Structure and Content of the Discourse on Climate Change in the Blogosphere: The Big Picture

3 authors, including:

Dag Elgesem
University of Bergen
35 PUBLICATIONS **697** CITATIONS

SEE PROFILE

Lubos Steskal
University of Bergen
5 PUBLICATIONS **90** CITATIONS

SEE PROFILE

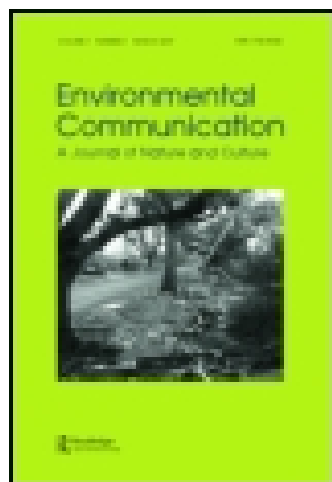**Some of the authors of this publication are also working on these related projects:**

Networks of texts and people View project

Structure and Content of the Discourse on Climate Change in the Blogosphere: The Big Picture

Dag Elgesem, Lubos Steskal & Nicholas Diakopoulos
Published online: 03 Dec 2014.

PLEASE SCROLL DOWN FOR ARTICLE

Routledge
Taylor & Francis Group

Structure and Content of the Discourse on Climate Change in the Blogosphere: The Big Picture

Dag Elgesem, Lubos Steskal & Nicholas Diakopoulos

Based on the texts of 1.3 million blog posts and the structure of the links between the blogs in which these posts appeared, this study presents an analysis of the discourse on climate change in the English-language blogosphere. Our approach combines community detection with probabilistic topic modeling to show how topics related to climate change are discussed across various parts of the blogosphere. We find that there is one community of predominantly climate skeptical blogs but several accepter communities. The topic analysis reveals a series of issues that are characteristic of the climate change discourse in the blogosphere. Two topics, one related to climate change science and one related to climate change politics, are particularly important for characterizing the discourse. We also find that the distribution of topics over the communities cuts across the divide between skeptics and non-skeptics (accepters) and that there are differences in the patterns of interactions between the skeptics and different groups of accepters.

Keywords: *climate change discourse; blogosphere; community detection; topic modeling; skeptics*

Introduction

The discussion of climate change is highly polarized (Dunlap & McCright, 2011; Hoggan, 2009; Mann, 2012; Washington & Cook, 2011), and the struggle has been particularly fierce in the blogosphere (Schäfer, 2012). The blogosphere has been a crucial outlet for climate change skeptics; for example, in the so-called "climategate"

Dag Elgesem is a Professor, Lubos Steskal is a Post-doctoral Fellow, and Nicholas Diakopoulos is a Research Fellow in the Department of Information Science and Media Studies, University of Bergen, Norway. Correspondence to: Dr Dag Elgesem, Department of Information Science and Media Studies, University of Bergen, P.O. Box 7802 Fosswinckelsgate 6, Bergen 5243, Norway. Email: dag.elgesem@uib.no

affair blogs played a key role in distributing the hacked e-mails (Hoggan, 2009; Pearce, 2010). Another, more recent illustration of the blogosphere's importance to the community of skeptics is their campaign to nominate the skeptical blog "WattsUpWithThat.com" for the yearly Best Blog Award (the Bloggies competition), resulting in the blog's selection as "Best Weblog of The year" in 2013 as well as the "Best Science Blog" for the third time in a row.[1] A number of science blogs chose to withdraw from the competition in protest at the process.[2] But skeptical voices are by no means the only ones in the blogosphere to be heard on the issue of climate change (see, e.g., DeSmogBlog.com, RealClimate.org, SkepticalScience.com).

There is extensive research on climate change coverage in the mainstream media (for overviews, see Boykoff, 2011; Schäfer, 2012), and an important recent study by Sharman (2014) investigates the structure and content of the climate skeptical blogosphere. Nevertheless, we are not aware of any attempts to chart the entire structure of the climate change blogosphere and the topics that get the attention of the bloggers in the different parts of this network. The present research aims to fill this gap by, first, mapping the hyperlink structure of the climate change blogs, then identifying what topics are discussed in the blogs, and, finally, showing which topics characterize the discussion in the various parts of the network. We address three research questions:

(1) How is the hyperlink network of (English) climate change blogs structured?
(2) How does this structure relate to the differences of opinion, as well as the presumed divide, between those bloggers who accept the consensus view on anthropogenic warming and those bloggers who are skeptical of this view?
(3) How is the topic of climate change discussed in the different communities in the blogging network?

As a first step in our investigation, we crawled approximately 3000 English-language blogs discussing climate change, downloaded their links, and extracted the texts of some 1.3 million blog posts. In a second step, we undertook a community detection analysis of the network and visualized it. We then manually classified the blogs as climate skeptics, accepters, or neutral. Afterwards, we subjected the corpus of texts from the blog posts to a probabilistic topic analysis to generate a model of the corpus topics. Upon presenting the results of these analyses, we then show how the topics are distributed over the communities in the climate change blogosphere. Finally, on the basis of the patterns that emerge, we discuss what inferences can be made about the relationships between the structure of links and the topics.

Overview of the Literature

The media coverage of climate change varies considerably between countries and over time (Boykoff, 2011) with respect to the release of Intergovernmental Panel on Climate Change (IPCC) reports and the yearly Conference of the Parties (COP) meetings (Eide & Kunelius, 2013). Analyses have shown that media with a

conservative profile, versus those with liberal or left-leaning positions, give more room to skeptical voices (Painter, 2011). There are also differences between countries in the styles of climate change coverage (Eide & Kunelius, 2013) and in the presence of skeptical voices in the media (Boykoff, 2011; Painter, 2011).

Political affiliation has been shown to be a strong predictor of attitude to climate change issues, with conservatives, versus liberals and leftists, as more likely to be climate skeptics. And it has been documented that conservative political organizations have actively lobbied against legislation to mitigate global warming (Dunlap & McCright, 2011; Oreskes & Conway, 2010).

The debate between the skeptical side and those that represent the majority view is polarized and characterized by hostility (Hoffman, 2011; Washington & Cook, 2011). This is also reflected in the tone of language used in the debates. In relation to the climategate affair, this polarization was driven to new heights with accusations of fraud and scientific misconduct (Pearce, 2010). An interesting recent study of the skeptical discourse online is Koteyko, Jaspal, and Nehrich (2012), which analyses data from readers' comments on online UK tabloids articles about the climategate affair. Using methods from corpus linguistics, they identify characteristic patterns in the negative representation that skeptics give of climate scientists, where the skeptical discourse in these forums is often strongly pejorative.

The term "skeptic" is in itself controversial. Washington and Cook (2011) argue that the label "skeptic" is misleading because it implies that non-skeptical climate scientists are not properly abiding by the scientific norm of rational skepticism. For the authors, those who oppose the theory of anthropogenic climate change should instead be called deniers. Still, in this study we will refer to those who accept the theory of anthropogenic climate change as "climate accepters" and those who reject this theory as "climate skeptics."

Rahmstorf (2004) distinguishes between three types of skeptics: trend skeptics who question global warning, attribution skeptics who question that human activity has significant effects on the climate, and impact skeptics who question the negative consequences from climate change. Sharman (2014) distinguishes between skepticism based on arguments against the science and skeptical arguments directed at political measures aimed at climate change mitigation. In her highly relevant study of skeptical climate change blogs, she finds that the most central skeptical blogs are largely concerned with climate science issues.

With respect to the methods used, several studies have analyzed the manifestation of polarization of political opinions in the online public sphere. Adamic and Glance (2005) charted the linking practices of the top Republican and Democrat blogs and found a strong tendency on both sides to link to other blogs within the same political community but not to the other side. The authors' illustration of a divided political blogosphere in the USA has become an icon of the polarized web. Sunstein (2006, 2007) finds evidence of the same tendency in the linking practices of the websites of political organizations and interprets that tendency as sign of a polarized public sphere. The same phenomenon has been found with communication in social media.

Himelboim, McCreery, and Smith (2013) identify clusters of people who use Twitter to communicate about politically charged topics, including climate change, and show that most of the clusters are quite politically homogenous. Adamic (2008) suggests that a possible explanation for why bloggers are reluctant to link to blogs with which they disagree is that by linking they would give attention and prominence to opposing views. Rogers and Marres (2000) chart the linking practices of websites of industry, government, and activist organizations concerned with climate change and find systematic differences in their linking practices. For example, they found few reciprocal links between industry and activist organizations and suggest that the absence of links in this case is a sign of a lack of recognition of the other party in the climate change discussion.

Several interesting approaches to statistical models of linking practices have been recently suggested. Gonzales-Bailon (2009), Shumate and Dewitt (2008), Ackland and O'Neill (2011), and Lusher and Ackland (2011) analyze the linking patterns among social movement websites and distinguish between links formed on the basis of content similarity (homophily) and links that must be explained with reference to the network's structural patterns. Ackland and O'Neill (2011) also provide a description of their network of environmental websites at the textual level and thereby are able to provide a richer context for interpreting the linking practices. They find that the texts on different webpages often contain the same key terms, even if the websites share few hyperlinks. Still, the exact relationship between the structure of hyperlinks and the relationships at the textual level remain unclear.

Methods

In this section we explain our methods, with an especially detailed focus on explaining probabilistic topic detection (LDA analysis).

Sampling, crawling, and extraction

One methodological goal was to maximize our collection of English-language, climate change blogs. A blog is a webpage with articles—blog posts—organized in a chronologically reversed order of their publication. A blog is connected to other webpages via hyperlinks from the blog's stable frame—blog rolls—or from links embedded in the blog post's text. We started with five seed blogs that were chosen because they were well connected in the climate change blogosphere and represented differing viewpoints in the climate debate. We harvested all posts from the seed blogs and extracted key terms from them, which were then used to determine topical relevance in our crawl. First, a frequency sorted word list was scanned to identify frequent words that were typical of the domain, resulting in terms like *climate, global, carbon, emissions, temperature, sea, solar, greenhouse*, etc. Afterwards, lists of the *n*-grams (word sequence length of $2 <= n <= 5$) containing each of these words were extracted and scanned for frequently occurring terms, e.g. *climate change, climate science, carbon dioxide, emissions trading, sea level*. We executed a breadth-first crawl

from the seed blogs, with the criteria that for a blog to be harvested it must be in English and have at least one key term appearing somewhere on its homepage. The crawl followed links from the homepage of each blog. If the next blog visited did not match our topic criteria, the crawl went one step further from the blog roll. Because of the effort needed to adapt programs to parse content from different blog platforms, we limited the crawl to Wordpress and Blogspot blogs. By searching blogs for "climate change" and "global warming" with different search engines, it appeared that there were roughly five times as many hits on these two platforms as on all others combined, which suggested that we would capture the bulk of the climate change blogs by crawling these platforms.

In a crawl carried out between June and September 2012, we downloaded the complete content of about 3000 English-language blogs (about 1.3 million posts) that included posts mentioning "climate change" or a related term. From the downloaded html files we extracted the text content of each post by using the Alchemy API[3] and stored this in a MongoDB database, extracted links between blogs, and stored these in a Neo4j graph database.

Community detection

Although the task of community detection is a well-studied challenge, the notion of what a community is has no single definition and thus no single method for addressing this challenge. Moreover, there is a plethora of methods for finding an underlying structure in a network (Fortunato & Barthelemy, 2007). This implies that even though these methods address the same problem when defined loosely, their outcomes might be similar for some input networks but very different for others.

We decided to use the modularity maximization algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Our reasons were threefold: modularity-based algorithms yield good results and belong to the most popular approaches in the literature (Fortunato & Barthelemy, 2007), they are computationally very efficient, and they are already implemented in the tool we used for network analysis (Gephi).

The intuition behind this method is that a community has more internal links than what would be expected in a random graph with the same properties (Newman, 2006). The modularity score of a set of nodes measures to what extent the number of links within this set is larger (or smaller) than expected in a random network with the same degree distribution. The larger the score, the more likely it is that the links were not generated by chance but by a process common for all nodes in the group. A modularity maximization algorithm tries to ascertain a network partition for which the sum of modularity scores for each partition is maximum.

It has been noted, however, that this approach suffers from resolution problems (Fortunato & Barthelemy, 2007) when the size of some of the actual communities is disproportionate to others (Lancichinetti & Fortunato, 2011). Since we are particularly interested in identifying the most dominant communities and we had no prior information on the scale of existing communities, we chose to disregard this

problem. That is why we chose a resolution parameter of 1, which is equivalent to the basic modularity optimization problem (Lambiotte, Delvenne, & Barahona, 2009).

Manual classification

We undertook a manual classification of the 1497 blogs in the seven largest groups in the central part of the graph identified from community detection. The blogs were classified as "accepting" the majority view on anthropogenic global warming (AGW), "skeptical" of this view, or "neutral" to it. The procedure used was as follows: (1) open the blog in the browser, (2) find the posts on the blog tagged with "climate change" either via the blog's search function or by clicking on the link to the category, (3) search through the result page for the word "climate," (4) inspect the sentences containing this word and its context, and (5) see if it is possible to identify statements that explicitly express the views of an "accepter" or a "skeptic." If no explicit endorsement of either position could be found, the blog was classified as neutral.

During the classification, we looked at more than one statement if we thought the first statement was not clear enough. The default classification of a blog was "neutral." We then checked if it could be classified as "skeptic": we classified a blog as "skeptic" if it explicitly rejected that global warming is happening (trend skeptic), questioned that human activity has an effect on climate (attribution skeptic), or that climate change has serious consequences (impact skeptic). If the blog was not found to be "skeptic" and expressed concern over, respectively, either global warming, that human activity has an effect on climate change, or the impact of climate change, it was classified as "accepter." Blogs that could not be classified as either "skeptic" or "accepter" were labeled "neutral." In a few cases the blog's policy was to represent both positions; these blogs were also classified as "neutral."

Most of the blogs classified as "skeptic" explicitly rejected mainstream climate science. Blogs using terms like "warmist," "alarmist," or "AGW religion" to characterize mainstream climate scientists or people arguing in favor of measures to mitigate climate change were classified as "skeptic." On the other side, several blogs classified as "accepter" expressed views on what should be done in the fight against climate change, even if it never actually mentioned its causes. For example, blogs that said things like "climate change is the most serious threat facing humanity today" were classified as "accepter." This method allowed us to classify the entire blog on the basis of statements in one or a few blog posts, with the assumption that the blogger is consistent and does not change his or her mind. One researcher performed the initial coding, and inter-coder agreement was then tested by letting a second coder classify a sample of 60 blogs. The agreement was 84.8% and the weighted Cohen's kappa was 0.72, which is considered to be sufficient (Fleiss, Levin, & Paik, 2003).

Topic detection

To identify the topics spanning over our corpus, we decided to use an unsupervised inference method called Latent Dirichlet Allocation (abbreviated LDA; see Blei, Ng, &

Jordan, 2003; Blei and Lafferty, 2009). LDA is a member of a class of methods for characterizing the hidden structure of a large corpus. It is an unsupervised method, meaning that the analysis is carried out automatically without any human support in the analysis. Furthermore, LDA is not a *classification* but a model of the corpus's topical structure. When the algorithm is run on a set of texts, the first output of the analysis is the number of topics where each topic is represented by a set of words that frequently occur together in the corpus. The basis for interpreting a group of words as a topic is that topics are regularly discussed with the use of a characteristic vocabulary. In some cases, the given group of words consists mainly of function words and does not give meaning as a topic.

In this analysis, the topics themselves are represented as hidden variables and are visible only in the form of groups of words from the corpus. Hence, the only manifested elements are the words from the texts. Note that two important assumptions are built into LDA. One is that a given corpus of texts can be modeled by a fixed number of topics. Hence, the number of topics must be given externally as a parameter to the process, and LDA does not provide an estimate of the optimal number of topics for modeling a given corpus. This is a limitation of the method and calls for experimenting with different numbers of topics. The other assumption that LDA makes is that each topic is present in every document to some degree. A second output from the analysis is therefore a distribution of all topics for each document. In practice a small number of topics, one to three, will make up more than 90% of the topics in a given document. For example, in one of our analyses of the climate change corpus, we ran the LDA analysis with 60 topics as a parameter and thus obtained 60 groups with 19 words in each as a suggested corpus model. Five of these topics contained the words "climate change." The result is displayed in Table 1.

We see that the groups of words can be interpreted as signifying various aspects of the climate change issue, i.e. as various topics related to climate change. We have suggested labels for the topics in the column to the right (Table 1). However, the topics generated by this model are not clearly distinct, and the model makes topical

Table 1. The 5 topics (out of 60) identified by the LDA algorithm containing the words "climate change."

	Topics	Label
1	temperature data climate surface model global models temperatures change al warming heat effect average period trend time years analysis	Climate change science
2	climate development change countries water world environmental research international sustainable food resources areas economic land work global issues environment	Climate change politics
3	climate carbon emissions change global countries china world greenhouse energy gas nations dioxide reduce copenhagen trade international kyoto year	Climate change, international negotiations
4	climate warming global change science scientists scientific ipcc evidence research years weather report gore world earth dr university scientist	Climate change science, IPCC
5	ice sea ocean climate years carbon arctic global warming earth water change scientists atmosphere level dioxide study levels rise	Climate change and global warming

distinctions that are not intuitively clear. In the analysis below we use 20 topics to obtain a less fine-grained model which we argue is more adequate.

To carry out the analysis, we used MALLET, a freely available Java implementation of LDA.[4]

Results

Our first research question was: How is the hyperlink network of the (English) climate change blogs structured?

We first ran the community detection algorithm on the network we harvested. This yielded 19 various-sized groups, as shown in Figure 1.

The size of areas in the graph is proportional to the number of blogs in the groups. Closeness represents density of links. The manual analysis (see below) revealed that the group of blogs colored with red is the predominantly skeptical community, which is also the largest community, while the rest are dominated by accepters. The group colored in yellow, which is one of the groups dominated by "accepters," is the group with which the skeptical community shares the most links.

Note that the size of the different areas represents the number of blogs in the group, not the number of blog posts. The distance between the nodes represents link density. So, for example, there are more links between the large red group and the

Figure 1. The network of climate change blogs, colored by community.
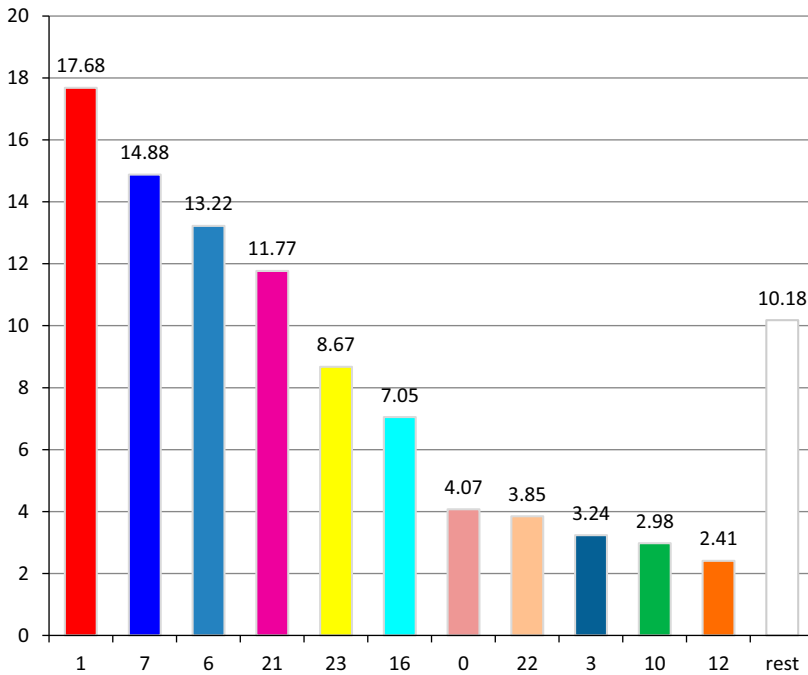
Figure 2. The percentage of blogs in each group. The color scheme is the same as in the one used in the graph in Figure 1.

group colored in yellow than between the red group and the purple group to the north. The proportion of blogs in each group is charted in Figure 2.

There are 9 groups with more than 50 blogs, which we have coded as accepting the majority view on climate change, as skeptical of the majority view, or as not expressing a view on the question (neutral). Two of these groups (groups 16 and 21 in Figure 2) consist mainly of blogs that are not classifiable as either skeptics or accepters. These groups are therefore of less interest to our analysis of blogging on climate change. This leaves us with seven groups: 1, 7, 6, 23, 0, 22, and 3.

Skeptics and accepters

Our second research question was: How does this structure relate to differences of opinion, as well as the presumed divide, between the bloggers who accept the consensus view on anthropogenic warming and those bloggers who are skeptical of this view?

The results from manually coding the blogs as skeptical, accepters, and neutral yielded the distribution depicted in Figure 3.

As Figure 3 shows, group 1 (colored with red in the graph in Figure 1) is the only one that consists of predominantly skeptical blogs, while the other groups are dominated by blogs that subscribe to the majority view on climate change. Note, however, in all groups there are some dissenters.
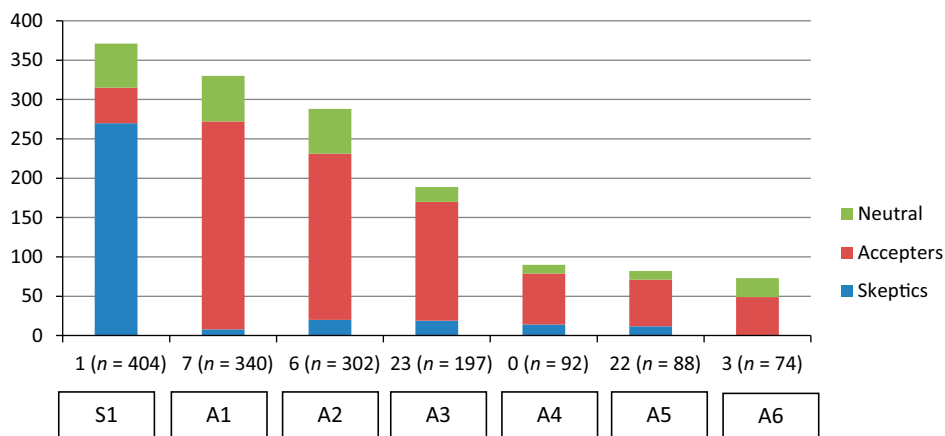
Figure 3. The distribution of skeptical, accepting, and neutral blogs in the seven largest among the central groups of blogs concerned with climate change.

In the rest of the study we will refer to the predominantly skeptical group 1 as S1, and in other communities as, respectively, A1 (group 7), A2 (group 6), A3 (group 23), A4 (group 0), A2 (group 22), and A6 (group 3).

We used PageRank (Brin & Page, 1998) as a blog centrality measure, since it is a centrality measure designed specifically for webpages, and it models the number of page views in a random walker model of the user. The advantage over pure degree centrality is that it models the effect of being endorsed by popular blogs. The skeptical blog "WattsUpWithThat.com" (in community S1) came out with the highest PageRank of all the blogs. The blog with the highest PageRank among the blogs in community A1 was "tamino.wordpress.com."

Topic analysis

Our third research question was: How is the topic of climate change discussed in the different communities in the blogging network?

We approach this question by using the LDA technique to analyze the texts of the 1.3 million blog posts in our corpus. As noted above, this is an unsupervised technique that automatically generates a set of topics represented by a set of words. However, the number of topics to be found has to be given as a parameter in the algorithm. For this reason, it is necessary to try out the algorithm with different numbers of topics to generate corpus models that can be compared and evaluated. There is no "correct" number of topics that characterize a given corpus: the question is whether the model is useful for the purpose of a given analysis.

Our study aims to compare the discourses in the various parts of the climate change blogosphere, and we ran the LDA algorithm with the number of topics set to 20 and 60. Table 2 reproduces the output from processing with 20 topics. We have also suggested labels for each topic. The labels do not serve any analytic purpose, but

Table 2. The outcome of 20-topics analysis in the corpus.

	Topic	Suggested label
0	0.10053 obama party president government people political election vote campaign media bush house bill democrats republican don mr time public	US politics
1	0.04269 city people road car community street local development urban good building traffic cities public london transportation cars council work	Urban
2	0.01052 information site public blog views responsible sites ehs opinions good contained linked understand facts website people material expressed posted	Blogging
3	0.05328 water weather river south area north sea service information coast disaster year storm people island high areas west region	Weather
4	0.04927 species birds garden food year time trees plants water bird small found plant tree day back good long ve	Wildlife
5	0.03707 space science research earth university nasa time system light life years scientists technology researchers field team theory work found	Science and technology
6	0.14584 ago time don people day back good ve year years ll make things days home week didn work hours	(incoherent)
7	0.07379 state law court public federal government police case act department states county mr rights justice information bill committee office	Government
8	0.06804 climate change world countries environmental international development global environment people africa food india resources national economic nations government sustainable	Climate change politics
9	0.00278 de da em para os es um se uma na mais dos por como sobre ncia das rio ao	(Portugese)
10	0.00682 att och det som en av om inte den med die jag till deals alstom har de du fr	(Swedish)
11	0.05081 climate global warming change science data ice carbon temperature years scientists sea earth year emissions temperatures scientific time ipcc	Climate change science
12	0.09517 people world life human god time society make good things fact power man political history don book social point	New age
13	0.0356 health study medical people children food cancer research care disease environmental public found women risk patients university drug treatment	Health
14	0.01038 love life di jesus sl playboy don time age la start change che model joy abortion meaningful universal loving	(incoherent)
15	0.06124 blog reading aud site google http news post email www phone information internet min web video uk free posts communication)	(incoherent/ online communication)
16	0.06057 energy oil power nuclear gas fuel coal solar production wind carbon electricity industry plant emissions renewable plants technology year	Energy
17	0.06955 government money economic tax economy market year people financial billion percent million pay debt bank years world system jobs	Economic policy
18	0.05946 war military iraq american world iran government united israel people states security country president bush al nuclear afghanistan forces	US military
19	0.00494 de la en el los del las se por es una para hace con le est les al lo	(Spanish)

The decimal number in front of each topic indicates the relative proportion of the topic in the corpus.

are included to make it easier to reference the topics and to remember the gist of each one.

Not all of these topics make equally good sense. Topics #14 and #15 are both useless and contain terms that should have been filtered out by the stop word list. The Portuguese topic #9, the Swedish topic #10, and the Spanish topic #19 have been identified because they have a different vocabulary than the rest of the corpus and,

therefore, are singled out as distinct topics. The documents were included in the corpus in the first place because some of the climate-related words we used in the crawls appeared on the front page of these blogs. The rest of the topics are sufficiently coherent to provide helpful characterizations of different discourse parts that exist in the corpus, and several of them seem to be somehow related to aspects of climate change.

Topics #8 and #11 are of particular interest, of course, since they both contain the words "climate change." They seem to be concerned with two different types of issues related to climate change: topic #11 is a set of terms related to the science of climate change while the words in topic #8 relate more to the politics of climate change (hence the suggested labels). To justify the labels for these topics, we have manually inspected the top 100 documents determined by the algorithm to have the strongest presence of the two topics. Intuitively, the topics of the two groups of blogs are quite distinct, and "climate change science" and "climate change politics" seem to capture the difference.

It seems impossible to give coherent interpretations of topics #6, #14 and #15. This does not invalidate the model but only goes to show that the method could find 15 rather than 20 coherent topics. The fact that the LDA method came up with some groups of frequently co-occurring words that are not intuitively coherent does not mean that the other topics can be used in the discourse analysis. The result of the LDA method is not the final truth about the corpus in question but a suggestion for a model that can be more or less useful. We will argue that the model described in Table 2 provides a useful perspective on the climate change discourses in the blogosphere.

Validation

It might be thought to be a problem for the analysis that the the topics #8 and #11 "overlap", since the words "climate change" appear in both topics. Yet that is not the case. First, LDA does not classify documents into mutually exclusive categories but assumes that all topics are present in every document to some degree. Second, LDA denotes topics by identifying groups of words that frequently occur together. That the words "climate change" show up in two topics—i.e. "climate change politics" (#8) and "climate change science" (#11)—only means that they appear frequently in two different contexts. This brings out an interesting and important property of the discourse, and it is therefore a feature, not a defect, of the method, namely its ability to identify different contexts for the use of this central concept.

The appearance of the terms "climate" and "change" in two topics suggests they are used in semantically different ways in two different contexts. To check this we identified for each topic the 100 documents where the topic in question accounted for the largest portion of the text. According to the LDA model, the respective topics accounted for more than 95% of the content in these documents. Hence, if the topics really referred to semantic differences—i.e. something like "climate change science" and "climate change politics"—the differences should be clearly visible in these documents. We tested this by using two methods from corpus linguistics: keyword

Table 3. The 18 most frequent keywords from the analysis of the top 100 documents for the topics "climate change science" and "climate change politics."

Top keywords in the top "climate science" documents	Top keywords in the top 100 "climate politics" documents
1 CLIMATE	1 CLIMATE
2 TEMPERATURE	2 CHANGE
3 GLOBAL	3 DEVELOPMENT
4 WARMING	4 COUNTRIES
5 OCEAN	5 MOUNTAIN
6 SURFACE	6 WILL
7 DATA	7 KNOWLEDGE
8 SEA	8 ADAPTATION
9 ICE	9 INTERNATIONAL
10 CHANGE	10 WORK
11 AL	11 RESEARCH
12 ET	12 WATER
13 HEAT	13 SUSTAINABLE
14 CHANGES	14 HIMALAYA
15 YEARS	15 SHARING
16 VARIABILITY	16 CONFERENCE
17 CARBON	17 ISSUES
18 MODEL	18 MEETING

Reference corpus: The British National Corpus, approximately 100 million words.

analysis of the two groups of documents, and computing the strongest collocates of the words "climate," "change," and "global" in the two groups. In addition, we read through the titles of the 200 documents. Table 3 shows the keyword analysis of the two groups of documents:

We see that the most keywords are quite distinct, and the analysis seems to support our interpretation and the labels.

Our second test was to check what words collocated most strongly with the terms "climate," "change," and "global" in the top 100 documents for "climate change science" and "climate change politics." The result for "climate" is displayed in Table 4. The same test with "change" and "global" showed the same distinction.

In this case we have computed the point-wise mutual information metric of the words with respect to the British National Corpus. The analysis shows a clear difference in the most distinct contexts in which the terms are used in the respective document groups. Moreover, a manual inspection of the document titles supports the assumption that the terms "climate" and "change" are used in semantically distinct contexts in the two groups of texts, thus strengthening our confidence that the model actually has picked out a significant difference in the way climate change is discussed in the blogosphere.

A polarized discourse

Studies of other media have found that discussion between skeptics and accepters is characterized by unfriendly language and mutual labeling (Koteyko et al., 2012), from

Table 4. The top 18 collocates computed with the point-wise mutual information metric.

Collocates of "CLIMATE" in top "climate science" documents	Collocates of "CLIMATE" in top "climate politics" documents
1 DECADAL	1 INNOVATES
2 VAPOR	2 HIMALAYA
3 ANTHROPOGENICALLY	3 DRR
4 INTERDECADAL	4 HIMALAYAN
5 SMERDON	5 CANCUN
6 KNMI	6 ORG
7 FEEDBACKS	7 MAINSTREAMING
8 SCHEFFER	8 REDD
9 ANTHROPOGENIC	9 BIODIVERSITY
10 ERBE	10 CLIMATE
11 IPCC	11 MITIGATION
12 GISS	12 THEMATIC
13 MODELED	13 IMPACTS
14 KELVINS	14 STARTER'S
15 THERMOHALINE	15 CSD
16 SKILLFULLY	16 ADAPTATION
17 DENIER	17 DESERTIFICATION
18 MODELING	18 DURBAN

Reference corpus: The British National Corpus, approximately 100 million words.

the skeptical side in particular. Our analysis did not produce a topic with these characteristics. However, as part of our corpus analysis, we computed the collocates of the terms "climate," "change," and "global" in all blog posts from each community. We computed what words have the strongest relation of mutual information with these terms as compared with their occurrences in a corpus representing the general language (The British National Corpus). From this analysis, we obtained the top 15 collocates for "climate" in the three largest communities (see Table 5).

As Table 5 shows, the labels that characterize this debate, like "skeptic," "alarmism," "deniers," "contrarians," and others, rank high on the list of strongly informative collocates of "climate." We see that skeptics discuss the labels given to them by accepters (e.g. "deniers") and vice versa (e.g. "alarmism"). These terms are not very frequent, and they do not appear with high regularity (which is why they do not appear in any of the topics), but the collocation analyses show them to be distinctive of how the word "climate" is used in these documents. The same terms are distinctive of the language of the other accepter communities we have discussed. This type of language, which has been found as a characteristic of the polarized debate over climate change, can thus be seen as salient in the climate change discourse in the blogosphere.

The distribution of topics throughout the groups of blogs

The question now is how these topics are represented at the blog and community levels. In analyzing this we will focus on the seven largest and most central communities. These are the communities containing the blogs that we have manually

Table 5. The top 15 collocates around "climate" in communities 1 (skeptic), 23 (accepter), and 7 (accepter) computed with the point-wise mutual information metric.

Top collocates of "CLIMATE" in the skeptical community S1	Top collocates of "CLIMATE" in the accepter community A3	Top collocates of "CLIMATE" in the accepter community A1
1 CLIMATE	1 DENIERS	1 POPPIN
2 SKEPTICS	2 SKEPTICS	2 DENIERS
3 ALARMISM	3 CLIMAT	3 SKEPTICS
4 DENIERS	4 DECADAL	4 OBAMA
5 IPCC	5 CONTRARIANS	5 WWW
6 DECADAL	6 OBAMA	6 EU'S
7 ALARMISTS	7 NOAA'S	7 CLIMATE
8 CLIMAT	8 AGW	8 YVO
9 CHANGE	9 WWW	9 NOAA'S
10 INTERGOVERNMENTAL	10 DENIER	10 WILDFIRES
11 OBAMA	11 CLIMATE	11 CHANGE'S
12 ANTHROPOGENIC	12 VAPOR	12 IPCC
13 AGW	13 ANTHROPOGENIC	13 ALARMISM
14 IPCC'S	14 ALARMISM	14 PACHAURI
15 WARMING	15 CONTRARIAN	15 DENIER

Reference corpus: The British National Corpus, approximately 100 million words.

coded as skeptics, accepters, and neutral. The LDA algorithm determines the presence of each topic in each document (blog post). To compare the blogs, we have to average the presence of all the topics in all the blog posts in each blog to get the relative "intensity" of each topic in the blog's total output. On the basis of this calculation of the degree of topic presence, we can then move up one level and calculate the presence of the topics in the communities.

Figure 4 shows how the topics "US politics," "climate change politics," "climate change science," "energy," and "economic policy" are distributed over the communities.
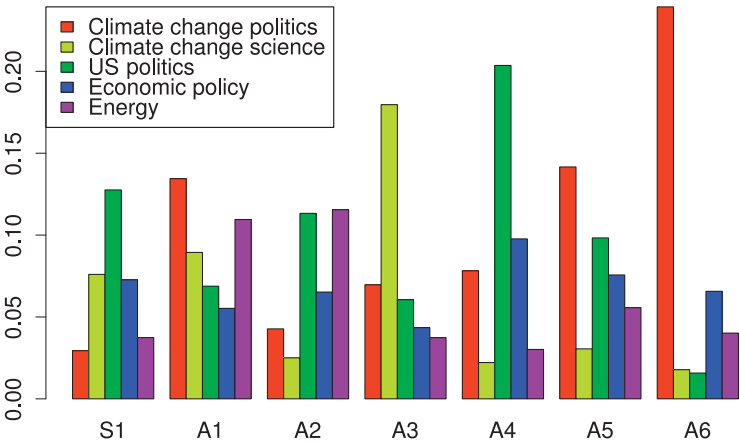
Figure 4. The distribution of selected topics across the seven communities.

Note that this is not a representation of the number of blog posts about climate change but the aggregated measure of the presence of the topics in the communities. The graphs show that the topics related to climate change are contextualized in different ways in the different communities. We see that that groups A1, A3, and S1 are the ones with the strongest presence of "climate change science," whereas "climate change politics" is dominating community A6 and strong in A1 and A5.

Discussion

Blogs are essentially networked texts that are connected both via hyperlinks and via references to other blogs in the texts. To analyze the networks of blogs about climate change we have crawled the hyperlinks of large set of blogs about this topic and harvested the posts published on these blogs. The analysis of the hyperlink structure showed the network to have a number of distinct communities, one of which turned out to be predominantly skeptical while the other communities were dominated by accepters. However, there are also important differences between accepter communities. One distinctive feature is that one of the accepter communities has a much higher level of mutual linking with the skeptical community than the other accepter communities, which suggests a more active engagement with the skeptics.

There are also important differences in the topic profiles between the accepter communities, and the relative emphasis on the scientific and political aspects of climate change varies considerably between the communities. Our analyses show that the central blogs in both the predominantly skeptical community and the central blogs in the accepter community, with which the skeptics engage the most actively, are publishing most about topics related to the science of climate change. This suggests that climate change science is the most central topic in the exchanges between the skeptics and the accepters. The other accepter groups also disagree with the skeptics, but they seem to engage less actively with them, as the links between these communities are sparse. Also, the accepter groups are different in the way they link with each other.

Charting the structure and content of the climate change blogosphere thus reveals distinctive patterns in the structural relationships among these discursive communities, patterns that would be difficult to detect with other methods. We see that the communities are not structured simply on the basis of disagreement between skeptics and accepters but also by various groups of accepters that focus on different aspects of the climate change debate, and these various accepter groups are often not linking to each other. Our analysis of the patterns of topics suggests that the discussion about the science of climate change is distinctively different from the discussion of climate change as a political issue in the blogosphere. We thus find two different types of climate skepticism, and both occur together in the skeptical community. Our analysis also reveals different ways of being an accepter, and suggests many of the accepter blogs are concerned more with issues related to the politics of climate change than with climate science. In addition, the accepter communities differ in the way their

writing about climate change brings in other topics like, for example, energy or developmental issues.

Our analysis also finds that not only is the topic "climate change politics" disproportionately prevalent in the skeptical community, but so too is the topic "US politics." The literature finds that climate skepticism often goes together with a conservative political view. While we have not examined the political leanings of the prominent skeptical documents, the present study provides a good starting point for further investigation into this issue. Our findings show that the relationships between the hyperlink structure and the content of climate change discussions are diverse and the disagreement over climate change manifests itself in the structure of links in complex ways. This enables us to go beyond the simple question of whether or not the discussion of climate change is "polarized" and, instead, makes visible a rich set of different relationships between the structure of links and the structure of topics.

Conclusions

The research reported in this study contributes to the essay of climate change discourse on the Internet in several ways:

- We have, for the first time, charted a major part of the English-language blogs devoted to the issue of climate change and made available for analysis both the hyperlink structure connecting the 3000 blogs and the texts of the 1.3 million blog posts published on them.
- We have identified and analyzed the discursive communities in the climate change blogosphere and found that there is one predominantly skeptical community of bloggers and a number of distinctively different communities of accepters. While there are some studies of the skeptical bloggers little research on the other side exists. The present study has characterized a large portion of the whole of the climate change blogosphere, including the different communities of accepters.
- We have combined network analysis with automated methods for the analysis of the topics of our corpus of blog posts and mapped the distribution of topics over the communities of bloggers. In this way we have characterized the discursive profiles of the different communities and the topical contexts in which climate change is discussed in them. In particular, we have shown the ways in which the emphasis on, respectively, the scientific and the political aspects of climate change differs between the communities.
- We have shown that the patterns of interaction between the communities are diverse and that climate change science is the central topic with the community of accepters that interact the most actively with the skeptical community.
- While our analysis did not produce a distinctively skeptical topic we did find that the skeptical community has a distinct topical profile. We also found that

the name calling often associated with the debate between skeptics and accepters was a salient feature of the discourses across the communities.

Limitations and future work

The present study has limitations at various levels. First, we have not been able to determine precisely the extent to which the units identified by the community detection algorithm constitute communities in a social sense. It is a plausible assumption that patterns of linking reflect relationships that are in some sense social, but it is difficult to know exactly why people choose to link or not to link with each other. The fact that we find that one community is predominantly skeptic seems correct in the light of what we know about the skeptics as a kind of counter-movement to the science community. Nevertheless, there are accepter blogs and neutral blogs in the skeptical community, but we do not know if this is a sign of diversity in the skeptical camp or whether it is an artifact of the community detection algorithm. A second limitation is that even though we have been able to verify that the two topics related to climate change are distinct in the sense that they point to real semantic differences in the top 100 documents, we have not been able to verify our interpretation of the other documents beyond manual inspection of a selection of documents. A central challenge for future work is to analyze how the discourse on climate change is evolving over time.

Funding

Notes

1. http://2013.bloggi.es/.
2. "Climate skeptics 'capture' the Bloggies' science category." Fred Hickman, *The Guardian*, 1 March 2013.
3. http://www.alchemyapi.com/.
4. http://mallet.cs.umass.edu/.

References

Ackland, R., & O'Neill, M. (2011). Online collective identity: The case of the environmental movement. *Social Networks*, *33*, 177–190. doi:10.1016/j.socnet.2011.03.001
Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. KDD'05: Proceedings of the 3rd International Workshop on Link Discovery, Palo Alto, CA, March 4 (pp. 36–43). doi:10.1145/1134271.1134277
Adamic, L. (2008). The social link. In J. Turow & L. Tsui (Eds.), *The hyperlinked society* (pp. 227–248). Ann Arbor: University of Michigan Press.

Blei, D. M., & J. Lafferty, J. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71–93) Boca Raton, FL: Taylor & Francis.

Blei, D. M., Ng, Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning research*, 3, 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008. doi:10.1088/1742-5468/2008/10/P10008

Boykoff, M. T. (2011). *Who speaks for the climate? Making sense of media reporting on climate change*. Cambridge: Cambridge University Press.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Proceedings of the Seventh International World-Wide Web Conference (WWW 1998), Brisbane, QLD, April 14–18. Retrieved from http://www7.scu.edu.au/

Dunlap, R. E., & McCright, A. M. (2011). Organized climate change denial. In J. Dryzek, R. Norgaard, & D. Schlossberg (Eds.), *The Oxford handbook of climate change and society* (pp. 144–160). Oxford: Oxford University Press.

Eide, E., & Kunelius, R. (Eds). (2013). *Media meets climate*. Gothenburg: Nordicon Press.

Fleiss, J. L., Levin, B., & Paik, M.C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), 36–41. doi:10.1073/pnas.0605965104

Gonzales-Bailon, S. (2009). Opening the black box of link formation: Social factors underlying the structure of the web. *Social Networks*, 31, 271–280. doi:10.1016/j.socnet.2009.07.003

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer Mediated Communication*, 18(2), 40–60. doi:10.1111/jcc4.12001

Hoffman, A. (2011). Talking past each other? Cultural framing of skeptical and convinced logics in the climate change debate. *Organization & Environment*, 24(1), 3–33. doi:10.1177/1086026611404336

Hoggan, J. (2009). *Climate cover up: The crusade to deny global warming*. Vancouver: Greystone Books.

Koteyko, N., Jaspal, R., & Nehrich, B. (2012). Climate change and 'climategate' in online reader comments: A mixed methods study. *The Geographical Journal*, 179(1), 74–86. doi:10.1111/j.1475-4959.2012.00479.x

Lambiotte, R., Delvenne, D.-C., & Barahona, M. (2009). Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv:0812.1770. Retrieved from http://arxiv.org/abs/0812.1770

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E*, 84(6), 066122. doi:10.1103/PhysRevE.84.066122

Lusher, D., & Ackland, R. (2011). A relational hyperlink analysis of an online social movement. *Journal of Social Structure*. Retrieved from http://www.cmu.edu/joss/content/articles/volume12/Lusher/

Mann, M. (2012). *The Hockey Stick and The Climate Wars*. Columbia University Press.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103, 8577–8582. doi:10.1073/pnas.0601602103

Oreskes, N., & Conway, E. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. New York, NY: Bloomsbury.

Painter, J. (2011). *Poles apart. The international reporting on climate scepticism*. Oxford: Reuters Institute for the Study of Journalism.

Pearce, F. (2010). *The climate files: The battle for the truth about global warming*. New York, NY: Random House.

Rahmstorf, S. (2004): *The climate sceptics*. Potsdam: Potsdam Institute for Climate Impact Research.

Rogers, R., & Marres, N. (2000). Landscaping climate change: A mapping technique for understanding science and technology debates on the World Wide Web. *Public Understanding of Science, 9*, 141–163. doi:10.1088/0963-6625/9/2/304

Schäfer, M. (2012). Online communication on climate change and climate politics: A literature review. *WIREs Climate Change, 3*, 527–543.

Sharman, A. (2014). Mapping the climate change blogosphere. *Global Environmental Change, 26*, 159–170. doi:10.1016/j.gloenvcha.2014.03.003

Shumate, M., & Dewitt, L. (2008). The north/south divide in NGO hyperlink networks. *Journal of Computer Mediated Communication, 13*, 405–428.

Sunstein, C. (2006). *Infotopia*. Oxford: Oxford University Press.

Sunstein, C. (2007). *Republic.com 2.0*. Princeton, NJ: Princeton University Press.

Washington, H., & Cook, J. (2011). *Climate change denial*. Oxford: Earthscan.