Creating and Using Topically-Focused Blog Corpora:
experiences in the NTAP project (2012-2017)

Andrew Salway

Uni Research, Bergen, Norway

Motivation

Blogs, along with other social media, are seen to be changing the public sphere, raising questions about:

- Democractic participation
- Information diffusion
- Polarization and the fragmentation of the public sphere
- Relationship with traditional news media

➔ Create a corpus containing all posts from all blogs in a chosen language that relate to a chosen topic: text, link and date data

Concepts / definitions

Blog

- a website that mentions a specified blog platform in its url
- platforms chosen by examining search engine results

Topic

- specified by a few generic terms
- a blog is considered topical if it has at least a small number of posts containing one or more of these terms

Language

- we seek to control the language of each corpus but not to associate blogs with nationalities or language varieties
- e.g. an English-language corpus contains US, Australian, Indian varieties, etc., and bloggers of any nationality including non-native speakers

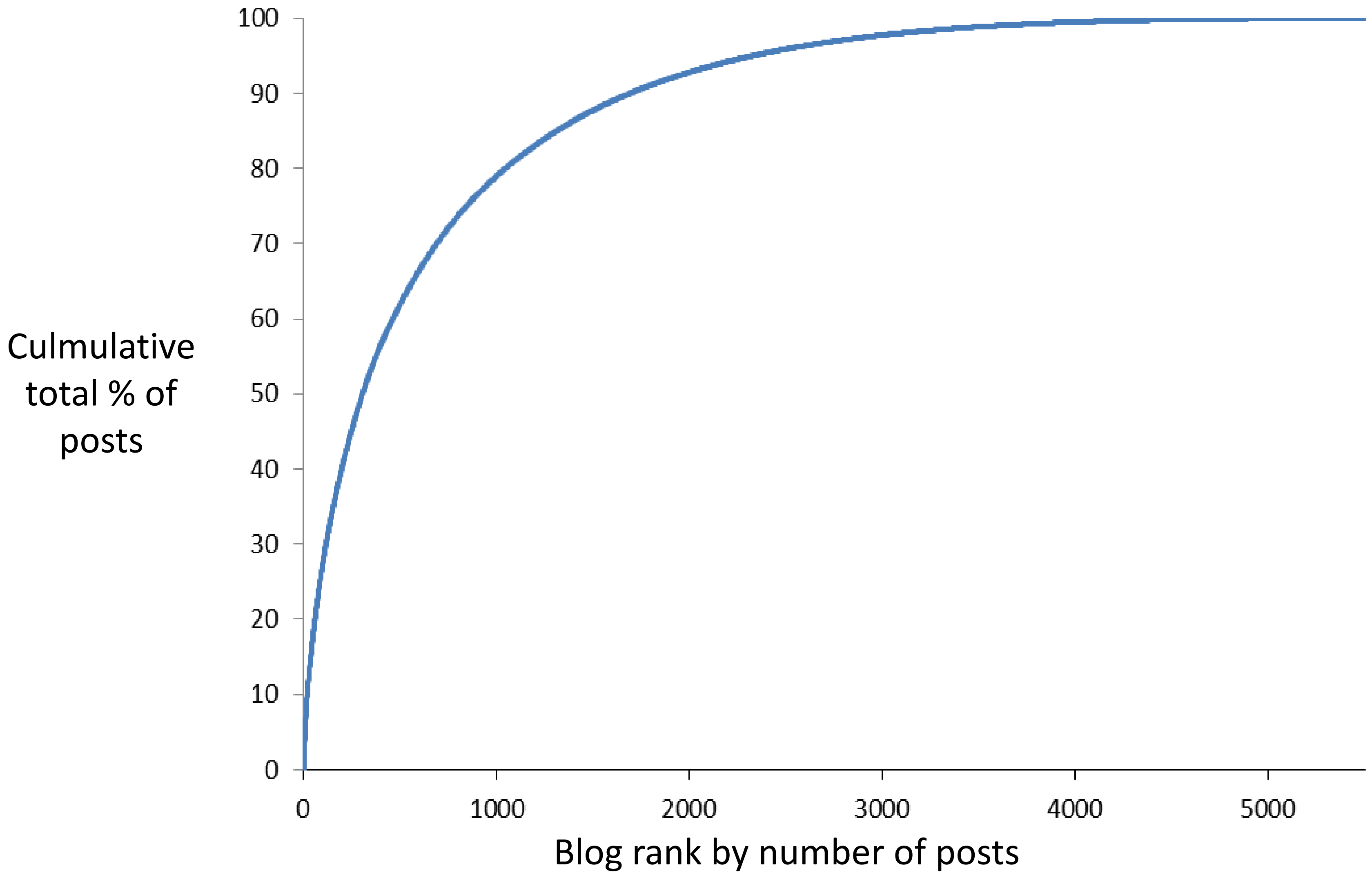
Pipeline

1. Identify relevant blogs
2. Harvest and de-duplicate blog posts
3. Text extraction from html blog posts
4. Boilerplate removal from extracted text
5. Identify language of each post
6. Link extraction (text links and blog roll links)
7. Date extraction

Three “climate change” blog corpora

	Blogs	Posts	Words
English	5497	10,539,575	4,837,481,377
French	2033	2,335,174	1,224,657,286
Norwegian	126	46,775	21,212,686

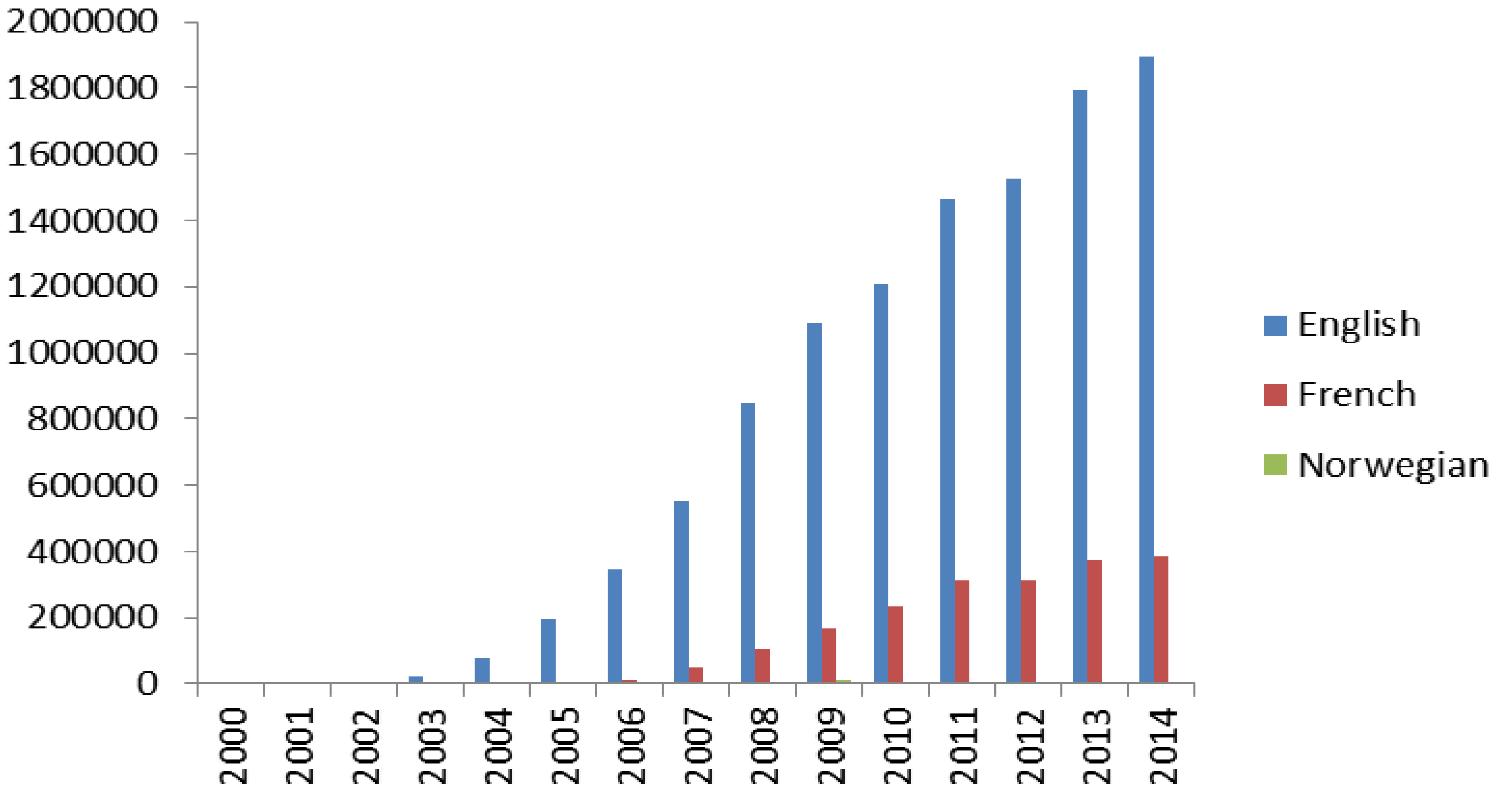
A few blogs with lots of posts



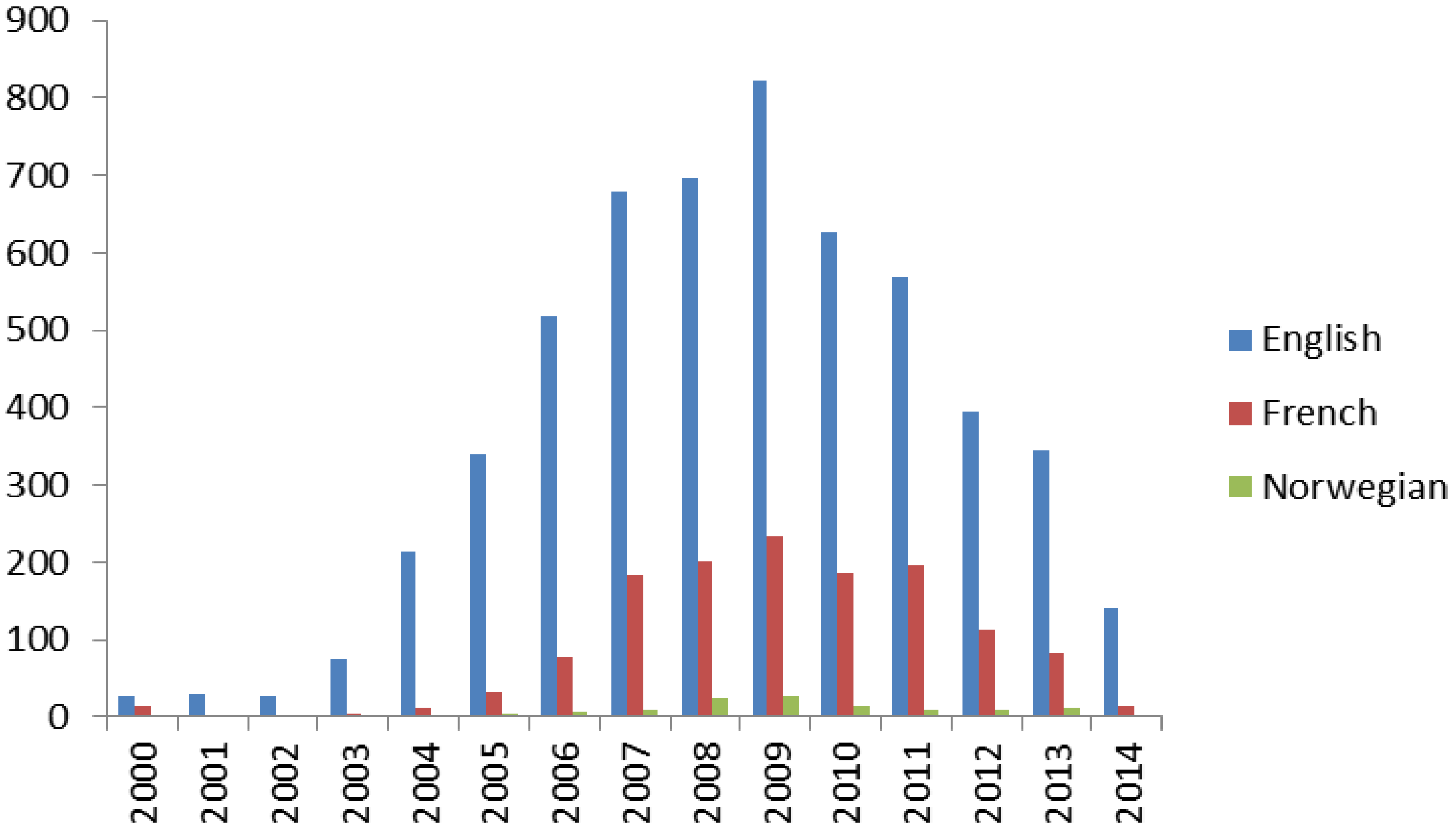
Topical content

	Frequency	% blogs	% posts	% pwc
100 climate terms	**6,837,623**	**99.55**	**11.83**	**25.26**
“climate change”	1,486,549	96.5	4.8	11.6
“global warming”	900,918	96.1	3.3	8.7
“greenhouse effect”	28,129	47.6	0.1	0.6

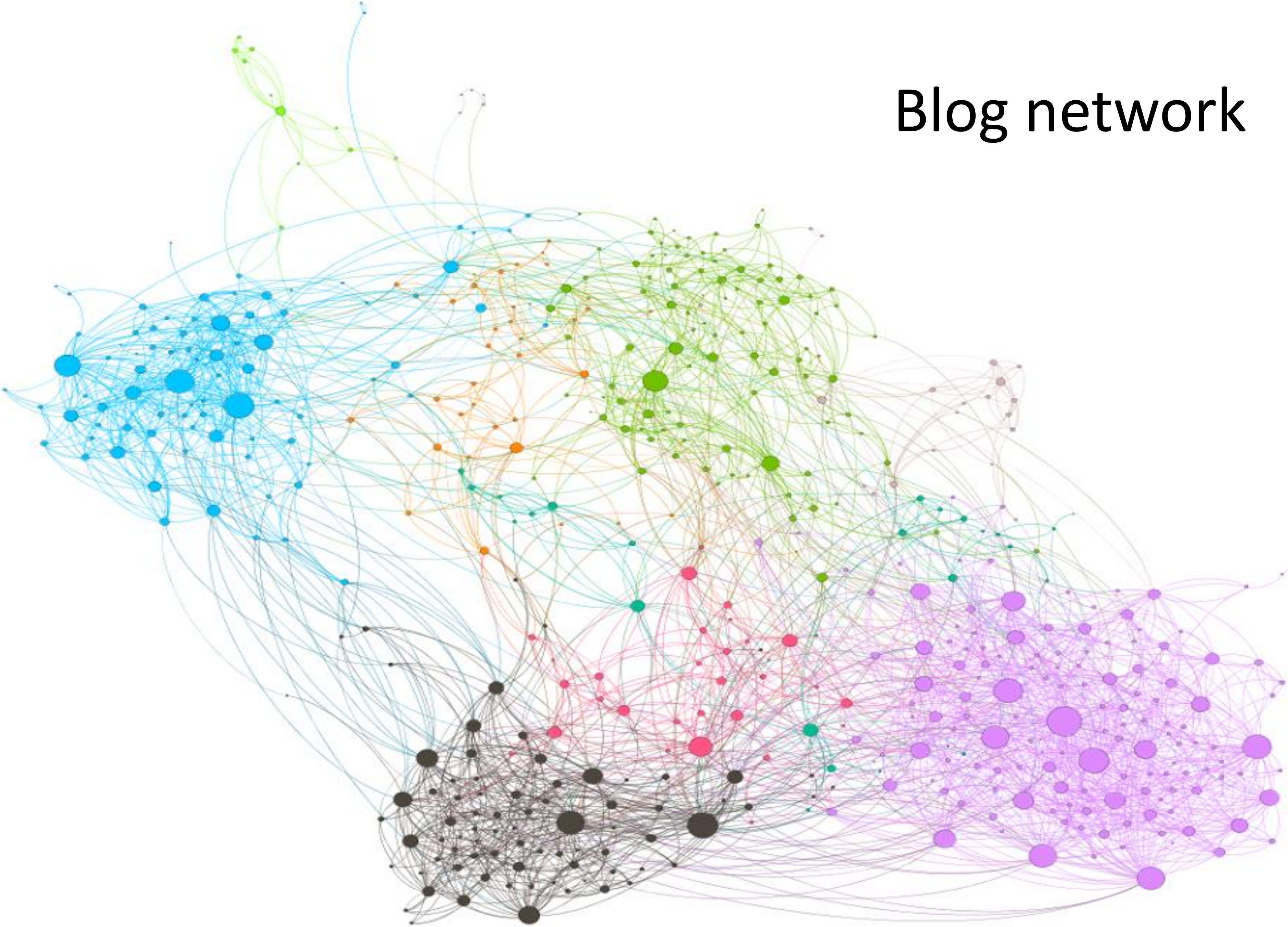
Temporal distribution of posts



Year of earliest post



Blog network



Use of the blog corpora

- To analyze the degree of polarization in the climate change blogosphere (Elgesem et al., 2015)
- To induce representations of the future, and compare between acceptor and sceptic blogs (Fløttum et al., 2014; Salway, 2017)
- To compare the use of quotations in acceptor and sceptic blogs (ongoing).

References

Salway, A., Elgesem, D., Hofland, K., Reigem, Ø. and Steskal, L. 2016. **Topically-focused Blog Corpora for Multiple Languages.** Procs. 10th Web as Corpus (WAC-X), ACL 2016, 17-26. http://aclweb.org/anthology/W16-26

Salway, A. 2017. **Data-driven approaches to climate change discourse.** In K. Fløttum ed., *The Role of Language in the Climate Change Debate*, Routledge.

Elgesem, D., L. Steskal and N. Diakopoulos (2015). **The structure and content of the discourse on climate change in the blogosphere: the big picture.** *Environmental Communication, 9 (2)* Special issue on climate change communication on the Internet, pp. 169-188.

Fløttum, K., Gjerstad, Ø., Gjesdal, A.M., Koteyko, N. and Salway, A. (2014). **Representations of the future in English language blogs on climate change.** *Global Environmental Change* 29, 213–222.

Language technology vs social science
- experiences from NTAP, exaggerated for effect

Creating the blog corpora

- Get as many blogs as possible **vs** clear definition of concepts (blog, topic, language)
- Systematic/algorithmic selection of blogs **vs** adding missing blogs that are known to be important
- No 100% solution for duplicate posts and boilerplate **vs** impact on research results
- Many non-trivial technical issues **vs** impatience for data

Using the blog corpora

- What can be operationalized for automatic analysis **vs** all the concepts that the researchers would like to analyse

Further remarks

- Corpora to be made available “as they are”; including data about suspicious 5-grams, posts in wrong language, and blogroll links, so researchers can adapt for their own purposes
- Can’t make strong claims about getting “all” blogs
- >1 >1 criterion is too permissive?
- A blog must have platform in name!?
- Role of search engines, w.r.t. replicability and transparency
- Use links to crawl? Requires manual intervention – because lots of other sites with have high in-degree; or, intensive harvesting and analysis of many irrelevant sites

How much of the climate change blogosphere was captured?

Assume in-links reflect a blog's importance...

Based on links from harvested corpus:

For blogs with specified platforms in names, the corpus contains 22 of the 25 most important blogs ranked by number of in-links; missing 3 are about politics in general

Based on all blogroll links from 6 blogs known to be important: 71 blogs not in corpus; 7 of these have blog platform in name; of the other 64, 17 have in-degree from corpus > 25; seems that relatively more important blogs do not have platform in name