“Big data” workshop

Digital Reading Network
6th March 2014, Sheffield

Andrew Salway, Uni Research, Bergen
Daniel Allington, The Open University

Overview

- What does big data mean (for social science and humanistic research)?
- Examples of big data research
- How can we use big data for research on digital reading?

What does “big data” mean?
(for social science and humanistic research)

Daniel's perspective…

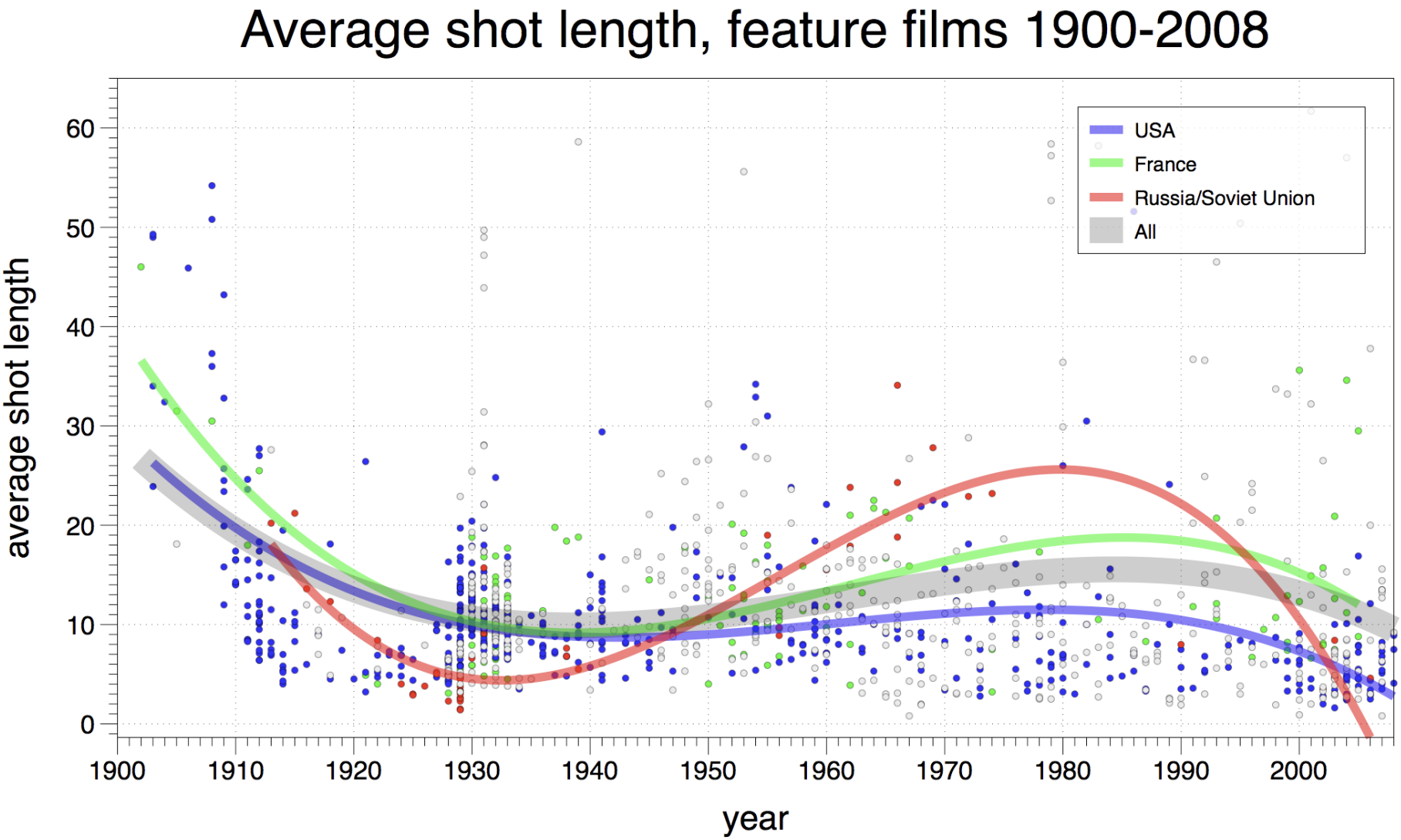
- In computer science: any dataset that exceeds the limits of commonly used tools
  - Cloud computing, Hadoop clusters etc
- In the humanities: 'any study with n > 50 is in danger of being described as an example of "big data."' (Underwood, 2013)
  - Note that he's talking about literary studies; in history, 50 is not necessarily a big number
- In the social sciences: whole population studies instead of e.g. sample surveys

Andrew's perspective…

- Emphasis on use of computers for analysis, rather than for databases and text annotation
- Exploratory, data-driven, inductive
  - Observe phenomena that are not present, or detectable, in smaller samples
  - Observations prompt new lines of theorising and explanation
  - Often involves visualization; (new ways of seeing, cf. microscopes and telescopes?)
- Size of data set necessitates automated analysis, but always a researcher "in the loop"
- May integrate the analysis of different kinds of data: text, network, image, numeric…

Examples of big data research

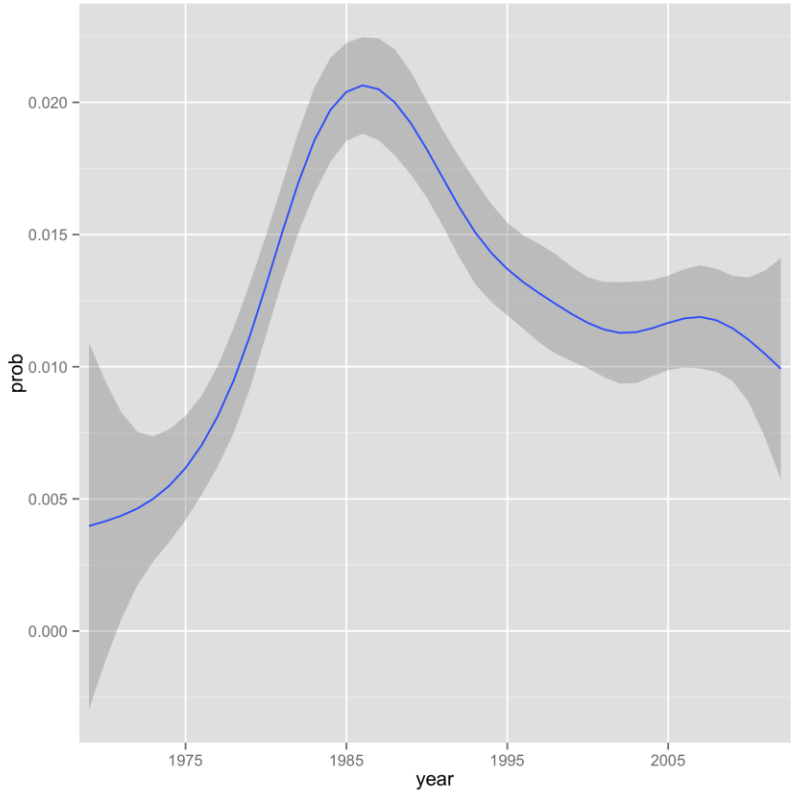
“Cultural analytics” (Manovich et al)



Distant reading

- Term coined and promoted by Franco Moretti
- Implied opposition to the ‘close reading’ of the New Critics
- Studying large volumes of texts without ‘reading’ them in the usual sense
- Analysis of titles, bibliographic data, genres (as identified by ‘experts’)

Distant reading of readings



'social class theory ideology political production ideological historical marxist marx bourgeois capitalist society capitalism marxism economic labor relations capital' in *New Literary History*, *Critical Inquiry*, *boundary 2*, *Diacritics*, *Cultural Critique*, and *Social Text*

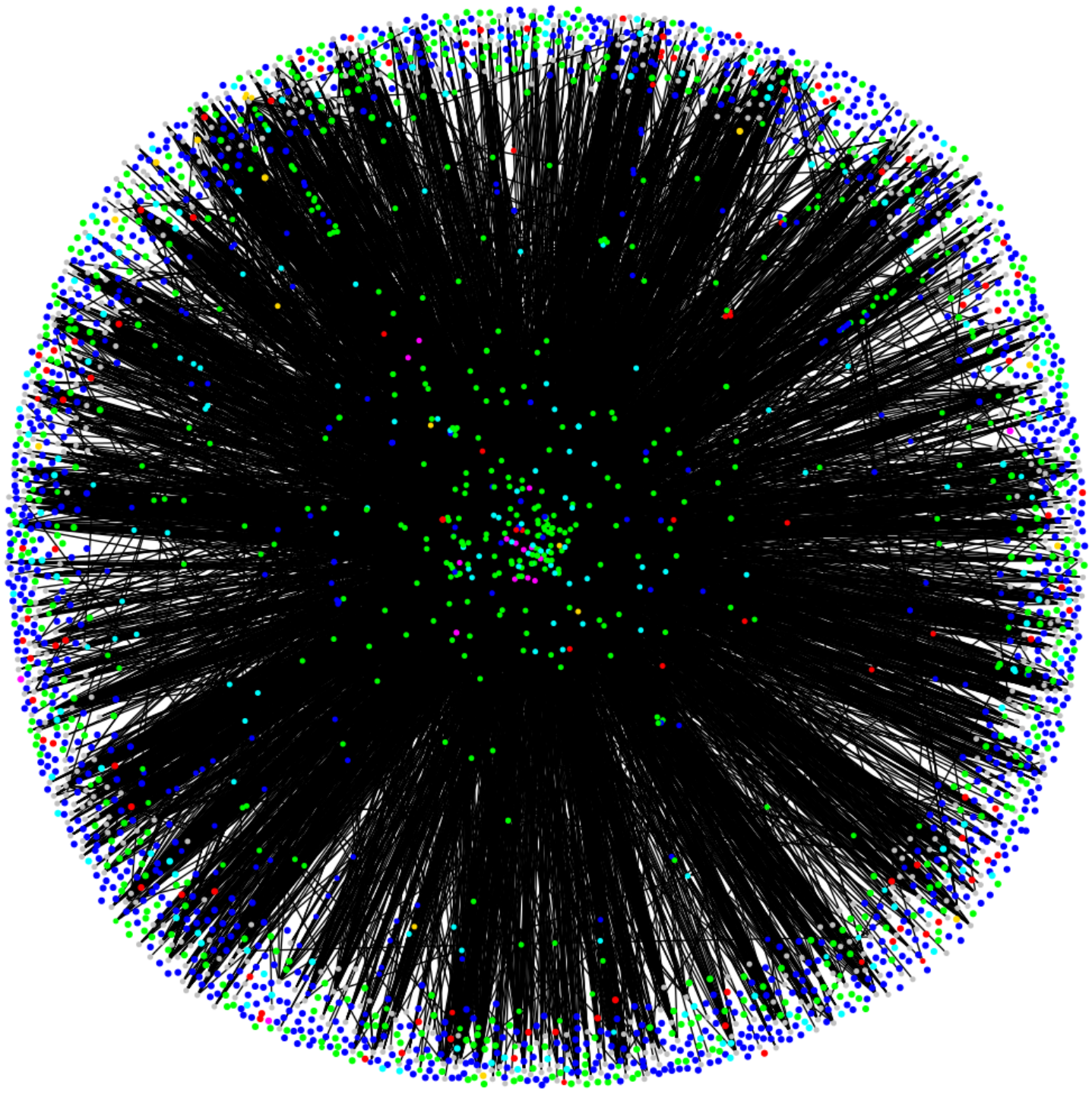
Jonathan Goodwin (2012)

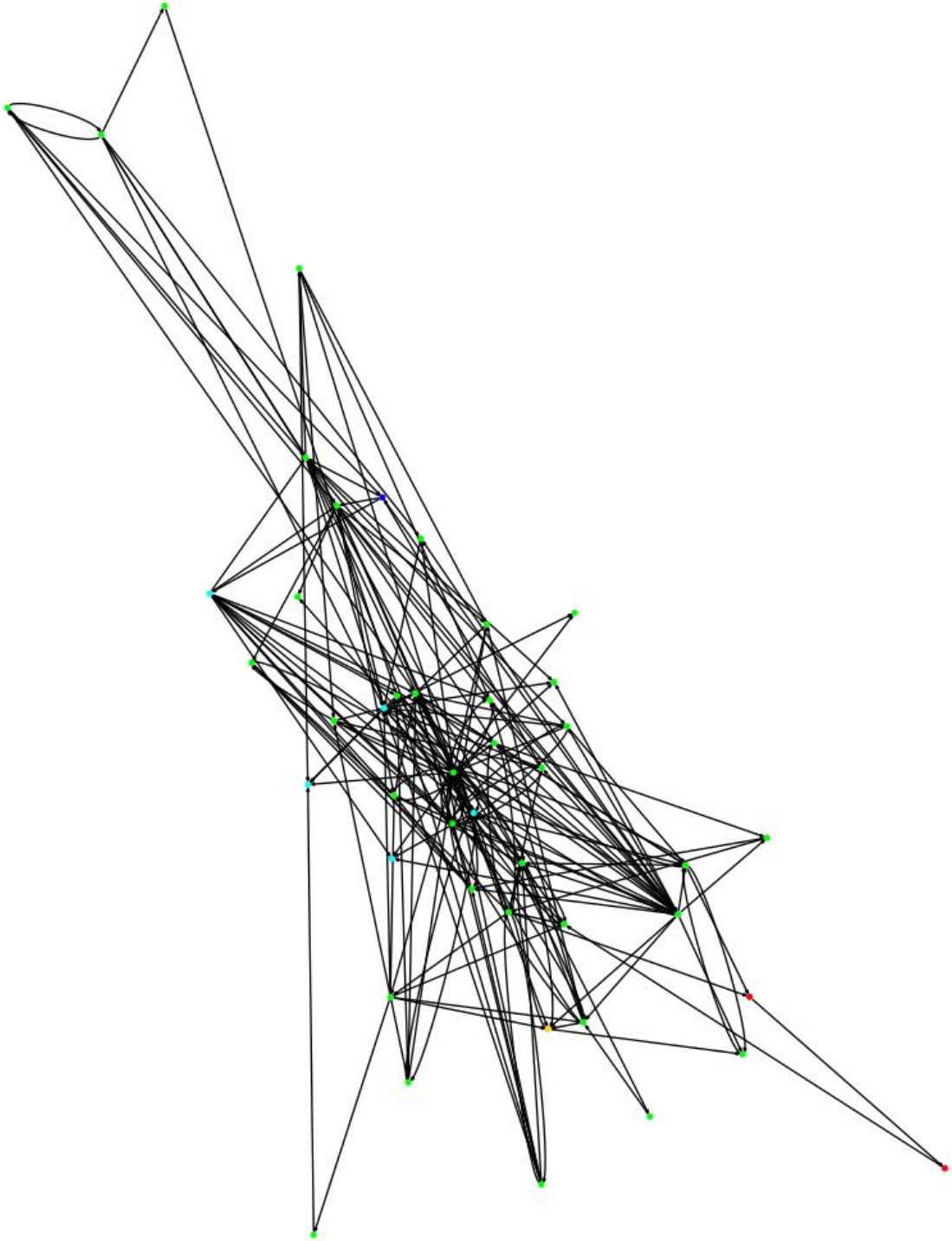
Social network analysis of users of the Interactive Fiction Database

- Qualitative impression that there were three ‘star’ authors, and that each had a ‘fan club’
- One authoring system had its own community of authors and readers, and was disparaged by IF enthusiasts outside that community

Social network analysis of users of the Interactive Fiction Database

- Expected to find a network with well-defined clusters
- Instead, found a core-periphery structure
- Three ‘stars’ at the very centre (not within individual clusters)
- Sub-component organised around disparaged authoring system was weakly connected and peripheral
- Parallels with Anheier and Gerhards (1991)





Climate change discourse in the blogosphere

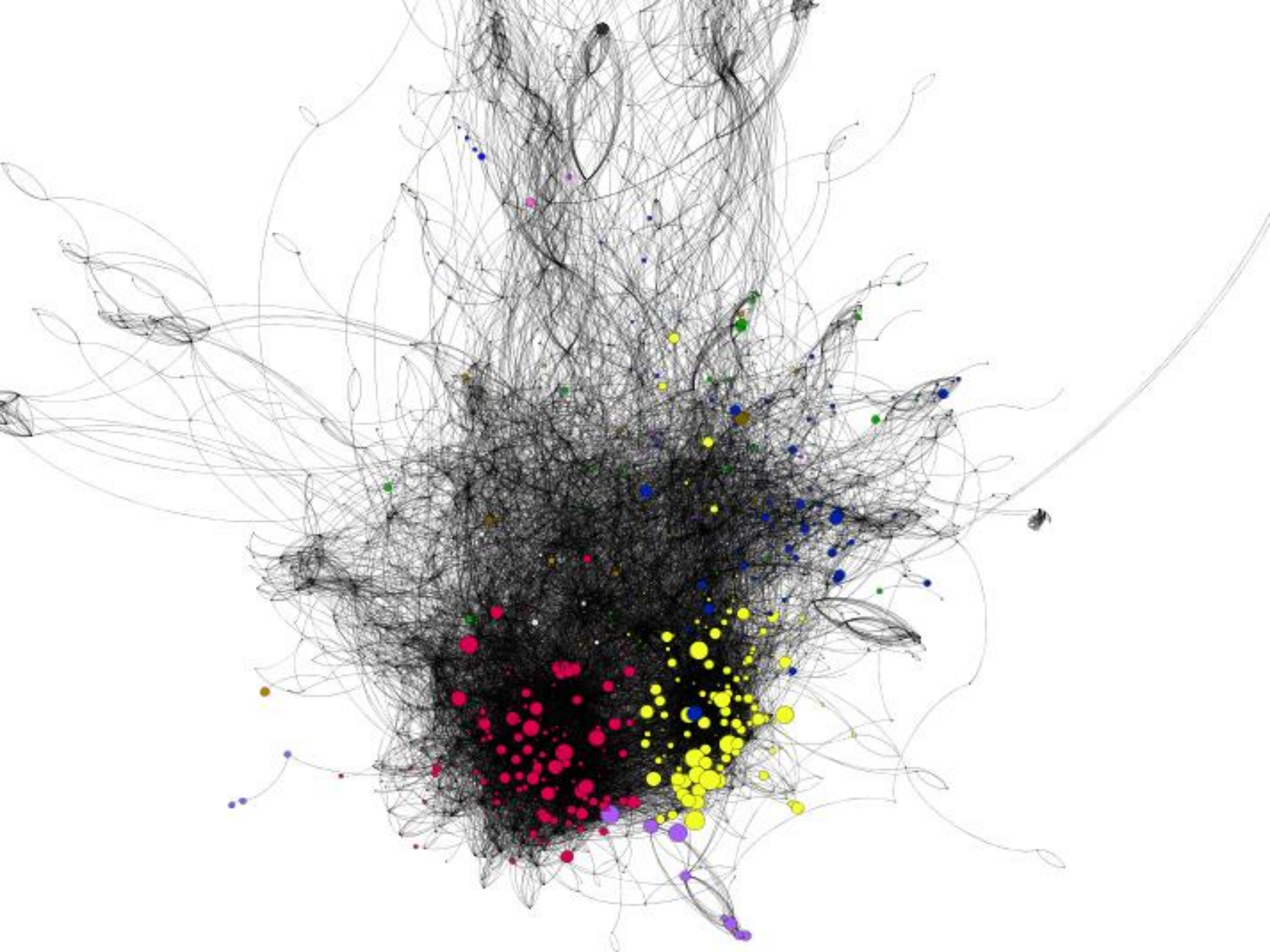
- In the NTAP project we created a corpus of English-language blogs that mention broad climate change issues across science, politics, environment, etc.
- Harvested the complete content of about 3,000 English-language blogs, about 1.5m blog posts and 100,000's links between blogs (crawl carried out June-September 2012)

Inducing blog communities with network analysis

- Used a community detection algorithm based on modularity maximization
  - blogs grouped to maximize inter-group hyperlinking (Louvain method, implemented in the Gephi tool)
- This suggested 11 major communities in the corpus, accounting for about 60% of all blogs

Characterising language use with topic modeling

- Latent Dirichlet Allocation (LDA) used to identify topics within the corpus, using MALLET tool
- Two (out of 20) topics related strongly to climate change:
  - "climate change science": *climate, warming, global, change, ice, data, temperature, years, science, scientists, carbon, sea, earth, year, ocean, time, temperatures, scientific, research*
  - "climate change politics": *climate, change, countries, world, environmental, international, development, global, emissions, carbon, india, environment, people, government, nations, policy, china, issues, sustainable*
- Other topics: "energy", "wildlife", "legal", "education", "economic policy", "transportation", "American politics", "storms and floods", "farming", "health", "new age", and some noise (incoherence topics and non-English words)



Sub-corpora for two communities

- A selection of blogs were inspected to manually code each community as “accepting”, “skeptical” or “neutral” regarding anthropogenic global warming

➔ sub-corpora

  – “accepting” 204 blogs, 69k posts, 27m words
  – “skeptical” 417 blogs, 290k posts, 127m words

Sub-corpora for two communities

	"accepting" (204 blogs, 69k posts)		**"skeptical" (417 blogs, 290k posts)**	
	no. of blogs	no. of posts	no. of blogs	no. of posts
acidification	64	1782	75	412
coral	70	939	117	701
ocean	122	4627	206	6550
species	112	3455	196	5319
tax	100	2339	213	17,180
Gore	84	845	200	10,669

Sub-corpora for two communities

	"accepting" (204 blogs, 69k posts)		**"skeptical" (417 blogs, 290k posts)**	
	no. of blogs	no. of posts	no. of blogs	no. of posts
climate science	99	2115	155	3551
anthropogenic c c	55	349	79	360
human-caused c c	37	124	57	201
human-induced c c	33	144	57	273
man-made c c	31	73	97	566
climate change denial	38	165	41	109
climate alarmist	6	6	47	216

Sub-corpora for two communities

Mentions of causes of climate change

	"skeptical" (35k instances of climate change)	**"accepting" (22k instances of climate change)**
cause [verb forms] climate change	147	49
cause(s) of climate change	117	34
contribute(s|d) to climate change	68	34
affect [verb forms] climate change	18	7
lead to [verb forms] climate change	6	5
result in [verb forms] climate change	3	2
TOTAL	**359**	**131**

Sub-corpora for two communities

Mentions of effects of climate change

	"skeptical" (35k instances of climate change)	**"accepting" (22k instances of climate change)**
result|effect(s)|impact(s)|consequence(s) of climate change	1,412	1,034
due to climate change	179	148
climate change cause|affect|lead to|result in|contribute to [verb forms]	68	34
TOTAL	**1,659**	**1,216**

Corpuscle demo

- Corpuscle is a corpus management and analysis tool developed by Paul Meurer at Uni Research, Bergen
  - “Standard” corpus analysis functionality, i.e. wordlists, keywords, concordances, collocation
  - Use of metadata allows for comparisons according to text features, e.g. date, blog community, etc.

Automatic grammar induction

- An active field in computational linguistics
  - researchers are attempting to demonstrate that language can be learned with a general learning mechanism (vs. a universal grammar), e.g. ADIOS (Solan et al. 2005)
  - cf. the work of Zellig Harris (1954; 1988) who demonstrated how linguistic units that map to information structures can be identified through the formal analysis of word distributions

Automatic grammar induction

- Research in Bergen is adapting and applying grammar induction algorithms to characterise what is written about key concepts in a discourse
  - i.e. rather than a grammar per se, we seek to induce patterning that captures important information structures

Automatic grammar induction

1. (((to (combat|fight))| (to (battle|slow|minimise|mitigate|tackle))) **climate_change**)

2. (**climate_change** (summit|adaptation|talks|meetings|convention))

3. (((greenhouse gases)|emissions|gases|(carbon emissions)|pollution) blamed ((for|to) **global_warming**))

4. ((cause|causes) (of **global_warming**))

5. ((dangers|signs|effect|consequences|perils) (of **global_warming**))

6. (to (confuse|mislead|educate) the public) *// from global_warming snippets*

7. ((anthropogenic|manmade|(man made)) **global_warming**)

Automatic grammar induction

8. ((would|should|to|must) (control|reduce|regulate|regulating|release) **greenhouse_gases**)

9. ((source|emitter|emitters|producers) of **greenhouse_gases**)

10. (the (effects|impact) ((under|of) ((a|its|the) **carbon_tax**)))

11. (a (modest|$_NUMBER a tonne|global|simple) **carbon_tax**)

12. ((will|would|to) (push|raise|elevate) (**sea_levels** (around|by)))

13. (((due to)|(caused by)) ((climate change)|(global warming))) *//from sea_levels snippets*

14. ((((the|global|some|sophisticated|complex) **climate_models**) (hint|show|indicate) that)

How can we use big data for research on digital reading?

- Data sets:
  - Textual: posts to online reading groups, book reviews, book websites, …
  - Network: links between online participants,
  - Numeric: sales figures, …
  - Geographic data
  - Media coverage of books, including advertisting

- Techniques:
  - Corpus analysis (keywords, concordances, collocations)
  - Text mining (topic modeling, information extraction, grammar induction)
  - Thematic analysis, assisted coding

- Research questions:
  - Who reads what?
  - How do social relationships influence reading?
  - How do readers group/associate/classify books?
  - What language is used to write about reading?
  - What aspects of books are (most) discussed, and contested?
  - Variation over time, and between social groups

- Critical issues:
  - What data do Amazon et al. hold on reading behaviour, how do they use it, how can we access it?
  - The interpretation of statistics and visualizations
  - Privacy and ethics of digital data