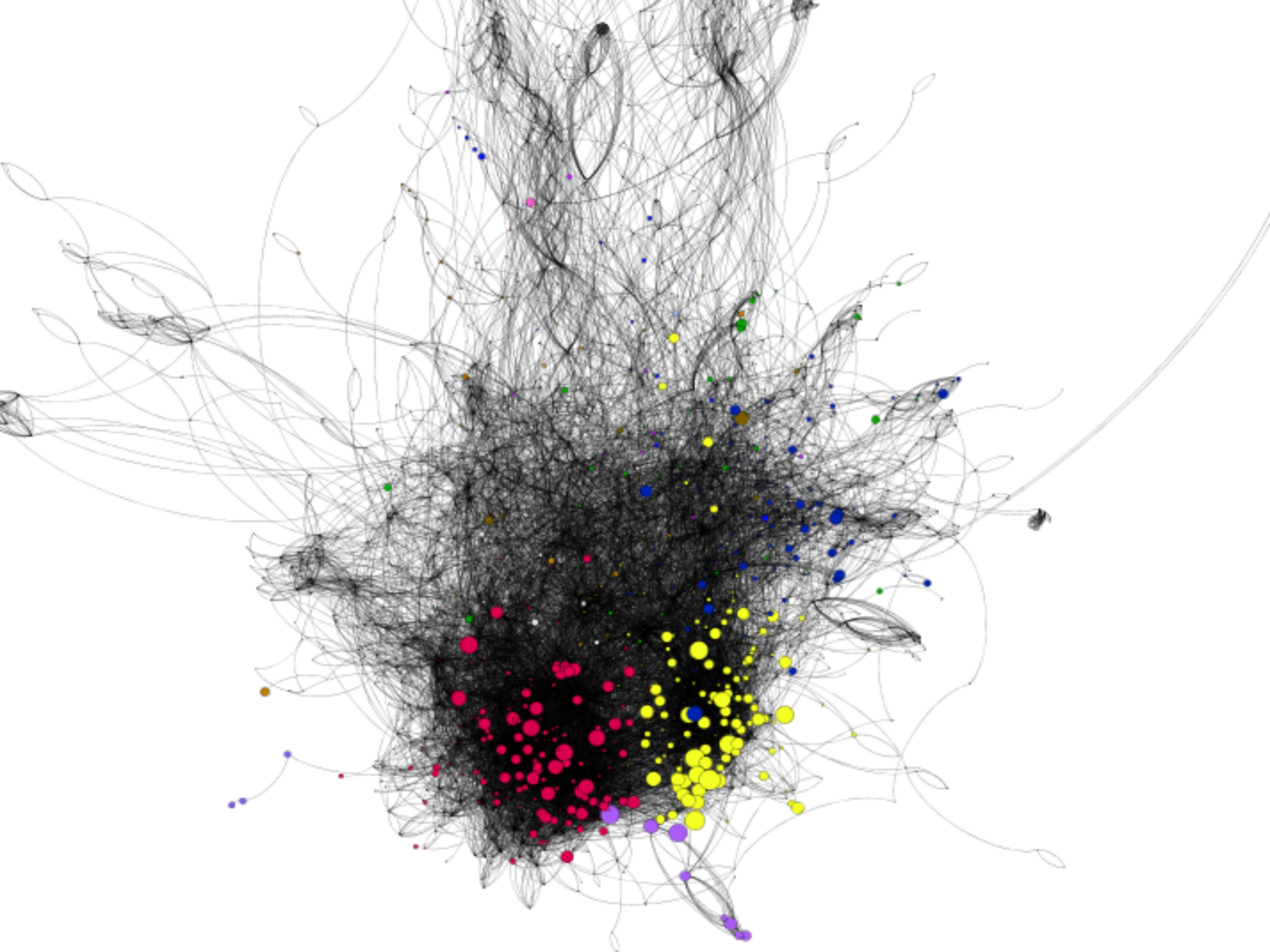
Topically-focused Blog Corpora for Multiple Languages

Andrew Salway[1], Dag Elgesem[2], Knut Hofland[1], Øystein Reigem[1], Lubos Steskal[2]

[1]Uni Research, Bergen, Norway

[2] University of 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

- PROBLEM: a lack of commonly available large-scale blog corpora to support empirically-grounded social science research

Task definition

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

Approach

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

Approach

- 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

Approach

- 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 may contain US, Australian, Indian varieties, etc., and bloggers of any nationality including non-native speakers

Pipeline

1. Identify relevant blogs
2. Harvest and de-duplicate
3. Text extraction
4. Boilerplate removal
5. Identify posts in wrong language
6. Link extraction
7. Date extraction

1. Identify relevant blogs

- Daily querying of search engine APIs for “TERM + BlogPlatform + Language”, for 12 weeks
- After 2 weeks expanded set of search terms with n-grams containing search terms and a function words, e.g. “of climate change”

English (WordPress, BlogSpot, TypePad) climate change, global warming, greenhouse effect → 95,662 posts
French (WordPress, BlogSpot, OverBlog) changement climatique, changements climatiques, réchauffement climatique, effet de serre, effets de serre → 68,853 posts
Norwegian (WordPress, BlogSpot, TypePad) drivhuseffekt, drivhuseffekten, global oppvarming, globale oppvarmingen, klimaendring, klimaendringen, klimaendringene, klimaendringer, klimaforandring, klimaforandringen, klimaforandringene, klimaforandringer → 8,973 posts *(after Danish removed)*

Search terms	Total posts	Blogs >0 post	Blogs >1 post	Blogs >2 posts	Blogs >3 posts
	English				
Total	84536	27873	7205	3995	2762
>0		25190	6515	3541	2391
>1		21231	**5563**	2998	2042
>2		18007	4633	2493	1730
	French				
Total	52029	13838	4552	2716	1931
>0		12732	3926	2217	1526
>1		6470	**2088**	1213	845
>2		4187	1318	754	512
	Norwegian				
Total	7194	613	505	293	224
>0		1393	337	172	119
>1		470	**128**	65	42
>2		268	67	26	18

2. Harvest and de-dupicate

- Harvesting script iteratively customised for each blog platform
  - Follow links to archive pages
  - Test all urls for characteristics of blog post
  - Use ftfy to resolve encoding issues
- Normalization for de-duplication
  - Standardize url format and use look-up table for alternative domain names
  - 3.3%-4.8% posts had duplicates in the three corpora

3. Text extraction

- High quality extraction important for social science research, so developed own tool that exploited blog platform-specific html characteristics
  - Iteratively developed heuritstics, based on html clues, to identify start and end of «main text»
  - Strip html tags and sequences except link, paragraph and break markers
  - Start and end clues matched for 99.7% of all posts
  - Evaluated 1463 English-language posts, all from different blogs: 11 posts (<1%) with part of main text missing; 72 (5%) with inappropriate text at start, 1 in middle, 48 (3%) at end

4. Boilerplate removal

- Within the main text there can be near-duplicate paragraphs on many posts in a blog, e.g. a slogan for the blog, contact details, request for donations
  - Identify suspicious 5-grams for a blog, i.e. all 5-grams appearing on >15% of that blog's posts
  - Any paragraph in which >50% of the words are part of suspicious 5-grams is marked up as likely boilerplate
  - Boilerplate was found in about 20% of all posts
  - Evaluation showed 23 (9%) of 258 posts with boilerplate marked up had some text incorrectly marked as boilerplate

5. Identify posts in wrong language

- Possible issues: a blogger quotes or cites from another language; bilingual blogs
  - Use langid.py to record a Boolean value for each post to show whether it is considered most likely to be in the language of its corpus
  - Posts in correct language: English 96%, French 81%, Norwegian 89%

6. Link extraction

- Article links, per post:
  - all links found in the «main text»; these are marked-up in text, normalized and stored in a set per post
- Blogroll links, per blog:
  - for all links occurring outwith main text, store the % of posts each occurs on within a blog

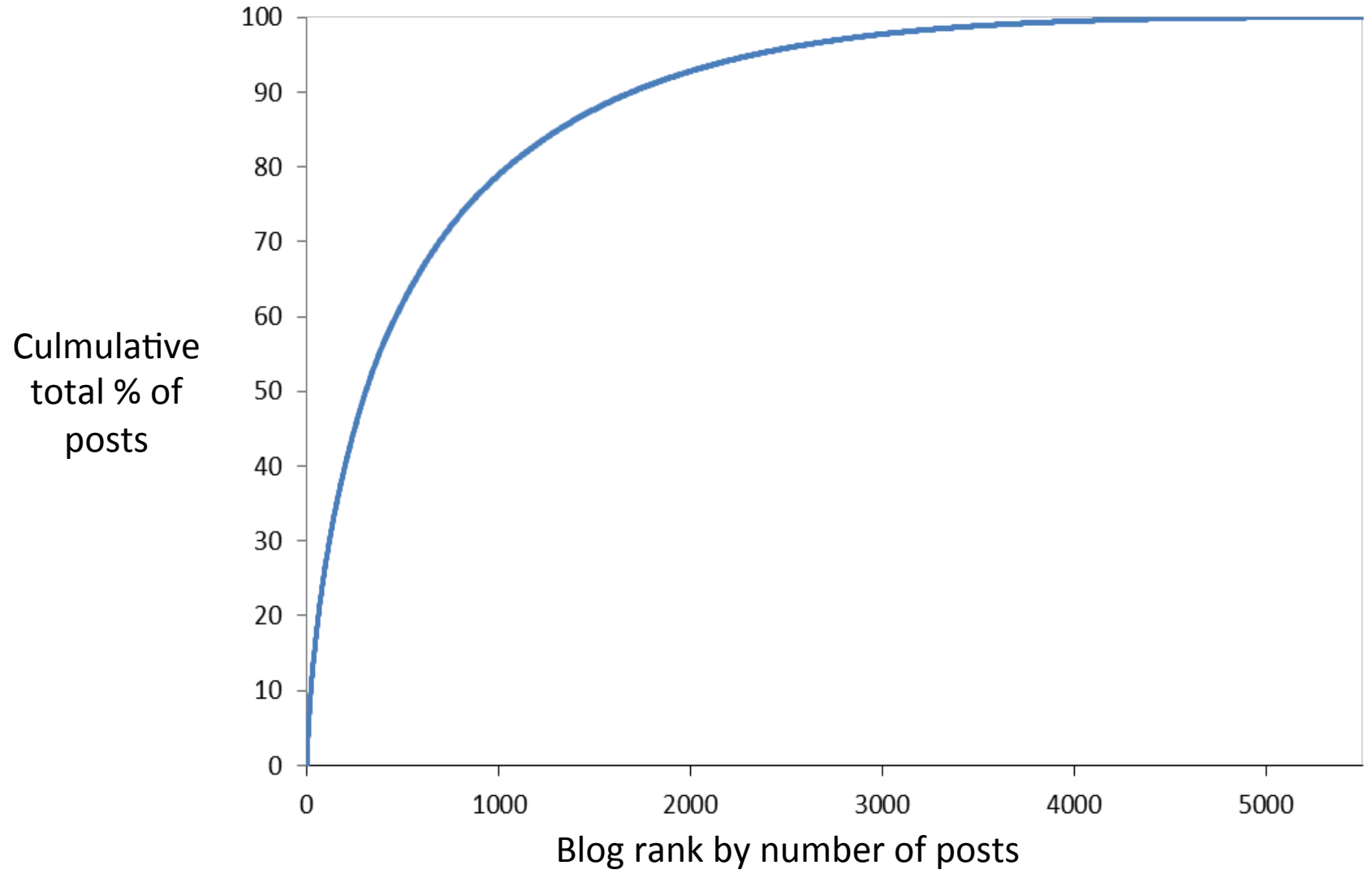
7. Date extraction

- Date data available, to some extent, from urls and from html:
  - MM and YYYY values available in the url of each post, except some OverBlog blogs; WordPress also gives DD in url
  - Developed heuristics to get date data from html for OverBlog blogs

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



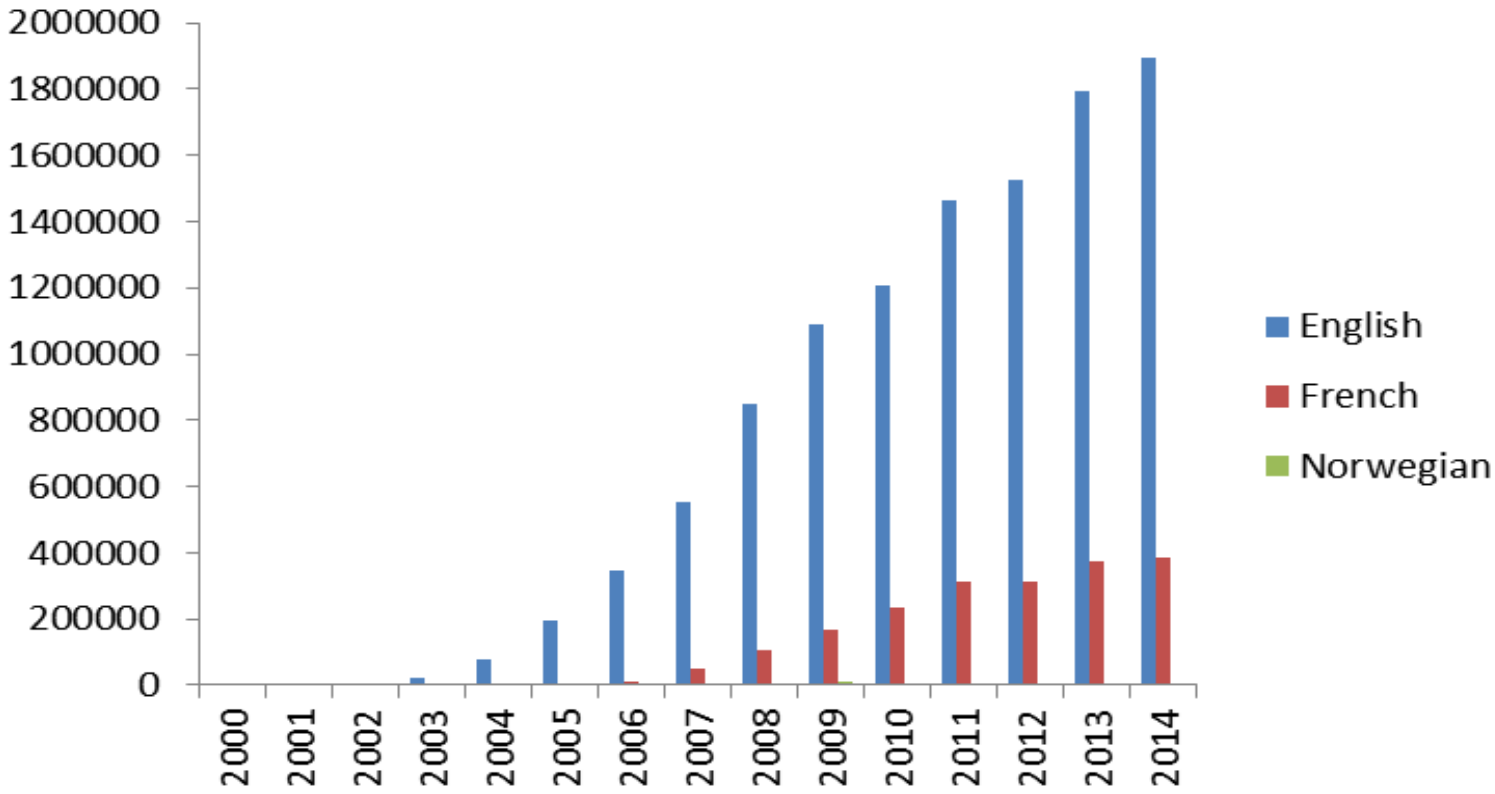
Topical content

al gore, antarctic ice, arctic ice, arctic sea, atmospheric co2, average temperature, carbon dioxide, carbon emissions, carbon footprint, clean energy, climate action, climate change, climate crisis, climate models, climate science, climate scientist, climate system, co2 emissions, co2 levels, computer models, coral reefs, developed countries, developing countries, dioxide emissions, earth's atmosphere, earth's climate, el nino, el niño, emissions trading, energy consumption, energy efficiency, energy efficient, energy policy, energy sources, environmental issues, environmental protection, extreme weather, fossil fuel, fossil fuels, framework convention, future generations, gas emissions, global climate, global cooling, global temperature, global warming, greenhouse effect, greenhouse gas, heat waves, hockey stick, human activity, human beings, ice caps, ice sheet, inconvenient truth, intergovernmental panel, ipcc report, kyoto protocol, nitrous oxide, nuclear energy, ocean acidification, ozone layer, polar bear, polar ice, population growth, reduce emissions, renewable energy, rising sea, rising temperatures , scientific community, scientific consensus, sea ice, sea levels, sea surface, sea-level rise, solar activity, solar energy, solar panels, solar radiation, surface temperature, sustainable development, tar sands, temperature data, temperature increase, temperature rise, un climate, warming trend, weather events, weather patterns, wind energy, wind farms, wind turbines

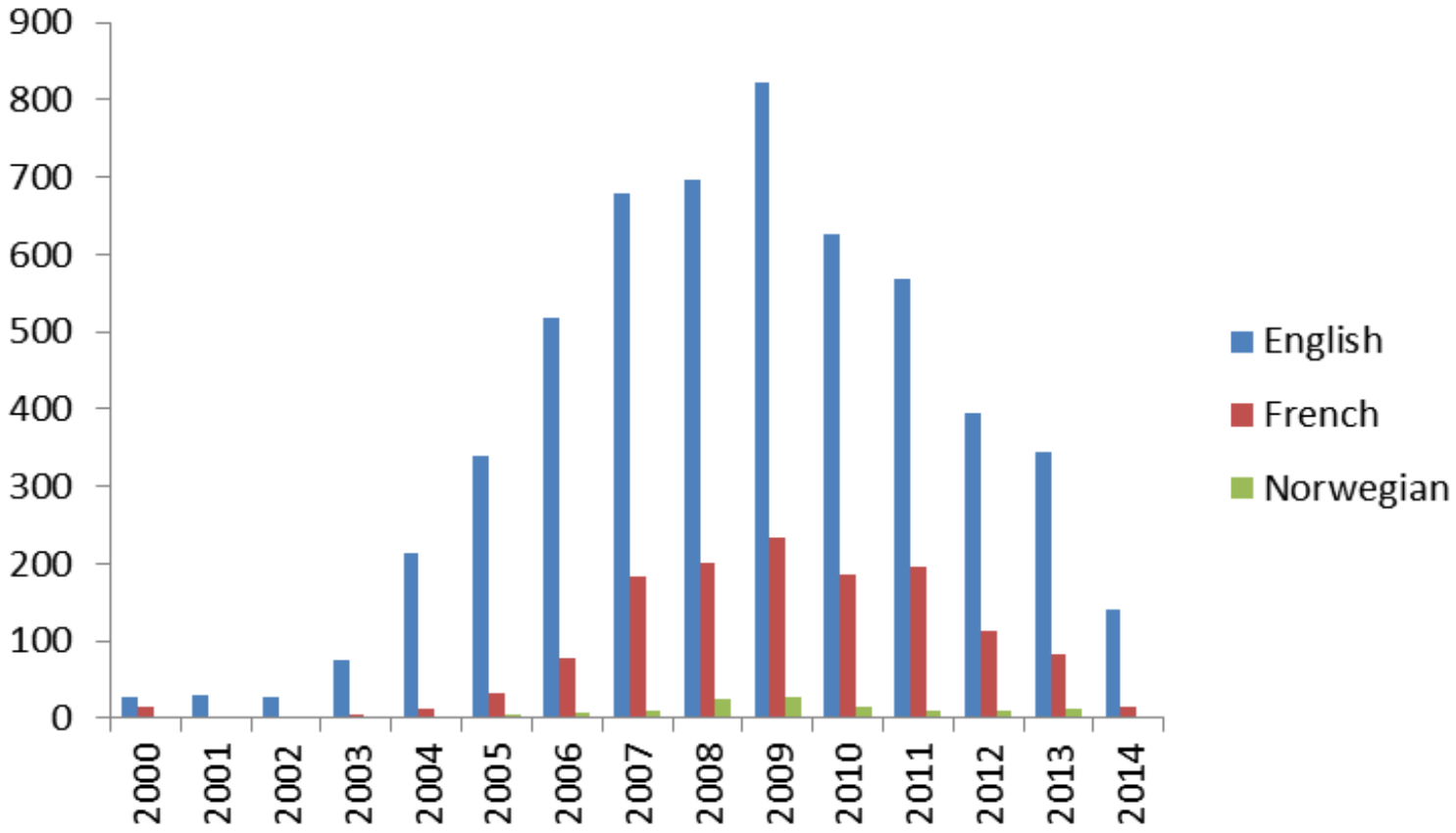
Topical content

	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

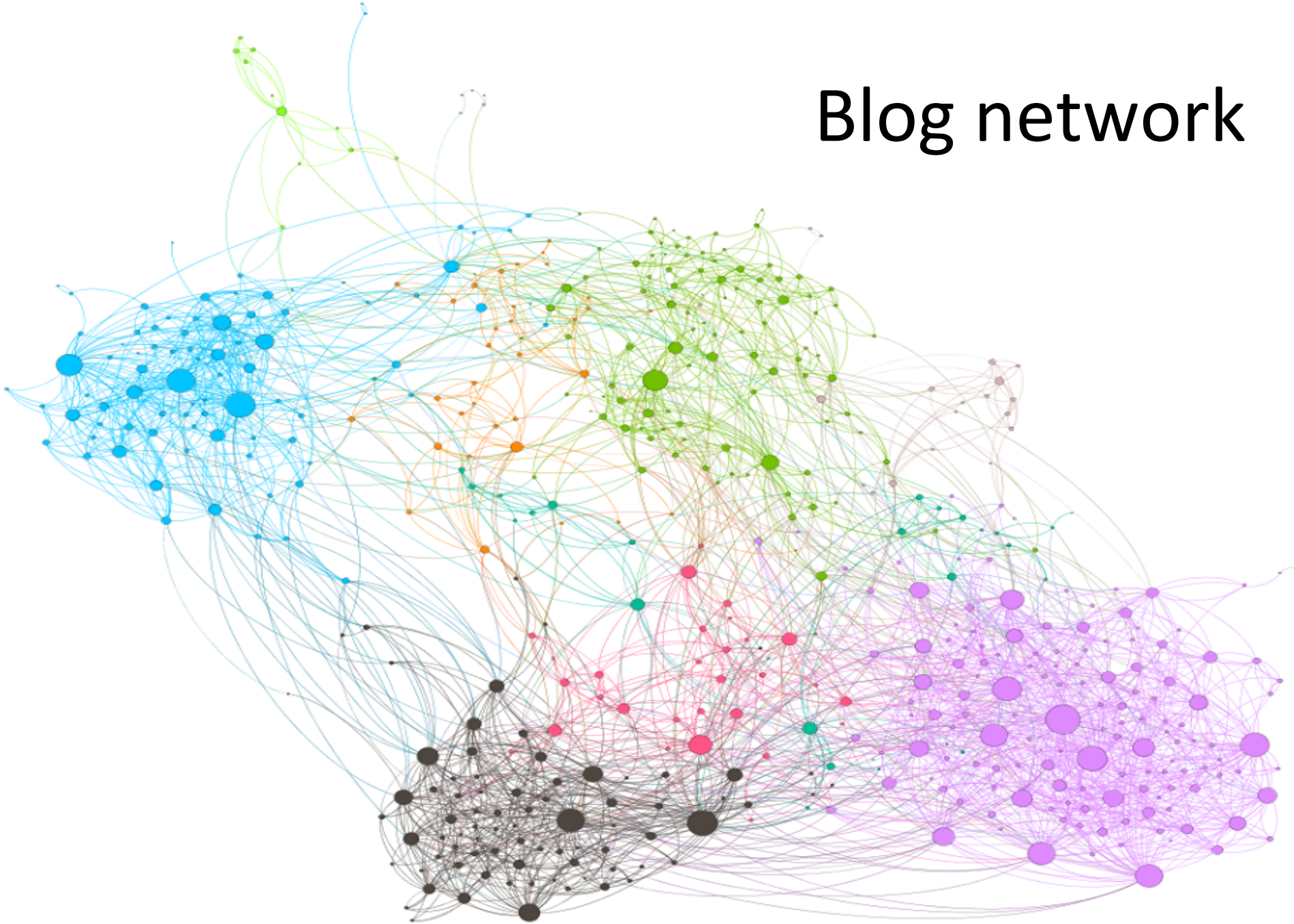
Temporal distribution of posts



Year of earliest post



Blog network



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

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

Closing remarks

- Corpora to be made available “as they are”; including data about suspicious 5-grams, wrong language, blogroll links, so researchers can adapt for their purposes

Closing remarks

- Corpora to be made available “as they are”; including data about suspicious 5-grams, wrong language, blogroll links, so researchers can adapt for their purposes
- Can’t make strong claims about getting «all»

Closing remarks

- Corpora to be made available “as they are”; including data about suspicious 5-grams, wrong language, blogroll links, so researchers can adapt for their purposes
- Can’t make strong claims about getting «all»
- >1 >1 criterion is too permissive?

Closing remarks

- Corpora to be made available “as they are”; including data about suspicious 5-grams, wrong language, blogroll links, so researchers can adapt for their purposes
- Can’t make strong claims about getting «all»
- >1 >1 criterion is too permissive?
- A blog must have platform in name!?

Closing remarks

- Corpora to be made available “as they are”; including data about suspicious 5-grams, wrong language, blogroll links, so researchers can adapt for their purposes
- Can’t make strong claims about getting «all»
- >1 >1 criterion is too permissive?
- A blog must have platform in name!?
- Role of search engines, w.r.t. replicability and transparency

Closing remarks

- Corpora to be made available “as they are”; including data about suspicious 5-grams, wrong language, blogroll links, so researchers can adapt for their purposes
- Can’t make strong claims about getting «all»
- >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