Supplementary Materials for "Deciphering the Decline: A Computational Analysis of Two Decades of Canadian Newspaper Op-Eds on Freedom of Information" published in the Canadian Journal of Communication

Alex Luscombe
University of Toronto

Kevin Walby
University of Winnipeg

STM Hyperparameter Optimization

To fit our corpus to a structural topic model with 20 topics (K), we relied on a number of data-driven diagnostics recommended by Roberts, Margaret E and Stewart, Brandon M and Tingley, Dustin (2014), in addition to our expert judgment. All programming was conducted in R using the RStudio integrated development environment.[1] Using the R library known as stm (Roberts, Margaret E and Stewart, Brandon M and Tingley, Dustin, 2014), we ran stm::searchK() to obtain a held-out likelihood estimate and conduct a residual analysis on a range of possible models (5 through 50 topics, in increments of 5). We also considered the mean semantic coherence and exclusivity of each of our models, which stm::searchK() also calculates. Qualitative assessments of each model were conducted by examining lists of the top-weighted words for each topic in each model specification, in some cases generating samples of top-weighted documents for a given topic for more in-depth analysis (we used the same process to construct labels for each topic in the final model). Based on the results of the above metrics, in combination with qualitative analysis, we determined an optimal model fit for our purposes to be 20 topics. As a final check, we assessed the results of models with a K of 16 through 24 (results not shown here). This process confirmed that 20 was indeed the best fit.

Held-out Likelihood

Held-out likelihood estimation involves training a model on documents that have had a proportion of words removed prior to training (Wallach, Hanna M and Murray, Iain and Salakhutdinov, Ruslan and Mimno, David, 2009). This allows the researcher to then assess

[1] R code available on GitHub at https://github.com/alexlusco/deciphering-the-decline.

how well a model performs on data it has not seen before (i.e., 'held-out' from training). stm::searchk() calculates held-out likelihood by constructing a data set in which 10% of the documents in the corpus have had half of their words removed. Held-out log likelihood can be defined as

$$\mathcal{L}(w) = \log p(w|\Phi, \alpha) = \sum_{d} \log p(w_d|\Phi, \alpha).$$

where w_d is a collection of unseen documents, ϕ is the topic matrix, and α is the model hyperparameter. In theory, the higher the likelihood the better the model (but see discussion of semantic coherence below). The results of our held-out likelihood estimations are shown in Figure 1.

Residual Checks

When conducting topic modeling analysis, it is helpful to take into account the dispersion of residuals. High residual variability (overdispersion) can be an indication that a higher *K* (topics) is needed to absorb more of the variance. The mathematical basis for this strategy is described in Taddy, Matt (2012). The results of our residual analyses are shown in Figure 1.

Semantic Coherence

Semantic coherence measures the degree to which the top-weighted words in a given topic co-occur with one another (Mimno, David and Wallach, Hanna and Talley, Edmund and Leenders, Miriam and McCallum, Andrew, 2011). Semantic coherence correlates well with human judgement of topic quality, and as such is widely preferred over metrics like held-out likelihood (which does not correlate well with human judgement) (Mimno, David and Wallach, Hanna and Talley, Edmund and Leenders, Miriam and McCallum, Andrew, 2011). Letting $D(v)$ be a count of documents D containing at least one instance of word type v, $D(v, v')$ be a count of how many times word types v and v' co-occur in a document D, the semantic coherence of each topic or topic model can be defined as

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} log \left(\frac{D(v_i, v_j) + 1}{D(v_j)} \right)$$

where $(v_i, ...v_m)$ is a list consisting of each of the most probable words M in each topic k. We took the mean semantic coherence for each model and cross-compared the results. The results of our semantic coherence analysis are shown in Figure 1.

Exclusivity

Since high semantic coherence can be easily achieved by having a handful of topics dominated by the most common words in the corpus, it is helpful to also consider topic exclusivity, which can be used as a counterpoint to semantic coherence (Roberts, Margaret E and Stewart, Brandon M and Tingley, Dustin, 2014). The two metrics exist in a trade-off relationship. We used FREX to identify words that were frequent but also exclusive to each topic in a model. Words receive a high FREX score when they are both highly frequent and highly exclusive. The formula used to calculate FREX is

$$FREX = \left(\frac{w}{F} + \left(\frac{(1-w)}{E} \right)^{-1} \right)$$

where w is the word, F is the frequency score for that word, and E is the word's exclusivity, calculated by determining the conditional probability of the word appearing in a given topic after column normalizing the beta matrix of a fitted topic model (for more on calculating FREX, see Roberts, Margaret E and Stewart, Brandon M and Tingley, Dustin, 2014, p. 11). We took the mean exclusivity for each model and cross-compared the results. Select results of our exclusivity analysis (*K*-scores 15, 20, 25, and 30 only), plotted against semantic coherence, are shown in Figure 2. A list of the top 10 most probable FREX words for each topic in our final model is available in Table 1.

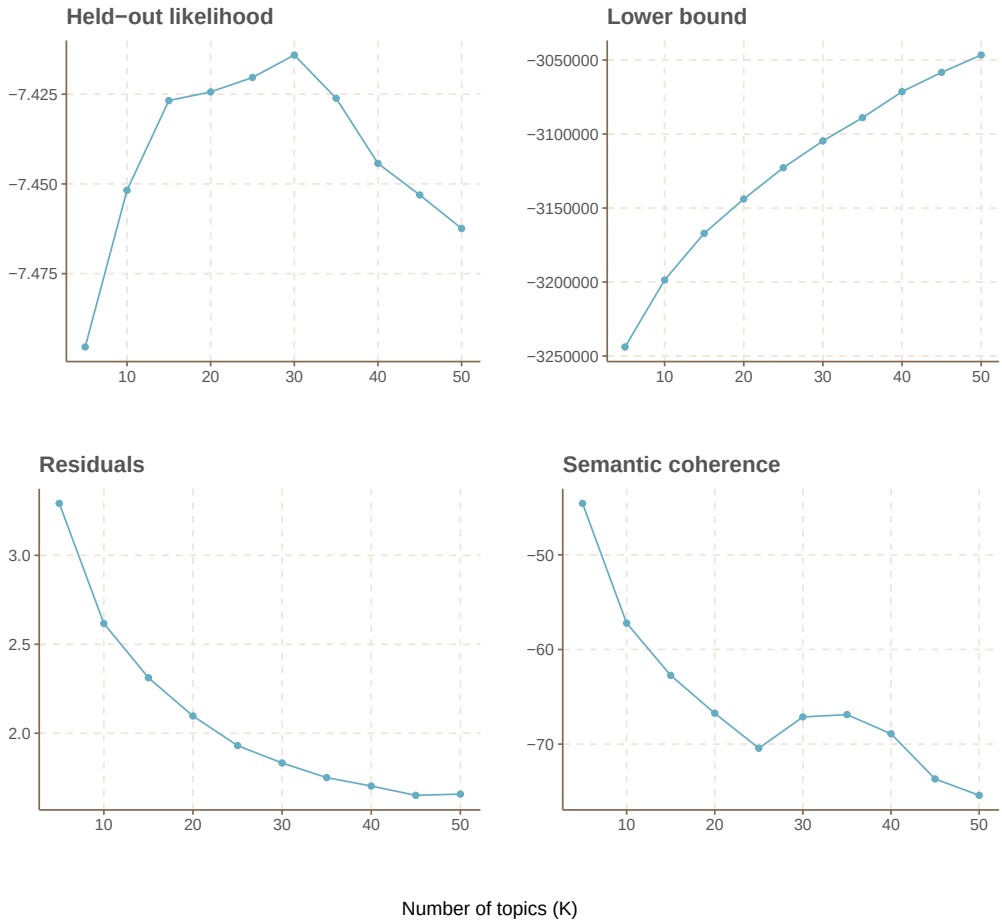

Figure 1: Held-out likelihood, residuals, lower bound, and semantic coherence estimates for STM models with 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 topics.

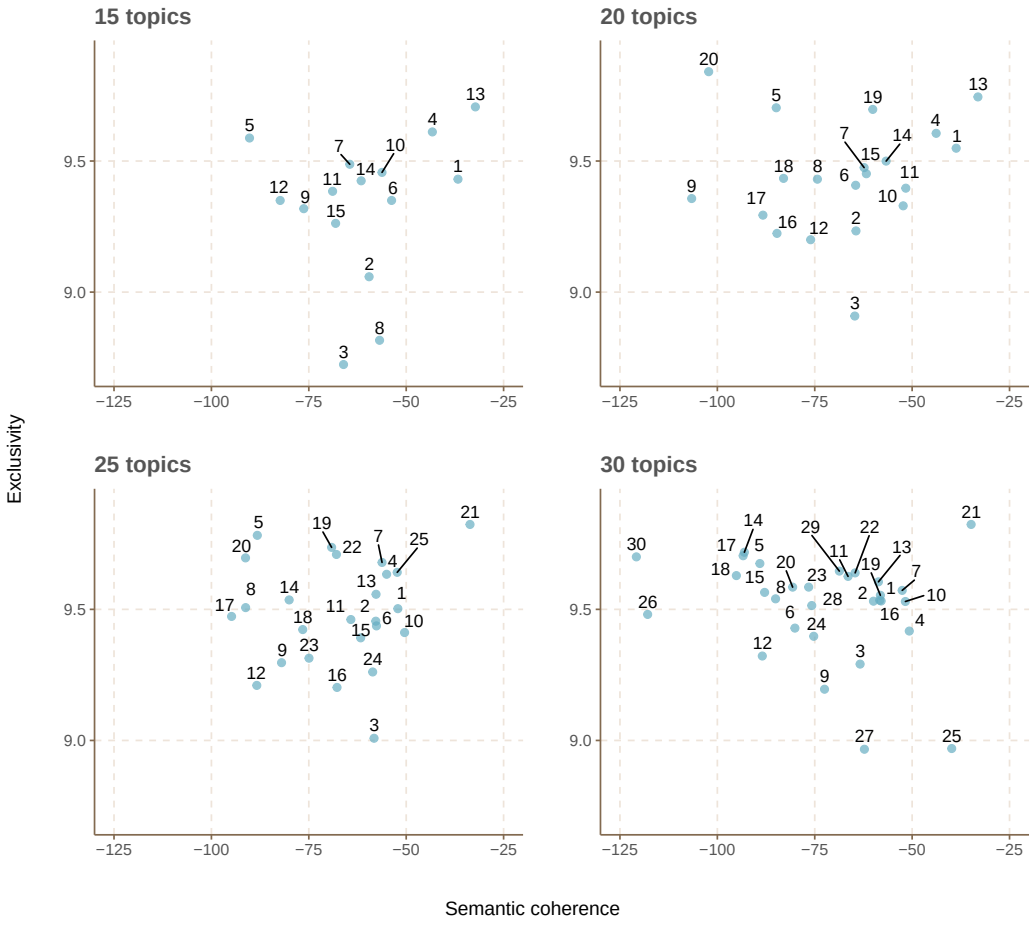

Figure 2: Comparing exclusivity and semantic coherence on models with 15, 20, 25, and 30 topics.

Table 1: 20 Topics by top 10 most probable words and FREX

	Topic Labels	Most Probable Words	FREX (FRequency and EXclusivity) Words
1	Information, Law and Governance	court, right, inform, law, privaci, govern, public, protect, freedom, bill	court, suprem, charter, privaci, constitut, mclellan, terror, appeal, claus, judg
2	Online Data and Internet Technology	inform, access, internet, librari, use, will, peopl, comput, can, technolog	librari, internet, comput, googl, web, technolog, buffer, onlin, zone, mosquito
3	Environment, Agriculture, and the Development	water, govern, park, will, one, industri, report, environment, feder, health	anim, oil, pipelin, gas, food, pollut, water, cattl, park, fuel
4	Harper Conservatives and Federal Politics	govern, harper, elect, conserv, parti, will, polit, minist, liber, promis	harper, voter, stephen, vote, conserv, elect, poll, parti, tori, scientist
5	Diversity, Education, and Citizenship	school, children, student, parent, educ, women, famili, one, immigr, univers	parent, school, student, children, teacher, gay, women, girl, abort, kid
6	Public Police and Accountability	rcmp, polic, inform, report, investig, public, releas, forc, use, death	taser, rcmp, dougla, dziekanski, mounti, death, die, incid, video, csis
7	Health Care and Information Governance	health, per, cent, hospit, care, year, provinc, govern, servic, patient	physician, patient, hospit, doctor, medic, health, care, cancer, drug, cent
8	Municipal Politics	citi, council, will, mayor, ferri, communiti, municip, meet, new, one	ferri, citi, mayor, hall, council, municip, brandon, councillor, hahn, ncc
9	Foreign Affairs and Trade	countri, govern, world, state, china, american, trade, will, refuge, nation	edc, china, refuge, chines, trade, american, saudi, iraq, bush, islam
10	Liberal Scandals and Federal Politics	minist, govern, prime, liber, chretien, offic, committe, public, will, parliament	chretien, gomeri, dingwal, jean, martin, ethic, bryden, sponsorship, pbo, mps
11	Provincial Information Commissioners and Compliance	govern, inform, privaci, law, liber, premier, person, bill, public, act	denham, foi, clark, loukid, delet, campbel, dickson, ministri, legislatur, gordon
12	Gun Control and Database Politics	polic, registri, gun, offic, law, crimin, inform, use, will, one	registri, firearm, gun, longgun, iiu, regist, polic, licens, licenc, registr
13	Federal Information Commissioners and Compliance	inform, govern, public, access, request, act, open, report, law, commission	reid, secreci, commission, bureaucrat, request, open, legault, delay, inform, servant
14	Law and Justice Procedures	victim, crime, system, will, justic, bill, crimin, case, time, inform	victim, trial, jail, montpelli, neglect, sentenc, convict, mental, guilti, crime
15	Journalism and News Production	media, journalist, report, news, press, public, freedom, time, can, stori	journalist, media, journal, star, stori, cjfe, press, news, newspap, daili
16	Intergovernmental Relations	govern, manitoba, minist, public, premier, report, polit, will, pallist, provinc	pallist, manitoban, manitoba, fippa, winnipeg, trudeau, anglophon, census, bilingu, redford
17	Military and Defence	militari, defenc, report, soldier, afghanistan, forc, foreign, year, govern, public	militari, detaine, soldier, afghan, hillier, afghanistan, mission, defenc, diplomat, troop
18	Security and Corrections	inform, agenc, secur, releas, one, prison, communic, intellig, document, can	dfo, parol, spi, toew, intellig, airlin, csis, supervis, prison, inmat
19	Fiscal Policy	taxpay, public, money, govern, million, expens, spend, feder, dollar, corpor	dollar, taxpay, spent, expens, cbc, money, corpor, gazett, contract, spend
20	Indigenous Politics	nation, first, alberta, govern, chief, feder, aborigin, reserv, band, salari	reserv, aborigin, nativ, alberta, band, ctf, salari, calgari, indian, chief

References

Mimno, David and Wallach, Hanna and Talley, Edmund and Leenders, Miriam and McCallum, Andrew. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).

Roberts, Margaret E and Stewart, Brandon M and Tingley, Dustin. (2014). stm: R package for structural topic models. *Journal of Statistical Software*, *10*(2), 1–40.

Taddy, Matt. (2012). On estimation and selection for topic models. In *Proceedings of the 15th international conference on artificial intelligence and statistics* (pp. 1184–1193).

Wallach, Hanna M and Murray, Iain and Salakhutdinov, Ruslan and Mimno, David. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112).