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

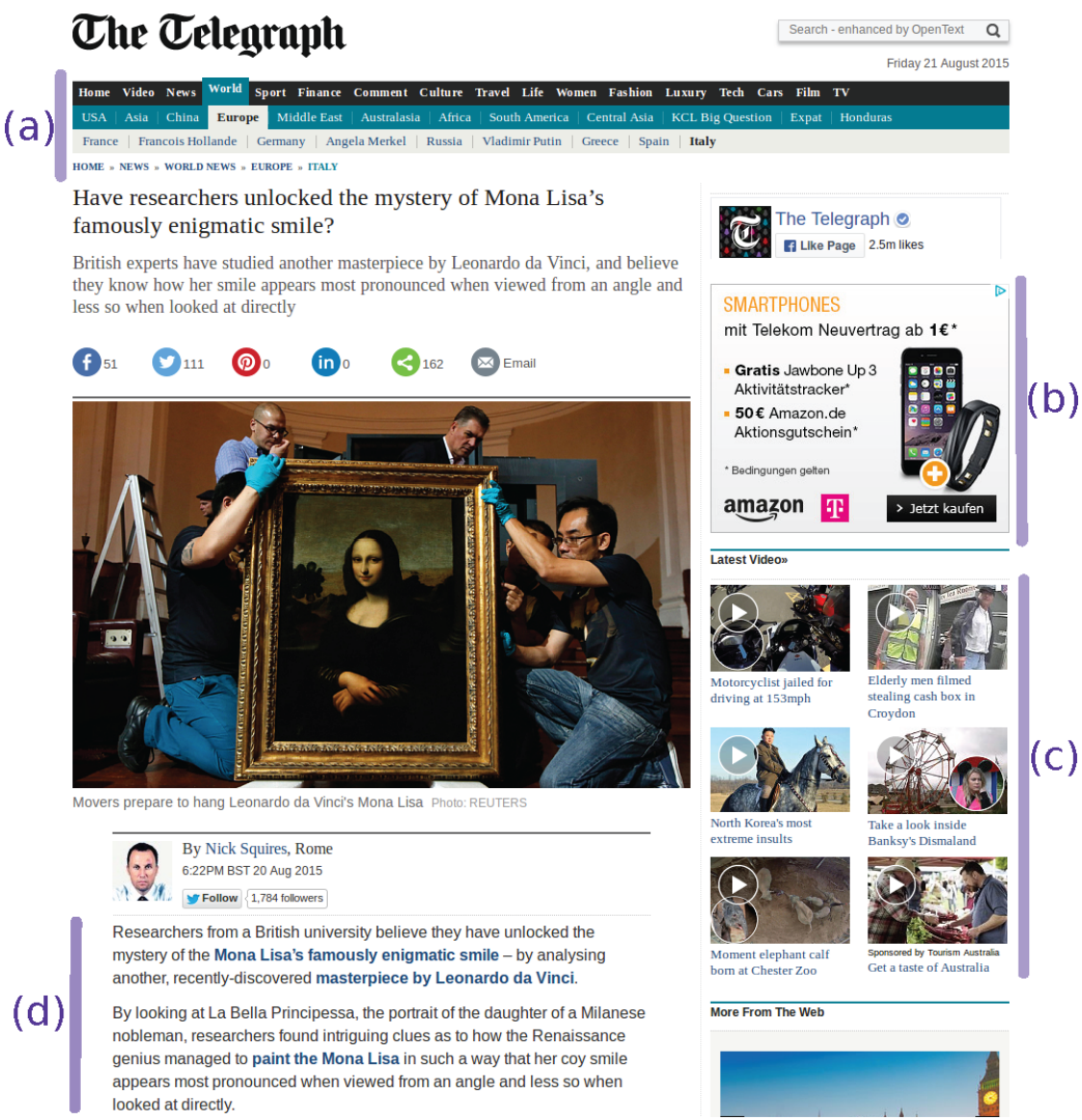
Andreas Spitz and Michael Gertz

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

gertz@informatik.uni-heidelberg.de

ASONAM
Paris, August 27, 2015

News Citation Networks



Classification of links by location and target:

a) navigational links

b) advertisement

c) internal links

d) **anchored references**

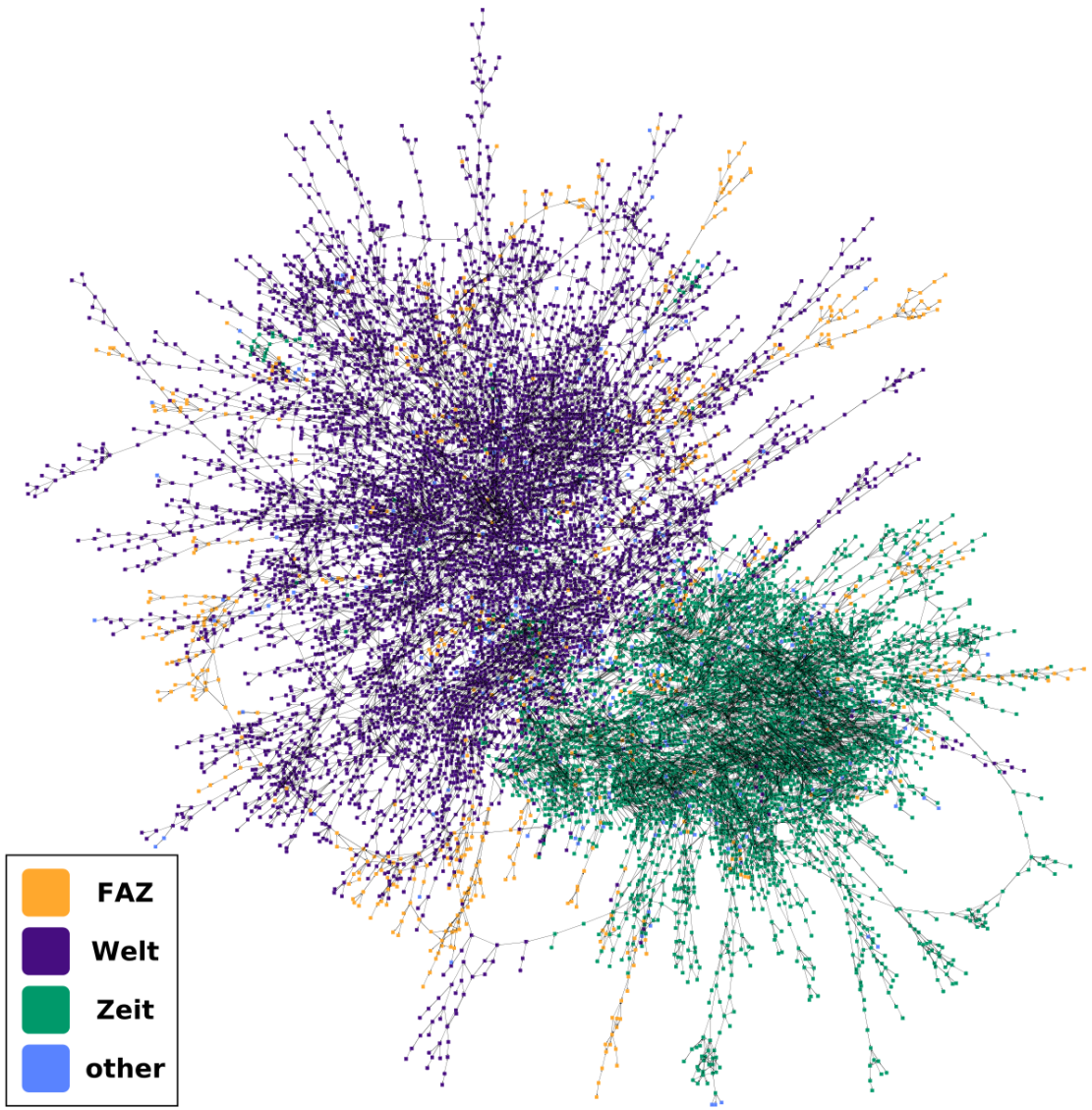
Objectives

- Construct news citation network from several news outlets, exploting anchored references ("semantic links") occurring in the main text of articles
- Investigate similarities and differences to "traditional" citation networks
- Develop and evaluate model for news citation network

Constructing the News Citation Network

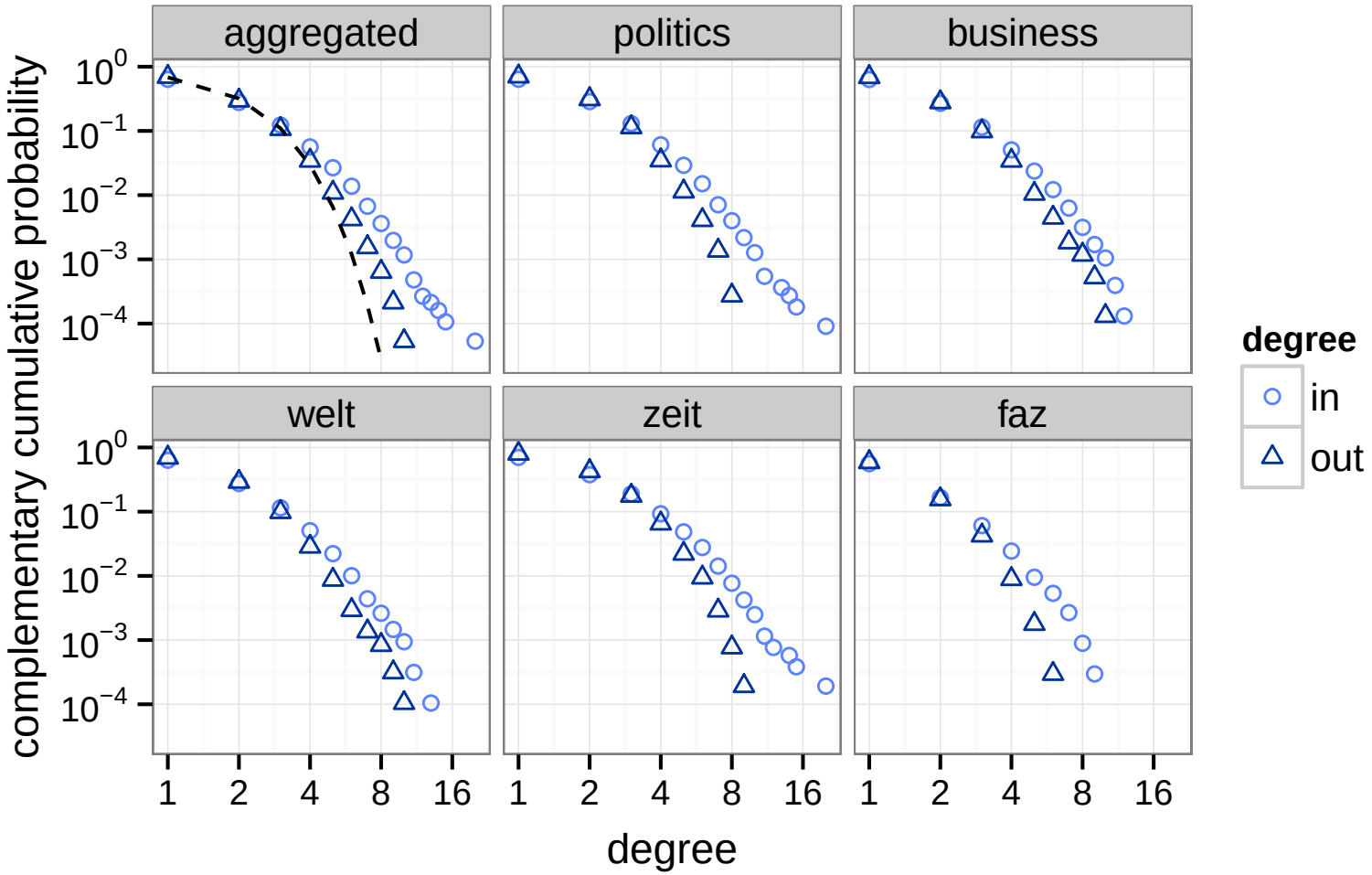
- Select a number of news outlets (*Zeit*, *FAZ*, *Welt*, *Spiegel*, *Tagesschau*) and categories (politics and business) during timeframe 6/2014-3/2015
- Employ RSS-feeds to obtain full articles
- Use outlet-dependent rules to extra article text and links within the texts as edges
- Record metadata, in particular article publication time
- Resulting network consists of 18,782 nodes (articles) and 21,581 directed edges

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

Degree Distribution



Structural Measures

network	$\|V\|$	$\|E\|$	cc	$\emptyset_d$	$\emptyset_u$	$\langle l_d \rangle$	$\langle l_u \rangle$
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 $\emptyset_u$, $\emptyset_d$ (un/directed) and average path lengths $\langle l_u \rangle$, $\langle l_d \rangle$.

Modularity and Assortativity

network	Q_{cat}	Q_{ol}	r	r_{ii}	r_{io}	r_{oi}	r_{oo}
aggreg.	0.39	0.57	0.25	0.13	0.16	**0.52**	0.19
politics		0.56	0.23	0.13	0.15	**0.51**	0.18
business		0.49	0.31	0.10	0.19	**0.53**	0.16

Modularity by category Q_{cat} and news outlet Q_{ol}, 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

Models for Citation Networks

Models and applications for citation networks are well established (e.g., de Solla Price (1965), Garfield (1972) and Hirsch (2005), Barabási and Albert (1999), Dorogovtsev and Mendez (2000))

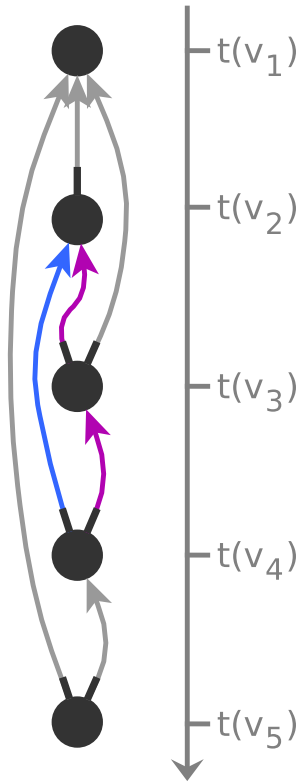
Models usually include:

- 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

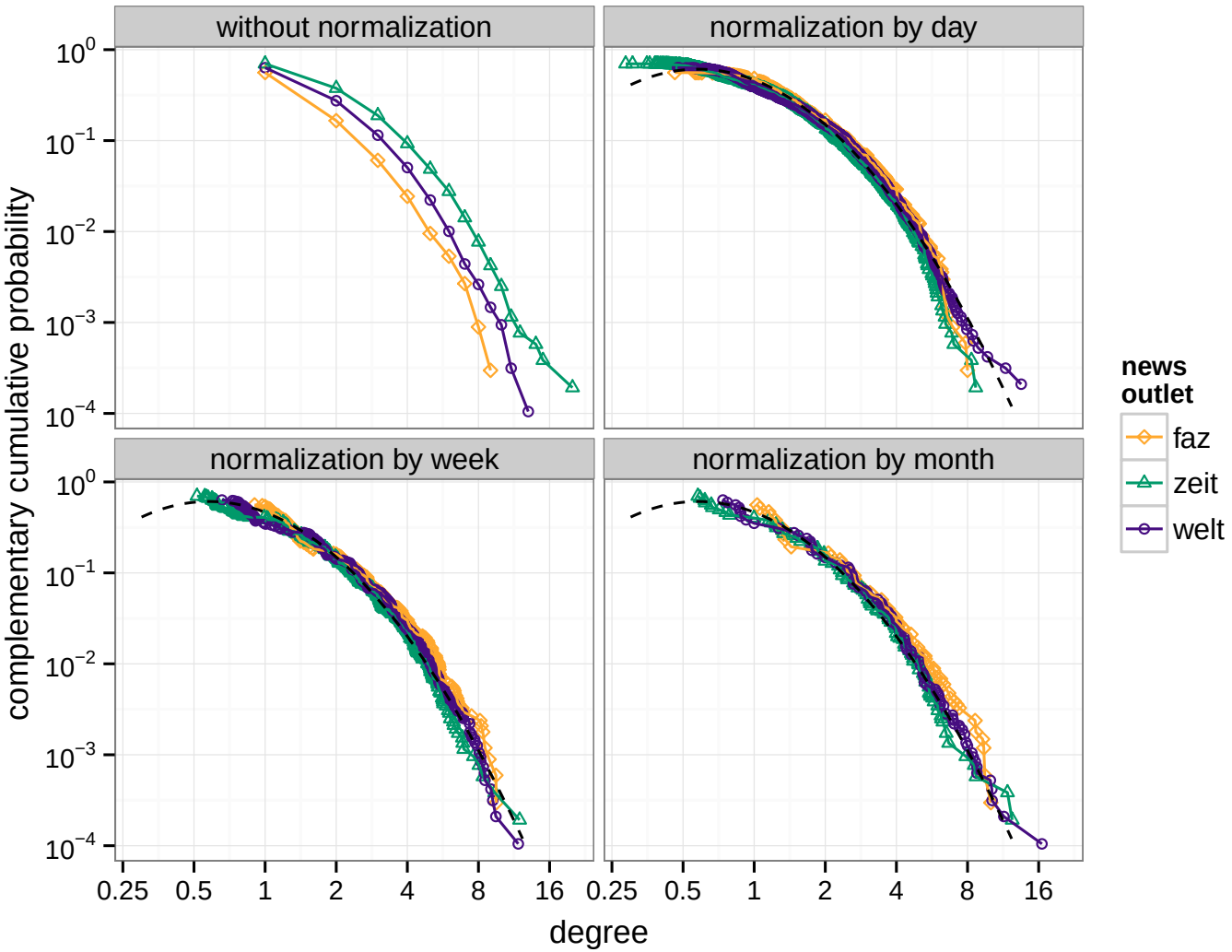
The nodes are sorted topologically. Outgoing degrees are fixed and parameters $\alpha \in \mathbb{R}$, $\beta \in [0, 1]$ are selected. New edges are then generated for each node v_i, starting with $i = 1$:

- **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)

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

Centrality in citation networks typically measures

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

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

Comparison to Crawled Networks

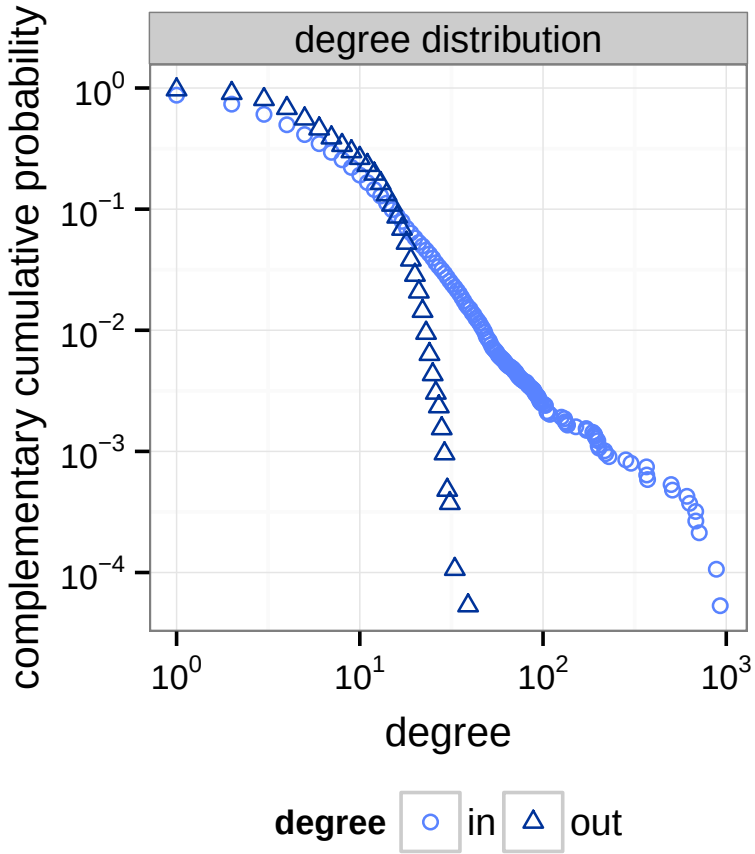
Construction of a traditional, crawled network

- over the same set of nodes (article pages)
- include all links, not just anchored references in articles' text

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 for a Crawled Network



Conclusions and Ongoing Work

- 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
- DAG-structure of the network allows for efficient analysis

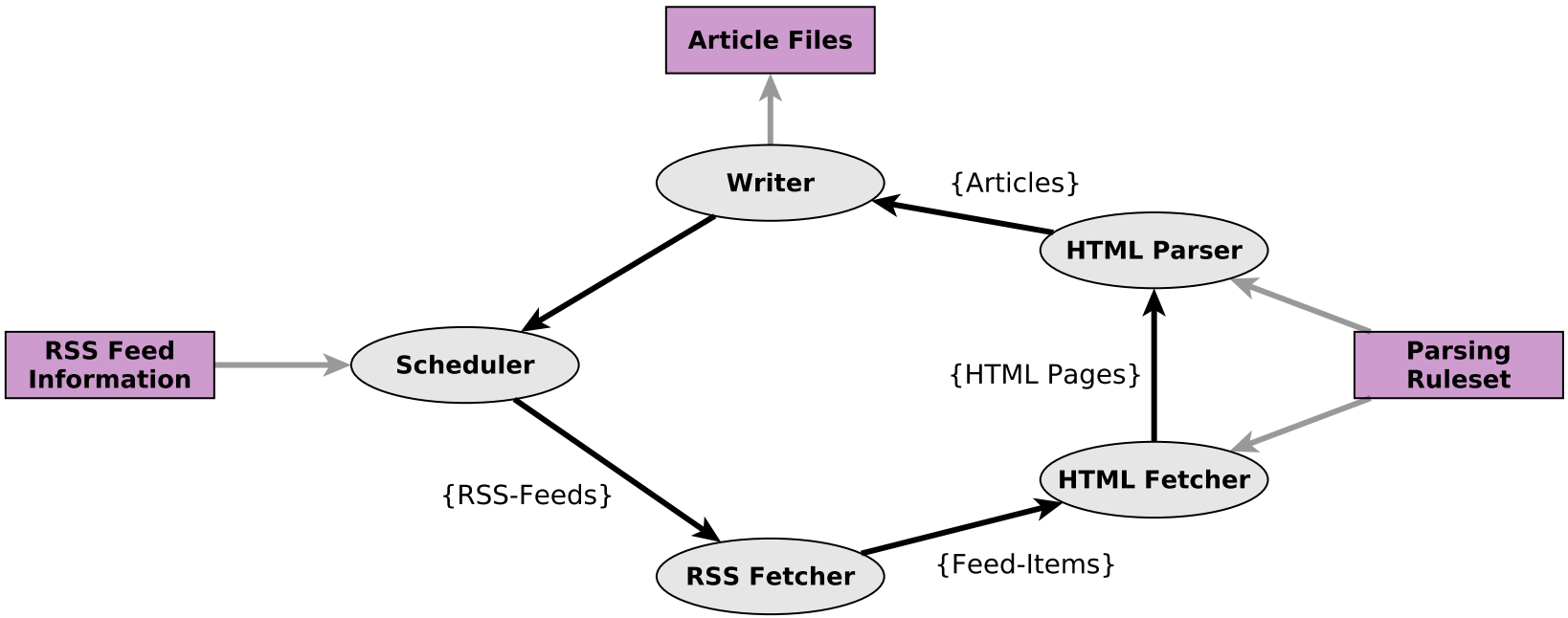
What's next?

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

Data:

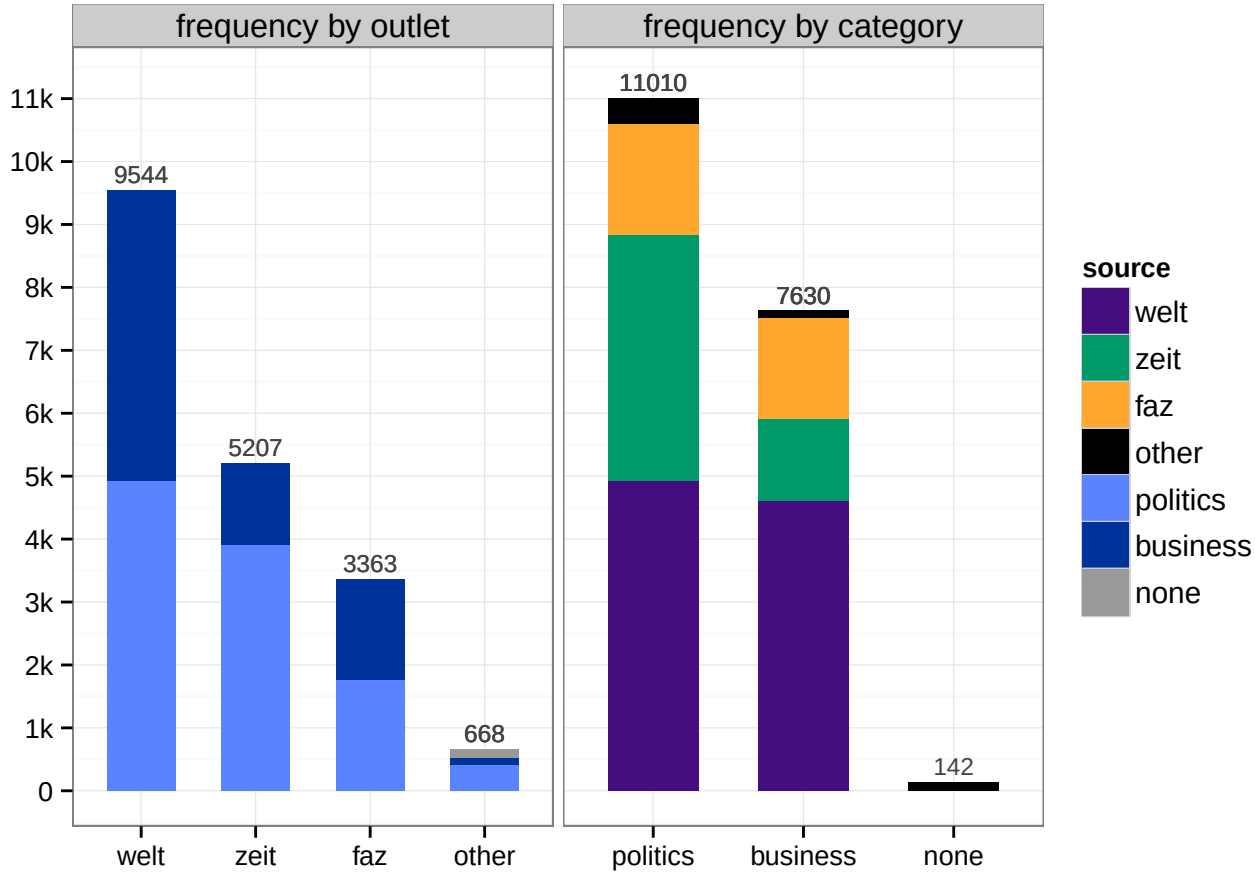
`http://dbs.ifi.uni-heidelberg.de/index.php?id=data`

RSS Aggregator



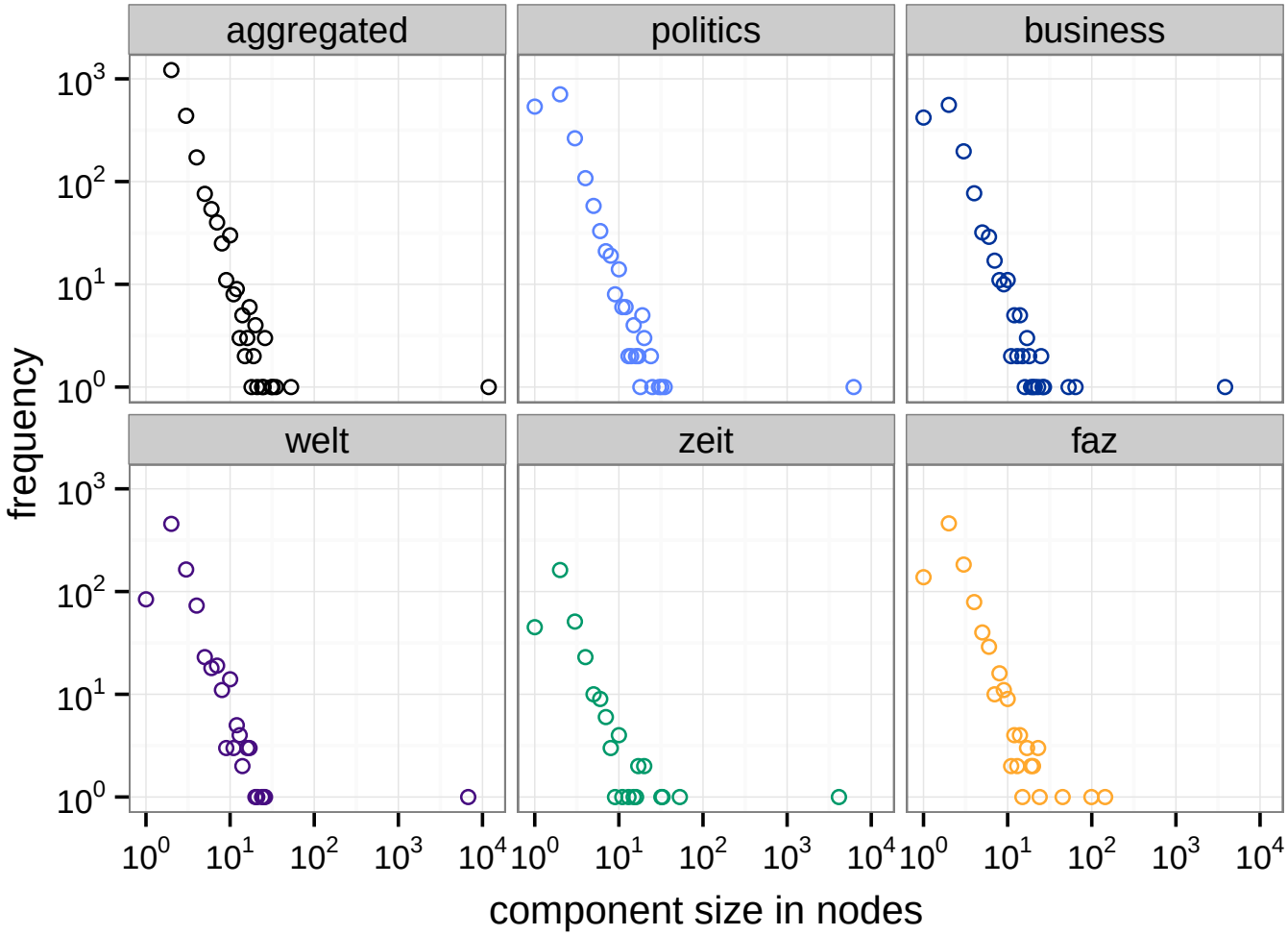
The News Citation Network

Data collected from 6 German news outlets from 6/2014-3/2015



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

Component Size Distribution

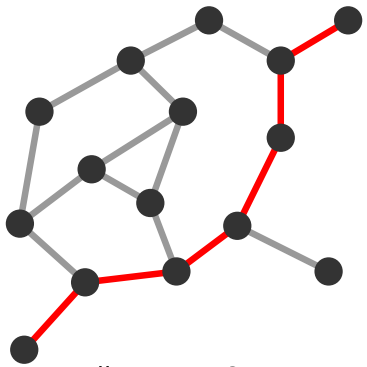


Structural Measures (Definitions)

Structural measures for a network:

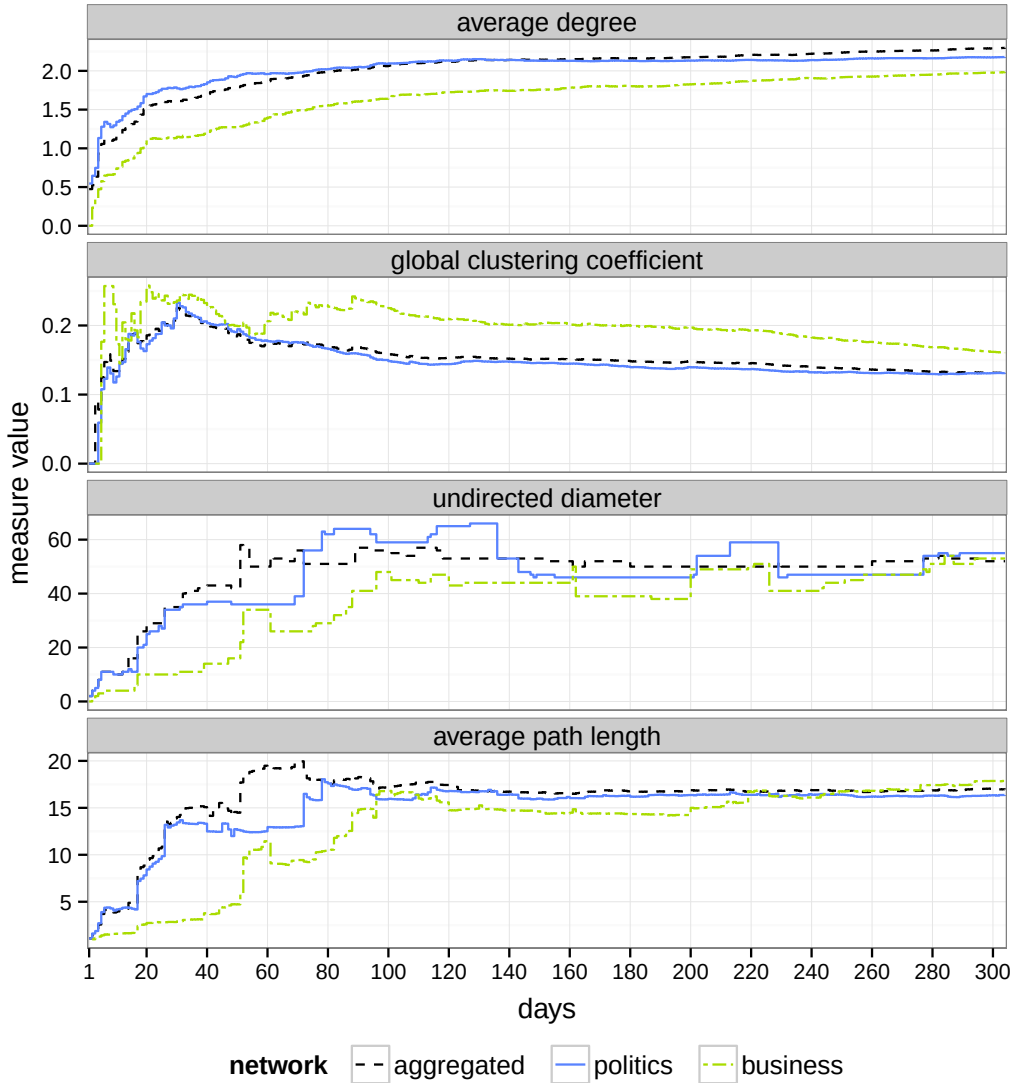
- **Average degree**: mean number of neighbours of a node in the network
- **Clustering coefficient**: $cc = \frac{3\Delta}{T}$
  Δ is the number of closed triangles
  T is the number of connected triples.
- **Diameter** $\varnothing$: the longest shortest path between any two nodes
- **Average path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

connected triple closed triangle



diameter: 6

Network Evolution



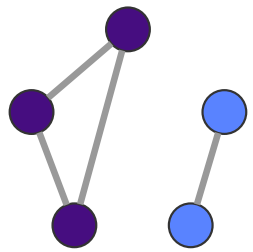
Modularity and Assortativity (I)

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

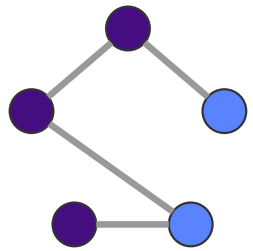
Where:

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

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



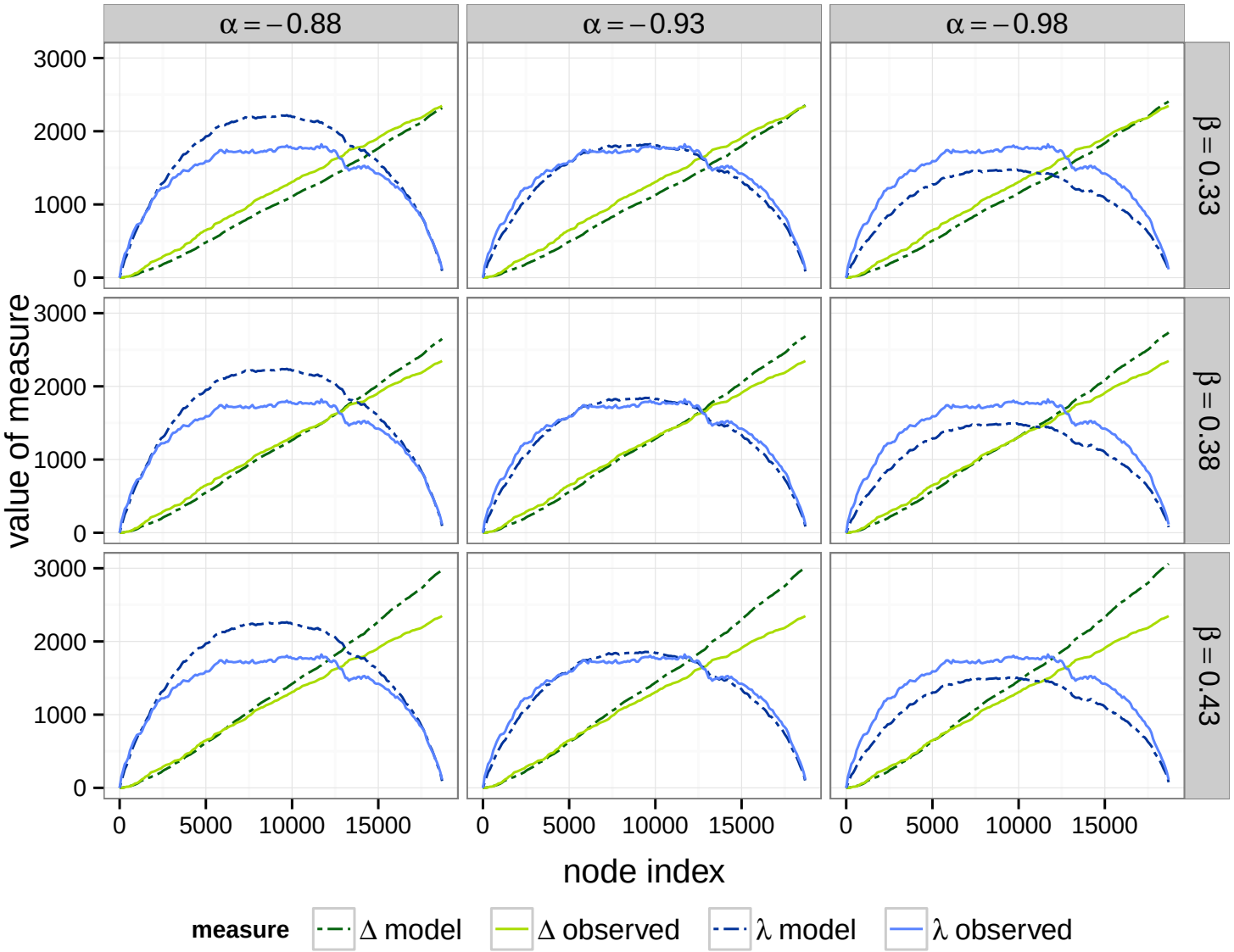
modular



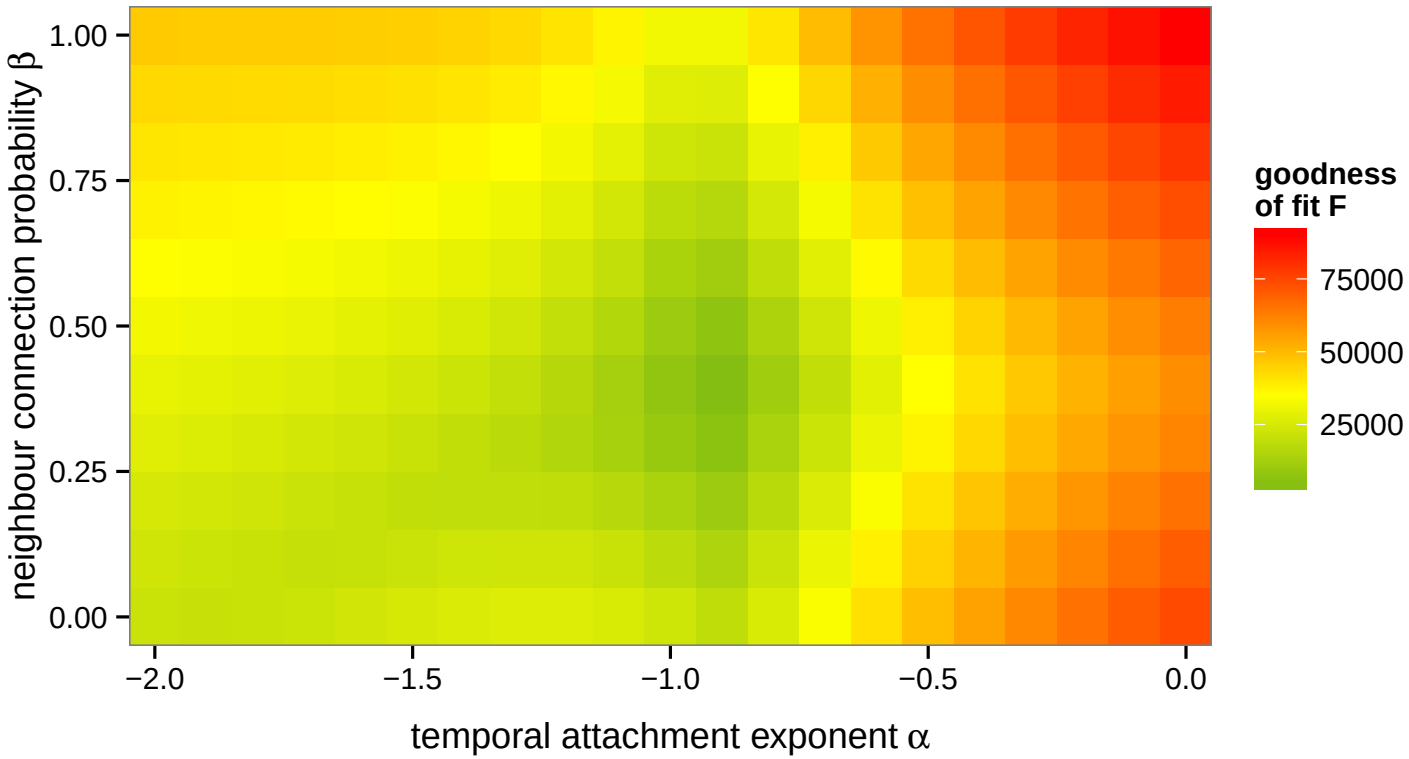
not modular

Newman (2003)

Fitting the Model (I)



Fitting the Model (II)



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

$\Rightarrow$ Attachment probability decays linearly with age

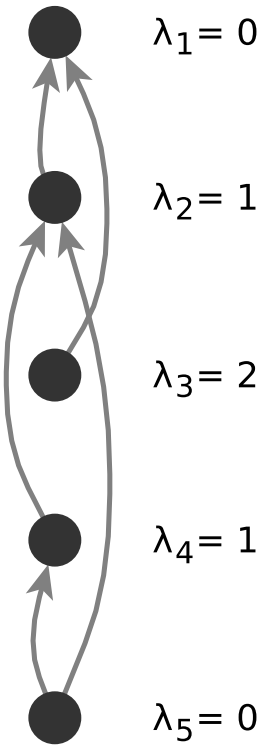
Goodness of Fit

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.



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