Big Data, Big Noise:

The Challenge of Finding Issue Networks on the Web

Annie Waldherr, Daniel Maier, Peter Miltner & Enrico Günther

Published: Waldherr, A., Maier, D., Miltner, P. & Günther, E. (2017). Big data, big noise: The challenge of finding issue networks on the Web. *Social Science Computer Review, 35*, 427-443. https://doi.org/10.1177/0894439316643050

Abstract

In this paper, we focus on noise in the sense of irrelevant information in a data set as a specific methodological challenge of web research in the era of big data. We empirically evaluate several methods for filtering hyperlink networks in order to reconstruct networks that contain only web pages that deal with a particular issue. The test corpus of web pages was collected from hyperlink networks on the issue of food safety in the United States and Germany. We applied three filtering strategies and evaluated their performance to exclude irrelevant content from the networks: keyword filtering, automated document classification with a machine-learning algorithm, and extraction of core networks with network-analytical measures. Keyword filtering and automated classification of web pages were the most effective methods for reducing noise whereas extracting a core network did not yield satisfying results for this case.

Keywords

big data, noise, hyperlink network, issue network, web crawler, document classification, machine learning, data cleaning

Introduction

The focus of this paper is noise as a specific methodological challenge of web research in the era of big data. We refer to *noise* as unnecessary, irrelevant information in a data set. Today, many scholars in social sciences are interested in analyzing public online discourses on specific issues. In the digital age, online public spaces are widely acknowledged as important venues for public debate on matters of common interest. Online public spaces not only extend the traditional mass-mediated public sphere but also form alternative arenas and counter-publics (Benkler, Roberts, Faris, Solow-Niederman & Etling, 2015; Castells, 2008; Toepfl & Piwoni, 2015) and can even function as a fifth estate in society (Dutton, 2009).

To study web discourses, a growing number of researchers rely on web crawling methods to collect samples of web documents (e.g., Bennett, Foot, & Xenos, 2011; Bruns, Burgess, Highfield, Kirchhoff, & Nicolai, 2011; Wallis & Given, 2016). Regardless of the specific strategy for collecting data, scholars need to address the problem of noise, because noisy data might result in biased analyses and invalid conclusions. To gather a population of web documents that treat a particular issue of interest, irrelevant content must be filtered out.

The need for categorizing and cleaning data sets is not new to empirical, quantitative communication research. However, in the era of big data, the scale of this challenge has grown exponentially. Specifically, document corpora collected with web crawlers such as Issue Crawler[1] (Marres, 2005; Rogers, 2010) tend to be noisy as these tools include a diverse range of web sources in the process of tracking hyperlink networks. Therefore, the aim of this paper is, first, to discuss the dimensions of the noise problem in big data analyses of online content and, second, to evaluate several filtering strategies that can be applied to filter web documents gathered by a web crawler.

In the first part of the paper, we introduce the problem of noise for the analysis of online discourses theoretically and methodologically. Specifically, we discuss the problem in

the study of issue networks, which are defined as "a set of Web pages that are connected by hyperlinks and that all treat a particular issue" (Marres, 2005, p. 97). In the second part, we illustrate this noise problem with data from an ongoing research project.[2] Within the scope of this project, we collected hyperlink networks on the issue of food safety in the United States and Germany with Issue Crawler (Rogers, 2010). To identify sub-issues and frames in the online debate, we combined network data with manual content analysis. However, when we drew random samples of the web pages in the hyperlink networks and prepared samples for content analysis, we found a substantial number of irrelevant pages. To exclude these irrelevant pages from the network, we developed and empirically evaluated several filter strategies. We report and discuss the results of these inquiries.

The Problem of Noise in Big Data Analyses of Online Discourses

As defined in computer science, the term *big data* technically refers to data that are too big to be adequately stored and processed by current software and hardware (Manovich, 2012). In the context of social sciences, the term is used for amounts of data that pose enormous challenges to traditional methods of empirical research (boyd & Crawford, 2012; Mahrt & Scharkow, 2013; Savage & Burrows, 2007). Subtracting the noise from big data sets might lead to less challenging handling. In the following section, we first discuss why online discourses are noisy. Second, we focus on the specific methodological challenge of noise in issue networks collected by web crawlers.

Noise as a Fundamental Characteristic of Public Online Discourses

Public discourses are public communication processes of reasoning, including all forms of discussions, debates, and negotiations where arguments and counterarguments are exchanged (Habermas, 1995). Compared to the traditional, mass-mediated public sphere, the Internet has opened up new spaces for such interactions to take place. These spaces differ

fundamentally from their traditional counterparts, particularly concerning the openness and interconnectedness of public debates.

Compared to traditional mass media, access to web communication is a lot more open. There are only low thresholds for participation and no spatial restrictions or fixed publication schedules. Consequently, discourses on the web are diverse, dispersed, and unrestricted. They are diverse because they give voice to a variety of actors, which are welcomed for providing high potential for participation and criticized for creating a cacophony of voices (Dahlgren, 2005; Friedland, Hove, & Rojas, 2006; Sunstein, 2001). The discourses are dispersed, because they unfold in a variety of communication spaces of differing reach (Papacharissi, 2002). Finally, unlike traditional mass media, the web's carrying capacity is unrestricted (Hilgartner & Bosk, 1988), offering a seemingly endless and continually growing reservoir of ideas, statements, and opinions on a wide spectrum of issues.

Also in contrast to traditional mass media, online content is organized nonlinearly via hyperlinking (Park, 2003; Thelwall, 2006). Consequently, public discourses on the web are interconnected horizontally in many ways. Notions of the "networked public sphere" work best to describe the nature of this new form of public communication (Benkler et al., 2015; Castells, 2008; Friedland et al., 2006). An important consequence is the effect of blurring boundaries between different social spheres: Political news is only one click away from a shopping portal. This makes it particularly difficult to actually disentangle public discourses on the Internet.

Although the openness of the web leads to decentralization and diversity of web discourses, the growing connectivity has yielded new hierarchical structures. Several scholars have shown that the distribution of attention on the Internet is strongly skewed (power-law distributed) in favor of only a few prominent websites (e.g., Adamic & Huberman, 2000; Barabási & Albert, 1999; Pastor-Satorras & Vespignani, 2007).

These characteristics of web discourses are closely tied to the challenges of big data in general. First, openness leads to a growing amount of data, an increasing speed at which new data are produced, and a diversity of data formats and structures. These key challenges of big data are defined as the three Vs: volume, velocity, and variety (Laney, 2001). Second, interconnectedness has led to increasingly complex data sets, that is, increasing dimensions, overlaps, and links of data (Booz Allen Hamilton, 2013; Wu, Zhu, Wu, & Ding, 2014). A challenging consequence is growing noise. This is mirrored by increasing concern about the veracity (Booz Allen Hamilton, 2013) and validity (Boyd & Crawford, 2012; Vis, 2013) of big data analyses. Cleaning and reducing data is thus one of the biggest challenges in current web research (Marres & Weltevrede, 2013).

Noise in Issue Networks Collected by Web Crawlers

Web crawlers such as Issue Crawler (Rogers, 2010) take advantage of the network characteristic of the web by automatically collecting hyperlinks between web pages. Studies of hyperlink networks are based on the idea that hyperlinks are essential structural elements of online communication that are not set randomly but intentionally (Park, 2003; Thelwall, 2006). However, hyperlinks can be interpreted in many ways, which might be "ties of affinity, paths of communication, tokens of mutual aid in achieving public recognition, and/or potential avenues of coordination" (Burris, Smith, & Strahm, 2000, p. 215). Regardless of the specific reasons that lead to setting a hyperlink between two pages, a link always communicatively integrates the connected pages and their authors (Zimmermann, 2006).

Thus, communication scholars study hyperlink networks to learn about actor relationships and social structures online (De Maeyer, 2013; Pilny & Shumate, 2012), often also inferring conclusions about real-world connections between actors. Many scholars focus on organizational advocacy networks online and their role in social movements (e.g., Bennett & Segerberg, 2011; Carpenter & Jose, 2012; Shumate, 2012). Another group of studies aims

at mapping national or transnational "blogospheres" (Bruns et al., 2011; Etling, Kelly, Faris, & Palfrey, 2010; Moe, 2011). Recently, a growing number of scholars combined web crawling with content analysis techniques to learn more about what issues and narratives are actually communicated in these hyperlink networks (e.g., Benkler et al., 2015; Bennett et al., 2011; Bruns et al., 2011; Carpenter & Jose, 2012; Haider, 2014; Wallis & Given, 2016).

Assuming thematic relationships between interlinked pages, Marres and Rogers understand hyperlink networks collected by Issue Crawler as issue networks, that is, networks of hyperlinked web pages that all treat a particular issue (Marres, 2005; Marres & Rogers, 2005). However, given the operating mode of Issue Crawler, the resulting networks should be interpreted very cautiously. The tool starts from a given number of seed Uniform Resource Locators (URLs) and follows a specified number of internal links on the seed domains, as well as outgoing, external links to other domains. This technique draws on the assumption that—given the relevance of the seed pages for a specific issue area—the resulting network of web pages is spinning around the starting pages not only structurally but also thematically in the sense that it facilitates "the articulation of issues as public affairs" (Marres, 2005, p. 95).

A huge advantage of the use of web crawling is its operating range across different kinds of web sources. All web pages, forums, and blogs linked to the seed URLs are included in the network.[3] However, in the context of the open and interconnected web, this inclusive approach comes at the expense of noise. Bruns et al. (2011) emphasized that web crawlers treat every link on a web page equally. Thus, not only links within the text, but also navigation buttons or advertising links are followed. These might lead to pages that do not treat the issue of interest.

Depending on a study's aims, web pages that are thematically unrelated to the rest of the network may create problems concerning validity. If a researcher is interested in very general reference structures regardless of their respective social nature, mere hyperlink

networks might be appropriate. In this view, a page that is thematically unrelated to the issue under study but is linked to several relevant pages might be a functionally relevant part of a very loose issue network. However, if the focus is on the structures and content of issue-specific public discourses (e.g., on climate change, fair trade, food safety), thematically unrelated pages are usually considered noise. Not filtering out these irrelevant web pages might lead to biased or invalid conclusions about the structures of the issue-specific online discourse in question. The problem worsens if the web crawler is used as the first step of sampling issue-specific pieces of discourse on the web for a subsequent content analysis.

What specifically constitutes noise and what must be judged relevant in a study is closely connected to a particular research question and thus is left to specification by the researchers. Usually, researchers provide an issue definition that delineates the area of interest. In traditional content analysis, potentially relevant documents (that are, for instance, identified by a keyword search in electronic databases of press articles) are checked manually to see whether they match the issue definition and thus can be included in the sample. In the era of big data, this method is no longer feasible. Within only a few hours, web crawlers may collect networks of several thousand web pages. Not all of these pages can be manually checked for issue relevance. As we usually want to study the structure of issue networks as a whole, checking only a small sample of web pages will not suffice. Instead, the whole hyperlink network has to be automatically cleaned and filtered for issue relevance.

To sum up, due to the nature of Internet communication, noise is inevitable when web discourses are analyzed. At the same time, noise is problematic and should be taken into account as it might lead to invalid conclusions about issue networks and the structures of public online discourses.

Evaluating Filtering Strategies

In the following section, we evaluate three strategies for cleaning noisy hyperlink networks. The first two strategies (keyword filtering and machine-learning classification) focus on the thematic content of web pages. We chose these strategies because they correspond well to the traditional procedure of quantitative manual content analysis. Keyword search in electronic databases is a classic approach to defining a (potentially) relevant population of documents for content analysis. Machine-learning classification is an automated technique that supports and eventually substitutes manual relevance checks by human coders. The third strategy (extracting core networks), in contrast, uses network parameters to reduce data. This strategy is currently applied most often in studies that use Issue Crawler.

We tested each strategy on hyperlink networks on the issue of food safety collected with Issue Crawler for the United States and Germany. Due to recurring food scandals in the United States and in Europe, food safety has become an emotional issue of high news value (Anderson, 2000). The issue has been increasingly politicized and publicly discussed, engaging a growing number of civil society actors (Lang, Barling, & Caraher, 2009) who are heavily communicating online. Thus, we expected to find hyperlink networks that constitute issue networks surrounding this field of public interest.

The aim of the filtering process is to identify and delete irrelevant web pages from the network. This is a prerequisite (1) to be able to draw valid conclusions about the social structures of the online discourses on food safety in the two countries and (2) to generate an issue-specific population of online documents from which we can easily draw samples for manual content analysis.

We defined a relevant web document as containing at least one actor talking about a food safety problem. Defined in a broad sense, food safety includes all kinds of health problems and risks connected to food. The issue comprises many sub-issues such as food-

borne diseases, obesity, chemical additives, use of antibiotics in stock breeding, risks of genetically modified organisms, etc. However, the mere mention of the term food safety or a related food problem is not sufficient for our purposes of studying online discourse. Framing the issue according to Entman (1993, p. 52) implies promoting a particular problem definition, interpreting causes or consequences, and/or recommending treatments to the problem. We thus considered documents relevant if they contain an actor that formulates a problem definition and uses at least one of the other frame elements. Using this definition, we excluded from the analysis non-discourse items such as shops, login pages, privacy statements, calendar sites, and link lists.

These criteria are quite demanding concerning the filtering task. Another challenge is that food safety problems blur with neighboring issues such as other public health aspects, food security, or sustainability, adding extra potential for noise. Thus, the issue constitutes a particularly strong test case for our aim of finding effective data-filtering strategies.

Comparing the two countries' networks allows us to vary not only the language but also the cultural context. The countries are similar in that they are Western democracies with active civil society organizations and guaranteed freedom of speech. However, German online public spheres, particularly the blogosphere, have been found to be less developed and less active than in the United States (Bross, 2008). This difference could have a relevant effect on the suitability of the chosen filtering strategies.

In our analysis, we proceed as follows: First, we describe how we generated the hyperlink networks on food safety for the United States and Germany. Then, we apply each filtering strategy to the networks and evaluate the strategy's effectiveness in reducing noise.

Generating Hyperlink Networks

We used the web crawling tool Issue Crawler to generate snowball hyperlink networks on which we base our further analyses. The crawler automatically identifies and collects all hyperlinks starting from predefined URLs.

Identifying source seeds. We chose eight seed URLs per country that represent the websites of civil society organizations that actively engage in the public discourse on food safety. To identify the seed URLs, we performed Google searches for 10 search terms in English and German in May 2011 (we used google.com for the United States, google.de for Germany; see Table 1 for the URLs and the search terms). Additionally, we consulted literature and food experts to include the entire spectrum of important actors. We eventually identified a list of possible actors and their domains (52 for Germany, 68 for the United States). For each website, we checked whether it was technically available, food safety was an important topic, and the respective section concerning food safety was up-to-date. Only source seeds that fulfilled all three criteria were selected.

[Table 1 about here]

Crawling mechanisms. The hyperlink networks were collected in the beginning of November 2012. We chose the snowball method as the crawling procedure since it is the least restrictive crawling method: Without any other preconditions, the method includes any website that is linked to by a website already in the network. Snowball networks tend to become vast and barely manageable within only a few crawling steps. To avoid this, we chose low values for the crawl parameters: a depth of two levels and one degree of separation. The crawl depth designates the vertical dimension of crawling, i.e., internal links within a website whereas the degree of separation corresponds to the horizontal dimension of crawling, dealing with links that point to external websites.

The crawling sequence with our specified settings proceeded in the following way: First, starting from the source seed pages, the crawler followed all pages on the same website's domain and included them in the network within the reach of two links (depth = 2). Then, the crawler collected every link from these pages to pages external to the starting websites' domains, to web pages of other domains. Starting from these pages, the related domains were also vertically crawled, meaning that all internal links were followed up to a depth of two and added to the network. Finally, all of the links pointing back to any other web page already in the network were also included in the network. For visualization and interpretation, the hyperlink networks, which were gathered at the page level, were then aggregated to the level of domains; all web pages on the same website were merged in one node in the network.

Resulting hyperlink networks. The two raw hyperlink networks are depicted in Figure 1. The unfiltered networks contained 1,506 domains that included 17,331 pages for the United States and 1,112 domains that included 16,206 pages for Germany. Issue Crawler identified and listed the URLs of the pages in the network. We then archived these web pages with the tool wget[4] directly after the crawl.

Keyword Filtering

The first of the filtering strategies was straightforward. To track web pages with relevant content, we performed a keyword search on the web pages in the network using the indexing software Visual Web Spider.[5] Following our broad definition of food safety, we searched generally for mentions of the term food safety itself and—following de Jonge, Van Trip, Renes, and Frewer (2010)—for a more specific combination of keywords that denominate food and possible associated problems (for the list of keywords, see Table 2). The program accessed all URLs listed in the raw networks and registered the occurrence of the predefined keywords.[6] The tool could be adjusted in order to avoid indexing certain types of

non-text content (e.g., PDF documents) or specific websites (e.g., search engines such as google.com or social networking platforms such as facebook.com)—options that we chose. Finally, we deleted web pages (and their domains) from the network that did not contain the keywords.

[Table 2 about here]

Classifying Documents with Machine-Learning Algorithms

To automatically classify the corpus documents as either relevant or irrelevant, we developed a text categorization procedure using RapidMiner.[7] From a set of pre-classified documents, the software autonomously derives probabilistic rules that replicate the given categorizations. The classifying algorithm is capable of recognizing relevant text characteristics and applying them to a set of unclassified text documents (Sebastiani, 2002). The development of the algorithm is very similar to the process of training a human coder (Scharkow, 2013). Human coder decisions serve as the training basis as well as the benchmark against which the classifier's performance is tested.

Training sample. First, the text-mining tool had to be familiarized with four distinct text categories: relevant German, irrelevant German, relevant English, and irrelevant English texts. To generate a sample of pre-classified texts, four coders manually checked random samples of web pages regarding their issue relevance. The pages belonged to keyword-filtered networks that we collected in June, July, and August 2012. We applied the keyword filter to make sure that we found enough relevant examples since the raw networks mainly contained noise (see Table 3). To increase the number of positive examples and to enhance the naïve classifier's ability to identify potentially relevant documents, we also included documents in the relevant training set that reported a food problem but did not entail a complete actor-frame sequence as defined. The final training sample comprised a total of 593 relevant texts (269 in German, 324 in English) and 1,214 irrelevant texts (424 in German, 790 in English).[8]

To ensure intercoder reliability, we randomly selected 30 English and 30 German pages from the networks collected in June 2012. We checked reliability against a previously defined gold standard, the best coding decision defined in consensus coding sessions of the two researchers involved in the development of the codebook. The average percent agreement of the four coders with the gold standard reached .92 (Holsti coefficient).

Training of the classifier. We imported the manually classified HTML files, each labeled with one of the four categories, in RapidMiner and then preprocessed the files by applying natural language processing (NLP) methods.[9] The software then created a term document matrix that represented the term frequencies for each document in the training corpus.[10]

Based on the manually classified texts, RapidMiner created a statistical model that predicted whether the thus far unclassified texts belonged to one of the four categories. The categorization model was based on the naïve Bayes algorithm. This fast and robust algorithm is often used for topic classification (Hillard, Purpura, & Wilkerson, 2008; Scharkow, 2013) and differentiates between multiple categories. For every page, the algorithm calculated probability values that indicated the confidence in the attribution of a page to each category. The probability thresholds were set by convenient default at 50%: If the classifier's confidence value for one of the four classes was > 50% for a document, then the document was attributed to that class. Considering the binary nature of the decision, this threshold was reasonable: A document in a specific language is either relevant or not.

We applied ten-fold cross-validation. The training corpus was split into 10 subsets of equal size. Nine subsets were used as training data, and one subset was used as testing data. Based on the training data, the machine predicted classifications of the testing data and compared them to the true classifications. The process was repeated 10 times, with each subset serving exactly once as testing data. The overall accuracy of the final model was 80%;

that is, 80% of the predicted classifications were correct. The mean class precision reached 79% (*SD* = 6.5%) and the mean class recall 78% (*SD* = 10%).

Document classification. Finally, we performed the actual classification process and applied the statistical model to the total corpus. It was crucial to preprocess the unclassified texts with the same NLP methods as during the training of the algorithm. After we had classified all pages, we deleted from the networks all pages the classifier identified as irrelevant.

Extracting Core Networks

Instead of considering the content of the web pages, another filtering strategy uses network measures to extract a core network from the original snowball network. We refer to core networks as densely connected components of the raw snowball networks. The underlying theoretical assumption is that these core networks are also thematically more homogeneous than larger and less dense networks. The logic of this approach is based on the inherent hierarchical structures of online networks. Researchers have shown that in most online networks only a few pages or posts receive many links from others (Adamic & Huberman, 2000; Barabási & Albert, 1999; Pastor-Satorras & Vespignani, 2007).

Obviously, many different network measures can be applied. Most researchers studying hyperlink networks define some threshold of in-degree as a measure of visibility or authority for the nodes in the network (Thelwall, 2006). For example, co-link networks require a node have at least two incoming links from other nodes to be included in the core network. Issue Crawler offers a procedure that applies the co-link logic to the level of pages. Consequently, most studies that rely on Issue Crawler use this strategy (e.g., Bennett & Segerberg, 2011; Carpenter & Jose, 2012; Haider, 2014; Wallis & Given, 2016). In other studies, researchers defined stricter thresholds (e.g., Etling et al., 2010) or considered the top sites or pages with the highest number of in-links (e.g., Rogers & Ben-David, 2008). Other

centrality measures such as the page rank algorithm could also be applied (Brin & Page, 1998).

We checked whether the logic of creating core networks would also lead to reduced noise in the issue networks. For reasons of processing capacity, we applied the co-link strategy directly to the level of websites, which implies considerably fewer nodes than the level of web pages. For every website in the hyperlink networks, we calculated the incoming links and filtered out websites that received fewer than two in-links from other sites. As a result, we obtained core networks composed of websites with an in-degree of at least two. At the lower hierarchical level, the networks included all pages from these websites; the co-link rule did not apply to this level of web pages. Thus, compared to the Issue Crawler mechanism, the core networks were less strictly defined.

Empirical Evaluation of the Filtering Strategies

We applied each filtering strategy to the raw snowball network collected by Issue Crawler in November 2012. From each of the raw and filtered networks, we drew random samples of 100 web pages. We then checked the pages manually for issue relevance as defined. Table 3 shows the network characteristics and corresponding noise rates for each crawl and filtering step.

[Table 3 about here]

Raw networks. The snowball crawl resulted in hyperlink networks of 17,331 web pages on 1,506 websites in the U.S. network and 16,206 pages on 1,112 sites in the German network. In both crawls, due to crawler-blocking tools and other technical reasons, around 15% of the web pages were not downloaded properly by the web-archiving software wget and thus could not be classified by the coders. The noise rates of the manually checked pages were very high but differed considerably between the two countries: For the U.S. network,

72% of the sampled web pages were rated irrelevant whereas for the German network 90% of the web pages were rated irrelevant.

The high amounts of noise occurred mainly because the crawler follows hyperlinks in text not only on web pages but also on advertising banners and navigation bars. This leads to irrelevant web sources in the snowball network, such as web shops, link lists, calendars, and registration sites.

Keyword-filtered networks. As a result of the keyword check, the number of sites and pages in the networks was significantly reduced. In the U.S. case, 54% of the original sites and 58% of the original pages remained in the network. In the German case, only 35% of the sites and 19% of the pages passed the keyword filter. The number of irrelevant pages was reduced to 55% in the U.S. network and 67% in the German network. For both cases, this equals a decrease of about 25% compared to the raw networks. A positive side effect of the keyword check was that the number of un-retrievable web pages in the archives decreased to approximately 5% for both countries.

Classified networks. For both networks, automated classification was equally effective, reducing noise rates by about 40% compared to the raw networks. The coders rated 43% of the pages in the U.S. network and 55% of the pages in the German network as irrelevant. The number of sites and pages in the networks also decreased heavily compared to the unfiltered snowball network. After classification, only 32% of the original sites and 26% of the original pages remained in the U.S. network. In the German case, only 19% of the sites and 8% of the pages remained. Only web pages that had been downloaded correctly remained in the network.

Co-link networks. Reducing the keyword-filtered networks to co-link networks on the site level did not yield the same outcome as the content-based filtering strategies. The mechanism shrunk both networks considerably to approximately 50% of the original sites and

around 60% of the original pages. However, the noise rates were much higher: For the U.S. case, 72% irrelevant pages were registered, which is the same percentage as in the unfiltered network. For the German network, the noise rates improved only slightly to 82%. For both networks, more than 20% of the sampled pages were not retrievable for manual checks.

With pairwise chi-square tests, we verified whether the percentages of relevant pages differed significantly among all four networks, the raw network and the three filtered networks. Results are reported in Table 4. The keyword-filtered network and the classified network contained significantly more relevant pages than the unfiltered network and the co-link network. However, the differences between these two content-based filtering strategies were not statistically significant. The co-link filter, in contrast, did not significantly increase the percentage of relevant pages compared to the raw network.

[Table 4 about here]

Figure 1 illustrates how the network structures changed with each filtering strategy. The images map the sites in each crawl as nodes and the links between the nodes as edges.[11] For the U.S. network, keyword filtering and classification reduced the size of the network but not the fundamental structure. The basic pattern of a core network in the center of the graph with several clusters reaching out to the periphery can be recognized even in the greatly diminished keyword-filtered and classified networks. In the co-link network, the peripheral clusters decreased considerably in size compared to the core network in the center, complying with the original aim of applying the co-link strategy.

In the German case, the structure of the raw network with a noise rate of 90% fundamentally changed through the filtering processes. Given that the keyword-filtered network contained only 35% and the classified network only 19% of the initial sites, it is not surprising that only rudimentary leftovers of the original network patterns were found. As in

the U.S. case, the German co-link network differed from the raw network in the expected way—the original network was condensed to a smaller core network.

[Figure 1 about here]

Discussion

This paper dealt with the problem of noise in the analysis of web discourses in the era of big data. After providing a general theoretical introduction to the problem, we specifically addressed the challenge of extracting issue networks from hyperlink networks. We evaluated three filtering strategies in order to reduce as much noise as possible: keyword filtering, machine-learning document classification, and extraction of the core networks. The test cases for empirically evaluating these filtering strategies were hyperlink networks for the issue of food safety in Germany and the United States.

We discovered very high noise rates in the original, unfiltered networks of both countries: 72% of the crawled web pages in the United States and 90% of the web pages in the German network were found to be irrelevant, i.e., not containing an actor-frame statement on food safety. Consequently, the unfiltered snowball networks are not issue networks in the strict sense but should be interpreted as general interlinking structures, reaching out from civil society actors in an issue field for whatever reason.

As we showed, the differences in the noise rates between the countries are striking. The unfiltered U.S. network contained considerably more relevant web pages than the German network. These differences remained irrespective of the filtering strategy applied—a hint of the stronger and more discursive nature of the U.S. food safety network. As previous studies have shown, in the United States more civil society actors and blogs are explicitly specialized on the food safety issue, whereas in Germany the issue is mainly pursued by multiple-issue groups such as environmental or consumer organizations. In addition, U.S.

challengers in food safety more actively connect to each other and link up with online media (Pfetsch, Maier, Miltner, & Waldherr, 2016).

Ultimately, which strategy is best suited to extract issue networks that can be interpreted meaningfully? Although automated classification was the most effective filtering strategy for our samples, the differences in noise rates were not statistically significant compared to the keyword filtering strategy. Thus, both content-based filtering strategies should be considered valuable approaches for extracting issue networks. Furthermore, both strategies were equally powerful for the German and U.S. networks and reduced the original noise by about 25% through keyword filtering and 40% through automated classification. The main advantage of machine-learning classification is that the algorithm inductively identifies distinguishing features between relevant and irrelevant content. This makes the algorithm a powerful data-cleaning tool for fuzzy issues (such as food safety) when it is difficult to grasp the issue with only a limited number of keywords. Additional training, with a bigger training corpus representing all subtopics in the total corpus, could further improve results. However, there are inherent limits to these efforts. Scharkow (2013) compared human and automated classification by the gold standard and found that automated classification is generally about 15% less reliable than human coders. In our case, this means that the best automatic classifier would probably still produce noise rates of about 25%. In addition, much of the remaining noise in the networks may be due to the strict definition of issue relevance that we applied.

However, training a machine-learning algorithm is a time-consuming effort whereas checking for keywords is a quick and easily applicable filtering strategy. For more clear-cut issues, a keyword filter thus might be the better choice. The keyword check with Visual Web Spider was crude; all keywords, no matter where on the page they were found, were indexed. For instance, Visual Web Spider does not differentiate whether the keyword is part of the text

or of an advertisement, a navigation element, or the source code. This might explain most of the noise that remained in the network. Adding scraping tools to identify the relevant textual parts on the web page could further improve results. However, since the web crawler collects a variety of web pages that are all structured differently, this is a tricky task. Another disadvantage of the keyword filter is that it might exclude relevant web pages that comply with our issue definition but do not contain the keywords.

The strategy of extracting co-link networks at the level of websites was not effective for our case. Although this method helped reduce the size of the networks and focus on a more densely connected center of websites, it did not lead to networks with significantly more relevant pages. Other than expected, issue-specific discourse is not bound to the core networks but is scattered all over the center as well as the periphery. This held for the networks in both countries. Nevertheless, more strict specifications of core networks might yield better results, such as co-linking on the page level or defining higher in-degree thresholds. This should be tested with future research.

This study is also limited in that it encompassed only one issue and two languages. Future research should apply the test to additional issues with different characteristics in the context of more languages. Nevertheless, with the broad and fuzzy issue of food safety and with our strict definition of issue relevance, we chose a hard test setting.

The main message to be taken from our study is that hyperlink networks must be filtered in order to interpret them as issue networks in the sense of web discourse about an issue. Of course, with the snowball crawl we chose the least restrictive crawler setting that was likely to produce noise in the collected hyperlink networks. However, we have also shown that we cannot necessarily expect less noise in the core networks. Considering the high requirements for issue relevance, it is also clear that even the most effective filtering strategies are not perfect and still must be complemented by manual checks of the documents.

However, automatically pre-filtering potentially relevant documents facilitates and speeds up this process considerably.

Finally, we have to critically unpack the basic assumption that researchers know best a priori how the relevance of an issue can be defined and that human coders can best assess relevance according to this definition. Thus, researchers define the gold standard against which the filtering algorithms are benchmarked. This approach entails the danger of a priori excluding other associations with food safety from the analysis and missing relevant parts of the actual discourses in the networks. An alternative approach would be to apply more inductive, data-driven methods, such as probabilistic topic models to identify common topics in our text corpus (Blei, 2012), or analyzing word co-occurrences on the pages in the hyperlink networks (de Bakker & Hellsten, 2013). This knowledge could then be used to choose the parts of the corpus (and the network) that are of interest for further in-depth analysis.

Author Biographies

Annie Waldherr is a postdoctoral researcher at the Institute for Media and Communication Studies at the Freie Universität Berlin. She received her PhD from the Freie Universität Berlin in 2011 for her dissertation on the dynamics of media attention. In 2005, she graduated with a degree in communication sciences from the University of Hohenheim in Stuttgart. Her research interests include mediated public spheres, political online communication, science and technology discourses, and social simulation. Email: annie.waldherr@fu-berlin.de

Daniel Maier is a research associate and PhD candidate at the Institute for Media and Communication Studies at the Freie Universität Berlin. He graduated with a degree in political science from the University of Passau in 2009 and received a master's degree in communication science from the Freie Universität Berlin in 2012. His research interests include network analysis methods. Email: maier@zedat.fu-berlin.de

Peter Miltner is a research associate and PhD candidate at the Institute for Media and Communication Studies at the Freie Universität Berlin. He graduated with a degree in communication sciences from the University of Hohenheim in Stuttgart in 2009 and holds a master's degree in European interdisciplinary studies from the College of Europe in Warsaw (2010). His research interests include political (online) communication and network analysis. Email: peter.miltner@fu-berlin.de

Enrico Günther is a master's degree candidate in the Department for EU International Relations and Diplomacy Studies of the College of Europe in Bruges, Belgium. He graduated from the Freie Universität Berlin in 2012 and holds a master's degree in European studies (2014) from the Viadrina European University in Frankfurt Oder, Germany, and the Institute for Political Studies in Strasbourg, France. His research interests include European public spheres, EU foreign policy, and content analysis. Email: enrico.gunther@coleurope.eu

References

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the World Wide Web. *Science, 287*(5461), 2115. doi:10.1126/science.287.5461.2115a

Anderson, W. A. (2000). The future relationship between the media, the food industry and the consumer. *British Medical Bulletin, 56*, 254–268. Retrieved from http://bmb.oxfordjournals.org

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Benkler, Y., Roberts, H., Faris, R., Solow-Niederman, A., & Etling, B. (2015). Social mobilization and the networked public sphere: Mapping the SOPA-PIPA debate. *Political Communication, 32*, 594–624. doi:10.1080/10584609.2014.986349

Bennett, W. L., Foot, K., & Xenos, M. (2011). Narratives and network organization: A comparison of fair trade systems in two nations. *Journal of Communication, 61*, 219–245. doi:10.1111/j.1460-2466.2011.01538

Bennett, W. L., & Segerberg, A. (2011). Digital media and the personalization of collective action. *Information, Communication & Society, 14*, 770–799. doi:10.1080/1369118X.2011.579141

Blei, D. M (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77–84. doi:10.1145/2133806.2133826

Booz Allen Hamilton. (2013). *The field guide to data science.* Retrieved from http://www.boozallen.com/insights/2013/11/data-science-field-guide

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*, 662–679. doi:10.1080/1369118X.2012.678878

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*, 107–117. doi:10.1016/S0169-7552(98)00110-X

Bross, J. F. M. (2008). Weblogs, a promising new form for E-democracy? *Proceedings of the Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference* (Vol. 3, pp. 667–671). doi:10.1109/WIIAT.2008.268

Bruns, A., Burgess, J., Highfield, T., Kirchhoff, L., & Nicolai, T. (2011). Mapping the Australian networked public sphere. *Social Science Computer Review, 29*, 277–287. doi:10.1177/0894439310382507

Burris, V., Smith, E., & Strahm, A. (2000). White supremacist networks on the Internet. *Sociological Focus, 33*, 215–235. doi:10.1080/00380237.2000.10571166

Carpenter, R. C., & Jose, B. (2012). Transnational issue networks in real and virtual space: The case of women, peace and security. *Global Networks, 12*, 525–543. doi:10.1111/j.1471-0374.2012.00363.x

Castells, M. (2008). The new public sphere: Global civil society, communication networks, and global governance. *The ANNALS of the American Academy of Political and Social Science, 616*, 78–93. doi:10.1177/0002716207311877

Dahlgren, P. (2005). The internet, public spheres, and political communication: Dispersion and deliberation. *Political Communication, 22*, 147–162. doi:10.1080/10584600590933160

De Bakker, F. G. A., & Hellsten, I. (2013). Capturing online presence: Hyperlinks and semantic networks in activist group websites on corporate social responsibility. *Journal of Business Ethics*, *118*(4), 807–823. doi:10.1007/s10551-013-1962-1

De Jonge, J., Van Trijp, H., Renes, R. J., & Frewer, L. J. (2010). Consumer confidence in the safety of food and newspaper coverage of food safety issues: A longitudinal perspective. *Risk Analysis, 30*(1), 125–42. doi:10.1111/j.1539-6924.2009.01320.x

De Maeyer, J. (2013). Towards a hyperlinked society: A critical review of link studies. *New Media & Society, 15*, 737–751. doi:10.1177/1461444812462851

Dutton, W. H. (2009). The fifth estate emerging through the network of networks. *Prometheus, 27*(1), 1–15. doi:10.1080/08109020802657453

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*, 51–58.

Etling, B., Kelly, J., Faris, R., & Palfrey, J. (2010). Mapping the Arabic blogosphere: Politics and dissent online. *New Media & Society, 12*, 1225–1243. doi:10.1177/1461444810385096

Friedland, L. A., Hove, T., & Rojas, H. (2006). The networked public sphere. *Javnost - The Public, 13*(4), 5–26. Retrieved from http://javnost-thepublic.org

Habermas, J. (1995). *Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft* [Structural transformation of the public sphere: Inquiry into a category of bourgeois society]. Frankfurt am Main, Germany: Suhrkamp.

Haider, J. (2014). Taking the environment online: Issue and link networks surrounding personal green living blogs. *Online Information Review, 38*, 248–264. doi:10.1108/OIR-03-2013-0052

Hilgartner, S., & Bosk, C. L. (1988). The rise and fall of social problems: A public arenas model. *The American Journal of Sociology, 94(*1), 53–78.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research, *Journal of Information Technology & Politics, 4*(4), 31–46. doi:10.1080/19331680801975367

Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety.* Stamford, CT: META. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lang, T., Barling, D., & Caraher, M. (2009). *Food policy: Integrating health, environment and society*. Oxford, England: Oxford University Press.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*, 20–33. doi:10.1080/08838151.2012.761700

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis: University of Minnesota Press.

Marres, N. (2005). *No issue, no public: Democratic deficits after the displacement of politics* (Doctoral dissertation). University of Amsterdam, Netherlands. Retrieved from http://dare.uva.nl/document/17061

Marres, N., & Rogers, R. (2005). Recipe for tracing the fate of issues and their publics on the Web. In B. Latour & P. Weibel (Eds.), *Making things public: Atmospheres of democracy* (pp. 922–935). Cambridge, MA: MIT Press.

Marres, N., & Weltevrede, E. (2013). Scraping the social? Issues in live social research. *Journal of Cultural Economy, 6*(3), 313–335.

Moe, H. (2011). Mapping the Norwegian blogosphere: Methodological challenges in internationalizing internet research. *Social Science Computer Review, 29*, 313–326. doi:10.1177/0894439310382511

Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New Media & Society, 4*, 9–27. doi:10.1177/14614440222226244

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the Web. *Connections, 25*, 49–61. Retrieved from https://insna.org/connections_archives.html

Pastor-Satorras, R., & Vespignani, A. (2007). *Evolution and structure of the Internet: A statistical physics approach.* New York, NY: Cambridge University Press.

Pilny, A., & Shumate, M. (2012). Hyperlinks as extensions of offline collective action. *Information, Communication and Society, 15*, 260–286. doi:10.1080/1369118X.2011.606328

Pfetsch, B., Maier, D., Miltner, P., & Waldherr, A. (2013, September). *Online networks of challengers in food policy: A comparative study of structures and coalitions in Germany, UK, US and Switzerland.* Paper presented at the 7th ECPR General Conference, Bordeaux, France.

Rogers, R. (2010). Mapping public Web space with the Issuecrawler. In C. Brossard & B. Reber (Eds.), *Digital cognitive technologies: Epistemology and knowledge society* (pp. 115–126). London, England: Wiley.

Rogers, R., & Ben-David, A. (2008). The Palestinian—Israeli peace process and transnational issue networks: The complicated place of the Israeli NGO. *New Media & Society, 10*, 497–528.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology, 41*, 885–899. doi:10.1177/0038038507080443

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity, 47*, 761–773. doi:10.1007/s11135-011-9545-7

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*, 1–47. doi:10.1145/505282.505283

Shumate, M. (2012). The evolution of the HIV/AIDS NGO hyperlink network. *Journal of Computer-Mediated Communication, 17*(2), 120–134. doi:10.1111/j.1083-6101.2011.01569.x

Sunstein, C. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology, 57*, 60–68. doi:10.1002/asi.20253

Toepfl, F., & Piwoni, E. (2015). Public spheres in interaction: Comment sections of news websites as counterpublic spaces. *Journal of Communication, 65*, 465–488. doi: 10.1111/jcom.12156

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday, 18*(10). doi:10.5210/fm.v18i10.4878.

Wallis, J., & Given, L. M. (2016). #digitalactivism: New media and politics. *First Monday, 21*(2). doi:10.5210/fm.v21i2.5879

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97–107. doi:10.1109/TKDE.2013.109

Zimmermann, A. C. (2006). *Demokratisierung und Europäisierung online? Massenmediale politische Öffentlichkeiten im Internet* [Democratization and europeanization online? Mass-mediated political public spheres on the Internet] (Doctoral dissertation). Freie Universität Berlin, Germany. Retrieved from http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000003532

Notes

[1] For a detailed documentation on the tool, please consult: http://www.govcom.org/

[2] The research design of the project is comparative and longitudinal, assessing issue-specific online networks regarding climate change and food safety in four countries (Germany, Switzerland, United Kingdom, and the United States) over a period of 30 months and analyzing under what conditions issues and frames spill over from online issue networks to traditional print media and to the official political agenda.

[3] However, there is an important restriction concerning social media. In a hyperlink crawl with the Issue Crawler, social networks such as Twitter or Facebook appear as a single web domain. Therefore, single pages or profiles that are part of these social networks cannot be identified.

[4] Information and download are available at: http://www.gnu.org/software/wget/

[5] For further information, see: http://www.newprosoft.com/web-spider.htm

[6] Some pages could not be indexed for several reasons, e.g., if they were not accessible, could not be found, or were password-protected (United States: 2.1%, Germany: 1.8%).

[7] For more information on the software, visit: http://rapid-i.com

[8] The networks were collected within the framework of an ongoing research project. The data stem from networks in four countries under study (Germany, Switzerland, the United Kingdom, and the United States). For the pre-classification process, we recurrently drew random samples of 50 pages each according to a probability proportional to size logic: web pages with many in-links had a proportionally higher chance of becoming part of the sample. The sampling and manual relevance checks continued until a sufficient number of relevant texts were assembled that led to an acceptable outcome of the automated classifier (cf. following footnote).

[9] The following operators were used to eliminate artificial differences that were not related to differences in content: transform cases to lower cases, exclude stop words (most common, short function words without a distinctive meaning), and reduce terms to stem terms (e.g., “dangerous” transformed to “danger”).

[10] In addition, the frequencies of the combinations of two consecutive terms were analyzed. To avoid extensive computing times, only the terms and consecutive combinations of two terms that appeared at least five times in all the documents were processed.

[11] The network visualizations were produced using Yifan Hu’s algorithm as implemented in open-source software Gephi (https://gephi.org/).

Acknowledgements

This publication originated in the context of the Research Unit Political Communication in the Online World (1381), sub-project 7, funded by the German Research Foundation (DFG) and the Swiss National Science Foundation (SNSF).

Table 1. Starting URLs and Google Search Terms for Their Identification.

Starting URLs	Google Search Terms
United States	
http://www.centerforfoodsafety.org/ http://www.cspinet.org/foodsafety/ http://www.foodandwaterwatch.org/food/ http://www.organicconsumers.org/foodsafety.cfm http://notinmyfood.org/newsroom http://barfblog.foodsafety.ksu.edu/barfblog http://www.greenpeace.org/international/en/campaigns/agriculture/ http://www.pewhealth.org/topics/food-safety-327507	food safety, safe + food, food scandal, GM foods, food + consumer protection, food + consumers, food + risk, food safety + campaign, food + labelling, food safety + control
Germany	
http://www.aid.de/verbraucher/lebensmittelsicherheit.php http://www.vzbv.de/Ern%C3%A4hrung.htm http://www.foodwatch.de http://www.verbraucher.org/verbraucher.php/cat/3/title/Ern%E4hrung http://www.greenpeace.de/themen/landwirtschaft/ http://www.verbraucher-papst.de/category/essen-und-trinken/ http://www.meine-landwirtschaft.de/ http://www.slowfood.de/	Lebensmittelsicherheit, Sicher + Lebensmittel, Lebensmittelskandal, Genfood, Lebensmittel + Verbraucherschutz, Lebensmittel + Konsumenten, Lebensmittel + Risiko, Lebensmittelsicherheit + Kampagne, Lebensmittel + Kennzeichnung, Lebensmittelsicherheit + Kontrolle

Table 2. Search Terms for Visual Web Spider.

Issue Label (A)	Food Terms (B)	Food Safety Problems and Food Regulation Bodies (C)
United States		
Food safety	Food, aliment, feed	Germ, epidemic, scare, illness, health, infected, borne, contagious, contaminated, polluted, GM food, genetical, hazard, bioengineer, harmful, scandal, hygiene, risk, EFSA, FDA, FSA
Germany		
Lebensmittel-sicherheit	Lebensmittel, Nahrung, Futter	Erreger, Keim, Epidemie, Seuche, Krankheit, Gesundheits, Infiziert, Verunreinig, Kontamin, Belast, Gentechni, gefähr, Gefahr, Skandal, Hygien, Risiko, EFSA, BVL, BAG

Note. The search was performed in English and German. Some terms were deleted after a certain point to ensure that related terms were tracked, e.g., a search for "genetical" may include "genetical" or "genetically." Pages that contained a term in the first column or a combination of terms from the second and third columns were considered relevant; thus, A OR (B AND C).

Table 3. Network Characteristics and Corresponding Noise Rates.

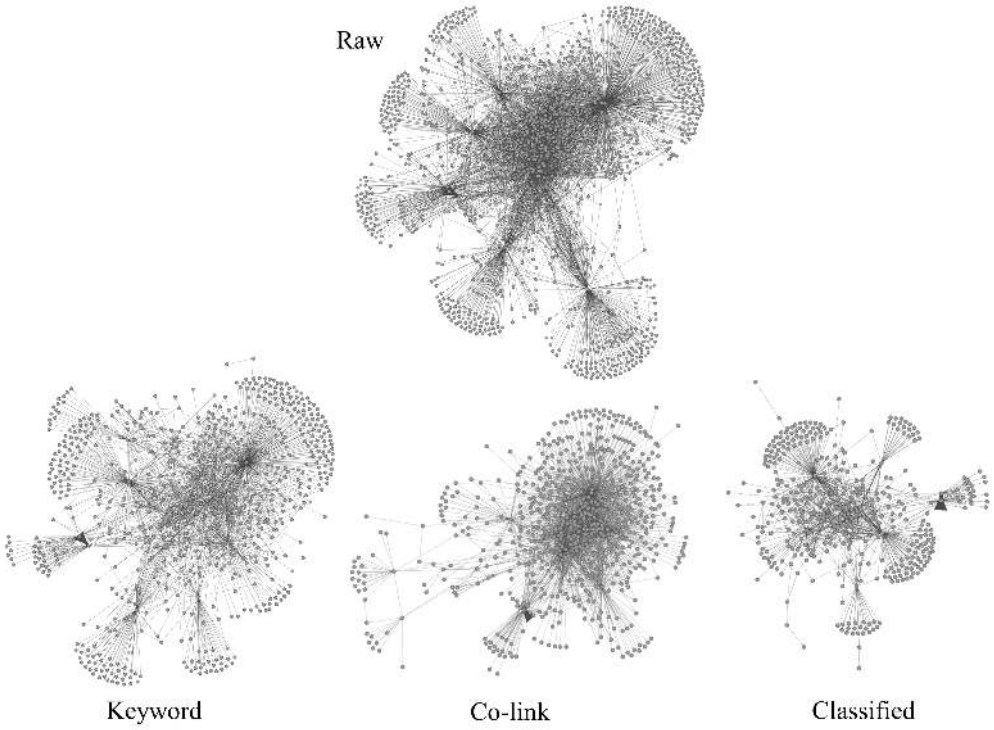
	United States					Germany				
	#		%			#		%		
	Sites	Pages	**Noise**	True	n	Sites	Pages	**Noise**	True	n
Raw	1,506	17,331	**72**	28	85	1,112	16,206	**90**	10	84
Keyword	818	10,225	**55**	45	94	387	3,114	**67**	33	95
Classified	489	4,565	**43**	57	100	206	1,339	**55**	45	100
Co-link	707	10,866	**72**	28	74	562	10,043	**82**	18	79

Note. Each sample comprises 100 randomly chosen web pages. # Sites = number of collected website domains in the network, # Pages = number of web pages in the network; % Noise = percentage of irrelevant documents in the manually checked sample of web pages; % True = percentage of relevant documents in the sample; n = number of retrievable pages. We calculated the percentage of noise by dividing the number of relevant pages by the number of retrievable pages. Pages that we were not able to open or find were not regarded as noise.

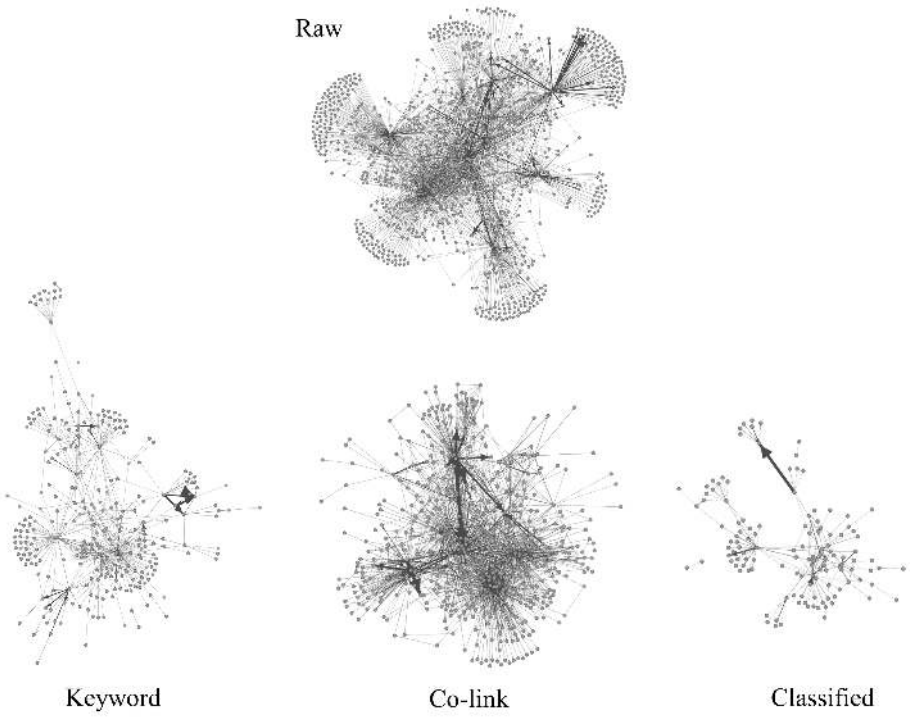
Table 4. Probability Values of Chi-Square Tests for Significant Differences between Filtering Strategies.

	United States				Germany			
	Snow-ball	Key-word	Classi-fied	Co-link	Snow-ball	Key-word	Classi-fied	Co-link
Raw	-	.03*	<.01**	n.s.	-	<.01**	<.01**	n.s.
Keyword	-	-	n.s.	.04*	-	-	n.s.	.03*
Classified	-	-	-	<.01**	-	-	-	<.01**
Co-link	-	-	-	-	-	-	-	-

Note. Pearson's chi-square tests with Yates' continuity correction were calculated. P values of less than .05 were regarded as indicators of statistically significant differences between the numbers of relevant pages in the samples that resulted from the respective filtering strategies.



(a) United States



(b) Germany

Figure 1. Network Structures Before and After Filtering.