Minimizing Conflicts of Interest

Optimizing the JSM Program

Luca Frigau[1], Qiuyi Wu[2], David Banks[3]

University of Cagliari[1], University of Rochester[2], Duke University[3]

Aug 12, 2021

Joint Statistical Meetings

Every year, ASA has to convene 47 section program chairs and other representatives for a two-day in-person meeting to finalize the JSM schedule. Normally, the sessions of the program are manually assigned to time bands by the program committee. Assignment is a labor-intensive and time-intensive process.

Participant Satisfaction Survey

> *"Each year we send a questionnaire out to a subset of JSM attendees asking questions about their overall satisfaction with the program. ... I read through the results and comments every year and find the biggest complaint to be about subject conflicts. With so many concurrent sessions, this is nearly impossible to avoid. As you well know, the program committees work hard each year to minimize these conflicts as much as possible."*
>
> *— Ms. Kathleen Wert, Director of Meetings for the ASA*

Structure of JSM

Structure of JSM is governed by the Joint Agreement among the four founding societies: the ASA, the Institute of Mathematical Statistics, the International Biometric Society, and the Statistical Society of Canada.

- In even years, there are 181 invited sessions; in odd years there are 209 invited sessions as IMS holds its annual meeting at the JSM.
- 7 categories of sessions: plenaries, invited sessions, topic contributed sessions, contributed sessions, introductory overview lectures, late-breaking sessions, and memorial sessions.
- Up to 44 rooms available for parallel sessions.
- Three 110 minute time bands on Monday, Tuesday and Wednesday and two 110 minute time bands on Sunday and Thursday

Sunday	Monday	Tuesday	Wednesday	Thursday
110 min	110 min	110 min	110 min	110 min
110 min	110 min	110 min	110 min	110 min
	110 min	110 min	110 min	

Who?

The JSM program is set by the Program Committee, which consists of

- program chairs of each section
- representatives from sister societies participating in the JSM
- three overall program chairs (past, present, and future)
- three associate chairs

In 2020

It had 47 members who worked with the American Statistical Association(ASA) staff over the course of the year, with a two-day meeting at the ASA headquarters in February to finalize the schedule.

A key activity during the meeting was to minimize overlapping content in the same time band. The in-person meeting was time consuming and expensive to the ASA.

How?

- Goal: Minimize the overlap in topics within the same time slot.
- Data: Our information on the 2020 JSM program consists of the title, keywords, and abstract text for each talk in each session. To boost signal, we enrich the data by representing the title of each talk and the associated key words three times.
- Approaches:
  - Use an extension of LDA (Latent Dirichlet Allocation) to automatically identify broad topics (e.g., environmental statistics, time series analysis, survey methodology, and so forth).
  - Use a scheduling algorithm to assign sessions to time bands such that the amount of conflict among parallel sessions is minimized.

Data Cleaning

1. remove stop words
2. stem the words
3. n-gramming
4. remove non-statistical text
5. manually removed the tokens corresponding to "statistician" and "statist"

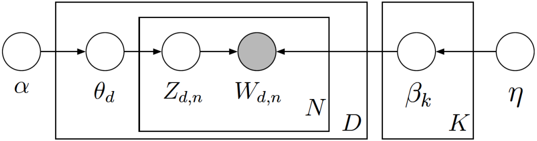
Number of Reduced Words

The first step reduced the number of words to 21,813 tokens. The second step and third steps decreased the number of tokens to 17,471, of which 14,017 are ngrams. The fourth step, using the Trayvon Martin corpus, reduced the number of tokens to 15,755.

Method

- We use seeded LDA to find which tokens have high probability within partially prespecified topics that correspond to distinct subfields of statistics at the JSM.
- Our work used the governance structures of the founding societies to guide the seeding. In particular, we created topic seeds for nearly every section of the ASA.

Graphical Model of LDA



1. For each topic $k \in \{1, ..., K\}$
   - Draw $\beta_k \sim \text{Dirichlet}(\eta)$
2. For each document $d \in \{1, ..., D\}$
   - Draw $\theta_d \sim \text{Dirichlet}(\alpha)$
3. For each word w_n in document d, $n \in \{1, ..., N\}$
   - Draw topic $z_n \sim \text{Multinomial}(\theta_d)$
   - Draw word $w_n | z_n \sim \text{Multinomial}(\beta_k)$

In applications, MCMC reverses the generative process, estimating the distributions corresponding the topics and the probabilities for all documents. In our application, a document corresponds to a JSM session.

Seeded LDA

- Seeded LDA is one of the variants of LDA which incorporates the user's understanding of the corpus and bias the topic formation process with the help of representative word of each topics.
- Our work used the governance structures of the founding societies to guide the seeding. In particular, we created topic seeds for nearly every section of the ASA.
- In order to seed LDA we identified words that are reasonably specific to topics we want to define. We looked at the stemmed words' frequencies and assigned fairly specific but relatively common tokens to topics.

Seeded LDA

- This process excluded extremely frequent stemmed words, such as bayesian, infer, asymptot, gaussian and causal, because they were insufficiently specific, and naturally appear in many different topics.
- But we did use them when they appeared in n-grams that corresponded to more narrow topics, such as bayesian-nonparametr, asymptot-distribut, gaussian-process and causal-infer.
- Additionally, we added ten unseeded topics, to account for areas we may have overlooked, and to allow the data to speak for itself.
- We performed LDA with $K = 51$ topics: 41 seeded and 10 unseeded.

Distinctivity

Distinctivity: quality of the resulting topic estimates

The distinctivity of the jth token for the kth topic is the posterior probability of the kth topic given that the jth token appears in that session's text, for a uniform prior over the topics.

- Highly distinctive tokens are specific to a single topic
- Tokens with smaller posterior probabilities are associated with multiple topics

Astrostatistics	prob	**Causal Inference**	prob	**Risk Analysis**	prob	...
astronom	1.0	counterfactu	1.0	cox_model	1.0	...
astrostatist	1.0	causal_infer	1.0	risk_predict	0.9	...
astrophys	1.0	causal_mediat	0.9	cox_proport	0.9	...
astronomi	0.9	unmeasur_confound	0.9	risk_factor	0.7	...
stellar	0.8	unmeasur	0.9	compet_risk	0.6	...

Table 1: five most distinctive tokens for each of the seeded topics

Distinctivity

We probably had not overlooked too many topics, and this guess was supported by the fact that seven of the unseeded clusters were reasonably coherent, but three were not.

- Unseeded topics 1, 3, and 10 seem difficult to identify, and probably have absorbed noise rather than signal
- Unseeded topic 2 relates to the environment, topic 4 concerns cluster analysis, topic 5 is about additive regression trees, 6 is about MCMC, 7 is about false discovery rates, and 8 is about regression.

Topic 4	prob	**Topic 6**	prob	**Topic 10**	prob	...
trio	0.8	bayesian_infer	0.9	rasch_tree	0.6	...
prompt	0.8	posterior	0.9	demand_elast	0.5	...
model_bas_cluster	0.8	variational_infer	0.9	hospic	0.5	...
latent_class_model	0.7	conjug	0.9	rasch	0.5	...
parent_child	0.7	mcmc	0.9	dif	0.5	...

Table 2: five most distinctive tokens for each of the unseeded topics

Seeded LDA implementation

- Within this framework of seeded and unseeded topics, we applied seeded LDA to the pooled abstracts in each session. The traceplot raised no concerns about convergence (total computing time was 70 secs).
- The calculation estimated the percentage of each session that is “drawn” from each of the available topics. In nearly all cases, a session has substantial percentages from a handful of topics, but small percentages from many topics.
- Regularize outcome: set the percentage of the topic having the smallest non-zero percentage to zero and reallocate its weight proportionally to the remaining topics.
- The reallocation process terminates when all remaining topics have percentages at least equal to 20%.

Sensitivity Analysis

As usual with LDA, the analysis depends upon a large number of parameters, and it would be problematic if the results were sensitively dependent upon those choices.

Parameter Choices

- Minimum number of times a token had to appear: 3, 4, **5**, 6
- Length of n-gram: 3, 4, **5**, 6
- Significance probability cutoff needed for declaring an n-gram: 0.05, 0.01, **0.005**, 0.001

Sensitivity Analysis

- Our focus is upon interpretability of the topics in terms of the statistical discipline.

Thus, we compared the 20 most distinctive tokens in the each of the 51 topics across the 64 experimental conditions in our experiment.

- Looked at the the number of top-twenty distinctive tokens in common with our analysis among the topics found in these new variations
- Averaged the number of common distinctive tokens over all 51 categories and all 63 runs

The average and the standard deviation were 10.21 and 0.380, indicating much agreement among the distinctive word lists. This outcome shows that our results are fairly insensitive to the parameter choices.

Session Assignment Issue

Recall

- The JSM schedule has 13 110-minute time bands when parallel sessions may occur.
- The program chairs from all the ASA sections meet for two or three busy days, and that builds upon much prior work by the JSM program chair and three ASA staff members.
- At all steps of the process, everyone works together to try to minimize overlap in content.

Issue

Nonetheless, many JSM attendees still complain about being forced to choose between two similar sessions in the same time band.

Session Assignment Solution

An optimal solution is not unique:

- Interchanging the assignments for any two time bands having the same number of parallel sessions provides an equally good solution.
- Interchanging two sessions in different bands that both have the same amount of participation in the same topics provides an equivalent solution

Session Assignment Constraint

- A soft constraint that sections with two or fewer guaranteed invited sessions not be scheduled on Sunday or Thursday, but this is often waived (based upon speaker availability, organizer request, or to avoid double-booking) and was not used in our scheduling.

Time band	Day	From	To	# of parallel sessions
1	08-02-2020	14:00	15:50	38
2	08-02-2020	16:00	17:50	38
3	08-03-2020	08:30	10:20	40
4	08-03-2020	10:30	12:20	43
5	08-03-2020	14:00	15:50	43
6	08-04-2020	08:30	10:20	40
7	08-04-2020	10:30	12:20	43
8	08-04-2020	14:00	15:50	44
9	08-05-2020	08:30	10:20	41
10	08-05-2020	10:30	12:20	41
11	08-05-2020	14:00	15:50	40
12	08-06-2020	08:30	10:20	37
13	08-06-2020	10:30	12:20	39

- An additional constraint on the scheduling is made to ensure that no one is double-booked in the same time band.
- Two additional constraints in random assignment: no time band has two or more introductory overview lectures (IOLs), and IOLs should be scheduled on Monday, Tuesday or Wednesday.

Session Assignment

Let $\sigma = \{s_1, ..., s_N\}$ be the set of sessions that must be assigned to a band. Let $\Gamma = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1K} \\ \vdots & \ddots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NK} \end{bmatrix}$ be parameters of the posterior topic distribution for each document (the extent to which session i participates in topic j), $\gamma_{ij} \geqslant 0.2$.

Total variation distance for sessions s_i and s_j:

$$\delta_{ij} = \frac{1}{2} \sum_{k=1}^{K} |\gamma_{ik} - \gamma_{jk}|$$

Small values of δ_{ij} imply the sessions have strongly overlapping content.

Measure topic overlap:

$$\rho = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} \theta_{ij}$$

where $\theta_{ij} = 1$ when s_i and s_j are assigned to the same time band, and otherwise 0. A good assignment produces a large value of ρ.

Session Assignment Procedure

Procedure

1. Assign people with over one role to sessions in different time bands
2. Randomly assign the remaining sessions; calculate ρ
3. Greedily optimize the assignment
    - randomly picks two time bands and swaps two sessions at random between them, then calculate the new ρ_{new}.
    - If $\rho_{new} > \rho$, keep the swap; otherwise the swap reverts and a new swap is tried.
    - Terminate the algorithm when 10,000 attempted swaps have not produced a larger ρ.

Steps 2 and 3 are repeated 100 times, and then we use the best local optimum found. More repetitions might find slightly better optima, but negligible.

Optimizing the 2020 JSM Schedule

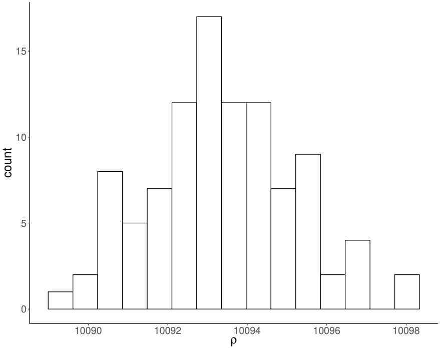
- The average ρ from 100 random allocations with no search for improvement was $9,891.31$, considered as minimum.
- The original assignment made by the 2020 JSM organizing committee had $\rho = 9,923.74$
- The best assignment obtained by the algorithm had $\rho = 10,098.06$
- If there's no overlapping content in any time band, ρ reaches its maximum as $\rho = 10,488$

Improvement $= (\rho - \text{min})/(\text{max}-\text{min})100\%$

	ρ	Improvement
Original Assignment	9,923.74	5.8%
Best Assignment	10,098.06	37.1%

Table 3: Improvement of the hill-climbing searches

Scheduling Algorithm



The histogrom shows the histogram of the local maxima found in our 100 random restart hill-climbing searches. It is approximately normal with mean 10093.38 and standard deviation 1.825. Our largest local maximum is 2.56 standard deviations above the mean, which suggests that further search would not find a meaningfully better maximum.

JSM 2020 schedule

Time band	# of parallel	Best assignment		JSM 2020	
	sessions	Sum	Average	Sum	Average
1	38	677.46	0.964	649.19	0.923
2	38	677.79	0.964	650.13	0.925
3	40	754.69	0.968	748.33	0.959
4	43	875.87	0.970	869.37	0.963
5	43	876.90	0.971	841.00	0.931
6	40	750.13	0.962	746.71	0.957
7	43	877.27	0.972	863.91	0.957
8	44	917.35	0.970	894.50	0.946
9	41	793.49	0.968	789.14	0.962
10	41	793.17	0.967	776.94	0.947
11	40	751.87	0.964	742.40	0.952
12	37	638.11	0.958	639.27	0.960
13	39	713.96	0.964	712.85	0.962
		10098.06		9923.74	

Comparison of ρ values from our best schedule to that planned for the 2020 JSM.

- The mean for the average column is larger for the best schedule than that for 2020 JSM schedule.
- The assignment averages are more than three times less dispersed than the 2020 JSM averages (mean $\pm$ sd is 0.966 ± 0.004 for the best assignment and 0.950 ± 0.014 for the 2020 JSM assignment).

Summary

- A general methodology for minimizing overlapping content in complex professional society meetings with multiple tracks
- Methods:
  1. Use seeded topic models to identify overlapping content in the same time band
  2. Use an optimization algorithm to reschedule sessions so as to reduce that overlap
- Obtained a much improved schedule: the original program showed about a 6% improvement, whereas our method achieved at least a 37% improvement
- ASA could improve the JSM experience and save a significant amount of money by adopting the method

Thank you for you attention!

We gonna have a better JSM experience next year!

Reference

Frigau, L., Wu, Q., Banks, D. (2021). Optimizing the JSM Program (submitted to JASA A&CS, minor revision)