Analyzing Senator Community Structure from 2012-2015 Roll Call Data

Weiwei Li[1], **Zhengling Qi**[1], **Yuzhou Sun**[2], **Wenting Hu**[2]
[1]Department of Statistics and Operations Research, the University of North Carolina, Chapel Hill, NC 27708
[2]Departments of Electrical and Computer Engineering, Duke University, Durham, NC 27708

Abstract

In this project, we analyzed the 111th-114th roll calls in the US Senate (2012-2015), by applying unsupervised learning methods to assess the association between pairs of senators based on the votes they cast. For pairs, we used similarity-based methods, including hierarchical clustering and multidimensional scaling. We also applied information entropy to find the most influential senators during these periods. By doing so, we observed the gap between Democratic Party and Republican Party increases with time. The senators? goal and behavior is more and more consistent to their representative party.

1 Introduction

Roll call data is popular in politics that is well - amenable to detect community structure among senators by data mining techniques. In our research, we downloaded roll call voting results from voteview.com, which is supported by researchers from Berkeley, Princeton and San Diego. For each roll call, the txt file provides a list of votes cast by each of the senators. For each of those, the vote of every senator is recorded in three ways: 'Yea', 'Nay' and 'Not Voting'.

Throughout our project, we are particularly interested in digging out time-varying community structures (aside from the obvious segmentation via Democracy party and Republican Party) behind senators from 2012-2015, which can better describe the voting behaviors of senators.

Basically, we use Rajski's distance[1] to evaluate pairs of senator's distance which enjoys the universal metric property. Based on the distance matrix, we first make use of hierarchical clustering without using prior knowledge including party information. In addition, multi-dimensional scaling was conducted to visualize all the senators in Euclidean space, which helps us to further understand the gap between senators. After analyzing the overall relationship between senators, we focus on senators individual influence on voting results by three criterions such as agreement ratio and information share rate.

Rajski's distance is widely used in information science, particular for categorical data since entropy estimation for continuous variable is hard to approximation. Rajski's distance enjoys universal metric property, which means that when two points are closed in other non-trivial metric distance, they will be also close under Rajski's distance. Multidimensional scaling is a technique of visualizing the level of similarity of objects with high dimensional features. It refers to a set of related coordination techniques used in information science by using distance matrix. In our project, we apply metric measure to multidimensional scaling instead of non-metric measure.

In Section 2 we define Rajski's distance and its property. In section3, we perform hierarchical clustering and 2-D multidimensional scaling to detect hidden community structure behind senators. Influential analysis is used to quantify senators' individual impact on issues in Section 4.

2 Rajski's Distance

2.1 Definition of Rajski's Distance

Since voting result for every senator is a binary vector, we use the information-theoretic Rajski's distance to measure all pairs of senators. It is a metricized version of mutual information which obeys the triangle inequality and other requirement for a metric. Obviously, the larger the Rajski's distance, the more separate the two senators. The Rajski's distance is defined as:

$$d(X,Y) = 1 - \frac{I(X;Y)}{HX,Y}$$

where $H(X,Y)$ is the joint entropy of two random variable X and Y.
The mutual information and joint entropy are define as:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$H(X;Y) = -\sum_{x}\sum_{y} P(x,y)\log_2 P(x,y)$$

The assumption behind Rajski's distance is that each senator is regarded as random variable. Besides, by using mutual and joint entropy, independent assumption is made on each issue vote by senators. This strong assumption may be replaced by weakly dependency in future work.

2.2 Displaying Dissimilarity Matrices

Rajski's distance as plain numbers provides little insight. However, we can provide the distances between all pairs of senators in the form of a graphical matrix. The symmetric dissimilarity matrix graphically illustrates the information-theoretic Rajski's distance between all pairs of senators, based on their votes. If two senators voted similarly, the square at the intersection of their names is dark. As we can see from the graph, two large clusters can be identified visually from this graph, and one

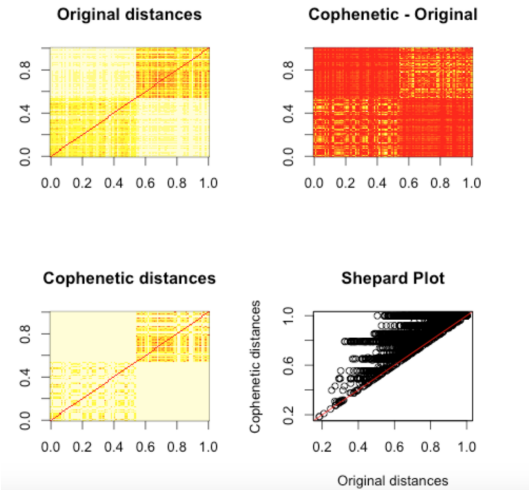

Figure 1: Dissimilarity Matrices for 111th-114th Senators

group of moderate senators in each party. The major clusters correspond to the political parties even if the party information was not used in the computation of distance.

3 Clustering Analysis

3.1 Hierarchical Clustering

In the hierarchical agglomerative clustering[2] of senators based on their pair-wise Rajski's distance, we can identify the two major clusters: the Republican(green) and the Democratic(red), senators with no party affiliations are marked as blue. An interesting fact is, independent senators seem to have the same preference with Democracy side.

From Figure 2, we observed that as time goes from 2012-2015, the boundary between two parties becomes more and more clear. For example, during 2012-2014, we can still find some senators votes opposite to their belong parties, like Mark Kirk from Illinois in 2012.

Figure 2: Hierarchical Clustering for 111th-114th Senators

3.2 Multi-Dimensional Scaling[3]

To distinguish similarity, which is in our case quantified information-theoretically with a probabilistic model, from distance that is defined geometrically.

If each senator is denoted with a point in some k-dimensional space, we can try to place these points so that the Euclidean distances between the points would match the distances as specified by the

dissimilarity matrix. By using 2-D multi-dimensional scaling, we could visualize the senators in 2-D coordinate(Figure 3).

The senators whose votes were similar, also appear close in the resulting diagrams. From 111th to 114th 2-D multi-dimensional scaling, we again validate our conjecture that the gap between two parties is constantly becoming larger.

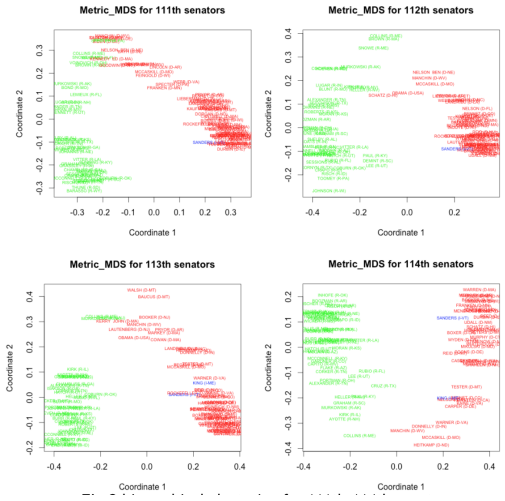

Figure 3: Multi-Dimensional Scaling for 111th-114th Senators

4 Influential Analysis

4.1 Individual Senator Influence Analysis

We may define influence as the similarity between a vote cast and the outcome. We can interpret the process of voting under the framework of information theory[1]. Each senator can be considered as an information source, while the outcome of the vote is the destination. Therefore, mutual information can be interpreted as a measure of how influential is a senator with respect to the outcome of the vote.

Table 1: Top 5 Most Influential Senators

	I(X;Y)/H(Y)	I(X;Y)/H(X,Y)	Agreement	NotVotingP
MURKOWSKI (R-AK)	0.497598664894792	0.273265644364143	0.855263157894737	0.0427631578947368
COCHRAN (R-MS)	0.419582376565703	0.239057683547413	0.855263157894737	0.00986842105263158
CAPITO (R-WV)	0.419130425638844	0.247081733551818	0.868421052631579	0.00657894736842105
ALEXANDER (R-TN)	0.41438661703442	0.220710476818469	0.835526315789474	0.0328947368421053
AYOTTE (R-NH)	0.40000077106231	0.246194491337452	0.878289473684211	0.00328947368421053

In order to find the influential senators, we use three criterion to evaluate their voting impact. The first one is the percentage of outcome entropy eliminated by the senator's vote, the second one is the percentage of mutual information between outcome and vote eliminated by senator's vote and the last one is the percentage of issues when the vote and the outcome matched.

Interestingly, from above table, top 5 most influential senators are all from Republican Party. Overall, aggressive assertion could be made that republican senators are more influential than democratic senators. However this criterion should be further investigated based on such as graphical model.

5 Conclusion

Roll call data represents senators' political ideologies, which can be used to detect senators' community structure. Therefore, we have investigated the 111th-114th Senate from both a global perspective viewing pairs of senators and viewing voting a local pair-wise perspective within the Senate. That senators from the same party tend to vote similarly and the senators from Republican Party are more influential. In the meantime, we observe a time-varying community structure during the 2012-2015. Basically, the gap between two main parties are becoming larger and larger both from hierarchical clustering and 2-D multidimensional scaling. We found that data analysis methods developed for natural and social sciences were useful also in political science.
Our future work could combine voting results and bill questions via topic modeling methods, which could further detect the voting behavior or ideology behind senators.

Acknowledgments

We would like to thank our supervisor, Prof. Sayan Mukherjee, Duke University, for support and tutoring us to work on analyzing senators' roll call data. We also thank the the Voteview Website (http://voteview.com/) providing the data.

References

[1] A. Jakulin, W. Buntine, T. M. La Pira and H. Brasher *Analyzing the U.S. Senate in 2003: Similarities, Clusters, and Blocs Political Analysis* 2009 17 (3): 291-310.

[2] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[3] J. de Leeuw. *Applications of convex analysis to multidimensional scaling*. In J. R. Barra, F. Brodeau, G. Romier, and B. van Cutsem, editors, Recent developments in statistics, pages 133-145, Amsterdam, The Netherlands, 1977. North Holland Publishing Company.

Appendix

Person	**Matr. No.**
Hans Maier	12345
Anna Huber	23456
Werner Weisbaeck	34567

114th influential senators.csv

Senetor	**I(X;Y)/H(Y)**	**I(X;Y)/H(X,Y)**	**Agreement**	**NotVotingP**