Table of contents

SI 608 Final Report

William Godley, Lauren Lunsford, Ayaz Mammadov, Jovana Paripovic
Date: December 10, 2020

Research Question and Motivation

The tumultuous events of 2020 have led political affairs to the forefront of conversations. Especially in an election year, attitudes feel more polarized than ever before. Our team is interested in analyzing the reading behavior of Amazon reviewers before and after the 2016 election to see if polarization of readership occurs. We believe that the change in network of political affiliations over time, and after politically charged events such as the 2016 US presidential elections, should be salient. We sought to find a way to analyze that polarization and see if large political events had detectable effects on behavior. With access to the Amazon reviews dataset, we decided to focus on purchasing/readership behavior, specifically analysing the similarities or differences in left and right leaning readership.

We are therefore interested in exploring the following research question: how do book purchasing behaviors change based on political affiliation after inflammatory political events? In other words, when there are external or internal political shocks within a country, do those that are politically aligned exhibit certain changes in their book purchasing behavior or not?

Related Work

The motivation behind this topic stemmed from our curiosity regarding purchasing behaviors among politically motivated individuals and local and global political events that might influence such behaviors. In his 2017 article, which looks at profiles of conservatism and liberalism that is relevant to understanding consumer choice and behavior, John Jost argues that certain descriptors in advertisements directed at consumers are associated with political ideology (Jost 2017). These identity-relevant cues, when properly noticed, affect consumer behavior in significant ways. For example, "associates of conservatism are tough-mindedness, individualism, respect and deference to tradition and authority, while associates of liberalism are tolerance, compassion, flexibility, and openness to new experiences" (Oyserman and Schwartz 2017). Similarly, researchers have found that political affiliations have indeed shifted consumer ideology and expressiveness in the marketplace (Jacobs 2018, Kozinets and Handelman 2004, Ordabayeva and Fernandes 2018). Consumers who identify as conservative tend to prefer to differentiate themselves as better than others whereas liberal consumers tend to prefer to differentiate themselves based on uniqueness (Ordabayeva and Fernandes 2018).

Further research has also confirmed that consumer product preferences go beyond aesthetics and can stem from subconscious beliefs even when purchasing decisions "may be seemingly unrelated to politics" (Carney et al. 2008, Jihye and Vikas 2020). Most importantly, scholars link the emergence of political consumerism to the shifts in the political landscape, and argue that concerns unraveled by these changes motivate citizens to use their power as consumers to influence policy (Micheletti 2003). Finally, sociologists have found that contemporary consumption in the United States "is a primary arena in which political ideology is expressed and constructed" (Crockett and Wallendorf 2004). For our project, we hope to continue to expand this research and identify if parallels can be found through Amazon purchases and product reviews.

Data

Our dataset contains two different types of data: customer reviews and product information.

Customer reviews, also referred to as review data, consists of information pertaining to individual reviews left on a specific product page. The dataset consists of over 233.1 million Amazon reviews. Among these reviews, over 51.3 million pertain specifically to books found on the platform. Review data includes user ID, product ID, book rating, and review text.

Product information, also referred to as metadata, is information that pertains to a specific product. This includes, but is not limited to, product ID, product description, price, sales rank, products also bought, and products also viewed. There is about 12 gb worth of total metadata. 2.9 million products correspond directly to books.

The dataset can be found at https://nijianmo.github.io/amazon/index.html (Ni, Li, & McAuley 2019).

Data Processing

After providing academic credentials, both datasets, review data and metadata, were downloaded directly from the website where they are located. We then used PySpark to reduce the size of our review data in order for analysis to occur. Unnecessary columns, such as 'review_text' and 'reviewerName', were removed. The time frame of the data was reduced to January 2014 - October 2018 and only verified purchasers were selected for analysis as this ensured reliable reviews. For initial exploratory analysis and debugging purposes, random samples of 0.01% and 1% were respectively used. To reduce processing time and computational limits for additional analysis, a 10% random sample of the data was further extracted. From the resulting review dataset, the metadata was filtered to only include books that appeared in the reduced review dataset. Additionally, to get a better understanding of the political affiliation of books, a list of political books were manually extracted from composed liberal and conservative booklists (30 Books Every Young Republican Should Own)(Cain, 2020).

Methodology

Our team took a three step approach to our methodology, which included an initial exploratory analysis, classification methods, and a final evaluation of our results. Each step has been broken down into further steps as explained in the following sections.

Initial Analysis

After cleaning and sampling data, we created a bipartite graph where reviewers related to the books they reviewed. From that bipartite graph, we then created a one mode projection graph where books are the nodes and edges are shared reviewers. The one-mode projection is shown in Figure 1 below.

To validate that this method would produce edges that show similarity, we looked at the top 50 heaviest edges and identified the books within the clusters manually to see if they logically made sense as a group. Results can be found in Figure 2. Of course, with only 10% of the data, there is loss, however, for the most part, this step showed logical groupings of books. For example, the biggest grouping were books written by JRR Tolkien.

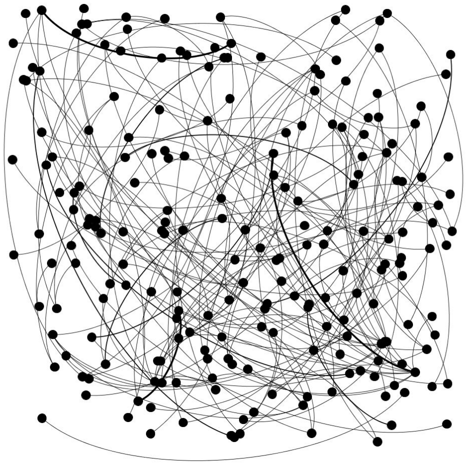


Figure 1: One Mode Projection

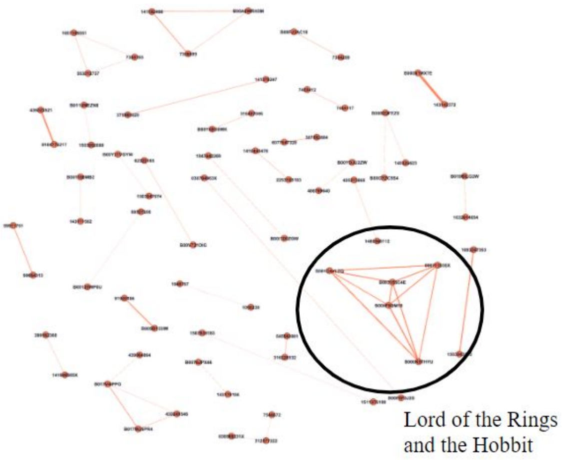


Figure 2: Methods prove that clusters of books are logically related.

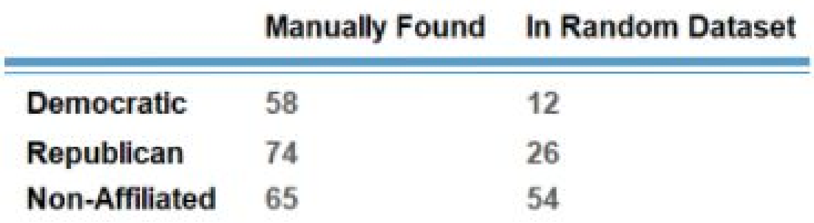
As a result of this initial exploratory analysis, our team was ready to start classifying our network for political affiliation analysis.

Classification

In order to classify books, our team went out and manually identified over 200 books as either democratic, republican, or non-affiliated. To do so, we used a variety of online databases and articles to identify left or right leaning books, went to Amazon to ensure the product was being sold in the time frame of our analysis, and collected the product id from the HTML of the Amazon webpage. Once completed, we created a dataframe and cross validated it with our random dataset. While we were able to identify books within our dataset, not all of the manually classified books could be found within our random network. As a result, the number of labels we intended to have dropped as seen in Table 1.

Table 1: Number of books manually classified and those found in the dataset.

Number of Labeled Books

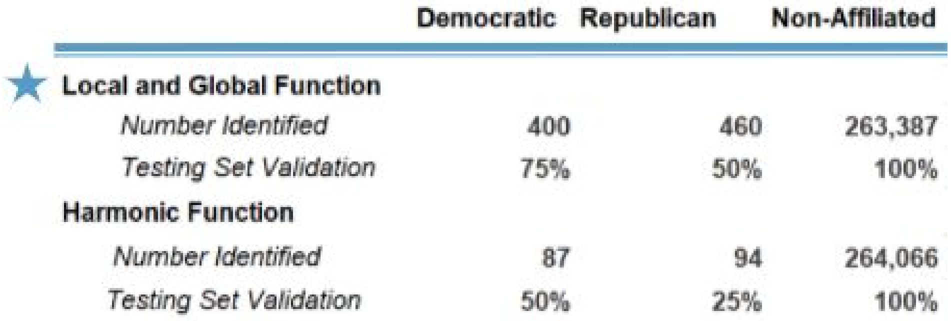


	Manually Found	In Random Dataset
Democratic	58	12
Republican	74	26
Non-Affiliated	65	54

Once we had gathered and labeled the books within our network, we turned to two different classification methods. The first method, known as the local and global consistency function, is similar to a random walk approach, however, operates under a smoothing algorithm. This means that noise is filtered out across the entire global series or over a smaller local series during classification (Zhou 2004). The second method used is known as the harmonic function. Once again, this classification method is similar to a random walk approach, however, keeps the state of the system at an equilibrium rather than iterating over time t. Additionally, the harmonic function keeps the probability of each node fixed during its execution (Zhu 2003). Each method was used on our network to test which approach would be best. As seen in Table 2, each function was able to successfully classify and accurately predict some of the books. When compared, however, the local and global consistency function outperformed the harmonic function. Not only was it able to classify more books, but based on the testing set used for validation, the local and global function was more accurate in its predictions.

Table 2: Local and global classification function outperformed the harmonic function.

Performance of Each Classification Model



	Democratic	Republican	Non-Affiliated
★ **Local and Global Function**			
Number Identified	400	460	263,387
Testing Set Validation	75%	50%	100%
Harmonic Function			
Number Identified	87	94	264,066
Testing Set Validation	50%	25%	100%

From this analysis, we continued to use the results from the local and global consistency classification to label reviewers and evaluate our results.

Evaluating Results

To label our reviewers, we remade the graph into a bipartite graph, now with the labeled books. From there, we used the labeled books to compute a political affiliation score to label reviewers as left leaning or right leaning. Republican books were given a score of -1, Democratic books were given a score of 1 and apolitical books were given a score of 0. A viewer's affiliation was taken by averaging the scores of all the books they read. Any positive values were considered left leaning or Democratic, and negative values were considered right leaning or Republican. Since our focus has been on testing political polarization within book purchasing behavior, we wanted to focus on reviewers we could label as political. Therefore, any reviewers who did not read a political book were dropped from the graph for analysis.

In order to understand the change of our network over time, our team split the data into five 13 month splits. Each time interval started and ended in the month of January allowing for a one month overlap between the previous and future split (i.e. January 2014 - January 2015). Metrics of this network were then calculated and graphed over the time period. Subgraphs of each of the two political parties were also made to visualize independent changes over time.

We calculated three metrics, assortativity, average clustering coefficients, and edges within versus between parties, over each of the 5 time periods.

i) *Assortativity* is a metric used to identify how many nodes are connecting to other similar nodes. The equation below, describes the assortativity coefficient such that j_i and k_i are the degrees of nodes connected by an edge, i, in a graph where there are M total edges (Newman 2003).

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}$$

The coefficient value is between 1 and -1 such that:

r =1 is perfectly assortative. This means nodes with high degrees are connected to other nodes of high degree.
r=0 is non-assortative. This means there is no pattern to the structures by which highly connected and slightly connected nodes are connected to each other.
r=-1 is perfectly disassortative. This means nodes with low degrees are connected to nodes of high degrees

ii) *Average clustering coefficients* reflect the average probability across all nodes that any two neighbors, j_i and k_i, of a given node i, are connected to each other. This can be seen mathematically below. Where $\{e_{ik}\}$ are the set of edges between connected neighbors and k_i is the degree of node i.

$$C_i = \frac{2\,|\,\{e_{jk}\}\,|}{k_i(k_i - 1)}$$

A high clustering coefficient indicates a dense graph. In the context of political purchasing networks, a high clustering coefficient of a party's network indicates readers have very similar reading patterns and are likely more polarized.

iii) *Edges within versus between parties* identifies the interconnectivity of groups as well as their affinity for cross-exposure. There are three components to this analysis, counting the edges within Democratic nodes, counting the edges within Republican nodes, and counting the edges between Democratic and Republican nodes. The "within" analyses provide insight into the density of each party. The "between" analyses tell us how much cross-party readership is occurring. The relationship between "within" and "between" describe the level of polarization. If most reviewers only are reading books that people of their own party read, they are more polarized and that "within" score will be higher than the "between score".

Results

The first metric our team was interested in visualizing was assortativity. As seen in Figure 3, the overall assortativity for our network was fairly high, suggesting that similar people are connected and reading similar books. It is also interesting to note the dip in 2017. While this was following the year of the U.S. Presidential election, it is not concluded whether or not the event was a result of this dip.

Assortativity Across Each Time Period

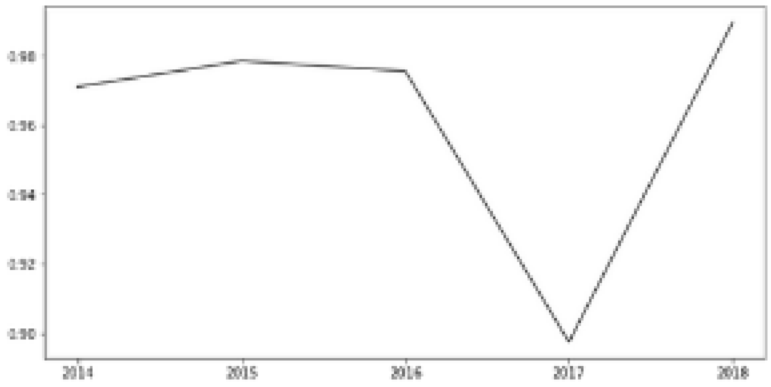


Figure 3: Total assortativity dipped in the 2017 time period.

Assortativity was also mapped for each of the two political parties. This was done to gain a better understanding of the trends within each party. As a result, as seen in Figure 4, we found that both Republicans and Democrats had fairly high assortativity, but that both dropped in the year 2016. This dip could potentially suggest that readers were reading books less similar to one another. As seen in the Democratic time series, however, Democratic assortativity fluctuated year over year, which made it difficult to form definitive conclusions from the results.

Assortativity Across Each Time Period

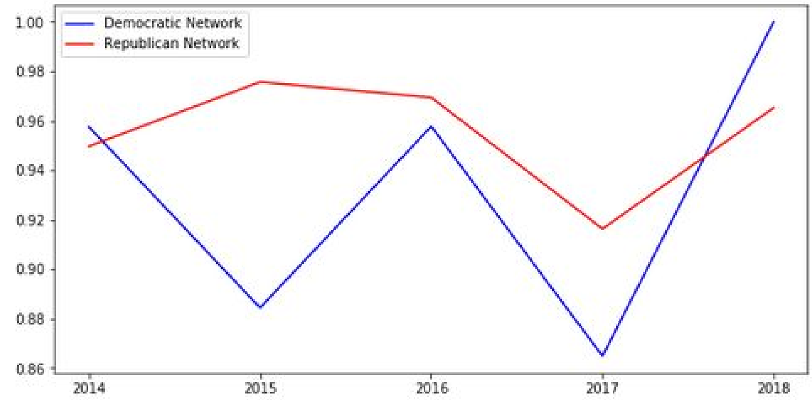


Figure 4: High assortativity was found among each political party.

Following our analysis of assortativity, we were interested in looking at how the average clustering coefficient may have changed over time. As seen in Figure 5, the average clustering coefficient in the overall network, decreased quite a bit over the years. This decrease could potentially suggest that reviewers are reading similar material amongst one another.

Average Clustering Coefficient Across Each Time Period

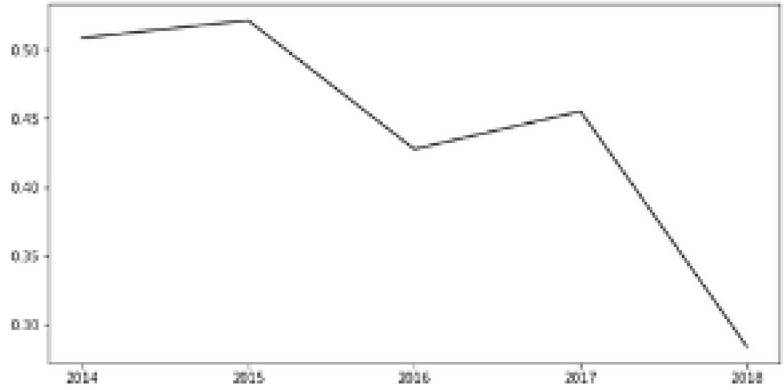


Figure 5: Total average clustering coefficient decreased over time

When we split the network into both republican and democratic networks, we were able to compare trends for each of the political parties. As seen in FIgure 6, we found that the decrease was particularly prevalent in republican reviewers. While there was a decrease in the first couple years of our analysis for the democratic party, after 2016, democratic reviewers significantly increased the amount of material reviewed amongst one another. Once again, while we cannot completely attribute this to the 2016 U.S. Presidential election, there is fluctuation as a result of this event.

Average Clustering Coefficient Across Each Time Period

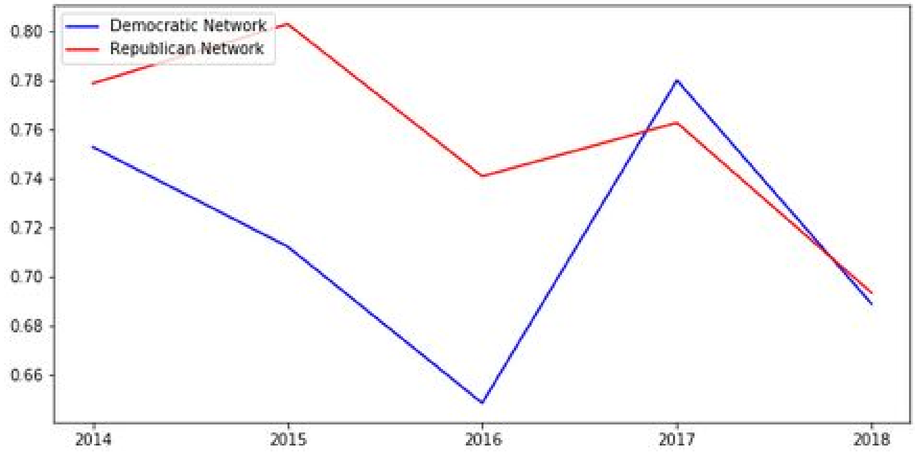
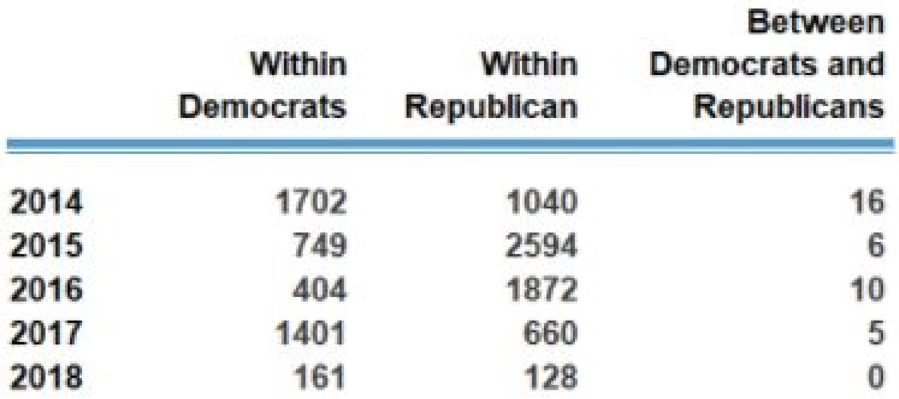


Figure 6: Clustering coefficient fluctuates over time.

The final metric our team chose to evaluate were the number of edges present within and between each political party. As seen in Table 3, edges present within both democratic and republican parties were extremely prevalent. Edges between parties, on the other hand, were few and far between. The lack of edges indicated that there was very little interconnectivity between democrats and republicans. It is important to note that these results could be due to the use of readership as a classification input resulting in more edges within parties rather than between.

Table 3: Number of edges between parties show little to no interconnectivity.

Number of Edges Within and Between Parties



	Within Democrats	Within Republican	Between Democrats and Republicans
2014	1702	1040	16
2015	749	2594	6
2016	404	1872	10
2017	1401	660	5
2018	161	128	0

Results from Table 3 were normalized and visualized as seen in Figure 7. As seen by the stagnant line represented by the Total Within Groups, no significant changes were seen in total betweenness vs withinness across time.

Percent of Edges Within and Between Parties

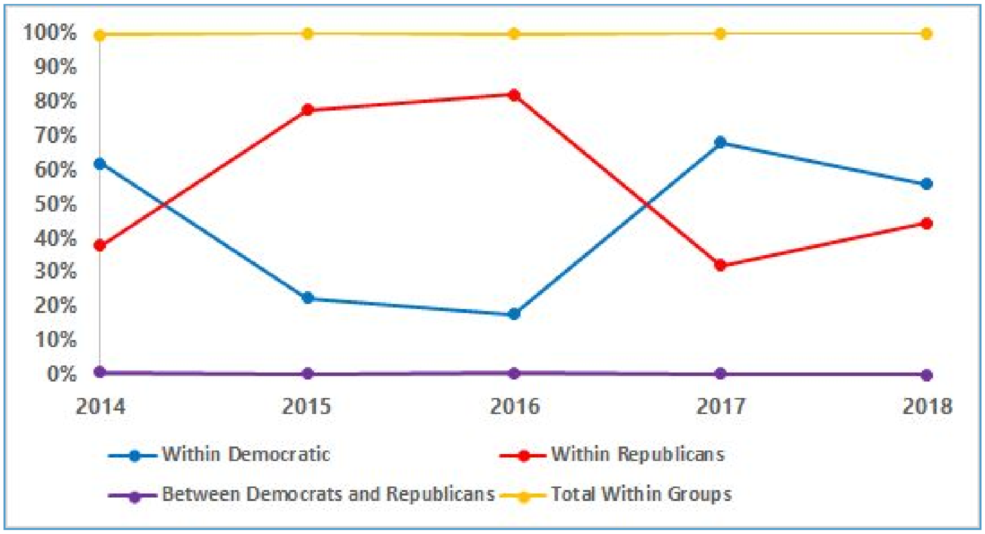


Figure 7: Percent of edges over time show little fluctuation among total within groups.

Given our analysis, we were able to identify trends within and between each political party. Each metric revealed new findings and gave us insight as to how the political network changed over time.

Challenges

Our team experienced a few challenges during the course of this project. First, the vast size of the data provided to be difficult to work with and often resulted in our systems crashing. As a result, we were able to reduce the size into a 10% random sample, however, more data would have allowed for more robust results. As we worked to classify books, the manual task of labeling how books were affiliated proved to be difficult as it was not guaranteed that the books found manually would be in the random sample set. This led us into our third challenge of not having enough training and testing labels when computing classification. Having more labels would have led to more accurate classification results as there would be more data for the algorithm to use when being trained. Additionally, books that did not have common reviewers were dropped since they were not able to be classified which may have led to less accurate findings. Finally, while we were able to successfully classify books and reviewers, we do not have a way to verify the political affiliation of these nodes for evaluation.

Conclusion

As a result of our analysis, we were successfully able to classify ~265,000 books and ~2,400 reviewers from 2014 - 2018. Based on this classification, our main finding was that similar people review similar books. While this could have been due to the result of our random walk classification method, it was ultimately a finding that our results showed. As seen in the number of connected edges within and between each party, we found very little interconnectivity between Democrats and Republicans. Little interconnectivity suggests that reviewers do not read books from the opposite political affiliation. Additionally, as described from the clustering coefficient over time, Republicans have become less polarized while Democrats have been seen to become more polarized. While we were able to compute metrics and identify trends for each political party, we found no evidence in support of our hypothesis. In other words, no evidence was found to prove that readership became more polarized as a result of inflammatory events.

References

1. Carney R. Dana, John T. Jost, Samuel D. Gosling, Jeff Potter. 2008. “The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind.” Political Psychology, 29 (6): 807-840.

2. Cain, H. (2020, November 07). 27 of the Best Political Books to Read to Process the 2020 Election. Retrieved November 01, 2020, from https://www.oprahmag.com/entertainment/books/g23804630/best-political-books/

3. Crockett, David, and Melanie Wallendorf. 2004. "The role of normative political ideology in consumer behavior." Journal of Consumer Research, 31 (3): 511-528.

4. Jacobs, Tom. How Politics Can Influence What You Buy. 17 Feb. 2018, theweek.com/articles/755274/how-politics-influence-what-buy.

5. Jihye Jung, Vikas Mittal. 2020. “Political Identity and the Consumer Journey: A Research Review.” Journal of Retailing, 96 (1): 55-73.

6. Jost, John. 2017. “The marketplace of ideology: “Elective affinities” in political psychology and their implications for consumer behavior.” Journal of Consumer Psychology, 27 (4): 502–520.

7. Kozinets, Robert V., and Jay M. Handelman. 2004. "Adversaries of consumption: Consumer movements, activism, and ideology." Journal of Consumer Research, 31(3): 691-704.

8. Micheletti, Michele. 2003. Political Virtue and Shopping: Individuals, Consumerism, and Collective Action: Pelgrave Macmillan.

9. Newman, M. E. J. “Mixing Patterns in Networks.” Physical Review E, vol. 67, no. 2, 2003, doi:10.1103/physreve.67.026126.

10. Ni Jianmo, Jiacheng Li, Julian McAuley. 2019. “Justifying recommendations using distantly-labeled reviews and fine-grained aspects.” Empirical Methods in Natural Language Processing (EMNLP).

11. Ordabayeva, Nailya, Daniel Fernandes. 2018. “Better or Different? How Political Ideology Shapes Preferences for Differentiation in the Social Hierarchy.” Journal of Consumer Research, 45 (2): 227–250.

12. Oyserman Daphna, Norbert Schwarz. 2017. “Conservatism as a situated identity: Implications for consumer behavior.” Journal of Consumer Psychology, 27 (4): 532-536.

13. Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. Advances in neural information processing systems, 16(16), 321-328.

14. Zhu, X., Ghahramani, Z., & Lafferty, J. (2003, August). Semi-supervised learning using gaussian fields and harmonic functions. In ICML (Vol. 3, pp. 912-919).

15. 30 Books Every Young Republican Should Own. (n.d.). Retrieved November 10, 2020, from https://www.bestdegreeprograms.org/best-books-for-young-republicans/