Lost in Discussion? – Tracking Opinion Groups in Complex Political Discussions by the Example of the FOMC Meeting Transcriptions

Cäcilia Zirn, Robert Meusel and Heiner Stuckenschmidt

{caecilia, robert, heiner}@informatik.uni-mannheim.de Data and Web Science Group

University of Mannheim Germany

Abstract

The Federal Open Market Committee (FOMC) is a committee within the central banking system of the US and decides on the target rate. Analyzing the positions of its members is a challenge even for experts with a deep knowledge of the financial domain. In our work, we aim at automatically determining opinion groups in transcriptions of the FOMC discussions. We face two main challenges: first, the positions of the members are more complex as in common opinion mining tasks because they have more dimensions than pro or contra. Second, they cannot be learned as there is no labeled data available. We address the challenge using graph clustering methods to group the members, including the similarity of their speeches as well as agreement and disagreement they show towards each other in discussions. We show that our approach produces stable opinion clusters throughout successive meetings and correlates with positions of speakers on a dove-hawk scale estimated by experts.

1 Introduction

In many discussions, participants can easily be divided into two opposing groups, for example people who support democrats versus people who support republicans, or people who are *pro* or *contra* towards the discussed topic.

Having a closer look at the argumentation why people support or defend something however might reveal various stances even within one such group. People might have different opinions and reasons why they support or oppose something.

Some discussions have a subject so complex that participants cannot be simply divided in a supporting and an opposing group, like the discussion whether abortion should be legal and if yes, up to which status of the pregnancy. In that case the discussants should be grouped by similar positions rather than into *pro* or *contra* partitions.

In this paper, we are analyzing the discussions of the Federal Open Market Committee (FOMC). The FOMC is a committee within the central banking system of the US and decides on the target rate. The committee meetings are not public, however their transcriptions are released five years later. Understanding the several hundred pages long transcriptions requires a deep knowledge of the financial ecosystem. Apparently even for experts the analysis of those documents is intricate and time-consuming.

Our goal is to develop a robust approach to reveal the opinion groups present in discussions where positions are complex to detect and the content is difficult to understand for non-experts. Furthermore, we want to assist human readers with a fast automatic system to avoid reading those immense amounts of text.

There are two major reasons that make it difficult to directly learn a model for the different opinion groups hidden in the discussions. First, the data is not labeled. This is mainly due to the small number of people having sufficient knowledge of this particular domain. In order to overcome this issue, the votes at the end of each meeting, where the discussion members finally decide on the target rate, seem to be a valid starting point to serve as labels. Those votes, however, do not reveal the position of the individual speakers, as they agree on consenting votes. Thus the voting records cannot serve to learn the opinions of the committee's members. The second issue is that topics discussed in the meetings might vary. To address those issues, we choose to cluster opinion groups in each discussion dynamically, using an unsupervised approach.

There are some experts analyzing the FOMC's

members and discussions. They usually place the discussants on a scale between *doves* and *hawks*. Doves aim at higher employment, whereas hawks focus on a low inflation rate. However, this classification might not be appropriate to capture opinion groups: although two people might tend to behave rather hawkish, they still can have different views on how the discussed problem should be solved. We also have to keep in mind that political positions are not limited to one dimension only, but span over several ones, like left-right or liberal-conservative, to mention only a few.

To find opinion groups, we focus on two things: First, we compare the terms used to express a position among the speakers. According to political science, if speakers use the same terms, they share a similar position, as described by Laver et al. (Laver et al., 2003) and also Slapin et al. (Slapin and Proksch, 2008), among others. We will hence analyze the pairwise overlap of the speakers' vocabulary. Second, we investigate how they address each other throughout the discussion. Do speakers agree to their predecessor? Do they disagree and argue against each other's arguments?

In the rest of the paper, we will present our method to cluster positions of speakers in complex political discussions when neither labels are provided nor the underlying opinion groups are known in advance.

2 FOMC Data

The FOMC is a committee within the central banking system of the United States and decides on the target rate. The committee consists of members of the Federal Reserve Board and Federal Reserve Bank presidents. Twelve of the members have voting rights, while the rest is only allowed to attend and participate in the discussions. The meetings are non-public and the members know each other well, which allows an open and direct dialogue. As described by Havrilesky and Gildea in (Havrilesky and Gildea, 1991) and also by Adolph in (Adolph, 2013), the committee decides with consenting votes – dissenting votes appear rarely, although the members do have different goals and positions.

The transcriptions of the meetings are only released after five years and comprise several hundred pages in PDF format. In our work, we analyze the transcriptions of the FOMC meetings between 2005 and 2008. This includes 41 meetings, each containing 43 600 words and 24 speakers on average. The total number of different speakers for the selected meetings is 96.

FOMC members are considered to act *dovish* or *hawkish*. Doves aim at higher employment, whereas hawks focus on a low inflation. Domain experts thus classify the FOMC members into *doves, moderate doves, centers, moderate hawks* and *hawks*. We were able to retrieve this classification for 19 members only, as we could not find information dating back earlier than 2009. We collected estimations from various sources we found on the Web¹

3 Distinguishing Statements from Discussion Elements

Browsing through the transcriptions, we figured out that there are two types of contributions in the following called turns - to the discussion. In the first type of turns, the speakers elaborate on their opinion. Presumably they have prepared their argumentation in advance. Following those statements, other speakers ask questions or comment on the speaker's statement; discussions might arise. The contributions to those discussions are shorter and seem to be of a more spontaneous nature. We consider those turns as the second type. We think that the content of those two types of turns - statements and discussion elements - need to be analyzed with different techniques. Statements are prepared and reflect the general position of the speaker. According to research in political science, the political position of a speaker is determined by the topics he speaks about (Grimmer and Stewart, 2013; Hillard et al., 2008; Laver et al., 2003). The speaker will expand on the topics he considers important.

The shorter **discussion elements** are spontaneous reactions to the previous statement. They contain an attitude towards previous turns: the

editorial/outside-the-box-musical-chairs\
\-at-the-fomc/

¹http://graphics.thomsonreuters.com/F/ 10/US_HAWKOMETER1010.html,

http://graphics.thomsonreuters.com/F/10/
scale.swf

http://cib.natixis.com/flushdoc.aspx?id=
54743,

http://www.mauldineconomics.com/

http://www.ritholtz.com/blog/2009/11/
fed_bawks_vs_doves/

http://blogs.wsj.com/economics/2010/09/

^{30/}balancing-the-feds-hawks-doves/



Figure 1: Manually annotated discourse contributions and their lengths (word count).

speaker often expresses agreement or disagreement, as in "*I can see why you assume that, but* …" or "*To be honest, I don't think* …".

We manually annotated one meeting, classifying each discourse contribution either as statement or as discussion element. The sequence of those contributions and their word length together with their assigned class is shown in Figure 1.

From the diagram, we can see that the threshold between the statements and the discussion elements is around five hundred words. We use this number as a shallow heuristic to automatically classify the discourse contributions into statements and discussion elements. Using this straightforward approach, we correctly label 98% of the speaker turns.

3.1 Analyzing Statements

As mentioned before, the positions are expressed through the topics mentioned in the speeches, which are mainly determined by nouns. We conclude that if two speakers share similar views, they are likely to use the same vocabulary. Therefore, we access the closeness of speakers by calculating the similarities between their speeches.

As observed in Section 3, the speakers' positions are represented by the longer statements rather then the short discussion elements, so we only use the former to compare positions. In the spontaneous discussion elements, speakers tend to repeat the vocabulary of their previous speakers, for example by phrases like "*I do not agree with your view on unemployment.*", which would influence our similarity calculation. In natural language, topics are mainly determined by nouns. So we keep nouns only, lemmatize them and represent every speaker for each meeting as a word vector. Then we pairwise compare the vectors of each meeting using cosine similarity.

As we do not have a gold standard to evaluate the similarity calculation, we investigate whether the similarity between two speakers is pertained across all meetings. Two speakers having close positions in one meeting should have close ones in further meetings, too, as they are not likely to change their position while being on the committee. For each speaker pair, we calculate the standard deviation of the similarities across all meetings they both attended. It ranges from 0 to 0.37 (0.08 on average). For two thirds of the speaker pairs the standard deviation is below 0.1. Hence, this approach can be considered as being very robust.

To evaluate our hypothesis that the longer statements are more relevant for determining the speakers' positions, we compare the above described results to the similarities calculated using all utterances of a speaker including spontaneous discussion elements. The standard deviations range up to 0.46 with an average of 0.1. For better comparability, we plotted the standard deviations for all meetings of both experiments in Figure 2 sorted in descending order. We can clearly see that the standard deviations for the similarities calculated using statements only is continuously below the standard deviations based on both utterance types.



Figure 2: Comparison of the standard deviations of similarities for each speaker pair sorted in descending order calculated on the long statements only vs. calculated on both statements and discussion elements.

3.2 Analyzing Discussion Elements

While we used statements for similarity analyses, we are interested in agreement and disagreement among the speakers within discussion elements. In (Misra and Walker, 2013), Misra and Walker analyze disagreement and rejection in dialog. They generate a set of cue phrases like has always been, you don't understand or yeah, correct to classify types of agreement and disagreement and achieve 66% accuracy. We use their cue phrases to detect (dis)agreement within discussion elements. Whenever we find a cue of (dis)agreement within a turn, we consider this as a (dis)agreement of the speaker with his predecessor. We have to consider one special case: discussions are moderated by a Chairman. He gives the speakers the floor, like in: "Other questions for Mr. Kos? President Minehan." or "Thank you. President Moskow.". So if the predecessor of a turn is the Chairman, the (dis)agreement might actually be towards the Chairman's predecessor. Considering the Chairman only as the moderator is not quite appropriate, however, as he is also a participating member of the committee and thus representing his own position, too. To find the correct predecessor of a disagreement statement, we therefore have to distinguish between a call for the next speaker or the Chairman's personal contribution. We use a simple heuristic: if the Chairman mentions the following speaker's name, he is considered as moderator. We then treat his predecessor as the aim of the (dis)agreement and ignore the Chairman's turn.

4 Clustering Opinion Groups

In Subsection 3.1, we explained how we calculated the similarity between two speakers' positions based on their statements. In Subsection 3.2, we described how we detect agreement and disagreement among the speakers. To determine opinion groups in the FOMC discussions, we make use of both properties. The interactions and similarities describe the relations between the speakers. Hence it seems reasonable to model the speakers as nodes in a graph with their relations constituting the edges.

4.1 Graph Clustering

Blondel et al. (Blondel et al., 2008) introduced a novel fast and efficient community detection method for large graphs – called *Louvain Clustering* – which outperforms existing community detection methods. This method is based on optimization of the so called modularity of a network as described by Newman (Newman, 2006). The modularity of a graph or network is a measure of its structure and measures the degree of division of the network into clusters. Networks with a high modularity haven dense clusters with a minimal number of links between the clusters. The method of Blondel et al. works in a two step approach. Within the first phase all nodes are assigned to different communities and the possible gain of modularity is calculated for each node under the premise that it is removed from its own community and assigned to the community of one of its neighbors. Then, the community with the maximal, positive gain is chosen. The phase stops, when a local maximum is reached an no node can be assigned to another community to increase the modularity. In the second phase, a new network is built where a node represents a single community of the original network after phase one. The weights of the links between the new nodes are calculated by summing up all existing weights of links of old nodes between those two communities. After phase two has finished it is possible to reapply phase one until the network does not change any more.

An alternative for community detection is the *VOS Clustering* introduced by Waltman et al. (Waltman et al., 2010). This technique combines VOS mapping with a weighted and parametrized variant of the modularity function of Newman and Girvan (Girvan and Newman, 2002).

4.2 Graph Construction Methodology

We cluster the discussants of every FOMC meeting between 2006 and 2008. For each meeting, we create one graph with the speakers constituting the vertices and the relations between them constituting the edges.

Similarity. Similarity is modeled as undirected edges between two speakers s_1 and s_2 using their cosine similarity, normalized between -1 and 1:

$$sim(s_1, s_2) = norm_{-1,1}(cos(s_1, s_2))$$
 (1)

Agreement / Disagreement. Agreement and disagreement are in the first place directed relations: one speaker (dis)agrees with his predecessor. However, we can make the assumption that if a speaker disagrees with his predecessor, the predecessor also disagrees with him. For this reason, we will experiment with directed and undirected (dis)agreement. In order to measure the agreement or disagreement we first count the number

of agreements $c_{ag,dir}(s_1, s_2)$ and disagreements $c_{disag,dir}(s_1, s_2)$ of a speaker s_1 towards his predecessor s_2 . For the undirected case this count is calculated as shown in the following two equations:

$$c_{ag,undir}(s_1, s_2) = c_{ag,undir}(s_2, s_1) = c_{ag,dir}(s_1, s_2) + c_{ag,dir}(s_2, s_1)$$
(2)

$$c_{dis,undir}(s_1, s_2) = c_{dis,undir}(s_2, s_1)$$

$$= c_{dis,dir}(s_1, s_2) + c_{dis,dir}(s_2, s_1)$$
 (3)

We than flatten the total counts by using their square roots which we further scale between 0 and 1 as formalized in the following two equations:

$$ag(s_1, s_2) = norm_{0,1}(\sqrt{c_{ag}(s_1, s_2)})$$
 (4)

$$dis(s_1, s_2) = norm_{0,1}(\sqrt{c_{dis}(s_1, s_2)})$$
 (5)

We merge agreement and disagreement between speakers by subtracting disagreement from their agreement:

$$agDis(s_1, s_2) = ag(s_1, s_2) - dis(s_1, s_2)$$
 (6)

This results in agDis being scaled between -1 and 1.

4.3 Experiments

We assess the quality of our results in two ways. First, we want to track the robustness of our clusters. We expect opinion groups in one meeting to be retained in the next meeting, as the topics of the meetings are not supposed to have changed completely, neither should the opinions of a speaker have changed so fast. By pairwise comparing the clusters of one meeting to the clusters of the following one, we use the *Rand index ri* introduced by Rand in (Rand, 1971). It is a measure for the similarity between two clusterings of a set of elements, in our case the speakers:

$$ri = \frac{a+b}{a+b+c+d} \tag{7}$$

where *a* refers to the amount of speaker pairs being within the same cluster in both meetings (*true positives*), *b* refers to the amount of speaker pairs belonging to different clusters in both meetings (*true negatives*), *c* refers to the amount of speaker pairs who belong to the same cluster in the first meeting, but not in the second (*false positives*), and *d* refers to the amount of speakers who belong to different clusters in the first meeting, but

Edges	Rand index
Similarity	0.634
(Dis-)Agreem. (dir.)	0.701
(Dis-)Agreem. (undir.)	0.742
Similarity + (Dis-)Agreem. (undir.)	0.625

Table 1: Louvain Clustering

Edges	Rand index
Similarity	0.621
(Dis-)Agreem. (dir.)	0.783
(Dis-)Agreem. (undir.)	0.839
Similarity + (Dis-)Agreem. (undir.)	0.651

Table 2: VOS Clustering

to the same cluster in the second meeting (*false negatives*). The Rand index can be interpreted as the accuracy of the clustering.

The list of average Rand indexes comparing all pairs of successive meetings is shown in Table 1 (applying Louvain Clustering) and in Table 2 (applying VOS Clustering). Clustering speakers based on similarity edges only, both algorithms reach a Rand index of about 0.6. The results are more stable throughout the meetings when clustering based on the directed (dis-)agreement relations only (0.7 for Louvain, 0.78 for VOS) and even improve using undirected (dis-)agreement, achieving a Rand index of 0.74 for Louvain and 0.84 for VOS. If we combine both edge types, we do not gain stability: With 0.62 and 0.65 respectively the results are worse than using one of the edges types only. It is remarkably however that both edge types being based on completely independent dialog parts and approaches still achieve comparable performance.

In a second experiment we want to verify whether our hypothesis holds that speakers in the same opinion group should have a similar position on the dove-hawk scale. We compare the opinion group clusters to the clustering of the speakers given their dove-hawk labels (*dove, moderate dove, center, moderate hawk, hawk*) calculating the Rand index. The results for Louvain are shown in Table 3, for VOS in Table 4. The results range between 0.62 and 0.77. Like in the pairwise meeting comparison, there is only little difference between directed and undirected (dis-)agreement, the differences spanning from 0.001 to 0.05 only. Again, using one type of edges only outperforms their combination. Instead of increasing perfor-

Edges	Rand index
Similarity	0.711
(Dis-)Agreem. (dir.)	0.651
(Dis-)Agreem. (undir.)	0.656
Similarity + (Dis-)Agreem. (undir.)	0.628

Table 3: Louvain Clustering

Edges	Rand index
Similarity	0.666
(Dis-)Agreem. (dir.)	0.743
(Dis-)Agreem. (undir.)	0.768
Similarity + (Dis-)Agreem. (undir.)	0.675

Table 4: VOS Clustering

mance, we receive the average performance of both information sources – the algorithm seems to suffer from contradictory information. We will further investigate how to combine information sources in an appropriate way. In general, the results show that the dove-hawk positions are correlated with the opinion groups we derive.

5 Related Work

Common approaches for position analysis in political science scale texts based on word frequencies and co-occurrences as described by Grimmer (Grimmer, 2010), by Quinn et al. (Quinn et al., 2010), and by Gerrish and Blei (Gerrish and Blei, 2011). Approaches developed in the field of computational linguistics usually classify speakers or texts as pro and contra towards discussed topic. Anand et al. (Anand et al., 2011), Somasundaran and Wiebe (Somasundaran and Wiebe, 2009), and Walker et al. (Walker et al., 2012) all classify stance in on-line debates. While Anand et al. use a supervised learning approach, Somasundaran and Wiebe mine opinions and opinion targets from the web. Then, they combine the thereby learned stance with discourse information formulating an Integer Linear Programming problem. The approach of Walker et al. makes use of same author links and rebuttal links to model posts as a graph, cutting it into two parts (pro and contra) with MaxCut. These methods are hardly applicable to our complex discussion data for reasons we elaborated in Section 1.

A similar idea to our approach is described by Thomas et al. (Thomas et al., 2006). Their goal is to label congressional floor-debate speeches as supporting or opposing the discussed topic. In contrast to our approach, where speakers are the nodes, they model speech turns as nodes connected by *same label* relations. They then find minimum cuts in the resulting graph.

Abu-Jbara et al. (Abu-Jbara et al., 2012) explore the dialog structure in on-line debates with the goal of subgroup detection. They represent each discussion participant as a vector consisting of the polarity and the target of their opinionated phrases, combining it with the information about who replies to whom. In a final step, they cluster the vectors. They point out that the reply feature needs further investigation since they cannot tell whether speakers tend to agree or disagree when they answer each other.

6 Conclusion

We presented a completely unsupervised approach to cluster opinion groups in the complex political discussions of the FOMC using two independent types of information. On the one side, we made use of the similarity between the speakers' statements, on the other hand we integrated their behavior towards each other within discussions. For this, we detected agreement and disagreement using cue phrases. Both types of information turned out to be comparably useful for clustering the speakers. Our simple strategy to distinguish between statements and discussion elements - the two sources of information - is straightforward and effective. We showed that the results are stable throughout successive meetings and correlate with the dove-hawk positions for speakers estimated by experts.

Regarding further challenges, we have to investigate how we can improve the combination of various information sources, e.g. by weighting them. In addition, we plan to add further sources like political party adherence, background of a speaker or their function in the FOMC, such as member of the Federal Reserve Board or Federal Reserve Bank president.

References

- Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 399–409, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Christopher Adolph. 2013. Bankers, Bureaucrats, and Central Bank Politics: The Myth of Neutrality. Cambridge University Press.
- Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Sean Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 489–496.
- Michelle Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June.
- Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*.
- Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.
- Thomas Havrilesky and John Gildea. 1991. The policy preferences of fomc members as revealed by dissenting votes: Comment. *Journal of Money, Credit and Banking*, 1:130–138.
- Dustin Hillard, Stephen Purpura, and John Wilkerson. 2008. Computer assisted topic classification for mixed methods social science research. *Journal of Information Technology and Politics*.
- Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02):311–331.

- Amita Misra and Marilyn A. Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Conference of the Special Interest Group on Discourse and Dialogue*, page 920.
- M. E. J. Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.
- Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228, January.
- W.M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.
- Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, July.
- Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 Volume 1, ACL '09, pages 226–234, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 327–335, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the* 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 592–596. Association for Computational Linguistics.
- Ludo Waltman, Nees Jan van Eck, and Ed C.M. Noyons. 2010. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4):629 – 635.