Getting to know your corpus: applying Topic Modelling to a corpus of research articles

Paul Thompson
University of Birmingham
p.thompson@bham.ac.uk

Akira Murakami
University of Cambridge
am933@cam.ac.uk

Susan Hunston
University of Birmingham
s.e.hunston@bham.ac.uk

Background

- A challenge in corpus linguistics is to develop bottom-up methods to explore corpora without imposing pre-existing distinctions such as the genre or the author of the text.

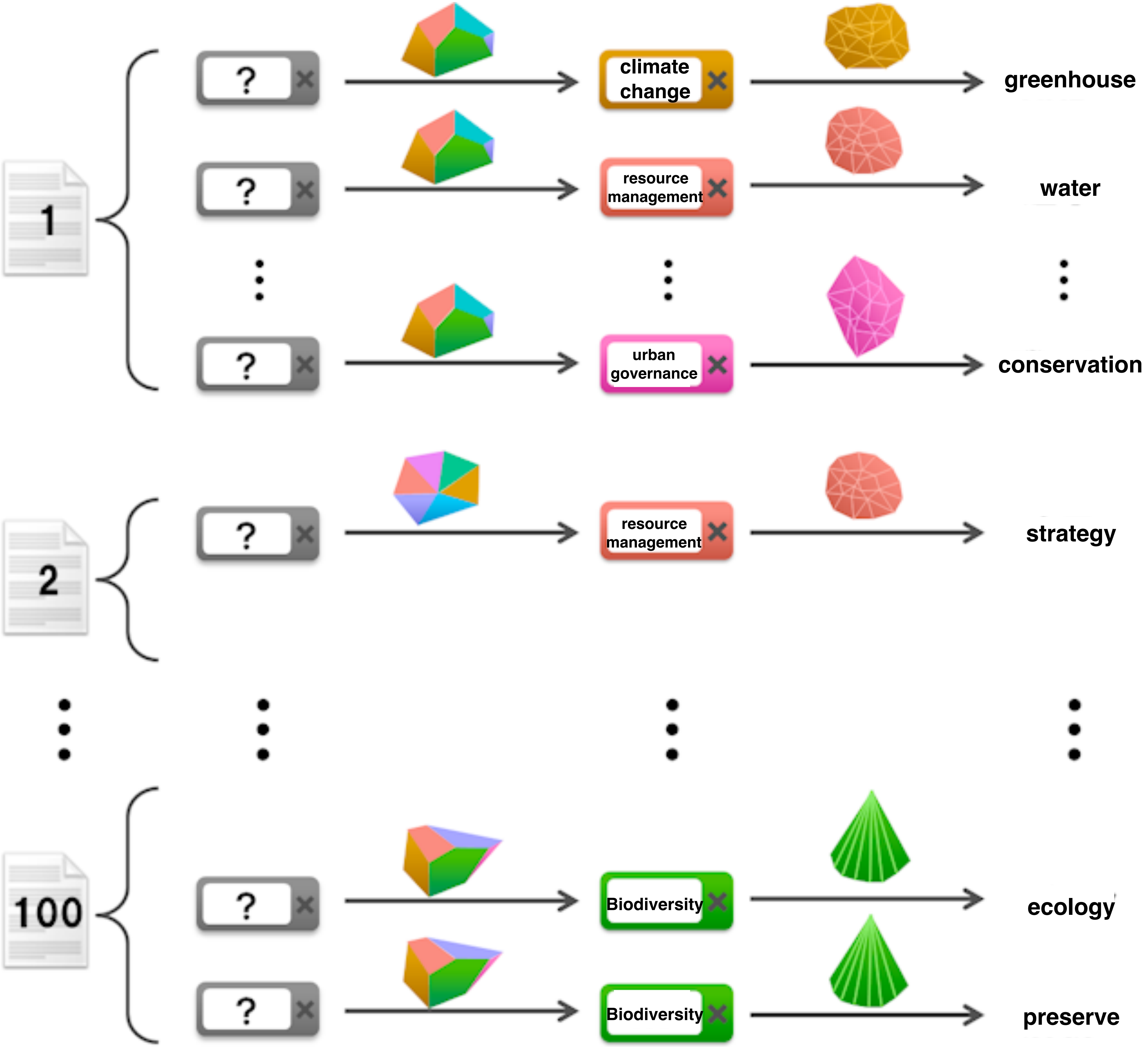
- In this talk, we will introduce the use of topic modeling (Blei, 2012), a machine-learning technique that automatically identifies “topics” in a corpus.

Brief Overview of Topic Models

Features of Topic Models

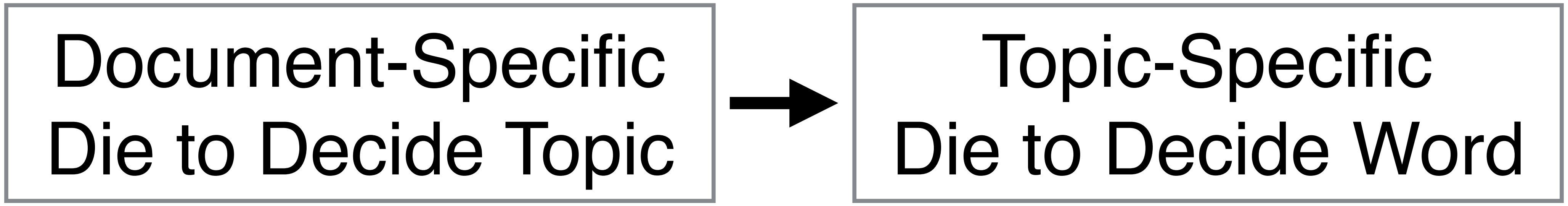
- Latent Dirichlet allocation (LDA)
- Automatically identifies “topics” in a given corpus
  - keywords in each topic
  - distribution of topics in each document
    - A document consists of multiple topics
- Topic
  - probability distribution over words
  - characterised by a group of co-occurring words in documents
- Methodologically,
  - latest technique to analyze document-term matrices.
  - Bag-of-words approach → single words

Assumed generative process of each word.



Adapted from http://heartruptcy.blog.fc2.com/blog-entry-124.html

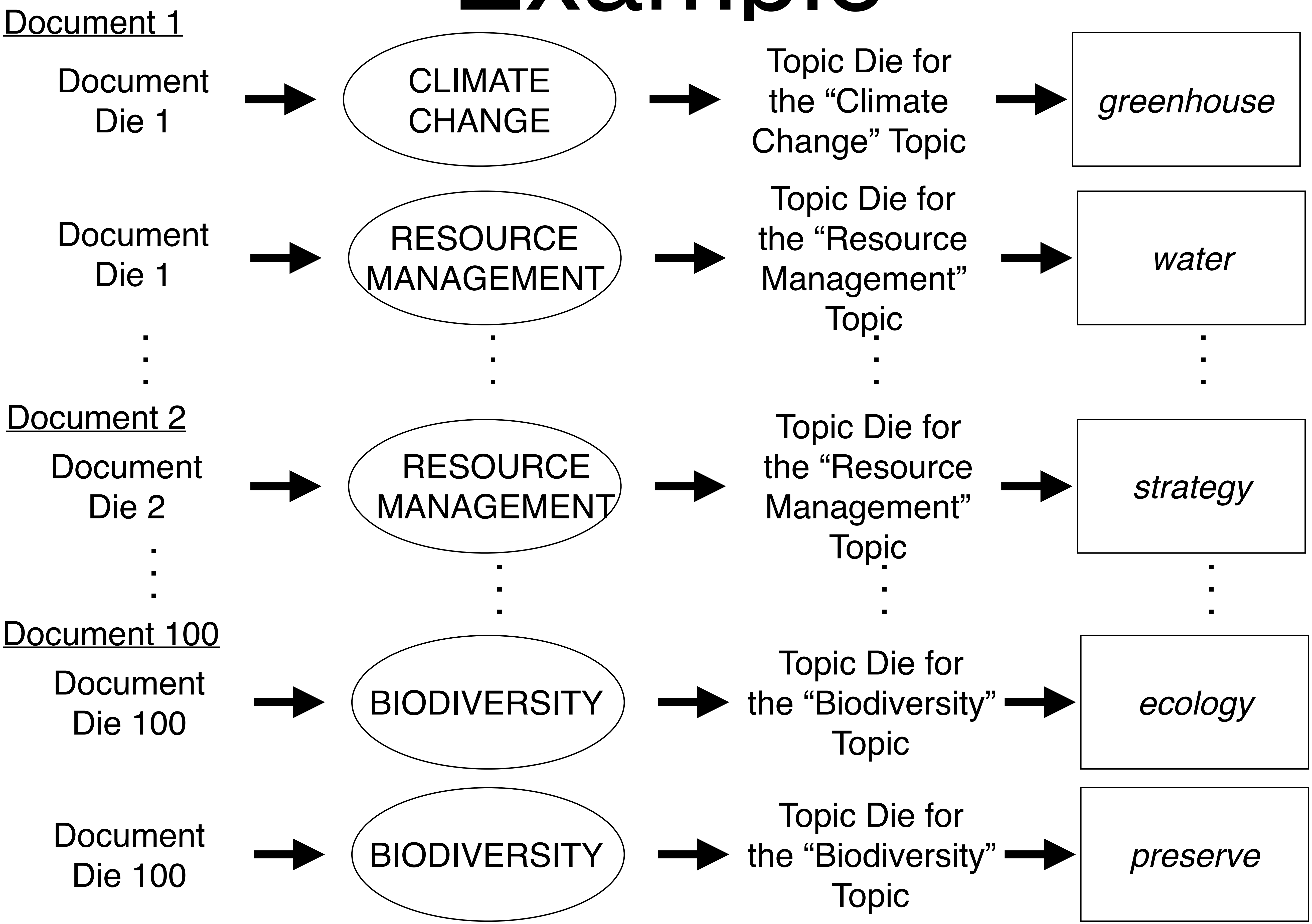
Assumed Generative Process of Each Word



Document-Specific Die to Decide Topic

Topic-Specific Die to Decide Word

Example



Document 1

Document Die 1 → CLIMATE CHANGE → Topic Die for the “Climate Change” Topic → *greenhouse*

Document Die 1 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *water*

. . .

Document 2

Document Die 2 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *strategy*

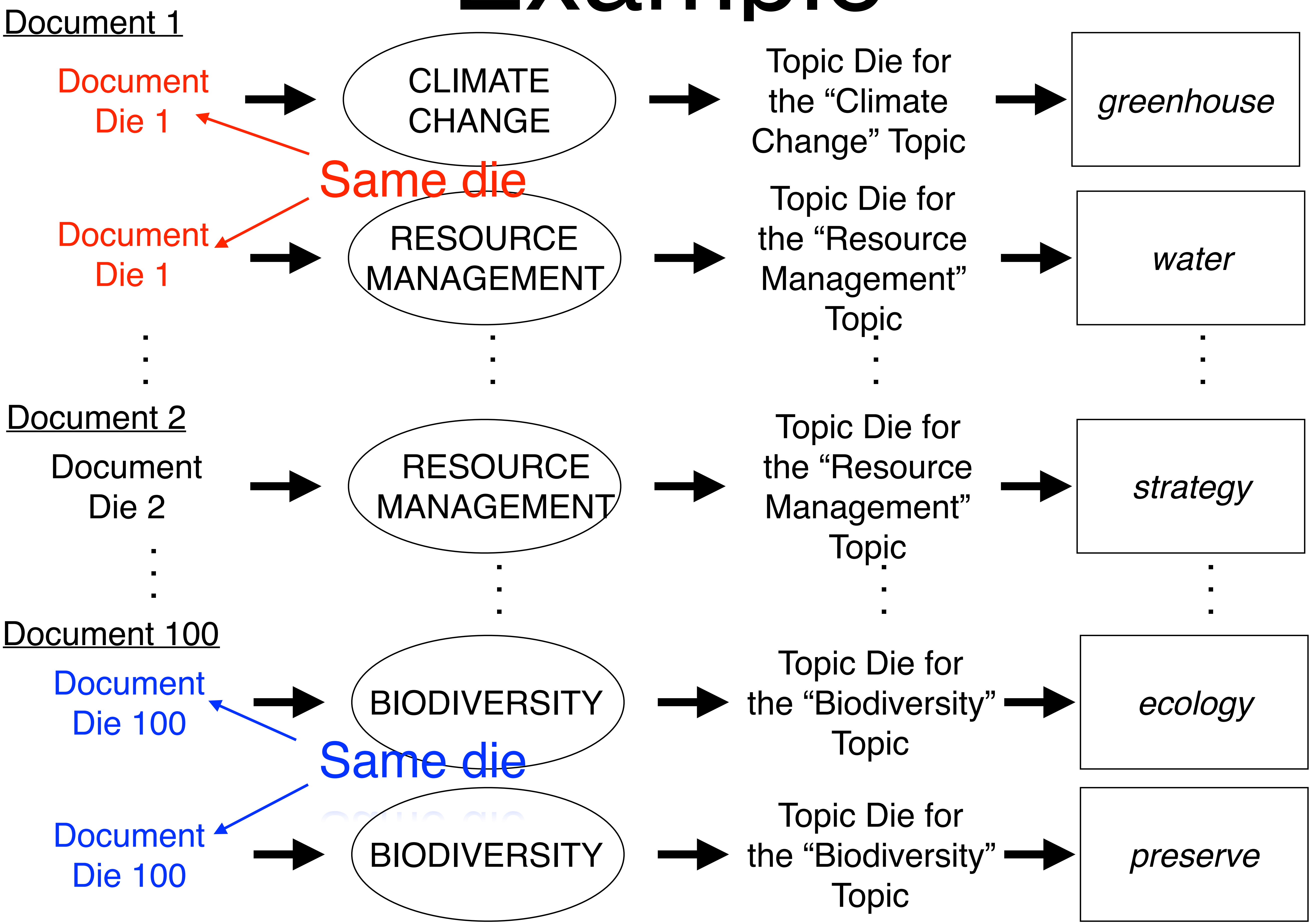
. . .

Document 100

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *ecology*

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *preserve*

Example



Document 1

Document Die 1 → CLIMATE CHANGE → Topic Die for the “Climate Change” Topic → *greenhouse*

Same die

Document Die 1 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *water*

⋮

Document 2

Document Die 2 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *strategy*

⋮

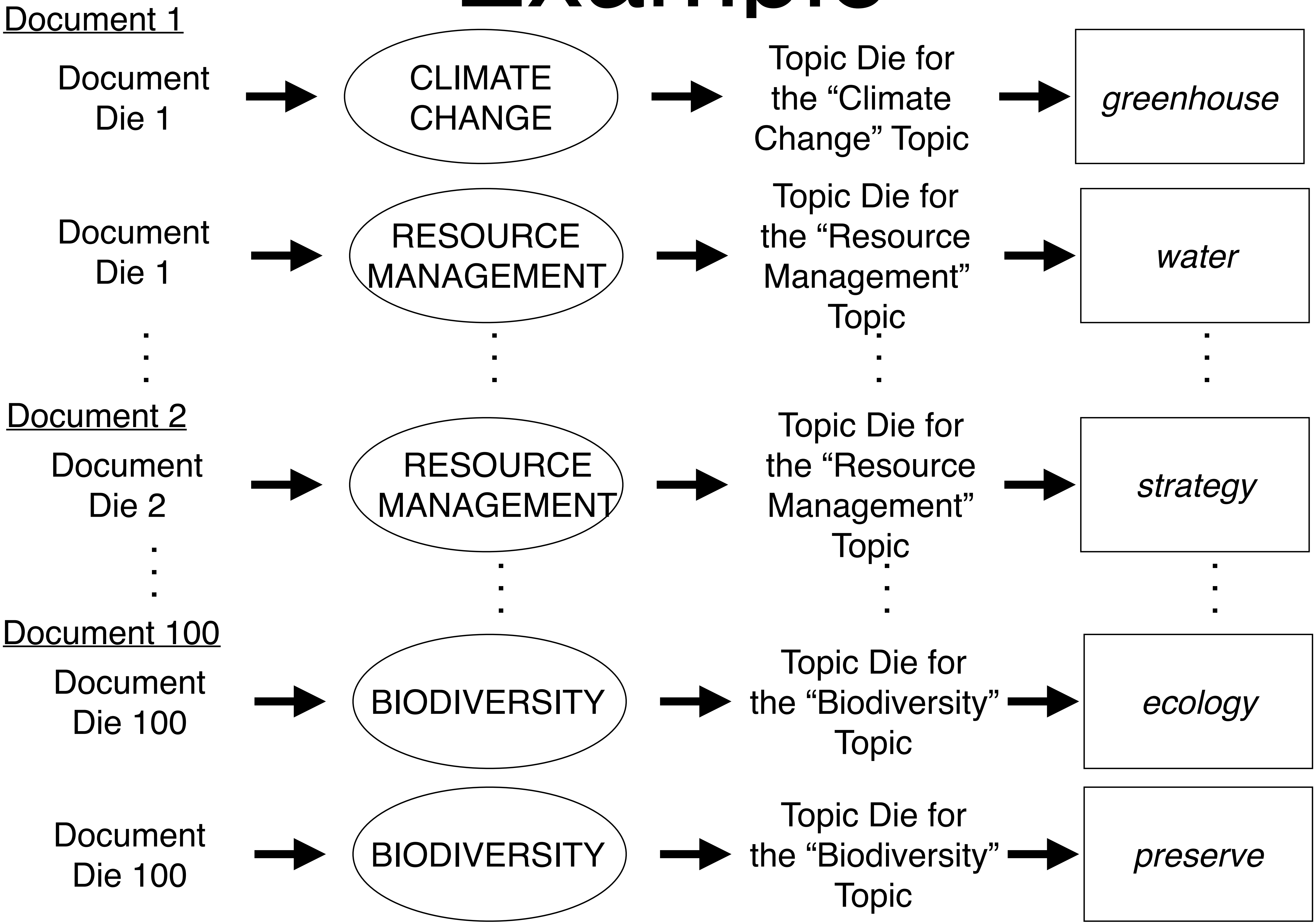
Document 100

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *ecology*

Same die

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *preserve*

Example



Document 1

Document Die 1 → CLIMATE CHANGE → Topic Die for the “Climate Change” Topic → *greenhouse*

Document Die 1 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *water*

. . .

Document 2

Document Die 2 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *strategy*

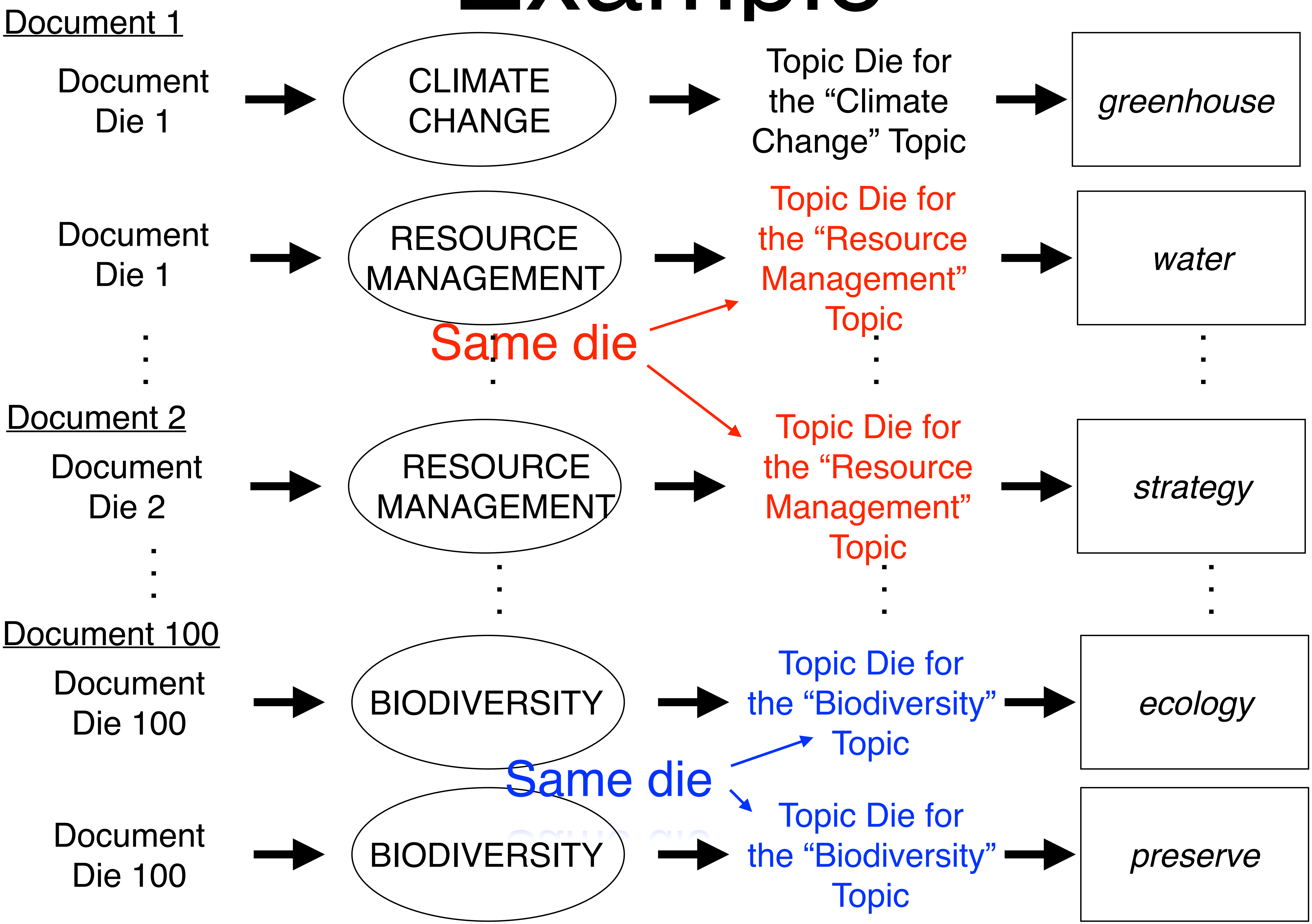
. . .

Document 100

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *ecology*

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *preserve*

Example



Document 1

Document Die 1 → CLIMATE CHANGE → Topic Die for the “Climate Change” Topic → *greenhouse*

Document Die 1 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *water*

Same die

Document 2

Document Die 2 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *strategy*

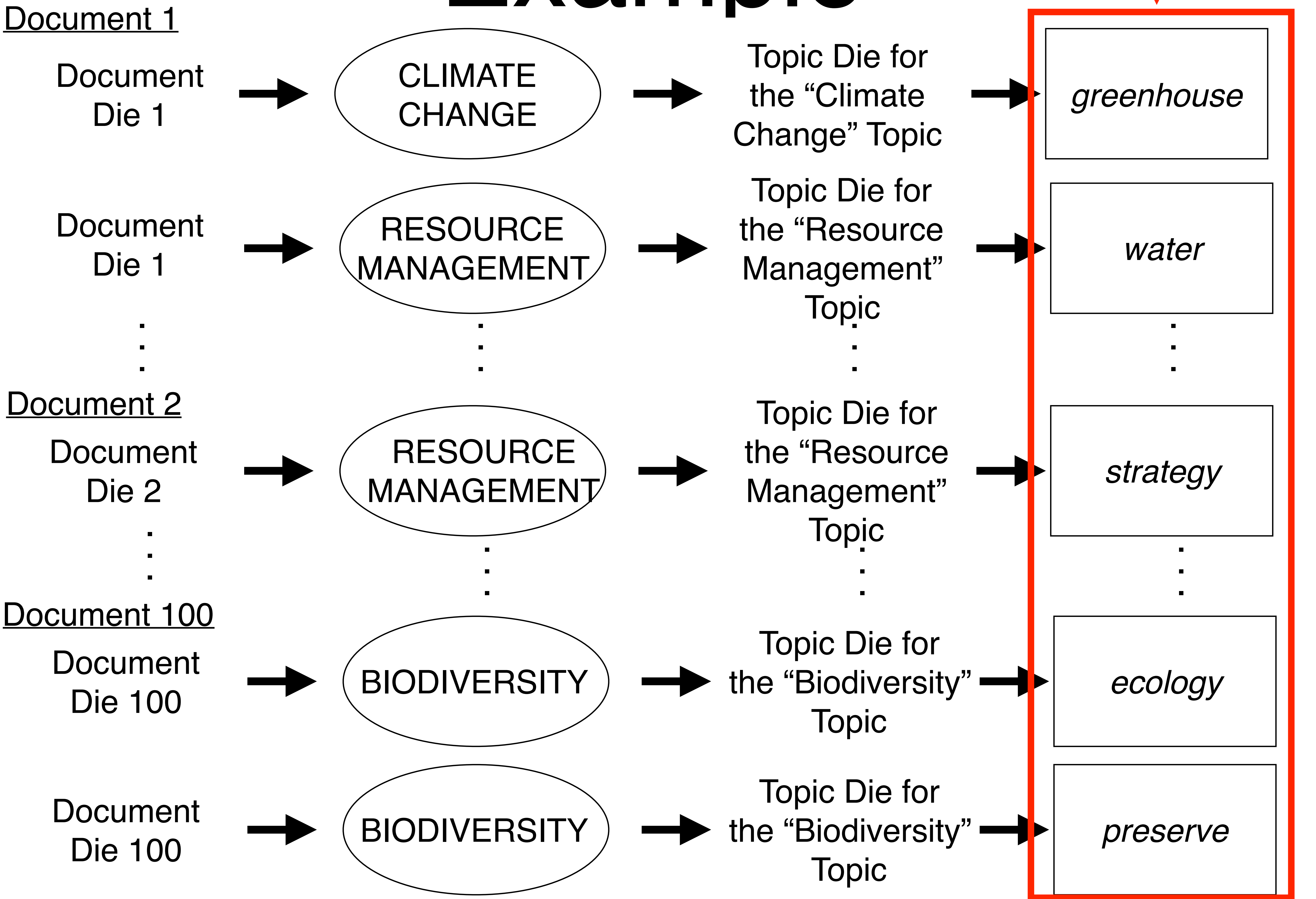
Document 100

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *ecology*

Same die

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *preserve*

Example



what we observe

Document 1

Document Die 1 → CLIMATE CHANGE → Topic Die for the “Climate Change” Topic → *greenhouse*

Document Die 1 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *water*

⋮

Document 2

Document Die 2 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *strategy*

⋮

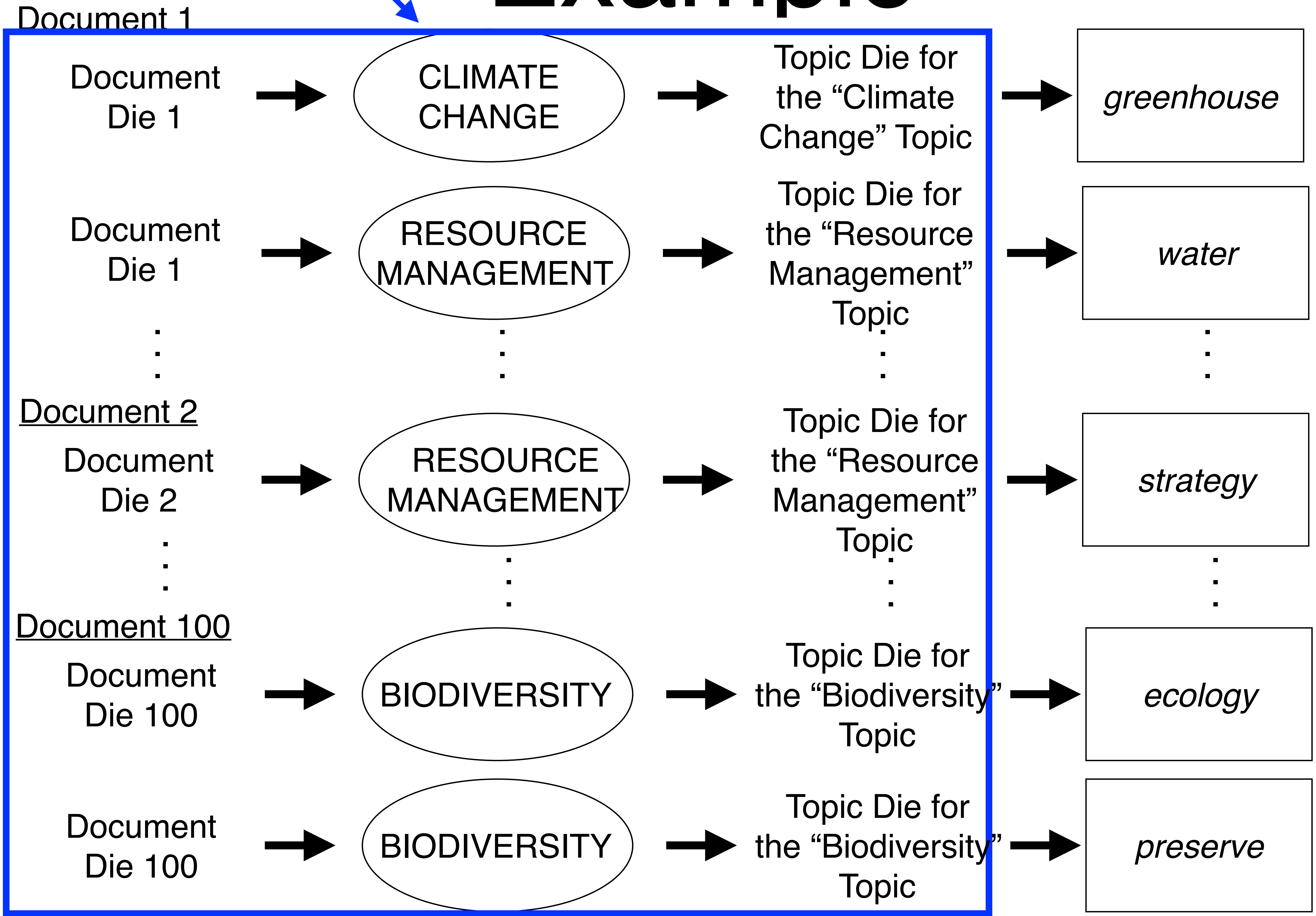
Document 100

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *ecology*

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *preserve*

Example

what we are interested in



Document 1

Document Die 1 → CLIMATE CHANGE → Topic Die for the “Climate Change” Topic → *greenhouse*

Document Die 1 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *water*

⋮

Document 2

Document Die 2 → RESOURCE MANAGEMENT → Topic Die for the “Resource Management” Topic → *strategy*

⋮

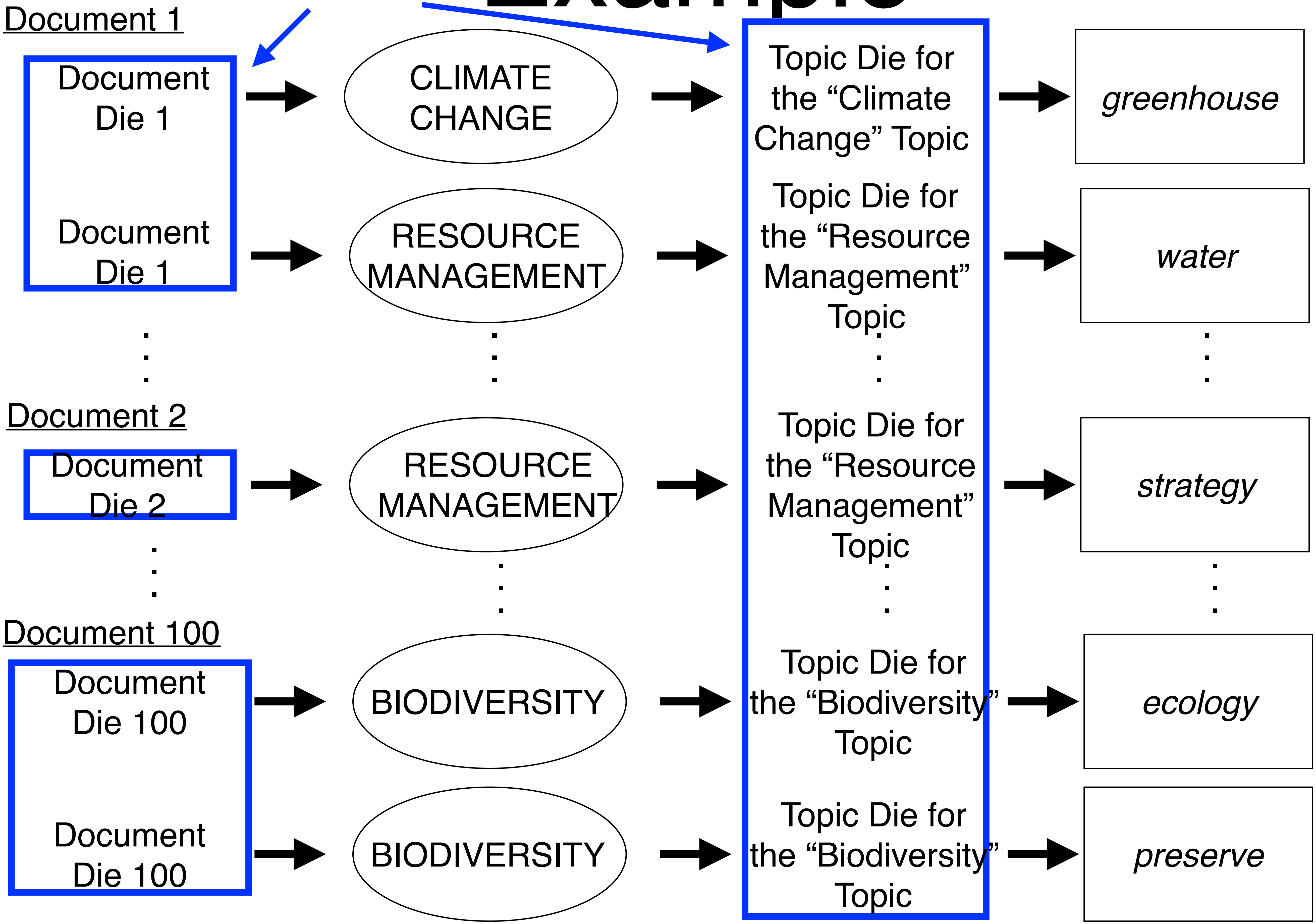
Document 100

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *ecology*

Document Die 100 → BIODIVERSITY → Topic Die for the “Biodiversity” Topic → *preserve*

Example

what topic modeling reveals



Document 1

Document Die 1 → CLIMATE CHANGE → Topic Die for the "Climate Change" Topic → *greenhouse*

Document Die 1 → RESOURCE MANAGEMENT → Topic Die for the "Resource Management" Topic → *water*

⋮

Document 2

Document Die 2 → RESOURCE MANAGEMENT → Topic Die for the "Resource Management" Topic → *strategy*

⋮

Document 100

Document Die 100 → BIODIVERSITY → Topic Die for the "Biodiversity" Topic → *ecology*

Document Die 100 → BIODIVERSITY → Topic Die for the "Biodiversity" Topic → *preserve*

Shape of Dice

- We are interested in the shape of each irregular dice.
- For instance,
  - How likely that we get Topic 5 in Document 1?
  - How likely that we get the word *water* in Topic 8?
- This is what topic modeling does.

Estimating the Shapes of the Dice (or the Latent Variables) Given a Corpus

- An estimation method for the topic model is Gibbs sampling (Griffiths & Steyvers, 2004), a form of Markov Chain Monte Carlo (MCMC).
- Intuitively (Wagner, 2010),
  - “Once many tokens of a word have been assigned to topic *j* (across documents), the probability of assigning any particular token of that word to topic *j* increases”
  - “Once a topic *j* has been used multiple times in one document, it will increase the probability that any word from that document will be assigned to topic *j*”

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Illustration

Document 1	Word X Word X Word Y
Document 2	Word Y Word Z Word Z
Document 3	Word Z Word Z Word Z

Our Study

Aim

- We explore the use of topic models in a corpus of academic discourse.
- We target research papers published in the journal, *Global Environmental Change (GEC)*.

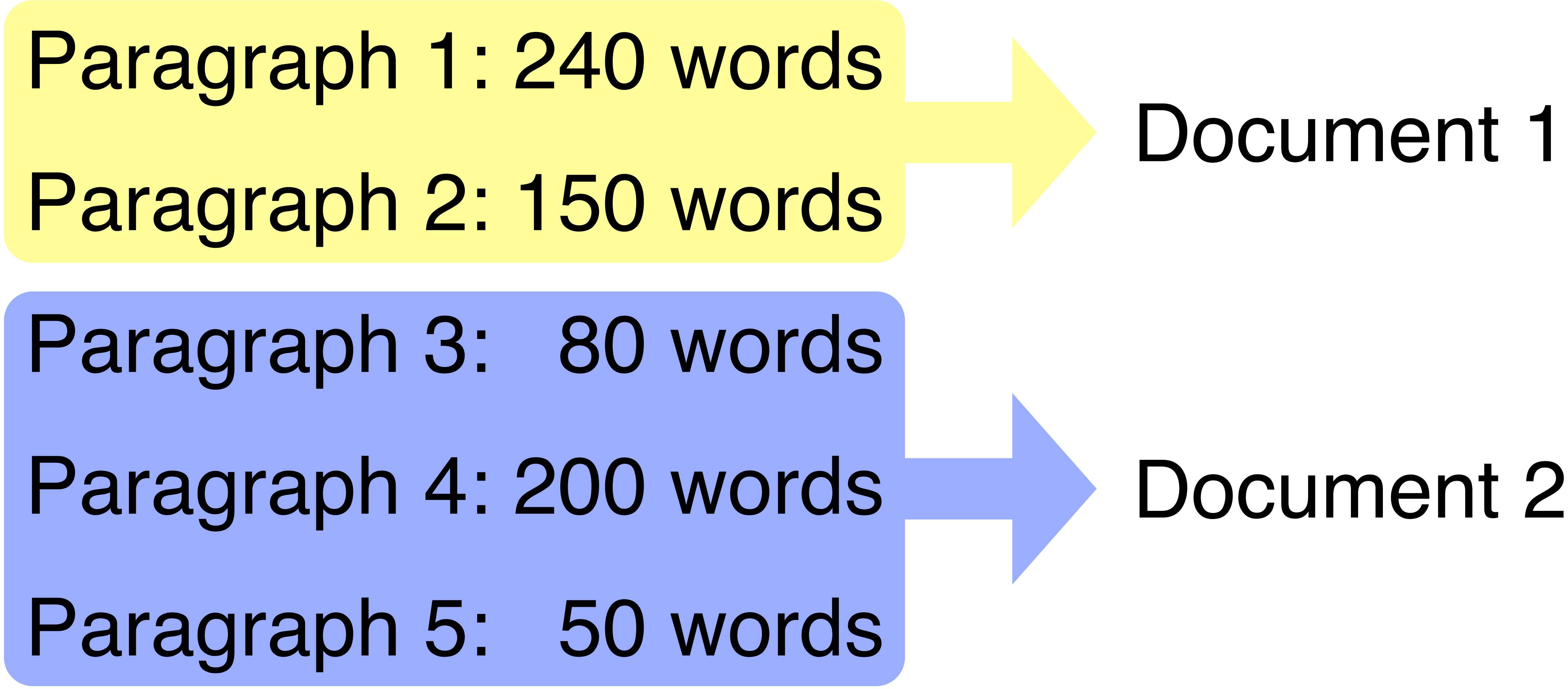
GEC Corpus

- All the full papers in the journal (1990-2010)
- Main text only
- 675 papers
- 4.1 million words

Division of Papers

- A decision we need to make is what to conceive as a document. A document should be
  - short enough to be topically (relatively) uniform and
  - long enough to reliably identity word co-occurrence patterns.
- A research paper
  - is longer than a typical document targeted in topic models
  - can contain multiple topics
- Better to divide papers into multiple parts
- This allows the investigation of topic transition within papers as well.

Document Generation



Paragraph 1: 240 words

Paragraph 2: 150 words

Document 1

Paragraph 3: 80 words

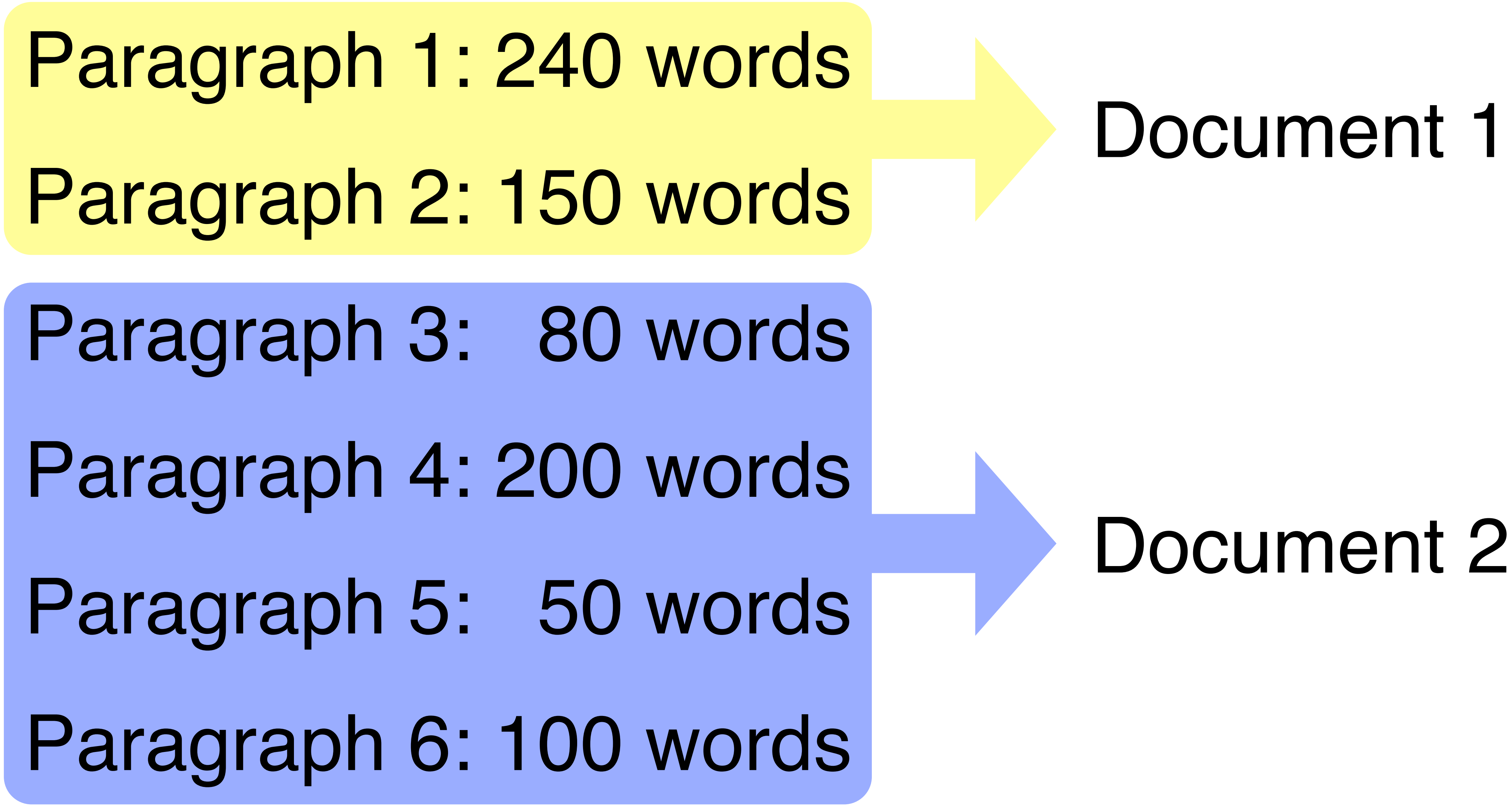
Paragraph 4: 200 words

Paragraph 5: 50 words

Document 2

Paragraph 6: 100 words

Document Generation



Paragraph 1: 240 words

Paragraph 2: 150 words

Document 1

Paragraph 3: 80 words

Paragraph 4: 200 words

Paragraph 5: 50 words

Paragraph 6: 100 words

Document 2

Details

- Only targeted the terms that
  - are not in the following stopwords: *BE*, *HAVE*, *DO*, articles, prepositions, *and*, *it*, *as*, *that*,
  - are equal to or longer than two letters, and
  - appear in at least 0.1% of all the documents.
- All the words were stemmed (e.g., *require* → *requir*, *analysis* → *analysi*).
- Each document was assigned with the information on where in the paper the paragraph(s) appeared.
  - e.g., 70% from the beginning of the paper
- 10,555 documents with the average length of 242 words (SD = 50)
- *topicmodels* package (Grün & Hornik, 2011) in R

Number of Topics

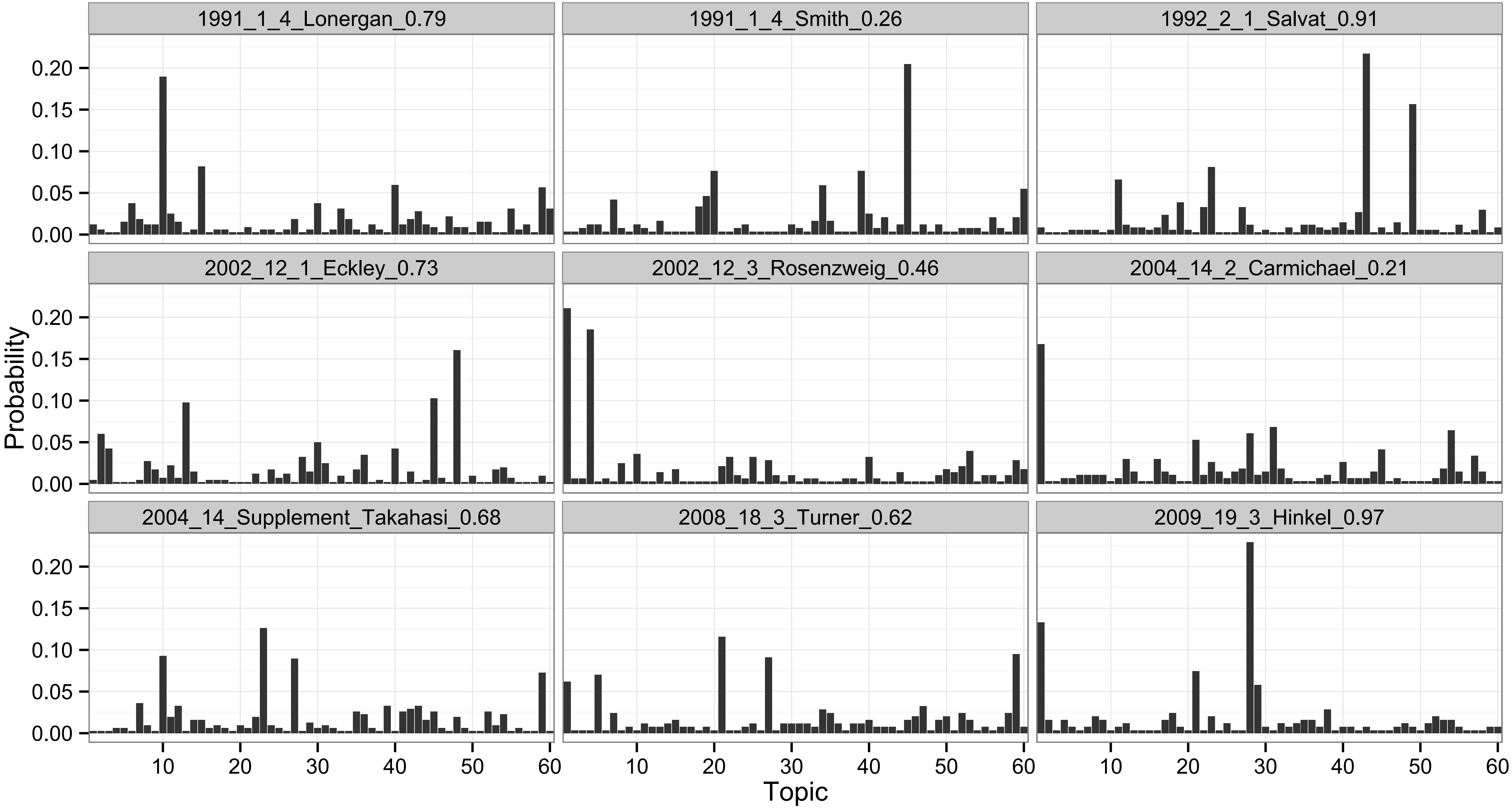
- No agreed way to automatically determine the number of topics.
- Built topic models with 40, 50, 60, . . . , 90,100 topics.
- 60 topics looked like the right level of granularity.
  → 60 topics

Results

&

Discussion

By-Document Topic Distribution

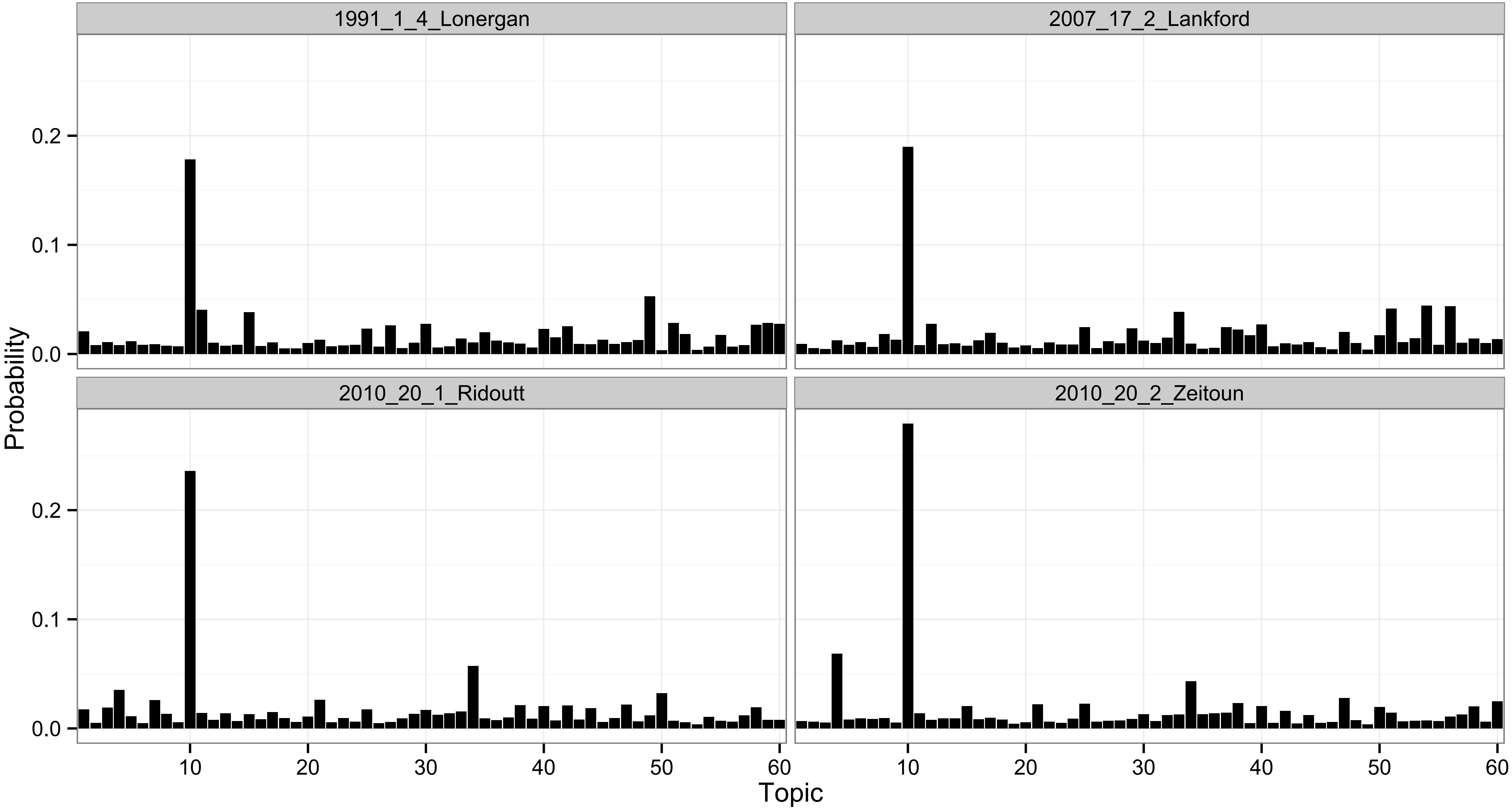


We can . . .

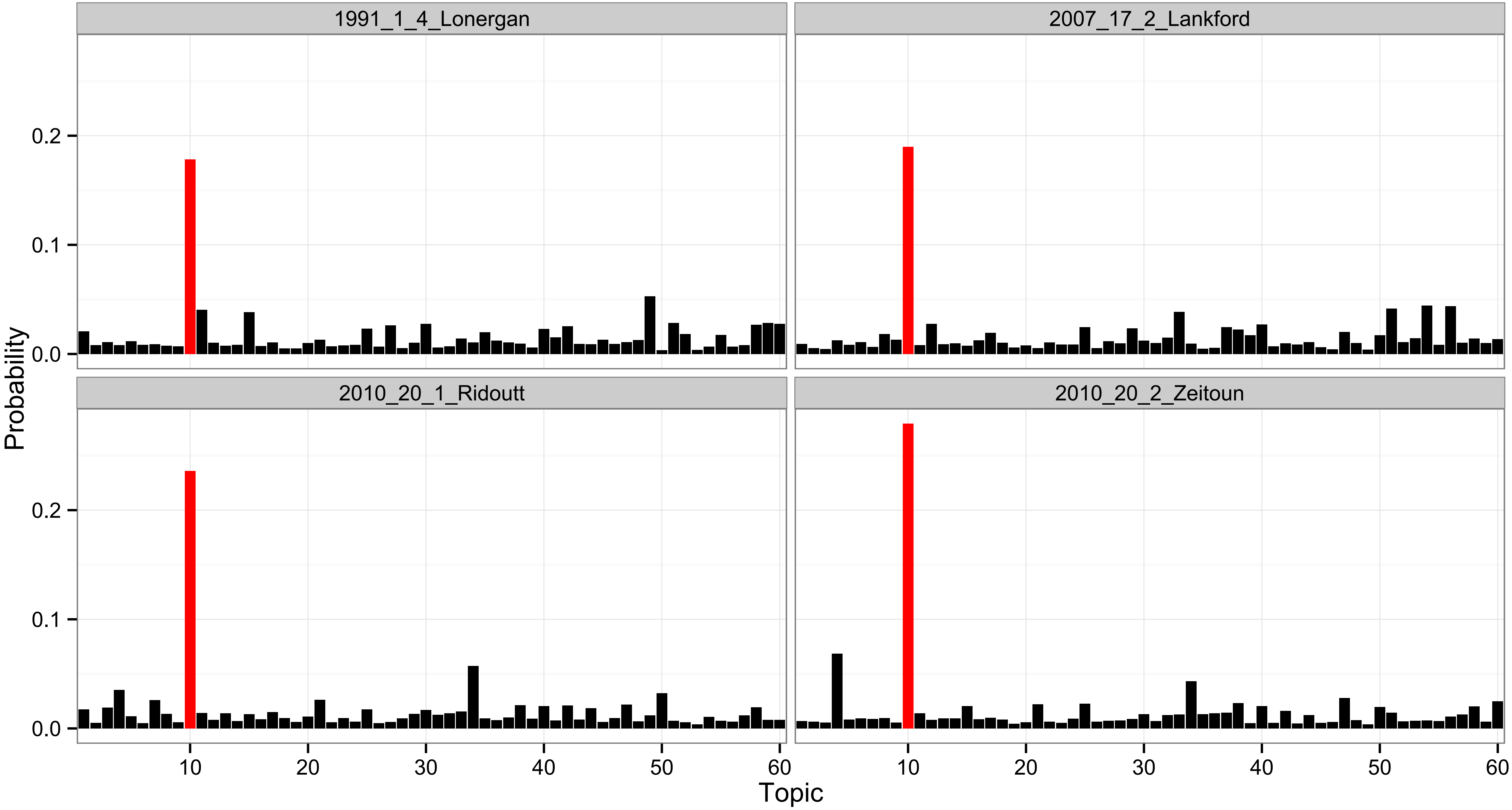
- Identify prominent topics at different positions of a paper.
- Identify prominent papers and documents of each topic.
- Cluster papers according to topic distribution,

etc.

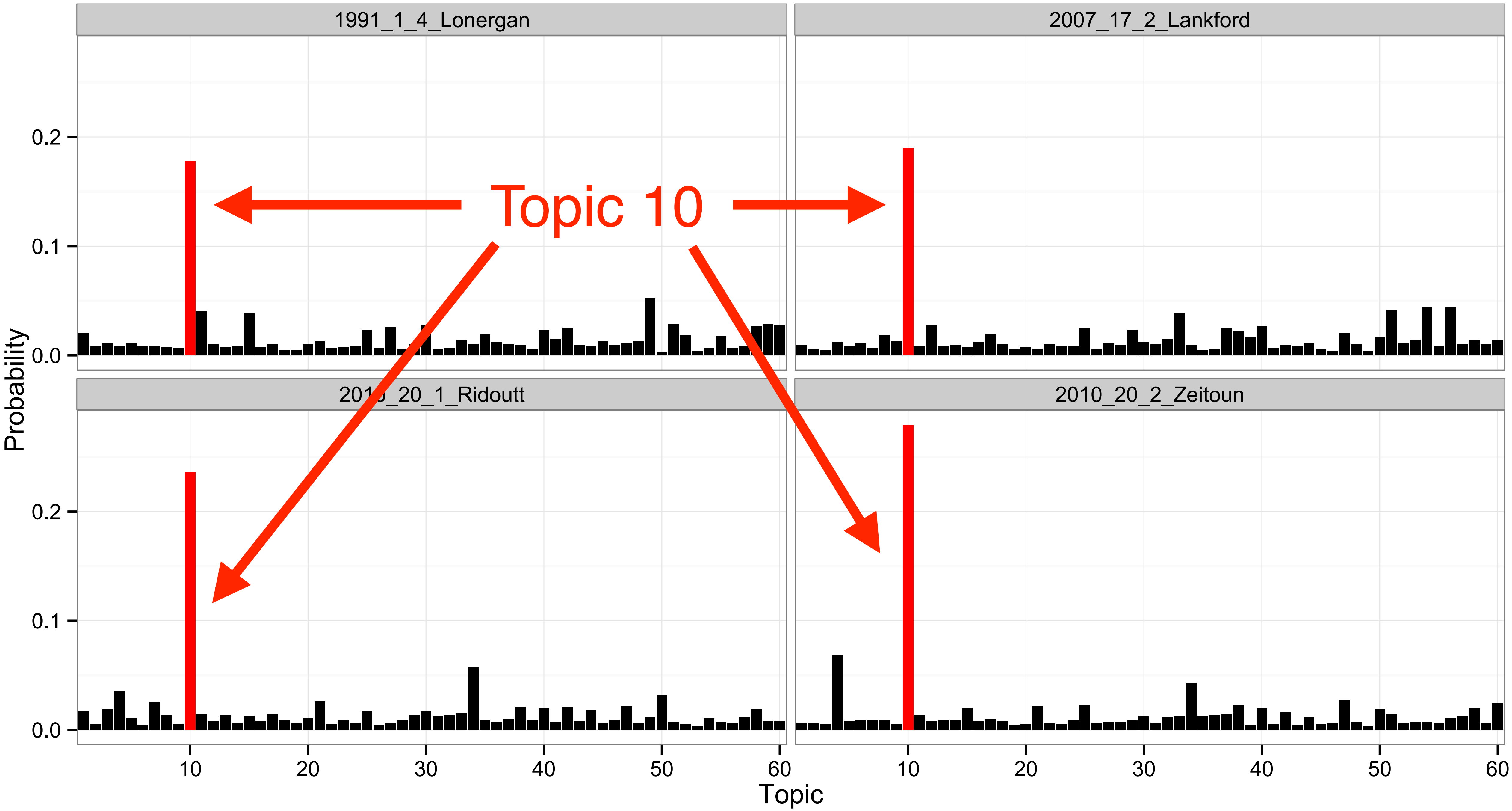
By-Paper Topic Distribution



By-Paper Topic Distribution



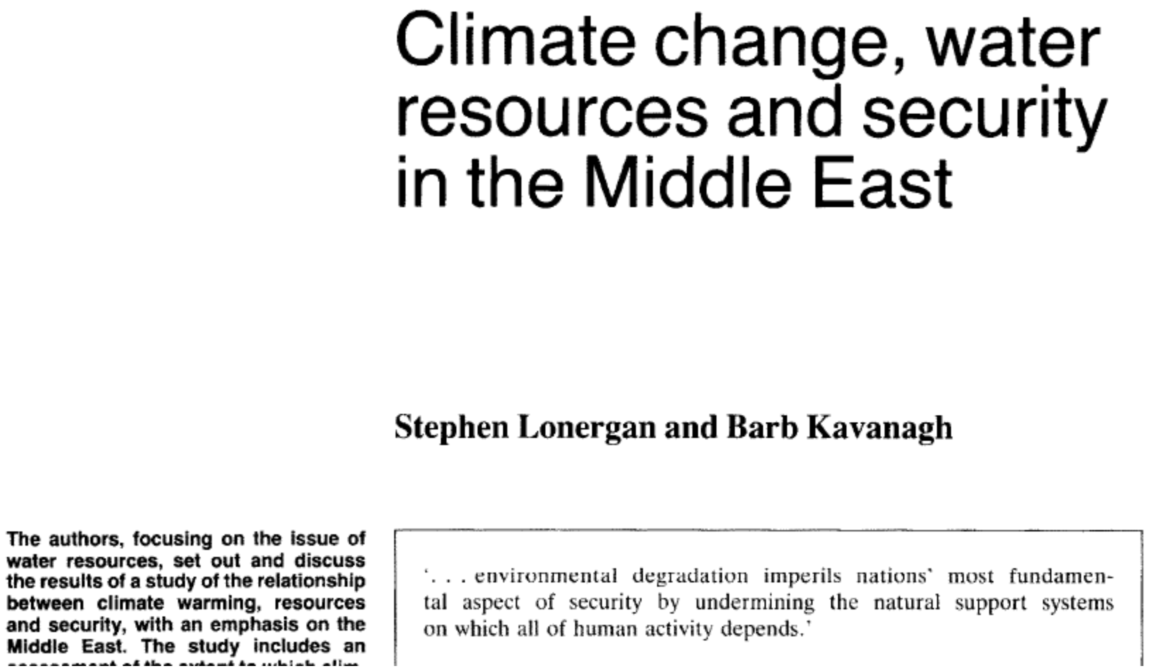
By-Paper Topic Distribution



Keywords of Topic 10

- water, river, basin, suppli, flow, irrig, resourc, avail, use, stress, demand, state, system, lake, manag, hydrolog, qualiti, virtual, groundwat, watersh
- The topic is labeled “water systems, supplies, trade”.

1991_1_4_Lonergan



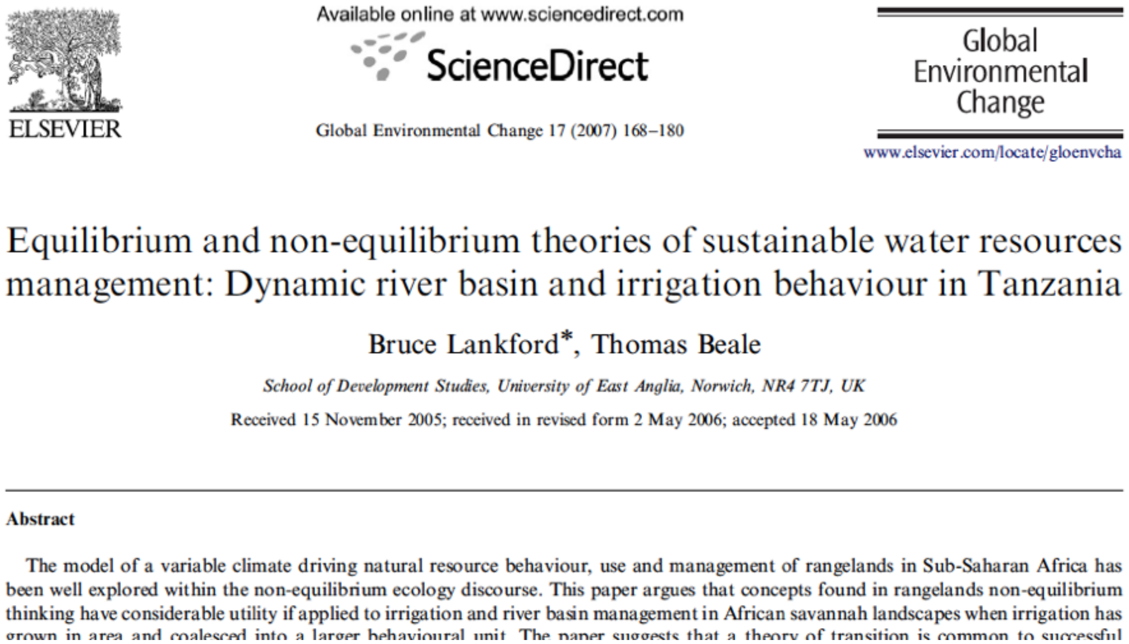
Climate change, water resources and security in the Middle East

Stephen Lonergan and Barb Kavanagh

The authors, focusing on the issue of water resources, set out and discuss the results of a study of the relationship between climate warming, resources and security, with an emphasis on the Middle East. The study includes an

> '. . . environmental degradation imperils nations' most fundamental aspect of security by undermining the natural support systems on which all of human activity depends.'

2007_17_2_Lankford



ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Global Environmental Change 17 (2007) 168–180

Global Environmental Change

www.elsevier.com/locate/gloenvcha

Equilibrium and non-equilibrium theories of sustainable water resources management: Dynamic river basin and irrigation behaviour in Tanzania

Bruce Lankford*, Thomas Beale

School of Development Studies, University of East Anglia, Norwich, NR4 7TJ, UK

Received 15 November 2005; received in revised form 2 May 2006; accepted 18 May 2006

Abstract

The model of a variable climate driving natural resource behaviour, use and management of rangelands in Sub-Saharan Africa has been well explored within the non-equilibrium ecology discourse. This paper argues that concepts found in rangelands non-equilibrium thinking have considerable utility if applied to irrigation and river basin management in African savannah landscapes when irrigation has grown in area and coalesced into a larger behavioural unit. The paper suggests that a theory of transition is common to successful

2010_20_1_Ridoutt

Global Environmental Change 20 (2010) 113–120

Contents lists available at ScienceDirect

Global Environmental Change

journal homepage: www.elsevier.com/locate/gloenvcha

ELSEVIER

A revised approach to water footprinting to make transparent the impacts of consumption and production on global freshwater scarcity

Bradley G. Ridoutt[a,*], Stephan Pfister[b]

[a] *CSIRO Sustainable Ecosystems, Private Bag 10, Clayton, Victoria 3169, Australia*
[b] *ETH Zurich, Institute of Environmental Engineering, 8093 Zurich, Switzerland*

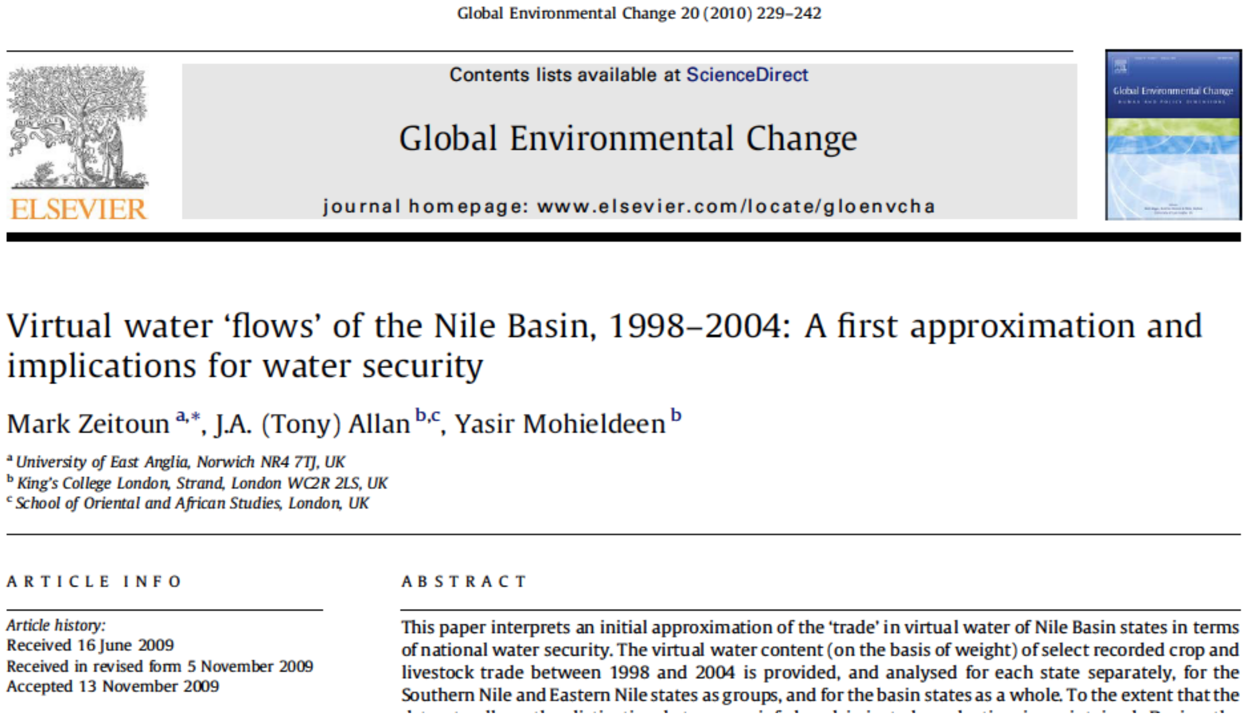
ARTICLE INFO

Article history:
Received 14 April 2009
Received in revised form 27 July 2009
Accepted 20 August 2009

ABSTRACT

Through the interconnectedness of global business, the local consumption of products and services is intervening in the hydrological cycle throughout the world to an unprecedented extent. In order to address the unsustainable use of global freshwater resources, indicators are needed which make the impacts of production systems and consumption patterns transparent. In this paper, a revised water footprint calculation method, incorporating water stress characterisation factors, is presented and

2010_20_2_Zeitoun



Global Environmental Change 20 (2010) 229–242

Contents lists available at ScienceDirect

Global Environmental Change

journal homepage: www.elsevier.com/locate/gloenvcha

ELSEVIER

Virtual water 'flows' of the Nile Basin, 1998–2004: A first approximation and implications for water security

Mark Zeitoun[a,*], J.A. (Tony) Allan[b,c], Yasir Mohieldeen[b]

[a] *University of East Anglia, Norwich NR4 7TJ, UK*
[b] *King's College London, Strand, London WC2R 2LS, UK*
[c] *School of Oriental and African Studies, London, UK*

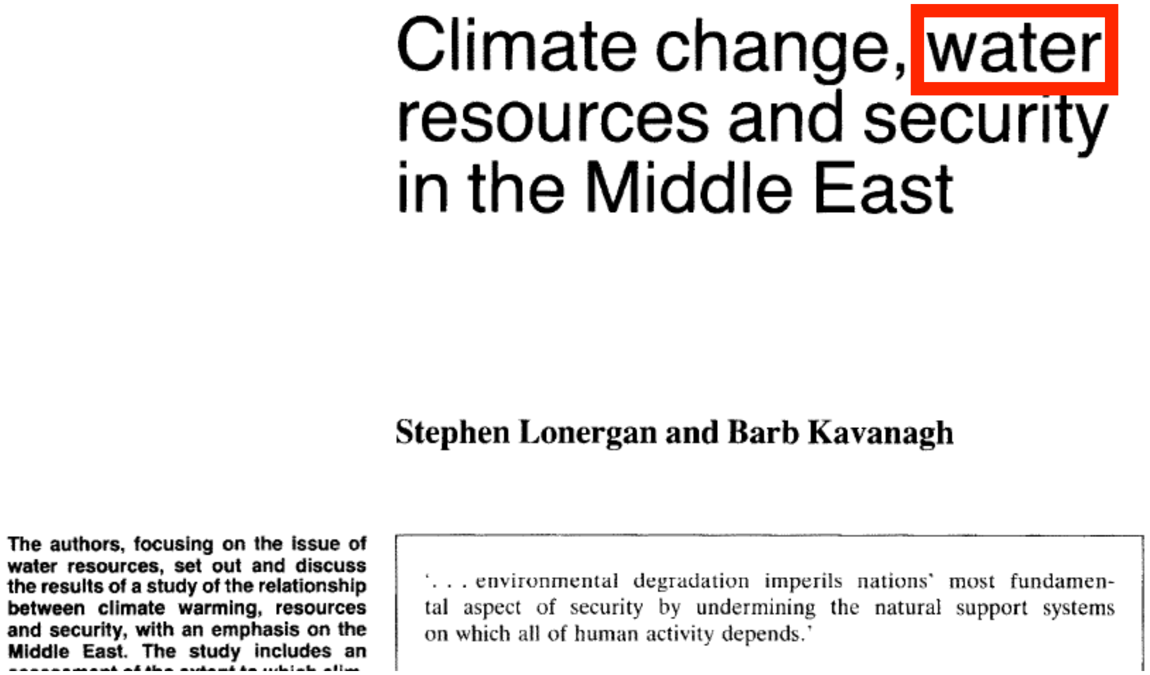
ARTICLE INFO

Article history:
Received 16 June 2009
Received in revised form 5 November 2009
Accepted 13 November 2009

ABSTRACT

This paper interprets an initial approximation of the 'trade' in virtual water of Nile Basin states in terms of national water security. The virtual water content (on the basis of weight) of select recorded crop and livestock trade between 1998 and 2004 is provided, and analysed for each state separately, for the Southern Nile and Eastern Nile states as groups, and for the basin states as a whole. To the extent that the

1991_1_4_Lonergan



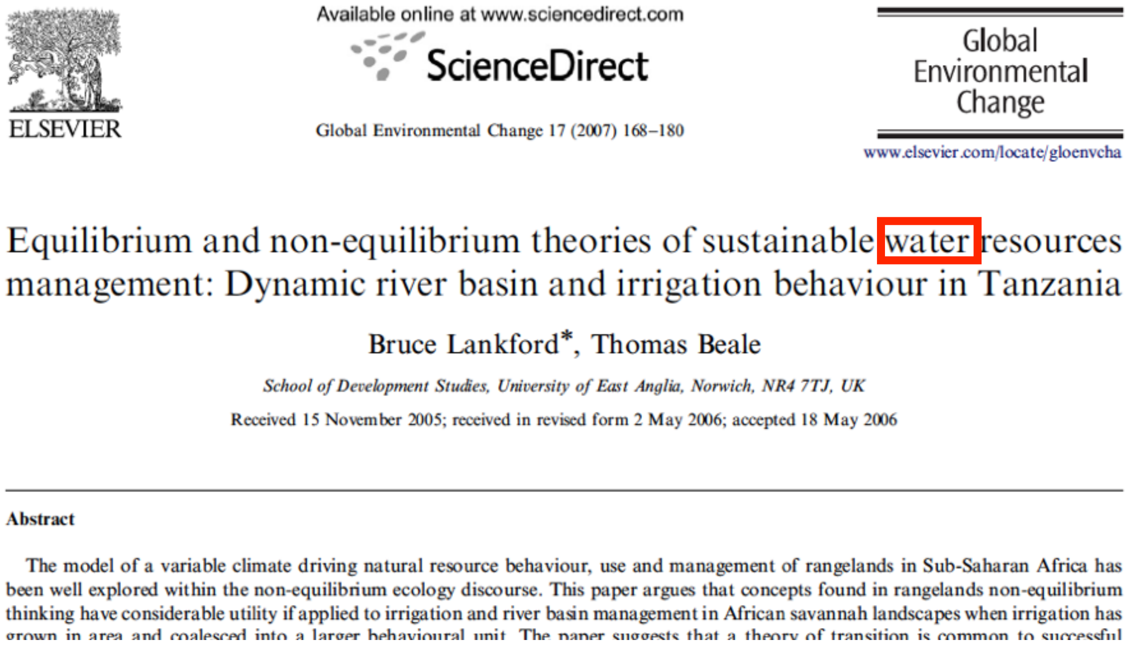
Climate change, water resources and security in the Middle East

Stephen Lonergan and Barb Kavanagh

The authors, focusing on the issue of water resources, set out and discuss the results of a study of the relationship between climate warming, resources and security, with an emphasis on the Middle East. The study includes an

'. . . environmental degradation imperils nations' most fundamental aspect of security by undermining the natural support systems on which all of human activity depends.'

2007_17_2_Lankford



ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Global Environmental Change 17 (2007) 168–180

Global Environmental Change

www.elsevier.com/locate/gloenvcha

Equilibrium and non-equilibrium theories of sustainable water resources management: Dynamic river basin and irrigation behaviour in Tanzania

Bruce Lankford*, Thomas Beale

School of Development Studies, University of East Anglia, Norwich, NR4 7TJ, UK

Received 15 November 2005; received in revised form 2 May 2006; accepted 18 May 2006

Abstract

The model of a variable climate driving natural resource behaviour, use and management of rangelands in Sub-Saharan Africa has been well explored within the non-equilibrium ecology discourse. This paper argues that concepts found in rangelands non-equilibrium thinking have considerable utility if applied to irrigation and river basin management in African savannah landscapes when irrigation has grown in area and coalesced into a larger behavioural unit. The paper suggests that a theory of transition is common to successful

2010_20_1_Ridoutt

Global Environmental Change 20 (2010) 113–120

Contents lists available at ScienceDirect

Global Environmental Change

journal homepage: www.elsevier.com/locate/gloenvcha

ELSEVIER

A revised approach to water footprinting to make transparent the impacts of consumption and production on global freshwater scarcity

Bradley G. Ridoutt [a],*, Stephan Pfister [b]

[a] *CSIRO Sustainable Ecosystems, Private Bag 10, Clayton, Victoria 3169, Australia*
[b] *ETH Zurich, Institute of Environmental Engineering, 8093 Zurich, Switzerland*

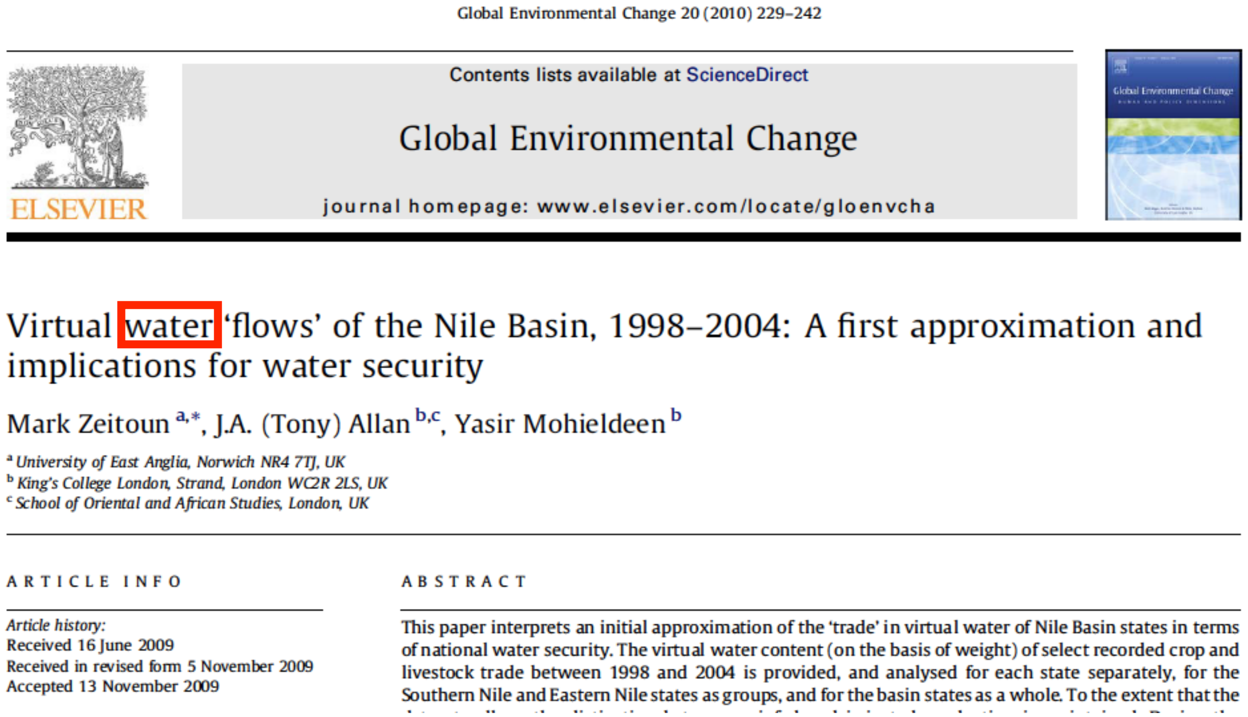
ARTICLE INFO

Article history:
Received 14 April 2009
Received in revised form 27 July 2009
Accepted 20 August 2009

ABSTRACT

Through the interconnectedness of global business, the local consumption of products and services is intervening in the hydrological cycle throughout the world to an unprecedented extent. In order to address the unsustainable use of global freshwater resources, indicators are needed which make the impacts of production systems and consumption patterns transparent. In this paper, a revised water footprint calculation method, incorporating water stress characterisation factors, is presented and

2010_20_2_Zeitoun



Global Environmental Change 20 (2010) 229–242

Contents lists available at ScienceDirect

Global Environmental Change

journal homepage: www.elsevier.com/locate/gloenvcha

ELSEVIER

Virtual water 'flows' of the Nile Basin, 1998–2004: A first approximation and implications for water security

Mark Zeitoun [a],*, J.A. (Tony) Allan [b,c], Yasir Mohieldeen [b]

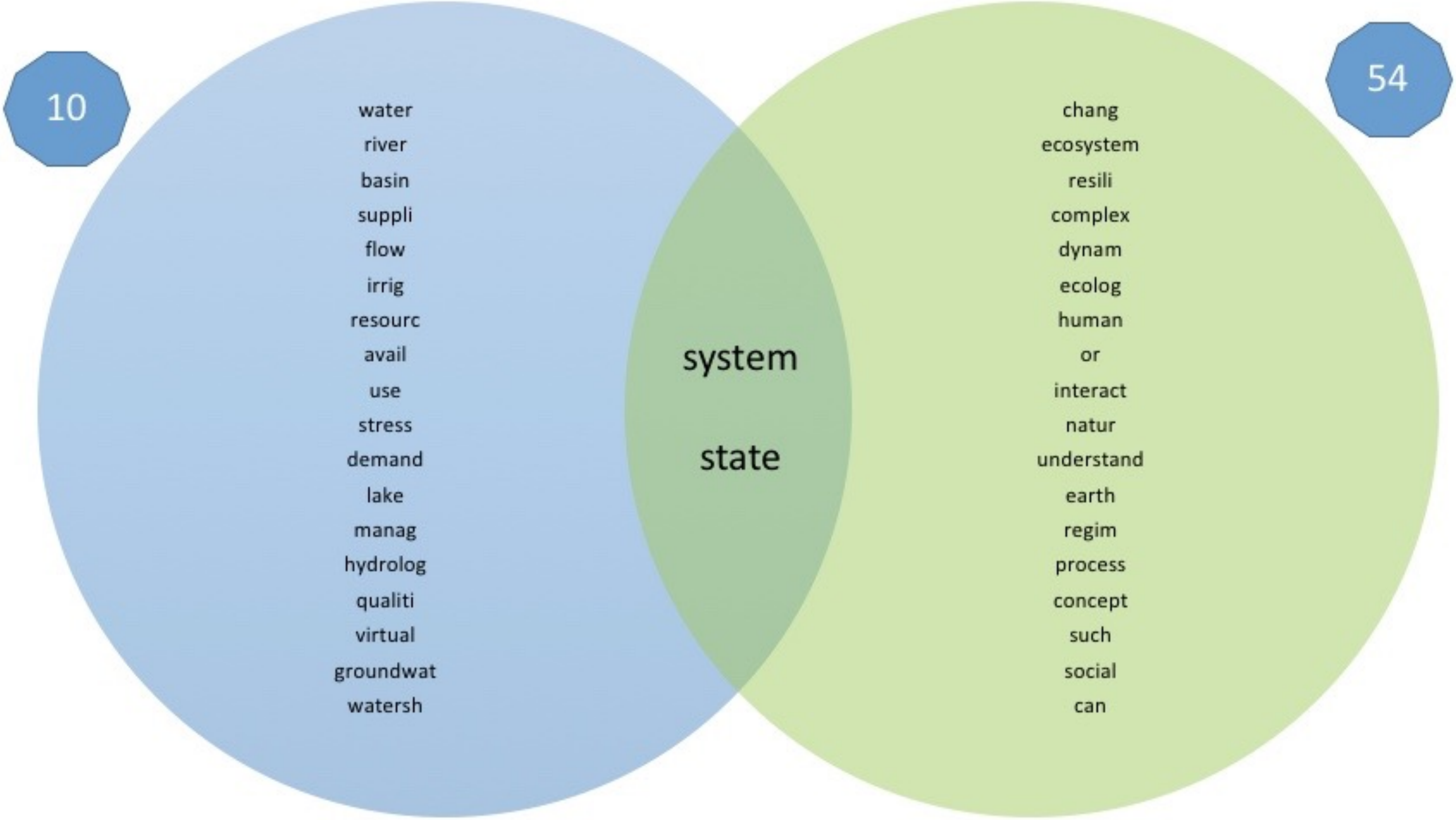
[a] *University of East Anglia, Norwich NR4 7TJ, UK*
[b] *King's College London, Strand, London WC2R 2LS, UK*
[c] *School of Oriental and African Studies, London, UK*

ARTICLE INFO

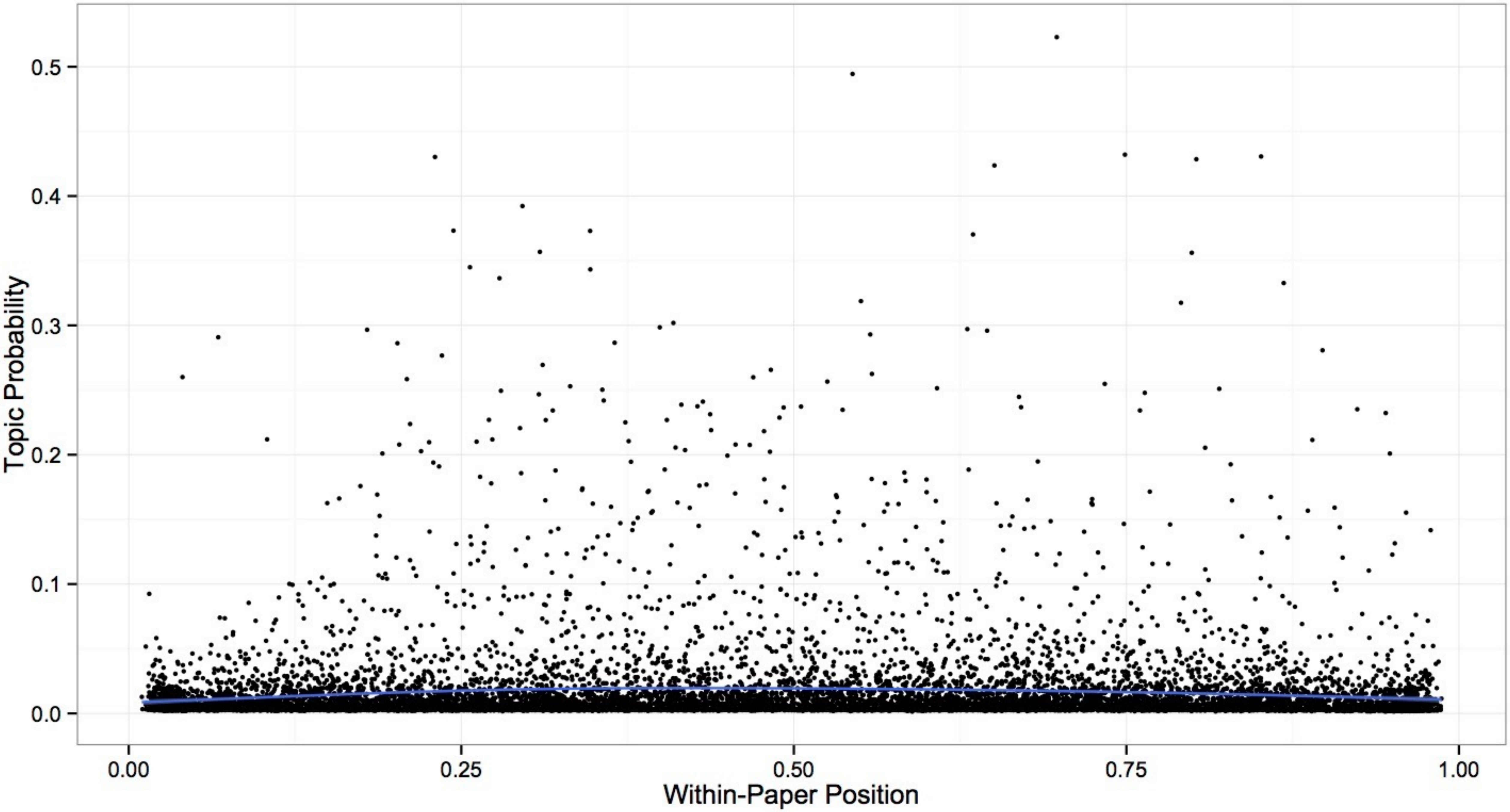
Article history:
Received 16 June 2009
Received in revised form 5 November 2009
Accepted 13 November 2009

ABSTRACT

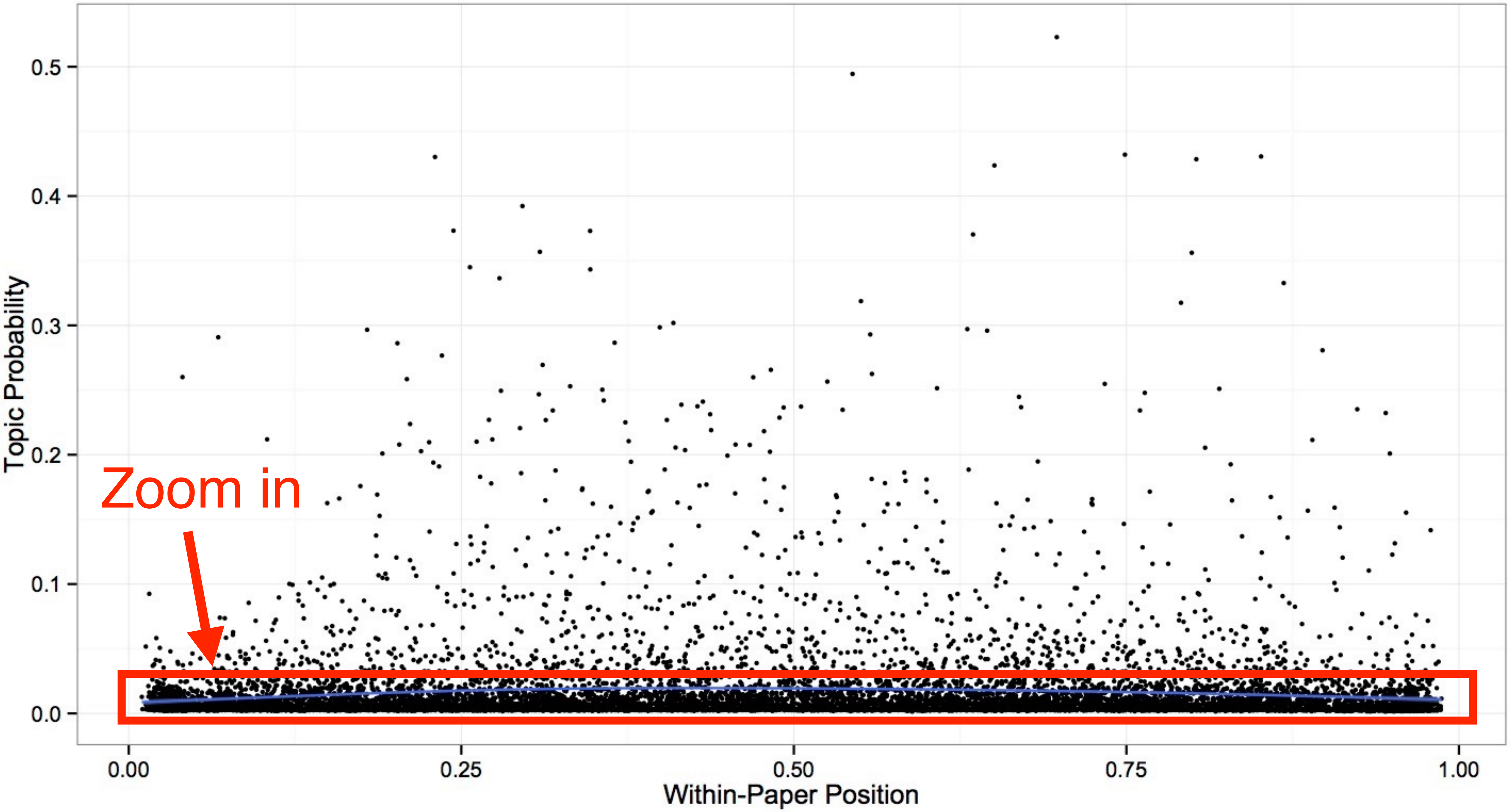
This paper interprets an initial approximation of the 'trade' in virtual water of Nile Basin states in terms of national water security. The virtual water content (on the basis of weight) of select recorded crop and livestock trade between 1998 and 2004 is provided, and analysed for each state separately, for the Southern Nile and Eastern Nile states as groups, and for the basin states as a whole. To the extent that the


10
54
water
river
basin
suppli
flow
irrig
resourc
avail
use
stress
demand
lake
manag
hydrolog
qualiti
virtual
groundwat
watersh
system
state
chang
ecosystem
resili
complex
dynam
ecolog
human
or
interact
natur
understand
earth
regim
process
concept
such
social
can

Within-Paper Topic Distribution of Topic 26

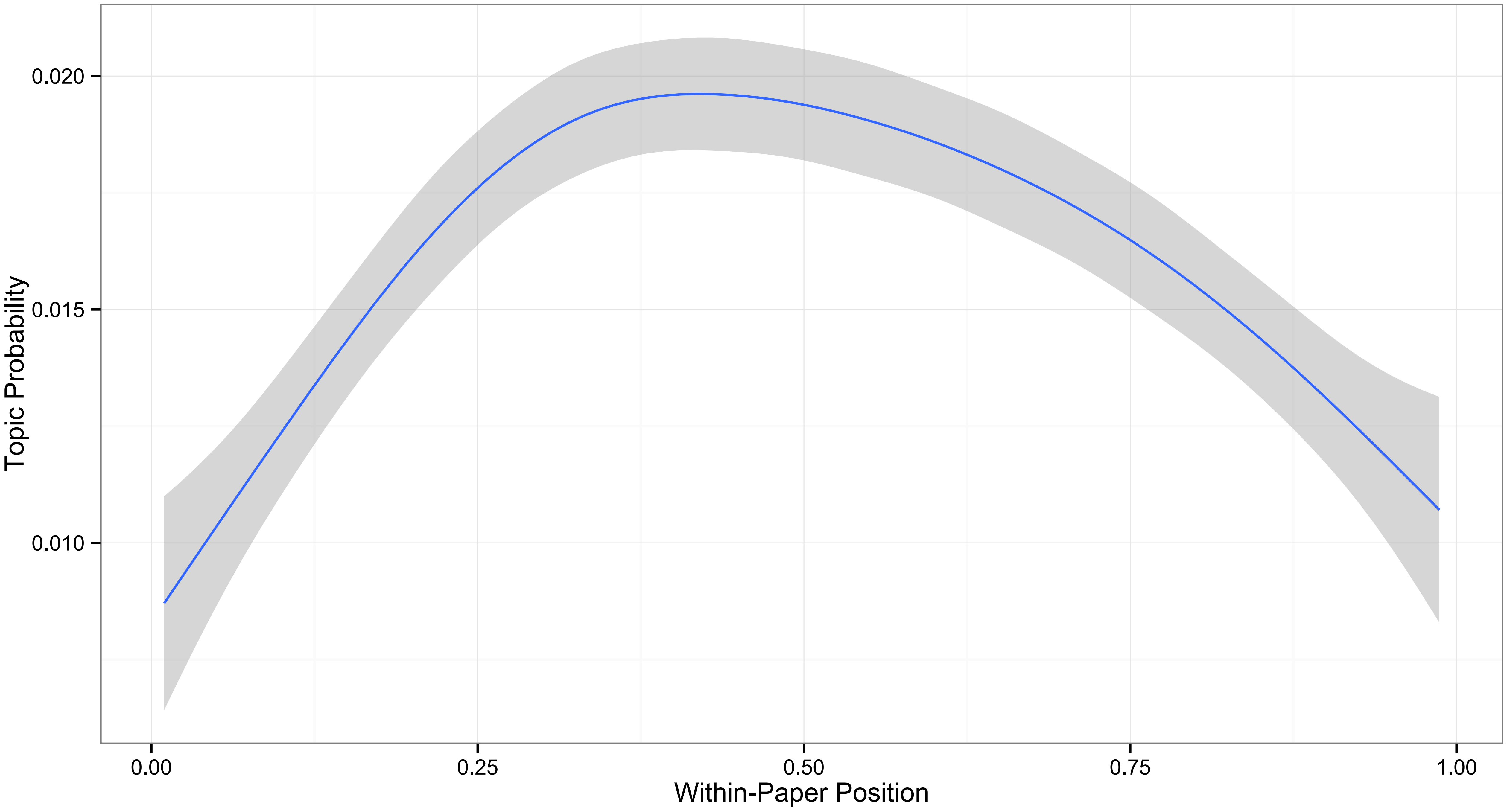


Within-Paper Topic Distribution of Topic 26

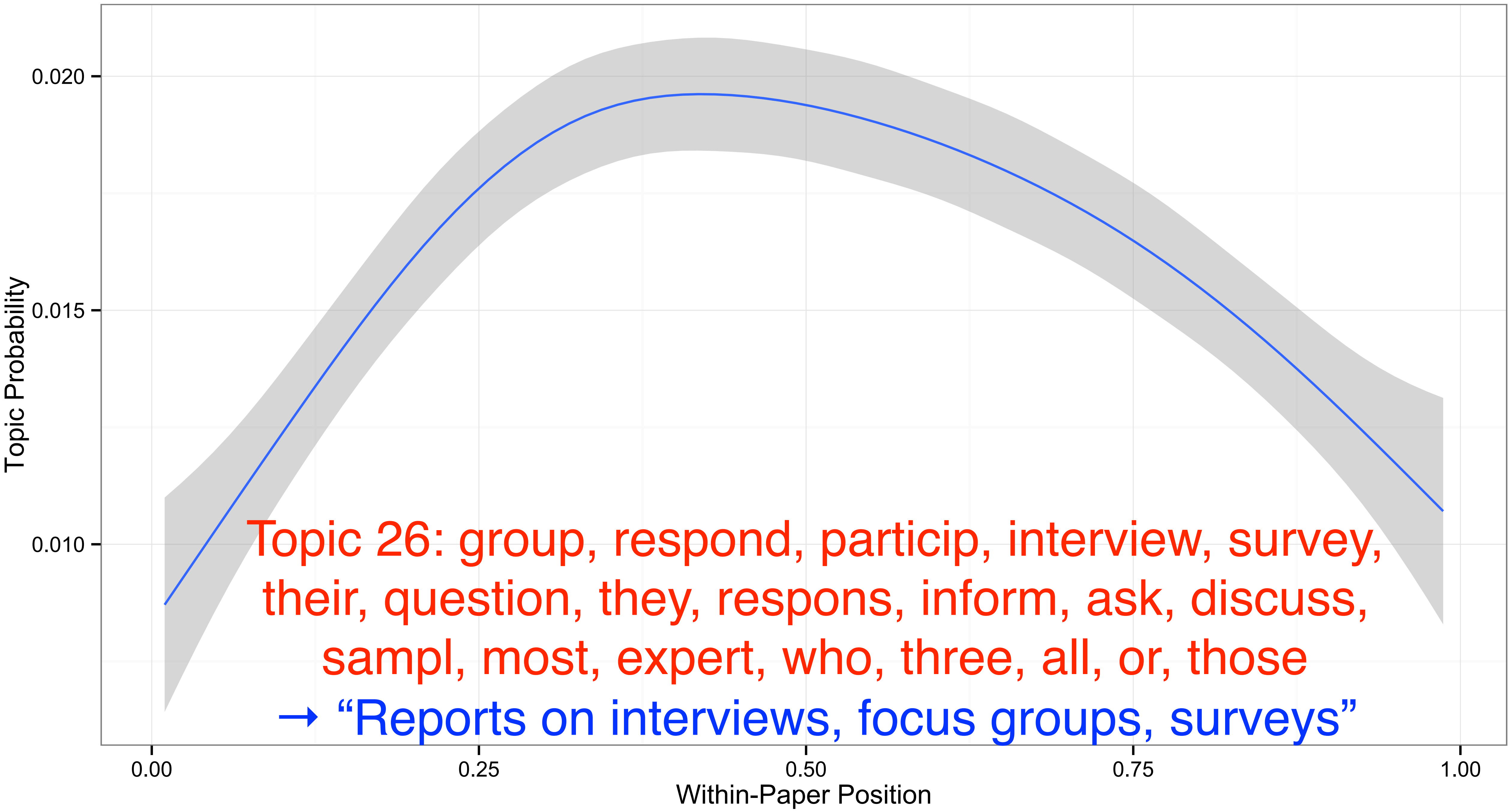


Zoom in

Within-Paper Topic Distribution of Topic 26



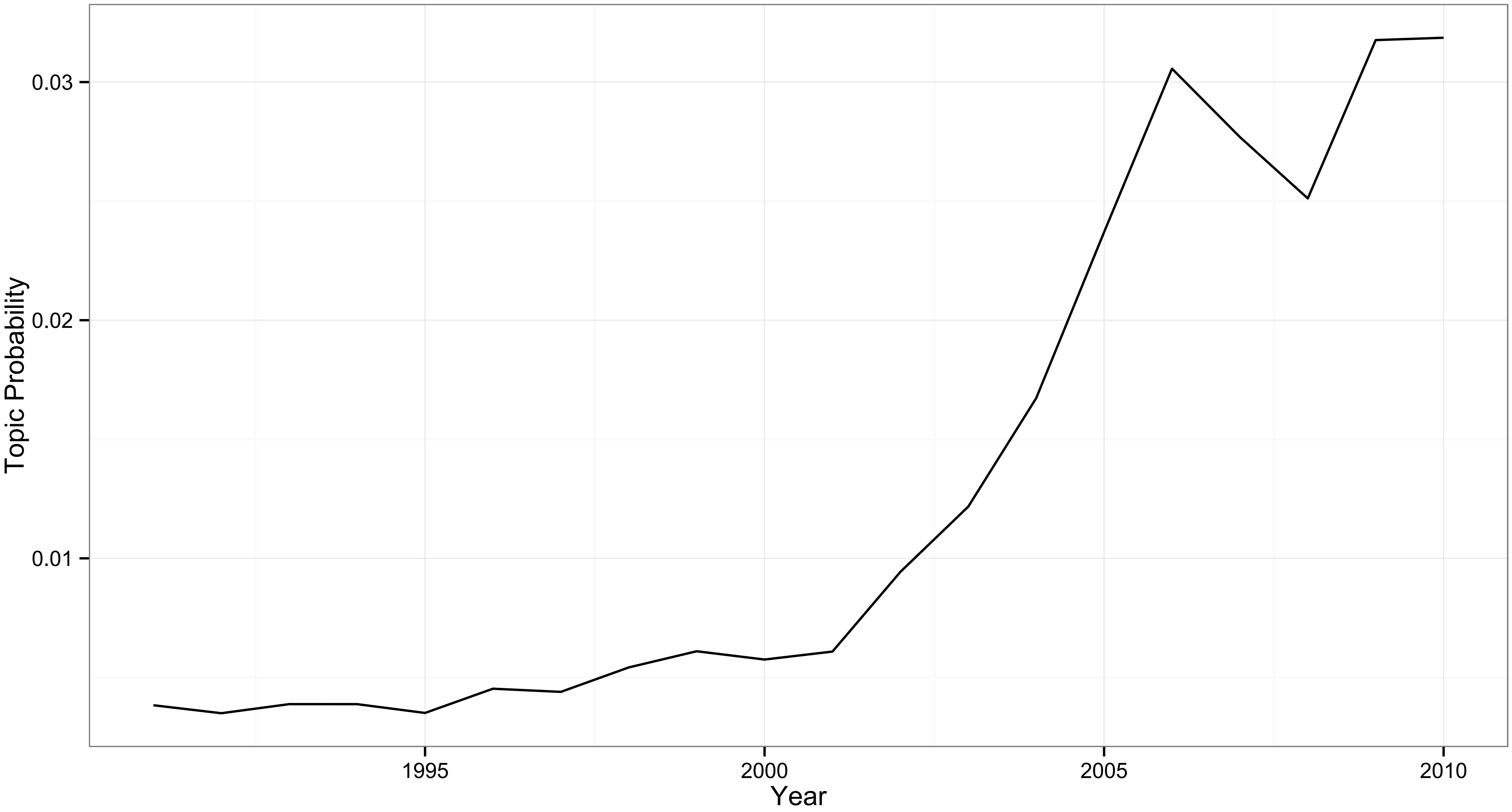
Within-Paper Topic Distribution of Topic 26



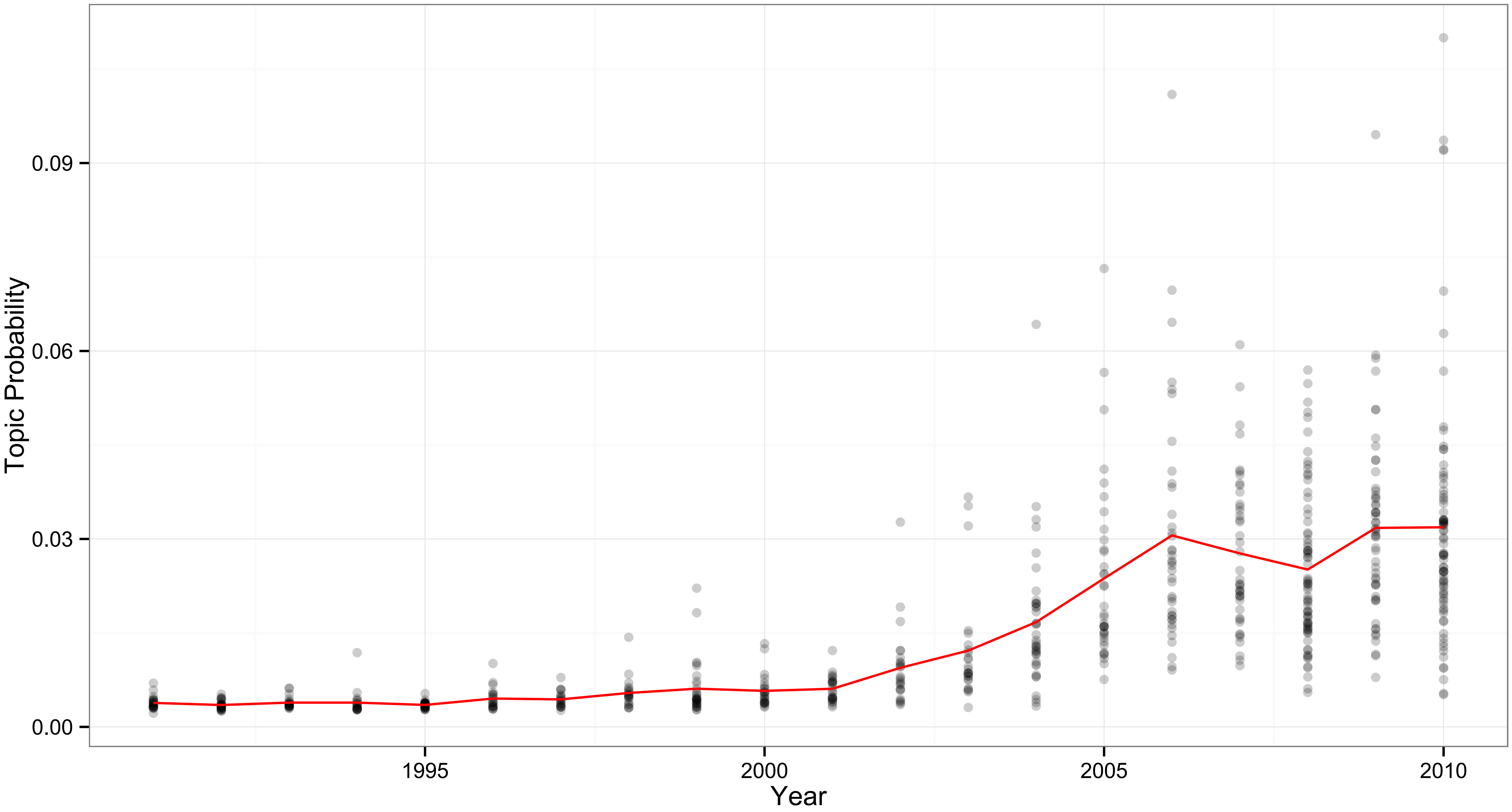
Topic 26: group, respond, particip, interview, survey, their, question, they, respons, inform, ask, discuss, sampl, most, expert, who, three, all, or, those

→ “Reports on interviews, focus groups, surveys”

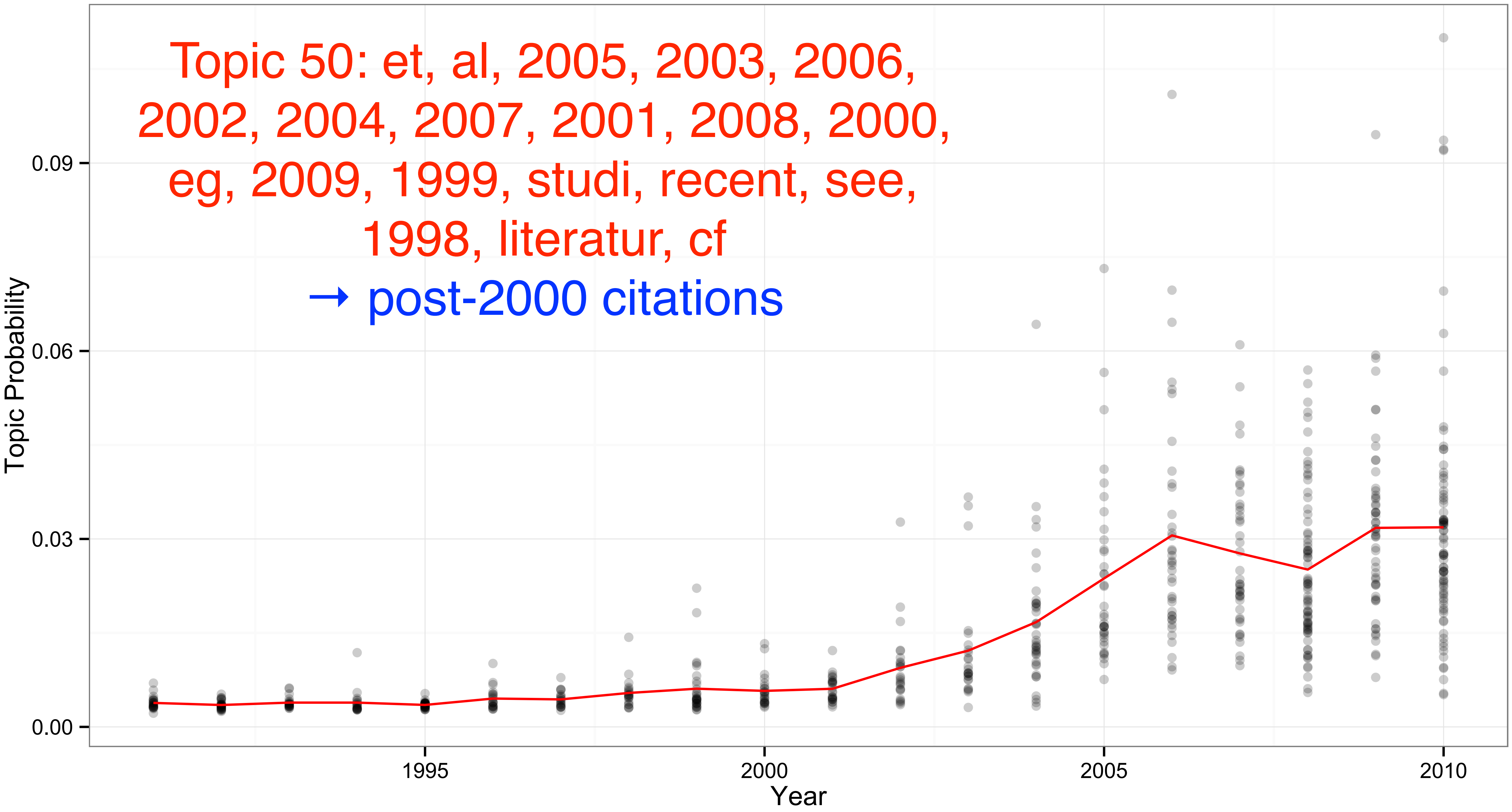
Chronological Change of Topic 50



Chronological Change of Topic 50



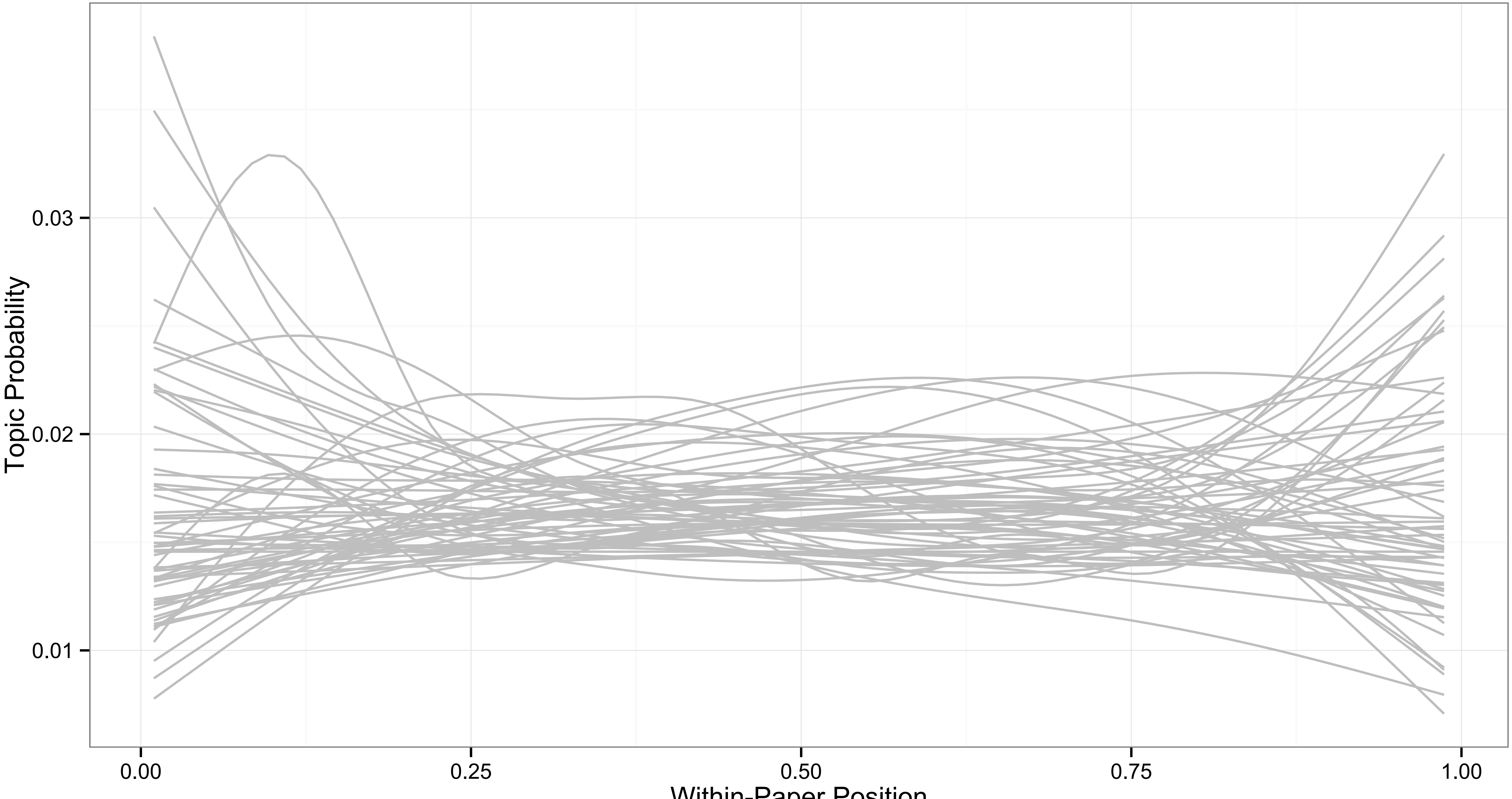
Chronological Change of Topic 50



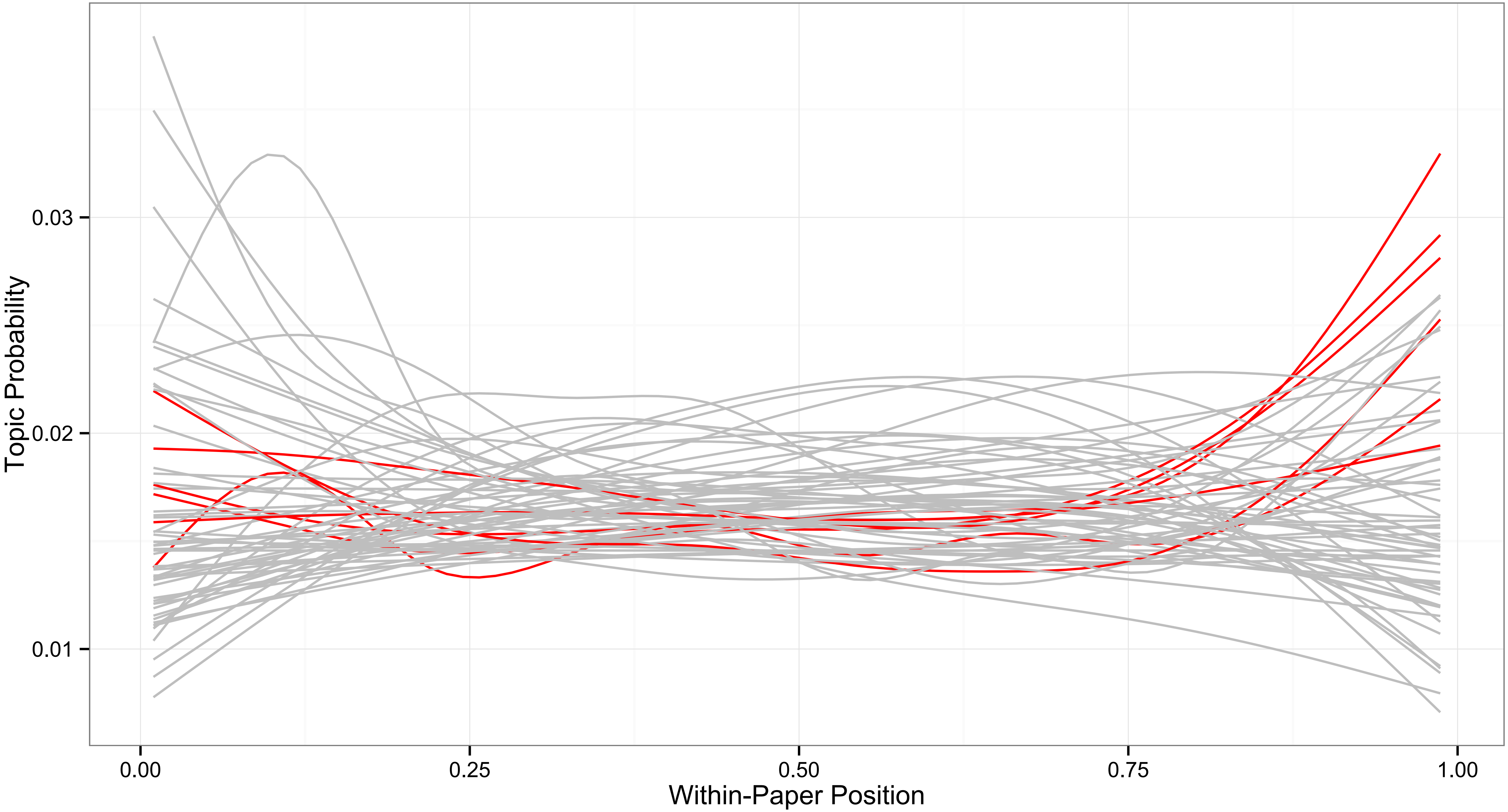
Topic 50: et, al, 2005, 2003, 2006, 2002, 2004, 2007, 2001, 2008, 2000, eg, 2009, 1999, studi, recent, see, 1998, literatur, cf

→ post-2000 citations

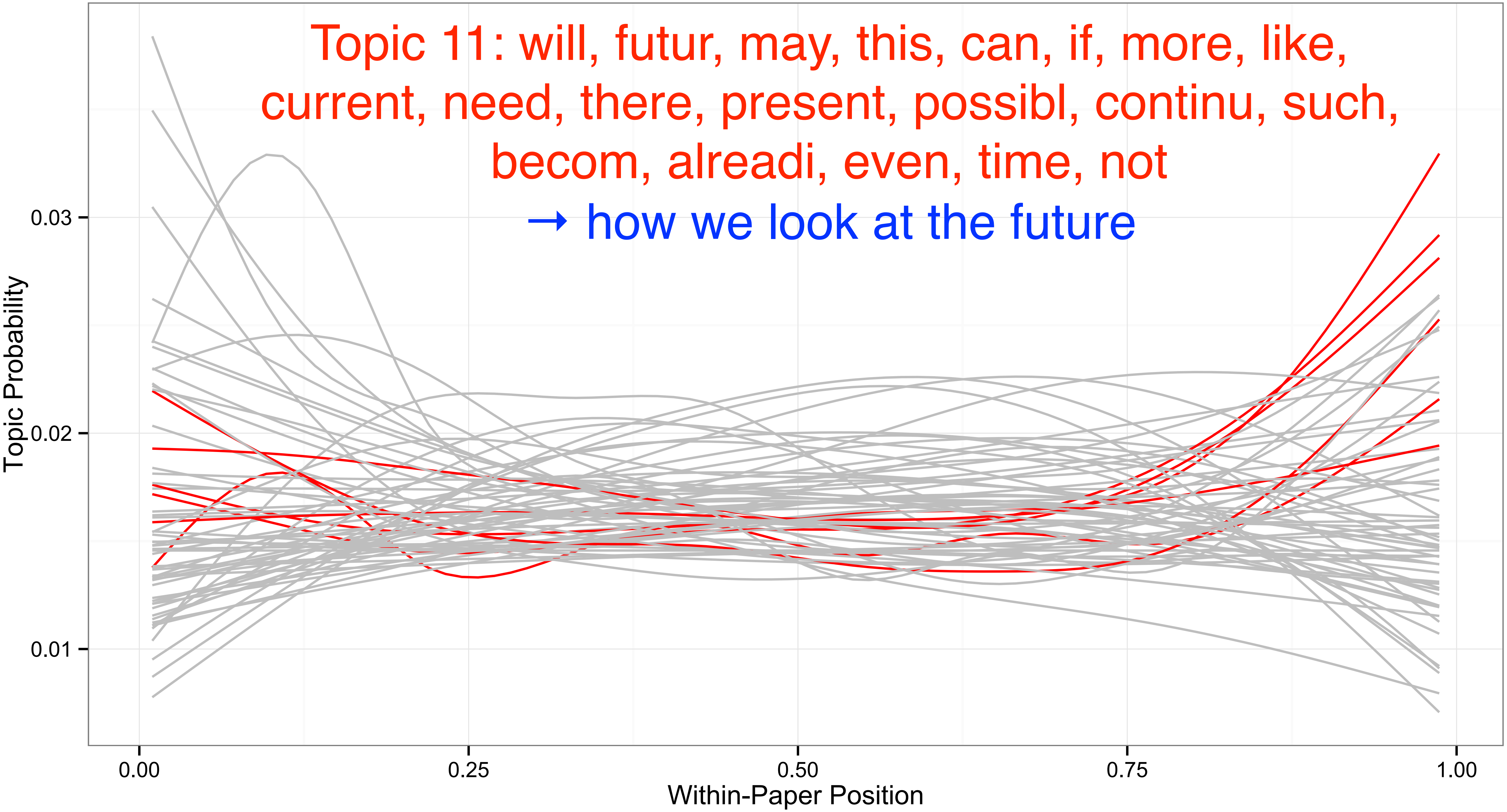
Within-Paper Topic Distribution



Within-Paper Topic Distribution



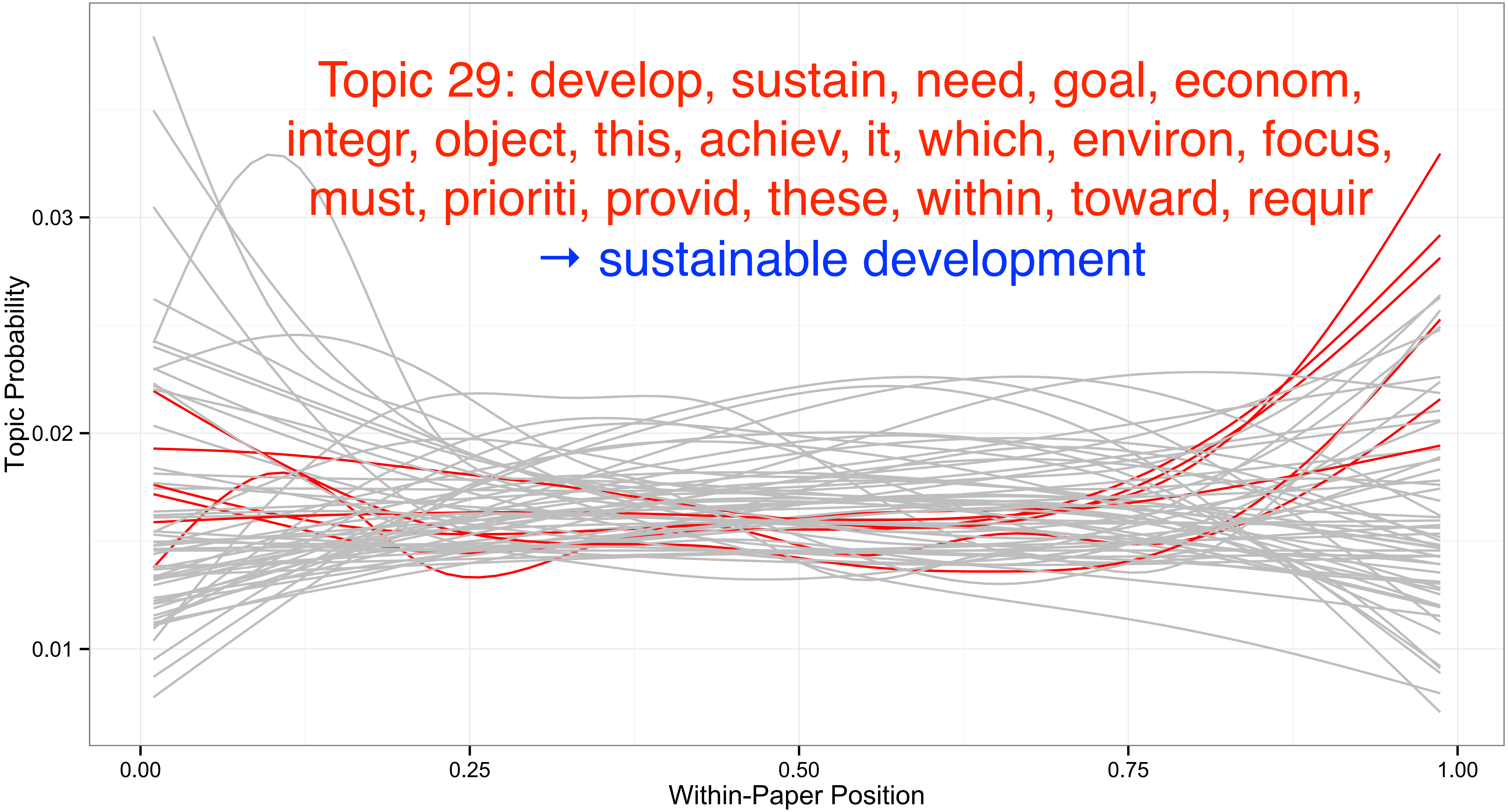
Within-Paper Topic Distribution



Topic 11: will, futur, may, this, can, if, more, like, current, need, there, present, possibl, continu, such, becom, alreadi, even, time, not

→ how we look at the future

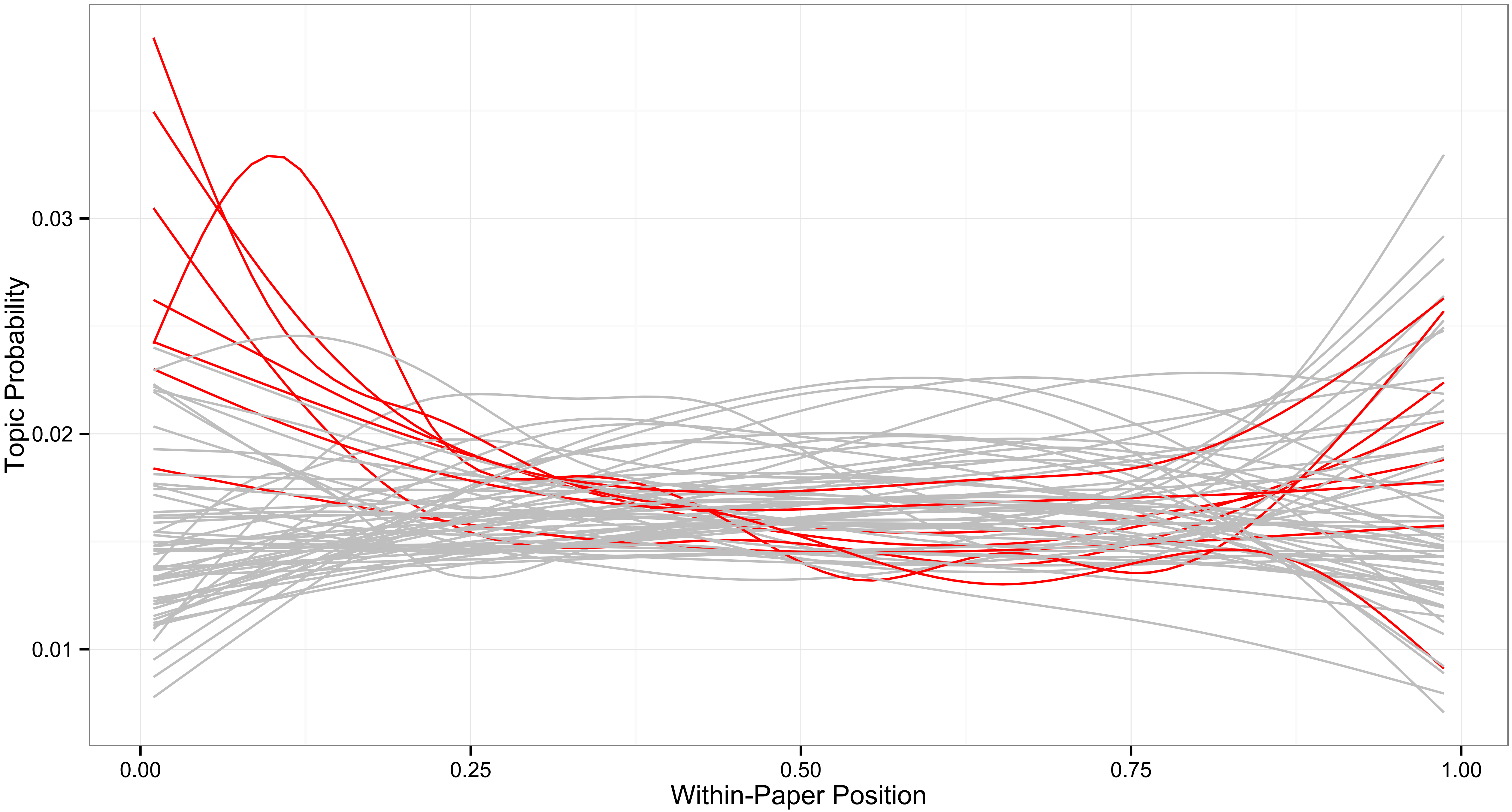
Within-Paper Topic Distribution



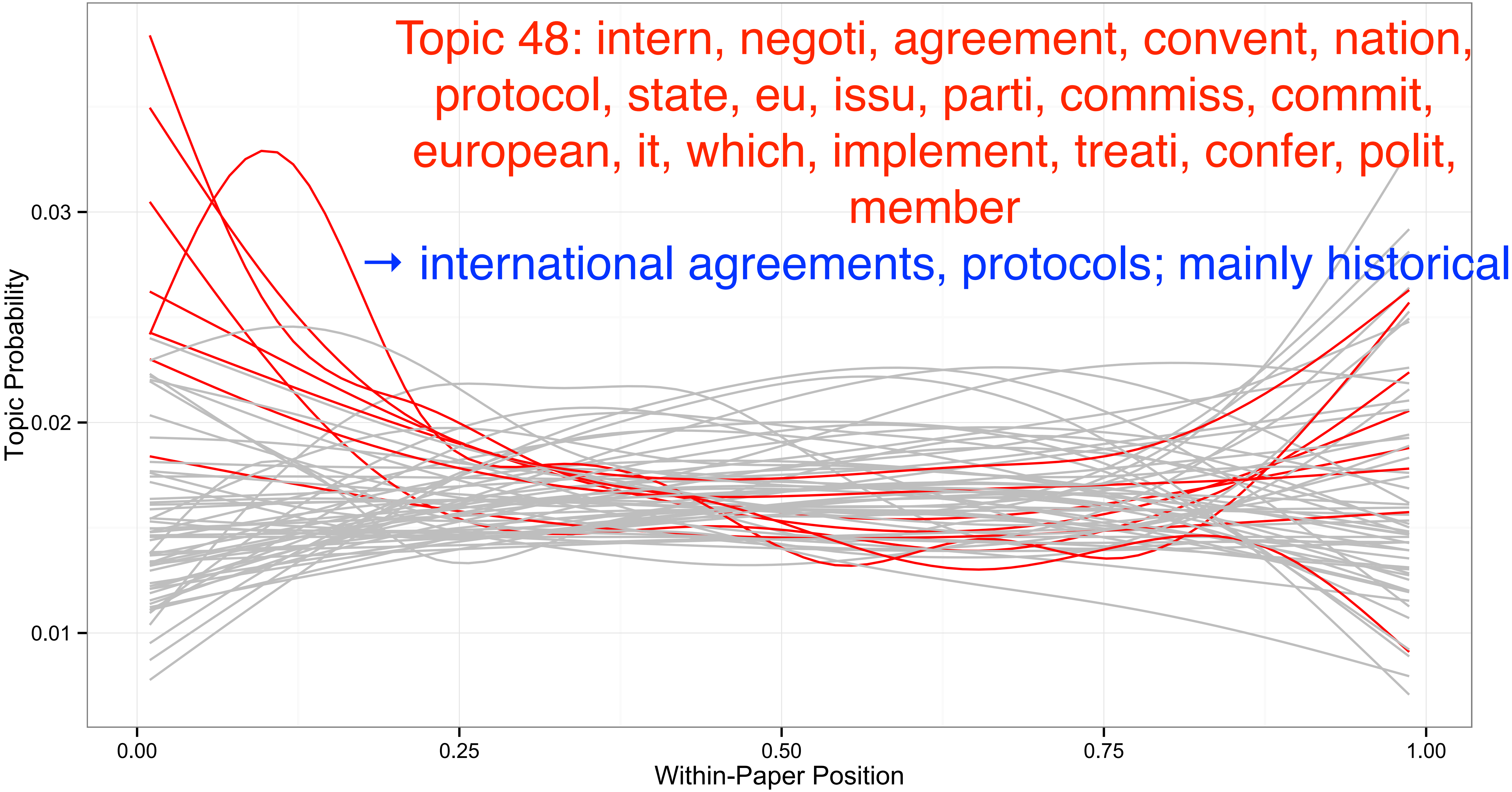
Topic 29: develop, sustain, need, goal, econom, integr, object, this, achiev, it, which, environ, focus, must, prioriti, provid, these, within, toward, requir

→ sustainable development

Within-Paper Topic Distribution



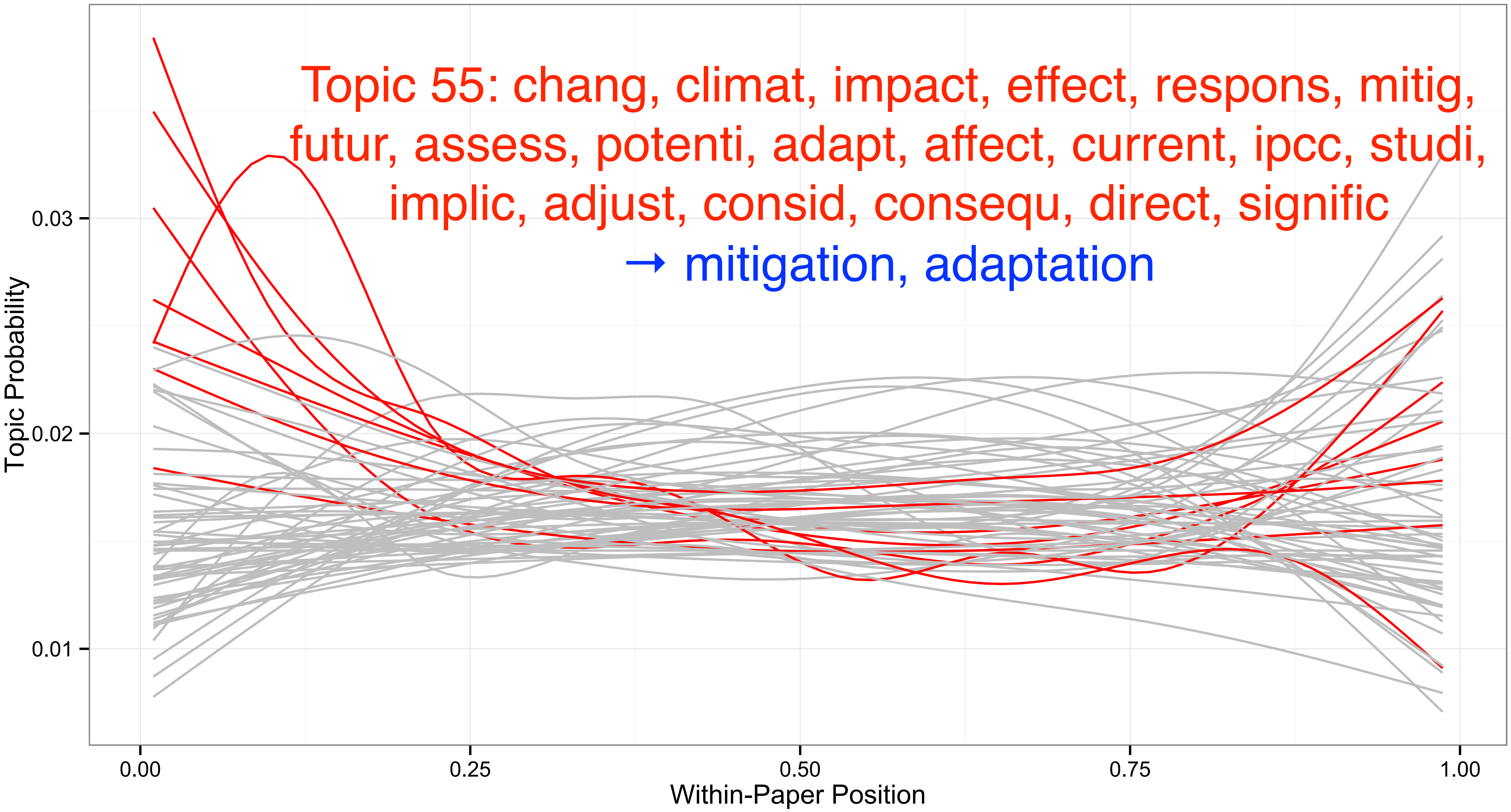
Within-Paper Topic Distribution



Topic 48: intern, negoti, agreement, convent, nation, protocol, state, eu, issu, parti, commiss, commit, european, it, which, implement, treati, confer, polit, member

→ international agreements, protocols; mainly historical

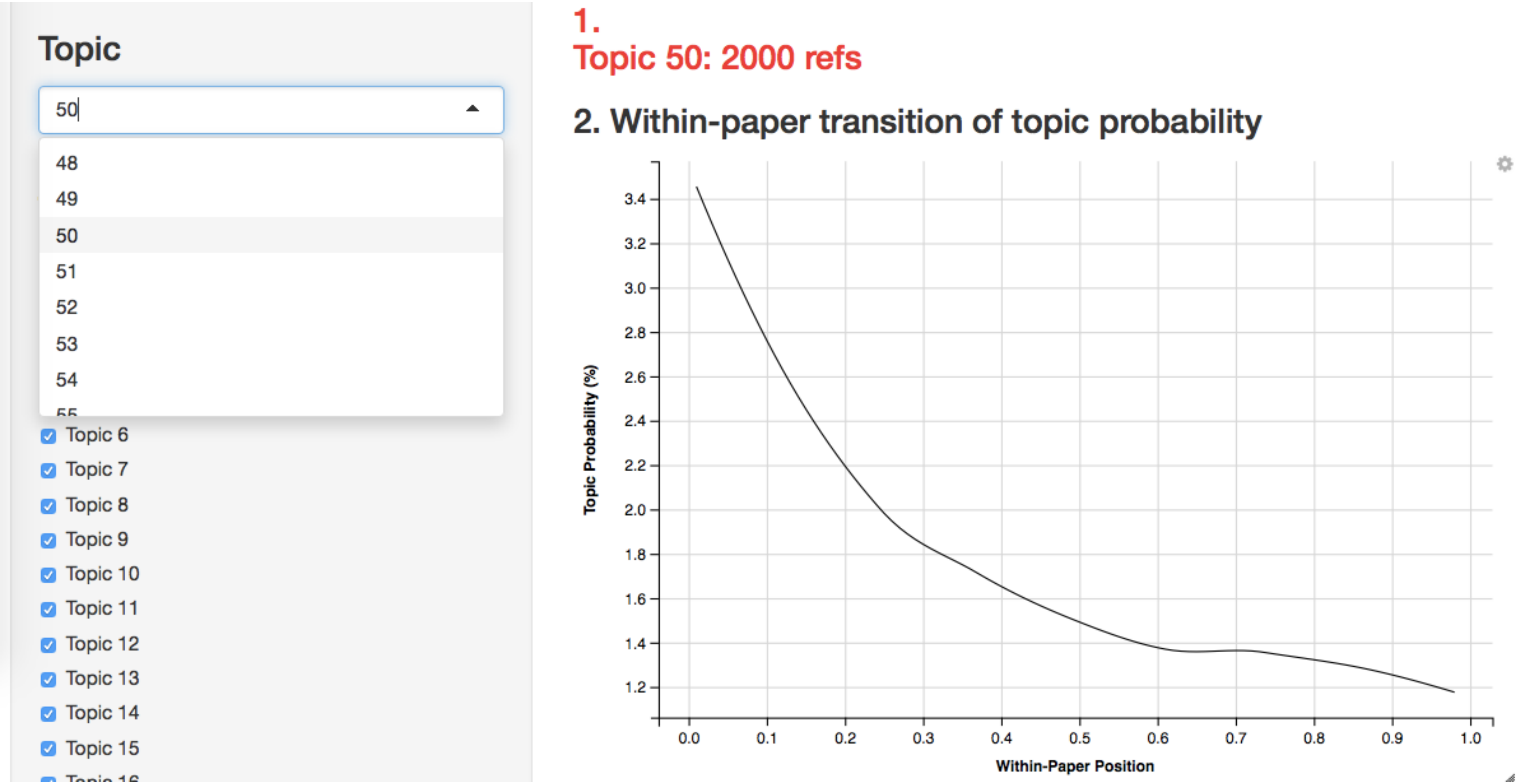
Within-Paper Topic Distribution



Topic 55: chang, climat, impact, effect, respons, mitig, futur, assess, potenti, adapt, affect, current, ipcc, studi, implic, adjust, consid, consequ, direct, signific

→ mitigation, adaptation

Interactive Visualization Tool



Topic

50

☑ Topic 6

☑ Topic 7

☑ Topic 8

☑ Topic 9

☑ Topic 10

☑ Topic 11

☑ Topic 12

☑ Topic 13

☑ Topic 14

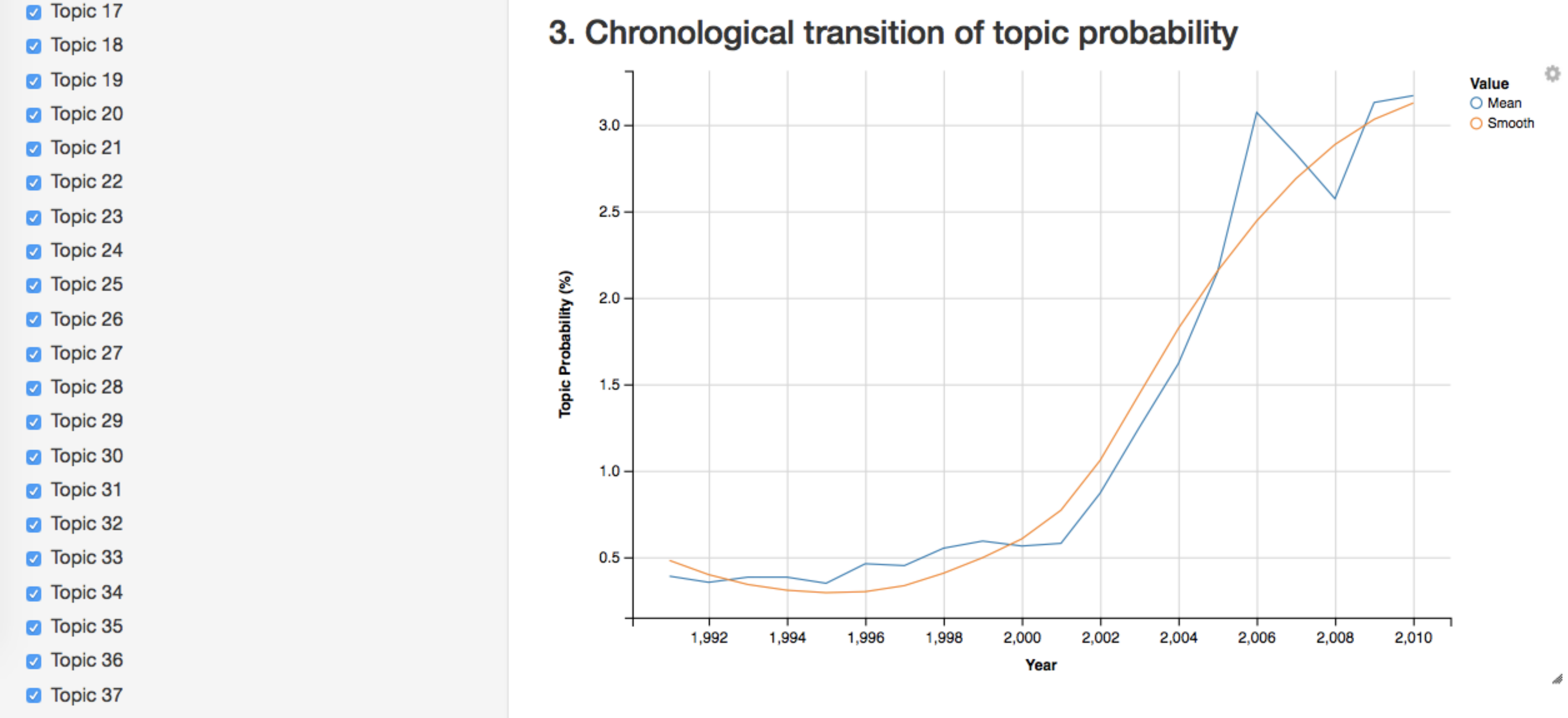
☑ Topic 15

☑ Topic 16

1.
Topic 50: 2000 refs

2. Within-paper transition of topic probability

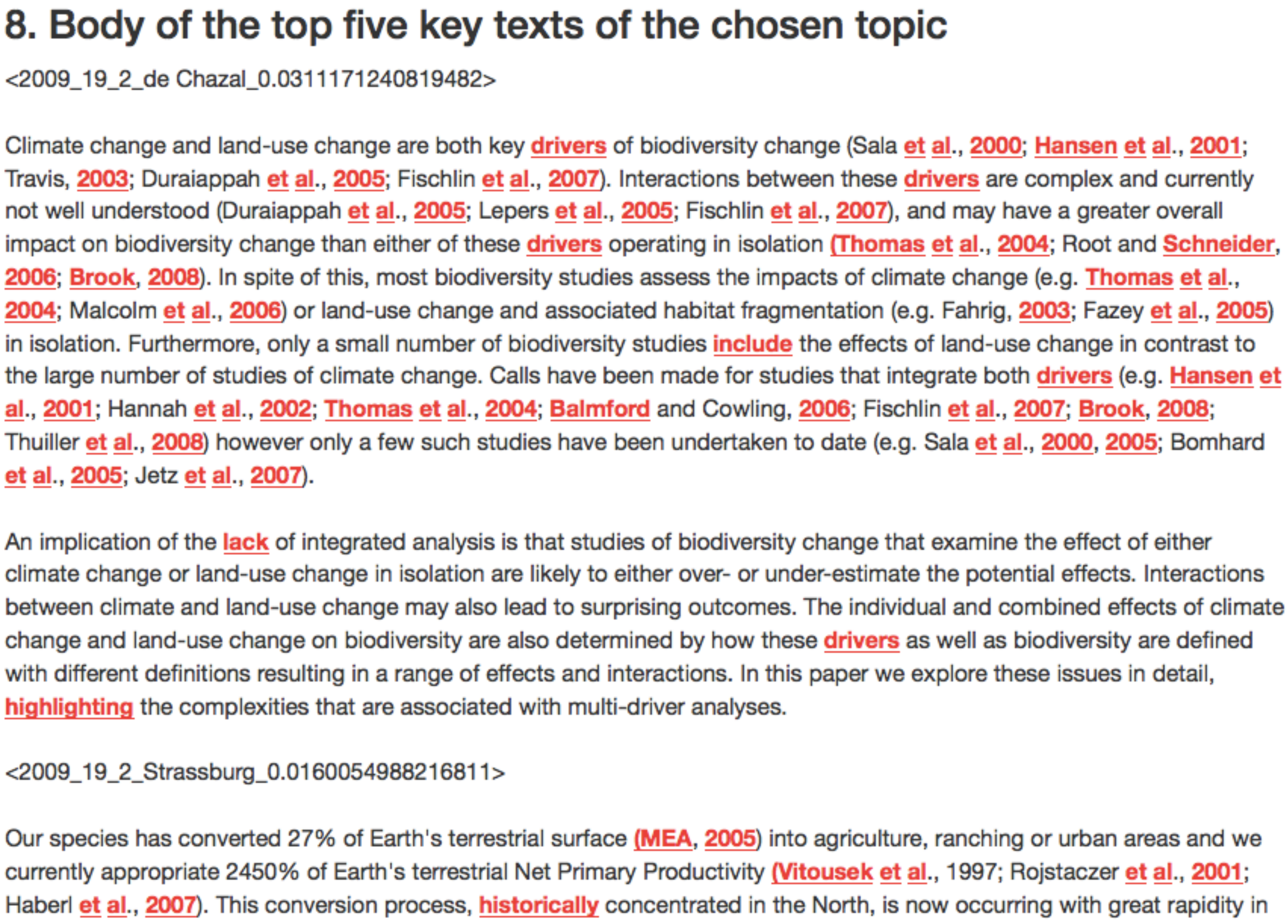
Interactive Visualization Tool



- ☑ Topic 17
- ☑ Topic 18
- ☑ Topic 19
- ☑ Topic 20
- ☑ Topic 21
- ☑ Topic 22
- ☑ Topic 23
- ☑ Topic 24
- ☑ Topic 25
- ☑ Topic 26
- ☑ Topic 27
- ☑ Topic 28
- ☑ Topic 29
- ☑ Topic 30
- ☑ Topic 31
- ☑ Topic 32
- ☑ Topic 33
- ☑ Topic 34
- ☑ Topic 35
- ☑ Topic 36
- ☑ Topic 37

3. Chronological transition of topic probability

Interactive Visualization Tool



8. Body of the top five key texts of the chosen topic

<2009_19_2_de Chazal_0.0311171240819482>

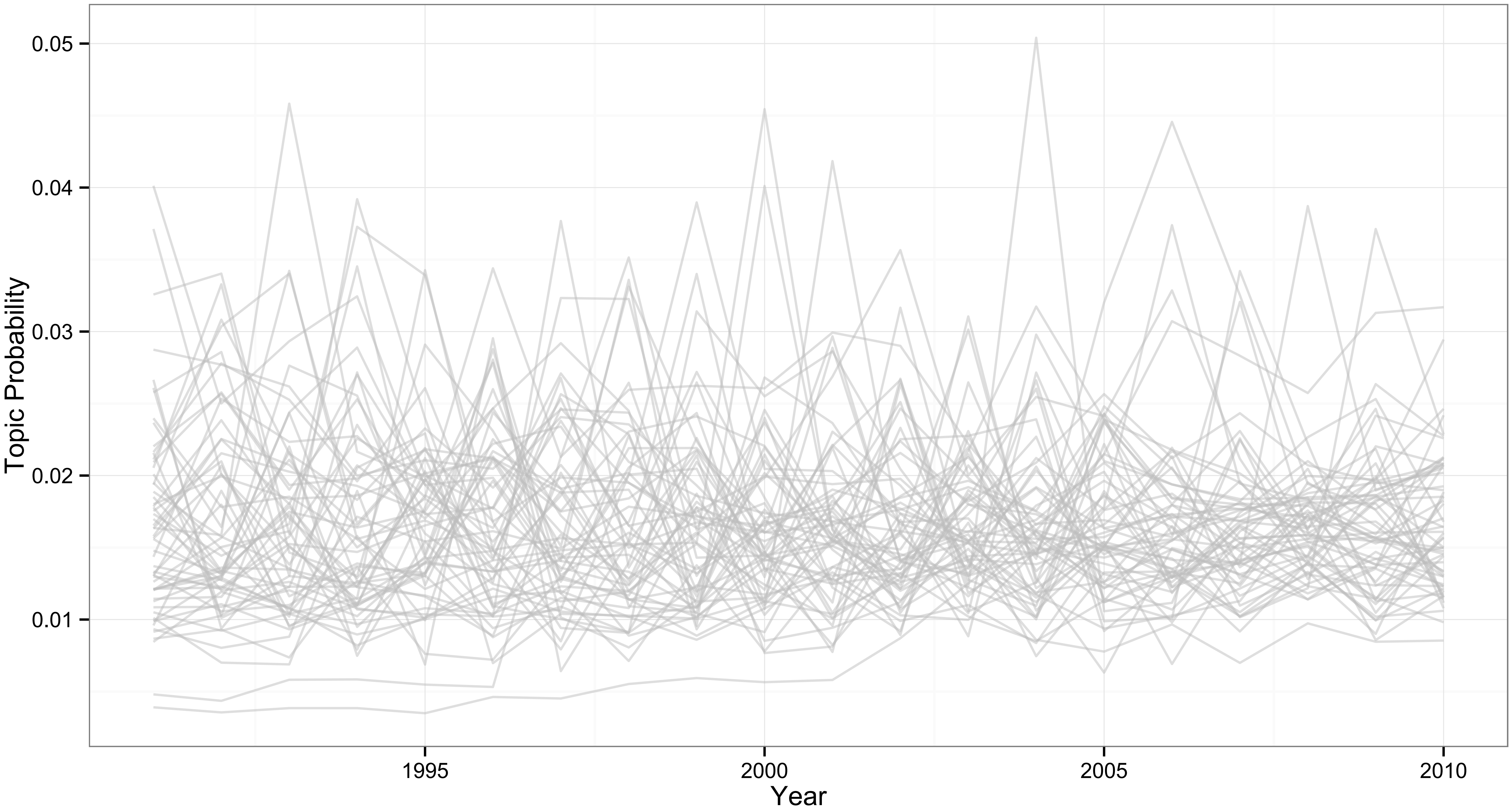
Climate change and land-use change are both key drivers of biodiversity change (Sala et al., 2000; Hansen et al., 2001; Travis, 2003; Duraiappah et al., 2005; Fischlin et al., 2007). Interactions between these drivers are complex and currently not well understood (Duraiappah et al., 2005; Lepers et al., 2005; Fischlin et al., 2007), and may have a greater overall impact on biodiversity change than either of these drivers operating in isolation (Thomas et al., 2004; Root and Schneider, 2006; Brook, 2008). In spite of this, most biodiversity studies assess the impacts of climate change (e.g. Thomas et al., 2004; Malcolm et al., 2006) or land-use change and associated habitat fragmentation (e.g. Fahrig, 2003; Fazey et al., 2005) in isolation. Furthermore, only a small number of biodiversity studies include the effects of land-use change in contrast to the large number of studies of climate change. Calls have been made for studies that integrate both drivers (e.g. Hansen et al., 2001; Hannah et al., 2002; Thomas et al., 2004; Balmford and Cowling, 2006; Fischlin et al., 2007; Brook, 2008; Thuiller et al., 2008) however only a few such studies have been undertaken to date (e.g. Sala et al., 2000, 2005; Bomhard et al., 2005; Jetz et al., 2007).

An implication of the lack of integrated analysis is that studies of biodiversity change that examine the effect of either climate change or land-use change in isolation are likely to either over- or under-estimate the potential effects. Interactions between climate and land-use change may also lead to surprising outcomes. The individual and combined effects of climate change and land-use change on biodiversity are also determined by how these drivers as well as biodiversity are defined with different definitions resulting in a range of effects and interactions. In this paper we explore these issues in detail, highlighting the complexities that are associated with multi-driver analyses.

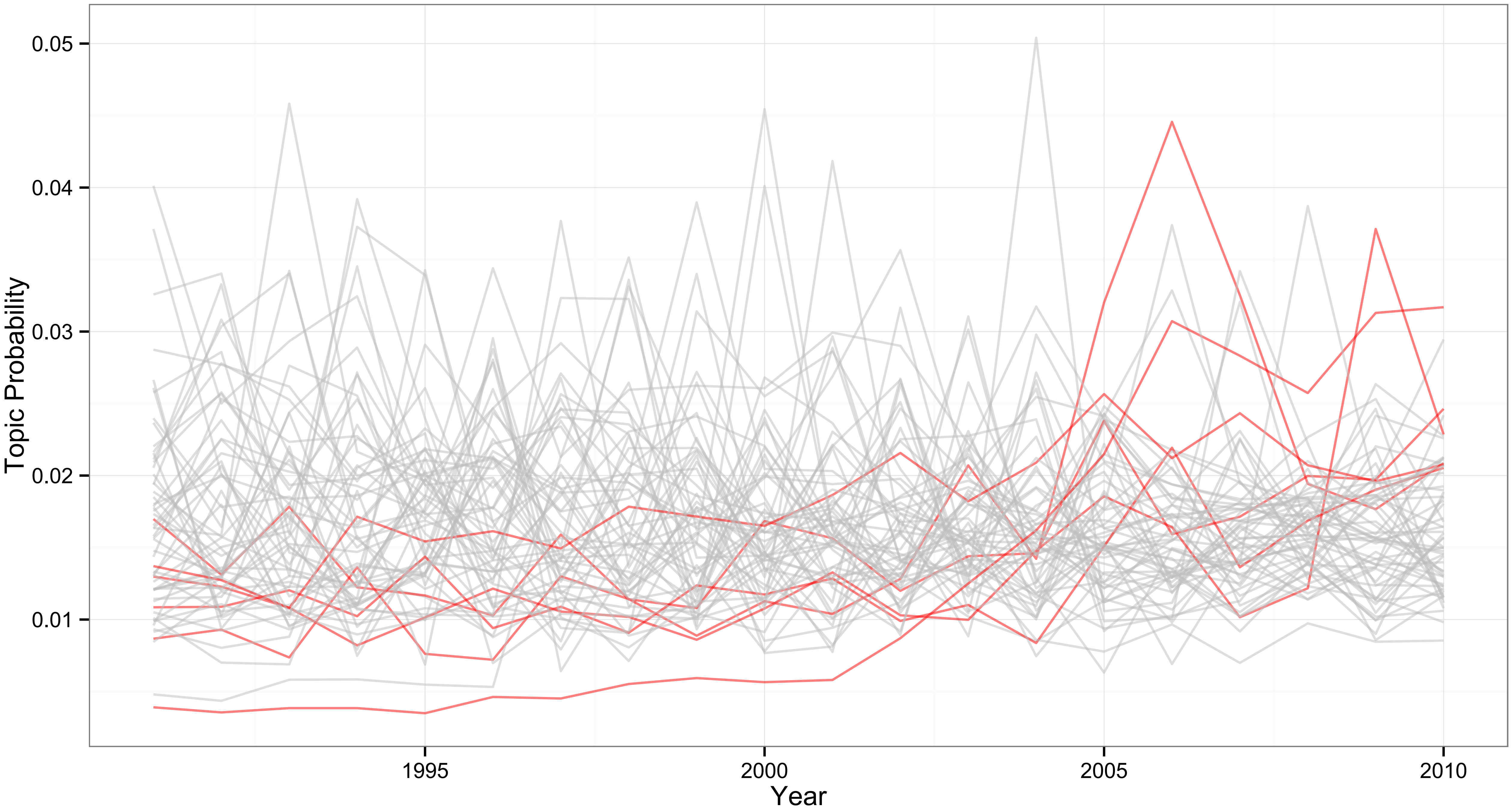
<2009_19_2_Strassburg_0.0160054988216811>

Our species has converted 27% of Earth's terrestrial surface (MEA, 2005) into agriculture, ranching or urban areas and we currently appropriate 2450% of Earth's terrestrial Net Primary Productivity (Vitousek et al., 1997; Rojstaczer et al., 2001; Haberl et al., 2007). This conversion process, historically concentrated in the North, is now occurring with great rapidity in

Chronological Topic Transition



Chronological Topic Transition



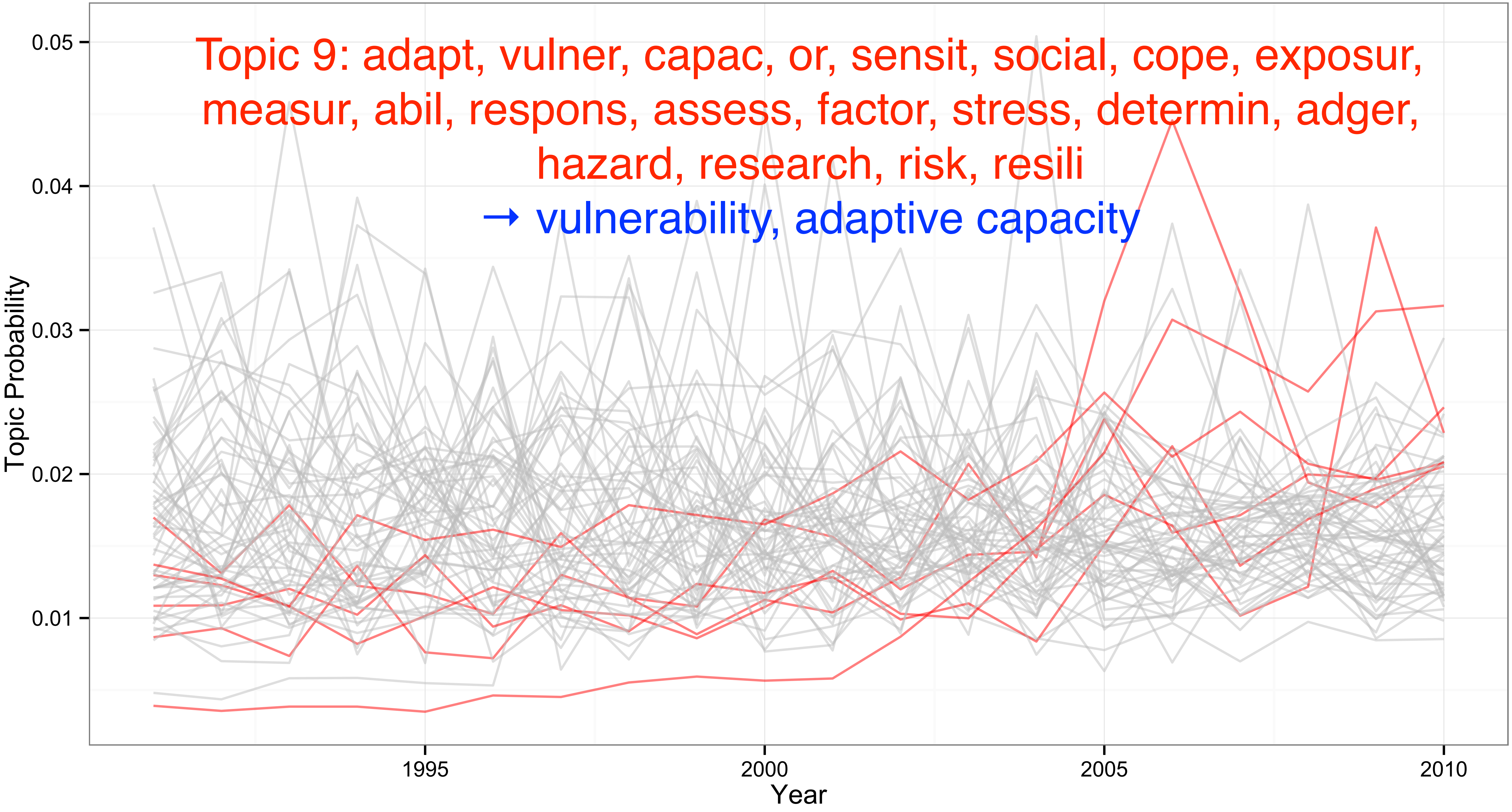
Increasing Topics

- Topic 9
    - adapt, vulner, capac, or, sensit, social, cope, exposur, measur, abil, respons, assess, factor, stress, determin, adger, hazard, research, risk, resili

      → vulnerability, adaptive capacity
- Topic 24
    - discours, point, articl, this, media, public, report, issu, frame, us, debat, coverag, such, 96, new, scientif, influenc, 2008, time, 2007

      → media and public discourse, and reviews of scientific literature

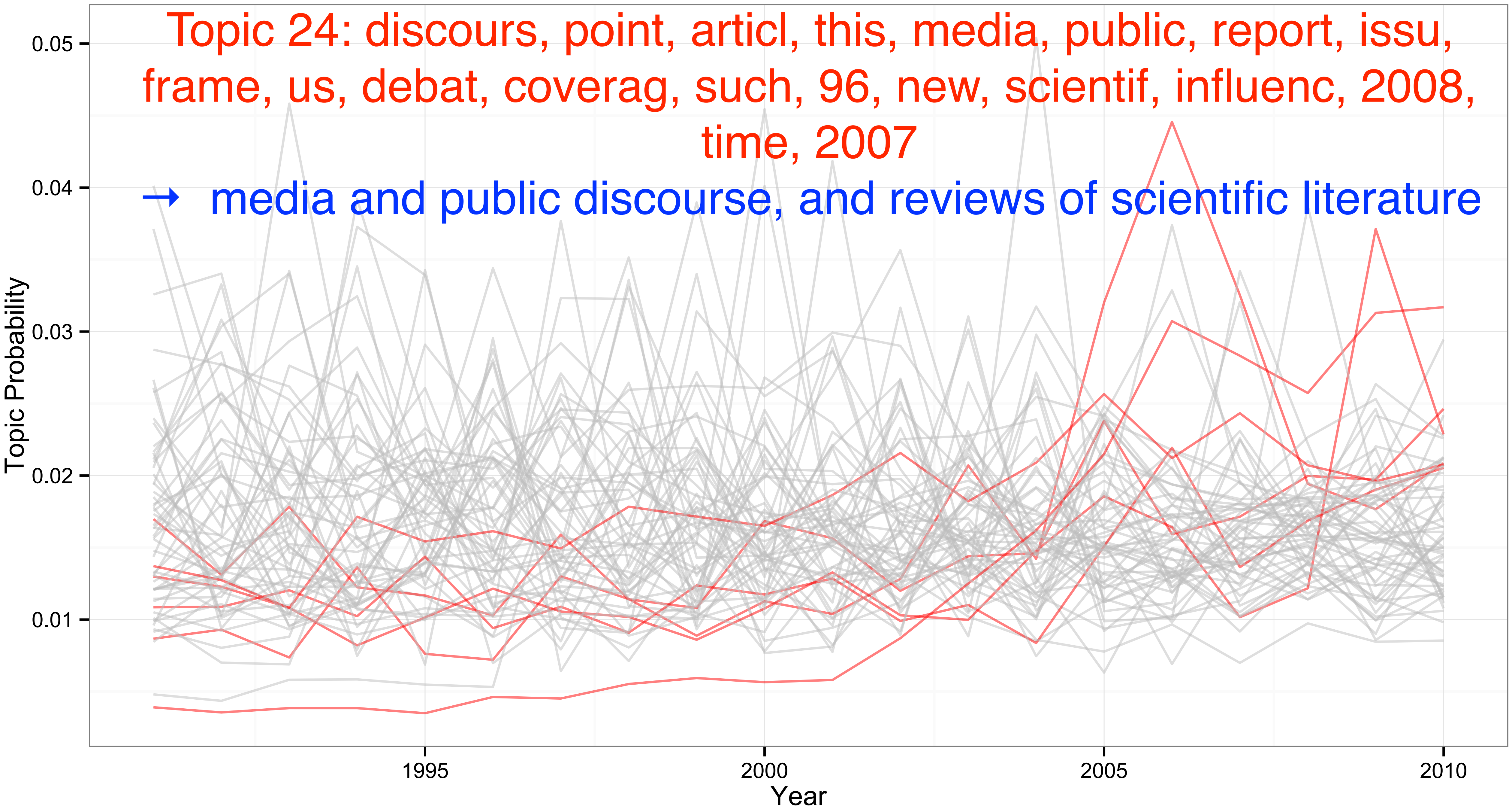
Chronological Topic Transition



Topic 9: adapt, vulner, capac, or, sensit, social, cope, exposur, measur, abil, respons, assess, factor, stress, determin, adger, hazard, research, risk, resili

→ vulnerability, adaptive capacity

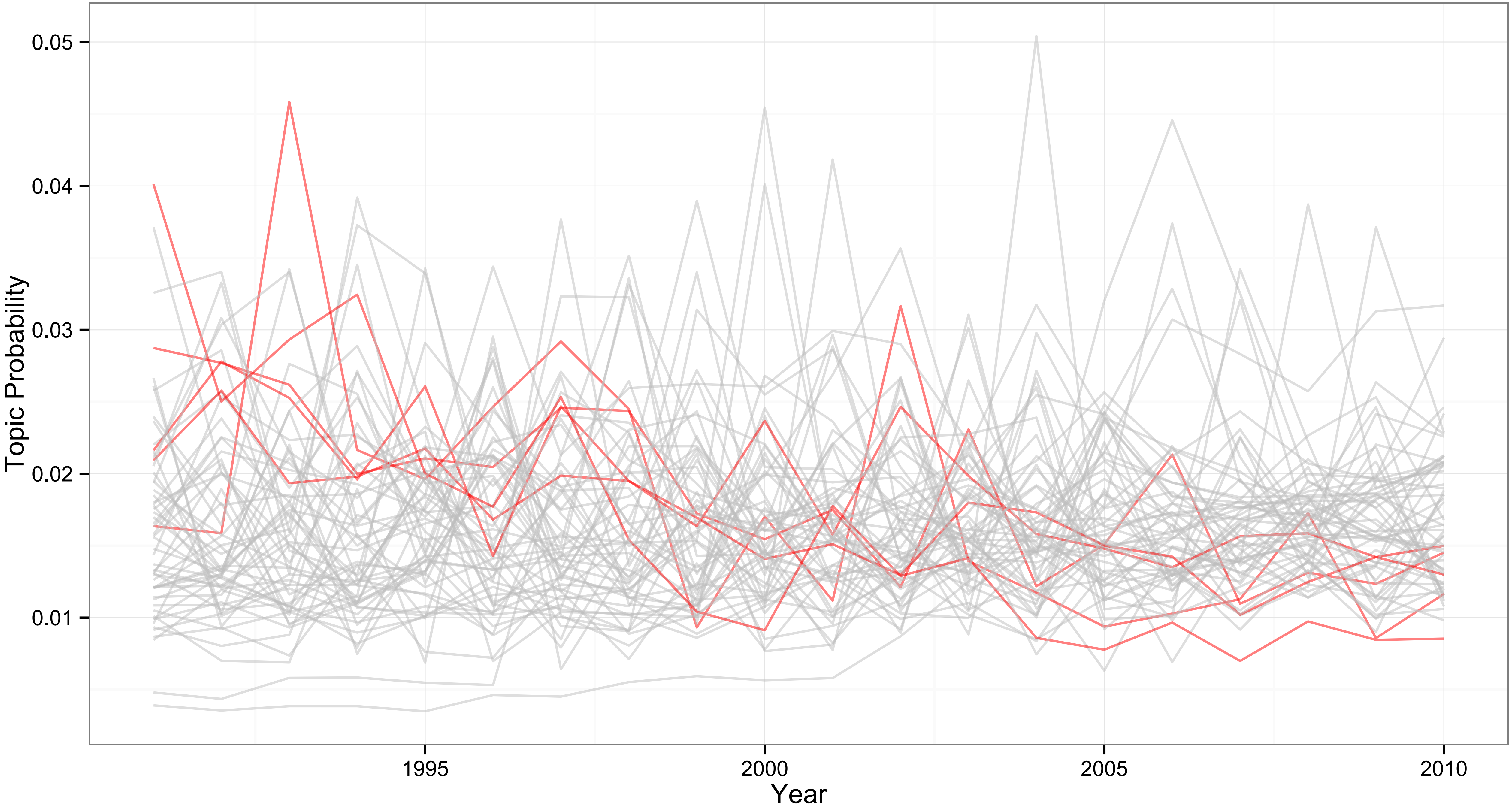
Chronological Topic Transition



Topic 24: discours, point, articl, this, media, public, report, issu, frame, us, debat, coverag, such, 96, new, scientif, influenc, 2008, time, 2007

→ media and public discourse, and reviews of scientific literature

Chronological Topic Transition



Decreasing Topics

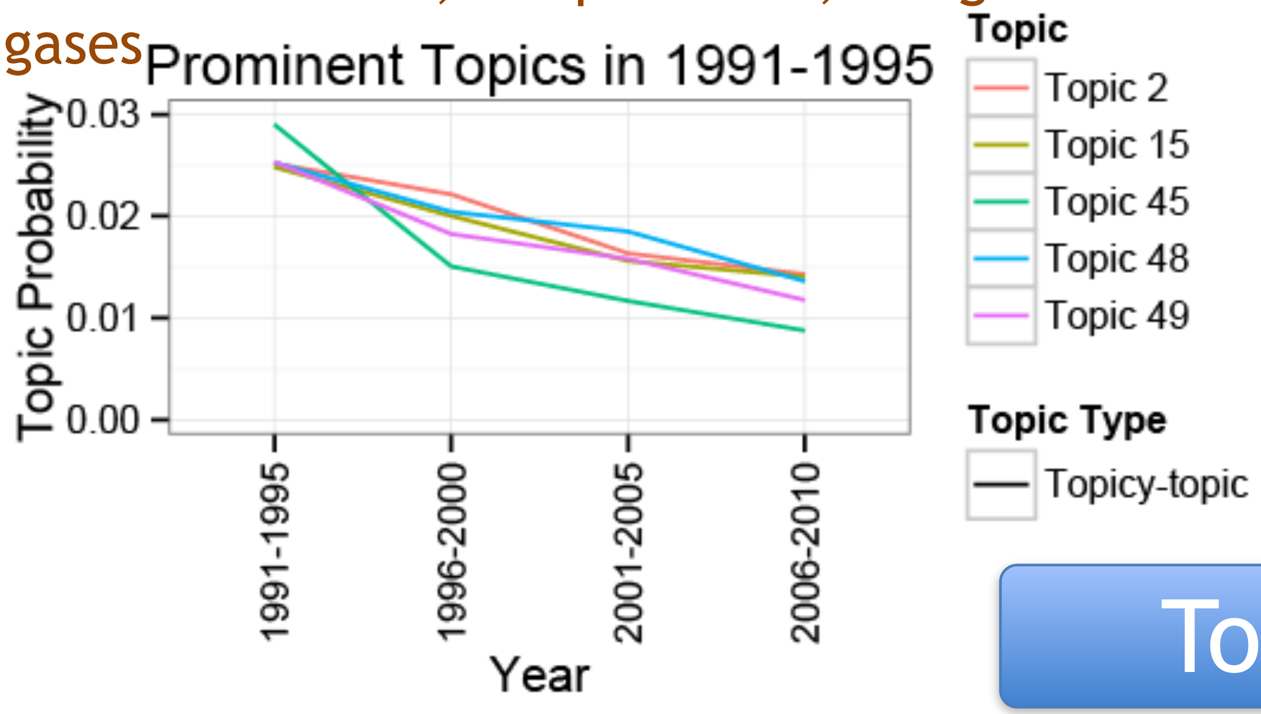
- Topic 15
  - environment, global, problem, environ, econom, concern, issu, chang, secur, polit, human, world, such, degrad, intern, conflict, activ, address, solut, ecolog

    → global environmental security and other problems
- Topic 45
  - pollut, control, air, ozon, environment, wast, effect, deplet, which, problem, industri, use, most, or, sourc, this, chemic, cfcs, qualiti, layer

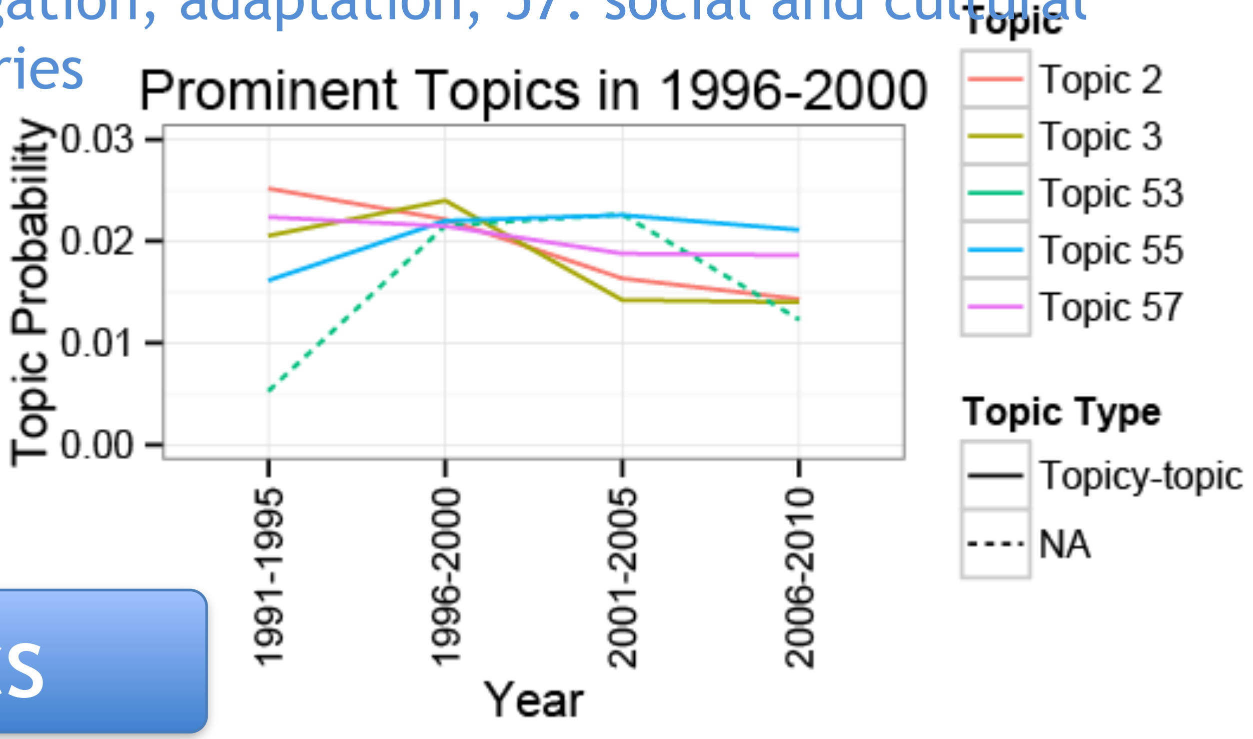
    →toxic substances and pollution management

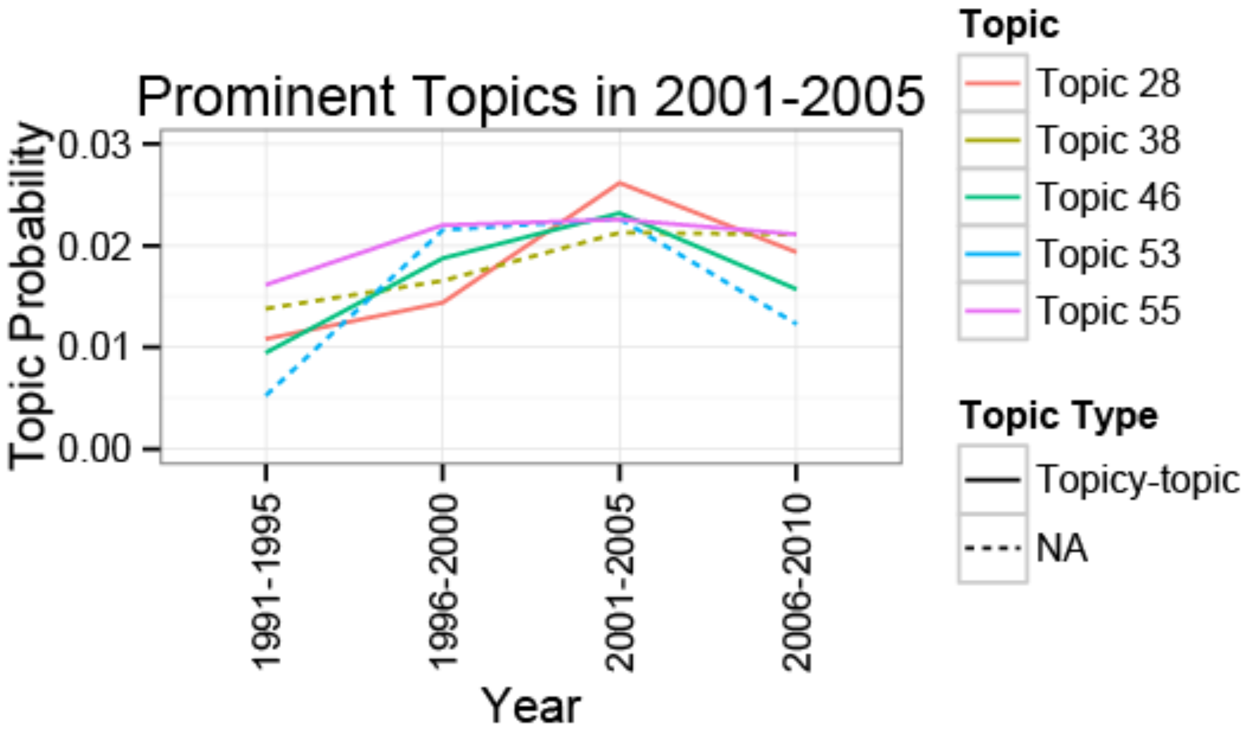
Topics

2. planning, agenda / 15. GE security etc, 45. toxic substances, 48. protocols, 49. greenhouse gases

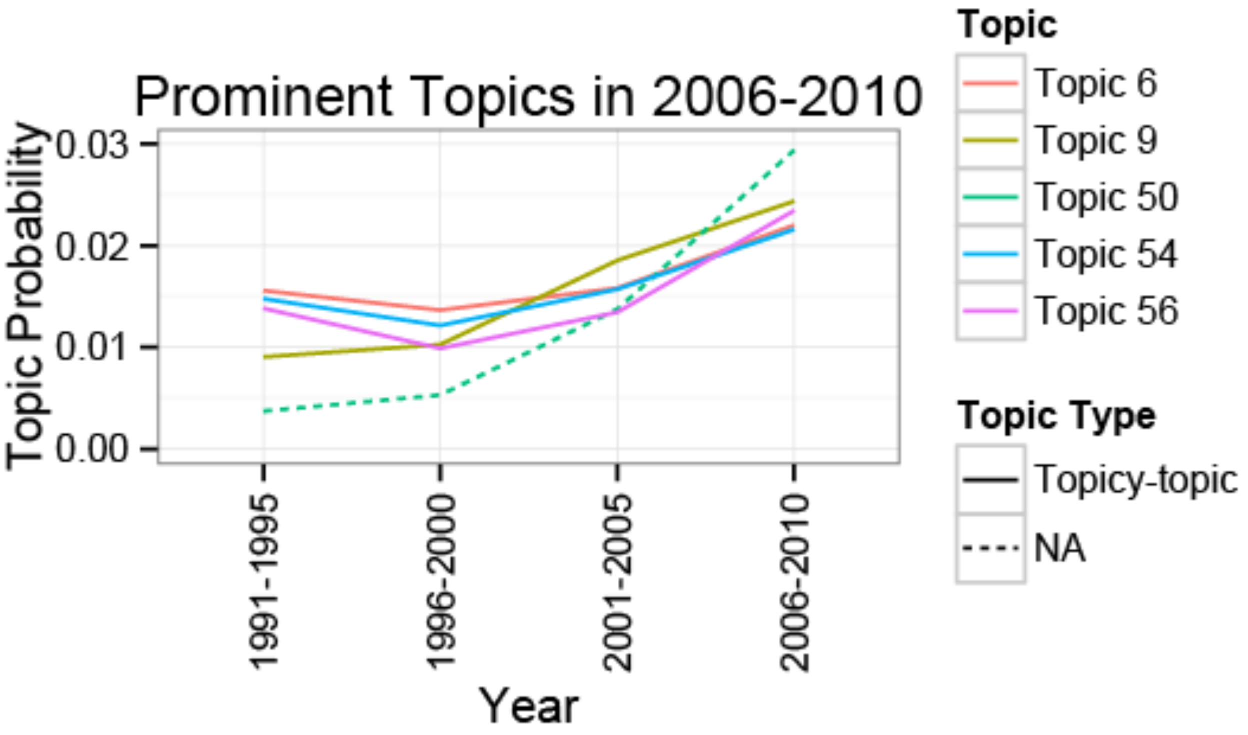

Prominent Topics in 1991-1995

2. planning, 3. emissions regulations, 55. mitigation, adaptation, 57. social and cultural theories


Prominent Topics in 1996-2000


Prominent Topics in 2001-2005

28. Assessment processes, participatory, 38. meta-analyses & case studies, 46. comparing scenarios, 55. mitigation, adaptation


Prominent Topics in 2006-2010

6. Network actor analysis, 9. vulnerability, 54. ecological systems and resilience, 56. households, village level

Trends in GEC

Increasing trend

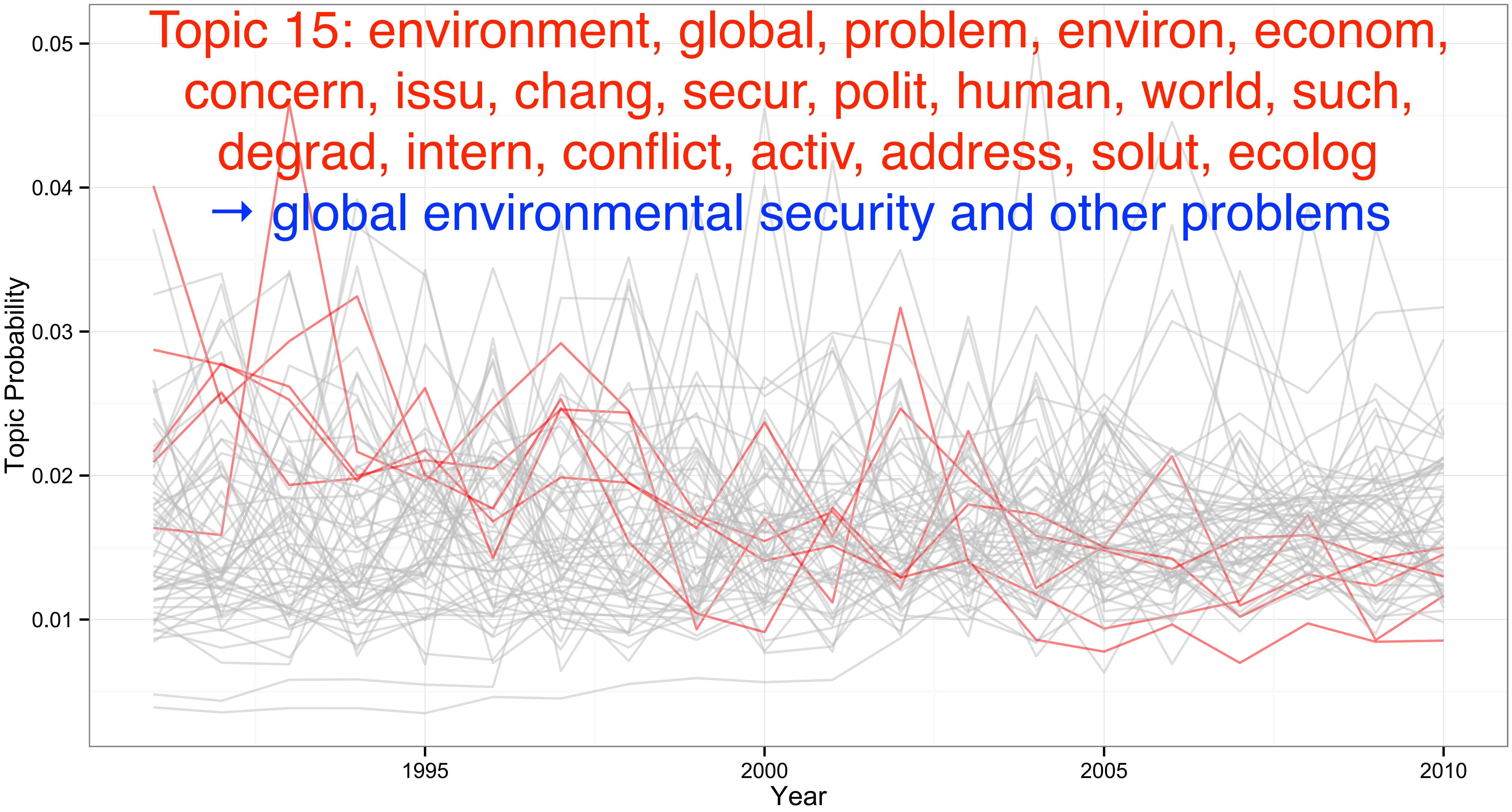
Topic	Label
9	vulnerability, adaptive capacity
12	learning & management
18	local knowledge, traditions, culture
24	media and public discourse, and reviews of scientific literature
38	metatext, meta-analyses and case-studies
50	2000 refs

Decreasing trend

Topic	Label
5	energy use, efficiency
15	global environmental security and other problems
30	Hypothetical discussion
35	Developing and developed countries
45	toxic substances and pollution management

GEC is moving away from discussion of energy, global environment, developed vs developing countries, and pollution, and moving towards the issues of vulnerability, management, culture preservation, media and public discourse, and empirical studies.

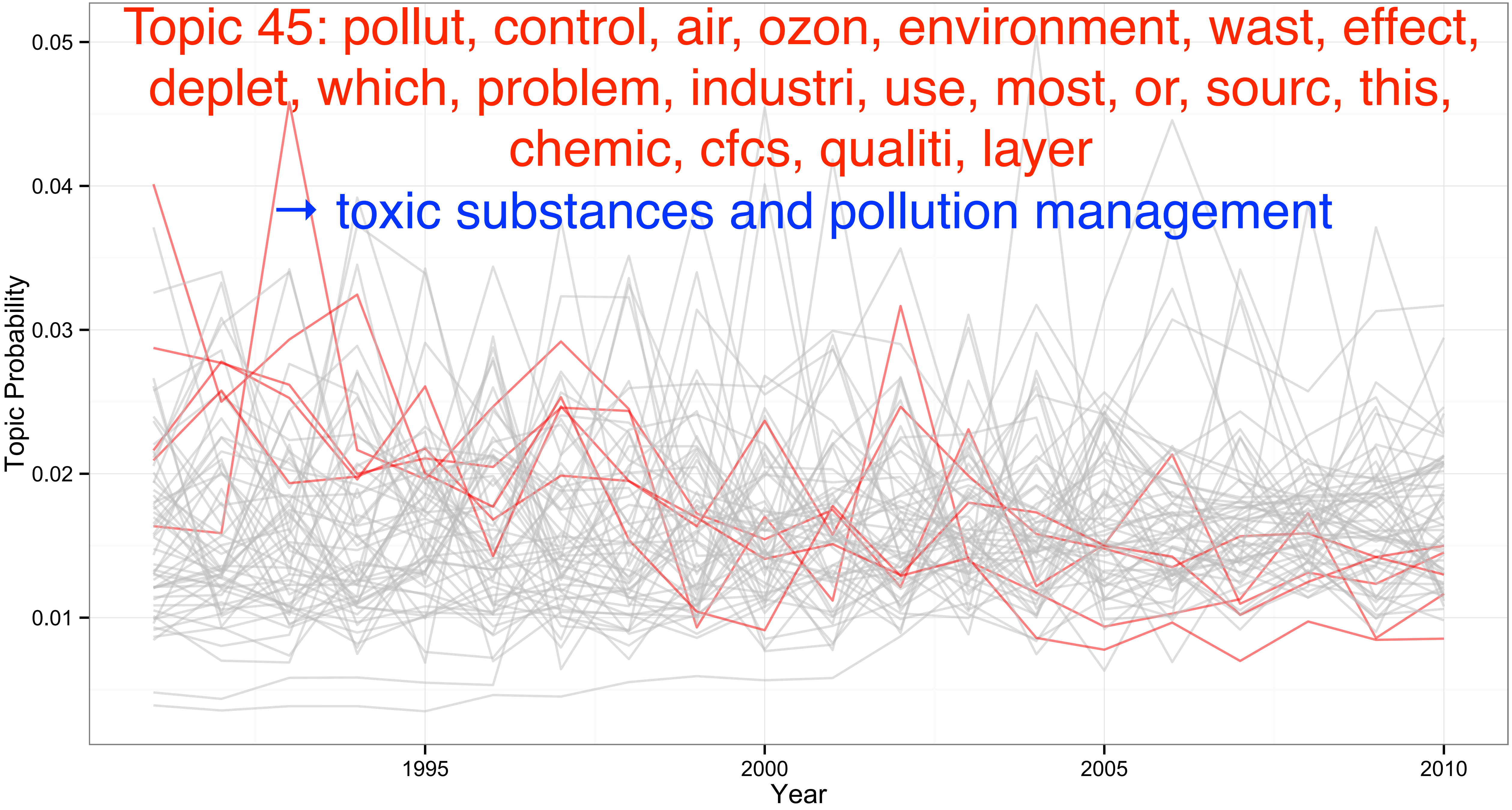
Chronological Topic Transition



Topic 15: environment, global, problem, environ, econom, concern, issu, chang, secur, polit, human, world, such, degrad, intern, conflict, activ, address, solut, ecolog

→ global environmental security and other problems

Chronological Topic Transition



Topic 45: pollut, control, air, ozon, environment, wast, effect, deplet, which, problem, industri, use, most, or, sourc, this, chemic, cfcs, qualiti, layer

→ toxic substances and pollution management

“Topic” in Topic Modeling

- The “topic” in topic modelling does not necessarily correspond to the topic in its usual sense of the word.

- We divided the topics into two types:

  1. thematic topics

  2. rhetorical topics

“Topic” in Topic Modeling

- The “topic” in topic modelling does not necessarily correspond to the topic in its usual sense of the word.

- We divided the topics into two types:

  1. thematic topics

  2. rhetorical topics

Rhetorical Topics

- Topic 8: ‘We’ as researchers & our intention, evaluation and procedures
  - Keywords: we, our, this, these, can, which, not, import, both, first, term, use, time, how, point, then, differ, where, see, us
- Topic 30: Hypothetical discussion
  - Keywords: would, could, not, if, might, or, this, but, ani, should, such, some, one, possibl, more, suggest, potenti, even, then, other

Conclusion

- Topic models are useful in exploring large-scale specialized corpora in a bottom-up way.
- This leads to insights into
  - how they change over time
  - how they change within papers, and
  - how each text is characterised in terms of topics.

Conclusion

- In this talk, we have introduced only the most basic type of topic models.
- Topic models have been extensively researched in machine learning and computational linguistics, and a number of extensions have been proposed;
    - topic models using n-grams (e.g., El-Kishky, Song, Wang, Voss, & Han, 2014)
    - correlated topic models that allow correlation between topics (Blei & Lafferty, 2007)
    - dynamic topic models that account for the chronological change of keywords within topics (Blei & Lafferty, 2006)
    - automated ways to identify the optimal number of topics (Ponweiser, 2012)
    - automated ways to compute coherence of each topic (Lau, Newman, & Baldwin, 2014)

Further Illustration

Murakami, A., Hunston, S., Thompson, P., & Vajn, D. (forthcoming). ‘What is this corpus about?’ Using topic modeling to explore a specialized corpus. *Corpora*.

To Follow the IDRD Project

- Visit
  - www.idrd-bham.info
- Twitter
  - @IDRD_bham

References

Blei, D. M. (2012). Probabilistic topic models: Surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM, 55*(4), 77–84. doi:10.1145/2133806.2133826

Blei, D., & Lafferty, J. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.

Blei, D., & Lafferty, J. (2007). A correlated topic model of *Science*. *Annals of Applied Statistics, 1*(1), 17–35. http://doi.org/10.1214/07-AOAS114

El-Kishky, A., Song, Y., Wang, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment, 8*(3), 305–316.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(supplementary 1), 5228–5235. doi:10.1073/pnas.0307752101

Grün, B., & Hornik, K. (2011). topicmodels : An R Package for fitting topic models. *Journal of Statistical Software, 40*(13). Retrieved from http://www.jstatsoft.org/v40/i13

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.

Ponweiser, M. (2012). *Latent Dirichlet allocation in R*. Vienna University of Business and Economics.

Wagner, C. (2010). *Topic models*. Retrieved from http://www.slideshare.net/clauwa/topic-models-5274169