IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts[1]

Hector Carcel-Villanova,
International Monetary Fund

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Text Data Analysis using Latent Dirichlet Allocation, an Application to FOMC Transcripts[1]

Hali Edison (Williams College)

Hector Carcel-Villanova (International Monetary Fund)

Abstract

This short paper explains Latent Dirichlet Allocation (LDA), a machine learning algorithm, and uses it to analyse the content of U.S. Federal Open Market Committee (FOMC) transcripts covering the period 2003–2012, including 45,346 passages. The results of this exercise show that discussions on economic modelling were dominant during the Global Financial Crisis (GFC), with an increase in discussions on the banking system in the years following the GFC. Discussions on communication also gained relevance towards the end of the sample. LDA analysis could be further exploited by researchers at central banks and institutions to identify topic priorities in relevant documents such as the FOMC transcripts.

Keywords: FOMC, Text data analysis, Transcripts, Latent Dirichlet Allocation.

JEL classification: E52, E58, D78.

Contents

[1]This version is an adaptation from: Edison, H. and Carcel, H. (2021) Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. *Applied Economics Letters* 28: 1, 38-42.

The views expressed in this paper are those of the author and do not necessarily represent the views of the IMF, its executive board or management. Special thanks to C. Schwarz for providing the Stata LDA codes.

1. Introduction

Text data analysis can be a useful tool to analyse the main topics addressed in central bank official documents. The aim of this note is to explain the LDA methodology as presented in Schwarz (2018), and show some results obtained using the ldagibbs Stata command. We analysed 45,346 entries of the Federal Open Market Committee (FOMC) during the period of 2003-2012, with the goal of detecting the evolution of the different topics discussed by the members of the FOMC.

Overall, we detected that discussions on economic modelling played an important role during the Great Financial Crisis (GFC), followed by a considerable increase in discussions on the banking system in the following years, and discussions on communication gained importance at the end of the sample.

2. FOMC Meetings

The FOMC decided in 1976 to release and make publicly available a detailed memorandum of all the discussions taking place at its meetings. The Federal Reserve Act states that the objectives of monetary policy enhanced by the FOMC shall "promote effectively the goals of maximum employment, stable prices and moderate long-term interest rates". There exists considerable debate among economists on how to translate these goals into a coherent description of U.S. monetary policy. This is the reason why a detailed and precise account on the discussions taking place during the FOMC meetings can result useful to understand the evolution in the conduct of U.S. monetary policy.

The FOMC meets eight times in a year to formulate monetary policy and determine other Federal Reserve policies. It is composed of nineteen members comprising seven Governors of the Federal Reserve Board located in Washington D.C., of whom one is the Chairperson of both the Board of Governors and the FOMC, and twelve Presidents of the Regional Federal Reserve Banks with the President of the New York Fed as Vice-Chairman of the FOMC.

The main policy variable of the FOMC is a target for the Federal Funds rate, as well as potential guidance on future monetary policy. At every meeting, all the seven governors have a vote, together with the president of the New York Fed and four of the remaining eleven Fed Presidents who vote on a rotating basis.

Most FOMC meetings last a single day except the meetings that take place before the Monetary Policy Report for the President, which last for two days. During each meeting, every member participates in the discussions independently from their voting right. In this note we analyse the transcripts of these meetings focusing on the conversations held between the FOMC members.

3. Methodology

As explained in Schwarz (2018), Latent Dirichlet Allocation (LDA) consists of two parts. The first is based on a probabilistic model describing the text data as a likelihood function. In the second part, given the unfeasibility of maximizing the likelihood function of text data, LDA utilizes an inference algorithm.

The probabilistic model of LDA considers that every document *d* of the *D* documents in the whole text can be assumed as a probabilistic mixture of *T* topics. These probabilities can be found in a document vector θ_d of length *T*. The value of T, that is, the number of topics, is decided by the user according to the preciseness required. The output of LDA is a $D \times T$ matrix θ containing the probabilities $P(t_t|d_d)$, of each document *d* belonging to topic *t*:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix} = \begin{pmatrix} P(t_1|d_1) & \cdots & P(t_T|d_1) \\ \vdots & \ddots & \vdots \\ P(t_1|d_D) & \cdots & P(t_T|d_D) \end{pmatrix}$$

Each topic $t \in T$ is defined by a probabilistic distribution over the vocabulary (the set of words in all documents) of size *V*. In this paper, documents related to forecasting will have a high probability of containing words such as "expectations" or "market-based", while in documents related to economic modelling there will be a higher probability of finding terms such as "model", "standard errors" or "shocks". The word probability vectors of the topics can be represented in a matrix φ of dimensions $V \times T$:

$$\varphi = (\varphi_1 \ldots\ldots, \varphi_T) = \begin{pmatrix} P(w_1|t_1) & \cdots & P(w_1|t_T) \\ \vdots & \ddots & \vdots \\ P(w_v|t_1) & \cdots & P(w_v|t_T) \end{pmatrix}$$

The probabilities $P(w_v|t_T)$ in φ_t describe how probable it is to observe word *w* from the vocabulary conditional on topic *t*. Hence, the φ_t vectors permit to decide the content of each topic and how each topic can eventually be named, since LDA does not produce concrete topic labels. These need to be decided by the users according to their knowledge on the subject.

With parameters θ and φ, the LDA probabilistic model infers that the whole data text is generated by the following process. First, a word probability distribution is drawn following $\varphi \sim Dir(\beta)$. For each document *d* in the text, topic proportions are drawn following $\theta_d \sim Dir(\alpha)$. For each of the N_d words w_d, a topic assignment is drawn such that $z_{d,n} \sim Mult(\theta_d)$ and each word $w_{d,n}$ is drawn from $p(w_{d,n}|z_{d,n},\varphi)$. In this model, α and β are hyperparameters required for the Gibbs sampling process. The overall likelihood of the whole text with respect to the model parameters is:

$$\prod_{d=1}^{D} P(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}} P(z_{d,n}|\theta_d) P(w_{d,n}|z_{d,n},\varphi) \right)$$

$P(\theta_d|\alpha)$ denotes how likely it is to obtain the topic distribution of θ_d of document *d* conditional on α. $P(z_{d,n}|\theta_d)$ determines how likely the topic assignment $z_{d,n}$ of word *n* in document *d* is conditional on the topic distribution of the document. Finally, $P(w_{d,n}|z_{d,n},\varphi)$ is the probability of having a concrete word conditional on the topic assignment of the word and the word probabilities of the given topics that are in φ. By calculating the sum over all possible topic assignments, the product over all words in a document and the product over all documents in the text, we obtain the likelihood of observing the texts in the documents.

The LDA procedure is based on finding the optimal topic assignment $z_{d,n}$ for every word in each document and the optimal word probabilities φ for each topic that maximizes this likelihood. Maximizing this likelihood would need adding over all possible topic assignment for all words in all documents, which would result in being computationally unfeasible. Thus, alternative methods such as the Gibbs sampler have been developed for this purpose. In this work, such method is used following Griffiths and Steyvers (2004), based on the ldagibbs Stata command introduced by Schwarz (2018).

Gibbs sampling consists of a Markov Chain Monte Carlo (MCMC) algorithm based on repeatedly drawing new samples conditional on all other data. In the case of LDA, the Gibbs sampler relies on iteratively updating the topic assignment of words conditional on the topic assignments of all other words. As Gibbs Sampling is a Bayesian technique, it requires priors for the values of the hyperparameters α and β, which lie within the unit interval. The prior for α is chosen based on the number of topics *T* while the prior for β depends on the size of the vocabulary. The higher the number of topics or the larger the vocabulary, the smaller the priors for α and β will be chosen. In general the choice of the priors will not influence the outcome of the sampling process.

Firstly, the ldagibbs algorithm splits the document into single words or word tokens. These are randomly assigned to one of the *T* topics with equal probability. This gives an initial assignment of words and thereby documents to topics for the sampling process. Later ldaggibs samples new topic assignments for each of the word tokens, with the probability of a word token being assigned to topic *t* being:

$$P(z_{d,n} = t|w_{d,n},\varphi) \propto P(w_{d,n}|z_{d,n} = t,\varphi) \cdot P(z_{d,n} = t)$$

The Gibbs Sampler makes use of the topic assignment of all other tokens in order to acquire approximate values for $P(z_{d,n} = t|w_{d,n},\varphi)$ and $P(z_{d,n} = t)$. $P(w_{d,n}|z_{d,n} = t,\varphi)$ is calculated by the number of of words which are identical to $w_{d,n}$ and assigned to topic *t* divided by the total number of words assigned to that topic.

4. Analysis and Results

We used a total of 80 FOMC meeting transcripts covering all the meetings that took place between 2003 and 2012. A full set of minutes for each FOMC meeting is published three weeks after each regular meeting but complete transcripts are published only five years after the meeting. It is precisely these complete transcripts that we used in our analysis. We introduced the text of the FOMC transcripts into the Stata software database dividing each of the transcripts into data text entries consisting of sentences or paragraphs mentioned by the Governors during the meetings.

A total of 45,346 discussion entries were analyzed covering all the conversations that took place between FOMC members. Staff explanations were not included, putting thus an emphasis on the predominant topics discussed by the Governors during the meetings. The LDA algorithm was then implemented with the goal of splitting the whole text data into 8 distinguishing topics. After a careful analysis of the data texts with highest probability of belonging to each topic, we decided that the topics corresponded to the following themes: Forecasting, Economic Modelling, Statement Language, Risks, Banking, Voting Decisions, Economic Activity and Communication.

The average evolution of the probability of each of the topics being addressed during this period at each of the meetings is graphically shown in Figure 1. Discussions on economic modelling played a major role during the GFC, followed by an increase in the discussion of the banking system in the following years, and in the most recent years discussions on communication have gained relevance. Figure 2 shows the evolution of the number of data entries assigned to each topic. A clear rising upward trend can be detected in the amount of text of the transcripts, showing that FOMC meetings have become more extensive.

Figure 1: Average Topic Share of FOMC Transcripts (2003-2012)

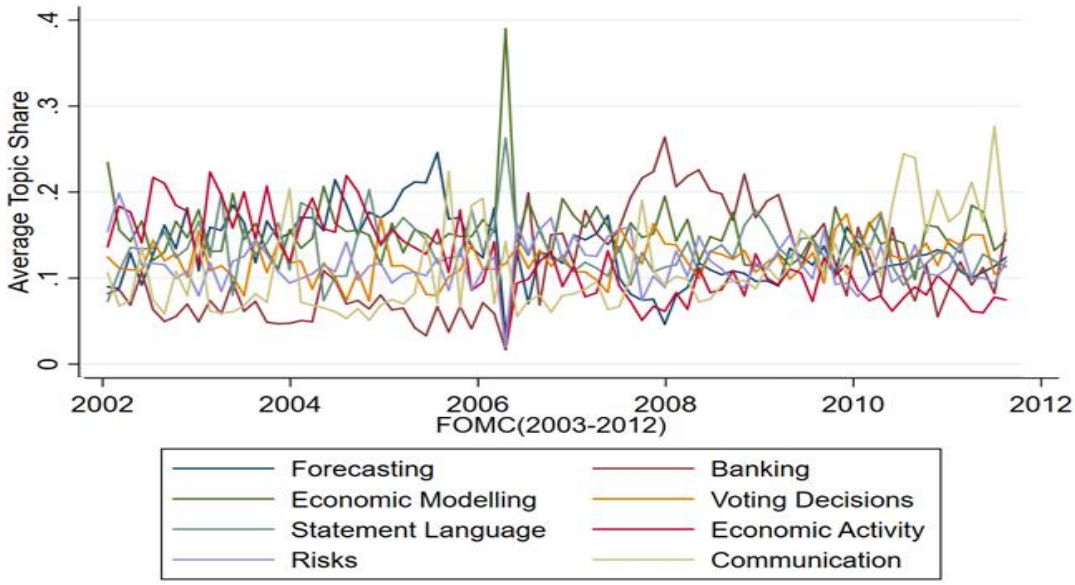
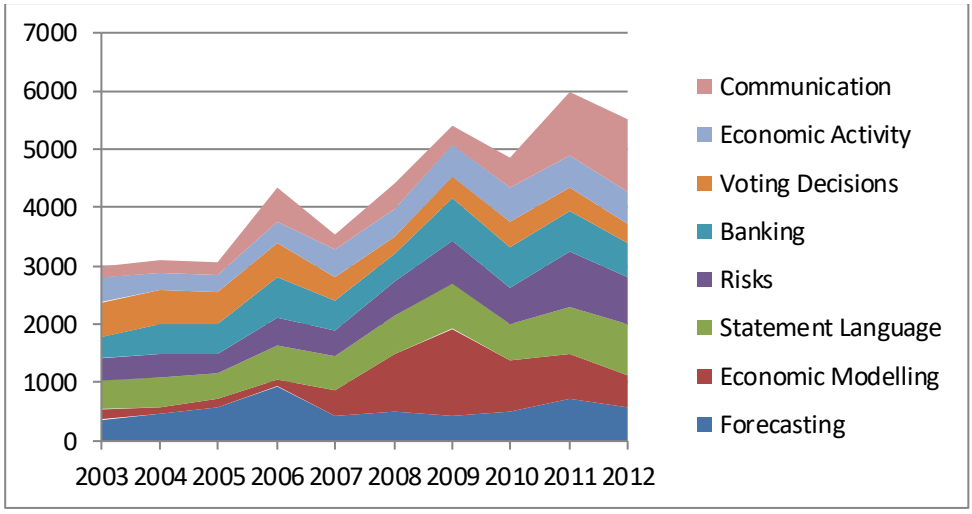


Figure 2: Annual number of data entries (sentences and paragraphs) in FOMC Transcripts (2003-2012)



5. Concluding Comments

The LDA algorithm can be easily implemented to analyze the different themes and their corresponding evolution in terms of use throughout time. In this note we have explained the algorithm, its implementation and estimation and we have provided an empirical example by analyzing the FOMC transcripts covering the meetings that took place during the period 2003-2012.

The use of the LDA algorithm and in particular the Stata command ldagibbs introduced by Schwarz (2018) can be easily implemented to detect which topics are addressed within an abundant number of documents. In this note, we have presented the case of the FOMC transcripts, applying the algorithm to more than 45,000 text data entries and obtaining the evolution of eight identified topics. We observed that discussions on economic modelling played a major role during the GFC, followed by an increase in the discussion of the banking system in the following years, with discussions on communication gaining relevance at the end of the sample. Such type of analysis could be further exploited and employed by researchers at central banks or institutions aiming at determining topic priorities in their official documents.

References

Griffiths, T.L. and M. Steyvers (2004) Finding scientific topics, *Proceedings of the National Academy of Sciences* 101, 5228-5235.

Schwarz, C. (2018) ldagibbs: A Command for Topic Modelling in Stata using Latent Dirichlet Allocation, *The Stata Journal* 18, 1, 101-117.

STATISTICS

Text Data Analysis using Latent Dirichlet Allocation: an Application to FOMC Transcripts

IFC WORKSHOP ON DATA SCIENCE IN CENTRAL BANKING
OCTOBER 21, 2021

Hector Carcel-Villanova
Financial Institutions Division, Statistics Department

The views expressed in this presentation are those of the author and do not necessarily represent the views of the IMF, its executive board or management.

What is Text Data Analysis and its goal

- A useful tool for disentangling and analyzing main topics in different kinds of documents.
- Detect changes in topics and subjects discussed in committees, public institutions, etc.
- Our contribution: a) Explain the LDA algorithm.
  b) Apply it to the analysis of the FOMC transcripts during the GFC.

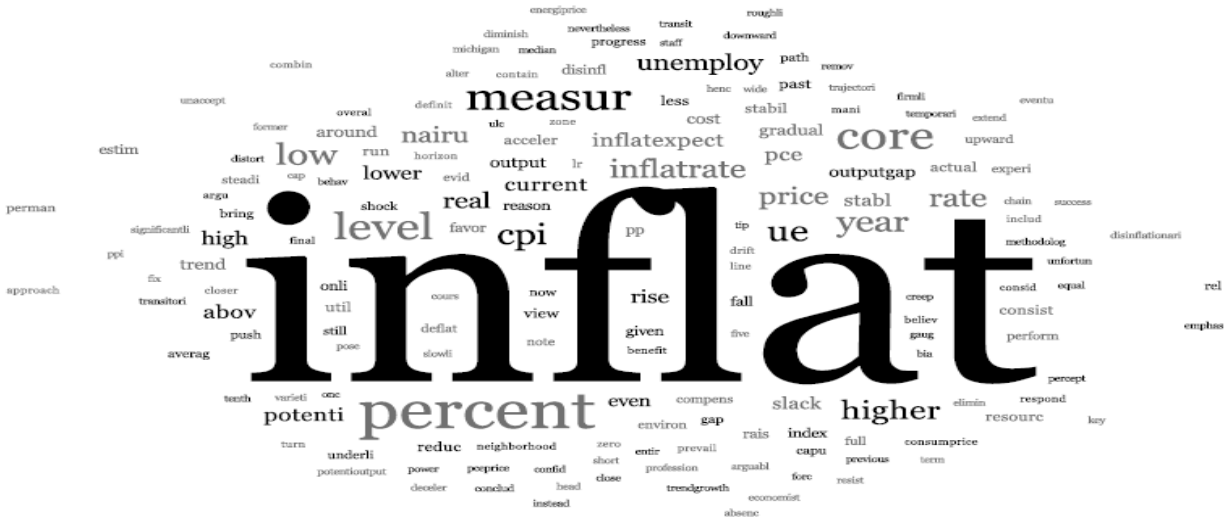


Figure 1: Topic 25—"Inflation"

Methodology

- Each document *d* of the *D* documents in the whole text can be described as a probabilistic combination of *T* topics.

- The outcome of LDA is a $D \times T$ matrix θ containing $P(t_t|d_d)$, with θ_1,.. θ_D being $1 \times T$ vectors, in such a way that the probability of document *d* belonging to topic *t* corresponds to:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix} = \begin{pmatrix} P(t_1|d_1) & \cdots & P(t_T|d_1) \\ \vdots & \ddots & \vdots \\ P(t_1|d_D) & \cdots & P(t_T|d_D) \end{pmatrix}$$

- Every topic $t \in T$ is determined by a probabilistic distribution over the vocabulary (the set of words in all documents) of size *V.*

- The word probability vectors of each of the topics can be represented in a matrix φ of dimensions $V \times T$:

$$\varphi = (\varphi_1 \ldots\ldots, \varphi_T) = \begin{pmatrix} P(w_1|t_1) & \cdots & P(w_1|t_T) \\ \vdots & \ddots & \vdots \\ P(w_v|t_1) & \cdots & P(w_v|t_T) \end{pmatrix}$$

Methodology

- Given the parameters θ and φ, the LDA probabilistic model considers that the whole data text is created by the following procedure:

    1. A word probability distribution is drawn following $\varphi \sim Dir(\beta)$.

    2. For each document *d* in the text, topic proportions are drawn following $\theta_d \sim Dir(\alpha)$.

    3. For each of the N_d words w_d, a topic assignment is drawn such that $z_{d,n} \sim Mult(\theta_d)$ and each word $w_{d,n}$ is drawn from $p(w_{d,n}|z_{d,n}, \varphi)$.

- The likelihood of the whole text with respect to the model parameters is:

$$\prod_{d=1}^{D} P(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_d,n} P(z_{d,n}|\theta_d)P(w_{d,n}|z_{d,n},\varphi)\right)$$

- LDA is based on finding the optimal topic assignment $z_{d,n}$ for each word in each document and the optimal word probabilities φ for each topic that maximizes this likelihood.

Methodology

- This would require adding up all possible topic assignments for all words in all documents, which is computationally impossible.

- Alternative methods such as the Gibbs sampler have been developed for this purpose.

- ldagibbs Stata command introduced by Schwarz (2018).

C. Schwarz (2018) *ldagibbs: A command for Topic Modelling in Stata using Latent Dirichlet Allocation. The Stata Journal 18,1, 101-117.*

Study set-up

- We used a total of 80 FOMC meeting transcripts, covering all the meetings between 2003 and 2012.
- How discussions at the FOMC meetings evolved leading to the GFC, at its height and thereafter.
- Transcripts divided into data text entries consisting of sentences or paragraphs stated by the governors during the meetings.
- A total of 45,346 discussion entries analyzed, covering all the conversations between FOMC members.
- Staff explanations were not included.

LDA Output

Data Text:

41293	One still might argue that because high unemployment is very costly and we are uncertain about the effect of more n	2012	6
41294	Will that be disruptive to markets? We won't know until we face that situation. If our policy is not very effective at in	2012	6
41295	and will be in uncharted waters. That's one reason I strongly urge us to be prudent. To my mind, at this point, costs o	2012	6
41296	MR. FISHER. Mr. Chairman, just as President Lacker was a straight man for President Lockhart, in a way, President Plc	2012	6
41297	President Williams made a very important point. He talked about how uncertainty has paralyzed most businesses and	2012	6
41298	VICE CHAIRMAN DUDLEY. First, I want to make a comment on President Fisher's last remark. My understanding is tha	2012	6
41299	MR. FISHER. At least in the drafts of our statement, we were saying that business fixed investment was weak. The qu	2012	6
41300	VICE CHAIRMAN DUDLEY. My point was that I don't think very many people in the room would debate the point that	2012	6
41301	As far as the outlook is concerned, since the last meeting, I think there's been very little change with respect to the U	2012	6
41302	There are also two other negative developments that I think are really worth highlighting. First, I think the external er	2012	6
41303	basket. Also, as many other people have noted, the risks in early 2013 are tilted to the downside given what's going t	2012	6
41304	So to me, the economic outlook calls for us to do more. Now, I agree that the tools we have are not that powerful. E	2012	6
41305	Which creates the greatest disappointment? Surely the latter. I would be very happy if we did another round of LSAPs	2012	6

LDA Output:

Content	Year	Meeting	FOMC2003201	topic_prob1	topic_prob2	topic_prob3	topic_prob4	topic_prob5	topic_prob6	topic_prob7	topic_prob8
CHAIRMAI	2009	4	52	0.01590909	0.02045455	0.025	0.86590909	0.01590909	0.01590909	0.01590909	0.025
CHAIRMAI	2009	4	52	0.03888889	0.02777778	0.07222222	0.66111111	0.07222222	0.03888889	0.02777778	0.06111111
MR. STOCI	2009	4	52	0.03888889	0.09444444	0.03888889	0.55	0.10555556	0.03888889	0.06111111	0.07222222
As for the	2009	4	52	0.01785714	0.005	0.17928571	0.00928571	0.07785714	0.68214286	0.015	0.01357143
CHAIRMAI	2009	4	52	0.02272727	0.02272727	0.02272727	0.82272727	0.02272727	0.03181818	0.03181818	0.02272727
8 The mat	2009	4	52	0.03571429	0.03571429	0.03571429	0.73571429	0.03571429	0.05	0.03571429	0.03571429
MR. PLOSS	2009	4	52	0.04459459	0.00945946	0.02837838	0.07702703	0.01486486	0.73378378	0.08243243	0.00945946
The most	2009	4	52	0.22905405	0.00608108	0.01959459	0.04391892	0.01283784	0.66013514	0.01148649	0.01689189
The most	2009	4	52	0.46071429	0.03214286	0.01785714	0.01785714	0.03214286	0.28928571	0.11785714	0.03214286

Topic selection

TOPIC 1: Forecasting

-Turning to inflation, I have nudged my forecast for both core and headline PCE inflation down a little since April ...

-When I compare the Board staff's forecast with ours, I find that the Greenbook projection, even the most updated one ...

TOPIC 2: Banking System

-Wells, Goldman, Bank of New York, Sun Trust, and BB&T, for example—opted out. Whenever a fee is assessed on assets or ...

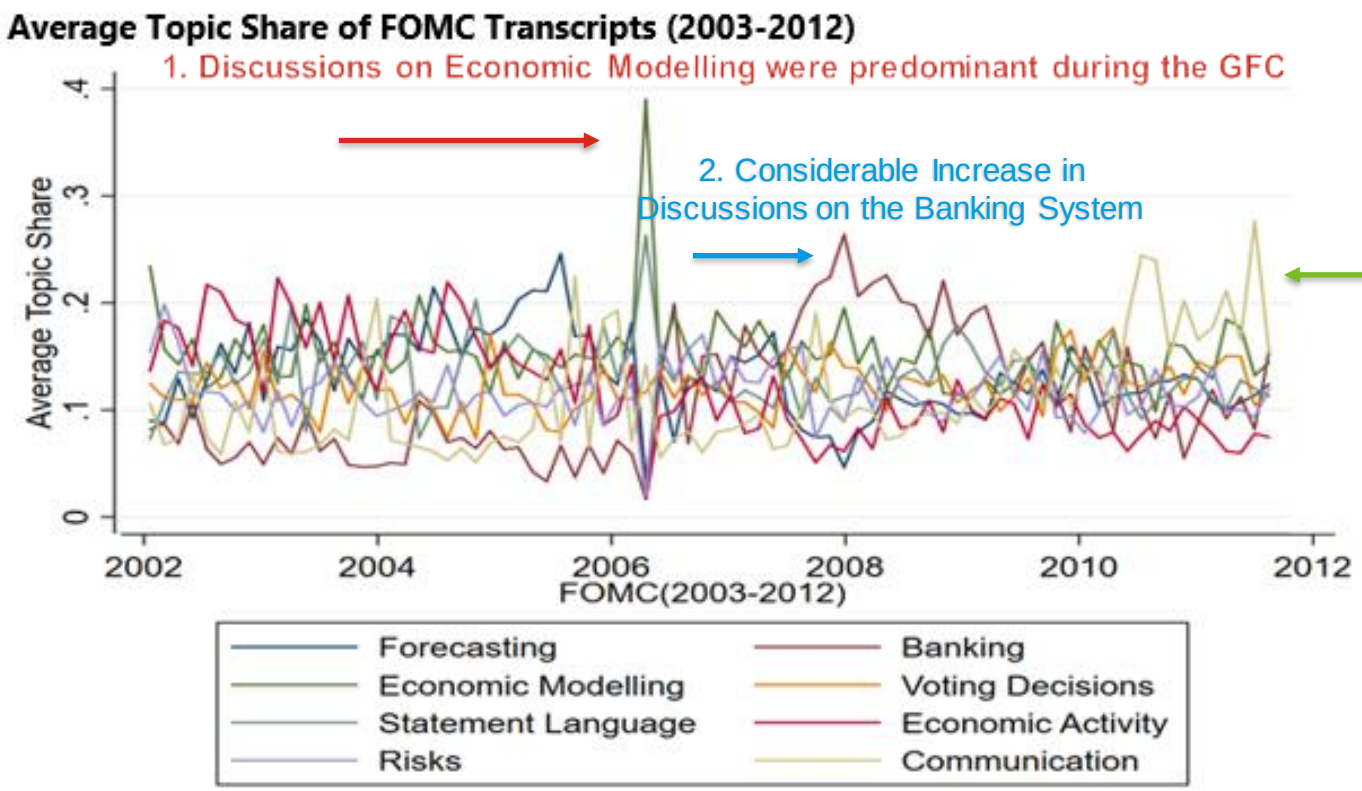
-The first thing is that if we had a floor system, there would be more reserves in the banking system, and that might ac …

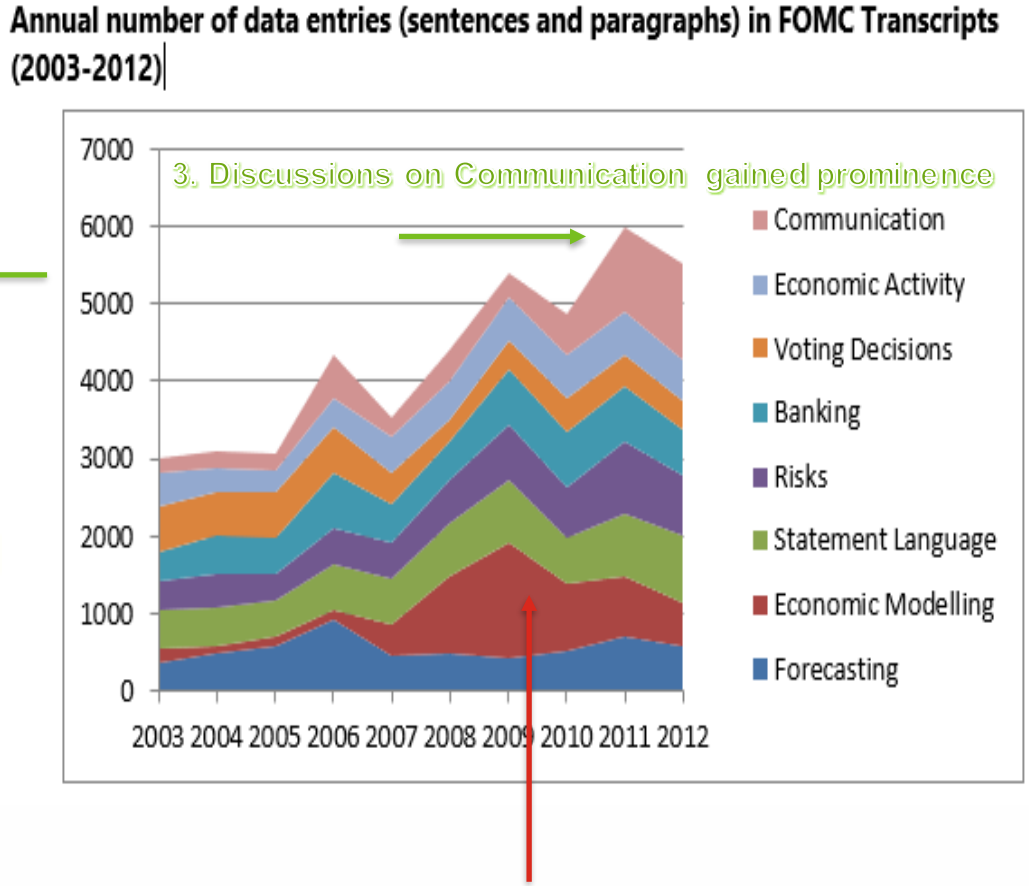
TOPIC 3: Economic Modelling

-Your second question about standard errors is a really good one, and it is hard. There are lots of different models that …

-If you ask whether a DSGE model would tell the story differently from, let's say, FRB/US, the answer is "maybe—it depends ...

Results: FOMC transcripts (2003-2012)





- **LDA could be further used by researchers at central banks and institutions to determine topic priorities in relevant documents.**

- **Future aim: carry out further research to investigate the evolution of concrete economic models (e.g., Phillips curve, Taylor rule, etc.)**

Thank you!