Welcome

11th CODH Seminar

人文学オープンデータ共同利用センター
Center for Open Data in the Humanities

Text Mining for Analyzing Research Communities: *Sociological Topics and Socio-Technical Imaginaries*

http://codh.rois.ac.jp
https://www.nii.ac.jp

Program

Time	Presentation	Presenter
14:00-14:30	Making Inferences with Topic Modeling: The Effects of Sociological Topics on Citation Impact	Radim Hladik (JSPS Postdoctoral Fellow at National Institute of Informatics)
14:30-15:00	The Trends of Research Methods in Japanese Sociology, 1952-2018	Hiroshi Tarohmaru (Kyoto University)
15:00-15:30	Understanding of Socio-technical Imaginary and ELSI: an Application of Social Science Perspectives and Quantitative Text Analysis."	Ryuma Shineha (Seijo University)
15:30-16:00	Open Floor Discussion and Closing	All

Opportunities for collaboration

- Radim Hladík
  - 2-year postdoctoral fellowship sponsored by JSPS
  - hosted by Asanobu Kitamoto-sensei's lab at the National Institute of Informatics
  - Centre for Science, Technology and Society Studies at the Institute of Philosophy of the CAS
  - text analysis, scholarly communication, research funding, academic disciplines
- JSPS sponsors the Japanese side in other projects of bilateral collaboration between the Czech Academy of Sciences and Japanese HE and research institutions (MPP)

Mobility Plus Projects (MPP)

The aim of the joint 2- or 3-year mobility plus projects is to establish or intensify scientific cooperation, which should result in preparation of larger joint international projects. Partner research teams submit a joint project proposal and aspire to reach joint outputs while being able to have long-term access to unique research infrastructures and equipment of the cooperating team and complimentary techniques and methods.

Czech Academy of Sciences

JSPS

https://www.avcr.cz/en/academic-public/international-affairs/bilateral-cooperation/partner-organizations/#Japan

Making Inferences with Topic Modeling: *The Effects of Sociological Topics on Citation Impact*

11th CODH Seminar 2019

Tokyo (JP), September 25, 2019

http://codh.rois.ac.jp/

Radim Hladík

National Institute of Informatics (Tokyo, JP)
Institute of Philosophy of the Czech Academy of Sciences (Prague, CZ)

Contact

radim.hladik@fulbrightmail.org
@hlageek

Acknowledgement

Grant-in-Aid for JSPS Fellows no. 17F17769.

Outline of the presentation

1. Motivation
2. Data
3. Topic modeling
4. Citation counts modeling

Motivation

Citation patterns are field-dependent

- varying size of research communities
- varying publication preferences
- varying citation cultures and practices

- ➔ need for normalized and/or robust indicators, e.g.
  - mean normalized citation score (MNCS)
  - source normalized impact per paper (SNIP)
  - integrated impact indicator (I3)

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. Journal of Informetrics, 4(3), 265–277. https://doi.org/10.1016/j.joi.2010.01.002
Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics, 5(1), 37–47. https://doi.org/10.1016/j.joi.2010.08.001
Waltman, L., & van Eck, N. J. (2018). Field normalization of scientometric indicators. ArXiv:1801.09985. Retrieved from http://arxiv.org/abs/1801.09985

What are the fields anyway?

- the problem of the **granularity of classification**
  - at what level should we aggregate by subject?
  - Ruiz-Castillo & Waltmann (2015) suggest thousands
  - Klavans & Boyack (2017) suggest tens of thousand
- problematic **identification** of research subjects
  - often-used journal-level classification unreliable Shu et al. (2019): 46% of articles do not match journal's discipline
- ➔ **algorithmically constructed subject classes** on paper level can be helpful

Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. Journal of Informetrics, 13(1), 202–225. https://doi.org/10.1016/j.joi.2018.12.005
Klavans, R., & Boyack, K. W. (2017). Research portfolio analysis and topic prominence. Journal of Informetrics, 11(4), 1158–1174. https://doi.org/10.1016/j.joi.2017.10.002
Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. Journal of Informetrics, 9(1), 102–117. https://doi.org/10.1016/j.joi.2014.11.010

Task

extract topics from full-text articles
score articles on their strength of association with particular topics
infer the effects of topics on citations

Challenge

allow for an article to belong to more than one topic
avoid treating topics in articles as compositional data

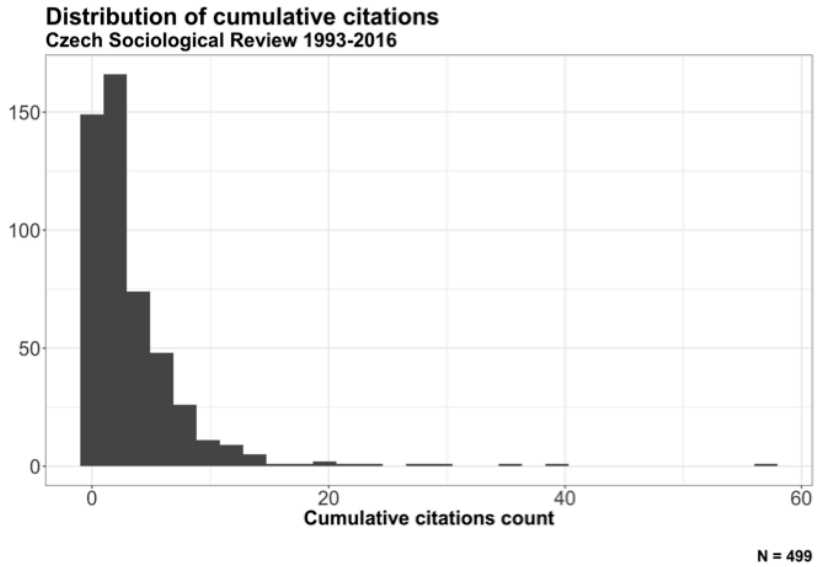
Data

Czech Sociological Review

- "core" journal of Czech sociology
- the only Czech sociological periodical indexed in *Web of Science*

- Corpus of selected research articles and essays published in *Czech Sociological Review* from 1993 to 2016
  - Originally Czech, non-translated material only
  - 522 documents and 3138072 tokens in total
  - Parts-of-speech tags
  - Published in Lindat-CLARIN repository
  - URI: http://hdl.handle.net/11372/LRT-2703
- 499 items matched with citation records in *Web of Science*
- number of authors
- sex of lead authors

Distribution of citations



- low IF of the journal (0.554 in 2018)
- small number of highly cited papers
- large number of poorly cited papers
- in such cases, Poisson distribution makes a better approximation than negative binomial

Pudovkin, A. I., & Bornmann, L. (2018). Approximation of citation distributions to the Poisson distribution. COLLNET Journal of Scientometrics and Information Management, 12(1), 49–53. https://doi.org/10.1080/09737766.2017.1332605

Topic modeling

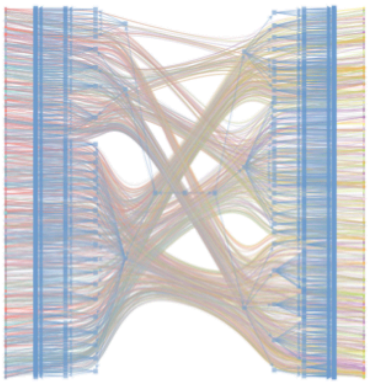
Why not LDA?

- LDA (Blei et al. 2003) is good for descriptive purposes and suitable for machine learning, but not for inference
- LDA topics are mixtures (compositional data)
  - perfect multicollinearity
- correlated topic models (CTMs) (Blei & Lafferty 2007)
  - perfect multicollinearity
  - allow for topics to correlate
- structural topic models (STMs) (Roberts et al. 2014)
  - perfect multicollinearity
  - correlate with metadata
    - good for description
    - cart before the horse in inference

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.
Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. The Annals of Applied Statistics, 1(1), 17–35. https://doi.org/10.1214/07-AOAS114
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. https://doi.org/10.1111/ajps.12103

TopSBM topic model

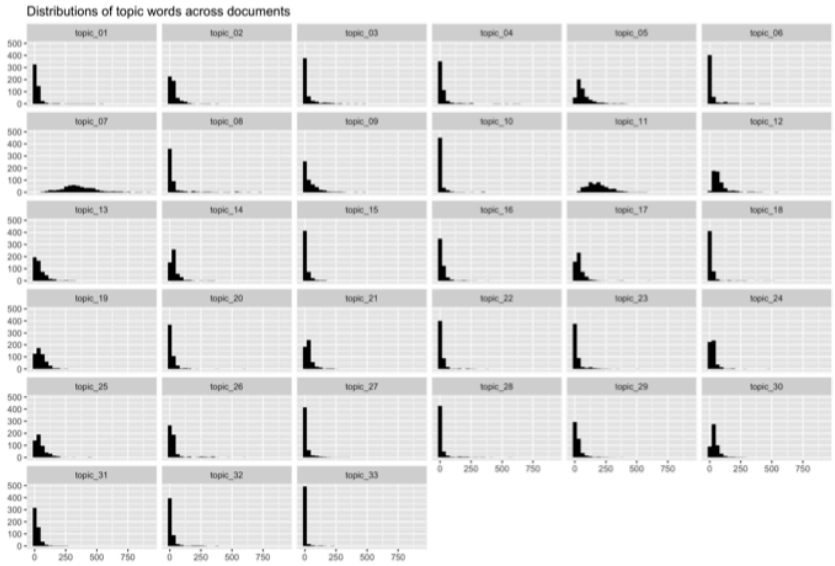
- Gerlach et al. 2018
- network approach to topic modeling
- stochastic block modeling (SBM) of communities in a bipartite network of documents and words (types)
- hierarchical (non-parametric), i.e. automatic selection of the number of topics
- words in topics do not have mixed membership
  - disadvantage compared to LDA and the like
  - allows to treat topics as dictionaries



Documents Word types

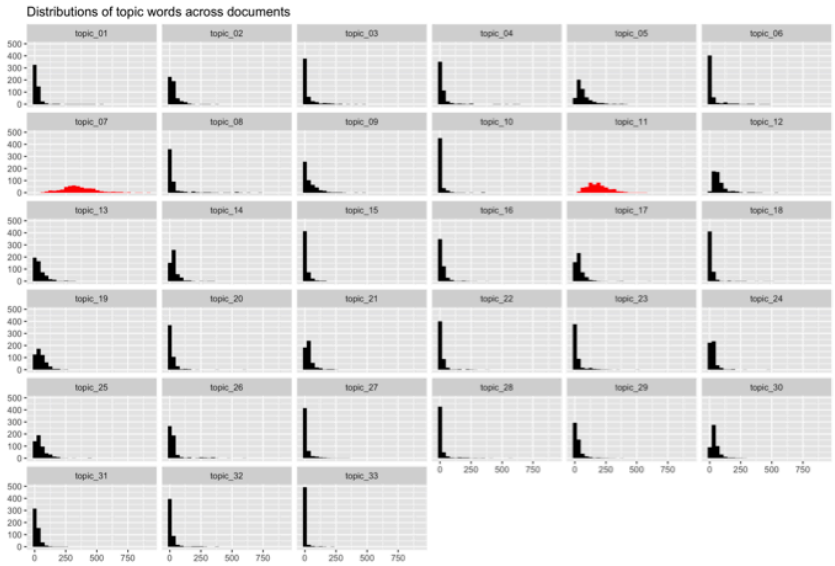
Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, 4(7), eaaq1360. https://doi.org/10.1126/sciadv.aaq1360

Distribution of topics in documents



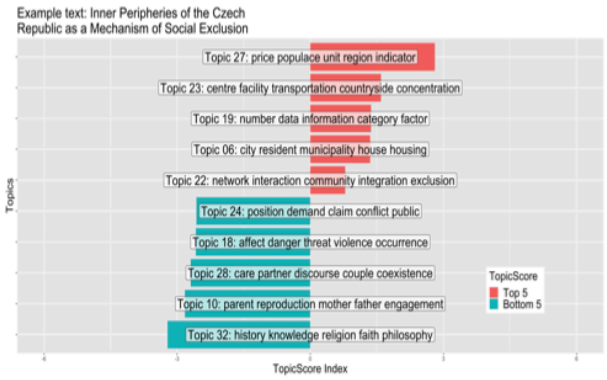
- in most documents, topic words are scarce
- only in a small number of documents are topic words highly frequent
- bins of 20

Distribution of topics in documents



- Gerlach et al. (2018): topics made of words whose distributions accross documents approaches normal distributions are semantically less informative, aka stopwords

Scoring of documents



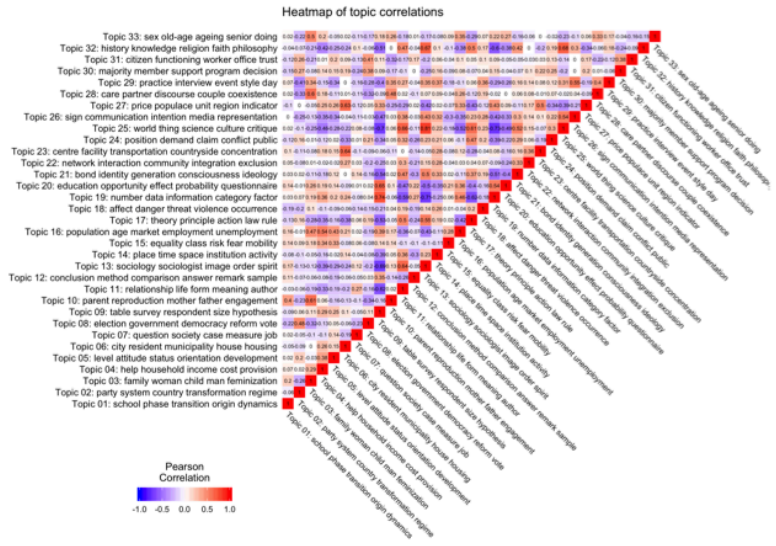
In the Czech Republic inner peripheries are usually the peripheral zones of metropolitan areas and regional centre areas. In the mid-1990s population numbers stopped declining in some peripheries as a result of suburbanisation processes, but in other peripheries depopulation processes continued. This last category of inner peripheries can be described as the hard core of Czech peripheral regions and in the authors' opinion they warrant the development of specific regional policy measures, stressing the creation of new jobs, the improvement of public transport, greater accessibility of service centres, and co-operation among communities.

To avoid treating topics as document mixtures (LDA-style), we measure the strength of association between topics and documents in the corpus with bootstrapped "keyness" statistic (log odds ratio).

- Word (token): w
- Document i (collection of words): $D_i = \{w_{1i}, w_{2i}, \ldots, w_{ni}\}$
- Corpus (collection of documents): $C = \{D_1, D_2, \ldots, D_n\}$
- Corpus complement to document i: $C'_i = \{D_j \in C : j \neq i\}$
- Topic words from topic k: $T_k = \{w_{1k}, w_{2k}, \ldots, w_{nk}\}$
- Non-topic words (word complement to topic k): $T'_k = \{w \in C : w \notin T_k\}$

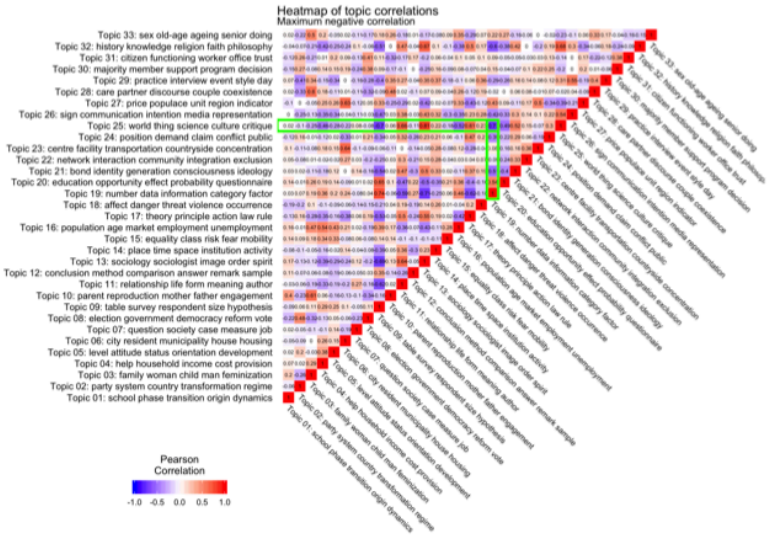
$$TopicScore_{TkDi} = log\left(\frac{\frac{|T_k \cap D_i|}{|T'_k \cap D_i|}}{\frac{|T_k \cap C'_i|}{|T'_k \cap C'_i|}}\right)$$

Correlation matrix of topics



- topic scores can correlate
- aid in interpretation

Correlation matrix of topics



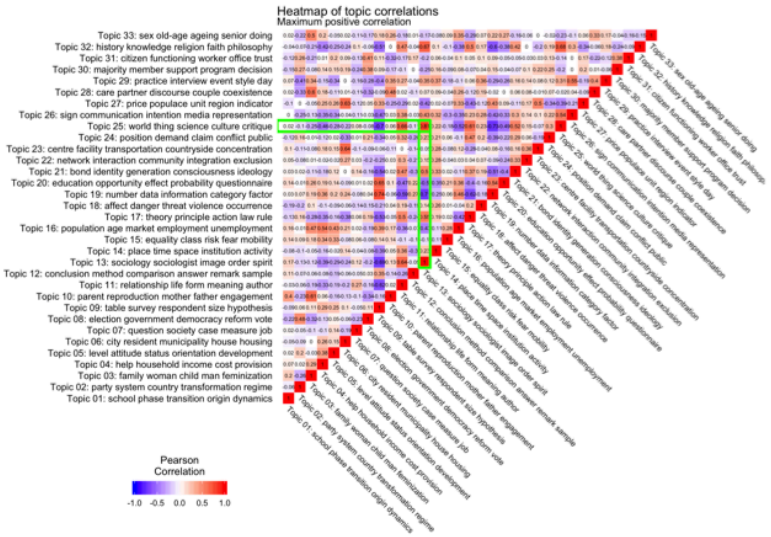
Maximum negative correlation **-0.73**

Topic 25: world thing science culture critique

Topic 19: number data information category factor

....

Correlation matrix of topics



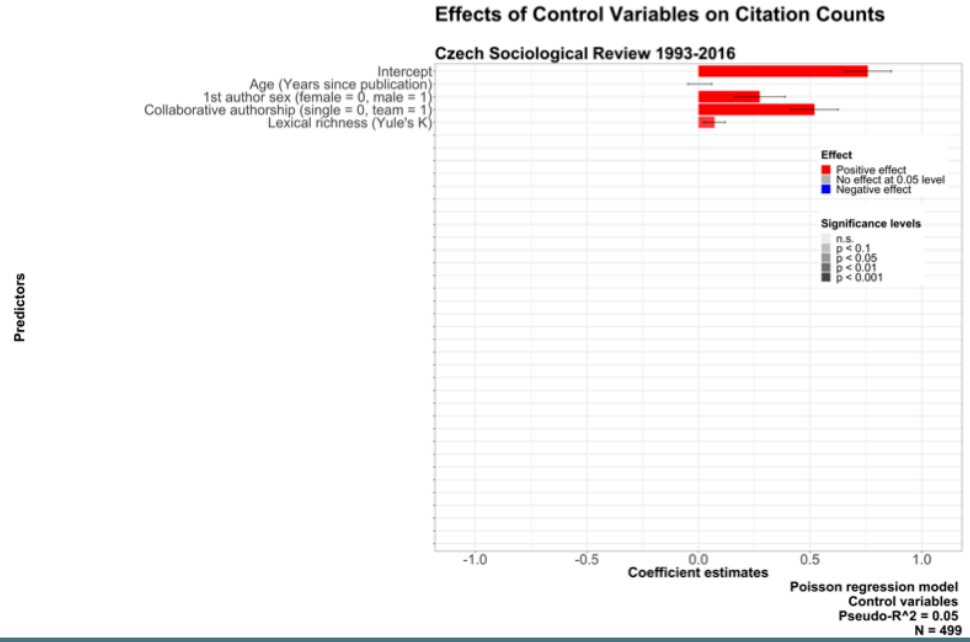
Maximum positive correlation **+0.81**

Topic 25: world thing science culture critique

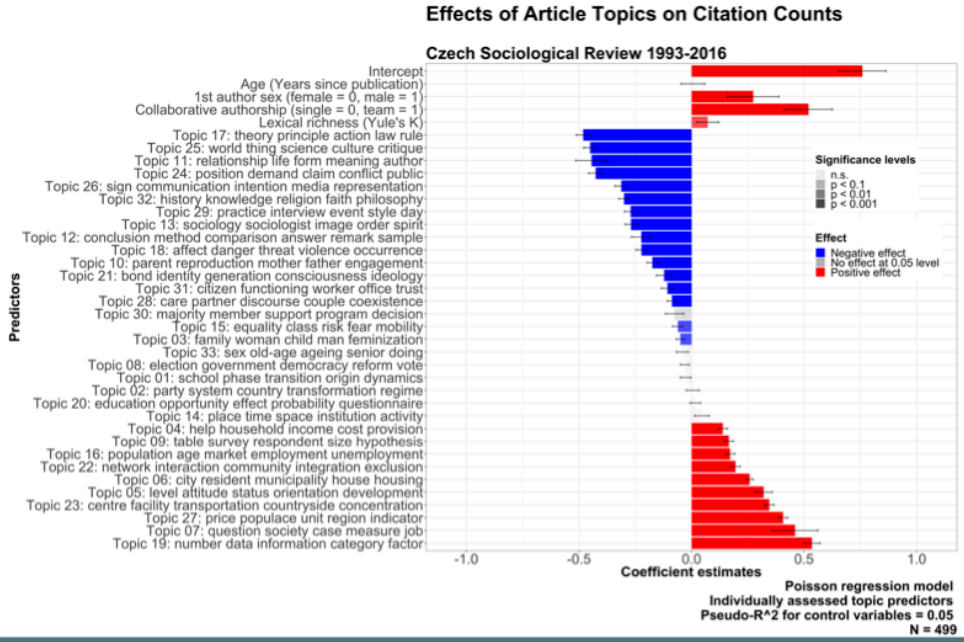
Topic 13: sociology sociologist image order spirit

Citation counts modeling

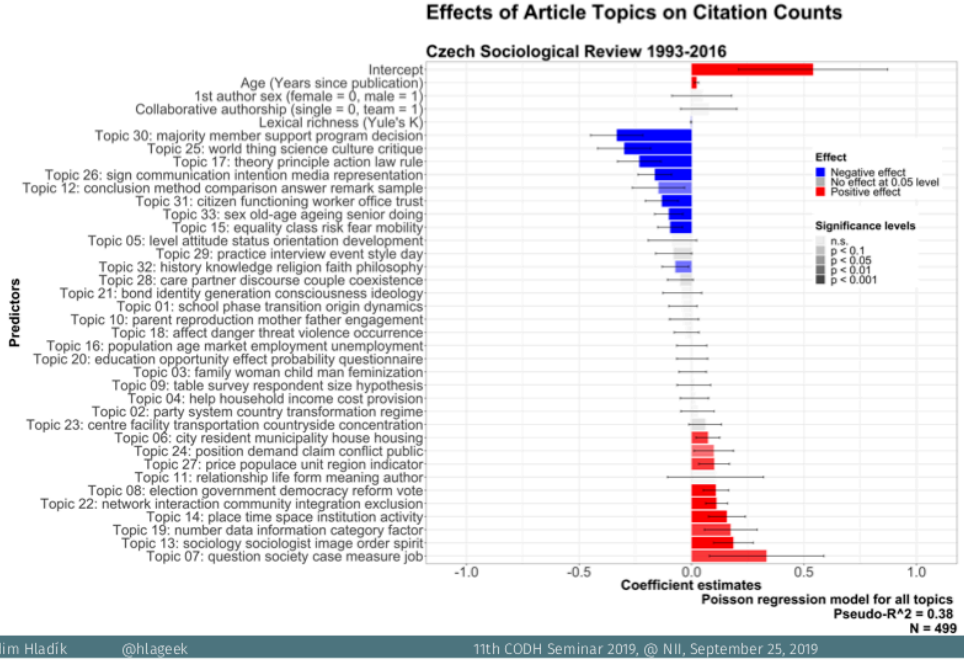
Controls



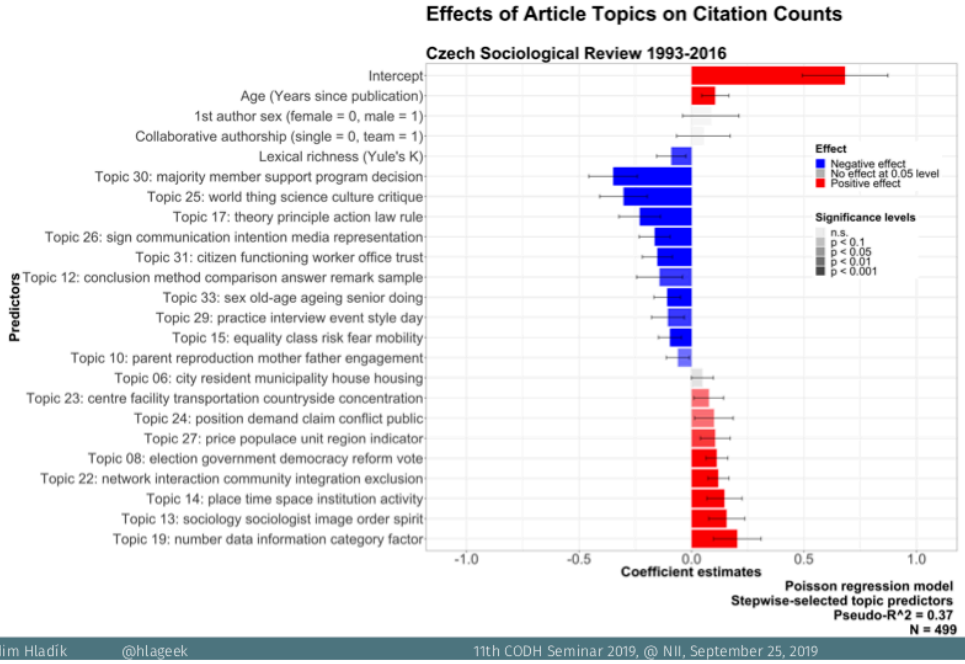
Topics effects modeled in segregation



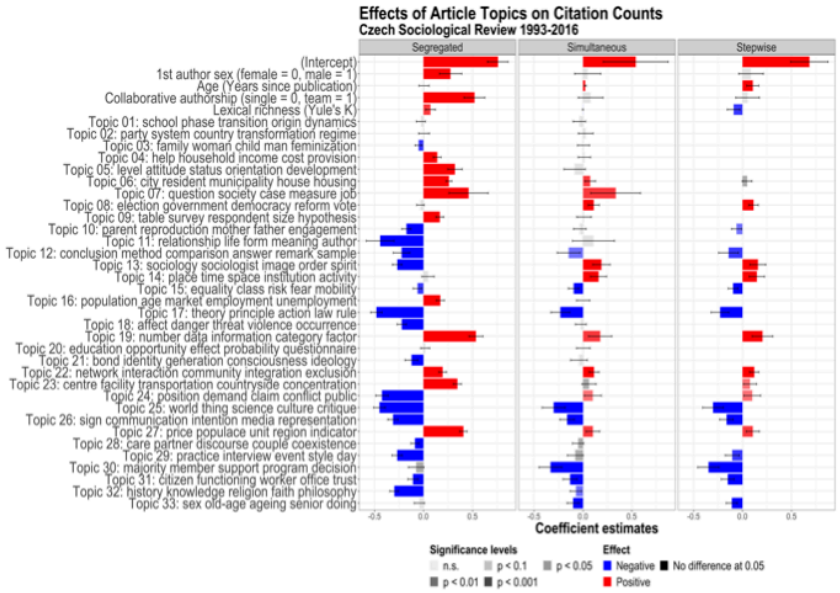
Topic effects modeled simultaneously



Topic effects modeled with stepwise selection



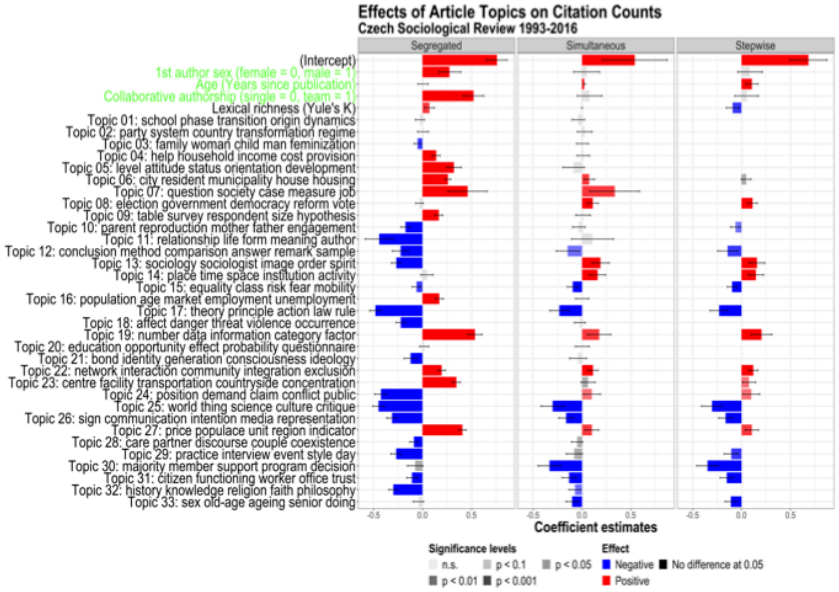
Comparison of effects



Interpretation

- Classification whereby an article belongs to one topic only may overestimate effects
  - Effects are stronger when topics are modeled separately
- Undercited topics cumulate disadvantages when compared to the advantages of topics with citation bonuses.
  - More topics carry citation penalty than premium
  - Penalties are stronger than premiums
  - However, the distinction is not prominent

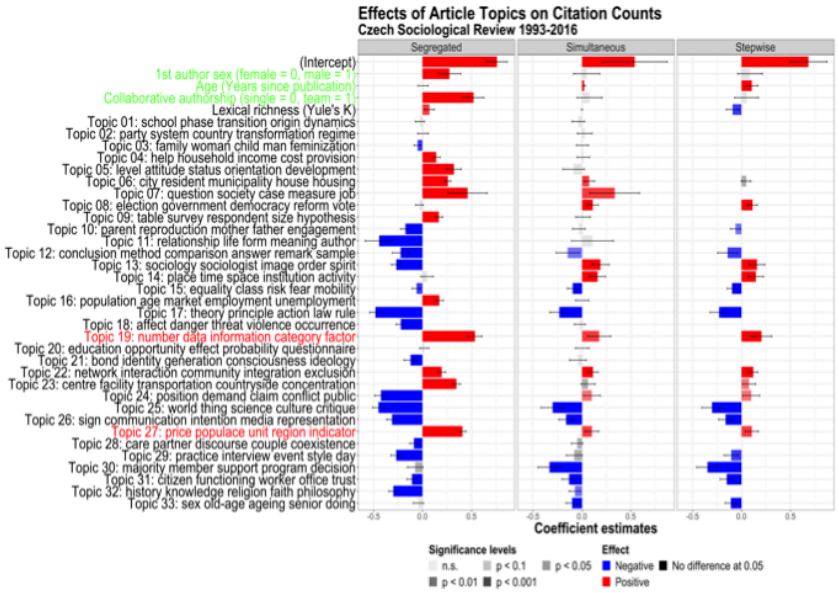
Comparison of effects



Controls

- Simpson's paradox for collaborative authorship and lead authors' gender
  - Significant effects disappear when the data is disaggregated
- The age of publication reappears as significant predictor when topics are accounted for

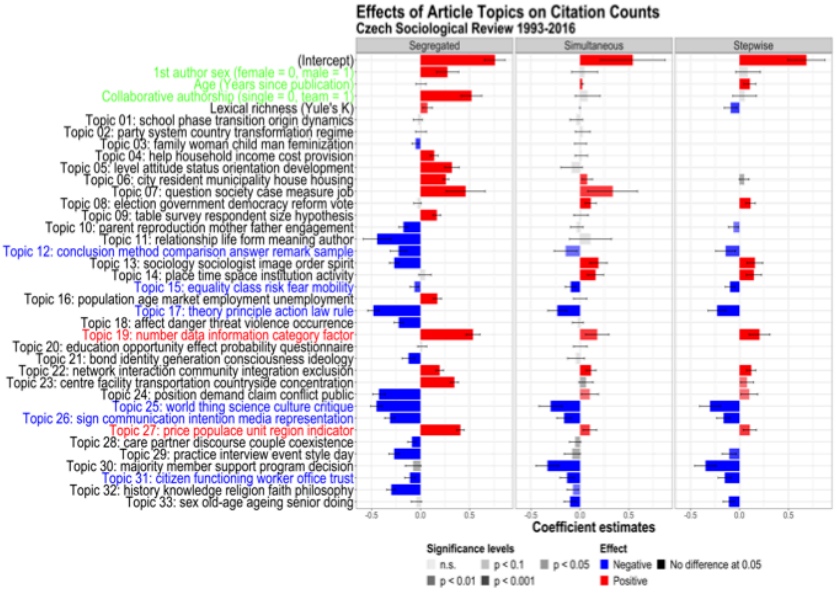
Comparison of effects



Consistent premium

- Topic 19: number data information category factor
  - ***quantitative sociology***
  - qualitative methods (topic 29) tends to be penalized
- Topic 27: price populace unit region indicator
  - ***social geography***
  - other topics in spatial social sociology, e.g. urban sociology and housing (topic 6), rural and regional sociology (topic 23) also tend to bring premiums

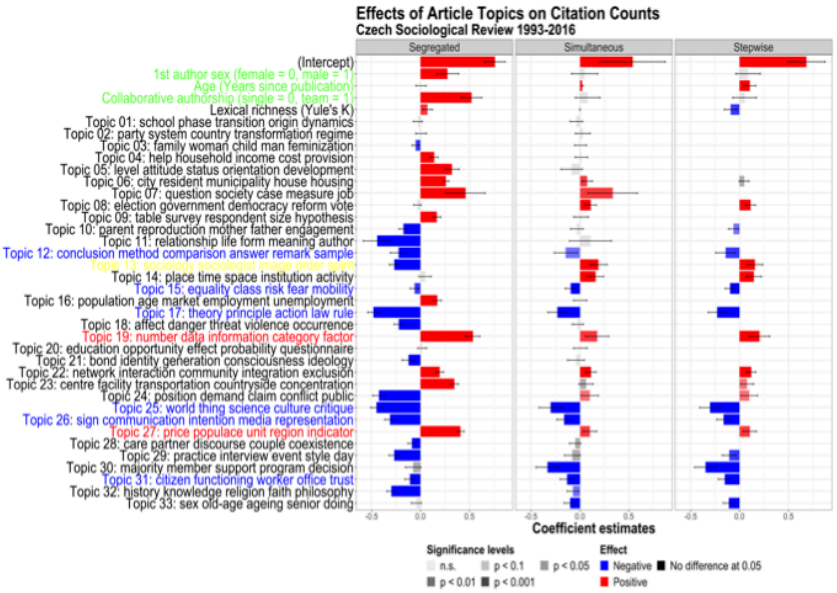
Comparison of effects



Consistent penalties

- Topic 12: conclusion method comparison answer remark sample
  - ***methodology***
- Topic 15: equality class risk fear mobility
  - ***stratification***
- Topic 17: theory principle action law rule
  - ***social theory***
- Topic 25: world thing science culture critique
  - ***philosophy***
- Topic 26: sign communication intention media representation
  - ***media studies***
- Topic 31: citizen functioning worker office trust
  - ***organizational sociology***

Comparison of effects



Sociology as a topic

- Topic 13: sociology sociologist image order spirit
  - as a topic of its own, ***reflexive sociology*** comes with a ***penalty***
  - congruent with the observed penalties for theory and methodology
  - when other topics are accounted for, ***sociological relevance*** adds citation ***premium***

Conclusions

Insights

- (some) topics can have effect on citations
  - penalties more prominent than premiums
- (some) factors affecting citations can be moderated by topics
  - gender of authors, collaboration
- classification of papers into single topics may overestimate the latter's effects

Applications

- journal IFs can be scored on their topical consistency
  - the lesser the model fit, the higher the consistency
- effect sizes can be used to calculate normalized, topic-adjusted citation counts
- Open Access can facilitate algorithmic extraction of topics as an alternative to classification systems based on information retrieval

The end

We have shown that algorithmically extracted topics are useful predictors of citations in a sociological journal. Building on a novel network approach to topic modeling, we introduced a topic scoring formula that avoids multicollinearity issues that come with the established approaches. Specific sociological topics were discussed.

11th CODH Seminar 2019

Tokyo (JP), September 25, 2019

Radim Hladík

National Institute of Informatics (Tokyo, JP)
Institute of Philosophy of the Czech Academy of Sciences (Prague, CZ)

Contact

radim.hladik@fulbrightmail.org
@hlageek

Acknowledgement

Grant-in-Aid for JSPS Fellows no. 17F17769.