Mixing Social Network Analysis with Structural Topic Modeling:

The case of Internet regulation coverage in the Russian media

Olga Silyutina
oyasilyutina@gmail.com
Anna Shirokanova
a.shirokanova@hse.ru
National Research University Higher School of Economics

Saint-Petersburg, 2018

Outline

1. Mixing methods in text analysis
2. STM:
   - Searching for the K
   - Correlation between topics
   - Effect estimation

2. Covariates:
   - Entity extraction
   - Case of countries and years

3. Networks
   - Social networks
   - Clusterization

4. Empirical case: Internet regulation coverage in Russia

Mixing Methods in Text Analyses

Traditional way to go:

Quantitative content analysis + interpretation of meanings (qual)
OR: development of categories (qual) + content analysis (quant)

Problem:

Human coding (workload / time / reliability)

One solution:

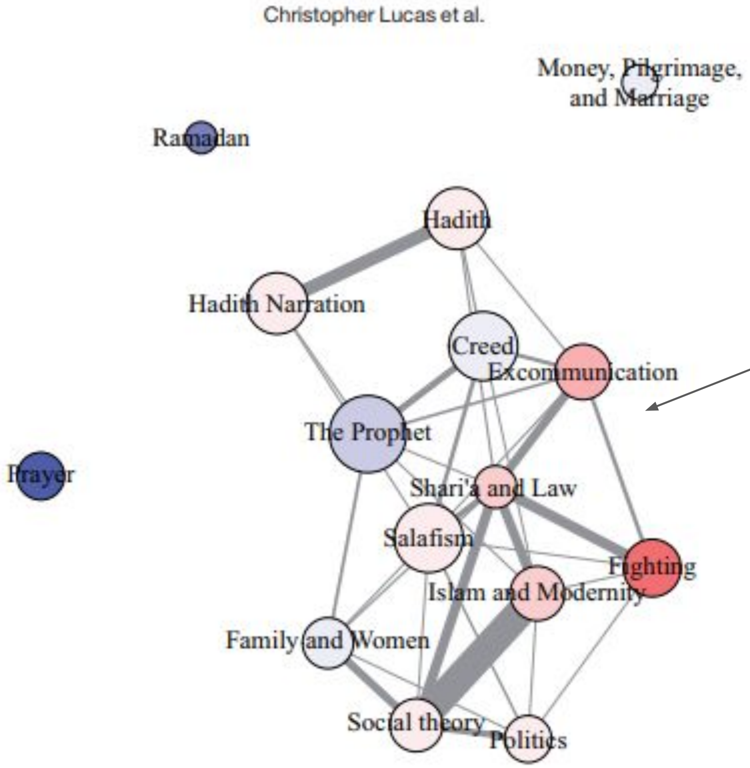
Semi-automated structural topic modeling, STM (Roberts et al., 2013)

Topic is a vocabulary representing semantically interpretable 'themes'

- STM infers topics from the documents taking into account document properties, e.g. author's gender, date of publishing, etc.
- STM discovers topics from texts rather than assuming them in advance (no pushing of categories)

Applications: mapping multilingual reactions to political event across countries, processing open-ended questions, digital news archives, etc.

Example of STM application



Texts: 27,248
Arab Muslim cleric writings with (non-)Jihadi as covatiate (report-based)

Topics' correlation network:
edge width ~ correlation strength;
node size ~ # of words in corpus/topic;
blue-red palette ~ effect of covariate (direction, strength)

(Lucas et al., 2015)

Topic models are a framework of statistical-based algorithms used to identify and measure latent topics within a corpus of text document (Wesslen)

Structural topic modeling (Roberts et al. 2013)

LDA
- Document = mixture of topics

- Document = mixture of topics
- User-specified covariates -> topical prevalence
- Topics are correlated
- Each document has its own prior distribution over topics
- Words in topics depends on covariates

Goal:
to allow researchers to discover topics and estimate their relationship to document metadata

Advantages for social science:
- Analysis of a large number of unstructured texts (Wesslen)
- Provision of hard evidence even for politicized topics' coverage (Shirokanova, Silyutina)

Limitations:
Usefulness depends on the correspondence between topics and the constructs of theoretical interest (Jacobi et al.)

Where Methodological Logics Clash

- Topics are extracted automatically, but how many topics to choose? (defined by researcher)

- Naming the topics (based on frequency-exclusivity metrics)

- Interpreting the correlations of topics and their ‘communities’

Data

Data source: Integrum (private digital archive 30 years deep, covers 64,000 media outlets)

Time span: 2009 to 2017

Sample: 7,240 texts, final sample after clearing: 6,140 texts

Covariates for different models:

- 48 countries co-occurred in pairs 100+ times
- political or non-political source

Searching for the Right K*

*number of topics

There is no "right" answer to the number of topics which is appropriate for a given corpus (Grimmer, Stewart)

There is a strong positive relationship between the number of topics and the probability of topics being nonsensical (Mimno) more -> worse

stm::searchK()

var$results:
held out likelihood
residual analysis
exclusivity
semantic coherence

K	exclusivity	semantic coherence	heldout	residual	bound	lbound	num its
30	9.62	-57.5	-9.1	6.47	-22487226	-22487151	86
50	9.74	-60.3	-9.11	5.16	-21954014	-21953866	86
70	9.8	-65.3	-9.08	4.88	-21607059	-21606828	67
90	9.83	-65.4	-9.14	5.89	-21348428	-21348112	67

Entity extraction

Automatically revealing important information from the text (dates, places, organizations)

Our case:

countries

years

Solution:

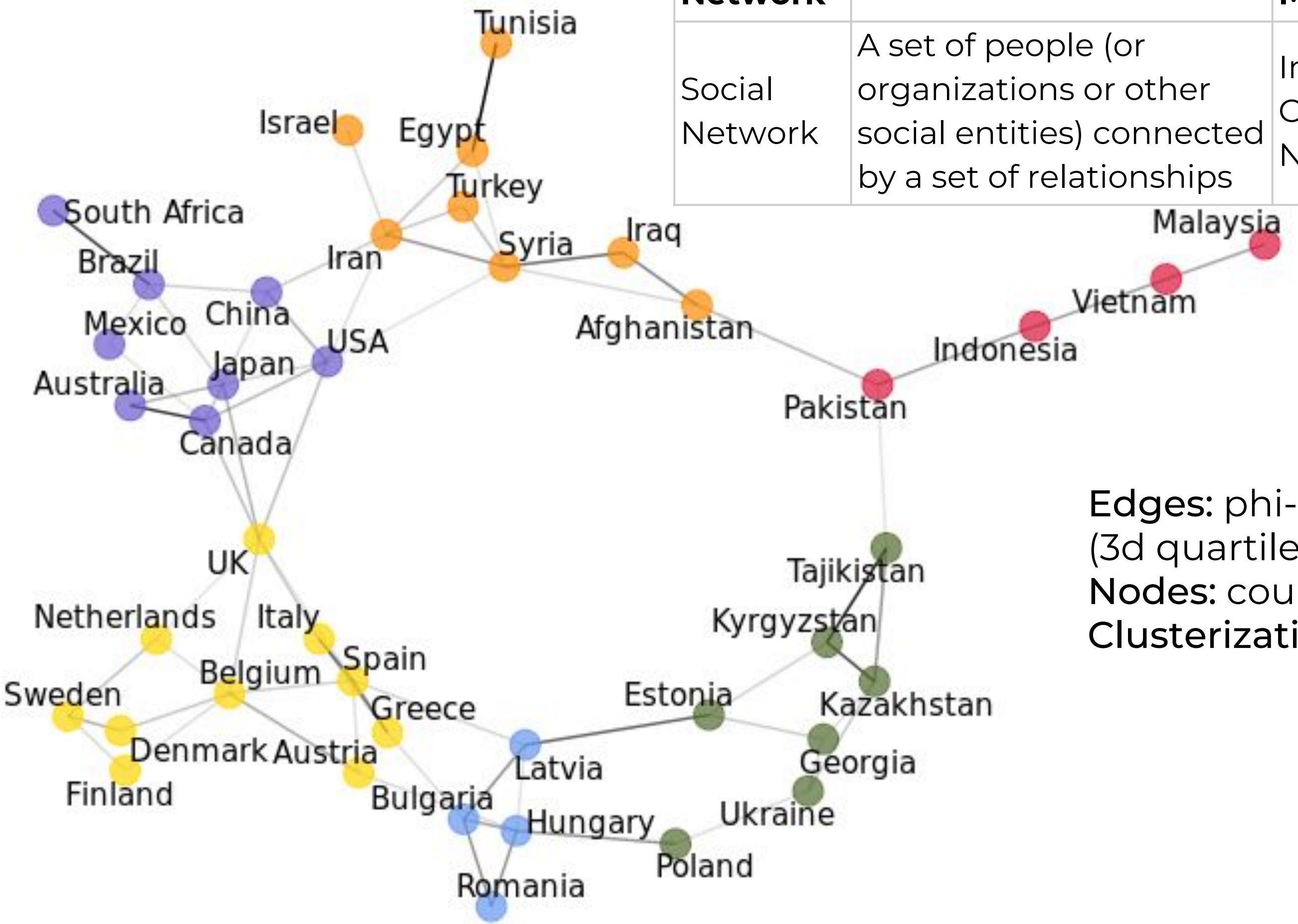
dictionary

Assign the most frequent country in document as meta data -> example of covariates

Correlation network of countries

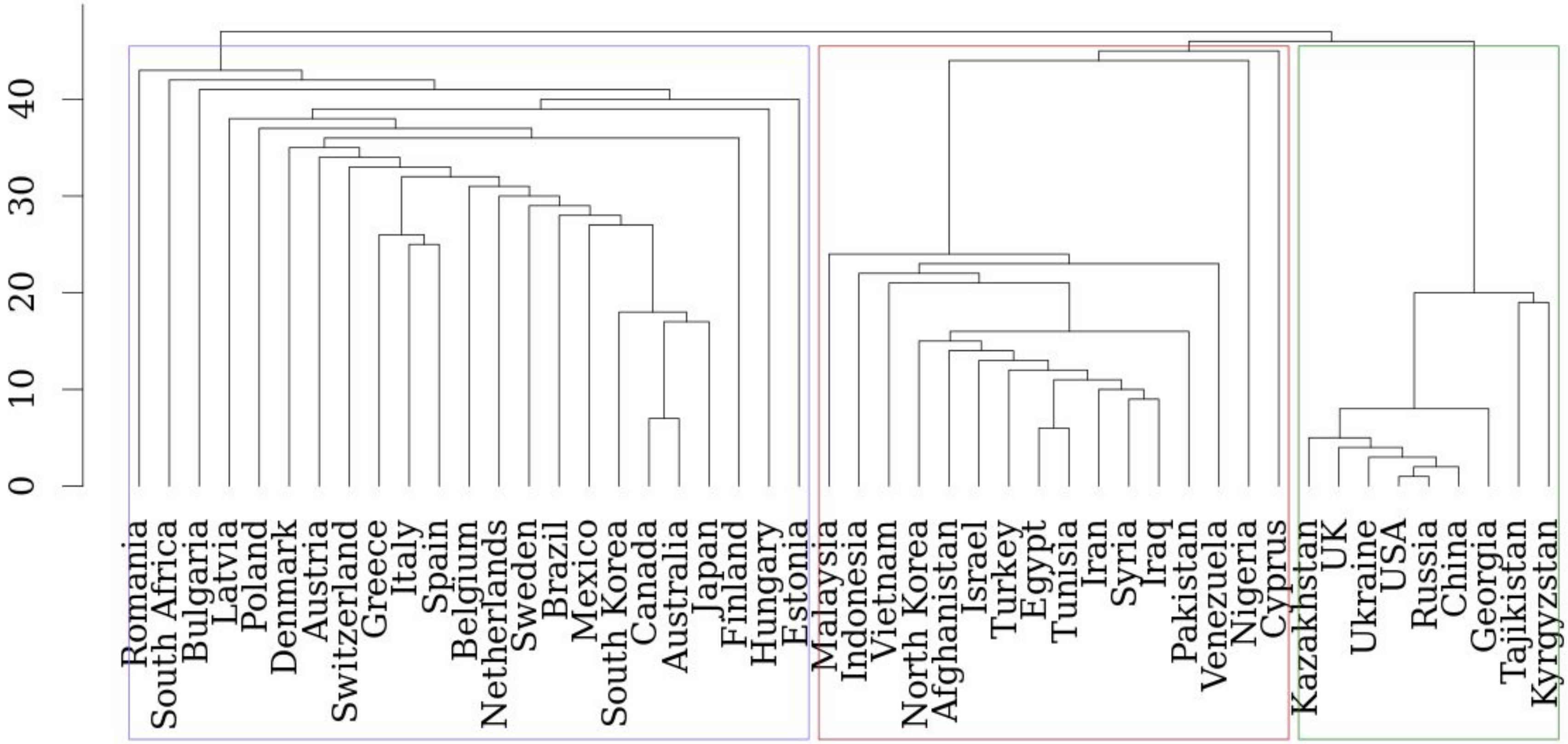
Type of Network	Conceptual Definition	Operational Measure	Content of Relation/Link
Social Network	A set of people (or organizations or other social entities) connected by a set of relationships	Individual, Group, Organization, Nation-State	Any Kind of Social Relation

(Park, 2003)



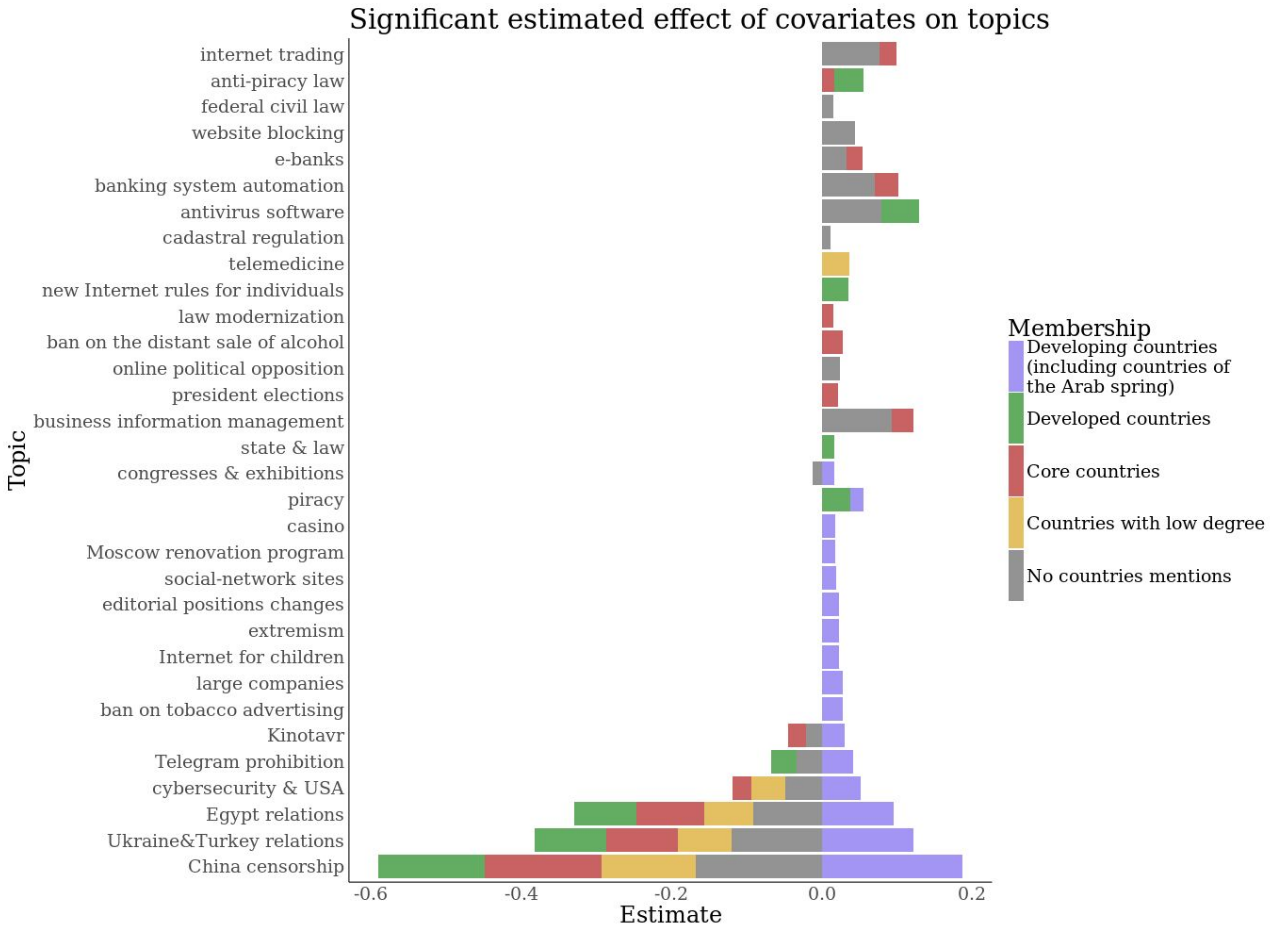
Edges: phi-coefficients higher than +0.2 (3d quartile)
Nodes: countries >100 docs)
Clusterization: fast-greedy

Co-occurrence network



co-occurrence of countries in texts + fast-greedy

Estimation of covariate effect



`stm:estimateEffect()`

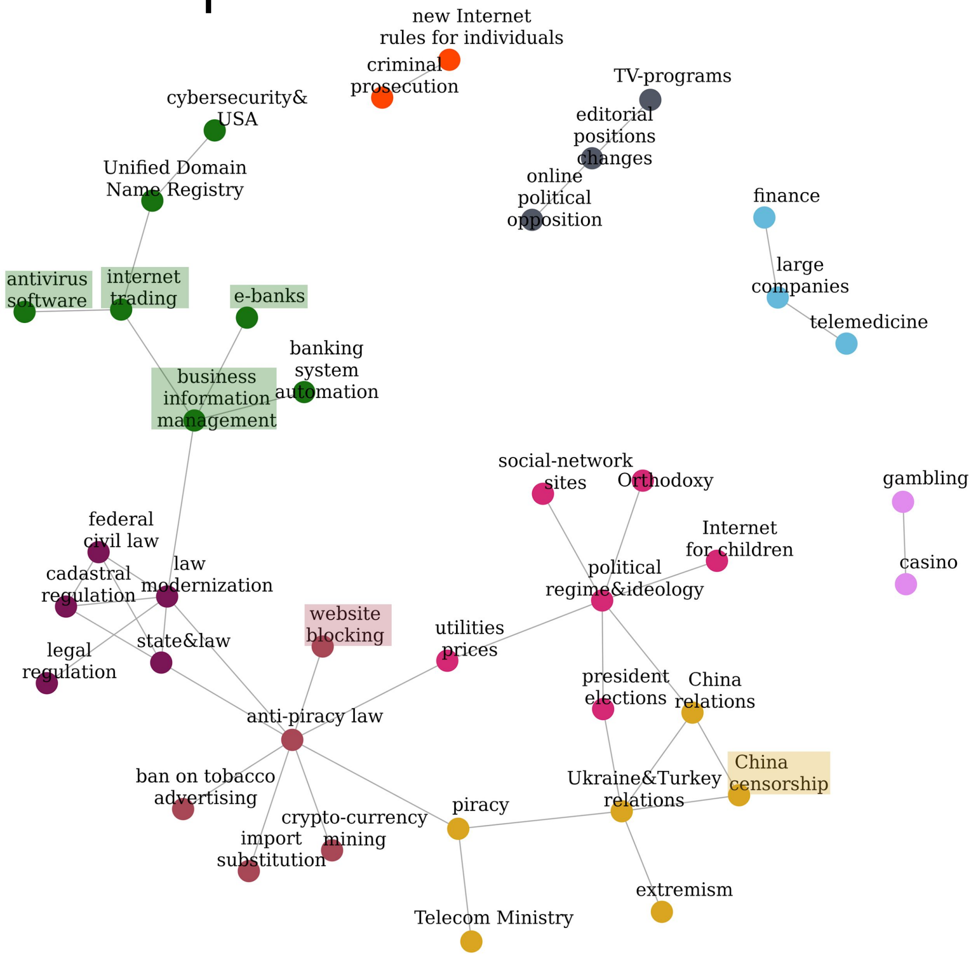
Units: documents

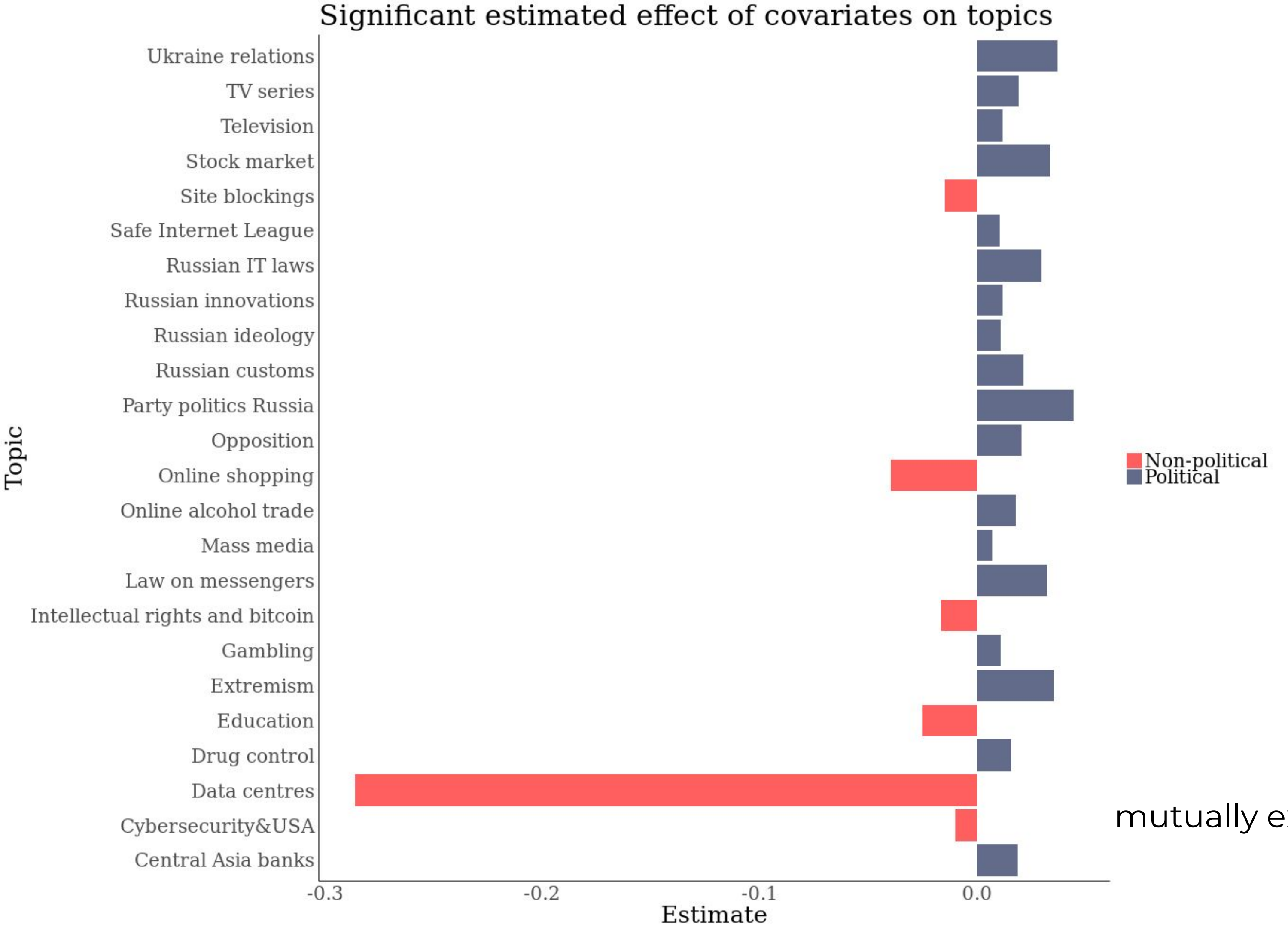
Outcome: the proportion of each document about a topic in an STM model

Covariates: document meta data

Correlation network of topics

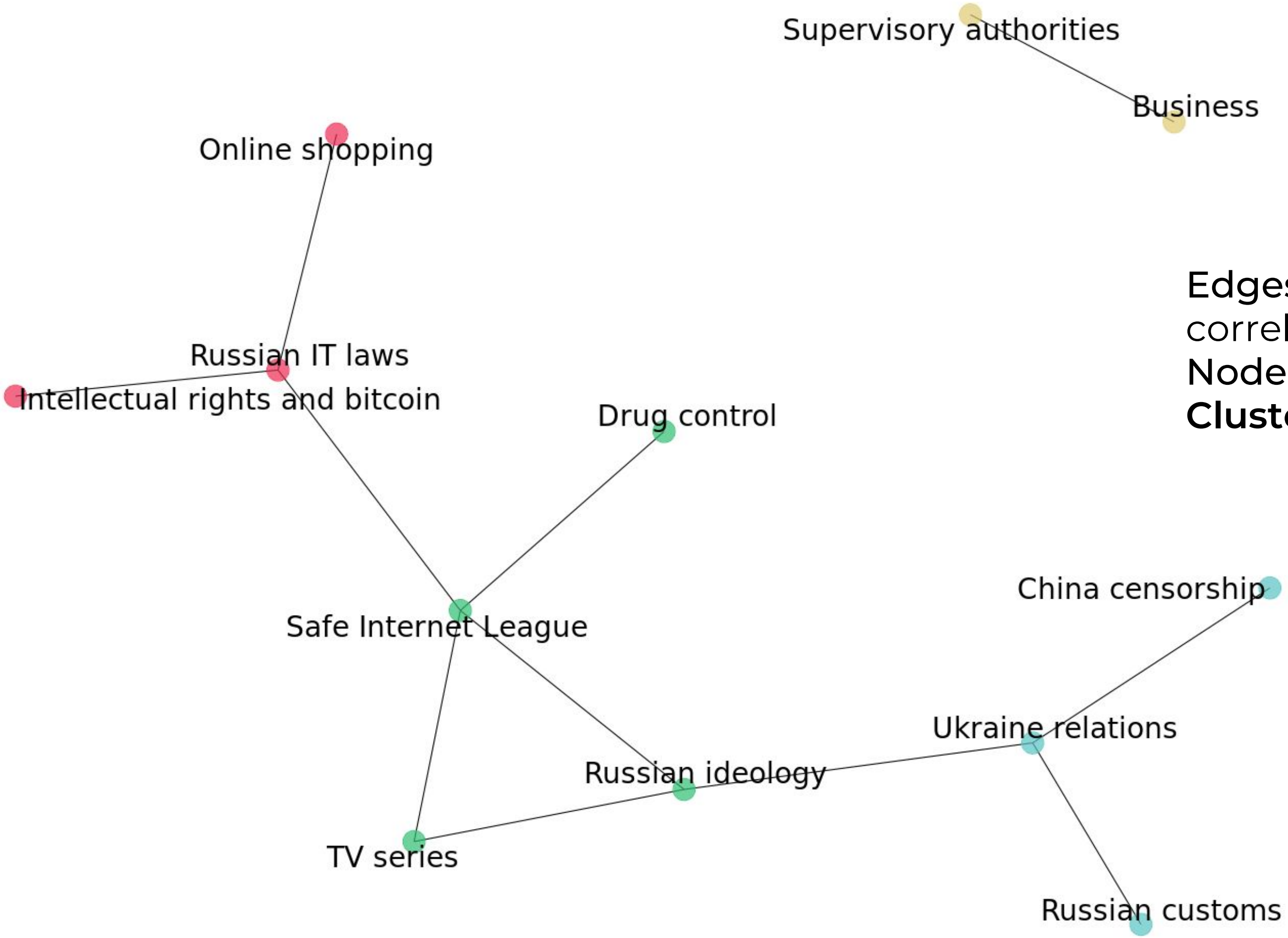
Edges: positive, significant correlations
Nodes: topics
Clusterization: fast-greedy



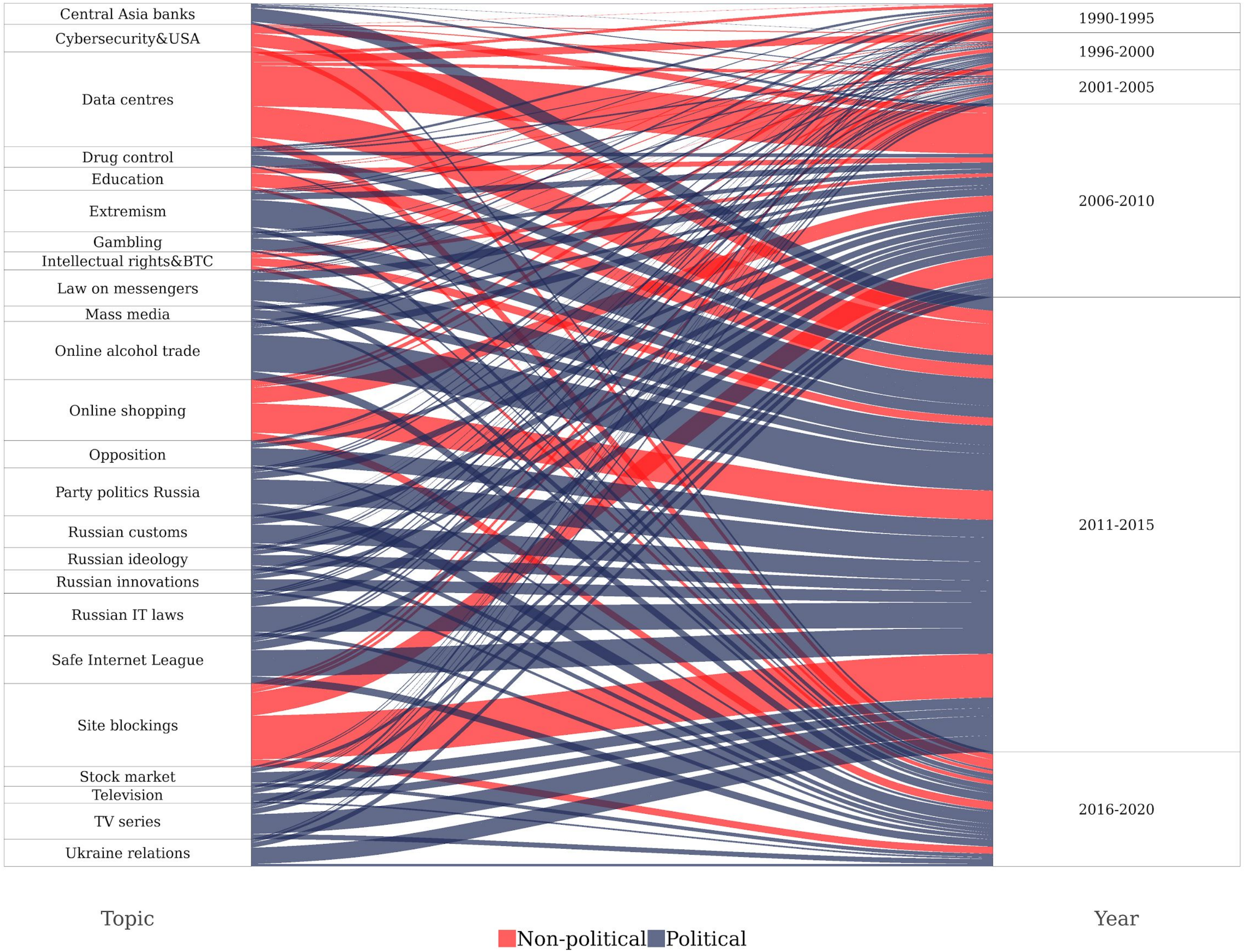


mutually exclusive

Correlation network of topics



Edges: positive, significant correlations
Nodes: topics
Clusterization: fast-greedy


Central Asia banks
Cybersecurity&USA
Data centres
Drug control
Education
Extremism
Gambling
Intellectual rights&BTC
Law on messengers
Mass media
Online alcohol trade
Online shopping
Opposition
Party politics Russia
Russian customs
Russian ideology
Russian innovations
Russian IT laws
Safe Internet League
Site blockings
Stock market
Television
TV series
Ukraine relations
1990-1995
1996-2000
2001-2005
2006-2010
2011-2015
2016-2020
Topic
Year
Non-political
Political

Conclusions

Topic segmentation is part of natural language processing tasks. It can help substantive goals in social/behavioral sciences

Inferring (correlated) topics from texts, STM largely improves text coding experience, reliability and reproducibility of results

“Mixing” resides in the iterations of finding model solutions / interpreting the topics and their correlations in ‘communities’

Pros: ready-made software, time-saving, coder-independent
Requirements: understanding of data for choice of covariates, K-number, and interpretation

References

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064-1082.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2), 254-277.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1), 89-106.

Wesslen, R. (2018). Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond. *arXiv preprint arXiv:1803.11045.*

Thank you for your attention!