Web archives and hyperlink analyses: The case of videnskab.dk 2009–2022

Niels Brügger, Katharina Sølling Dahlman

Abstract: This chapter demonstrates how the use of a national web archive in hyperlinked network analyses may prove an indispensable source when conducting not only historical but also contemporary analyses of a given website. Our analyses are based on the case of videnskab.dk, a Danish journalistic website disseminating research-related knowledge to the public. Focus is on the examination of hyperlinks related to videnskab.dk in the years of 2009, 2014, 2018, and 2022, followed by a network analysis of videnskab.dk in relation to similar transnational websites. Our results showcase what insights may be gained when conducting analyses with and without access to a national web archive, respectively, highlighting the impact and importance of data collections when studying the online web..

Keywords: web archive, hyperlink network analysis, actor types, historical analysis, contemporary analysis.

1. Introduction

This chapter investigates how the holdings of a national web archive can be used to shed light on the hyperlinks related to one individual website. The study explores the case of the Danish science website videnskab.dk, and it is primarily based on content from the national Danish web archive Netarkivet. Videnskab.dk is a journalistic website that disseminates research-related knowledge to the wider public, like scientificamerican.com in the US, futura-sciences.com in France, and scinexx.de in Germany, and it has been chosen as a case because historical hyperlink network analyses of the website were conducted as part of a larger evaluation project of the many activities of the website (explained in more detail below).

The aim of the following is twofold. Firstly, to provide empirical results about the historical development of the hyperlinks related to the website videnskab.dk, with a focus on the changing main actor types to which it is connected. Secondly, to showcase that the archived web is not only useful for historical studies, but it is also an indispensable source type for contemporary analyses, in particular hyperlink analysis because web archives are (probably) the only place where in-links to any given website can be found, in contrast to out-links that are known to the website owner, and that can be collected from the website itself on the online web. As part of the latter aim a transnational hyperlink network analysis of the online web is included to highlight what could be done if national web archives

Niels Brügger, Aarhus University, Denmark, nb@cc.au.dk, 0000-0003-1787-1980
Katharina Sølling Dahlman, Aarhus University, Denmark, katharina@j-p.dk

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Niels Brügger, Katharina Sølling Dahlman, *Web archives and hyperlink analyses: The case of videnskab.dk 2009–2022*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.19, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 201-222, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

could be combined.

Thus, the overall research question is: What characterizes the changes of actor types in the hyperlink network of videnskab.dk?

1. Context of the study

To better understand the following hyperlink analyses some context is needed, including information about why this study was made, where the data came from, what characterizes network analysis of hyperlinks and the archived web, and finally how the available data were prepared for analysis.

2.1 The starting point: Evaluating videnskab.dk

Videnskab.dk was founded in 2008 to promote and communicate research-related knowledge to the wider public, and in 2023 videnskab.dk had 18 employees, 12 full-time and 6 part-time (Degn et al. 2023, 22). In 2022, after 15 years of the website's existence, the Danish Agency for Higher Education and Science, which provided funding for the website, sought its evaluation. The Centre for Cultural Evaluation at Aarhus University was commissioned to perform this evaluation, and the authors of this chapter were invited to contribute analyses of the hyperlink structure around videnskab.dk.

To cover as many facets of the evaluation of the website's activities as possible, a very broad analytical design was chosen, including (1) an analysis of the website and its content, with a focus on genre, design, functionality, and journalistic communication, (2) an analysis of social media presence and communication strategies (Twitter, Facebook, Instagram, and LinkedIn), (3) a field study and interviews with management and staff, (4) interviews with researchers who have contributed to articles on videnskab.dk, (5) interviews with science journalists/editors from other media who published articles based on content from videnskab.dk, (6) questionnaires and interviews with teachers and pupils (elementary and high school), and the two elements that constitute the basis of this chapter, (7) analyses of hyperlinks extracted from Netarkivet, from the period 2009–2022, and (8) a network analysis of outgoing hyperlinks from international websites of similar type, that is journalistic websites that disseminate research.

The study was conducted in 2022 by researchers with different backgrounds to cover the various approaches, and the final evaluation report was published in March 2023 (Degn et al. 2023). In the following, focus is only on the hyperlink analyses (points (7) and (8) above), and the results that did not find room in the final report (Degn et al. 2023, 32–36) are unfolded in more detail. The report was written in Danish, but a brief

summary in English can be found on page 4 in the report.

2.2 Getting the data: The national Danish web archive Netarkivet

Since 2005, the Danish web has been collected by the national Danish web archive Netarkivet at the Royal Danish Library (see http://netarkivet.dk). Netarkivet collects the entire Danish web domain .dk four times each year, along with a limited amount of Danish web material on other web domains. In recent years, Netarkivet has enabled researchers to extract and obtain content for research purposes. Based on this service, data containing all links to and from videnskab.dk for the years 2009, 2014, 2018, and 2022 was extracted. The raw data files contained between 100,000 and 200,000 links each year: 141,903 in 2009; 233,886 in 2014; 127,234 in 2018; and 113,706 in 2022.

A few limitations that may influence the results have to be addressed. Firstly, in contrast to Netarkivet's collection of the Danish web that is almost complete, social media platforms such as Facebook, Twitter, and YouTube have not been collected in a systematic and exhaustive manner. This implies that links from the web to social media are present, whereas the opposite may not be true. Secondly, the following hyperlink analyses do not place videnskab.dk in the complete link graph of all links on the Danish web, which amounts to 10–12 billion links. Rather, videnskab.dk is positioned within its immediate context, defined as links one iteration away from the website. This includes links from videnskab.dk to other websites, links from other websites to videnskab.dk, and links in and out of all these websites. While this approach makes the analysis more focused, it comes at the expense of completeness (a few examples of studies of the Danish web exist, e.g. Brügger et al. 2017; Brügger et al. 2020).

2.3 Hyperlink (network) analyses and the archived web

The methodological history of network analysis dates back to the 1930s (Moreno 1934) and has been used to study diverse topics (refer to Wasserman and Faust 1994, 5–6, for an extensive list). In the mid-1990s, the advent of the web as a media platform offered new opportunities to study networks, because the web is characterized by concrete connections manifested as hyperlinks. This led to the inception of hyperlink network analysis around 1997, one of the first articles in this new sub-field being Jackson (1997). In the following years, hyperlink network analysis gained prominence in internet studies (see early overviews in Foot et al. 2003, 4–8; Park and Thelwall 2003) and within the software industry, with Google's PageRank playing a pivotal role (Brin and Page 1998). By the early 2010s, the widespread availability of web archives gave rise to a new branch of

hyperlink network analysis: hyperlink network analysis of the archived web. Weltevrede and Helmond's historical study of the Dutch blogosphere stands as one of the first examples, based on the holdings of the Internet Archive (Weltevrede and Helmond 2012). Shortly thereafter, discussions on how the specificities of the archived web as a source affect network analysis are added to this literature (Brügger 2013). However, as of today, the number of network analyses studying the archived web remains limited, predominantly adopting a historical perspective (e.g. Meyer et al. 2017; Weber 2017; Cowls and Bright 2017; Ackland and Evans 2017; Webster 2017; Brügger 2021, 2022; Fage-Butler et al. 2022), whereas the archived web is not studied as a source that can shed new light on contemporary hyperlink networks (for a brief introduction to network analysis, hyperlink network analysis, and the archived web, see Stevenson and Ben-David 2018). This chapter aims to bridge this gap by studying both the contemporary web and the past web with the archived web as a source.

The network analysis of videnskab.dk's hyperlinks is based on standard network analytical concepts (Wasserman and Faust 1994), where the value of an entity is a function of its relations to other entities in the network. The nodes of the network are entire websites (and not individual web pages), while a hyperlink constitutes the edge, and the number of concrete hyperlinks between two nodes determines the weight of the edge. In addition, as outlined below, websites are categorized into actor types, which then serve as an attribute of the node. Since hyperlinks point from one website to another website, the network is directed. The analysis focuses on three ways of measuring centrality: in-degree centrality (the number of edges pointing to a given website), out-degree centrality (the number of edges pointing from a given website to other websites), and betweenness centrality (how often a node is present on the shortest path between two nodes, in other words how often it functions as a bridge). It is important to note that a website can control its out-degree, but not its in-degree or its betweenness centrality.

However, when using the archived web as the source for hyperlink (network) analyses, two limitations related to the nature of the archived web must be considered, in contrast to conducting hyperlink network analyses of the online web (Brügger 2013—for a general introduction to the archived web as a source, see Brügger 2019). Firstly, due to the method of web archiving and the organization of the collection, the same web page is likely to exist in the archive more than once, even within a limited period of time. In some cases, it may be an identical copy, while in others, it may be a version, that is two web pages with the same URL but different content from different points in time. Therefore, versions of the same web page from two different points in time are excluded if they link to precisely the

same websites, even though their content may differ, thus retaining only one version of each web page in the dataset. This curation approach aims to reflect what the web actually looked like in the past (and not what it looked like in the web archive) while reducing the number of links considerably.

Secondly, since Netarkivet collects the entire .dk web domain four times a year, material meeting the criteria for this analysis (in- and out-going links from videnskab.dk + one iteration) can be archived at different times during a calendar year. Consequently, the analysis of each of the four years has a temporal extension of one year, wherein links that were not simultaneously present online are analyzed as if they were. In other words, each annual link graph becomes temporally inconsistent. This inconsistency is accepted because it provides a more comprehensive link graph compared to a link graph based on only one week or one month per year (for a discussion on the balancing of temporal inconsistency and completeness, see Brügger 2019, 22–25).

2.4 Preparing the data for analysis

The csv files extracted from Netarkivet were processed in Excel to prepare them for analysis using the network analysis software Gephi (gephi.org) and to perform certain descriptive statistics. To simplify the network, a cut-off level of 100 was applied to the edge weight, that is: edges connecting two websites with fewer than 100 links were excluded from the dataset. Upon initial test analyses, it became clear that further data cleaning was necessary, and additional information needed to be incorporated.

Initially, the dataset contained nodes with very high weights, raising questions about whether the numerous links were actual links to videnskab.dk (and other websites) or if they were the result of recurring website construction elements, (menu, navigation, footer, and the like). These components could lead to a high weight even if all links were, strictly speaking, identical. To remove these types of links, the edge table was manually checked for suspicious edges, including the following:

Table 1: Suspicious rows in the edge table.

source	target	weight
ronniandersson.dk	facebook.com	8305
ronniandersson.dk	google.com	8305
ronniandersson.dk	twitter.com	8305
ronniandersson.dk	upworth.dk	8305

Table 1 illustrates instances where five different edges from the same node display an identical number of links. This clearly indicates that the links were found in a footer or a similar element present on all pages of the

websites. Manual checks were conducted on such cases by examining the website in Netarkivet's browser view. When evaluating the link types, inspiration was drawn from the categorization proposed by Ryfe et al. (2016), which distinguishes four types of links:

> [...] navigation, commercial, social, and citation. Navigation helps users find content. Commercial involves linking practices for earning money from others. Social includes sharing content via social media feeds and/or offering users opportunities to share content. Citation directs users to information in an effort to establish the credibility of news reports. (Ryfe et al. 2016, 42)

Since our analysis primarily focuses on 'citation' links related to videnskab.dk, web features such as share buttons to social media and similar elements are not considered links and are thus excluded from the dataset.

Secondly, the dataset had to be enriched with information on the actor types within videnskab.dk's link graph, since this could not be immediately deduced from the web addresses. Therefore, the top 50 websites (measured by weight) in the edge tables for in- and out-links were manually checked either in Netarkivet or on the online web, to determine their respective categories. The list of categories was developed through an iterative process, establishing a new category if a website on the list did not fit one of the already identified categories. To reduce complexity, websites were assigned to a single category only. This grounded approach led to the following list of categories:

- Science website (e.g. sciencenorway.no)
- Mainstream media (such as national daily newspapers and weekly magazines)
- Niche media (niche media *with* a journalistic/editorial approach)
- Alternative media (niche media *without* a journalistic/editorial approach)
- Research institution
- Library
- Education
- Association
- Official (e.g. Ministries, Health Care system)
- Publisher (academic publisher, either publishing house or journal)
- Academic portal (e.g. researchgate.net)
- Blog
- Encyclopedia (e.g. wikipedia.org)
- Discussion forum
- Other

3. Results: videnskab.dk in the hyperlink network 2022, and actor types in the past

The primary focus of the analysis of videnskab.dk centers on the website's link graph as it appeared when the evaluation report was drafted, i.e. in the year 2022. The analysis comprises (1) descriptive statistics based on the number of links to and from videnskab.dk, with a particular focus on actor types, and (2) a network analysis that includes videnskab.dk's position in the network. The difference between these two approaches lies in the fact that the statistical analysis gives an overview of actor types and individual websites but does not provide insights into the centrality of each website in the network—whether the websites connected to videnskab.dk are themselves central or not. This dimension is elucidated through the network analysis. Looking back from 2022, the study examines the website's hyperlinks in 2009, 2014, and 2018, followed by a brief outline of some of the major developments. This historical dimension specifically focuses on the actor types of the hyperlink and not on the network as such, aligning with the analysis presented in the published evaluation report. Finally, a brief analysis of videnskab.dk in the transnational web landscape is included. An Appendix with the figures that are not included in the following can be found in the Zenodo community 'Book chapter Web archives and hyperlink analyses' at https://zenodo.org/communities/resaw2023_chapter.

3.1 The hyperlink network 2022

The first analytical step is to examine the top50 actor types linked either through out-links or in-links from videnskab.dk. Most links from videnskab.dk point to either academic publishers or research institutions, comprising nearly half of the top-50. The remainder is primarily linked to mainstream media and other reputable scientific websites (Figure 1 in appendix). Libraries and research institutions dominate the in-linkers, constituting just below half of top-50. Blogs, discussion fora, and alternative media also contribute to in-links as much as mainstream media (Figure 2 in appendix).

Comparing linked-to and in-linking actor types, research institutions are prevalent in both cases, while mainstream media also hold significance, albeit to a lesser extent. Unsurprisingly, videnskab.dk does not link to more (scientifically) dubious websites from alternative media, but these websites link back to videnskab.dk.

The second analytical step examines the distribution of individual actors within the top50 concerning the number of concrete links (Figure 3 and 4 in appendix). Regardless of whether the focus is on out- or in-links to videnskab.dk, the structure remains the same: very few actors possess a very

high number of links, followed by a mid-group of approximately 10 actors with fewer links, and then a long tail of actors with very few links.

A closer look at in-linking actors reveals a number of characteristics: (1) the highest in-linking website is a Nordic science-related website (sciencenordic.com), similar to videnskab.dk and with which videnskab.dk collaborates; (2) many research institutions among the top in-linkers have a substantial number of concrete links (high edge weight); (3) among the top10 in-linkers, niche media like ing.dk (a journalistic website on technology and science) and two alternative media—a news aggregator (godt-nyt.dk) and a website about drug use (psychedelia.dk)—are noteworthy; (4) the rest of top50 includes a mix of mainstream media, two alternative media (uriasposten.net—an anti-elite website—and nomedica.dk—an anti-medical science website),personal blogs (e.g. lykkeliginoerdland.dk, a blog aimed at informing women about 'hard science'), and discussion fora such as ingeniordebat.dk (an engineering forum with 647 members) and musclezone.dk (a bodybuilding forum); (5) finally, it is worth noting that public libraries link to videnskab.dk, but generally with very few concrete links.

Moving on to the actual network analysis, our focus is on how videnskab.dk is positioned within its immediate hyperlink network and identifying the characteristics of other nodes in this network. A few network statistics: edges are only included if they have a weight above 100 (as previously mentioned); the network comprises 1,147 nodes and 1,679 edges; the network diameter is 4, indicating that 4 steps are needed to travel between the two farthest nodes; the average degree of nodes is 1.164, indicating that each node is connected to a little more than one other node; the average weighted degree is 371.983, representing the average number of edges weighted with the weight of each edge; and the graph density is 0.001, measured on a scale between 1 and 0, where 1 implies that all potential edges are realized, and 0 signifies none.

Figure 1 illustrates the network with a focus on out-degree that is the number of outgoing hyperlinks. Unsurprisingly, many of the actor types and specific actors already identified in the statistics of top-50 most linked to from videnskab.dk are visible, but new actors also emerge as central out-linking nodes, notably psyx.blogspot.com, a blog for a psychotherapist and sexologist. Also, it is worth noting that two major out-linkers are alternative media, 23.dk (likely due to its Wikipedia structure) and uriasposten.net, known for being a link central. Blogs, given their inherent nature, are also central out-linkers (ing.dk, version2.dk, minkusinemaria.dk, lykkeliginoerdland.dk). Finally, it is worth noting that mainstream media and libraries do not play a substantial role in the out-degree network contrasting with their prominence when focusing solely on links in

and out from videnskab.dk.

Figure 1: The near out-degree network of videnskab.dk. Nodes are sized according to their out-degree, edges according to weight (graph spatializsed with Fruchterman Reingold (area 5.000, gravity 1.,0), zoomed for better readability. For the full network visualizsation, see Figure 5 in appendix.

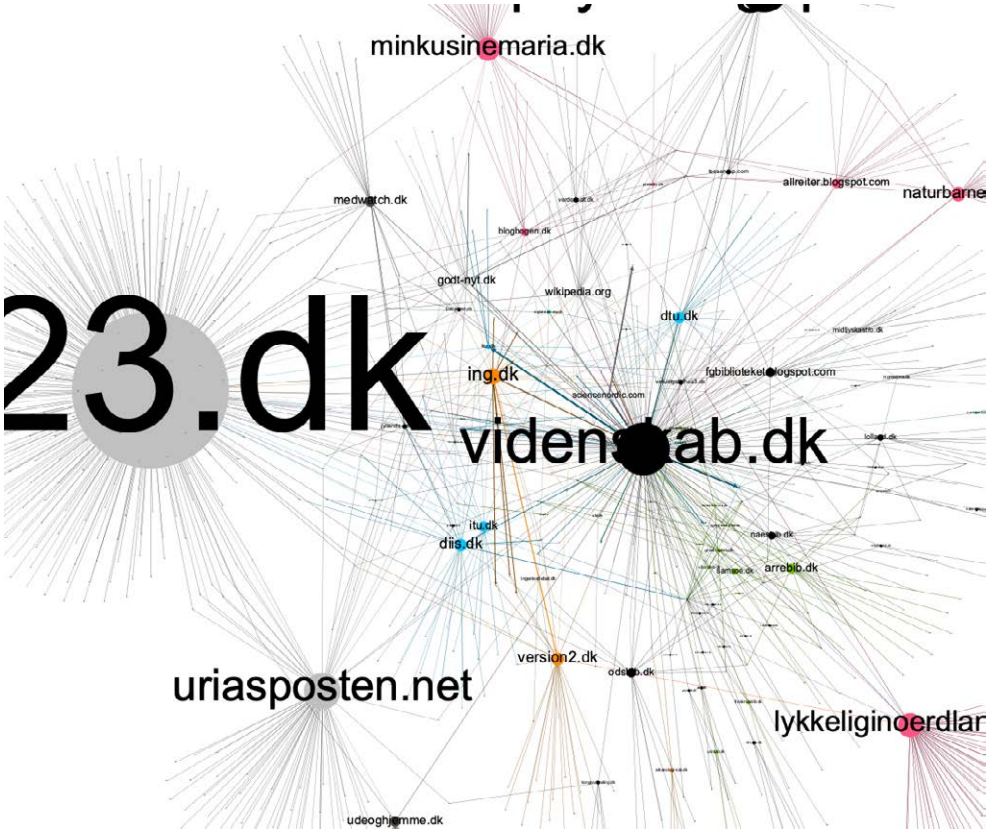


Figure 2: The near in-degree network of videnskab.dk. Nodes are sized according to their in-degree, edges according to weight (graph spatializsed with Fruchterman Reingold (area 5.000, gravity 1.,0), zoomed for better readability. For the full network visualizsation, see Figure 6 in appendix.

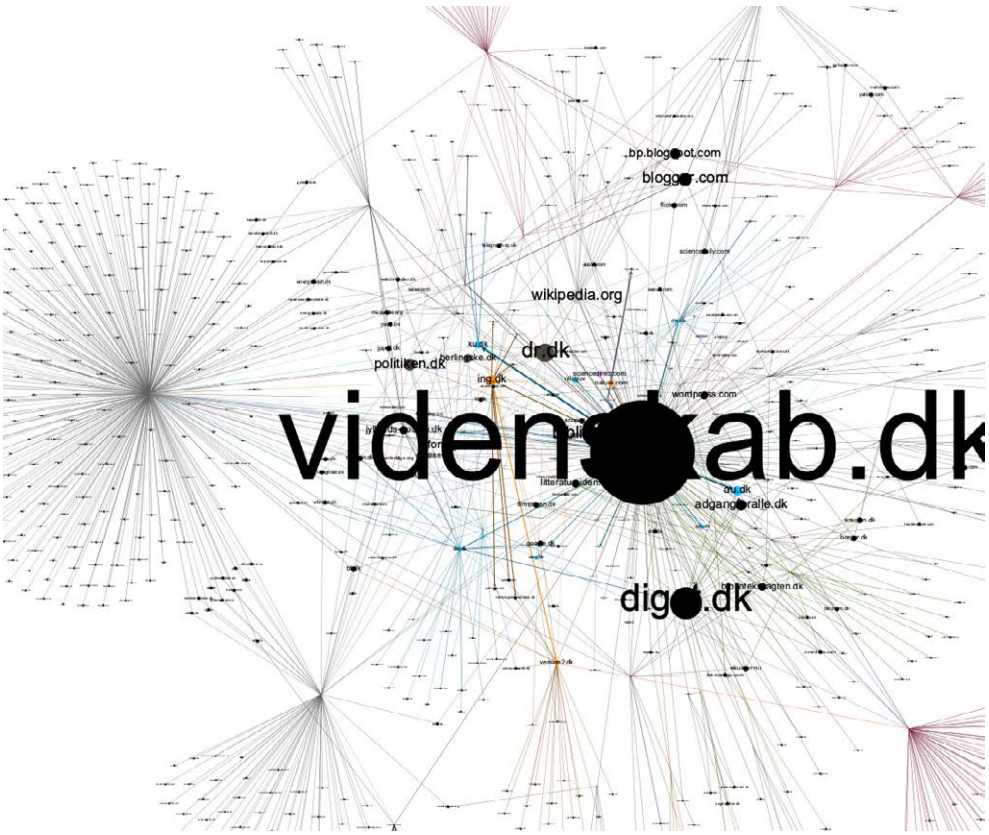
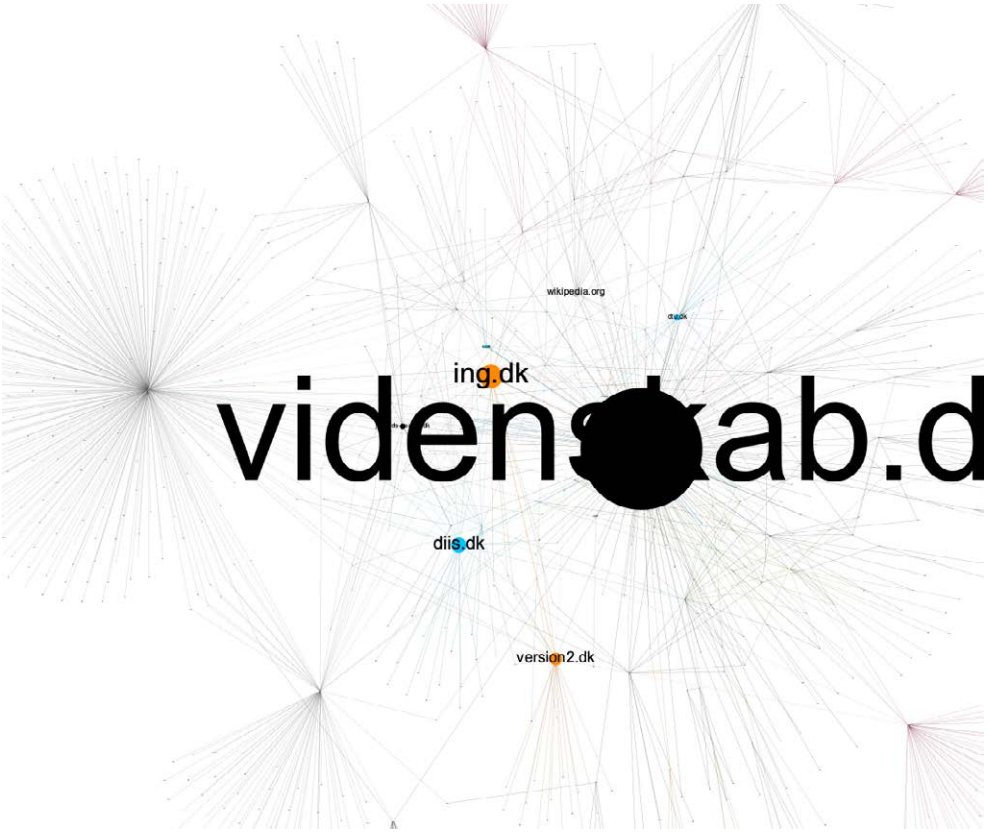


Figure 2 visualizes the network with a focus on in-degree that is the number of ingoing hyperlinks. When analyzing the in-degree network, it is important to bear in mind that any given website cannot, by itself, produce a high in-degree, since it does not decide who links to it. The in-degree thus provides insights into how popular a website is if a link is considered a popularity measure. Actors from the top-50 list of links related to videnskab.dk are also present in the in-degree network, with mainstream media and research institutions being central. New websites appear as central in-linked nodes, such as digst.dk (the Agency for Digital Government), labradorcms.com (a website for digital publishing), and blogger.com. Notably absent in the in-degree network are libraries. Finally, a stark difference observed when comparing the in-degree network to the out-degree network in Figure 1 is that nodes central in the out-degree network are not central in the in-degree network, except for ing.dk. This

suggests that the heavily out-linking websites may not be popular for incoming links, making them relevant only in the out-degree network if directly accessed (by typing their web address in the browser) because they are not likely to be visited by users who arrive at them through an in-link.

Figure 3: The near betweenness-centrality network of videnskab.dk. Nodes are sized according to their betweenness centrality (the bigger the node, the more it functions as a bridge in the network), edges according to weight (graph spatializsed with Fruchterman Reingold (area 5.000, gravity 1.,0), zoomed for better readability. For the full network visualizsation see Figure 7 in appendix.



The third segment of the network analysis focuses on betweenness centrality, identifying websites that play a central role as bridges between other nodes (Figure 3). Unsurprisingly, videnskab.dk emerges as the most central bridge, and apart from a handful of nodes (ing.dk, version2.dk, the research institutions diis.dk, dtu.dk, and ku.dk, and one mainstream media), there are minimal important bridges. This means that actor types such as publishers, libraries, mainstream media, alternative media, and blogs, do not serve as bridges enabling connections between the different actor types.

When comparing the out-degree, in-degree, and betweenness networks, it becomes evident that different actors take central roles depending on the network focus. Only a few actors are central in more than one network, notably ing.dk and version2.dk, along with a few research institutions, most notably ku.dk (the University of Copenhagen). Surprisingly, libraries are not central in any of the networks.

Determining the role of videnskab.dk in the link network in relation to other actors highlights the significance of those in top 50 in-linkers to videnskab.dk, that is the actors who deliberately point to videnskab.dk and potentially send their own users in that direction. However, their value for videnskab.dk grows with their centrality in the network. In other words: it is interesting when an actor links to videnskab.dk, but it is even more interesting if the linking node holds a central position in the network. To investigate this, one needs to consider actors in the top 50 of in-linkers with these nodes' centrality within each of the three network measures: out-degree, in-degree, and betweenness centrality. If a website not only links significantly to videnskab.dk (among the top 50 in-linkers) but also ranks high in one or more of these measures, it signifies that the website is particularly important for videnskab.dk. An analysis along these lines reveals four websites as the most crucial: ing.dk (the journalistic website on technology and science), diis.dk, dtu.dk (websites from two research institutions), and wikipedia.org. Others are also important, but to a lesser extent: e23.dk, lykkeliginoerdland.dk, minkusinemaria.dk, naturbarnet.dk (a blog about healthy living), version2.dk, udeoghjemme.dk (a weekly magazine), information.dk (a mainstream media), ku.dk, au.dk (research institutions). Notably, some strong in-linkers to videnskab.dk (and of whom there were many) do not play a significant role in the broader network: science websites, libraries, and discussion fora. Moreover, the absence of expected actor types that would either link to videnskab.dk or be part of the network is noteworthy, including local newspapers, NGOs, companies, and primary and high schools, which are key target groups for videnskab.dk.

In conclusion, the linking patterns in 2022 suggest that videnskab.dk supports its own ethos as a serious scientific publisher by linking to academic publishers and research institutions. However, the actors linking to videnskab.dk form a much more heterogeneous group: research institutions are still important players, but they are supplemented to various degrees by niche and mainstream media with an interest in videnskab.dk's topics, along with alternative media, blogs, and discussion fora. Thus, videnskab.dk is embedded in networks where the dissemination of scientific knowledge is key, but it also interacts with actors promoting views aligned with its ethos.

3.2 Hyperlinked actor types 2009, 2014, 2018

This section delves into videnskab.dk's out-links and in-links in 2009, 2014, and 2018, providing an overview of actors appearing across these years, followed by a historical analysis comparing these results with those from 2022 to identify and discuss historical developments.

3.2.1 2009: Unreciprocated attention

Looking at videnskab.dk's out-links in the year 2009 (Figure 8 in appendix), research institutions emerge as the predominant actor type, followed by mainstream media. Positioned in the middle are actors such as scientific websites and alternative media, while education, officials, blogs, and discussion forums occupy the lower positions. This suggests a consistent presence of both 'scientific' and 'non-scientific' sources throughout (with reference to differences regarding scientific engagement or association). Although the distribution of concrete links from these actors may exhibit some unevenness (Figure 10 in appendix), this use of both scientific and non-scientific sources appears prevalent here as well, since both feature prominently at either end of the distribution spectrum, as exemplified by the presence of both research institutions and alternative media among those with the highest link count.

Turning to in-links, blogs take the lead as the most prevalent type of actor, appearing more than twice as much as any of the other actor types (Figure 9 in appendix). Research institutions, which dominate out-links, shift to the bottom of the in-links, establishing a contrast between videnskab.dk's out-links and in-links. Furthermore, compared to the above, the consistent presence of scientifically and not scientifically engaged actors is small in both the distribution of actors and the distribution of concrete links (Figure 11 in appendix). Both are largely constituted by actors who may be considered less scientifically engaged, such as discussion forums, mainstream media, and, notably, blogs.

3.2.2 2014: Closing in

In the year 2014, research institutions continue to dominate as the most prevalent actor in videnskab.dk's out-links (Figure 12 in appendix). Mainstream media, however, has been surpassed by publishers, who were not present in the out-links from 2009. The appearance of yet another new actor, academic portals, accompanies the disappearance of discussion forums, blogs, and alternative media. Additionally, the size of education has doubled. The presence of actors more scientifically engaged now outweighs those who may be considered less so. This shift is also apparent in the

distribution of concrete links (Figure 14 in appendix), where research institutions and other scientific actors largely constitute the head of the distribution.

Examining the in-links (Figure 13 in appendix), blogs remain the most prevalent actor, but their size is no longer more than twice that of every other present actor. Mainstream media, niche media, and research institutions almost reach the same level of prevalence as blogs, somewhat evening out the top. Furthermore, several new types of actors appear, namely scientific websites, encyclopedias, libraries, and academic portals. Although these are positioned throughout the middle and the bottom, the contrast between videnskab.dk's out-links and in-links appears less pronounced compared to the figures from 2009. The presence of scientific actors has increased and expanded, as emphasized by the distribution of concrete links (Figure 15 in appendix), with research institutions holding some of the highest link counts.

3.2.3 2018: One step forward, two steps back

Largely the same actors present in 2014 are also present in the out-links from 2018, with research institutions now being preceded by publishers. Scientific websites have almost doubled in size, ranking as the fourth most prevalent actor (Figure 16 in appendix). While still partially uneven, the distribution of concrete links appears somewhat more flattened with an elongated tail (Figure 18 in appendix), possibly indicating a sharpened preference for certain types of sources, such as education and research institutions, which have the highest link counts.

Videnskab.dk's in-links in 2018 somewhat mirror the out-links from 2009, taking two steps back, as blogs have once again grown to twice the size of any other present actor, maintaining their position as the most prevalent actor (Figure 17 in appendix). While the appearance of research institutions remains largely the same as in 2014, this actor is now preceded by alternative media. Furthermore, there is an increase in discussion forums, while actors such as encyclopedias and libraries have disappeared. In other words, the contrast between videnskab.dk's out-links and in-links appears more significant when compared to 2014, with both the prevalence and presence of more scientifically engaged actors diminished. While research institutions still hold some of the highest link counts in the distribution of concrete links (Figure 19 in appendix), they are now accompanied by actors such as blogs and alternative media, emphasizing the aforementioned changes.

3.3 The development of actor types related to videnskab.dk

When comparing videnskab.dk's out-links and in-links over the years, a contrast emerges between who videnskab.dk links to, and who links to videnskab.dk. Videnskab.dk's out-links increasingly target more scientific actors, such as research institutions, publishers, education, and similar entities. On the other hand, actors who are not scientifically engaged, namely blogs, alternative media, and discussion forums, continue to appear as in-linkers to videnskab.dk—actors to whom videnskab.dk, apart from the year 2009, does not link (Figure 8 in appendix; Figure 12 in appendix; Figure 16 in appendix; Figure 1 in appendix).

However, the contrast appears to diminish, as blogs, which were the most prevalent actor in videnskab.dk's in-links throughout 2009, 2014, and 2018 (Figure 9 in appendix; Figure 13 in appendix; Figure 17 in appendix), are finally preceded by libraries and research institutions in 2022 (Figure 2 in appendix). Furthermore, there is a continuous increase in link counts from scientific actors. Building on Terveen and Hills' understanding of a website's hyperlink connectivity as a reflection of the website's credibility and perceived quality, described as a positive correlation (Park and Thelwall 2003, 13), these results may suggest a development in which videnskab.dk is increasingly recognized as a credible source of science—at least in the eyes of other scientific actors. At the same time, this may also point to a trend in which videnskab.dk is less cited by the general population, whether through discussion forums or blogs, which videnskab.dk itself has described as an important target group (Degn et al. 2023, 5–6).

The prevalence of certain actor types may also reflect societal and technological changes. Thus, the decrease in citations of videnskab.dk cited by the general population could be attributed to social media's partial takeover of blog-related activities, which, as stated earlier, has been excluded from the datasets. Likewise, the transformation of libraries from being minimally present to becoming the most prevalent actor in videnskab.dk's in-links in 2022 (Figure 2 in appendix) may mirror the institutional development libraries have undergone since digitalization, characterized by a growing demand for users' electronic access to scientific journals and papers (Povlsen 2016).

The continuing decrease of non-scientific actors in videnskab.dk's out-links over the years could also be time-related, reflecting videnskab.dk's gradual foothold since its establishment in 2008. This development may have reduced the need to produce content based on what is already popular among the general population. At the same time, this might also explain why videnskab.dk's in-links appear to be converging with its out-links

(actor-wise) over the years, shaping the identity of the videnskab.dk of today.

As demonstrated above, the use of archived web data has offered a deeper understanding of the 'whats' and 'whys' surrounding present-day videnskab.dk by providing insights into older versions hereof. As such, the analysis of 'what has been' can serve as the foundation for the analysis of 'what is', proving the archived web to be an invaluable source not only for historical but also contemporary analyses.

3.4 A transnational perspective

Having examined the actors present in videnskab.dk's out-links and in-links across the years 2009, 2014, 2018, and 2022, the decision was made to examine videnskab.dk from an international perspective. This involved conducting a network analysis of videnskab.dk alongside similar international scientific websites to identify potential differences or similarities in connections and use of sources.

The analyzed network is based on hyperlinked citations, referring to out-links found in the content of each website, which have been harvested—thus not collected from a web archive—in 2022 from videnskab.dk, and a corresponding English (newscientist.com), American (scientificamerican.com), French (futura-sciences.com), German (scinexx.dk), Norwegian (forskning.no), and Swedish (fof.se) website. These sites were chosen for their journalistic profile and orientation towards conveying scientific content, which resemble that of videnskab.dk.

Merging the edges allowed us to examine the in-degree of each node as a reflection of the number of unique connections, facilitating the identification of shared sources among the individual websites. The analysis showed that approximately one-tenth of the nodes in the network could be identified as shared sources, with each node having no more than 1.195 connections (degree value), resulting in a sparse network graph with a low density score of 0.002. Additionally, a positive modularity score suggested a tighter connection within the clusters (each representing the individual websites chosen and their dedicated out-links). Thus, one might describe the network as somewhat polarized (Smith et al. 2016), also visible in the overview of the graph (Figure 4 in appendix).

Most of the shared sources can be identified as scientific actors such as academic portals and publishers (Figure 21 in appendix). Notably, the ones with the highest in-degree tend to be engaged in various branches of the natural sciences, such as nature.com, ncbi.nlm.nih.gov, pubmed.ncbi.nih.gov, and nasa.gov, to name a few. The use of these specific sources may suggest a shared appreciation, potentially owing to

their accessibility for general research or their emphasis on content related to the natural sciences. This might indicate a perceived academic or international value, which could be of interest considering their shared Top-Level Domains (TLDs). Zooming in, it becomes apparent that videnskab.dk has a higher number of shared sources with the other Nordic websites, specifically forskning.no and fof.se, along with the American website, scientificamerican.com, than with the remaining websites. This may imply more similarities in content among these platforms.

Taking a broader view of sources or connections in general, the Nordic websites share structural similarities by having a large number of edges compared to the remaining websites, thus acting as larger hubs. Furthermore, upon merging the edges, videnskab.dk appears to have more distinct or individual connections than any other website in the network, indicating a wider use of different sources (Figure 5 in appendix).

Having examined what can be characterized as formal connections in the network, we decided to delve deeper into the Country Code Top-Level Domains (ccTLDs) of each source, as a means of exploring what Park and Thelwall refer to as the "trans-national knowledge flow" (2003, 12). This also allowed for an exploration of how we might understand the said notion of 'different' sources. Given our interest in the connectivity between the chosen websites, we concentrated on the ccTLDs associated with their respective nationalities, resulting in the following list of identified ccTLDs:

- .se (Swedish)
- .no (Norwegian)
- .us / .gov (American)
- .uk (English)
- .fr (French)
- .dk (Danish)
- .de (German)
- "other" (unidentified ccTLDs)

Results showed that videnskab.dk exhibited the largest variety of ccTLDs (Figure 6 in appendix). This was somewhat mirrored by the Norwegian website, where sources also demonstrated a wide variety of ccTLDs compared to the other websites. However, most of videnskab.dk's sources shared the website's own national origin (Danish), a pattern also observed in the other Nordic websites. For instance, forskning.no and fof.se mainly connect to Norwegian and Swedish sources, respectively, suggesting a shared 'favoring' of national sources among the Nordic websites.

While the use of national sources was also evident among the remaining

websites, sources representing neither the website's own nationality nor that of the others in the network, but instead those categorized as "other", were more prevalent. This might indicate a more widespread use of sources in terms of nationality, as national ccTLDs seemed to be less favored in these instances. On the other hand, since the specific ccDLTs in the "other" category have not been identified, it remains unclear whether the "other" category constitutes a broad range of ccTLDs (apart from those identified), or the same unidentified ccTLD—potentially representing a narrower, rather than widespread, use of (international) sources.

All of the aforementioned points indicate that videnskab.dk shares more similarities with the other Nordic websites, both in terms of structure and sources, effectively distinguishing itself from the American and the European websites. With only a few sources serving as common denominators, the graph presents a rather polarized network, portraying a sense of disconnectivity in an otherwise globalized world. Nevertheless, given the temporal limitations of the method used for data collection, along with the continuous evolution of online web content (Brügger 2019), the results might merely be a reflection of the temporary, emphasizing the need for further investigation.

4 Discussion

In this section, we will briefly discuss some potential implications of the results and the methods.

Analyzing hyperlinks, including actor types and their positions in the hyperlink network, provides an opportunity to uncover the structure that underpins one of the main communicative infrastructures in contemporary society—the web. Hyperlinks can be likened to the 'roads' that allow a web user to 'travel' from one communicative entity to another. However, this map of roads is not readily visible when navigating hyperlinks from one website to another. Users are embedded in the web landscape and cannot see its structure from above. And not only are all the potential roads not visible, so is the role of the interlinked entities, the websites, because their status, that is their centrality, cannot be fully understood as such while moving around on the web. Only a network analysis can provide an overview map of the interconnected websites and the distinct role each one plays on the entire map. However, mapping the roads does not reveal information about the content of websites, their creators, or the frequency of visits. To complete the picture, one has to include analyses of website content and user statistics. Nevertheless, providing a map of hyperlinks is a valuable first step in comprehending the nature of our communicative infrastructure.

The international hyperlink network analysis of videnskab.dk and the contemporary online web indicates numerous narratives when extending the scope to include websites and hyperlinks beyond the nation of origin for each website. However, given the often closed nature of national web archives, expanding this analysis to a comprehensive transnational analysis would require access to all relevant national web archives, including easy access to the hyperlinks. Unfortunately, as for now, such an analysis is not feasible due to the lack of transnational research infrastructure between national web archives.

Web archives such as Netarkivet can be valuable tools for mapping and identifying website content through their recordings of various hyperlink data, such as link paths or link positions. However, depending on the design and software used, certain tools may struggle to perform an accurate reading of the code embedded within the structure of specific hyperlinks. This can lead to inaccurate data and introduce uncertainty regarding the validity of any findings. Ensuring a tool's alignment with a chosen collection of hyperlinks can be challenging with large datasets constituted by hyperlinks with different structures. While this calls for practical solutions, questioning how a tool's interpretation of data differs from our own may be a valuable step in determining the usability and validity hereof. Thus, apart from demonstrating web archives to be an indispensable source in analyses of the web, historical as contemporary, we also encourage a critical reflection on the data and findings generated through the use of these.

5 Conclusion and next steps

As demonstrated in this chapter, the archived web is not only a valuable source for analyzing the past, but also enhancing our understanding of the present communicative infrastructure of the web and its hyperlinks, particularly in providing information about in-going hyperlinks to websites, which cannot be collected from the live web. It thus makes a plea that web archives are not only significant for historical studies, but also for contemporary investigations.

While the present study is limited to focusing on one website and one iteration of hyperlinks from this website, it serves as a model to inspire broader studies, such as those exploring entire national web domains as outlined by Brügger et al. (2020). It can—and should—encourage more focused analyses of the most central nodes.

Finally, to conduct an exhaustive analysis of the hyperlink network in which any given website is embedded, it is pivotal to create research infrastructures that extend beyond the borders of national web archives.

Acknowledgements

We would like to express our gratitude to the contributors to the evaluation report of videnskab.dk: Hans-Peter Degn, Christiane Særkjær, Line Hassall Thomsen, and Maja Sonne Damkjær. Special thanks to the IT developer at the Department of Media and Journalism Studies, Ulrich Karstoft Have, and Netarkivet for facilitating the extraction of the data.

References

Ackland, Robert, and Ann Evans. 2017. "Using the web to examine the evolution of the abortion debate in Australia, 2005–2015." In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 159–189. London: UCL Press.

Beaudouin, Valérie, Zeynep Pehlivan, Peter Stirling. 2018. "Exploring the memory of the First World War using web archives: Web graphs seen from different angles." In *The SAGE handbook of web history*, edited by Niels Brügger and Ian Milligan, 441–463. London: SAGE.

Brin, Sergey, and Lawrence Page. 1998. "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems* 30, no. 1: 107–117, https://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.pdf.

Brügger, Niels. 2013. "Historical Network Analysis of the Web." *Social Science Computer Review* 31, no. 3: 306–321. https://doi.org/10.1177/089443931245426

Brügger, Niels. 2019. "Understanding the archived web as a historical source." In *The SAGE handbook of web history*, edited by Niels Brügger and Ian Milligan, 16–29. London: SAGE.

Brügger, Niels. 2021. "Digital humanities and web archives: Possible new paths for combining datasets." *International Journal of Digital Humanities* 2, no. 1–3: 145–168.

Brügger, Niels. 2022. "Tracing a historical development of conspiracy theory networks on the web: The hyperlink network of vaccine hesitancy on the Danish web 2006–2015." *Convergence* 28, no. 4: 962–982. https://doi.org/10.1177/13548565221104989

Brügger, Niels, Ditte Laursen, and Janne Nielsen. 2017. "Exploring the domain names of the Danish web." In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 62–80. London: UCL Press.

Brügger, Niels, Janne Nielsen, Ditte Laursen. 2020. "Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web." *First Monday* 25, no. 3. https://firstmonday.org/ojs/index.php/fm/article/view/10384.

Cowls, Josh, and Jonathan Bright. 2017. "International hyperlinks in online news media." In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 101-116. London: UCL Press.

Degn, Hans-Peter, Christiane Særkjær, Line Hassall Thomsen, Maja Sonne Damkjær, and

Niels Brügger. 2023. “Evaluering af Videnskab.dk”. Aarhus: Center for Kulturevaluering, Aarhus Universitet. https://ufm.dk/publikationer/2023/evaluering-af-videnskab.dk.

Fage-Butler, Antoinette, Loni Ledderer, and Niels Brügger. 2022. “Proposing methods to explore the evolution of the term ‘mHealth’on the Danish Web archive.” *First Monday* 27, no. 1. https://firstmonday.org/ojs/index.php/fm/article/view/11675.

Foot, Kirsten, Steven M. Schneider, Meghan Dougherty, Michael Xenos, and Elena Larsen. 2003. “Analyzing Linking Practices: Candidate Sites in the 2002 US Electoral Web Sphere.” *Journal of Computer-Mediated Communication* 8, no. 4. https://doi.org/10.1111/j.1083-6101.2003.tb00220.x.

Jackson, Michele H. 1997. “Assessing the Structure of Communication on the World Wide Web.” *Journal of Computer-Mediated Communication* 3, no. 1. https://doi.org/10.1111/j.1083-6101.1997.tb00063.x.

Meyer, Eric T., Taha Yasseri, Scott A. Hale, Josh Cowls, Ralph Schroeder, and Helen Margetts. 2017. “Analysing the UK web domain and exploring 15 years of UK universities on the web.” In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 83–100. London: UCL Press.

Moreno, Jacob L. (1934). *Who shall survive? A New Approach to the Problem of Human Interrelations*. Washington, DC: Nervous and Mental Disease Publishing.

Park, Han Woo, and Mike Thelwall. 2003. “Hyperlink Analyses of the World Wide Web: A Review.” *Journal of Computer-Mediated Communication* 8, no. 4. https://doi.org/10.1111/j.1083-6101.2003.tb00223.x.

Povlsen, Karen Klitgaard. 2016. “BØGER! BØGER! BØGER!” In *Dansk Mediehistorie, vol 4.*, edited by Klaus Bruhn Jensen. Frederiksberg C: Samfundslitteratur.

Ryfe, David, Donica Mensing, and Richard Kelley. 2016. “What is the meaning of a news link?” *Digital Journalism* 4, no. 1: 41–54.

Smith, Marc A., Lee Raine, Ben Schneiderman, and Itai Himelboim. 2014. “Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters.” *Pew Research Center*, February 20, 2014. https://www.pewresearch.org/internet/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/.

Stevenson, Michael, and Anat Ben-David. 2018. “Network analysis for web history.” In *The SAGE handbook of web history*, edited by Niels Brügger and Ian Milligan, 125–137. London: SAGE.

Wasserman, Stanley, and Katherine Faust. 2009 [1994]. *Social network analysis: Methods and applications*. Cambridge: Cambridge UP.

Weber, Matthew S. 2017. “The tumultuous history of news on the web.” In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 83–100. London: UCL Press.

Webster, Peter. 2017. “Religious discourse in the archived web: Rowan Williams, Archbishop of Canterbury, and the sharia law controversy of 2008.” In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 190–203. London: UCL Press.

Weltevrede, Esther, and Anne Helmond. “Where Do Bloggers Blog? Platform Transitions within the Historical Dutch Blogosphere.” *First Monday*, February 2, 2012. https://doi.org/10.5210/fm.v17i2.3775.