2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

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

Andreas Spitz and Michael Gertz
Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
Email: {spitz, gertz}@informatik.uni-heidelberg.de

***Abstract*—Networks of online news articles and blog posts are some of the most commonly used data sets in network science. As a result, they have become a vital piece of network analysis and are used for the evaluation of algorithms that work on large networks, or serve as examples in the analysis of information diffusion and propagation. Similarly, scientific citation networks are part of the bedrock upon which much of modern network analysis is built and have been studied for decades. In this paper, we show that the backbone inherent to networks of online news articles shares significant structural similarities to scientific citation networks once the noise of spurious links is stripped away. We present a data set of news articles that, while it is extremely sparse and lightweight, still contains information relevant to the propagation of information in mass media and is remarkably similar to scientific citation networks, thus opening the door to the use of established methodologies from scientometrics and bibliometrics in the analysis of online news propagation.**

I. Introduction

A common saying has it that a journalist never reveals his source. The reality looks different of course, and most of the time the source is not a secret informant in need of protection. A vast majority of online news articles reveal sources of information in the form of press agencies, blogs, social media or other news reports. Frequently, such stories are not entirely original and have been reported by other news outlets before, akin to the phenomenon of *resharing* in social media. In other cases, articles build upon previously published articles as a news story evolves over time. In either case, this relationship information is often included in a link that is placed within the subsequent article. Such links are semantically anchored within the surrounding text, which differentiates them from other types of links that can be found on the websites of news outlets that might also point to other news articles. While networks of interlinked online news articles, blog entries and microblogging posts have been studied extensively in the context of information diffusion [1], [2], [3], event aggregation and detection [4], spatial clustering [5] and the dynamics of mass media [6], [7], these analyses have largely been performed with a focus on the quantity of news articles. Although this focus is often necessary for the observed effect to emerge, it also requires the automated collection of data from thousands of different news outlets. As a result, the distinction between the different types of links seems to have gone unnoticed.

In this article, we argue that the deliberate act of placing a reference to another article within the text provides this reference with the character of a citation. We propose a method for the extraction of a network consisting of only such links, thus giving it the characteristics of a citation network. Citation networks, which are well researched in disciplines such as Scientometrics, have been used to rank the scientific performance of both journals [8] and scientists [9]. In network analysis, they have fuelled a host of research into models behind citation patterns and their emergence (for an overview, see [10]). At their core, citation networks consist of a set of scientific articles that share connections based on references from one paper to another. As such, they are perhaps the classic example of so-called scale-free networks that have played a pivotal role in the revived interest in network science. Since the data of many citation networks is available and well curated, they are one of the best researched types of scale-free networks. However, there also exist numerous instances of citation networks outside of science, and it has been shown that they can be analysed with similar methods and approaches. Examples are as diverse as networks of patent citations used to measure the flow of knowledge behind innovations [11], or legal decisions citing precedent cases [12]. Even artistic inspiration can be considered in this context, such as the multiplex network of movies that include quotes, props, scenes or entire sequences of original footage of other movies [13]. Here, we provide evidence that the *news citation network*, which results from considering only semantically anchored references between news articles, is a natural addition to this list. The analysis of such a network of online news can benefit from decades of research into citation networks and avoid common pitfalls that have not been considered in this context. Furthermore, we show that the careful exclusion of spurious links drastically reduces the network's size while improving the results of analyses such as network centrality. In an exploratory step, we give examples of how the news citation network can be used to derive information that benefits both those who seek to obtain information from the news as well as the publishers who create it. As an added benefit, the resulting network contains no proprietary content and can be published to ensure reproducibility and enable further research.[1]

II. Related Work

We are unaware of any research into the structural properties of different types of links between online news articles or analyses of semantically anchored links. There are, however, several recent works on the propagation of news between articles through links or textual similarity. Leskovec et al. explore a corpus of crawled newspaper and blog articles to track the spread of news through the web [7]. By breaking

[1] We provide and update this network, including URIs to the original articles, at `http://dbs.ifi.uni-heidelberg.de/index.php?id=data`

ASONAM '15, August 25-28, 2015, Paris, France

ACM ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2809380

the articles down into memes and phrases, they are able to perform a quantitative analysis on the dynamics of the global news cycle. In a similar approach, Flaounas et al. work with a multilingual corpus of news articles with respect to the structure of the European media sphere [6]. In contrast to our work, they do not consider explicit links but create connections through textual similarities. Gamon et al. aggregate links from blog posts to news articles to add emotional and social context to the news articles [14]. In a model for information diffusion through social media, Myers et al. study the role of news media as external influences on the process [2]. Lastly, there is a host of research into the emergence and prediction of information cascades through social networks and microblogs that rely on mechanics that can reasonably be applied to networks of news articles of proper dimensions (for an introduction, see [15]). All of the above approaches utilize large-scale data sets for which an exact differentiation between semantically embedded links and noise would be a challenge in itself.

For the extraction, aggregation and processing of news data, there exist a number of similar applications that retrieve data either by crawling the websites of news outlets or by monitoring RSS feeds for news releases. Lydia is a large scale aggregation and analysis tool for news articles and blogs that relies on scheduled crawling [16]. The system behind the European Media Monitor retrieves RSS feeds to build and process a repository of multilingual news articles from European news outlets [4]. In a similar approach, News Stand monitors and retrieves a large number of RSS feeds to extract geographic content from articles for spatial clustering [5]. All of these approaches are aimed at a large throughput and are thus based on heuristics for the automated extraction of text from a large number of sources. This would not be well-suited for the exact extraction of semantically anchored links, which is why we employ a rule-based approach.

III. Data

Our methods for the collection of news articles and the resulting network differ from those that are frequently employed in similar settings, leading to a unique network of news references. In this section, we describe the data collection and give an overview of the network's structural properties.

A. Data collection

For the collection of references between news articles, we observe significant differences in the nature of hyperlinks on such web pages. Independent of the news outlet in question, we can identify four different types of links. *Navigational links* provide a menu structure and allow the user to browse the news outlet's web page. Links in *advertisement sections* lead to third-party web sites. *References* are found within the article text and point to sources or articles that are relevant to the article's topic, while *internal links* lead to a selection of other articles that were published by the same outlet. Of these, only references possess a semantic anchor to the topic of the news article, while the others are potential noise. Ideally, a network of news citations should thus consists exclusively of references. However, commonly employed approaches for the creation of news networks cannot accurately distinguish between the different types of links due to the sheer volume of collected articles. In an approach that emphasizes quality over quantity, we instead use a rule-based system that allows the extraction of an article's content and the identification of only those references to other news articles that are found within the text. We do this by subscribing to the RSS feeds of a selection of news outlets and applying a simple set of extraction rules that define the location of text content within the web page's HTML tree. Rules for identifying links to subsequent pages of multi-page articles ensure that no information is neglected. While this approach would also work with a crawler instead of an RSS-Listener, the extraction of accurate publication times would be problematic. Based on the article texts collected in this way, the extracted links and the publication time, we construct a network of articles in two steps. First, we iterate over the set of all files and map each article to its URI address on the web, which works well as unique identifier. These are now considered to constitute the set of nodes of the network. While doing this, we store the metadata for each node, including the time of publication, the title and the URIs of links located in the article text. In this step, we also remove articles with no relevant text (such as videos) and articles with identical text that are republished under a different URI. In a second step, we then iterate over all created nodes to check for each stored link if its URI matches another node in the network and create an edge if it does. In a final step, the set of nodes is reduced to nodes that are incident to at least one edge. As a result, we obtain a directed graph $G = (V, E)$, where V constitutes the set of news articles, and we include a directed edge $(v \to w)$ in E if the text of article v contains a hyperlink to article w. We include as node attributes the time of publication, the news outlet that released the article and the category of the RSS feed that was used to publish the article.

B. Data sources

We used the RSS aggregation strategy described above to collect data from six major German news sources between June 2014 and March 2015, namely *Zeit*, *Welt*, *Frankfurter Allgemeine Zeitung* (FAZ) and *Süddeutsche Zeitung* (SZ), which are influential newspapers, as well as *Spiegel*, a weekly news magazine, and *Tagesschau*, which is the best known news show on public broadcasting. In the following, we refer to all news sources as *news outlets*. For all listed news outlets, we retrieved the *politics* and *business* feeds. After the removal of duplicates based on article text or article URI and articles with no text, we collected a total of $58,753$ articles. From this set of articles, we generate the news citation network, which is significantly smaller than the entire data set and contains only about a third of the articles, resulting in a size of $|V| = 18,782$ nodes and $|E| = 21,581$ edges. This decline is due to two factors: on the one hand, some articles contain no anchored links or are not targeted by links. On the other hand, the data set is restricted to six news outlets and many articles reference data sources outside of our current network, because they were published either by international or smaller German news outlets or prior to June 2014. We also find that not all news outlets are represented equally, with articles from *Spiegel*, *SZ* and *Tagesschau* making up only a small fraction of nodes in the network (3.6%). In the following, we therefore focus on *Zeit*, *Welt* and *FAZ* as individual news outlets. A small number of edges are anachronistic (197) or self loops (6) and are removed from the data. The sizes of subnetworks by news outlet and category are shown in Table I.

TABLE I. GRAPH METRICS FOR THE NEWS CITATION NETWORK AND ITS SUBNETWORKS BY NEWS OUTLET AND CATEGORY. SHOWN ARE THE GLOBAL CLUSTERING COEFFICIENT cc AS WELL AS THE DIRECTED AND UNDIRECTED DIAMETERS $\varnothing$ AND AVERAGE PATH LENGTHS $\langle l \rangle$.

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

C. Network structure

With slightly more edges than nodes, the network is extremely sparse yet still mostly connected, since 63.1% of its nodes form a giant connected component. This is also the case for all subnetworks by article category and news outlet, with the exception of *FAZ*, which decomposes into small components. As a directed network of references between articles with publication times, the news citation network is a directed acyclic network and contains only weakly connected components. In Table I, we show basic graph metrics for all involved subnetworks. Due to the time ordering of the nodes, edge directions are implicit and we thus also include a number of undirected graph metrics. We find a clustering coefficient for all subnetworks that is remarkable, given the network's sparseness. Both the directed and undirected diameter are fairly large, indicating that components are only loosely connected. The directed diameter of the business subnetwork is significantly smaller than the undirected diameter, which attests for the absence of longer citation chains that can be found in the politics subnetwork. The average path lengths are very large when compared to similar networks [17]. In Figure 1, we show the distribution of in- and out-degrees, i.e., the number of received and given citations. The degrees are much smaller than one would expect from a scientific citation network, which is not surprising due to the faster pace of news reporting. Given the overall small degrees, it would be questionable to attribute a long tail to the network and we observe that the maximum degree is lower than what one would expect from a pure preferential attachment model by an order of magnitudes. In direct comparison to an Erdős-Rényi model, however, we find that the frequency of nodes with larger degrees is higher by a substantial margin. Similarly to scientific citation networks, the distribution of out-degrees drops more sharply than the distribution of in-degrees, which reflects constraints in article space and time investment by the author.

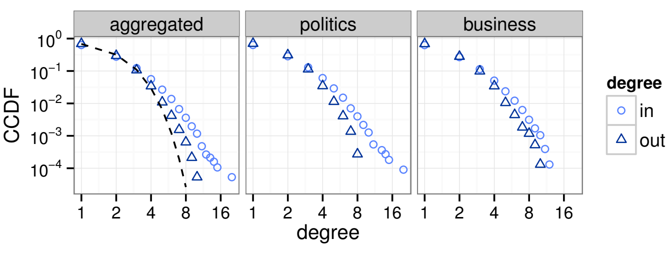

Fig. 1. Distribution of degrees for the aggregated network and the subnetworks that are induced by article feeds, given as the complementary cumulative degree distribution $P[d \geq d_x]$. The degrees of nodes in subnetworks include edges to and from nodes of different types. The dashed line represents the expected distribution for an Erdős-Rényi graph of the same size.

Lastly, we give a summary of the mixing of articles by external attribute and degree. We consider the modularity by the node attributes *category* and *news outlet* and the assortativity by degree [18]. Since the network is directed, we also include directed assortativity coefficients for the in- and out-degree sequences [19]. Assortativity equates to a Pearson correlation of degrees along edges, which means that there exists one coefficient for undirected networks and four coefficients that result from all possible combinations of in- and out-degrees in directed networks. In the emergence of assortativity, the network's underlying model is a major factor. In a lattice graph or network consisting primarily of cliques for example, one would not expect to find assortative mixing. It is therefore advisable to consider assortativity scores in relation to a suitable graph model. Here, we compare them to values obtained for a triadic closure model for citation networks (see Section IV) and show the results in Table II. We find both modularity scores to be high, indicating that most references occur within a given outlet or category. For the undirected assortativity, we observe a relatively high score across all subnetworks. For directed assortativity scores, we find almost no evidence of assortative mixing in the aggregated network, relative to the model. While the *out-in* assortativity is substantial, it is perfectly matched in the model, which indicates that this is an integral element in the network's growth. The overall lack of assortative mixing in the aggregated network is an indication that there are other driving forces behind edge creation, such as the evolution of the news stories that are being referenced. For the subnetworks by category, we find that assortative mixing is very different from the aggregated network and that there are large differences between the two subnetworks as well. While the observed values deviate only slightly from the aggregated network, the expected values show substantial variation. For both the *in-in* and *out-out* assortativities, the expected values are slightly disassortative, meaning that the observed networks show a much stronger correlation between nodes with similar in- or out-degree than one would expect. Also noteworthy is the *in-out* assortativity for the business network, which falls short of the model despite being substantial. As a result, the business network is the only network in which we find a tendency of popular articles referencing articles with few references and articles with few incoming references referencing survey articles with many references. Overall, the business network displays a lot less cliquish features than the politics network, which may indicate that news are more broadly interconnected here and less bound to specific news stories.

TABLE II. MODULARITY OF THE NETWORK BY CATEGORY Q_{cat} AND NEWS OUTLET Q_{ol} AS WELL AS ASSORTATIVITY BY DEGREE r AND THE FOUR DIRECTED ASSORTATIVITY COEFFICIENTS $r_{in,in}$, $r_{in,out}$, $r_{out,in}$ AND $r_{out,out}$. SHOWN ARE VALUES FOR THE OBSERVED SUBNETWORKS (obs), THE MODEL (mod) AND THE DIFFERENCE (δ) BETWEEN THE TWO.

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

IV. Citation Characteristics

We consider in detail a number of aspects of the news citation network that link it to citation networks. Key findings are a mechanism of decaying preferential attachment by age, the unique evolution of characteristic network measures as the network grows, and the universality of citation distribution.

A. *Citation age and preferential attachment*

A common factor in network evolution is preferential attachment, which has been linked to the age of articles in the case of citation networks (see, e.g., [20] and [21]). With the rapid turnover of news cycles, this is likely an important factor in the emergence of news references as well. Due to the limited number of expressed degrees, asymptotic models for citation networks with broader degree distributions are ill-suited for the news citation network. Furthermore, we would like to not only find a model that fits the network, but also to identify the rate of preferential attachment by age while preserving structural properties of the network. We therefore use a parametrized triadic closure model for citation networks that fixes the out-degree sequence and approximates the in-degree sequence by adding nodes according to a topological ordering [22]. In this model, nodes are added sequentially and the number of transient edges λ and closed triangles Δ are fitted at each step, where

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

for a topological ordering $v_1, \cdots, v_{|V|}$ of nodes. The model is fitted by evaluating a goodness of fit for Δ and λ values over a sample of graphs generated by varying two parameters α and β, where β denotes the probability of attaching a new edge to a neighbour of a previously selected node and α controls the preferential attachment probability with age between nodes v_i and v_j as

$$\Pi_{(i \to j)} \sim (time(v_i) - time(v_j))^{\alpha}.$$

By performing a grid search over the parameter space of $\alpha \in [-2, 0]$ and $\beta \in [0, 1]$, we obtain an estimate of the role that preferential attachment by age plays in the news citation network without having to rely on asymptotic models for traditional citation networks. For each parameter combination with a granularity of 0.01, we sample $1,000$ graphs from the model and compute the average goodness of fit as the sum of relative errors for each added node. For the best fit with a value of $\alpha = -0.93$, we find evidence of decaying preferential attachment with age that is almost linear and only slightly less pronounced than the decay that Wu and Holme originally obtained for a citation network of Physics papers ($\alpha = -1.0$) [22]. We observe that the tendency $\beta = 0.38$ to attach edges to neighbours of previous nodes and increase triangle formation is significantly smaller in the news citation network than the value of they obtained ($\beta = 0.99$). We attribute this to the low density of the news network, for which a high value of β would result in a disconnected network.

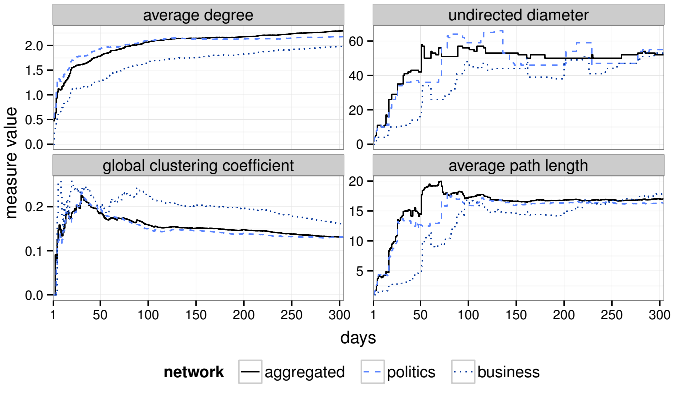

Fig. 2. Evolution of network metrics as a function of network age for the aggregated network and the subnetworks that are induced by article categories.

B. *Network evolution*

For time-ordered networks, network evolution is a natural topic of interest. In the case of networks for which preferential attachment plays a major role in their evolution, the clustering coefficient is known to increase over time [23]. The opposite is true for the diameter, which shrinks in the case of many evolving networks, including citation networks [24]. We therefore investigate the evolution of a number of measures that are frequently studied in this context and show the results in Figure 2. We observe that the average degree slowly increases with the addition of new nodes, yet the clustering coefficient, diameter and average path length are constant over most of the time frame. On the one hand, we attribute this unexpected result to the sparseness of the network, where new edges are introduced at a marginally faster rate than nodes. More importantly, however, in combination with the high modularity and assortativity, this is evidence of hierarchical structures around news events and news outlets in the network. Local citation structures around time-limited news events are responsible for the high clustering coefficient and assortative tendencies, leading to a hierarchical structure similar to that of research topics in scientific citation networks [25].

C. *Universality of citation distribution*

In Scientometrics, much research has been devoted to measuring the impact of scientific publications. One key finding in this debate is the universality of citation distributions [26]. While the distribution of received citations per article varies between different fields of science, this variance largely depends on the different rates of publication in these fields. As a result, a normalization of the in-degree d of a publication with the total number of received citations d_0 for any publication in the given field in the same year yields a homogeneous distribution of the indicator $d_f = d/d_0$ that can be approximated by a log-normal distribution. In Figure 3, we show that this finding is also true for the news citation network if we consider individual news outlets as fields and normalize by days. Similar results for a normalization by week or month (not shown) indicate that the time frame that is used for normalization has only a negligible effect on the result.

Combining the above observations, we find that the news citation network bears striking resemblance to known types of citation networks. Due to it's inherent sparseness, however, it differs in certain areas such as it's evolution, which makes it an interesting border case, not just for the study of information propagation in news but also for bibliometric methods.

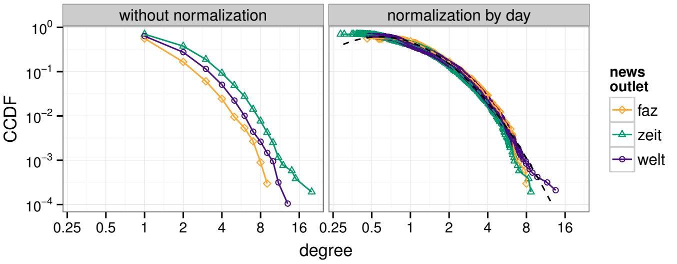

Fig. 3. Distribution of in-degrees (left) as $P[d \geq d_x]$ and the distributions of normalized degrees $d_f = d/d_0$ for individual news outlets over different time intervals, where d_0 is the total amount of received citations for the given news outlet during the day of release (right). The dashed line denotes the fit of a log-normal distribution with $\mu = 0$ and $\sigma = 0.86$.

V. Experimental Results

In the following, we investigate if useful information about the centrality and role of articles and news outlets can still be derived from the news citation network, despite its low density in comparison to traditional networks of online news articles. Based on Borgatti's criteria for selecting centrality measures [27], degree centrality and influence-based centrality measures are best suited for citation networks. Due to the small range of degrees, however, we find that degree-based centrality is only able to identify top-cited articles and quickly looses the ability to discern between articles of lower ranks. As an influence-based measure, we therefore use PageRank, which lends itself naturally to news articles and citation networks. We find that the top ranked articles consistently concern the prevalent topics of 2014 and 2015, such as the Ukraine crisis or events in the Middle East. Unsurprisingly, PageRank favours older articles due to the accumulation of references. It also favours political over business articles, which represent only 18% of the top ranked 100 articles, even though business articles make up 40.6% of all nodes. Given the more obvious influence that politics has on business news, this is reasonable.

A. Centrality profiles

To investigate the role of news outlets in the network, a direct comparison of article centralities is infeasible. Instead, we analyse the performance of outlets as a function of the number of highly ranked articles they publish. We order all articles by PageRank centrality and consider a growing set of the k highest ranked articles. For each news outlet, we compute the fraction of the outlet's articles that can be found among the first k articles. For a perfectly random ranking, this results in a linear advancement along the diagonal for all news outlets. Deviations from the diagonal thus correspond to a positive or negative performance for the respective outlet. The results are shown in Figure 4 for the aggregated network and subnetworks by article category. We observe a clear pattern across all three subnetworks, where *Zeit* performs well over the entire data. In the business network, the performance drops, likely due to the smaller number of business articles this outlet published overall, indicating that not only low influence articles are being cut. *Welt* performs averagely, while *FAZ* is trailing behind. Although these results are exploratory and objectively evaluating the true importance of news articles is infeasible, they indicate that the news citation network captures essential information about the flow of news between articles.

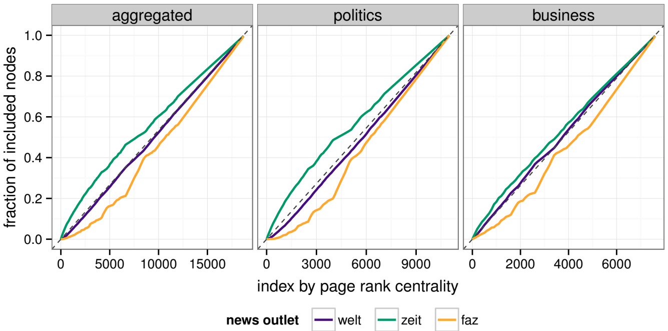

Fig. 4. Fraction of articles by news outlet that are included among the k highest ranked articles by PageRank centrality. The index k is increased by steps of 100 and all values are averaged over 100 runs to account for ties in the ranking. The dashed line denotes the performance of a random ranking.

B. Comparison to a traditional news network

As a final observation, we present a comparison of the news citation network to a more traditional network of news articles. To this end, we construct a network consisting of the same set of nodes V, but also include navigational and internal links as edges by parsing the entire HTML document instead of just the article text. After removing self-loops, this results in $|E'| = 128,364$ edges, thus forming a much denser network. Despite six times as many edges, the network's clustering coefficient increases only slightly to $cc = 0.182$. The directed diameter and average path length both increase drastically ($\varnothing_d = 112$ and $\langle l_d \rangle = 26.7$, respectively) while the undirected diameter and average path length drop ($\varnothing_u = 16$ and $\langle l_u \rangle = 5.9$). The distribution of in-degrees has a distinct long tail, while the out-degrees are clearly bounded (see Figure 5).

If we consider centrality scores on this network, we find that the results deviate substantially from those we obtain for the news citation network. While the overall correlation is fair with a Spearman coefficient of $\rho = 0.43$, we find only a 12.3% overlap between the two networks among the 1000 top ranked nodes by PageRank. As we show in Figure 5, this is also reflected in the centrality profiles of news outlets, in which *FAZ* and *Zeit* trade places and which show no meaningful distinction between news outlets for the first 2000 top ranked articles. The number of business articles among the top 100 ranked articles rises to 43% and we find many results of low relevance ranked highly, such as articles pertaining to the launch of Netflix in Germany (rank 3) or a general exposé about natural resources in Mongolia (rank 20).

We have to note that, while it is necessary for the sake of comparability, the restriction to links between nodes in V is artificial in comparison to a crawl of news websites that would include further articles. However, this restriction is likely to work in favour of the traditional approach, since the exact limitation to only news articles is not possible simply by crawling. Furthermore, a crawler would likely inject anachronistic edges into the data set when parsing an article with a publication date in the past, thus destroying the directed acyclic nature of the network that signifies the flow of information in only one direction. In summary, we find that there is no actual benefit in the inclusion of internal and navigational links as edges in a network of news articles, when the backbone of semantically anchored links can be extracted from article content.

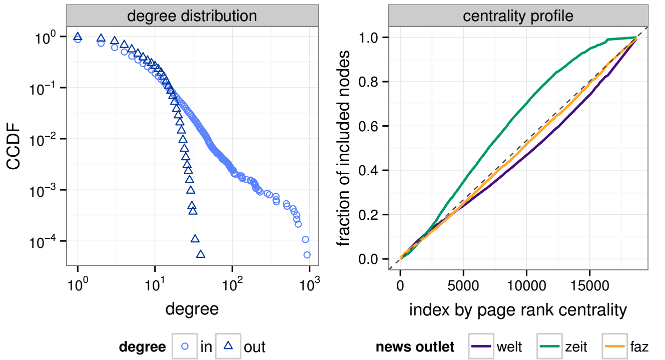

Fig. 5. Degree distribution and centrality profiles for a traditional network of news articles that also includes navigational and internal links over the same set of nodes as the news citation network.

VI. Conclusions and Ongoing Work

In this article, we have presented a network of news citations as an example of extremely sparse citation networks that can be extracted by focussing on semantically anchored links in the articles' texts. We have identified structural similarities to scientific citation networks and shown that networks of news articles can be treated and analysed as citation networks if properly pre-processed. We found evidence of preferential attachment by age that is only slightly less pronounced than in scientific citation networks. With regard to network evolution, we identified features such as the constant clustering coefficient, diameter and average path length that cannot be explained by preferential attachment alone and indicate a hierarchical graph structure. We have shown that the universality of citation distributions applies to the news citation network and that the time frame used for this normalization is of no relevance, an observation which cannot be easily obtained for the more coarsely grained scientific citation networks. We have provided empirical evidence that, despite it's sparseness, the network is still connected and allows for analyses with regard to the flow of information. Finally, we have shown that our method of network construction yields better results with respect to network centrality than traditional crawled networks of news articles resulting from automated extraction.

A central benefit of our approach, aside from the justification for using bibliometric methods on news networks, is the decrease in computational overhead for operations on the network due to its acyclic structure and sparseness. The extraction of further node attributes, such as authors, news agencies and user comments, can then provide information about the flow of information to both authors and users alike. Similarly, the possibility of an extension to other networks of information such as blogs or social media is obvious, given the ties that news outlets themselves already create between traditional news articles and the social media.

Currently, we are expanding the collection of articles to further and international news outlets, as well as reader comments and links to social media. We plan the inclusion of a semi-supervised approach at extracting candidate rules to include a larger number of news outlets for future studies. A comprehensive news citation network may then enable the analysis of information cascades through traditional media.

References

[1] M. Cha, J. Pérez, and H. Haddadi, "Flash floods and ripples: The spread of media content through the blogosphere," in *ICWSM '09*, 2009.

[2] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *KDD '12*. ACM, 2012, pp. 33–41.

[3] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *ICDM '10*. IEEE, 2010, pp. 599–608.

[4] M. Atkinson and E. Van der Goot, "Near real time information mining in multilingual news," in *WWW '09*. ACM, 2009, pp. 1153–1154.

[5] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling, "Newsstand: A new view on news," in *SIGSPATIAL '08*. ACM, 2008, p. 18.

[6] I. Flaounas, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, "The structure of the EU mediasphere," *PloS one*, vol. 5, no. 12, p. e14243, 2010.

[7] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *KDD '09*. ACM, 2009, pp. 497–506.

[8] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, vol. 178, no. 4060, pp. 471–479, 1972.

[9] J. E. Hirsch, "An index to quantify an individual's scientific research output," *PNAS*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[10] F. Radicchi, S. Fortunato, and A. Vespignani, "Citation networks," in *Models of Science Dynamics*. Springer, 2012, pp. 233–257.

[11] A. B. Jaffe and M. Trajtenberg, *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press, 2002.

[12] J. H. Fowler and S. Jeon, "The authority of supreme court precedent," *Social networks*, vol. 30, no. 1, pp. 16–30, 2008.

[13] A. Spitz and E.-Á. Horvát, "Measuring long-term impact based on network centrality: Unraveling cinematic citations," *PloS one*, vol. 9, no. 10, p. e108857, 2014.

[14] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. König, "Blews: Using blogs to provide context for news articles." in *ICWSM '08*, 2008.

[15] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[16] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *SPIRE '05*. Springer, 2005, pp. 161–166.

[17] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[18] M. E. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[19] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski, "Edge direction and the structure of networks," *PNAS*, vol. 107, no. 24, pp. 10 815–10 820, 2010.

[20] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks with aging of sites," *Phys Rev E*, vol. 62, no. 2, p. 1842, 2000.

[21] K. B. Hajra and P. Sen, "Modelling aging characteristics in citation networks," *Physica A*, vol. 368, no. 2, pp. 575–582, 2006.

[22] Z.-X. Wu and P. Holme, "Modeling scientific-citation patterns and other triangle-rich acyclic networks," *Phys Rev E*, vol. 80, no. 3, p. 037101, 2009.

[23] B. Bollobás and O. M. Riordan, "Mathematical results on scale-free random graphs," *Handbook of graphs and networks: from the genome to the Internet*, pp. 1–34, 2003.

[24] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *KDD '05*. ACM, 2005, pp. 177–187.

[25] E. Mones, P. Pollner, and T. Vicsek, "Universal hierarchical behavior of citation networks," *J. Stat. Mech. Theor. Exp.*, vol. 2014, no. 5, p. P05023, 2014.

[26] F. Radicchi, S. Fortunato, and C. Castellano, "Universality of citation distributions: Toward an objective measure of scientific impact," *PNAS*, vol. 105, no. 45, pp. 17 268–17 272, 2008.

[27] S. P. Borgatti, "Centrality and network flow," *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.