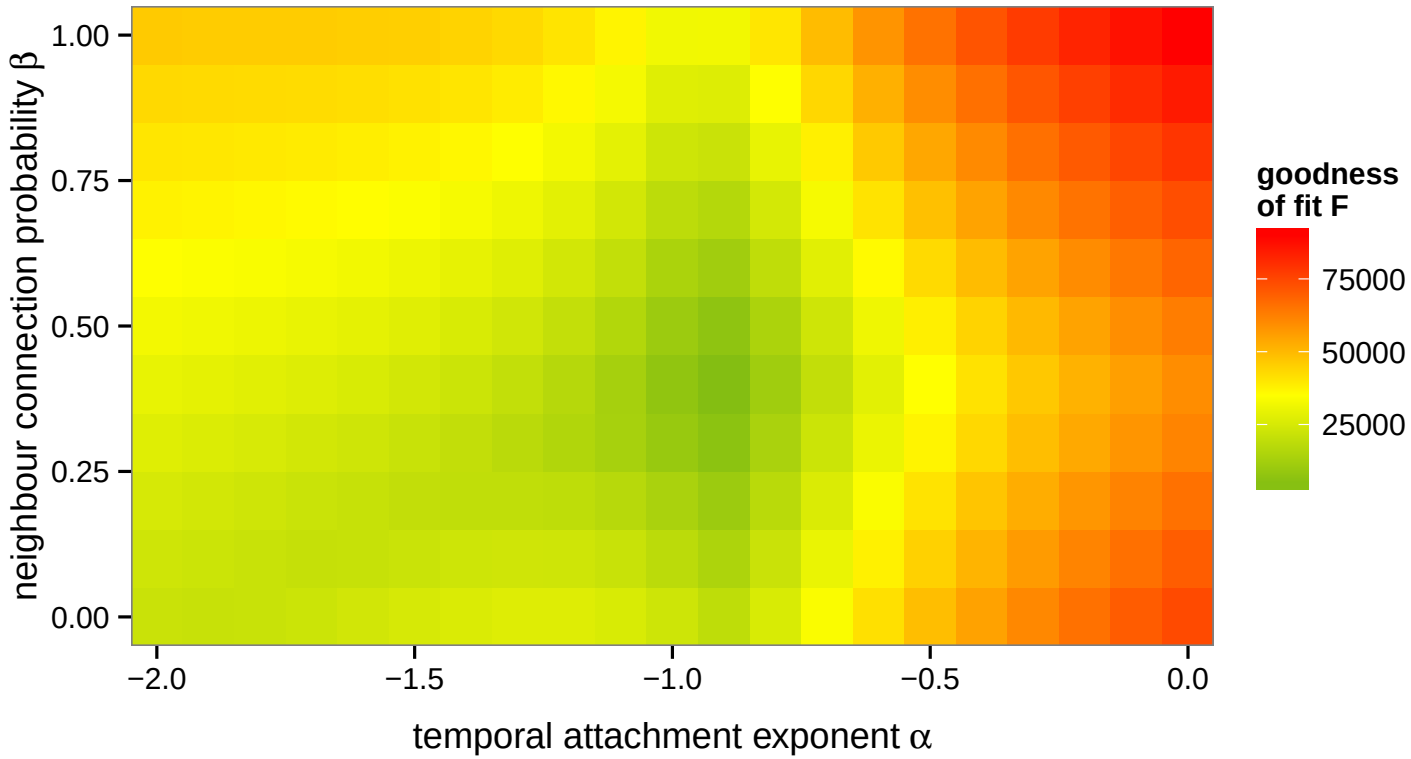


Fitting the Model (II)



Optimum at $\alpha = -0.93$ and $\beta = 0.38$

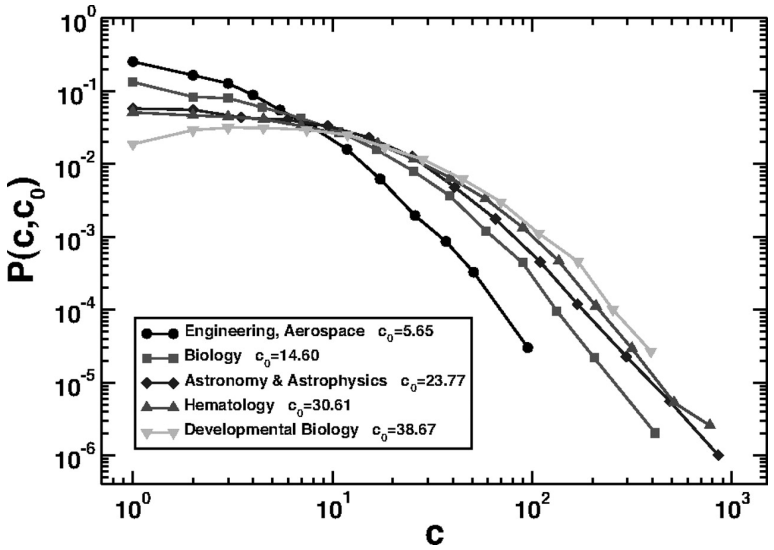
Fitting the Model (II)

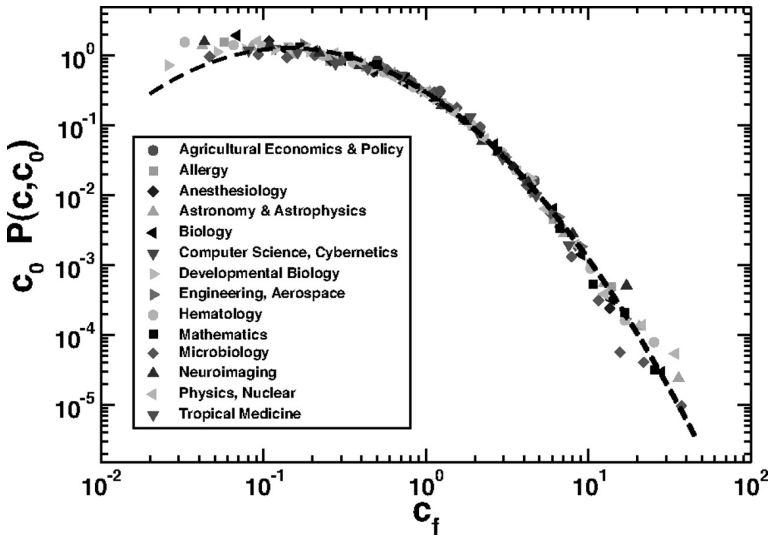


Optimum at $\alpha = -0.93$ and $\beta = 0.38$

$\Rightarrow$ Attachment probability decays linearly with age

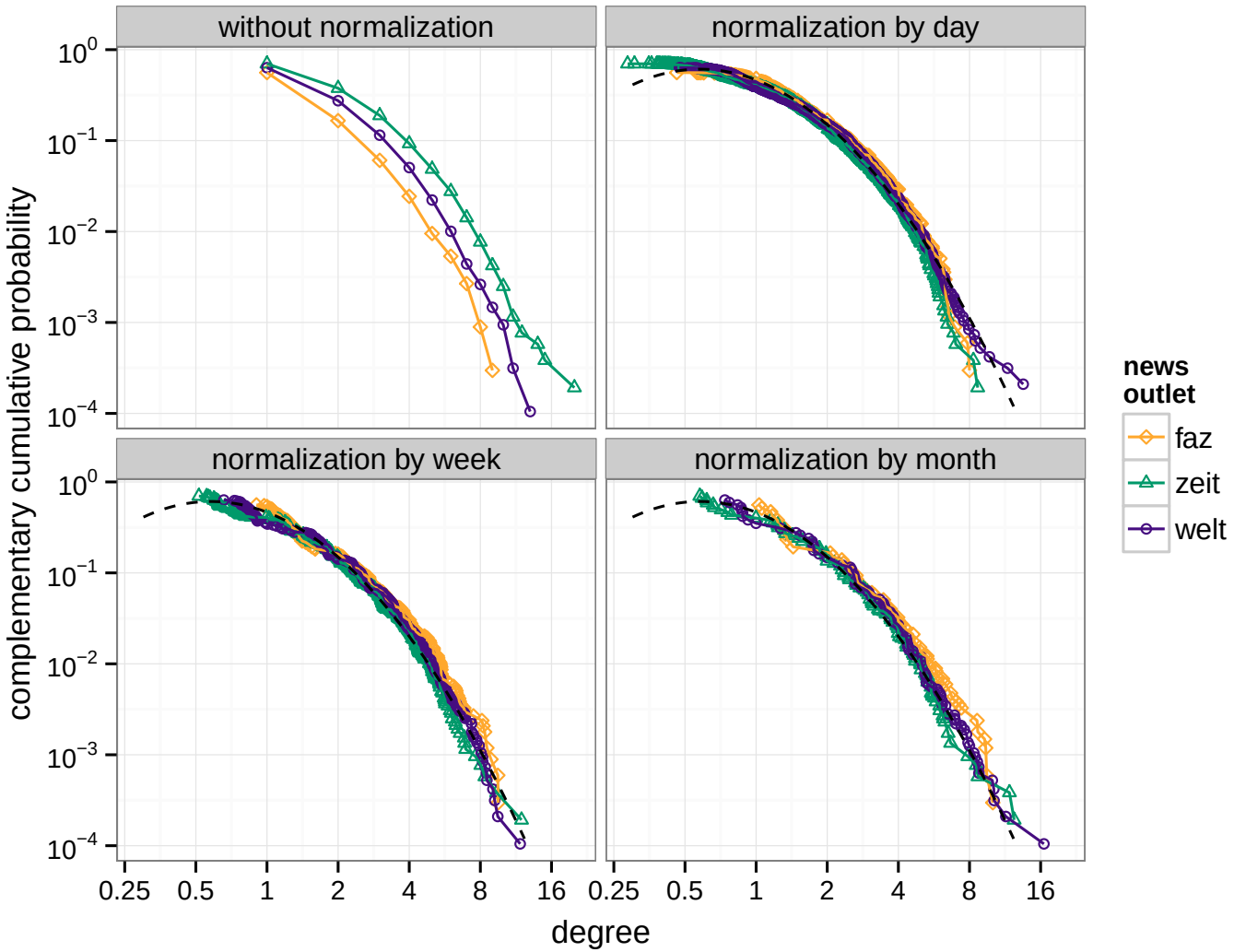
Universality of Citation Distribution





Radicchi, Fortunato and Castellano (2008)

Universality of News Citation Distribution



Summary of Citation Characteristics

In the News Citation Network

- preferential attachment is approximately linear with age
- the universal citation distribution is valid independent of the time frame

Centrality in Citation Networks

Network centrality

- measures the importance or influence of a node
- exists in many different forms based on
  - position within the network (path-based)
  - connectedness
  - information propagation

Centrality in citation networks typically measures

- article or author importance
- journal / newspaper influence
- connectedness and information propagation

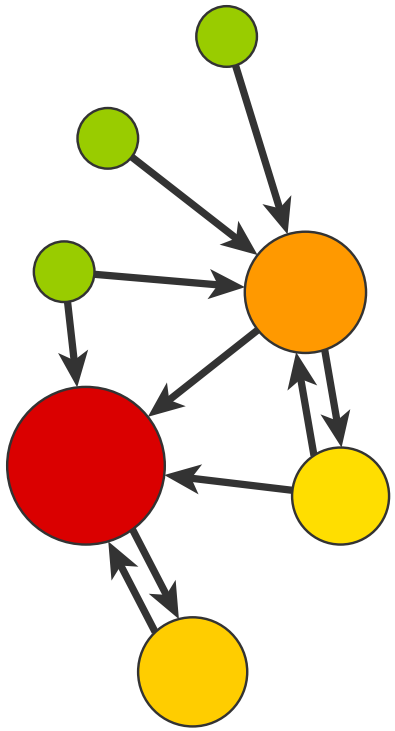
Page Rank Centrality

Page Rank is a measure of influence centrality and defined recursively for a node v:

$$pr_v := \alpha \sum_{w=1}^{|V|} A_{wv} \frac{pr_w}{deg_{out}(w)} + \beta$$

Intuitively:

- "credit" is propagated backwards where information is propagated forward in the network
- receiving "credit" from important neighbours is better than receiving credit from a nobody



Page et al. (1999)

Most Central Articles

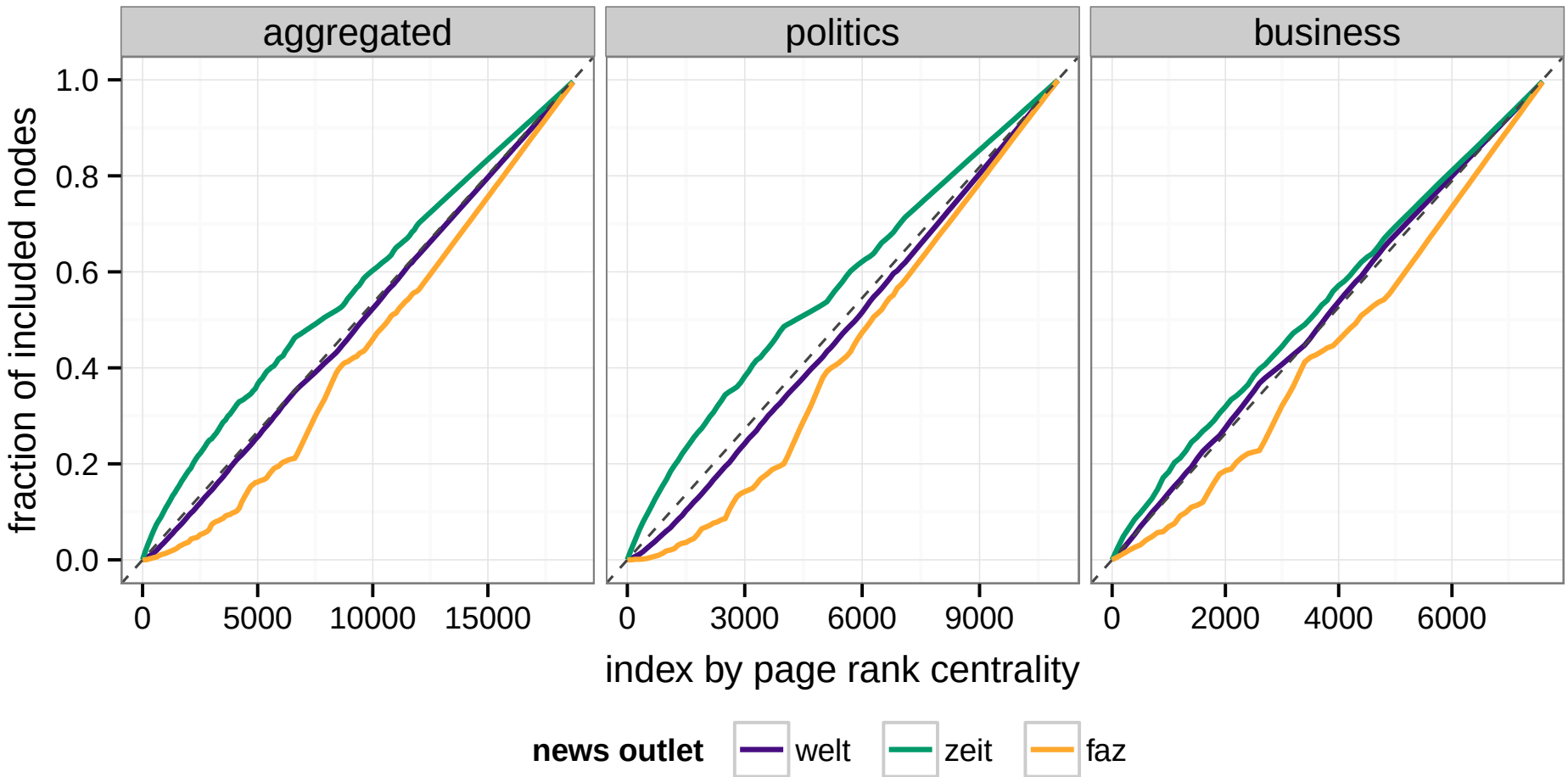
Top-ranked articles by in-degree centrality

d_{in}	pr-rank	outlet	category	date	headline
20	7	zeit	politics	2014.07.21	Ukraine – MH17-Absturz: was wann geschah
15	343	zeit	politics	2014.12.05	Ukraine-Krise – Wieder Krieg in Europa: Nicht in unserem Namen!
14	13	zeit	politics	2014.09.07	Ukraine – OSZE gibt Details des Minsker Abkommens bekannt
13	178	welt	politics	2014.10.15	Asylbewerber – Deutschland ist das Flüchtlingsheim Europas
12	312	zeit	business	2015.02.04	Yanis Varoufakis – "Ich bin Finanzminister eines bankrotten Staates"

Top-ranked articles by Page Rank centrality

d_{in}	pr-rank	outlet	category	date	headline
6	1	zeit	politics	2014.08.08	Erbil – Blitzvormarsch der Dschihadisten ließ USA angreifen
6	2	zeit	politics	2014.08.10	Irak – Zehntausende Jesiden bringen sich in Sicherheit
9	3	zeit	politics	2014.06.10	Irak – Aufständische besetzen Teile der Stadt Mossul
7	4	zeit	politics	2014.06.10	Al-Kaida in Mossul – Der Staat Irak schwindet
7	5	zeit	politics	2014.07.19	Irak – Tausende Christen fliehen aus Mossul

Centrality Profiles

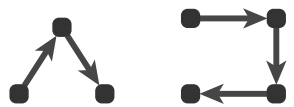


Referencing Patterns

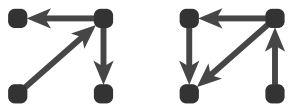
Network Motifs:

- Subgraphs that occur significantly more often in the observed network than in a random sample of graphs
- Significance is assessed by a z-score obtained from frequencies in the sample graphs

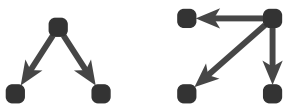
Milo et al. (2002)



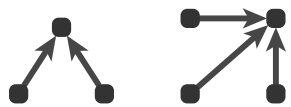
reference chaining



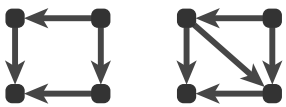
reference relaying



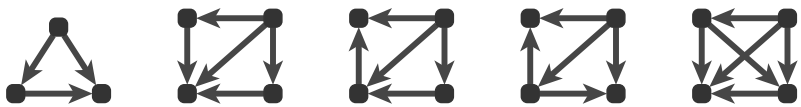
reference fanning



reference aggregation



reference recombination



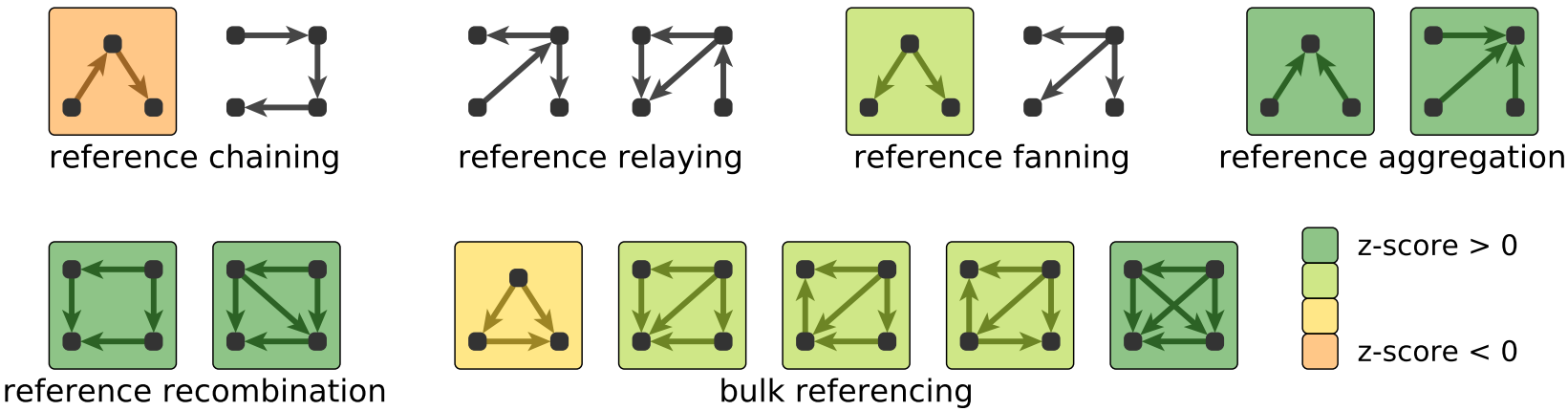
bulk referencing

Referencing Patterns

Network Motifs:

- Subgraphs that occur significantly more often in the observed network than in a random sample of graphs
- Significance is assessed by a z-score obtained from frequencies in the sample graphs

Milo et al. (2002)



Comparison to Crawled Networks

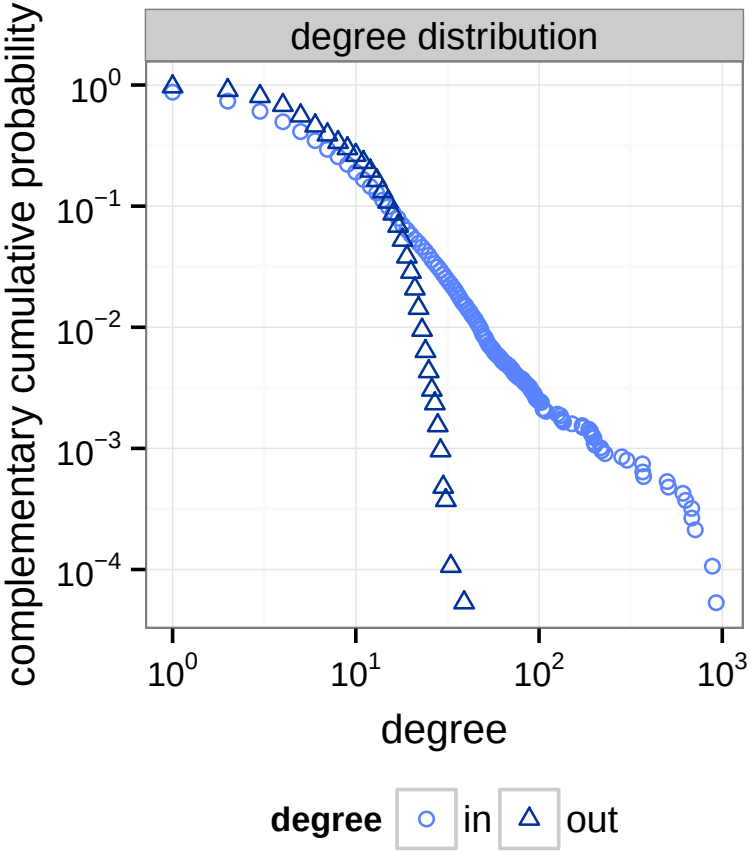
Construction of a traditional, crawled network

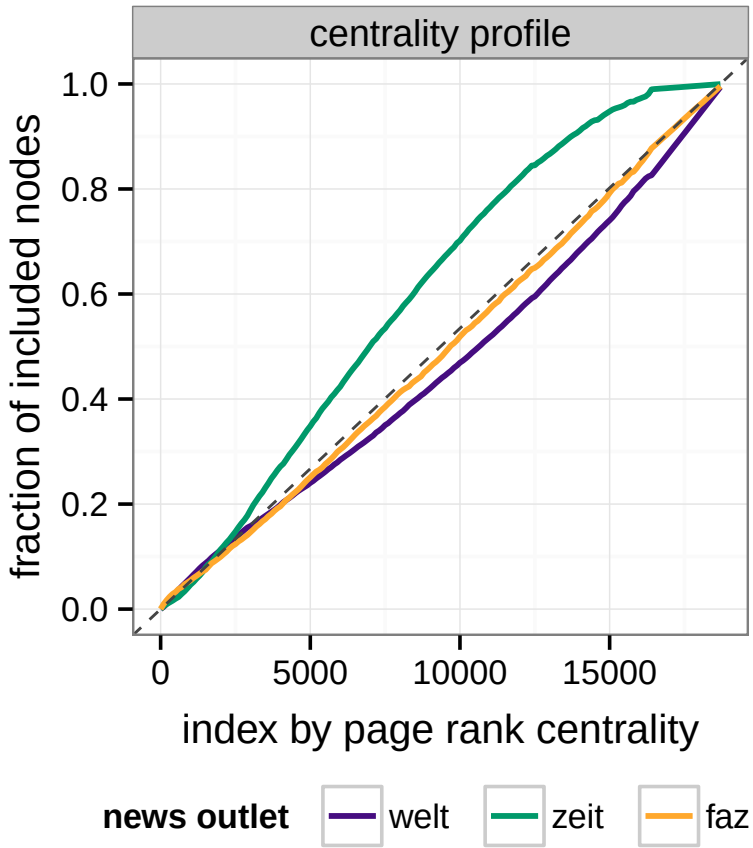
- over the same set of nodes (articles)
- include all links, not just anchored links

Structural measures of the traditional network

- much more dense with $|E| = 128,364$
- slightly higher clustering coefficient $cc = 0.182$
- higher directed diameter and average path length
- lower undirected diameter and path length

Degrees and Centrality for a Traditional Network





Summary

- Semantically anchored links are tied to network structure
- The News Citation Network is similar to scientific citation networks
- The universality of citation distribution is valid over multiple time frames
- The News Citation network has hierarchical structure
- DAG-structure of the network allows for efficient analysis

What's next?

- News citations between international news outlets
- Semi-automated rule extraction
- Ties to social media and user comments
- Analysis of information cascades in traditional media

What's next?

- News citations between international news outlets
- Semi-automated rule extraction
- Ties to social media and user comments
- Analysis of information cascades in traditional media

...or whatever you can think of!
The data is available.

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

Bibliography I

Albert-László Barabási and Réka Albert.
Emergence of scaling in random networks.
Science, 286(5439):509–512, 1999.

Sergey N Dorogovtsev and José FF Mendes.
Evolution of networks with aging of sites.
Physical Review E, 62(2):1842, 2000.

Eugene Garfield.
Citation analysis as a tool in journal evaluation.
Science, 178(4060):471–479, 1972.

Jorge E Hirsch.
An index to quantify an individual's scientific research output.
PNAS, 102(46):16569–16572, 2005.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon.
Network motifs: Simple building blocks of complex networks.
Science, 298:824–827, 2002.

Mark EJ Newman.
Mixing patterns in networks.
Physical Review E, 67(2):026126, 2003.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd.
The pagerank citation ranking: Bringing order to the web.
1999.

Bibliography II

Derek de Solla Price.

Networks of scientific papers.
Science, 149(3683):510–515, 1965.

Filippo Radicchi, Santo Fortunato, and Claudio Castellano.

Universality of citation distributions: Toward an objective measure of scientific impact.
PNAS, 105(45):17268–17272, 2008.

Zhi-Xi Wu and Petter Holme.

Modeling scientific-citation patterns and other triangle-rich acyclic networks.
Physical review E, 80(3):037101, 2009.

RSS Aggregator