Math^{Industry} Workshop

ANALYZING LEGISLATORS AND POLICY AREAS

DANA BERMAN, JILLIAN GLASSETT, DANYI LIU, DIAAELDIN TAHA, AND AARON (XIANG) ZHENG

Abstract. We used official legislative data on the Canadian and United States governments to see if politicians cluster around policy topics. Both data sets provided different information on legislators and legislation, resulting in two distinct methodologies and results. For the Canadian data set, we identified the distribution of times spoken in parliament per topic and visualized how often the topic came up in session from 2017 to 2019 using t-SNE. We saw that some topics were evenly spread over the three years while other topics clustered around a particular year. We provided insight into how this data can be used to analyze legislative performance. For the U.S. data set, we examined the percentage of bills that became law as a measure of political performance for current legislators. We also used various clustering techniques to analyzed politician's interest in different policy areas based on (co-) sponsored bills. We found that these groupings bear strong connections to both parties and location. We also performed some preliminary analysis on legislation passing rates by topic.

1. Introduction

Performance measurement in the the field of sports has achieved significant growth and development. Using data-based performance evaluation to make key decisions has played an important role in sports such as football, basketball, and ice hockey[1]. Analysis of complex aspects of the game requires a comprehensive mathematical tool to facilitate a continuous cycle of "question and answer" to provide detailed and flexible explanations.

What if we applied the same mathematical tools and concepts to political performance? IOTO International Inc. is a non-partisan analysis company that specializes in using AI to gain insights from political data. Through the PIMS Math^{Industry} workshop, we were given the opportunity to collaborate with this company. The IOTO team provided us with some data sets mined from official government websites. This report summarizes the work, which occurred over a two-week period.

Out of respect and confidentiality considerations, we will not provide the data sets nor the code used.

2. Problem Statement

IOTO provided us with two main data sets, one from Canada and one from the U.S. After cleaning up the data and conducting an initial analysis, we found that the the data sets were too different to make a direct comparison between the two countries a worthy endeavor. The Canadian data set was based on debate records on the floor of Parliament from 2017

[1]"Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer", Javier Fernández, Luke Bornn, Dan Cervone, `http://www.lukebornn.com/papers/fernandezsloan2019.pdf`

through 2019. Beyond basic information on the legislators, the data set contained the date, time, duration and topic that each legislator spoke on; it did not contain any legislative information. The U.S. data set contains the total bills sponsored or co-sponsored by each legislator; it also contains some basic information on each legislator. For each legislator, the record covered the legislator's entire congressional career, and was separated into thirty-two policy topics. Since the both data sets had information on policy topics, we focused on the following question:

Question. *Do politicians cluster around certain topics?*

In addition to this question, we performed a preliminary analysis of the percentage of (U.S.) bills that pass, and on the passing rate of legislator clusters. This was possible after the IOTO team mined some additional data on bill status. Given that we are studying the Canada and U.S. separately, we will split our report into two sections, one per country. Each section will detail our methodology and results. We will summarize all the results at the end.

3. Canada

Hansard is the name of the official reports of the Parliamentary debates in Britain and several other Commonwealth countries including Canada. It is named after the Hansards, a family of printers who worked with the Parliament at Westminster in the late 18th century.[2]

We received access to a data set that is derived from the Canadian Parliamentary Hansard speech data and were tasked with analysing the topics that were discussed in the debates. The data set included such information as MPs names, party affiliation, and timestamped summaries of the topics the MPs discussed on the floor of the Canadian Parliament. We restricted our analysis to the years 2017, 2018, and 2019. Our analysis involved 88,235 data points covering around 7,000 topics.

3.1. Motivating Questions.

For the sake of brevity, we showcase our work for only two questions:

(1) Is it possible to detect activity in the Parliament? Is it possible to tell when a particular subject is becoming a popular debate topic?
(2) Can we tell which topics are usually discussed with other topics or on their own?

3.2. Methodology and Results.

3.2.1. *Detecting activity and popularity.* We used the number of times an event of interest occurs in the Canadian Parliament during a specified time window as a sign of activity. This is a very natural indicator of activity that is prevalent throughout both the scientific and the popular literature. For instance, a sharp increase in a location of the number of cases confirmed to have contracted a particular disease is a sign that an outbreak might be taking place.

We created two plots showing the spikes in debates from 2017 to 2019 in the Canadian Parliament. Both plots are of a 30-day moving average to reduce the amount of noise. The

[2] *Encyclopaedia Britannica*, "Hansard," (accessed September 01, 2020), `https://www.britannica.com/topic/Hansard`

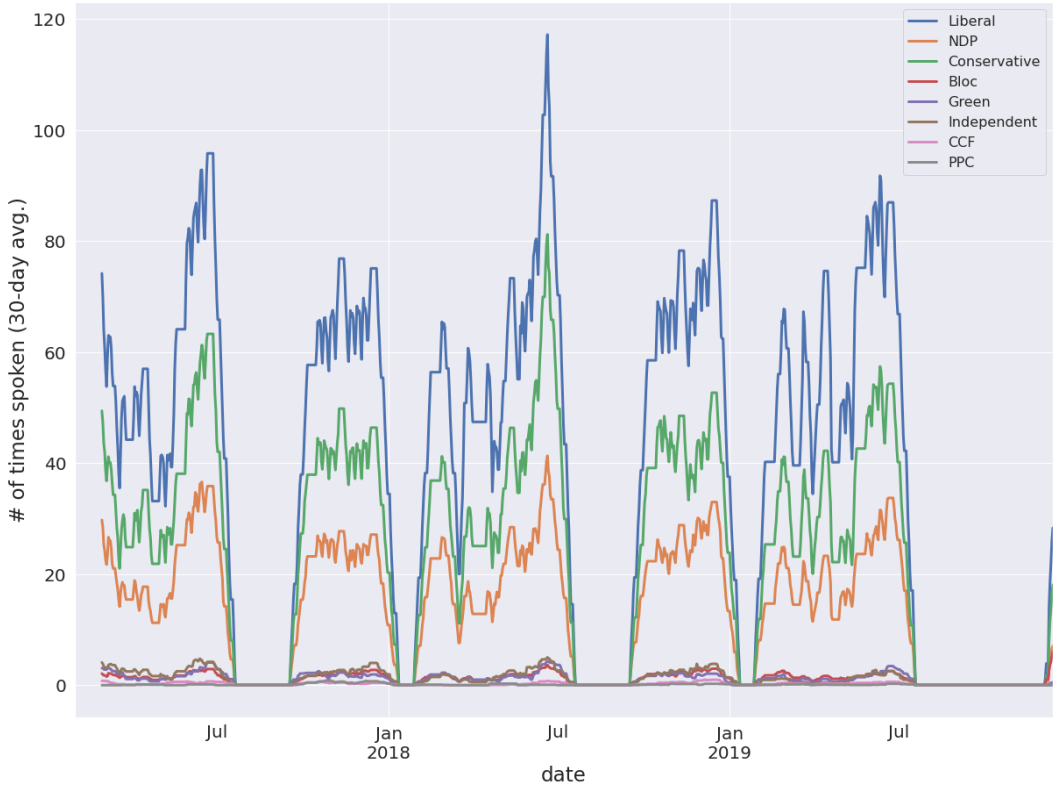


FIGURE 1. A plot of the 30-days moving average of the number of times each party spoke on the floor of the Canadian Parliament. (Years: 2017, 2018, and 2019.)

first plot, Figure 1, shows the number of times each party spoke; it shows how the Liberal, Conservative, and NDP parties dominate the speaking times (in order from greatest to least). All other parties overlap each other near the bottom of the graph.

Our second plot, Figure 2, took the ten most popular[3] topics discussed in the Canadian Parliament during 2018 and 2019. When considering some of these spikes in context of what was going on at the time, the results are not surprising. For example, two topics had a spike during the spring of 2019: Political Influence and SNC-Lavalin Group Inc.[4,5]

Taking these results, we decided to look into the co-occurrence of topics.

3.2.2. *Detecting co-occurrence of topics.* The debates in the Hansard dataset we analyzed discussed a little over 7,000 significant topics. We counted the number of times every topic was mentioned during each single hour the Parliament was in session. The dataset that resulted from this counting procedure was both high dimensional and large in size.

[3]based on amount time spoken on floor; more time equals more popular

[4]*Wikipedia*, `https://en.wikipedia.org/wiki/SNC-Lavalinaffair`

[5]When considering the top 20 topics, there is also a spike for the topic "Aboriginal People".

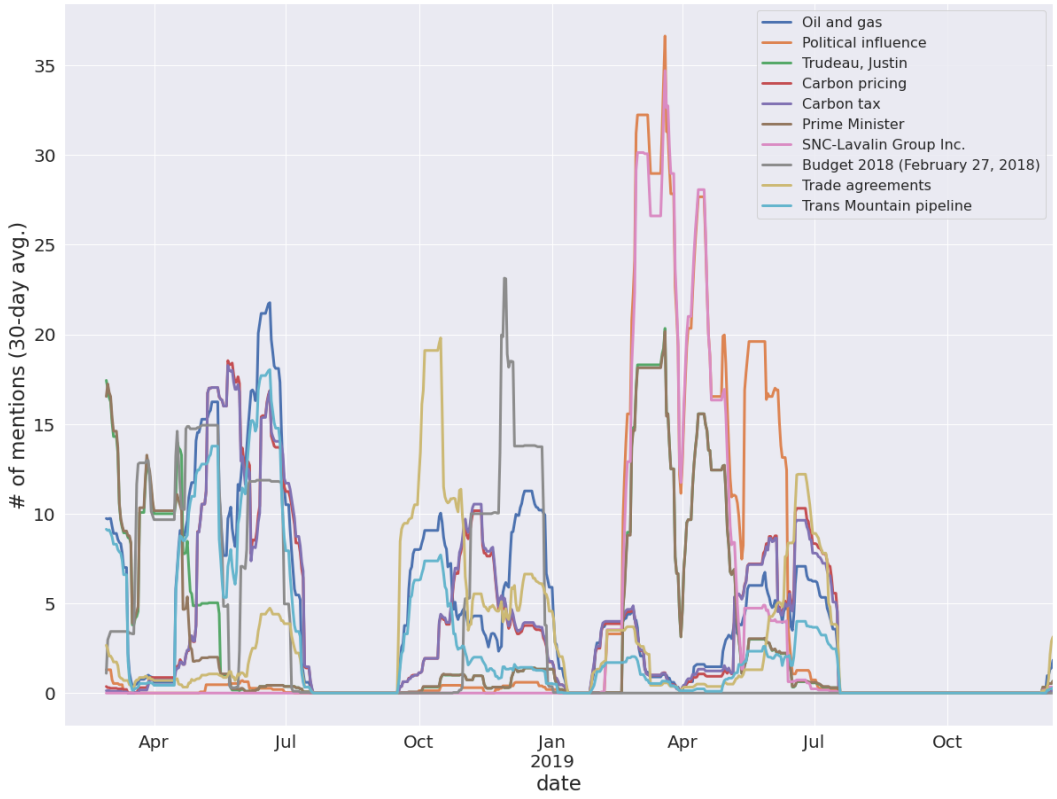
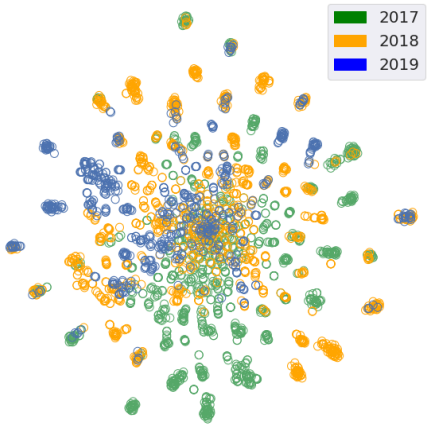


FIGURE 2. A plot of the 30-days moving average of the number of times each of the top-10 topics was mentioned on the floor of the Canadian Parliament. (Years: 2018, 2019.)

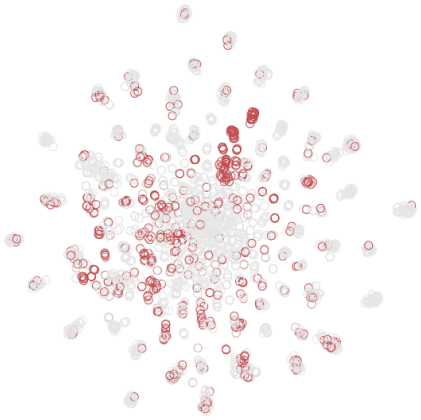
Due to time constraints and computational limitations, we restricted our attention to qualitatively analyzing the dataset through visual means. We used the t-distributed stochastic neighbor embedding algorithm (t-SNE) to create two-dimensional visualizations of our dataset that we could easily interact with. The t-SNE is a nonlinear dimensionality reduction technique that embeds high dimensional datasets in lower dimensions. The algorithm tries to keep close points in the original dataset also close in lower dimensional embedding. It is considered the state of the art in visualizing very high dimensional data.

Figure 3, which contains four plots, show some of our results from using t-SNE. Since the dataset was undersampled into hours for these plots, each dot in the plots below is a vector that has the number of times each (significant) topic was mentioned during one of those hours; multiple topics can be represented by the same dot. The top-left plot (Figure 3a) colors each dot by year. This top left plot will be used with each other plot to understand our results.

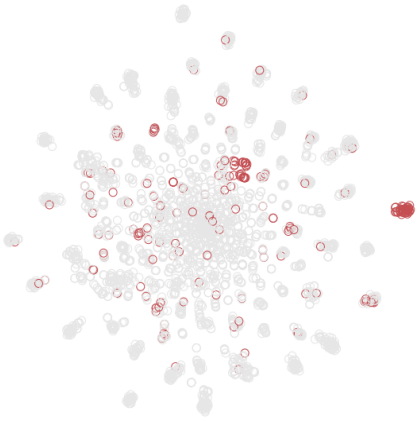
The topics will have different levels of "spread" over the three years. Figure 3b shows the trade agreements have some clustering but is relatively evenly distributed. However, Figure 3d almost entirely clustered in 2019, showing that it was a topic brought more



(A) The years 2017, 2018, and 2019 highlighted.



(B) The topic "trade agreements" highlighted.



(C) The topic "imprisonment and prisoners" highlighted.

(D) The topic "criminal prosectution" highlighted.

FIGURE 3. The t-SNE map of the topics discussed in the Hansard dataset of the Canadian Parliament for the years 2017, 2018, and 2019.

often in 2019 than the previous two years. Looking at all four plots, we can see not only the occurrence of specific topics, but also when two topics may co-occur. For instance, Figures 3b and 3c have a bit of overlap in right about the center, while Figures 3c and 3d has less.

3.3. Conclusion and Future Directions. Using both a line graph and a t-SNE plot to look at occurrence–and co-occurrence–gives us a better idea of topic trends in the Canadian Parliament. An interesting direction to take this is to compare these occurrences with the occurrences of bills introduced in Parliament around the same time; the same thing can be done with bills passed. This can give some perspective on how the speaking time on a topic may affect a bill being introduced and/or passing. This idea can be taken a step further by focusing on specific legislators who spoke on a topic. Is there a correlation with speaking time and the number of bills introduced? What about frequency a legislator spoke on a topic? Or number of bills passed? To continue this analysis, we need to gather more data about legislation from the Canadian Parliament from 2017-2019, making sure this includes the status of a bill, the (co-)sponsors of the bill, when the bill was introduced, and more.

4. U.S.

The Library of Congress was established in 1800 as a collection of books intended for the use of Congress. Despite its humble beginnings, haunted by a lack of funding, space shortages and fires, the Library has acquired a "symbolic role as a repository and promoter of the democratic tradition"[6]. In recent years, by collaborating with the House, the Senate and U.S. Government Publishing Office [7], the Library provides an official online source for legislative information[8]. The site includes a portrait of the current legislators in the form of aggregated data. Information on each member of Congress includes: name, party, number of sponsored/co-sponsored legislation, number of legislation by policy area, and more.

4.1. Motivating Questions. We attempt to use this compact data set in order to answer the following questions:

(1) How do politicians cluster around policy areas?
(2) How does this clustering, and other pertinent data, relate to a congress member's ability to pass bills to law.

4.2. Methodology and Results. As mentioned previously, our focus will be on policy area clustering. However, since the end goal is to study legislator performance, we begin by presenting some clear trends that were observed in our initial exploration. Recall that our data contained a portrait of the 437 representatives and 100 senators in the 116^{th} Congress. For each legislator, we were able to compute the percentage of bills that became law out of all the bills that have been sponsored or cosponsored by this given legislator. We will use this percentage as a measure of political performance.

A rapid overview revealed two factors which influence the percentage of bills that become law for each legislator. Firstly, we found that representatives[9] who worked in the house in 2019-2020 are more likely to (co)-sponsor bills that become law; this may need to

[6] *History of the Library*, `https://www.loc.gov/about/history-of-the-library/`

[7] *HLibrary of Congress to retire Thomas,*Adam Mazmanian `https://fcw.com/articles/2016/04/28/thomas-loc-retired.aspx`

[8] See `congress.gov`.

[9] representative here meaning a member of the House of Representatives

be compared to previous representatives to make further conclusions. Additionally, (co)-sponsorship from an experienced or "seasoned" legislator increases the probability of a bill becoming law. These findings can be visualized in Figure 4.

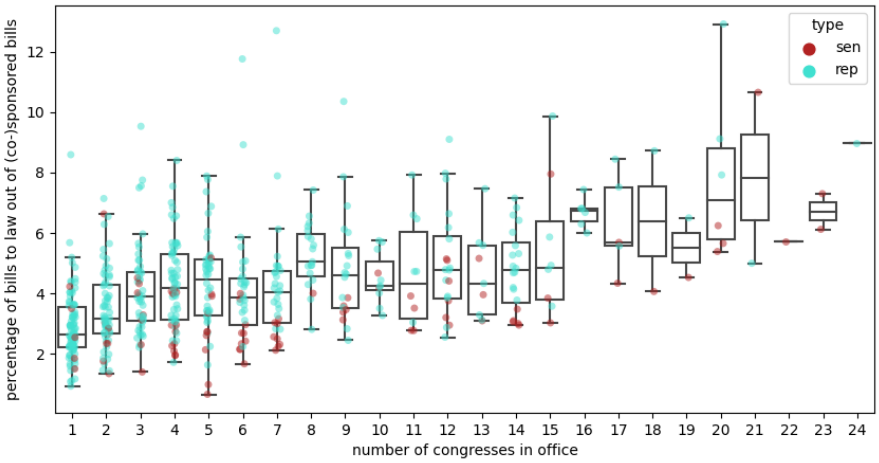


FIGURE 4. Percentage of bills that became law in terms of number of congresses spent in office.

Other features, such as party affiliation, do not appear to have an obvious effect on legislator efficiency. In order to find new trends, we must mine the data to find what lies beneath the surface.

4.2.1. *Legislators and policy areas.* When exploring the data, we chose to focus on the legislators' interest in various policy areas. To account for factors such as varying number of years in office, we consider the following ratio:

$$\frac{\textit{number of bills (co)-sponsored in a given policy area}}{\textit{total number of bills (co)-sponsored}}$$

Since some policy areas inherently require more bills than others, we then subsequently normalize these ratios.

Remark. We excluded 4 legislators from our data sets since they were newly elected and had therefore had little information. We set a cut-off off 100 bills total (co-)sponsored.

We begin to visualize our data by using t-SNE[10]. This method allows us to plot our data in a two dimensional space in such a way that preserves *nearness*. This plot is presented in Figure 5a, where we coloured the points by party. The separation between Democrats and Republicans demonstrates that policy area clustering is a reasonable indicator of political position.

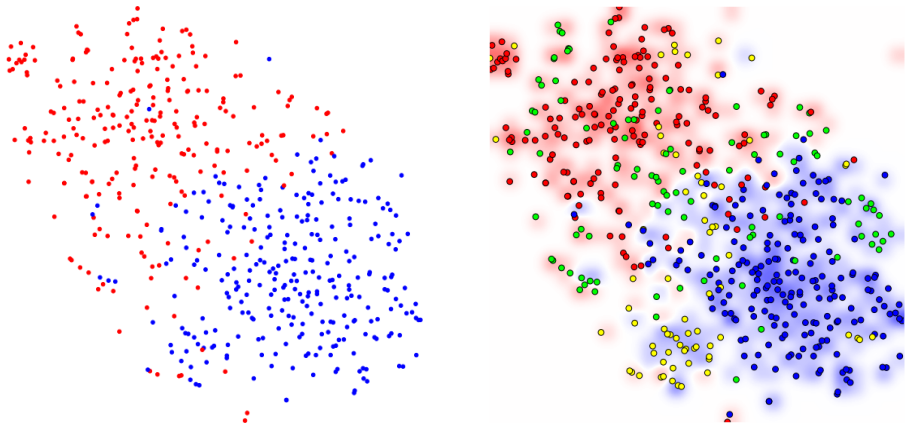
[10] `scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html`

We cluster our data using two methods: k-means[11] and gmm[12] (Gaussian mixture model). In order to maintain a balance between how much variation can be attributed to the clusters and our ability to interpret these clusters, we chose to cluster our data into four groups. After comparing our results for each method, we determined that the gmm yields better results. The kmeans algorithm produced clusters with a highly uneven number of legislators; there was a cluster containing only a single legislator.

In order to further refine our clustering, we chose to only consider a subset of policy areas. By reducing the number of features, we improved some clustering scores. This was done through feature selection.

4.2.2. *Feature Selection.* To determine which policy areas, we would re-run gmm while employing a greedy algorithm. Considering two clustering scores (Silhouette[13] and Calinski-Harabasz[14] score), we ran two greedy algorithms that, at any given step, pick the best feature to add. We plotted the scores against the added features at each step and select our desired features accordingly. Finally, we considered those policy areas that were selected with the greedy algorithm for both the Silhouette score and the Calinski-Harabasz score.

Clustering according to the selected data yields four groups of legislators. We obtained a sense of how well our data is clustered through another t-SNE visualization. In Figure 5b, we coloured our data to represent each cluster. The points are superimposed onto a heat-map based on Figure 5a.



(A) Democrats in blue Republicans in red

(B) Groups 1, 2, 3, 4 are respectively red, green, blue and yellow

FIGURE 5. t-SNE plots

[11] scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[12] scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

[13] scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouettescore.html

[14] scikit-learn.org/stable/modules/generated/sklearn.metrics.calinskiharabaszscore.html

In order to understand each group, we present Figure 6 which depicts the mean ratio of (co-)sponsored bills in each cluster.

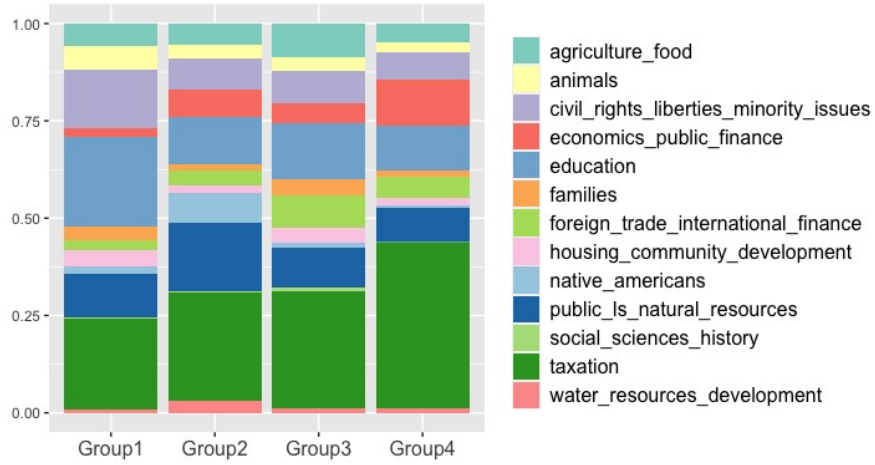


FIGURE 6. Average ratios of each policy area (feature selection)

By excluding senators, we can present the data geographically. In Figure 7, we colour each congressional district according to the cluster of the corresponding representative.

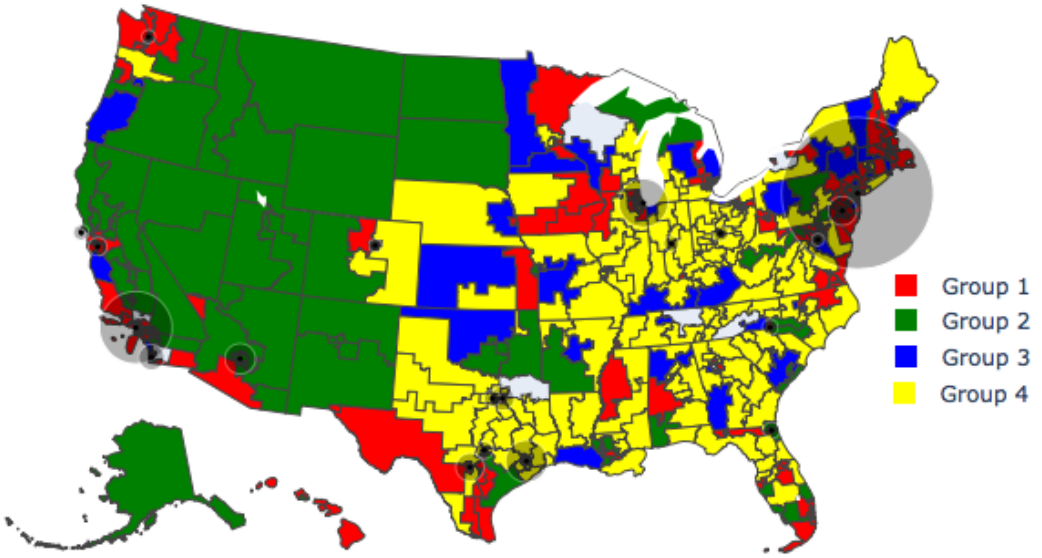


FIGURE 7. Map of House representatives by group

In Figure 7, we also included the location of the 20 largest U.S. cities with a translucent disk representing their population. The largest U.S. cities almost all fall within Group 3 (blue). This might be due to the fact that urban areas tend to have similar policy interests.

4.2.3. *Clusters and legislator performance.* Taking a look at the mean percentage of bills that became law for each cluster, we obtain minimal variability. However, it makes sense to evaluate cluster performance based on each policy area. To this end, we consider a data set containing information on bills introduced to congress since January 3rd 2019. Our data set contains information on almost 15 000 bills, including their sponsor, topic and current status (e.g. introduced, passed House, became law). For each topic, we group our data by clusters and evaluate the percentage of bills that became law. There are significant differences between the percentage of bills that became law in each cluster.

Out of the bills analyzed, only 156 have become law. Due to the small sample size, it is premature to draw any conclusions without considering information like the average number of bills passed in Congress. For this reason, we only summarize our results for what has historically been most popular policy area: health (see Table 1).

		Group 1	**Group 2**	**Group 3**	**Group 4**
Health	*total*	759	288	337	343
	to law	1	1	6	1
	percent	0.13	0.35	1.78	0.29

TABLE 1. Total number of bills sponsored in each group and amount that made it to law

Despite information on bill sponsorship in the health policy area not being included to produce our clustering, we observe a dramatic difference between Group 1 and Group 3.

4.3. CONCLUSION AND FURTHER DISCUSSION. Our results suggest that legislators' interest in policy areas is an indicator of multiple factors such as legislative performance, political position and even geographic location. We believe that our results are promising and that more work is in order. In particular, we would like to further refine our clustering by improving data preprocessing and considering other clustering methods. Moreover, we are interested in what can be found by analyzing bill specific information instead of aggregated data.

5. SUMMARY

We have shown two different approaches to answering the question if politicians cluster around topics. From our initial analysis, we do see indication that they do cluster around certain topics.

In the Canada data, you can see clusters around debate topics based on speaking time and when it occurred. The next steps is to start seeing how this data compares to Canada's legislation data, specific on bills presented and bills passed.

In the U.S. data, we see that there is clustering around parties, and from there some clustering around policy areas. However there are indicators that the clusters are not strictly down party lines and may be influenced by the region the legislator represents.

After our results were computed,[15], an article from the New York Times implies a similar trend[16].

There are many directions to go from based on our initial results. We have started looking at bills passed for U.S. data; this can be further explored by topic, legislator, region, etc. Additionally, we can use the Canada data with its legislative data to see if there are correlations. We could determine which topic debated was "most successful" and under that topic which legislator has the best record.

While analyzing legislator performance is further explored, it is important to consider the ethical implications. Detailed sports analysis has already affected players; some players in the NBA have avoided risky shots to keep good stats. We want to avoid this when analyzing legislators while still positively changing the how politics is currently viewed and discussed.

Acknowledgements

To succeed completion of this project is inseparable from the meticulous help of our mentors. We would like to express our deep gratitude to Dr. William Spat (the founder of IOTO International Inc.), Dr. Laleh Behjat (Professor of the University of Calgary), and Dr. Reza Peyghami (Professor of the University of York). All three gave us patient guidance, enthusiastic encouragement, useful comments, and interesting read materials to expand our view and deepen our understanding of the project topic.

We would also like to thank Justin Tendeck and the rest of the IOTO Team for our data support. They allowed use access to the data they mined, explained their API system, and provided us additional data when needed.

Finally, we want to thank to the PIMS Organization, with the help of organizations like Mitacs and Quansights, for creating this workshop. We especially appreciate Kristine Bauer and the rest of the Math$^{\text{Industry}}$ committee for organizing and running this successful workshop.

E-mail address: dana.berman@mail.mcgill.ca

E-mail address: jillian.glassett@wsu.edu

E-mail address: danyi2@ualberta.ca

E-mail address: diaaeldin.taha@aucegypt.edu

E-mail address: zhengx34@myumanitoba.ca

[15]On August 28th, 2020, We presented many of the results in this paper–plus some additional ones–in the PIMS Math$^{\text{Industry}}$ Showcase; this was a record session that may be posted on www.m2pi.ca

[16]*The New York Times*, "The True Colors of America's Political Spectrum Are Gray and Green", Tim Wallace and Krishna Karra, September 2, 2020, https://www.nytimes.com/interactive/2020/09/02/upshot/america-political-spectrum.html