Evaluating the effect of time and journal quality to topics: Structural topic models of scientific publications

Dr. Arho Suominen
Principal Scientist, VTT Technical Research Centre of Finland
Industry Professor, Tampere University, Finland
@ArhoSuominen

Dr. Arash Hajikhani
Research Scientist, VTT Technical Research Centre of Finland

BACKGROUND

- Studies have shown the applicability of topic models to classify textual data sourced from scientific publications and patents (e.g. Suominen & Toivanen 2015; Suominen, Toivanen, & Seppänen 2016; Yau et al. 2014)
- “Basic” topic models such as Latent Dirichlet Allocation are limited by capability to take into consideration of additional information, would that be term order of metainformation.
- Recent works have developed multiple targeted topic models, such as the author topic model (Rosen-Zvi et al. 2004) or dynamic topic model (Blei & Lafferty 2006), that allow for controlling one metadata attribute.

STRUCTURAL TOPIC MODEL

- Structural Topic Model (Roberts et al 2013) is a targeted topic models that allows discovering topics but also estimate their relationship to multiple metadata variables.

- Structural Topic accommodates corpus structure through document-level covariates affecting topic prevalence and/or topical content.

- Core idea is to specify the priors as generalized linear models then used to condition on data.

STRUCTURAL TOPIC MODEL

- Similar to other topic models, such as Latent Dirichlet Allocation:
  - Topic is a mixture over words where words has a probability to belong in a topic.
  - Document is mixture of topics, meaning that STM is a soft classification approach
- STM approach adds:
  - Topical prevalence covariates: metadata that explain topical prevalence
  - Topical content covariates: metadata that explain topical

STRUCTURAL TOPIC MODEL

- Metadata covariates
  - Topical prevalence: metadata affects the frequency with which a topic is discussed.
  - Topical prevalence: metadata affects the word frequency use within topic

- We use the R function tm to estimate the topical prevalence and content prevalence.

DATA – Pre-processing

- Data (N= 75478) is Web of Science data for Fuel Cell technology from 1991 to 2016.
- Preprocessing is done in Python:
  - Keep an merge AB and TI fields
  - Keep PY and ISBN
  - Create a feature JUFO (ISBN matching Finnish classification systems)
  - Remove short, under 500 character, AB+TI fields
    - Process reduces data from the original 75478 to 67490 documents
  - Consolidate fuel cell types using dictionary “Fuel_Cell_Types.csv”
  - Use POS tagging to clean others than NN, JJ, NNP
  - Remove punctuations and numbers
  - Data contained an erroneous field where PY was 173 – removed

DATA – Analysis

- Analysis is run using the R package stm
  - Topic number selection criteria K=0, approximates the number of topics
  - Retreive topic word probabilities and evaluate subjectively

    - highestProbability = highest probability terms
    - FREX = overall frequency and exclusivity to topic
    - Lift = weight of words in topic divided by frequency in other topics
    - Score = lift with log scaling

RESULTS

Topic 3 Top Words:

Highest Prob: hydrogen, reform, product, steam, gas, reactor, reaction

FREX: autotherm, reform, sorptionenhanc, nontherm, msr, lih, sesr

Lift: algraphit, allih, basestabil, butanetosynga, cellgrad, diammoni, diboran

Score: reform, hydrogen, steam, reactor, methan, product, ethanol

Topic 7 Top Words:

Highest Prob: reduct, oxygen, activ, reaction, orr, catalyst, graphen

FREX: orr, oer, fourelectron, pyridinicn, ndope, fenc, nitrogendop

Lift: agnr, carbonnanodiamond, coimidazol, concnt, copda, dioxidegraphen, doublevolcano

Score: orr, oxygen, reduct, graphen, activ, catalyst, reaction

Topic 10 Top Words:

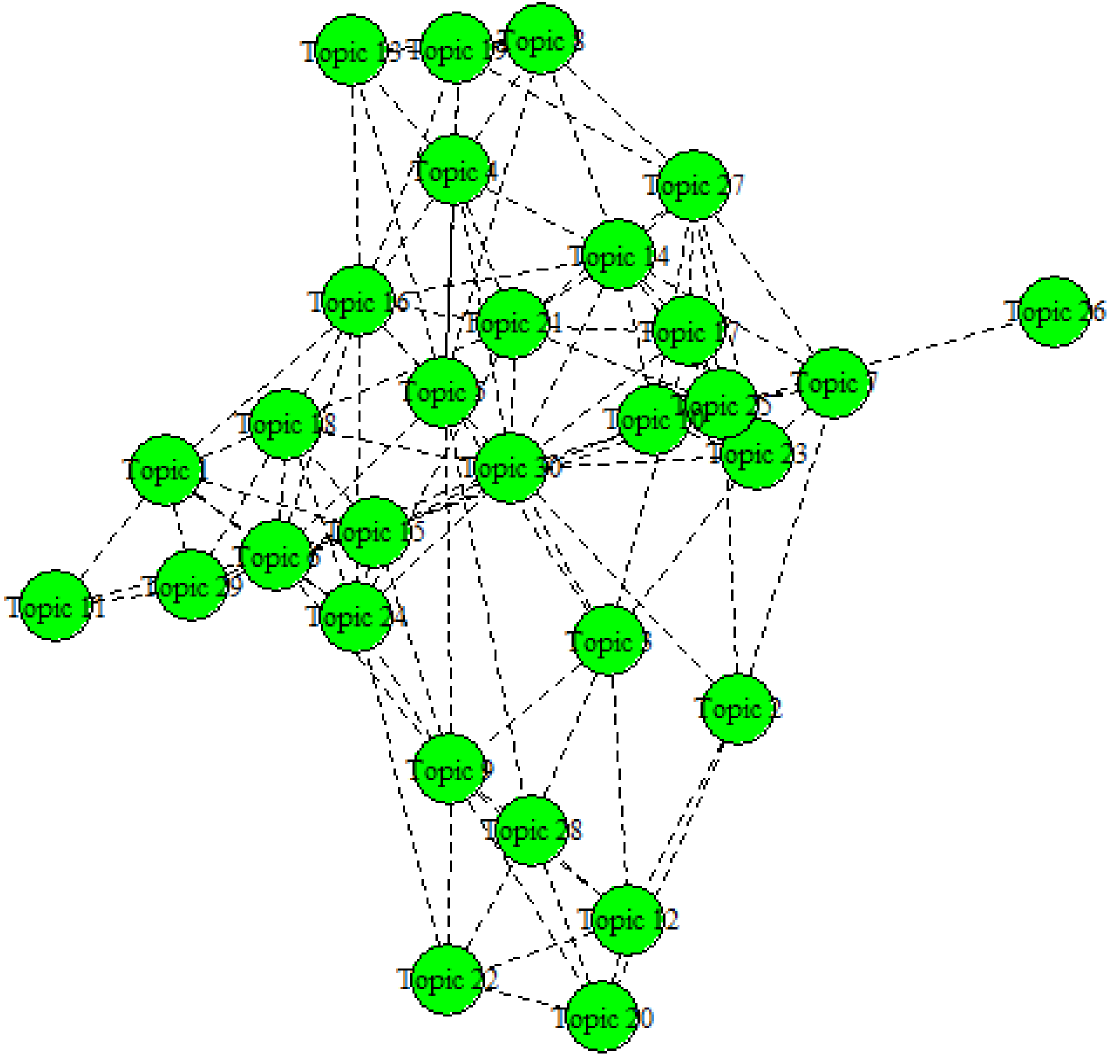
Highest Prob: oxid, activ, electrocatalyt, catalyst, acid, electrooxid, electrochem

FREX: ptruni, pdauc, pdnic, aupd, ptsnc, ifib, pdc

Lift: paani, palladiumlead, ptgns, absorptionnearedg, adgfc, aemdegfc, aeptm

Score: electrooxid, electrocatalyt, methanol, catalyst, ptru, formic, ptc

Topic 1: water, transport, liquid, gdl, diffus	Topic 16: electrochem, current, electrod, imped, polar
Topic 2: chemic, materi, metal, fuel, high	Topic 17: adsorpt, sulfur, adsorb, poison, remov
Topic 3: hydrogen, reform, product, steam, gas	Topic 18: membran, electrolyt, fuel, polym, catalyst
Topic 4: anod, nickel, nio, niysz, microstructur	Topic 19: conduct, oxygen, phase, temperatur, electr
Topic 5: sofc, solid, fuel, oxid, electrolyt	Topic 20: technolog, fuel, industri, commerci, process
Topic 6: flow, channel, gas, numer, mass	Topic 21: degrad, durabl, test, loss, cycl
Topic 7: reduct, oxygen, activ, reaction, orr	Topic 22: control, power, system, voltag, convert
Topic 8: process, deposit, film, powder, porous	Topic 23: catalyst, catalyt, activ, oxid, reaction
Topic 9: fuel, system, stack, design, power	Topic 24: pemfc, pem, fuel, temperatur, oper
Topic 10: oxid, activ, electrocatalyt, catalyst, electrooxid	Topic 25: carbon, electrod, electrochem, graphit, surfac
Topic 11: proton, acid, conduct, poli, sulfon	Topic 26: mfc, mfcs, power, cathod, electr
Topic 12: energi, hydrogen, electr, storag, fuel	Topic 27: electron, xray, spectroscopi, structur, microscopi
Topic 13: cathod, electrolyt, composit, temperatur, fuel	Topic 28: power, system, effici, gas, generat
Topic 14: surfac, oxid, alloy, metal, corros	Topic 29: membran, composit, conduct, nafion, exchang
Topic 15: methanol, dmfc, dmfcs, concentr, crossov	Topic 30: high, fuel, low, complex, combin


Topic 13
Topic 19
Topic 8
Topic 4
Topic 27
Topic 14
Topic 16
Topic 21
Topic 17
Topic 26
Topic 5
Topic 7
Topic 18
Topic 10
Topic 25
Topic 23
Topic 1
Topic 30
Topic 15
Topic 6
Topic 29
Topic 11
Topic 24
Topic 3
Topic 2
Topic 9
Topic 28
Topic 12
Topic 22
Topic 20

VTT

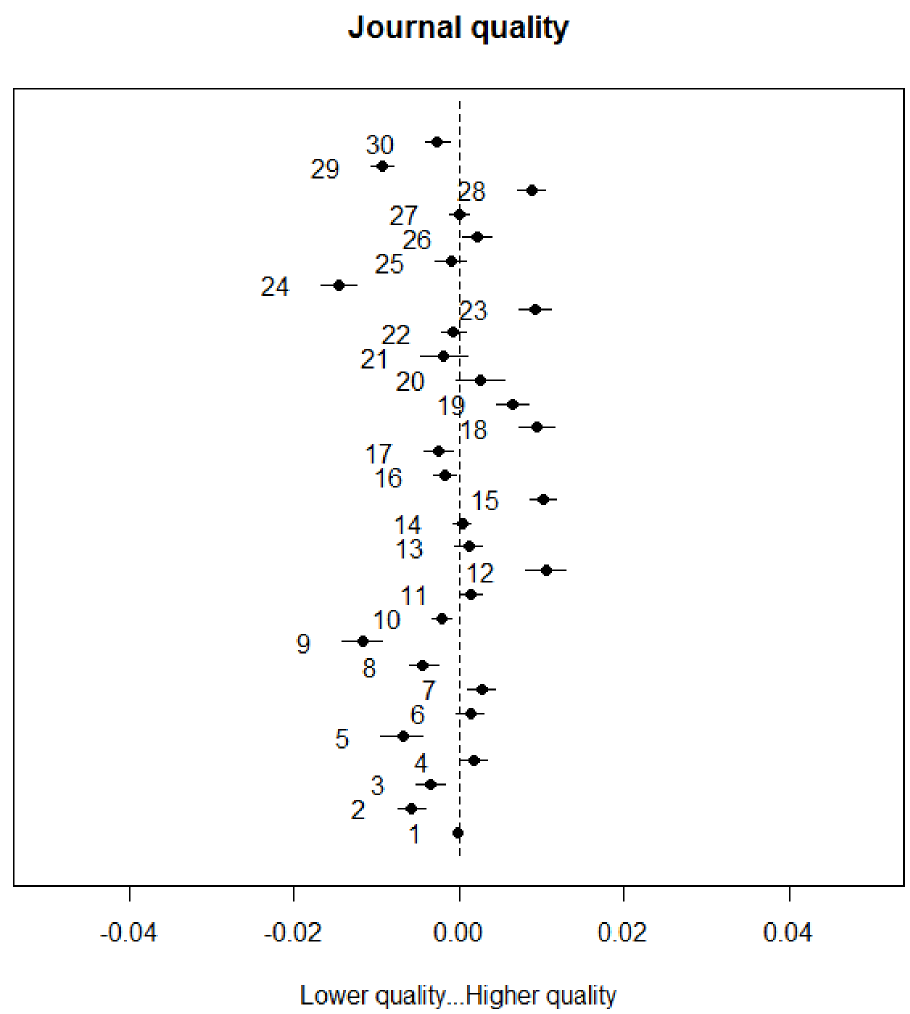
Calculated the effect of the dummy variable journal quality and a stepwise variable for time.

Example for Topic 29, time shown for 10 steps

Coefficients:					
Estimate Std. Error t value Pr(>|t|)					
(Intercept)	0.0198147	0.0096495	2.053	0.0400	*
JUFO2	0.0057500	0.0008371	6.869	6.52e-12	***
s(PY)1	-0.0210755	0.0182803	-1.153	0.2490	
s(PY)2	0.0034581	0.0083451	0.414	0.6786	
s(PY)3	0.0077393	0.0105535	0.733	0.4634	
s(PY)4	0.0105397	0.0096010	1.098	0.2723	
s(PY)5	0.0088766	0.0104084	0.853	0.3938	
s(PY)6	0.0108956	0.0107499	1.014	0.3108	
s(PY)7	0.0136161	0.0146565	0.929	0.3529	
s(PY)8	-0.0068462	0.0405122	-0.169	0.8658	
s(PY)9	0.0649827	0.0885411	0.734	0.4630	
s(PY)10	0.0170800	0.0099676	1.714	0.0866	.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

As with the majority of the topics, the variable time did not have a significant impact the topic. However, the variable JUFO, relating to journal quality did have a statistically significant impact.



FINAL NOTES

1. It seems that we could use the term evaluating approaches to identify emergence.
2. Adding covariates into topic models can highlight interesting relationships.
3. It was interesting to see how topics are impacted by journal quality, but not by time.

Questions?

Dr. Arho Suominen
Principal Scientist, VTT Technical Research Centre of Finland
Industry Professor, Tampere University, Finland
@ArhoSuominen

Dr. Arash Hajikhani
Research Scientist, VTT Technical Research Centre of Finland