Exploring issues in a networked public sphere

Combining hyperlink network analysis and topic modeling

{ABSTRACT}

The goals of this article are twofold: First, we present a methodological approach to analyze the content-structure of a hyperlink-network representing a specific segment of an issue public in the US. Using the case of the food safety movement we demonstrate how to generate a hyperlink network with the web crawling tool *Issuecrawler* and merge it with results of a *topic modeling* process that we applied to the networks' content data. Combining hyperlink network and content analysis allows us to interpret such a network in its entirety and with regard to different sub-issues. Second, our analysis is set up to explore issue-related mobilization potentials of the US food safety movement. By analyzing the online communication of NGOs and civil society actors in this sector we are keen to detect their specific communication strategies and on this basis develop hypotheses about issue-related mobilization potentials for further testing.

Key words: network analysis, topic model, food safety, movement, online communication

1. Introduction

For communication researchers, the Internet has not only brought up a variety of new and compelling research questions, but also put an immense pressure on the development of appropriate methods. When turning to analyze phenomena of web communication, online public spheres and social media networks, scholars had to adapt traditional methods such as content analysis or surveys as well as to develop new approaches. In this context, social network analysis and automated content analysis are probably the most fundamental and promising techniques for communication research. During the last years, tremendous efforts have been put into their further development and application to web research. Recently, more and more scholars also combine both approaches to study networks and content in conjunction (Ali-Hasan & Adamic, 2007; Bennett, Foot & Xenos, 2011; Carpenter & Jose, 2012; Gibson et. al, 2013; Hyun Kim, 2012; Tateo, 2005; Tremayne et al., 2006; Williams et al., 2005).

With our paper, we want to contribute to this ongoing challenge. Our goals are twofold: First, from a methodological perspective, we develop and test an approach, which allows for an insight into the semantic content dimensions of hyperlink networks. Specifically, we combine web crawling methods to generate an issue network with the computational text mining approach of topic modeling. This combination allows us to trace not only the social connections between actors online but also what messages they specifically put forth in which parts of the network.

Second, our empirical analysis is used to study online communication of civil society actors in the food safety movement sector. Our objective has been to scrutinize the issue network of food safety NGOs and develop hypotheses about their issue-dependent mobilization potentials in online public spheres. For this aim we analyze the hyperlink-network of NGOs and related actors of food safety advocacy in the United States (US). We focus on the simultaneous analysis of the network's structure and the synchronization of content within this network.

The broad issue of food safety comprises a variety of sub-issues which are supposed to bear differently on mobilizing potentials. Thus, by identifying the linkages between actors and discover in which specific sub-issues they work together we get a full picture on the multiple dimensions and layers of the debate in this community. Therefore, one goal of this paper has been to explore the multiplicity of sub-issues in the US food safety network. Based on this insight we further develop a hypothesis about sub-issue specific mobilization potentials.

Within a broader theoretical perspective the multi-layered and interlinked Internet communication of organizations and individuals in a specific issue sector can be discussed as a networked public sphere. The prevalence of certain sub-issues and structural connectivity can then be treated as indicators of the fragmentation of the online public sphere. In fact, one can assume that the more fragmented an issue network appears the less potential it develops for strong public mobilization of their cause. Notwithstanding the larger structure of online public spheres, and the food safety NGO community in particular, the main purpose of this paper is to demonstrate our methodological approach of combining network analysis and content analysis.

We proceed as follows: In section 2, we introduce the concept of the networked public sphere as the basic theoretical framework for our study and explicate why it is particularly appropriate for studying the engagement of civil society actors in online public spheres. We propose two separate dimensions on which integration vs. fragmentation of online public spheres can be measured and observed: content synchronization and structural connectivity. In section 3, we clarify our main concepts (topics, issues, frames and issue networks) and explain how we will measure them. The following sections are devoted to describing the methods of web crawling and topic modeling (section 4) and explaining our data structure and processing (section 5). In section 6, we present our findings. We first give an overview on the identified ‘topics’ and the generated network, before we analyze more in depth how two selected ‘topics’ which we consider to be sub-issues (“contaminated food & regulation” and “GMO”) are discussed across the network. Based on these analyses, we conclude in section 7 with formulating a hypothesis on the mobilization potentials of the selected sub-issues. We also discuss limitations and future implications of the study.

2. Theoretical foundation

Networked public spheres on the Internet

A public sphere is an abstract concept, which manifests itself in varying settings such as encounters, assemblies, or the mass media (cf. Gerhards & Neidhardt, 1990). According to the majority of scholarly definitions a public sphere is constituted by non-private communication within an openly accessible (metaphorical) space. This space is a precondition for the circulation of information and the discourse about public affairs (Dahlgren, 2005, p.148; Pappacharissi, 2010, p.115). With regard to the political system the public sphere fulfills the function of an intermediate communication system between political decision makers and citizens. Therefore, there needs to be some structural preconditions of connectedness, interaction, and integration to make this intermediation between citizens and decision makers work. In an ongoing scholarly debate these preconditions are contested with regard to the advent of new media and the Internet: Does the Internet constitute a public sphere? Habermas (2008), according to his normative stance is rather skeptical about this question. He argues that the Internet lacks of integrative entities and structures that bundle, process and synthesize the multiplicity of information. Therefore the public loses orientation on common political issues which become ever more intransparent. In other words, the public sphere becomes fractured.

Other scholars are more reserved as there have always been multiple forms of segmentation and stratification of public spheres (Hepp & Wessler, 2009; Imhof, 2013; Kim, 2012; Kleinen-von Königslöw, 2012). There is wide agreement among scholars that "our social world is composed of multiple, overlapping, and unequal publics" (Breese, 2011, p. 132). Therefore, Breese (2011) argues that "[i]t is more accurate to talk of [...] *publics* and *public spheres* than to refer to *the public sphere*" (p. 132). However, the Internet further fuels this structural diversity. It offers more public spaces with easier access than ever before and more personalized communication environments (Bennett & Manheim, 2006; Pariser, 2011). These public spaces differ in their communication modes, institutional settings, size and reach: from small chatrooms or fora of individuals to large networks of organizational actors, who communicate about one issue on the Web (so-called issue networks; cf. Marres & Rogers, 2005).

Neuberger (2009) argues that these different public spaces on the Internet are far from totally falling apart. On the opposite, the ability of the Web to interconnect the varying modes and genres of communication result in an "integrated networked public sphere" (p. 41; referring to Benkler, 2006, p. 212 ff.). Furthermore, he attributes online journalism an important role to integrate public spheres on the Internet. Castells (2008) and Van Dijk (2006) put forward similar notions of a networked public sphere.

Until now there is still poor empirical knowledge about whether and how the new digitized media environment brings up fragmented or sufficiently integrated public spheres. What can be stated is that fragmentation and integration of public spheres are as multifaceted as public spheres themselves. In the following, we refer to fragmentation and integration as two poles of a continuum. While the mechanisms of fragmentation have been extensively discussed, we still need to specify criteria of integration. In recent studies two conditions are considered as conducive to integrate communication within a broader public sphere: (a) if the same issues are discussed simultaneously from similar perspectives in various sub-spheres

(Eder & Kantner, 2000) and (b) if there are communicative relations and interactions between speakers of these different sub-spheres (Adam, 2007). Kleinen-von Königslöw (2010, pp. 49-64) summarizes these criteria as (a) similarity and (b) interconnectedness of debates (see also Wessler et al., 2008 for a similar terminology).

Transferring these concepts to online public spheres leads us to distinguish two analytical dimensions on which fragmentation vs. integration can be observed and measured. First, an online public sphere would be fragmented if the communicative content of the actors' web sites diverge substantially in terms of the issues they talk about and the interpretation patterns or 'frames' that they attach to these issues. On the contrary, it would be integrated if different actors converge on their issues and interpretations. We call this dimension *content synchronization*.

Second, an online public sphere would be fragmented if communicators fail to connect to one another. We name this dimension *structural connectivity*. *Structural connectivity* may be achieved for example via hyperlinks between actors' web sites. Hyperlinks are essential structural elements of online communication (Park, 2003) that enable actors to associatively refer to the communication of another actor (Zimmermann, 2006). A densely interconnected network of web sites could then be regarded as a structurally integrated public sphere on the Internet, whereas poor interlinking indicates structural fragmentation.

Civil society and the public sphere

From the perspective of civil society actors, scholars emphasize that the networked character of the "new" emerging public sphere enables citizens and civil society actors to actively participate in political life (Dahlgren, 2005). Civil society actors, which were marginalized in the mass media's public sphere (Gamson & Wolfsfeld, 1993) can now gain visibility and strengthen the salience of their positions via online communication (Pfetsch, Adam & Bennett, 2013). In other words, new media effectively changes and improves the discursive opportunity structure for social movements and civil society actors (Cammaerts, 2012).

The engagement of civil society organizations and NGOs in the public sphere is crucial when we talk about a publics' deliberation because the publics' interests form up in such kinds of nonprofit and nongovernmental organizations like citizens' initiatives or social movement organizations. Habermas (2006) writes that "[a]ssociational networks of civil society (…) translate the strain of pending social problems and conflicting demands for social issues into political issues" (p. 417).

In order to successfully mobilize their cause, civil society organizations which can also be regarded as driving forces of social movements (Gerhards & Rucht, 1992), are interested in building coalitions with other actors, increase awareness for their issues and thereby raise commitment of a critical amount of citizens. An online network which is well integrated both in terms of structural connectivity and content synchronization is a critical factor for their success as it boosts their discursive opportunities and visibility (Koopmans, 2004). We therefore assume that a low degree of fragmentation of an online public sphere induced by civil society can be regarded as an indicator for the potential to mobilize attention and support for its issues and frames. For these reasons one should expect that civil society actors are strategically interested in and actively working for the integration of online public spheres (cf. Castells, 2008). Their efforts of network building should act as counter-force against fragmentation. Studying online public spheres from the angle of civil society communicators we therefore concentrate on the integration pole of the fragmentation/integration continuum.

Summarizing our theoretical arguments, we conceive of the Internet as a networked public sphere consisting of multiple and connected public spaces or spheres (Benkler, 2006; Neuberger, 2009). From this perspective fragmentation as well as integration are inherent and intertwined characteristics of online public spheres (Breese, 2011), which need to be analyzed empirically along two dimensions: content synchronization and structural connectivity. As mobilization actors, civil society organizations act as drivers towards integration of the networked public sphere.

3. Main theoretical concepts

This section aims to explicate the more concrete concepts on which we base our empirical investigation. In particular we (a) refer to *issue networks* as a manifest concept of a networked online public sphere and (b) provide a theoretically guided assistance for understanding the ‘topics’ that a topic model produces. We deem this reflection necessary as we explicitly understand the mobilization potential of a movement as a consequence of *content synchronization* (see section 2), i.e. the degree of convergence of different actors on their issues and attached interpretations. Therefore we aim to give directions how ‘topics’ can be classified within the given framework.

Issue networks

The concept of the issue network within the context of online communication has been coined by Marres and Rogers (2005). According to them, an issue network is constituted by a set of different actors who treat a specific common topic on their web sites in different possible ways (e.g., documents, slogans, campaigns, etc.). There exist connections between these actors and their web sites through hyperlinks, which cannot be interpreted in an unambiguous way, since they may be set between both actors with common or with antagonistic positions (Rogers, 2010). What is more, these connections cannot be interpreted as a real conversation between actors (Marres & Rogers, 2005). However, these links signify whom or whose positions other actors assume to be relevant for an issue and therefore actors, which do not receive a link are excluded from an issue network (Rogers, 2002). In this regard, issue networks render public the configurations of actors around a common issue and give them the possibility of political articulation of an issue (Marres, 2006).

In conclusion, issue networks match nicely with the theoretical concept of the networked public sphere in that they constitute a set of actors with at least topical relationship through hyperlinks, by which meaningful communication is transported (Marres & Rogers 2005). This view is also shared by Bennett, Lang and Segerberg (2014), who see issue networks as representing ‘issue publics’, respectively public spheres. Like them, we make use of an issue network to empirically examine a thematically centered public sphere.

Topics, issues and frames

To avoid confusion about the terms topic, issue and frame, we deem it necessary to clarify the concepts we refer to.

Using topic modeling as part of our methodological approach, we have to be aware of what the term ‘topic’ in ‘topic modeling’ actually means. Van Atteveldt and his colleges (2014) point out that many applications directly interpret a topic models’ result as ‘topics’ in the common sense of the word (p. 1) disregarding that ‘topic’ is actually just a word for a latent variable that captures the “abstract notion” of a topic (Blei, Ng & Jordan, 2003, p. 995). For political communication research this abstract notion has to be interpreted in theoretical terms (cf. van Atteveldt et al., 2014). The provision of a theoretically grounded guidance is the aim of the following paragraphs:

What are topics in general? Topics are a basic precondition for interactional human communication and can be regarded as general categories[1] that (a) help us structuring the complexity of reality and (b) serve as points of reference for meaningful communication (Luhmann, 1971, p.13; Eilders, 2008, p. 40). Eilders (2008) develops an integrated and multilayered concept for topics and frames (p. 40) [2] where topics can be seen as general category (1st layer) and 'frames' (2nd layer) can be considered to be topical attachments that employ a more specific interpretive perspective.

In the context of research on contentious politics the term 'issue' is more common than 'topic'. In contrast to topics, issues are contentious by definition "with individuals and groups taking opposing positions" (Miller & Riechert, 2001, p. 108).[3] Reconsidering the multilayered concept of Eilders (2008), the contentious character of an 'issue' denotes a basic conflictive perspective (or 'superordinate contentious frame') of a topic.

Miller (1998) and Miller and Riechert (2001) who are initiators of computer-based frame analysis (cf. Matthes & Kohring, 2008) [4] deliver an instructive definition of frames which they base on Entmans' frequently cited article about the "fractured paradigm of framing". Entman (1993) argues (a) that framing involves selecting and highlighting some aspects of reality while omitting others (p. 53) and more importantly (b) that frames can be explored by examining "the presence or absence of certain key words" (p. 53). Miller and Riechert (2001) add that (in the context of political communication) frames become "manifest in the choice and range of terms that provide the context in which issues are interpreted and discussed" (p. 109). Stating this, they acknowledge that "[o]f course, these key words are not themselves the frames. Rather, the words are indicative of perspectives, or points of view, by which issues and events can be discussed and interpreted" (Miller & Riechert, 2001, p. 114).

This leads us to the question what the 'topics' from a topic models basically represent, 'topics'/ 'issues' or 'frames'? At this point, we can only generally state, that this depends on the topical specificity of the corpus. A general corpus of newspaper articles leads to more

[1] Kepplinger (2001) argues that topics are complexes of meaning that are subject to individual perspectives (what the term 'meaning' already implicates) and contemporary culture. Coming from agenda-setting research Rogers and Dearing (1988) suggest that topics can be composed of single events.

[2] Eilders (2008) actually formulates a theoretically integrated concept for topics and opinions arguing that opinons can be attached to topics but not the other way around. By exploring the relationship of topics and opinions she argues that opinions can be understood as a composition of frames and positions. Opinions are formed within frames that are attached to a topic (p. 40). Although a frame employs a perspective on the communicative topic it does not determine a position within that frame (Eilders, 2008).

[3] Miller and Riechert (2001) write, "[w]here there is no argument there is no issue" (p. 108). Issues can thus be regarded as contentious topics (also cf. Cobb & Elder, 1983, p. 82).

[4] Miller (1998) and Miller and Richert (2001) detect frames from press releases and news media coverage with a methodological approach which they call "frame mapping"; it is based on the co-occurrences of words.

general ‘topics’ than a corpus that consist of discussion texts from a special-issue forum about for example the health risks of genetically modified foods.

As the reader will notice, our findings corroborate that there is no standard interpretation for a topic models’ output. Instead every model and probably even every modeled topic should be interpreted separately within a predefined theoretical framework. However, we argue that ‘topics’ should not be interpreted as frames because the meaning of an interpretive perspective cannot validly be reconstructed from what a topic model gives us as a result, i.e. the conditional probabilities of word-topic assignments which do not implicate the semantically coherent meaning of a frame. In our view, the identification of frames needs a detailed analysis by human coders.

4. Methods

In order to detect the issue structure of a hyperlink network we have to gather two datasets. First, network data, i.e. information about the interlinking structure of a set of web sites has to be collected. Second, we have to gather the contents of the web pages of the network in order to further analyze them.

Gathering hyperlink network data and web site contents

We used the web crawling tool *Issuecrawler*[5] to retrieve the network data. Crawling tools such as *Issuecrawler* take advantage of the network characteristic of the web in that they automatically collect and follow hyperlinks between web pages. We applied a *snowball* procedure which is probably the most intuitional, inclusive and un-restrictive method to capture the interlinking structure of an a priori unknown population of web sites (cf. Maier et al., 2014).

{Table 1 about here}

The reconstruction of our hyperlink network began with the definition of ‘seed’ pages, i.e. the starting points of the crawling procedure.[6] In case of a snowball crawling technique the crawling algorithm enters the seed pages’ URLs and follows every embedded hyperlink to

[5] For further information on the tool, please visit: http://www.govcom.org/index.html.

[6] These starting points have to be defined carefully, as they determine the overall structure of the resulting network. We conducted Google searches, did a literature review and gathered expert opinions in order to choose the starting points. For the list of Google search terms and seed URLs, see table 1. Since we wanted to collect an issue network, all the organizations behind the seed URLs deal with the issue of food safety at least amongst other issues.

the subsequent pages. After entering these referred pages into the crawling population, the procedure recursively repeats itself. The crawling tool archives the history of visited web pages and hyperlink paths. In order to prevent the network of becoming too big to analyze, we determined specific values for the two relevant parameters "crawling depth" and "degree of separation", which restrict the inclusion of web pages into the network. Crawling depth refers to the *vertical* dimension of the crawling process (see figure 1) and restricts the algorithm to follow internal links *within* a web site up to a certain depth. We set the crawling depth to the value of two. In contrast, the degree of separation refers to the *horizontal* crawling dimension and affects the hyperlinks *between* web sites. We set the degree of separation to the value of one which means that the maximal distance of any web site in the network to a seed-site is a hyperlink path of length one.

{Figure 1 about here}

The crawling procedure ultimately results in a list of (a) web pages including the information to which web site they belong to (a list of nodes) and (b) information about how they are interlinked with each other (a list of edges) (cf. Park, 2003). In other words, the crawls' result is a directed hyperlink network.[7]

Although the term "crawler" is sometimes used to describe software that systematically downloads data from the Internet (Thelwall, 2001) the *Issuecrawler* does not download any contents. Moreover, the *Issuecrawler* simply retrieves structural meta-data for the reconstruction of the hyperlink networks. In order to download and archive the content data of the networks' web sites[8] we use the open-source-tool *wget*[9]. The outcome of this process is in turn a big corpus of files from which we extract the html-files for further analysis.

Topic modeling

This corpus of html-files is the corpus on which we apply a topic modeling procedure. The html-files are read with an html-parser and the plain text within the body of each file

[7] What might be confusing at this point is that both the concept of the *issue network*s well as its empirical realization (a hyperlink network) gathered with the *Issuecrawler* explicitly refer to the term *issue* in their names. Unfortunately there is by no means a guarantee for the resulting networks to treat common issues. Instead, the crawling procedure solely relies on the described rules, without regarding contents.

[8] O'Neill et al. (2001) argue that we can differentiate between web pages and web sites. "A Web site is the collection of all Web pages located at the same top-level […] URL" (p. 281).

[9] Information available at: http://www.gnu.org/software/wget.

constitutes the exploited textual material[10] on which we base our analysis, i.e. the set of documents that constitutes the corpus. Our tools of choice for further analysis are topic models.

A topic model is a Bayesian hierarchical probabilistic (graphical) model. It defines an artificial generative process for document generation, describing how the actually observable data (the words in the documents), get into their place. In a topic model this is controlled by two latent factors, the topics and the documents' topic proportions. A topic is defined as a probability distribution over the word simplex, i.e., each word in every topic has a certain probability and the probabilities in each individual topic sum up to one. The set of words with the highest probability is assumed to describe the individual topics thematically.

The second factor, the document's topic proportions, is again a set of probability distributions (one for each document), respectively defined over the topic simplex. Every topic gets some probability in a document and the probabilities of topics for a single document sum up to one. Simply put, the individual words that we see in a document are generated by first finding a topic through the document's distribution over topics and then finding a word from the chosen topic. Both choices are random draws from their respective distributions. During learning, we reverse this process in order to get approximations for the governing latent factors that best give rise to the observed words, i.e., we want to find the setting of the latent factors for which the observed words are highly likely. In our particular case, we use a non-parametric model, the Hierarchical Dirichlet Process topic model (HDP), to circumvent the problem of choosing a number of topics a-priori (cf. Griffiths and Steyvers, 2004). The number of topics is inferred from the data, the only choice left is that of an abstract granularity parameter β, ($0 < \beta < 1$), influencing this number. The lower beta is, the higher the granularity and the more topics we expect. More technically, the granularity governs the sparsity of word probabilities in topics. Higher granularity leads to less words with high probability in each topic (i.e. greater sparsity) and thus to a higher number of topics that is needed to fully explain the seen data.[11]

[10] English stopwords were removed from the text corpus.

[11] To be even more precise, our granularity parameter is the hyperparameter to the base measure of the top-level Dirichlet Process in the HDP model (cf. Teh and Jordan, 2009). For a technical description of the HDP model see also Teh and Jordan (2009).

5. Data and data-processing

We applied our methods to a hyperlink network which we gathered in November 2014. As seed pages of the crawling procedure, we selected pages on the web sites that are run by civil society actors in the US who engage in the field of food safety (see table 1). The hyperlink network comprises 17.881 web pages which belong to 3.755 web sites. Among these 3.755 web sites exist 8.148 hyperlinks[12]. We were able to download and archive 11.845 of these web pages[13], i.e. 66 per cent of the complete corpus of web pages. The downloaded web pages belong to 2.211 web sites (59 per cent), with 5.886 (62 per cent) hyperlinks connecting them. Every subsequent analysis refers to this 'reduced' hyperlink network (2.211 web sites, 11.845 pages, 5.886 hyperlinks, see footnote 11).

Description of the topic models

Two topic models with different granularity parameters (β = 0.1 and β = 0.5) were calculated. The model with high granularity resulted in 167 topics whereas for low granularity, an optimal number of 56 topics were estimated.[14] The content coverage of topics is strongly asymmetric for both models: For the model with high granularity, 52 out of 167 topics account for more than 95 per cent of the modeled content; for low granularity, 30 out of 52 make up 95 per cent of the content. We consider the residual topics to be negligible (a) because each of them accounts for less than 1 percent of the overall modeled content and (b) because the research team found them to be hardly interpretable. This matches an assessment of Mimno et al. (2011) who found that the smallest topics are almost always of poor quality (p. 262). Hence, these topics were disregarded from the results.

As described above, the documents of the model originate from the texts of the html-files' body-sections. The method of extracting the text from the body section of an html-file is a common way to approximate the usable content from html-files with a previously unknown page structure (Günther & Scharkow, 2014). As our html-files were gathered regardless of the

[12] The number of hyperlinks between the web *pages* is much higher (48.242). After the aggregation from the web page-level to web site-level we accounted only for the presence (1) or absence (0) of (dichotomized) hyperlinks between *web sites*.

[13] Many web sites prohibit an automated access and subsequent download with crawling programs like *wget*.

[14] Considering the model assumption that every topic has a specific probability in a document and the probabilities of topics for a single document sum up to 1, the overall sum of a document-topic matrix is equal to the number of documents (which is in our case 11.845). A document (d_i) –topic (t_j) matrix is a matrix in which the documents of a corpus over which a topic model was calculated are located in the rows and their estimated topics are located in the columns. An entry in cell d_i - t_j for document i and topic j shows the calculated fraction/proportion of topic j in document i.). A column sum (which is in other words the sum of topic proportion/fractions over all documents) divided by the overall sum can thus be interpreted as the topic's fraction/proportion over the corpus.

authorship, genres or communicative modes[15], the web pages' structures can be assumed to vary significantly across the corpus. The extraction method is therefore prone to deviate from what we would manually select as usable textual content for the description of 'topics'. Hence, so called 'boilerplate-content' including web site navigation, linking lists and ads et cetera also becomes part of the topic model's underlying 'bag of words'. Fortunately, this does not lead to a contamination of otherwise well interpretable topics. Instead, some of the resulting topics contain word clusters that obviously represent this boilerplate content (see e.g. topics 2 or topic 27, see Annex 1 on page 33-36). Boilerplate-topics are common phenomena in topic models (Mimno & Blei, 2011) and make up a considerable fraction of more than 40 per cent of the content in each of the calculated models. Although boilerplate-topics have no substantive meaning, their emergence sharpens the other meaningful topics "by segregating boilerplate terms in a distinct location" (DiMaggio, Nag & Blei, 2013, p. 586). The emergence of boilerplate-topics can therefore be evaluated positively as it facilitates the other topics' interpretations.

For the interpretation of the topics we used lists of the topics' 30 most probable words.[16] For some topics we had no unambiguous interpretation and it was questionable if for example a word like "twitter" in the context of "consumers" most probably indicates a link to the social media platform or a consumer action campaign via twitter or both. For this reason we analyzed the topic's specific proportion in each of the documents attributed to the different sources across the network and looked at how strong a topic is concentrated on the ensemble of networked actors, i.e. how many actors actually contribute to a topic.[17] To quantify a topics' concentration we calculated the Hirschman-Herfindahl-Index (HHI) (cf. Hirschman, 1964) where a value of one indicates absolute concentration, i.e. the documents which contribute to a topic stem from only one web site, and a value equal to zero indicates no concentration at all, i.e. every actor contributes to the same amount to a topic. Measuring a topic's concentration over the network can be very useful for the evaluation of a topic. The topic list (see Annex 1, p. 33-36) shows the top 30 topics of the low granularity topic model including their HHI score. Almost every topic that we interpreted to be a boilerplate-topic features a high HHI-value (HHI > 0.2). Although a high or above average HHI-value is not

[15] While the starting pages of the crawl are mainly text based organizational web sites, the network also comprises other types of content, e.g. political fora, media sites, videoclip-platforms and social media sites.

[16] The here given interpretation of the topics' word lists is the result of a discussion in the research team.

[17] For this calculation the document-topic matrix had to be transformed to a web site-topic matrix. Therefore the html-documents originating from the same web sites were aggregated by adding up the topic probabilities of the according documents. That means the "document rows"/"web page rows" were merged to "web site rows".

necessarily a sign for a boilerplate-topic, it can be a useful indication that an attentive interpretation is appropriate.

We interpreted 15 out of 30 topics of the low granularity model, and 28 out of 56 topics from the high granularity model to be boilerplate-topics and disregarded them from the lists of the finally remaining topics in table 2 (15 topics for the low granularity parameter model) and table 3 (28 topics for the high granularity parameter model).

The reason for calculating two models with different granularity parameters was to check whether a finer 'resolution' (high granularity) results in a more appropriate model regarding our research interest targeting the synchronization of sub-issues in the network. As a higher granularity results in more topics, it was unclear, however, whether the high granularity model also brings up more diverse issue-related aspects compared to a model with a lower granularity. Jacobi, Welbers and van Atteveldt (2014) present an instructive method to demonstrate the blending of different models' topics by presenting the models as a bipartite-graph (or two-mode network).

Figure 2 shows such a bipartite graph where the similarity of the model's topics is based on the cosine distance of their topic-document allocation (cf. Jacobi, Welbers & van Atteveldt, 2014, p.13; Niekler & Jaehnichen, 2012). The graph nicely illustrates that the higher granularity translates into a more detailed depiction of the issue-attached 'perspectives'. Focusing for example on topic 9 (from the low granularity model, termed "pollution of drinking water, fracking & water privatization"), we can see that it decomposes into two topics (topic 2 and 48) in the high granularity model. Here, both topics still treat the same issue of 'fracking' but with differing perspectives (topic 2: "privatization of water" and topic 48: "energy, fracking, toxic chemicals & climate change"). Topic 17 is another example for this pattern: The low granularity model results in a topic that treats the outbreak of foodborne infections like salmonella and other infectious diseases. In the high granularity model this topic splits up into two respective topics, where one topic (15) almost solely features words associated with foodborne diseases, whereas the other topic (52) refers to infectious diseases and outbreaks different from foodborne ones.

As we can see from the comparative analysis of the models, a higher granularity translates into more specific topics with a greater level of detail. These topics can likely be interpreted as hybrid types of issues and frames, we call sub-issues. In contrast, the topics from our low granularity model feature higher disjunction, i.e. they can be interpreted more independently from each other.

In the following section we focus on the results from the low granularity model, because we are rather interested in *content synchronization* in terms of sub-issues than in terms of frames.

{Table 2 about here}

{Table 3 about here}

{Figure 2 about here}

Integrating a hyperlink network with topic model data

As a result of the past procedures we yield a vast collection of information about (a) the structural composition of the hyperlink network and (b) the topic composition of the underlying content from the html-files that we gathered. We integrated these two datasets to a single comprehensive one. Due to the above described post-processing of the topic data, we already gained an insight about which web sites (the network's nodes) contribute to which topics. We use this information to integrate the datasets as we define a web site's topic fraction as a network's nodal attribute. Technically the topics' fractions are simply added as columns to the network's nodelist.

6. Findings

The hyperlink network, representing an empirical realization of an online public sphere induced by civil and nongovernmental organizations, like described above, is composed of 2.211 web sites with 5.086 hyperlinks connecting them. The network is composed of a densely interlinked core cluster and periphery clouds which are solely connected to the center via their connection with the seed sites. The density of the graph is rather low due to the high number of nodes which feature an average degree of 2.3.[18]

[18] The 'density' which is a common measure for structural cohesion, can be calculated as the ratio of the realized ties divided by all possible ties and for the realized network results in the value 0.001. As the number of all possible ties increases potentially with the number of actors in a network, the density measure is not a valid measure for structural integration.

Reconsidering our first aim to assess the spectrum of sub-issues in the online sphere, we can conclude from our data that the hyperlink network contains a diverse range of 'topics', that can be considered sub-issues of a contemporary food safety debate among civil society actors in the US.

This debate incorporates sub-issues relating to sustainability in agriculture (topic 16) like the use of genetically modified seeds (topic 10) or the production of organic food and sustainable products as well as adverse effects of the industrial conditions of food production, like the usage of antibiotics in animal feed (topic 18). These adverse effects are, of course, associated with consequential problems like terrifying outbreaks of infectious disease by means of contaminated food, like salmonella (topic 7) or mercury in fish (topic 19). All these effects are threatening human health (topic 22) either because bacteria or viruses cause infections (topic 17) or because unhealthy ingredients (like too much sugar) in highly processed food cause obesity (topic 33). Civil society actors who back up their arguments with scientific evidence seem to claim for governmental regulation of the food industry (topic 25) educational approaches for solving the existing problems and public health research programs (topic 20).

Agricultural issues like the use of toxic chemicals and pesticides as well as the energy issue 'fracking' (topic 9), are not only argued to threat human health but also the natural environment (topic 3).

We have to admit at this point, that the model's representation of the topics alone does not allow for a valid interpretation of the inner connection of the 'topics'. Moreover, the preexisting experience of our research team with the food safety debate proved essential to make sense of the topic model's results.

We are now focusing on two 'topics' which we regard as key aspects of the food safety debate (a) "contaminated food & regulation" (topic 7) and (b) "genetically modified seeds" (topic 10). As the present approach is rather explorative we prefer a visual-interpretive analysis in favor of generating testable hypotheses, as the main aim of this procedure. We align our analysis alongside the two theoretically deployed dimensions of the fragmentation of the public sphere, content synchronization and structural connectivity. Thereby, we rather focus on the content synchronization dimension, whereas the structural connectivity dimension is regarded more implicit. The connectivity dimension is reflected in the visualizations which in turn are based on the networks' relational dataset.

Contaminated food & regulation (topic 7)

Concentrating on the top words of the topic (see figure 4) leads immediately to the conclusion that the topic captures first and foremost the safety of meat products (*food, safety, products, meat, beef, animal, chicken*). The words *salmonella, coli, bacteria, outbreak* denote the pathogenic agents of infectious foodborne disease. All of these terms imply *health* risks for consumers. In combination with the terms *government, standards, USDA* and *industry* we can derive that most important regulatory actors like the US Department for Agriculture already play a decisive role in regulating food industry standards or that they at least are called to take such regulation.

"Contaminated food & regulation" is the third most prevalent modeled topic with a calculated fraction of around 4.3 per cent[19] (and an approximated corrected fraction of 8.3 per cent). The topic's Herfindahl-Index (0.041) is located close to the middle of a ranking list of all topics' HHIs, indicating that the report of topics is relatively equally distributed among the networks' actors.

Figure 4 depicts the network graph, with bigger nodes indicating a bigger amount of html-files contributing to the corpus and thus to the topic's results. The more intense the reporting of a web site on the sub-issue is, the more intense is the red color of a web site's node. What we can see is that a few web sites, particularly the starting points of the crawl, are the biggest nodes, contributing a larger amount of documents to the corpus than most of the other web sites of the network. The staring points of the crawl also contribute a considerable amount to the focused sub-issue, with four of them ranking on positions 3 to 10 of the top twenty list of contributors. Except from these actors which we ascribe an exceptional position as the leading civil society organizations in the field of food safety, governmental regulatory entities, such as the Food and Drug Administration (FDA), the Center for Disease Control (CDC) and their joint information platform (foodsafety.gov) are the most extensive contributors to topic 7. Regarding the fact that only a very small fraction of web sites (3.6 per cent, see Illustration 3) are run by governmental entities, this is noteworthy. Last but not least, the web sites of leading news media organizations like the New York Times, the Washington Post are actively reporting about the issue.

[19] Regarding the fact that the sum of the interpretable topics of the model account for 51.7 per cent of all contents, we calculated an approximate corrected topic fraction by dividing the original topic fraction (4.3%) by the total cumulative fraction of the interpretable topics (51.7%).

We can see that all of these actors enjoy well integrated network positions,[20] leading to the conclusion that the topic is not only literally rather located in the center of the public sphere. Moreover, it is a substantial issue that state entities apparently perceive as their duty. In addition the issue is also a moderately vivid part of the civil food safety debate.

{Figure 4 about here}

Genetically modified food (topic 10)

Topic 10 (see figure 5) is about the use of *genetically modified/engineered seeds* in *agriculture*, the *labeling of GM-foods*, like *corn* and potential adverse effects for flora and fauna (*bees*, *plants*, *environment*). The topic also features the predominant actors of the sub-issue, i.e. the industrial enterprise *Monsanto* which is the market's leader for seeds in agriculture, the *farmers* who are the potential users of *biotech*nology and the *USDA* as the most important governmental body in GMO-politics.

Genetically modified food is the 5th most prevalent sub-issue in the US food safety network, with a topic fraction of 3.5 per cent (corrected fraction: 6.8 per cent). The topic's Herfindahl-Index (0.022) features the 2nd/3rd lowest rank denoting a close to equal contribution of the network's actors. This also becomes visually apparent in figure 5 where we can see that the topic is more equally distributed among the actors in the center of the network, although its overall topic fraction is smaller than that of topic 7. Focusing on the peripheral areas of the network we can state that the debate about GMOs is not restricted to the center core of the network but reaches beyond to other, less well integrated areas of the network which can be located on the upper left side (connected via greenpeace.int.org).

The actor ensemble that mainly maintains the GMO-debate (see top contributing sites in figure 5) can be described as a compound of civil society organizations that are either specialized on the issue of GMOs (gmwatch.org, righttoknow-gmo.org,gmofreect.org), or agricultural affairs and food safety (cornucopia.org, ewg.org, centerforfoodsafety.org, foodandwaterwatch.org). These actors are intertwined with both the more general news media organizations (like nytimes.com, huffingtonpost.com) and specialized media sites (like civileats.com or motherjones.com). Interesting enough, we can observe that the central cluster

[20] The visualization algorithm (Yifan Hu), which is based on Force-algorithms, places mutually connected actors closer to another, which results in a densely interlinked 'core' of the network.

of governmental entities, in contrast to topic 7, rarely engages in the ongoing debate. Instead, they are rather silencing this sub-issue.

The data reveals that the sub-issue of genetically modified seeds and foods is by no means marginalized in the online network. Instead, we can conclude that the sub-issue itself not only brought up specialized organizations but also shares broad attention in the connected civil society community. There is a discernible tendency that the sub-issue is also apparent in different periphery areas of the graph which in comparison to topic 7 indicates a greater widespread.

{Illustration 4 about here}

Summary of the findings

How can we interpret the topic models' 'topics'? The resulting 'topics' of the topic model can be considered sub-issues of the rather vague field of 'food safety' because they suggest more specific matters of public and political dispute. These sub-issues include words that indicate contentious actor configurations as well as explanatory statements and/or treatment recommendation. This obtrusively points to Entmans' definition of frames (1993). But we still would not call these 'topics' frames as the conditional probabilities of the word-topic assignments cannot be regarded as valid indicators for the semantic coherence of the words, like for example the existence of boilerplate-topics demonstrates. Instead, the researchers' choice of a thematic compound of communicative points of reference, i.e. what we colloquially understand as a topic, is the only guidline for the interpretation of the topic models' 'topics'. In our case this is the intentionally selected topic of 'food safety'. In other words, we cannot validly make conclusions about the semantic dimension of the sub-issue just by focusing the word lists of the topic models' 'topics'. Therefore, we would not consider them 'frames' but 'sub-issues'.

Both of the here focused sub-issues feature a high prevalence indicating that there is a vivid debate going on in the civil society-induced online public sphere. But the topical debates are located in differing clusters of the network. While the "food contamination & regulation"-topic incorporates many regulatory entities, these actors are far less well integrated in the GMO debate, although GMO generally features a greater dispersion (lower HHI). In the "food contamination"-debate, on the contrary, there is only moderate participation of more general civil society organizations[21] (lower level of content synchronization dimension), while the GMO-topic unifies a more heterogeneous group of civil actors, excluding administrative entities.

Hence, from the perspective of the content synchronization-dimension we can conclude that “contaminated food” is a less well integrated sub-issue among the diversely orientated civil actors, compared to GMO. Furthermore, one could argue that the state obviously already claims to take responsibility in “food contamination”, which makes civil mobilization less pressing.

In contrast, our data indicates a lack of the state’s engagement regarding the GMO-issue which at the same time can be considered a more widespread sub-issue among the civil organizations. This is probably due to GMO’s multiple connections to the sustainability debate as well as to the food safety debate. Together these indications point to a stronger mobilization potential of the GMO-sub-issue.

Although the public sphere can be considered equally integrated/fragmented from a structural point of view our exploratory analysis of the content synchronization dimension consequently leads us to the hypothesis, that the mobilizing potential is much lower for “food contamination” than for “genetically modified food”.

7. Conclusion & Discussion

Summing up, the combined approach of hyperlink-network analysis and topic modeling can lead researchers to an enlightening insight about the topic structure of an online debate in general. Putting it in the style of Laswell’s model of communication (1948) we are able to reveal “who” reports about “what” to “what extent” referring to “whom” on large scale excerpt of the Internet.

Topic modeling proved to be a very valuable contribution to study the content-structure of the issue network. Both of the calculated models provided reasonable, well interpretable results which could easily be combined with the network data.

The approach is auspicious, but one also has to be aware of its’ pitfalls. It remains unclear, for example, how a topic models’ ‘topic’ or the list of its top-words, respectively, should be interpreted. A general recommendation for the interpretation cannot be given. Therefore, we construed a theoretically guided, double-layered concept for the classification

[21] Foremost multiple sub-issue organizations.

of the ‘topics’. The concept differentiates topics/issues (1st layer) and frames (2nd layer). For the above mentioned reasons, we locate the here produced topic models’ resulting ‘topics’ in between the 1st and the 2nd layer of the concept and consider them to be sub-issues (also see van Atteveldt et al., 2014).

In order to derive meaningful conclusions from such research a theoretically guided decision about the topic models’ input data set is crucial. As our input data comes from a hyperlink-network, we have to carefully think about every step in the reconstruction of that network. A random as well as an arbitrary choice of the crawls’ starting points, for example, would have possibly led to poor a representation of what we wanted to depict: A civil food safety network.

Our exploratory study aimed at the generation of hypothesis about the mobilization potential of specific (sub-)issues among the actors of that network that we regard to as a public sphere. We therefore examined *content synchronization* and *structural connectivity* to assess the mobilization potential of the network. While the structural connectivity dimension was for considered equal for both sub-issues, we found differing levels of content synchronization for two substantial aspects of the food safety debate. The empirical data suggests the following conclusion: Although both sub-issues are highly prevalent in the network, “genetically modified food” probably features a higher mobilization potential because it is a more widespread sub-issue among a more heterogeneous array of civil actors, in comparison to the sub-issue of “contaminated food”.

Our study has some limitations, of course. We have to admit, that a mere visual analysis is probably a relatively poor method for the simultaneous examination of content and structure. Further analyses should definitely make use of inferential network methods, like Exponential Random Graph Models. However, the visual analysis seemed to be a useful access to describe a topics distribution over the network and sufficient for the exploratory purpose of the generation of a hypotheses.

Another limitation certainly is the lack of validation for the topic models’ results which will be an upcoming challenge for following analyses, where extracted ‘topics’ could be validated against data from manual content analysis.

References

Adam, S. (2007). *Symbolische Netzwerke in Europa: Der Einfluss der nationalen Ebene auf europäische Öffentlichkeit. Deutschland und Frankreich im Vergleich*. Köln: Herbert von Halem.

Ali-Hasan, N. F. & Adamic, L. A. (2007). *Expressing Social relationships on the Blog through Links and Comments*. Retrieved from http://www.icwsm.org/papers/2--Ali-Hasan--Adamic.pdf, 2015/01/03.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven: Yale University Press.

Bennett, L. W., Lang, S., & Segerberg, A. (2014). European issue publics online: The cases of climate change and fair trade. In T. Risse (Ed.), *European Public Spheres. Politics Is Back* (pp. 108–137). Cambridge: Cambridge University Press.

Bennett, W. L., & Manheim, J. B. (2006). The One-Step Flow of Communication. *The ANNALS of the American Academy of Political and Social Science*, *608*(1), 213–232. doi:10.1177/0002716206292266

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*(2), 219–245. doi:10.1111/j.1460-2466.2011.01538.x

Blei, D. A. (2012). Topic Modeling and Digital Humanities. *Journal of Digital Humanities*. Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/, 2015/01/03.

Blei, D. A., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Mashine Learning Research*, *3*(1), 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944937, 2015/01/03.

Breese, E. B. (2011). Mapping the Variety of Public Spheres. *Communication Theory*, *21*(2), 130–149. doi:10.1111/j.1468-2885.2011.01379.x

Cammaerts, B. (2012). Protest logics and the mediation opportunity structure. *European Journal of Communication*, *27*(2), 117–134. doi:10.1177/0267323112441007

Carpenter, C. R., & Jose, B. (2012). Transnational issue networks in real and virtual space: the case of women, peace and security. *Global networks*, *12*(4), 525–543.

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*(1), 78–93. doi:10.1177/0002716207311877

Cobb, R. W., & Elder, C. D. (1983). *Participation in American politics: The dynamics of agenda-building*. Baltimore: Johns Hopkins University Press.

Dahlgren, P. (2005). The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation. *Political Communication*, *22*(2), 147–162. doi:10.1080/10584600590933160

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*(6), 570–606. doi:10.1016/j.poetic.2013.08.004

Eder, K., & Kantner, C. (2000). Transnationale Resonanzstrukturen in Europa. Eine Kritik der Rede vom Öffentlichkeitsdefizit. *Kölner Zeitschrift für Soziologie und Sozialpsychologie Sonderheft 40*, 306–331.

Eilders, C. (2008). Massenmedien als Produzenten öffentlicher Meinungen- Pressekommentare als Manifestation der politischen Akteursrolle. In S. Adam & B. Pfetsch (Eds.), *Massenmedien als politische Akteure. Konzepte und Analysen* (pp. 27–51). Wiesbaden: VS.

Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, *43*(4), 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x

Faust, K., & Wasserman, S. (1992). Centrality and Prestige: A Review of Synthesis. *Journal of Quantitative Antropology*, *4*(1), 23–78.

Gamson, W. A., & Wolfsfeld, G. (1993). Movements and media as interacting systems. *The ANNALS of the American Academy of Political and Social Science, 528*(1), 114–125.

Gerhards, J. & Neidhardt, F. (1990). *Strukturen und Funktionen moderner Öffentlichkeit - Fragestellungen und Ansätze*. Retrieved from http://www.polsoz.fu-berlin.de/soziologie/arbeitsbereiche/makrosoziologie/mitarbeiter/lehrstuhlinhaber/dateien/GerhardsNeidhardt-1990.pdf?1367713012, 2015/01/01.

Gerhards, J., & Rucht, D. (1992). Mesomobilization: Organizing and Framing in two Protest Campaigns in West Germany. *American Journal of Sociology, 98*(3), 555–595.

Gibson, R. K., Gillan, K., Greffet, F., Lee, B. J., & Ward, S. (2013). Party organizational change and ICTs: The growth of a virtual grassroots? *New Media & Society, 15*(1), 31–51. doi:10.1177/1461444812457329

Griffith, T. L., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences, 101*, 5228–5235.

Günther, E., & Scharkow, M. (2014). Automatisierte Datenbereinigung bei Inhalts- und Linkanalysen. In K. Sommer, M. Wettstein, W. Wirth, & J. Matthes (Eds.), *Methoden und Forschungslogik der Kommunikationswissenschaft: Vol. 11. Automatisierung in der Inhaltsanalyse* (pp. 111–126). Köln: von Halem.

Habermas, J. (2006). Political Communication in Media Society: Does Democracy Still Enjoy an Epistemic Dimension? The Impact of Normative Theory on Empirical Research. *Communication Theory, 16*(4), 411–426. doi:10.1111/j.1468-2885.2006.00280.x

Habermas, J. (2008). *Ach, Europa: Kleine politische Schriften XI.* Berlin: Suhrkamp.

Hepp, A., & Wessler, H. (2009). Politische Diskurskulturen - Überlegungen zur empirischen Erklärung segmentierter europäischer Öffentlichkeit. *M&K*, 174–197. doi:10.5771/1615-634x-2009-2-174

Hirschman, A. O. (1964). The Paternity of an Index. *The American Economic Review, 54*(5).

Hyun Kim, J. (2012). A Hyperlink and Semantic Network Analysis of the Triple Helix (University-Government-Industry): The Interorganizational Communication Structure of Nanotechnology. *Journal of Computer-Mediated Communication, 17*(2), 152–170. doi:10.1111/j.1083-6101.2011.01564.x

Imhof, K., Blum, R., Bonfadelli, H., & Jarren, O. (2013). *Stratifizierte und segmentierte Öffentlichkeit*. Wiesbaden: Springer.

Jacobi, C., Welbers, K., & van Atteveldt, W. (2014). *Quantitative Analysis of large amounts of Journalistic Text using Topic Modeling: Conference paper presented at "Political Context Matters: Content Analysis in the Social Sciences", 10th - 11th of October, Mannheim, Germany.*

Kepplinger, H. M. (2001). Der Ereignisbegriff in der Publizistikwissenschaft. *Publizistik, 46*(2), 117–139. doi:10.1007/s11616-001-0032-3

Kim, Y. M. (2012). The Shifting Sands of Citizenship: Toward a Model of the Citizenry in Life Politics. *The ANNALS of the American Academy of Political and Social Science, 644*(1), 147–158. doi:10.1177/0002716212456008

Kleinen- von Konigslöw, K. (2012). Europe in crisis? Testing the stability and explanatory factors of the Europeanization of national public spheres. *International Communication Gazette, 74*(5), 443–463. doi:10.1177/1748048512445153

Kleinen-von Königslöw, K. (2010). *Die Arenen-Integration nationaler Öffentlichkeiten: Der Fall der wiedervereinten deutschen Öffentlichkeit*. Wiesbaden: VS.

Koopmans, R. (2004). Movements and media: Selection processes and evolutionary dynamics in the public sphere. *Theory and Society, 33*(3/4), 367–391. doi:10.1023/B:RYSO.0000038603.34963.de

Laswell, H. D. (1948). The Structure and Function of Communication in Society. In L. Bryson (Ed.), *The Communication of Ideas. A Series of Adresses* (pp. 32–51). New York: Harper & Brs.

Luhmann, N. (1971). Öffentliche Meinung. In N. Luhmann (Ed.), *Politische Planung;. Aufsätze zur Soziologie von Politik und Verwaltung*. Opladen: Westdeutscher Verlag.

Maier, D., Waldherr, A., Miltner, P., Schmid-Petri, H., Häussler, T., & Adam, S. (2014). Stichprobenziehung aus dem Netz - Wie man themenspezifische Online-Inhalte erfassen kann. In K. Sommer, M. Wettstein, W. Wirth, & J. Matthes (Eds.), *Methoden und Forschungslogik der Kommunikationswissenschaft: Vol. 11. Automatisierung in der Inhaltsanalyse* (pp. 90–110). Köln: von Halem.

Marres, N. (2006). Net-Work Is Format Work: Issue Networks and the Sites of Civil Society Politics. In J. Dean, W. J. Anderson, & G. Lovink (Eds.), *Reforming Politics: Networked Communications and Global Civil Society* (pp. 3–18). London: Routledge.

Marres, N., & Rogers, R. (op. 2005). Recipe for Tracing the Fate of Issues and their Public on the Web. In P. Weibel & B. Latour (Eds.), *Making things public. Atmospheres of democracy, ZKM, Center for art and media Karlsruhe, 20.03.-03-10.2005* (pp. 922–935). Cambridge: (Mass.); MIT press.

Matthes, J., & Kohring, M. (2008). The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication, 58*(2), 258–279. doi:10.1111/j.1460-2466.2008.00384.x

Miller, M. M. (1997). Frame Mapping and Analysis of News Coverage of Contentious Issues. *Social Science Computer Review, 15*(4), 367–378. doi:10.1177/089443939701500403

Miller, M. M., Riechert, & B. P. (2001). The Spiral of Opportunity and Frame resonance: Mapping the Issue Cycle in News and Public Discourse. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *LEA's communication series. Framing public life. Perspectives on media and our understanding of the social world* (pp. 106–122). Mahwah, N.J: Lawrence Erlbaum Associates.

Mimno, D., & Blei, D. A. (2011). Baysian checking for topic models. *Proceeding EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 227–237.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 262–272.

Neuberger, C. (2009). Internet, Journalismus und Öffentlichkeit - Analyse des Medienumbruchs. In C. Neuberger, C. Nuernbergk, & M. Rischke (Eds.), *Journalismus im Internet. Profession, Partizipation, Technisierung* (1st ed., pp. 19–105). Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage, Wiesbaden.

Niekler, A., & Jaehnichen, P. (2012). Matching Results of Latent Dirichlet Allocation for Text. *Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling,* 317–322.

O'Neill, E. T., McClain, P. D., & Lavoie, B. F. (2001). A Methodology for Sampling the World Wide Web. *Journal of Library Administration, 34*(3-4), 279–291. doi:10.1300/J111v34n03_07

Papacharissi, Z. (2010). *A private sphere: Democracy in a digital age*. Cambridge, UK, Malden, MA: Polity.

Pariser, E. (2011). *The filter bubble: What the Internet ist hiding from you*. New York, NY: Penguin Press.

Park, H. W. (2003). Hyperlink network Analysis: A new Method for the Study of Social Structure on the Web. *Connections, 25*(1), 49–61.

Pfetsch, B., Adam, S., & Bennett, L. W. (2013). The critical linkage between online and offline media. an approach to researching the conditions of issue spill-over. *Javnost - the public, 2013*(3), 9–22.

Rogers, E. M., & Dearing, J. W. (1988). Agenda-Setting research: Where Has It Been, Where Is It Going? In J. A. Anderson (Ed.), *Communication yearbook. 11* (pp. 555–594). Newbury Park, London, New-Delhi: Sage Publications.

Rogers, R. (2002). Operating Issue Networks On The Web. *Science as Culture, 11*(2), 191–213. doi:10.1080/09505430220137243

Rogers, R. (2008). Mapping Public Web Space with the Issuecrawler. In C. Brossard & B. Reber (Eds.), *Digital cognitive technologies. Epistemology and knowledge society* (pp. 115–126). London: ISTE.

Rogers, R. (2012). Mapping and the Politics of Web Space. *Theory, Culture & Society*, *29*(4/5), 193–219. doi:10.1177/0263276412450926

Tateo, L. (2005). The Italian Extreme Right On-line Network: An Exploratory Study Using an Integrated Social Network Analysis and Content Analysis Approach. *Journal of Computer-Mediated Communication*, *10*(2), 0. doi:10.1111/j.1083-6101.2005.tb00247.x

Teh, Y. W., & Jordan (2009), M. I. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics*, (158).

Thelwall, M. (2001). A web crawler design for data mining. *Journal of Information Science*, *27*(5), 319–325. doi:10.1177/016555150102700503

Tremayne, M., Zheng, N., Lee, J. K., & Jeong, J. (2006). Issue Publics on the Web: Applying Network Theory to the War Blogosphere. *Journal of Computer-Mediated Communication*, *12*(1), 290–310. doi:10.1111/j.1083-6101.2006.00326.x

van Atteveldt, W., Welbers, K., Jacobi, C., & Vliegenthart, R. (2014). *LDA models topics ... But what are 'topics'?* Retrieved from http://vanatteveldt.com/wp-content/uploads/2014_vanatteveldt_glasgowbigdata_topics.pdf, 2015/01/02.

van Dijk, J. (2006). *The network society: Social aspects of new media*. London, Thousand Oaks, New Delhi: Sage.

Wessler, H. (2008). Investigating Deliberativeness Comparatively. *Political Communication*, *25*(1), 1–22. doi:10.1080/10584600701807752

Wessler, H., Brüggemann, M., Kleinen-von Königslöw, K., & Sifft, S. (2008). *Transnationalization of Public Spheres*. Basingstoke: Palgrave Macmillan.

Williams, A. P., Trammell, K. D., Postelnicu, M., Landreville, K. D., & Martin, J. D. (2005). Blogging and Hyperlinking: use of the Web to enhance viability during the 2004 US campaign. *Journalism Studies*, *6*(2), 177–186. doi:10.1080/14616700500057262

Zimmermann, A. C. (2006). *Demokratisierung und Europäisierung online? Massenmediale politische Öffentlichkeit im Internet*. Retrieved from http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000003532, 2014/12/05.

Table 1: *Starting URLs (source seeds) and Google search terms for their identification*

search terms	starting URLs (source seeds)
Food safety, safe + food , food scandal, GM foods, food + consumer protection, food + consumers, food + risk, food safety + campaign, food + labelling, food safety + control	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

Table 2: *Interpretable topics of the low granularity parameter topic model (β=0.5)*

Topic	Topics' share %	HHI	Description
Topic 20	9,009%	0,021	public health research & education
Topic 25	4,745%	0,011	governmental regulation of the industry
Topic 7	4,296%	0,041	contaminated food & regulation
Topic 18	4,136%	0,342	antibiotics & consumer action
Topic 16	3,547%	0,044	organic farming & sustainable agriculture
Topic 10	3,535%	0,022	genetically modified foods & seeds
Topic 5	3,535%	0,015	research in general
Topic 3	3,273%	0,022	climate change, renewable energy & the environment
Topic 17	2,982%	0,122	health & infectious, foodborne disease
Topic 9	2,897%	0,331	pollution of drinking water, fracking & water privatization
Topic 22	2,657%	0,062	health & disease
Topic 33	2,261%	0,048	obesity & sugary ingredients
Topic 19	1,902%	0,217	mercury in fish
Topic 36	1,827%	0,102	animal & dairy products
Topic 14	1,080%	0,225	organic clothing

Table 3: *Interpretable topics of the high granularity parameter topic model (β= 0.1)*

Topic	Topics' share %	HHI	Description
Topic 24	6,165%	0,004	public health research & education
Topic 18	2,797%	0,436	consumer action & antibiotics in animal
Topic 34	2,715%	0,070	contaminated food & outbreak of foodborne disease
Topic 29	2,695%	0,030	state regulation of food safety
Topic 11	2,633%	0,231	clinical & microbiological reports
Topic 43	2,561%	0,026	political regulation & law
Topic 47	2,297%	0,018	genetically modified foods and seeds
Topic 1	2,264%	0,013	health research
Topic 27	2,231%	0,100	local farming & nutrition programs
Topic 0	2,207%	0,017	healthy diet
Topic 2	2,198%	0,420	water privatization
Topic 3	1,953%	0,088	general health discourse
Topic 25	1,928%	0,011	general economic discourse about globalization and jobs
Topic 22	1,756%	0,058	climate change & energy
Topic 15	1,585%	0,286	foodborne disease & infections
Topic 30	1,575%	0,034	industrial meat production & antibiotics
Topic 52	1,523%	0,142	non-food-related infectious diseases
Topic 68	1,388%	0,040	justice movement & crisis politics
Topic 39	1,320%	0,036	beverage marketing & coca cola
Topic 56	1,262%	0,013	environmental pollution & toxic chemicals in agriculture
Topic 48	1,081%	0,066	energy, fracking, toxic chemicals & climate change
Topic 21	1,071%	0,384	organic clothing
Topic 9	0,972%	0,949	sugar industry & obesity
Topic 46	0,870%	0,025	pesticides, toxic chemicals & farming
Topic 26	0,706%	0,024	mercury in fish
Topic 73	0,673%	0,631	home produced fruits & vegetables
Topic 6	0,578%	0,284	agricultural regulation & guidance on organic farming
Topic 54	0,506%	0,486	fair trade coffee

Figure 1: *Logic of the crawling procedure*

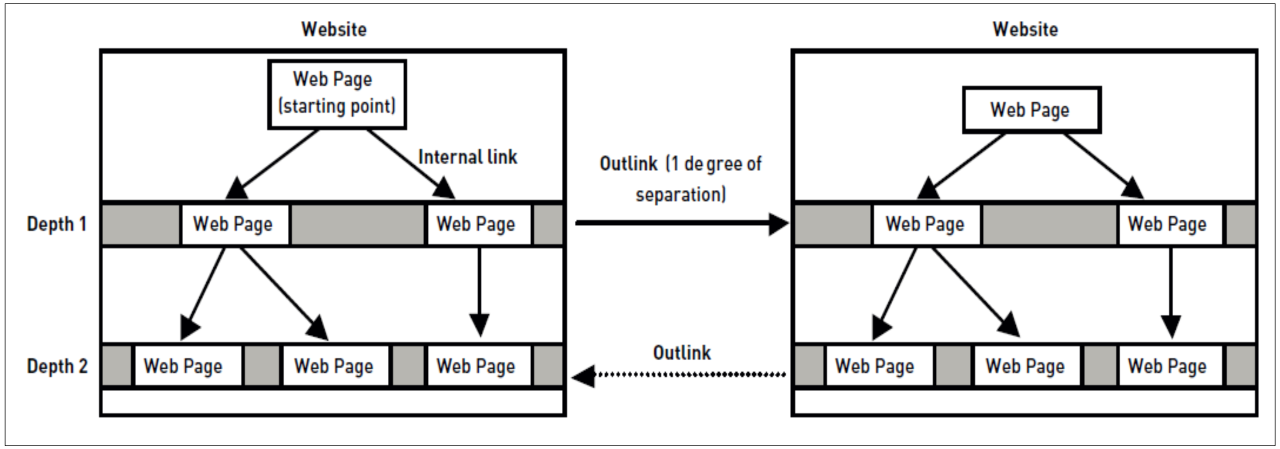


Figure 2: *Hierarchical blending of the two calculated topic models*

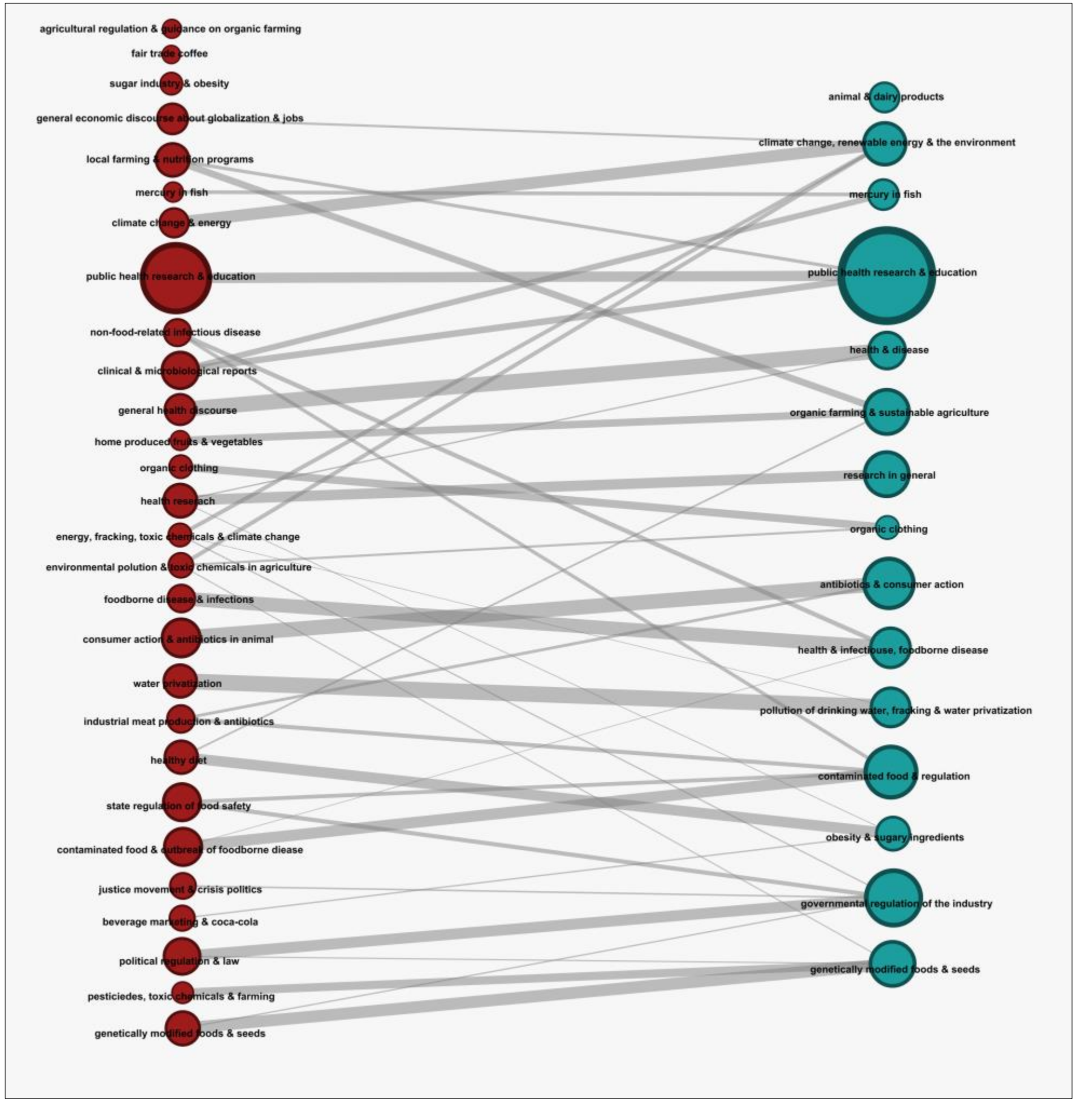
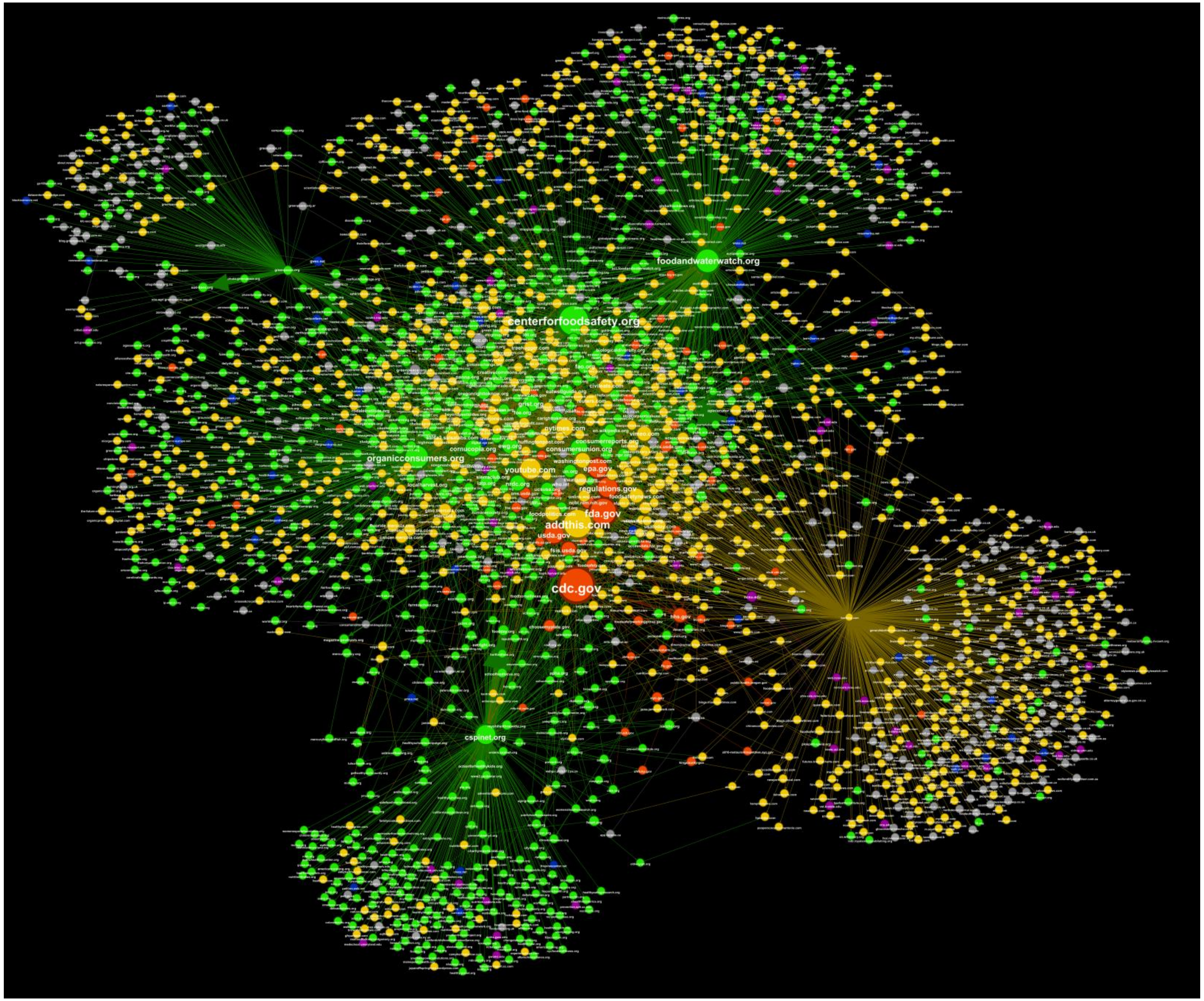

agricultural regulation & guidance on organic farming
fair trade coffee
sugar industry & obesity
general economic discourse about globalization & jobs
local farming & nutrition programs
mercury in fish
climate change & energy
public health research & education
non-food-related infectious disease
clinical & microbiological reports
general health discourse
home produced fruits & vegetables
organic clothing
health reseach
energy, fracking, toxic chemicals & climate change
environmental polution & toxic chemicals in agriculture
foodborne disease & infections
consumer action & antibiotics in animal
water privatization
industrial meat production & antibiotics
healthy diet
state regulation of food safety
contaminated food & outbreak of foodborne diease
justice movement & crisis politics
beverage marketing & coca-cola
political regulation & law
pesticiedes, toxic chemicals & farming
genetically modified foods & seeds
animal & dairy products
climate change, renewable energy & the environment
mercury in fish
public health research & education
health & disease
organic farming & sustainable agriculture
research in general
organic clothing
antibiotics & consumer action
health & infectiouse, foodborne disease
pollution of drinking water, fracking & water privatization
contaminated food & regulation
obesity & sugary ingredients
governmental regulation of the industry
genetically modified foods & seeds

Figure 2 depicts a bipartite graph, where the topics of the high granularity model (red nodes) are placed on the left side and the topics of the low granularity model are placed on the right side (blue nodes). The bigger the nodes are, the higher is the respective model fraction of topic. The existence of a connection among a red and a blue node indicates whether the cosine similarity of the topic-document allocation vectors is >0.175. Lower similarities were disregarded in favor of clarity. The thicker the ties are, the greater is the similarity value among the two topic-vectors. (cf. Jacobi, Welbers & van Atteveldt, 2014).

Figure 3: *Hyperlink network induced by civil society actors in the USA (8 starting points), November 2014*



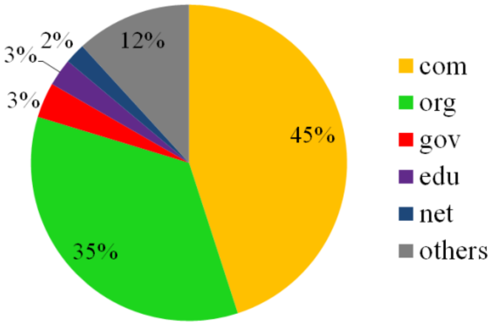
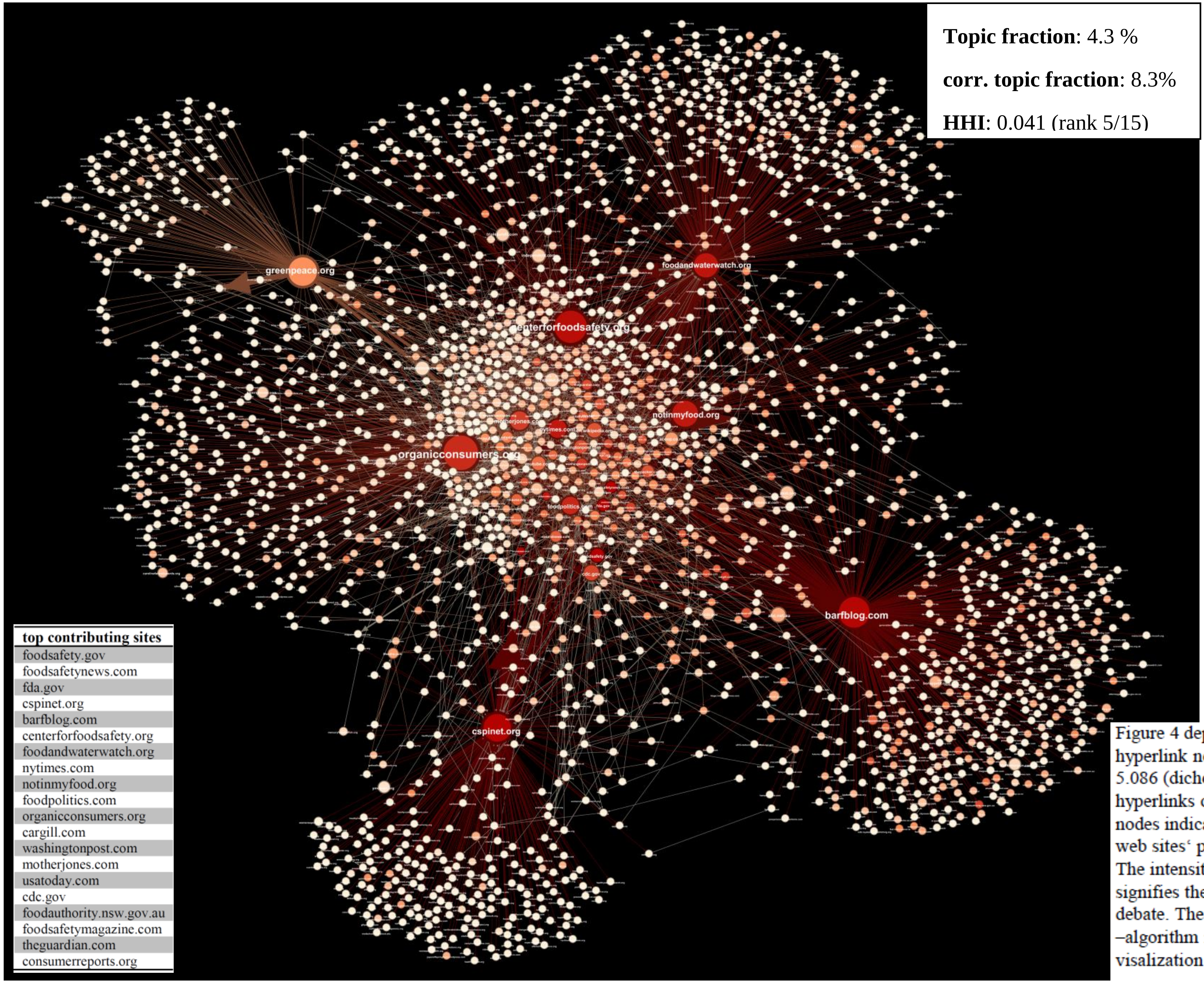


Figure 3 depicts the (directed) graph of the hyperlink network among 2.211 web sites with 5.086 (dichotomized, cf. footnote 11, p. 12) hyperlinks connecting them. The size of the nodes indicates the magnitude of their indegree. The indegree is a common measure for a node's „prestige" (cf. Faust & Wasserman, 1992). The color of the node signifies the node's top-level-domain. The color of the ties refers to the top-level-domain of the tie-/link-sender. The graph was drawn using the *Yifan Hu* – algorithm implemented in the open source graph visalization program *Gephi*.

Figure 4: *'Heat map' of topic 7"contaminated food & regulation"*



Topic fraction: 4.3 %

corr. topic fraction: 8.3%

HHI: 0.041 (rank 5/15)

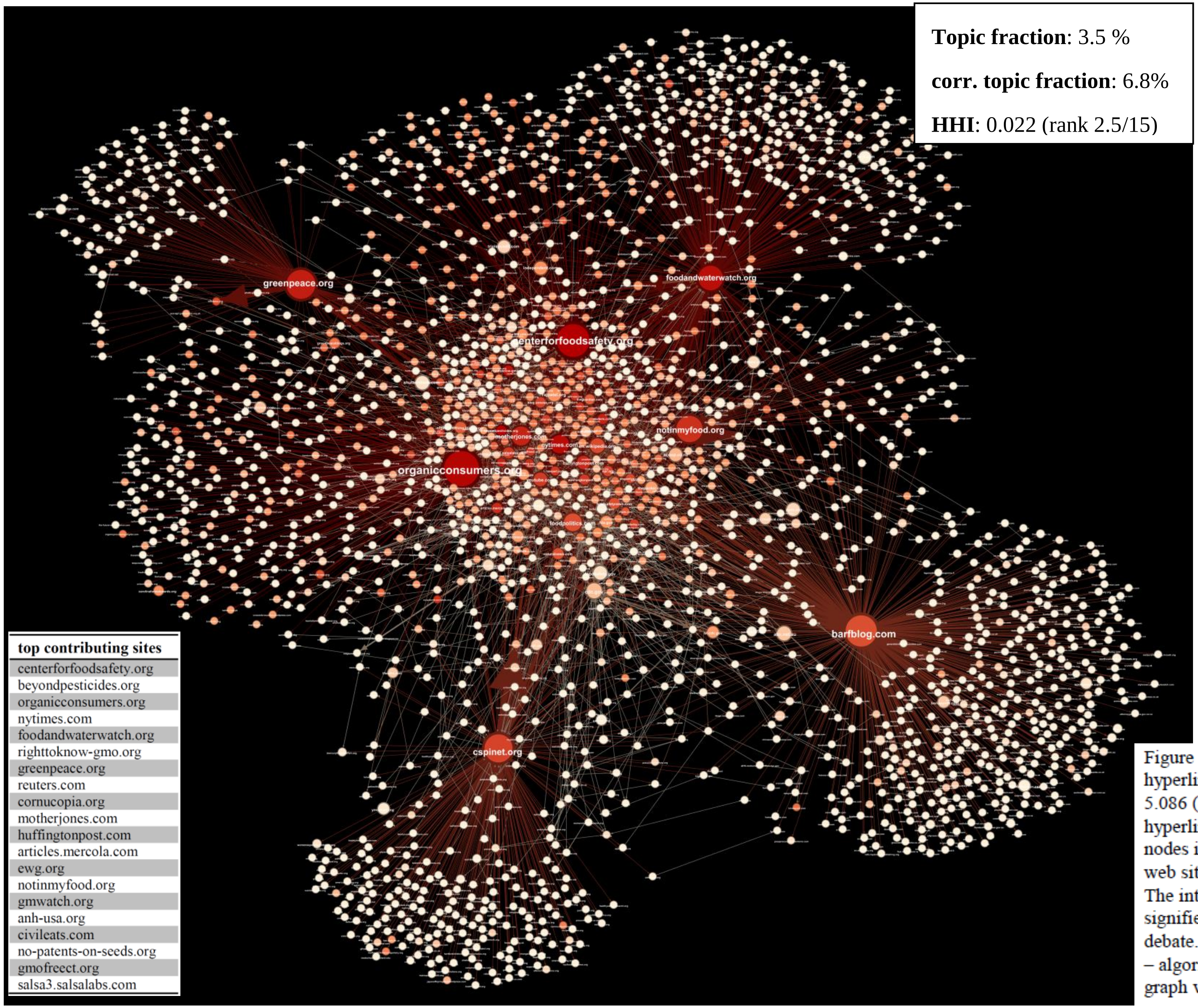
top contributing sites
foodsafety.gov
foodsafetynews.com
fda.gov
cspinet.org
barfblog.com
centerforfoodsafety.org
foodandwaterwatch.org
nytimes.com
notinmyfood.org
foodpolitics.com
organicconsumers.org
cargill.com
washingtonpost.com
motherjones.com
usatoday.com
cdc.gov
foodauthority.nsw.gov.au
foodsafetymagazine.com
theguardian.com
consumerreports.org

Top-words of topic 7

words	prob.
food	7,345%
safety	3,287%
products	1,425%
meat	1,347%
beef	0,798%
animal	0,788%
news	0,757%
chicken	0,753%
poultry	0,653%
salmonella	0,645%
industry	0,622%
health	0,615%
product	0,587%
recalls	0,538%
foods	0,535%
safe	0,512%
drug	0,483%
coli	0,473%
outbreak	0,468%
usda	0,459%
consumers	0,457%
animals	0,445%
consumer	0,436%
farms	0,418%
risk	0,413%
public	0,401%
government	0,385%
standards	0,372%
bacteria	0,346%
inspection	0,326%

Figure 4 depicts the (directed) graph of the hyperlink network among 2.211 web sites with 5.086 (dichotomized, cf. footnote 11, p. 12) hyperlinks connecting them. The size of the nodes indicates the number of html-files that the web sites' pages contribute to the content-corpus. The intensity of the red color of the nodes signifies the amount it contributes to the topical debate. The graph was drawn using the *Yifan Hu* –algorithm implemented in the open source graph visalization program *Gephi*.

Figure 5: *'Heat map' of topic 10"genetically modified foods & seeds"*



Topic fraction: 3.5 %

corr. topic fraction: 6.8%

HHI: 0.022 (rank 2.5/15)

top contributing sites
centerforfoodsafety.org
beyondpesticides.org
organicconsumers.org
nytimes.com
foodandwaterwatch.org
righttoknow-gmo.org
greenpeace.org
reuters.com
cornucopia.org
motherjones.com
huffingtonpost.com
articles.mercola.com
ewg.org
notinmyfood.org
gmwatch.org
anh-usa.org
civileats.com
no-patents-on-seeds.org
gmofreect.org
salsa3.salsalabs.com

Top-words of topic 10

words	prob.
food	2,307%
organic	1,903%
crops	1,857%
genetically	1,773%
farmers	1,372%
pesticides	1,274%
pesticide	1,137%
engineered	1,112%
monsanto	1,104%
agriculture	1,098%
seed	0,909%
corn	0,864%
foods	0,790%
modified	0,773%
seeds	0,744%
crop	0,727%
bees	0,649%
labeling	0,617%
gmos	0,613%
plant	0,502%
genetic	0,492%
agricultural	0,486%
environmental	0,479%
plants	0,476%
products	0,459%
farming	0,448%
usda	0,445%
herbicide	0,445%
biotech	0,434%
environment	0,400%

Figure 5 depicts the (directed) graph of the hyperlink network among 2.211 web sites with 5.086 (dichotomized, cf. footnote 11, p. 12) hyperlinks connecting them. The size of the nodes indicates the number of html-files that the web sites' pages contribute to the content-corpus. The intensity of the red color of the nodes signifies the amount it contributes to the topical debate. The graph was drawn using the *Yifan Hu* – algorithm implemented in the open source graph visalization program *Gephi*.

Annex

Topics of the low granularity topic model, ordered by topic proportion/fraction

Topic 20	prob.	Topic 23	prob.	Topic 11	prob.	Topic 25	prob.
health	1,72%	food	3,94%	news	2,29%	state	1,08%
public	1,02%	material	1,91%	sports	1,48%	public	0,91%
policy	1,02%	organic	1,51%	local	1,10%	bill	0,82%
research	1,00%	center	1,42%	weather	1,06%	government	0,75%
education	0,91%	site	1,38%	home	0,97%	states	0,71%
information	0,87%	safety	1,31%	county	0,74%	federal	0,59%
program	0,86%	issues	1,21%	health	0,73%	trade	0,58%
resources	0,74%	copyright	1,14%	business	0,68%	health	0,56%
national	0,72%	fair	1,03%	submit	0,66%	house	0,51%
news	0,71%	read	1,03%	sign	0,61%	congress	0,51%
programs	0,68%	action	1,01%	photos	0,57%	campaign	0,49%
events	0,65%	news	0,99%	facebook	0,53%	court	0,49%
members	0,63%	information	0,91%	click	0,52%	senate	0,47%
microbiology	0,61%	purposes	0,84%	santa	0,51%	rights	0,47%
center	0,60%	copyrighted	0,81%	chicago	0,51%	action	0,46%
community	0,57%	join	0,73%	back	0,48%	industry	0,46%
learn	0,50%	resources	0,71%	site	0,47%	president	0,44%
school	0,49%	legal	0,68%	mobile	0,45%	obama	0,43%
services	0,48%	home	0,65%	privacy	0,44%	support	0,41%
state	0,48%	publications	0,65%	close	0,44%	million	0,41%
join	0,47%	animal	0,63%	twitter	0,44%	report	0,39%
member	0,46%	videos	0,62%	entertainment	0,43%	vote	0,38%
media	0,45%	sludge	0,60%	read	0,42%	washington	0,36%
support	0,44%	foods	0,59%	address	0,42%	united	0,36%
board	0,43%	rights	0,58%	election	0,41%	national	0,35%
press	0,42%	south	0,58%	photo	0,40%	years	0,34%
issues	0,42%	section	0,57%	comments	0,38%	year	0,34%
membership	0,42%	room	0,56%	policy	0,38%	corporate	0,33%
search	0,40%	permission	0,56%	story	0,37%	groups	0,32%
american	0,39%	owner	0,56%	follow	0,37%	office	0,32%

Topic 31	prob.	Topic 7	prob.	Topic 18	prob.	Topic 15	prob.
news	2,37%	food	7,34%	food	3,28%	people	1,71%
world	0,84%	safety	3,29%	antibiotics	1,99%	time	0,99%
media	0,64%	products	1,43%	consumers	1,97%	it's	0,94%
comments	0,49%	meat	1,35%	share	1,90%	comments	0,93%
home	0,49%	beef	0,80%	meat	1,89%	make	0,86%
facebook	0,48%	animal	0,79%	cspi	1,68%	reply	0,72%
politics	0,46%	news	0,76%	labeling	1,50%	good	0,71%
latest	0,46%	chicken	0,75%	click	1,36%	years	0,70%
people	0,45%	poultry	0,65%	arsenic	1,13%	don't	0,69%
video	0,45%	salmonella	0,65%	consumer	1,10%	post	0,61%
environment	0,43%	industry	0,62%	safety	1,01%	service	0,58%
policy	0,41%	health	0,61%	union	0,92%	work	0,55%
action	0,40%	product	0,59%	trader	0,87%	posted	0,53%
science	0,39%	recalls	0,54%	newsroom	0,87%	things	0,50%
november	0,39%	foods	0,54%	posted	0,86%	back	0,46%
twitter	0,38%	safe	0,51%	label	0,81%	year	0,44%
follow	0,37%	drug	0,48%	drugs	0,81%	find	0,39%
stories	0,37%	coli	0,47%	joe's	0,74%	that's	0,36%
article	0,37%	outbreak	0,47%	public	0,67%	long	0,36%
culture	0,36%	usda	0,46%	health	0,65%	great	0,35%
rights	0,35%	consumers	0,46%	antibiotic	0,64%	made	0,35%
privacy	0,35%	animals	0,44%	reports	0,63%	thing	0,34%
education	0,34%	consumer	0,44%	gmos	0,61%	blog	0,33%
search	0,32%	farms	0,42%	action	0,58%	life	0,33%
social	0,31%	risk	0,41%	blog	0,57%	problem	0,31%
economy	0,31%	public	0,40%	eating	0,56%	money	0,30%
share	0,31%	government	0,38%	policy	0,56%	dogs	0,30%
health	0,30%	standards	0,37%	story	0,54%	real	0,29%
climate	0,30%	bacteria	0,35%	topics	0,51%	place	0,28%
america	0,29%	inspection	0,33%	foods	0,51%	dont	0,28%

Annex

Topic 16	prob.	Topic 10	prob.	Topic 5	prob.	Topic 1	prob.
food	7,30%	food	0,023067	study	1,18%	greenpeace	3,07%
farm	1,71%	organic	0,019028	health	0,97%	english	2,03%
local	1,16%	crops	0,018571	research	0,82%	arctic	1,32%
farmers	1,15%	genetically	0,017726	risk	0,82%	nuclear	1,10%
coffee	1,08%	farmers	0,013721	disease	0,75%	climate	0,88%
produce	0,86%	pesticides	0,01274	studies	0,70%	africa	0,82%
home	0,85%	pesticide	0,011374	article	0,65%	oceans	0,76%
fresh	0,82%	engineered	0,01112	data	0,65%	save	0,75%
policy	0,82%	monsanto	0,01104	levels	0,65%	español	0,75%
community	0,67%	agriculture	0,010985	found	0,61%	français	0,73%
market	0,65%	seed	0,00909	journal	0,60%	press	0,72%
canning	0,59%	corn	0,008638	human	0,56%	ship	0,69%
farming	0,59%	foods	0,007898	published	0,51%	ocean	0,67%
news	0,56%	modified	0,007733	exposure	0,50%	solutions	0,64%
hunger	0,54%	seeds	0,007437	science	0,49%	united	0,62%
organic	0,54%	crop	0,007272	effects	0,48%	asia	0,61%
city	0,52%	bees	0,006489	scientific	0,44%	international	0,61%
program	0,52%	labeling	0,006168	evidence	0,40%	rainbow	0,60%
agriculture	0,50%	gmos	0,006134	results	0,39%	esperanza	0,57%
sustainable	0,49%	plant	0,005022	animals	0,38%	warrior	0,57%
public	0,48%	genetic	0,00492	university	0,35%	republic	0,54%
farms	0,46%	agricultural	0,004857	researchers	0,34%	structure	0,53%
markets	0,45%	environmental	0,004789	environmental	0,34%	ships	0,53%
fruit	0,36%	plants	0,004764	issue	0,32%	islands	0,52%
vegetables	0,35%	products	0,004591	acid	0,32%	indonesia	0,51%
real	0,35%	farming	0,004481	samples	0,31%	forests	0,51%
school	0,34%	usda	0,004451	scientists	0,30%	change	0,51%
garden	0,34%	herbicide	0,004451	high	0,30%	sunrise	0,49%
nutrition	0,34%	biotech	0,004341	review	0,29%	centre	0,48%
grow	0,33%	environment	0,004003	articles	0,29%	victories	0,48%

Topic 3	prob.	Topic 27	prob.	Topic 30	prob.	Topic 17	prob.
climate	2,17%	october	5,12%	video	2,36%	health	1,88%
energy	1,72%	november	4,09%	music	0,59%	outbreak	1,55%
read	1,59%	september	3,35%	show	0,51%	case	1,41%
global	1,07%	august	3,15%	film	0,47%	salmonella	1,38%
change	1,05%	march	2,93%	series	0,45%	infections	1,29%
report	0,85%	april	2,84%	toronto	0,43%	information	1,09%
environmental	0,71%	june	2,76%	kids	0,42%	reported	1,08%
environment	0,69%	january	2,75%	back	0,40%	disease	1,05%
carbon	0,69%	july	2,69%	time	0,39%	cases	0,99%
power	0,68%	february	2,57%	review	0,35%	infection	0,93%
emissions	0,63%	december	2,46%	love	0,33%	count	0,93%
world	0,61%	food	1,51%	star	0,33%	page	0,92%
water	0,50%	comments	1,17%	home	0,32%	illness	0,87%
warming	0,45%	science	0,90%	movie	0,31%	persons	0,86%
natural	0,39%	video	0,83%	family	0,31%	outbreaks	0,80%
future	0,39%	recipes	0,69%	youtube	0,30%	file	0,78%
coal	0,38%	blog	0,53%	read	0,29%	symptoms	0,78%
interview	0,38%	policy	0,51%	watch	0,29%	public	0,75%
solar	0,38%	kitchen	0,49%	book	0,28%	linked	0,75%
years	0,37%	events	0,47%	play	0,28%	search	0,72%
land	0,37%	healthy	0,45%	canada	0,28%	states	0,66%
scientists	0,36%	tweet	0,42%	books	0,27%	foodborne	0,63%
pollution	0,34%	network	0,40%	baby	0,27%	skip	0,62%
green	0,33%	written	0,39%	night	0,27%	prevention	0,60%
development	0,32%	patel	0,36%	black	0,27%	human	0,58%
renewable	0,32%	categories	0,36%	canadian	0,27%	control	0,56%
forests	0,32%	archives	0,35%	awards	0,26%	curve	0,55%
plants	0,31%	clean	0,34%	videos	0,26%	state	0,53%
yale	0,31%	tags	0,34%	shows	0,26%	virus	0,50%
percent	0,29%	uncategorized	0,33%	story	0,26%	directly	0,49%

Annex

Topic 9	prob.	Topic 2	prob.	Topic 26	prob.	Topic 22	prob.
water	12,14%	food	4,86%	information	1,73%	cancer	3,19%
food	2,63%	safety	2,72%	sustainable	0,99%	health	2,72%
fracking	1,80%	posted	1,57%	site	0,86%	ebola	1,43%
service	1,29%	powell	1,27%	trade	0,83%	disease	0,86%
public	1,22%	reply	1,23%	content	0,74%	care	0,67%
watch	1,21%	june	1,11%	business	0,73%	skin	0,67%
works	1,00%	october	1,10%	social	0,70%	news	0,66%
quality	0,90%	september	1,09%	sustainability	0,68%	brain	0,65%
board	0,81%	april	1,06%	agreement	0,64%	natural	0,63%
bottled	0,68%	july	1,05%	services	0,62%	mercola	0,57%
resources	0,65%	coli	1,05%	policy	0,58%	medical	0,55%
moines	0,64%	august	1,05%	online	0,55%	article	0,51%
environmental	0,61%	march	1,03%	master	0,55%	heart	0,45%
local	0,56%	november	1,03%	fair	0,53%	food	0,45%
radiation	0,53%	restaurant	0,94%	register	0,51%	breast	0,45%
impacts	0,53%	doug	0,91%	privacy	0,44%	treatment	0,43%
issues	0,53%	february	0,90%	website	0,44%	vitamin	0,42%
clean	0,52%	december	0,90%	terms	0,42%	free	0,42%
privatization	0,50%	january	0,88%	provide	0,42%	body	0,41%
reports	0,50%	inspection	0,84%	subject	0,41%	information	0,40%
factory	0,49%	salmonella	0,74%	access	0,40%	site	0,38%
protect	0,49%	tagged	0,74%	corporate	0,40%	research	0,37%
customer	0,49%	norovirus	0,67%	management	0,39%	drug	0,36%
safe	0,45%	policy	0,65%	events	0,38%	vaccine	0,35%
drinking	0,44%	leave	0,59%	quality	0,37%	children	0,34%
people	0,43%	listeria	0,59%	security	0,37%	healthy	0,33%
control	0,41%	barfblog	0,56%	internet	0,36%	people	0,32%
services	0,41%	chapman	0,55%	personal	0,36%	patients	0,30%
policy	0,40%	culture	0,49%	copyright	0,35%	risk	0,29%
farms	0,40%	health	0,46%	including	0,34%	learn	0,29%

Topic 12	prob.	Topic 24	prob.	Topic 33	prob.	Topic 19	prob.
number	3,17%	data	1,60%	food	3,26%	fish	2,91%
volume	2,19%	business	1,05%	foods	2,19%	retrieved	1,84%
news	1,52%	news	0,98%	products	1,47%	salmon	1,78%
photos	1,20%	bloomberg	0,96%	sugar	1,05%	shutterstock	1,10%
home	0,82%	center	0,84%	nutrition	0,94%	page	1,00%
archives	0,80%	share	0,81%	trans	0,89%	seafood	0,86%
sports	0,66%	market	0,81%	coca-cola	0,80%	species	0,78%
edmonton	0,63%	markets	0,77%	diet	0,79%	mercury	0,76%
video	0,60%	home	0,76%	salt	0,71%	wikipedia	0,72%
today	0,57%	company	0,68%	health	0,65%	search	0,67%
entertainment	0,54%	video	0,60%	healthy	0,65%	stock	0,66%
comments	0,49%	facebook	0,59%	coke	0,64%	terms	0,59%
search	0,49%	google	0,58%	product	0,61%	create	0,57%
calgary	0,43%	industry	0,57%	ingredients	0,60%	wild	0,55%
sections	0,42%	reuters	0,57%	drinks	0,58%	images	0,54%
travel	0,41%	china	0,56%	drink	0,57%	account	0,52%
print	0,41%	jobs	0,52%	soda	0,54%	history	0,50%
articles	0,40%	percent	0,52%	water	0,48%	tuna	0,50%
business	0,40%	financial	0,48%	company	0,45%	english	0,49%
milwaukee	0,40%	report	0,48%	fruit	0,44%	fisheries	0,49%
opinion	0,39%	technology	0,45%	eating	0,43%	atlantic	0,48%
wisconsin	0,38%	global	0,43%	natural	0,42%	fishing	0,48%
page	0,38%	tweet	0,42%	juice	0,40%	navigation	0,47%
world	0,38%	energy	0,42%	obesity	0,40%	links	0,47%
alberta	0,37%	finance	0,42%	artificial	0,39%	privacy	0,44%
jobs	0,37%	companies	0,41%	read	0,38%	view	0,44%
life	0,36%	million	0,40%	vegetables	0,38%	found	0,43%
health	0,35%	billion	0,39%	beverage	0,38%	policy	0,42%
account	0,35%	sales	0,37%	free	0,37%	flight	0,42%
popular	0,34%	world	0,37%	weight	0,37%	articles	0,39%

Annex

Topic 36	prob.	Topic 0	prob.	Topic 21	prob.	Topic 14	prob.
milk	3,17%	opinion	1,55%	food	1,28%	organic	2,93%
farm	1,91%	times	1,43%	news	1,25%	cotton	1,78%
usda	1,56%	york	1,31%	huffpost	1,15%	clothing	1,40%
dairy	1,54%	business	1,24%	politics	0,92%	green	1,33%
animal	1,41%	hill	0,83%	arts	0,91%	natural	1,06%
cattle	1,38%	home	0,80%	science	0,89%	made	0,98%
organic	1,27%	op-ed	0,79%	voices	0,84%	fiber	0,95%
beef	1,08%	sports	0,78%	business	0,80%	fibers	0,93%
farmers	1,05%	health	0,77%	tech	0,78%	products	0,89%
cows	0,97%	arts	0,75%	search	0,78%	chemical	0,81%
agriculture	0,90%	video	0,70%	green	0,77%	sustainable	0,78%
animals	0,76%	search	0,70%	media	0,71%	chemicals	0,68%
livestock	0,69%	page	0,64%	edition	0,60%	fashion	0,64%
disease	0,65%	traffic	0,63%	world	0,59%	health	0,63%
producers	0,60%	real	0,61%	music	0,56%	skin	0,59%
letter	0,60%	estate	0,60%	life	0,56%	silk	0,57%
state	0,56%	subscribe	0,55%	follow	0,54%	facts	0,51%
food	0,55%	style	0,54%	post	0,52%	cleaning	0,48%
farms	0,55%	room	0,53%	live	0,51%	environmental	0,48%
products	0,55%	work	0,53%	healthy	0,48%	clothes	0,48%
sheep	0,53%	world	0,52%	entertainment	0,47%	high	0,47%
adobe	0,53%	travel	0,52%	huffington	0,45%	manufacturing	0,42%
national	0,53%	technology	0,50%	books	0,44%	process	0,39%
feed	0,50%	slow	0,50%	living	0,43%	plastic	0,37%
file	0,48%	nytimescom	0,49%	listen	0,43%	water	0,34%
consumers	0,48%	reading	0,47%	culture	0,42%	living	0,34%
program	0,48%	dining	0,44%	android	0,41%	care	0,33%
reader	0,48%	events	0,44%	morning	0,40%	trade	0,32%
bovine	0,47%	guide	0,44%	make	0,39%	toxic	0,31%
bill	0,43%	site	0,43%	health	0,39%	gold	0,30%

Topic 4	prob.	Topic 6	prob.
jones	1,94%	food	2,39%
mother	1,71%	health	0,73%
october	1,13%	comments	0,70%
august	1,12%	share	0,59%
july	1,07%	association	0,59%
june	1,04%	tags	0,56%
january	1,00%	foods	0,44%
september	1,00%	public	0,44%
february	0,98%	marketing	0,43%
april	0,98%	drinks	0,42%
march	0,97%	usda	0,40%
november	0,96%	kids	0,40%
december	0,95%	posts	0,40%
tweet	0,94%	sugars	0,40%
advertise	0,83%	meat	0,38%
subscribe	0,76%	marion	0,37%
motherjonescom	0,54%	nutrition	0,36%
heres	0,51%	calories	0,35%
mojo	0,50%	control	0,34%
year	0,46%	eating	0,32%
state	0,45%	corn	0,31%
percent	0,45%	safety	0,31%
city	0,42%	industry	0,31%
blue	0,38%	hfcs	0,29%
photo	0,36%	science	0,28%
follow	0,35%	move	0,27%
make	0,34%	agriculture	0,27%
marble	0,33%	fructose	0,27%
climate	0,33%	obesity	0,26%
home	0,33%	organics	0,26%