Appendix A

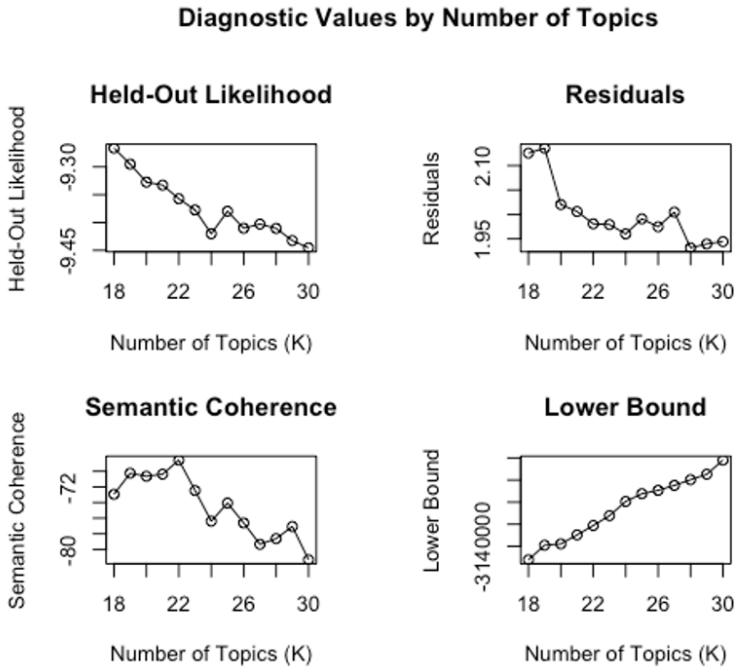
On K Resolution

To find appropriate K (number of topics) for the topics in the model and better understand their contents, the author applied several K to decide K resolution. 23 topics have been identified with the STM analysis. The Figure below illustrates how K is determined. According to Roberts et. al. (p.42) ," both the residual checks and held-out likelihood estimation are useful indicators of the number of topics that should be chosen."

A. Held-out likelihood: Hold out some fraction of the words in training and use the document-level latent variables to evaluate the probability of the heldout portion.
   The higher, the better.
   There are about two clusters: high and low. Topic Resolution between 18 and 23 belong to the HIGH cluster, and the rest are located in LOW cluster. The potential candidates are in the HIGH cluster, which are 18, 19, 20, 21, 22, 23.
B. Residual: Testing for overdispersion of the variance of the multinomial within the data generating process of STM.
   The Lower the better.
   There are about two clusters: high and low. Topic Resolution between 22 and 24, and between 28 to 30 are belong to the LOW clusters, and the rest are located in HIGH cluster. The potential candidates are in the LOW cluster, which are 22, 23, 24, 28, 29, 30.
C. Lower Bound: The lower bound is the approximation to the lower bound on the marginal likelihood. Once the bound has a small enough change between iterations, the model is considered converged.
   The Smooth, the better.
   There are about two clusters: STEEP and SMOOTH. Topic Resolution between 24 and 29 belong to the SMOOTH cluster, while the others are rather steeper. 24, 25, 26, 27, 28, 29.
D. Semantic Coherence: Semantic coherence is maximized when the most probable words in a given topic frequently co-occur together.
   The higher, the better.
   There are two clusters in the model: HIGH and LOW. Topic Resolution between 18 and 23 belong to the HIGH cluster, and the rest are located in LOW cluster. The potential candidates are in the HIGH cluster, which are 18, 19, 20, 21, 22, 23.

There ain't no "right" answer to the number of topics that are appropriate for a given corpus. Within certain boundaries, it seems that the choice of model is a matter of trade-offs. The author tested K=22, 23, 28, 29 and K=23 was finally appointed.
The below are diagnostic graph and table (and syntax) in regard to the "K resolution."

```
#diagnostic graph
set.seed(02139)
K<-c(18:30)
CRASresults <- searchK(docs, vocab, K, prevalence=~term+event, data=K3, max.em.its = 100)
#K3=CRAS2021.csv
plot(CRASresults)
```




```
> CRASresults$results
   K   exclus    semcoh   heldout residual    bound   lbound em.its
1  18 9.433207 -72.94311 -9.266471 2.125665 -3152192 -3152155    100
2  19 9.439343 -70.24578 -9.295111 2.135735 -3139033 -3138994     94
3  20 9.451491 -70.64297 -9.327455 2.020244 -3137831 -3137788     69
4  21 9.464768 -70.37094 -9.332847 2.005915 -3129698 -3129652     92
5  22 9.458395 -68.57491 -9.357336 1.980794 -3121201 -3121152    100
6  23 9.473822 -72.46399 -9.377472 1.979079 -3112286 -3112234     61
7  24 9.502889 -76.39287 -9.420251 1.960101 -3099472 -3099417     77
8  25 9.503575 -74.06635 -9.379818 1.990916 -3092443 -3092385     86
9  26 9.523902 -76.61665 -9.410353 1.974643 -3089388 -3089327     94
10 27 9.534919 -79.34516 -9.402912 2.004666 -3084918 -3084854     70
11 28 9.542182 -78.62027 -9.410603 1.931451 -3079779 -3079711     75
12 29 9.541866 -77.06611 -9.432613 1.939412 -3074816 -3074744     87
13 30 9.574973 -81.32405 -9.445554 1.944197 -3062052 -3061978     72
```

```
# Keywords in Topics (by frex)
library(stm)
processed <- textProcessor(K3$newcc, K3,wordLengths = c(2,Inf))
out <- prepDocuments(processed$documents, processed$vocab, processed$meta)
docs <- out$documents
vocab <- out$vocab
meta <- out$meta
out$meta$speech_date<- as.numeric(as.Date(out$meta$speech_date))
CRAS_stm.out_23 <- stm(out$documents, out$vocab, K=23, prevalence = ~ term+event,
data=out$meta, init.type="Spectral")
plot(CRAS_stm.out_23, type = "summary",labeltype="frex", family = "Songti SC", n=7, xlim =
c(0, .3))
```