Stochastic Blockmodels for Modelling Political Engagement

by

Cameron Raymond

A thesis submitted to the
School of Computing
in conformity with the requirements for
the course CISC 500 - Undergraduate Thesis

Queen's University
Kingston, Ontario, Canada
March 2020

Supervisor: Dr. Robin Dawes

Abstract

The advent of social media has enabled political parties to engage with the broader populous in new and unforeseen ways. This, coupled with rising levels of political polarization has prompted debates as to whether people care about policy anymore, or if they self-select into political bubbles online based on their chosen party leader. This thesis proposes a novel adaptation of stochastic blockmodelling to measure the degree to which political engagement on Twitter was driven by policy or party leaders. Building on a graph theoretical approach, measures of topic centrality are developed to give a metric for how efficient topics were at either rallying or spanning party leaders' bases. This is done in the context of the 2019 Canadian federal election.

Acknowledgments

I am very grateful for the guidance of my thesis supervisor, Dr. Robin Dawes, who has supported me throughout my this research project, and has been instrumental in my transition to graduate school. Though not my supervisors, I would also like to thank Dr. Adrienne Davidson and Dr. Jonathan Rose for their roles and help, ensuring the social sciences are not lost in the technicality. Finally, to my family, friends, and Miranda Dell'Anna – who have put up with my neuroticism over the past 10 months, thank you.

Contents

Introduction

1.1 Background

The advent of social media has enabled political parties to engage with the broader populous in new and unforeseen ways. The ability to bypass the traditional mediating forces of mass media allows for an unfiltered promotion of policy, ideology and party stances. This is specifically interesting in Canada's political system which has historically been defined by large brokerage parties. In order to win a diverse range of electoral districts across Canada, these "big tent" parties try to appeal to various political persuasions. Political adverts, and policy have traditionally been the conduits through which brokerage parties attempt to accommodate different ideologies, but social media allows for a direct, granular approach to political messaging which is completely novel. Social networks, formed through new media like Twitter, are inherently relational and thus lend themselves well to being represented as graphs. Therefore, as political strategy becomes increasingly digital, the use of graph theory can potentially illustrate how large brokerage parties organize and along what axes Canadians engage with political parties.

1.2 Motivation

The way information is distributed and received has changed significantly over the past decade. Cogburn and Espinoza-Vasquez argue that Barrack Obama's 2008 presidential campaign was a watershed moment in social media campaigning – and in the subsequent decade, from Macron to Brexit to the Five Star Movement, social media has played an increasing role in how politics is conducted [6]. Between 2013 and 2018, the share of Canadian federal media expenditure spent on digital advertising rose from 27% to 65%, a 140% increase, making the study of new media critical from a social science perspective [1]. Additionally, rises in political polarization, populism and a decline in trust in political institutions in the 21^{st} century has been a topic of popular debate. Ezra Klein argues in his 2019 book, *Why We're Polarized*, that this due to a shift in preferences for parties over policies [12]. If this is the case, then this preference to engage along party lines rather than choosing to engage with specific issues should pattern engagement. An empirical analysis of how users behave and engage with political parties online should privilege the relational aspect of social media. Social network analysis helps avoid the pitfalls of survey data, famously described by Allen Barton as "a sociological meat grinder, tearing the individual from [their] social context" [9].

Graph theory's use in social network analysis, also called network science, has already been applied to explore problems in marketing, sociology and epidemiology – but there is a gap in analysis of political engagement online. Therefore, the contribution of this thesis is a novel, robust mathematical process for analyzing different axes of political engagement online in a purely relational manner. A secondary outcome of making relationships a "first-class citizen" in this work will be an organic analysis of what issues produce the most engagement. This deviates from traditional survey data that ask test subjects which issues concern them – and instead uses observations of past behaviour to model variations in local connectivity to answer the question: "what do people actually care about?" All of this will be done in the context of the 2019 Canadian federal election and the tweets of Canada's five major, english speaking party leaders: Andrew Scheer, Elizabeth May, Jagmeet Singh, Justin Trudeau, and Maxime Bernier.

1.3 Research Question

The primary question concerning this project is: in the lead up to the 2019 Canadian federal election, did politically active users on Twitter engage with political elites along the axis of issues[1] or parties? If Ezra Klein is correct, then *who* produces the message will pattern engagement more than *what* the message is; this will act as the initial null hypothesis, with the alternate hypothesis being that who produces the message is equally to or less important than what the message is.

The secondary question to be explored is: during this period, what topics produced the highest level of engagement? Also, what topics spanned multiple party leaders' bases, indicating a bridging of different ideologies, and what topics rallied party leaders' bases?

[1] The terms policy, issue and topic will be used interchangeably to refer to categories of messages.

Social Media and Politics

2.1 Political Communication in the Digital Era

Technological innovation often precedes political disruption; Martin Luther's 95 theses could only be spread at scale by virtue of Gutenberg inventing his printing press first [10]. This example had obvious political ramifications and is a testament to the impact of the means of communication on the political sphere. McNair defines political communication as:

1. All forms of communication undertaken by politicians and other political actors for the purpose of achieving specific objectives.

2. Communication addressed to these actors by non-politicians such as voters and newspaper columnists.

3. Communication about these actors and their activities, as contained in news reports, editorials, and other forms of media discussion of politics. [13]

This is a broad, outward focused definition that includes most public, political discourse – verbal or otherwise. Social media has changed all three aspects of this definition in novel ways. Most notably, the rate at which politicians can communicate information to achieve "specific objectives" has increased as a result of rapid nature of social media, and the granularity at which actors can target the these messages has gotten smaller due to the massive swaths of user data available [16]. The increasing shift towards using social media for the purposes of political communication, and the ease with which these messages and the users who engaged with them can be collected and analyzed demonstrate the value of using computational methods to study political communication in the 21^{st} century.

2.1.1 Technology's Implications for Democratization

While fully exploring the impact of social media on democracy is out of the scope of this project, it is an important justification for why research in this area needs to be conducted. The traditional model of the media being a mediating force through their reporting, commentary and analysis is no longer valid. Previously, political actors needed to use the media to broadcast their messages to their desired audiences [13]. Social networking sites allow these same actors to reach audiences in the millions without having to gain access to the media first; in this effect, the second and third elements of McNair's definition are also being transformed. Yascha Mounk argues that this has given voice to political outsiders who would be shut out from mainstream platforms. Thus, social media may not be inherently democratic or undemocratic, as it has contributed to democratic backsliding and overturning authoritarian governments, but it can certainly have a destabilizing affect [14]. Therefore, a better understanding of new media's ramifications is critical, and empirical modeling can aide in this understanding.

2.2 Canadian Brokerage Politics

While it is clear that technology is changing how information is received, and thus must also be changing how politics is conducted, it may not be clear the role of Canadian politics in this context. However, Canada's political system is a fertile environment to test the importance of political messaging, because relative to most liberal democracies, the system is dominated by party politicians. As Carty put it:

> No obvious simple geographic reality, no common linguistic or religious homogeneity, no common revolutionary experience or unique historical moment animated [Canada] or gave it life. Canada was created when a coalition of party politicians deemed it to be in their interest to do so, and it has been continuously grown, reshaped and defended by its politicians. [4]

Thus, it is not surprising that Canada's electoral system encourages electoral pragmatism – and developed large, "big tent" parties that are among the most organizationally weak and decentralized of established democracies [4]. This system defines political parties as brokers of the often conflicting, weakly integrated electorate – as opposed to mobilizers of distinct communities, articulating claims rooted in their pre-existing interests. In this way, parties act as the principal instruments of national accommodation, rather than democratic division [4].

Rationale

The dominance of parties in Canadian politics, their amorphous ideological stances, and the many intersectional geographic, linguistic and religious cleavages have given birth to what's been coined the brokerage party system. [4]. The need to capture pluralities in a diverse range of electoral districts gives Canada a unique set of dynamics - making it an interesting case to explore.

Modelling Unstructured Text Data

3.1 Background

The majority of data currently being produced is unstructured and unclassified. As a result, there is a need for techniques that automatically organize big, unclassified corpuses of text. Topic modeling finds clusters of words that frequently occur together (topics), connects words with similar meanings, and distinguishes different uses of words with multiple meanings [2]. This is based on the underlying assumption that a document is generally concerned with a fixed set of topics, and that the frequency of words used is indicative of this latent structure [3]. Topic modeling has been used extensively to create recommendation systems, perform trending analysis, and segment text [2]. In this context, topic modeling is necessary to organize the tweets party leaders are promoting by their latent topics. The ultimate goal of evaluating along what axes the broader populous engages with political media (policies/topics/issues or party lines) necessitates a robust way of evaluating messages. To know if people engage based on various topics requires knowing what those topics are in the first place. Topic extraction approaches based on keywords are brittle, context specific and are unable to capture emergent topics. Using unsupervised machine learning techniques, like the latent Dirichlet allocation discussed in section 3.2, topics are able to be extracted in an autonomous manner - requiring little oversight.

3.2 Latent Dirichlet Allocation

Blei, Ng and Jordan describe latent Dirichlet allocation (LDA) as a "generative probabilistic model for collections of discrete data such as text corpora." The goal being to extract short descriptions of similar topics from a collection - and describe statistical relationships that are useful for classification, summarization, and describing similarity [3]. The underlying thought process behind the LDA is that each document in a corpus can be described as a distribution of topics, and that each topic can be described as a distribution of words.

In this processes, each word in all the text corpus is an element of the vocabulary, $Voc : \{1, ..., V\}$. Each word, $v \in V$, is represented by w, a unit-basis vector of dimension V where:

$$w[index = i] \begin{cases} 1 & \text{if } v = i \\ 0 & \text{if } v \neq i \end{cases}$$

A document is a sequence of N words denoted by $\boldsymbol{w} = (w_1, w_2, ..., w_N)$. A corpus is a collection of M documents denoted by $D = \{\boldsymbol{w_1}, \boldsymbol{w_2}, ..., \boldsymbol{w_M}\}$.

The input to LDA, M, will be the tweets aggregated and cleaned from the tweets of federal Canadian party leaders described in 1.2. Aside from the corpus and k - the number of topics - two other parameters are fed to the LDA: first, α, which is organizes the ground θ and acts as a concentration parameter for how documents are modelled as topics. Higher α values generally implies that documents will be viewed as a mixture of topics, whereas low α values imply that documents will be viewed as a belonging to a single topic. Similarly, to model words as topics the parameter η organizes the ground for β, a Dirichlet distribution.

The generative model for the LDA uses θ to choose a topic z_n of the k topics the next word will reside from, and chooses a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n. This processes is described graphically in figure 3.1

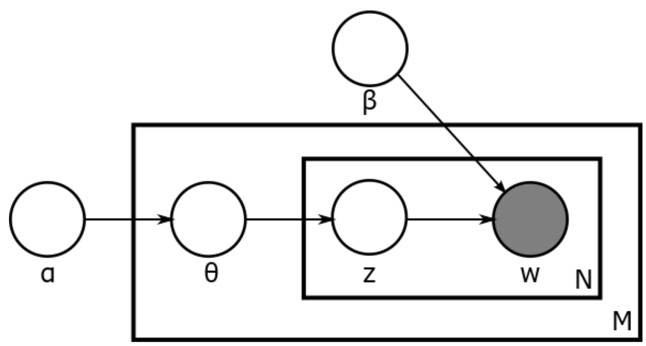


Figure 3.1: Graphical Model of the Latent Dirichlet Allocation

By using variational inference - θ, a distribution of topics for each document and β, a distribution of words (one for each topic) can be solved for giving the final equation $P(\theta_{1:M}, \boldsymbol{z}_{1:M}, \beta_{1:k}|D, \alpha_{1:M}, \eta_{1:k})$ [3].

Graph Theory and Computational Social Science

Graph theory is the study of mathematical structures, called graphs, which are used to model pairwise relations between entities. Graphs consist of a finite set of vertices, V, and a set of ordered pairs of vertices, E, called edges. A graph can be defined by the tuple, $G = (V, E)$. The graphs built in this project have added constraints and is defined as below:

- *Vertices*: Let $V_1 = \{v_1, v_2, ..., v_n\}$ be the set of party leaders; $V_2 = \{v_1, v_2, ..., v_m\}$ be the set of tweets by the party leaders described in section 1.2; and let $V_3 = \{v_1, v_2, ..., v_k\}$ be the set of users who retweet tweets. Let the total set of vertices, V, be equivalent to $V_1 \cup V_2 \cup V_3$.

- *Edges*: Let E be the set of edges. Allow the edge $(v_1, v_2) \in E$ if and only if $v_1 \in V_1, v_2 \in V_2$ or $v_1 \in V_3, v_2 \in V_2$. By this definition, we will only allow edges from a party leader vertex to a tweet vertex, or from a"retweeter" vertex to a tweet vertex.

Figure 4.1 shows the full graph – with 7,978 tweets, 36,450 generic users, and 113,293 retweet edges – built with these constraints.

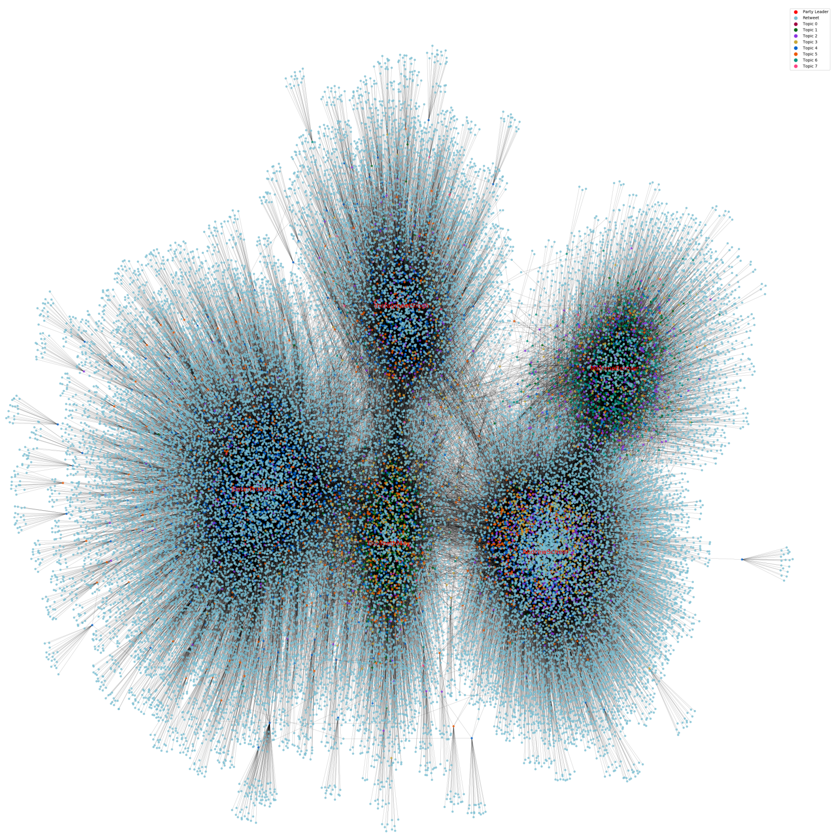


Figure 4.1: Complete Political Engagement Graph

4.1 Spectral Graph Theory

Spectral graph theory studies the structures of graphs via the eigenvectors of their adjacency matrix, Laplacian matrix[1], or some other variant of the two. The set of eigenvalues for a graph if size n, $\{\lambda_1, ..., \lambda_n\}$, is called the spectrum of a graph [20]. As Hammond et al. note, graph spectra are closely related to major graph invariants [5]. Graph spectra have been used in image segmentation and object recognition tasks, as well as in studying the stability of molecules; Elghawalby and Hancock demonstrated

[1]The Laplacian matrix of a graph is defined as its degree matrix minus its adjacency matrix.

how the euclidian distance between graphs' spectra track the edit distances between graphs [8, 5]. As such, spectral graph theory is useful in comparing the underlying structure of two graphs – allowing for nuanced distance and similarity metrics.

4.1.1 Network Laplacian Spectral Descriptor

While various distance metrics are tracked by graph spectra, most are not size invariant or scale adaptive. Size invariant similarity metrics capture two structurally similar graphs as close together, regardless of magnitude; for example, social networks like Facebook and Google Hangout likely have similar structural patterns despite the former being much larger. Scale adaptive similarity metrics would be able to capture both local and global features of a graph. Tsitsulin et al. developed the Network Laplacian Spectral Descriptor (NetLSD), a state of the art similarity measure, that creates extracts a compact heat trace signature from a graph's normalized Laplacian spectrum using the heat kernel [20]. This, in effect, models how heat diffuses throughout a network over; with local features being captured in the immediate time-steps after the nodes are "heated" (and only affecting adjacent nodes) and global features being captured as heat becomes further diffuse. Figure 4.2 shows how the two graphs can be similar at a global level and local level, but differ at an intermediate scale [20].

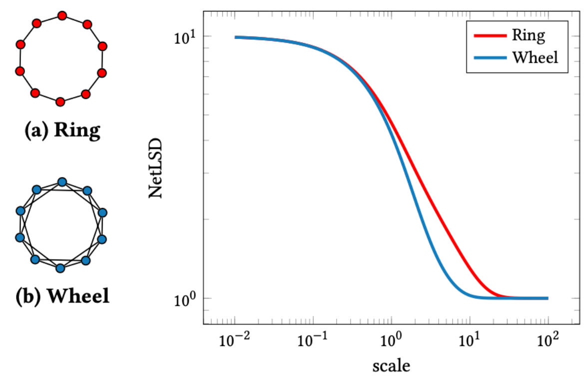


Figure 4.2: NetLSD Heat Trace Signatures for two Similar Graphs [20]

The heat trace (h_t) for a graph at time t is calculated by taking the eigenvalues of the graph's normalized Laplacian matrix, and summating their exponentiation multiplied by $-t$.

$$h_t = \sum_{j}^{n} e^{-t\lambda_j} \tag{4.1}$$

Given that a graph with n nodes will have n eigenvalues and thus have a higher heat trace value at any given time – the heat trace can be normalized against an empty graph, which has all zero eigenvalues [20]. The *heat trace signature* then is a vector of different heat traces at different times denoted by $h(G)$:

$$h(G) = \{h_t\}_{t \geq 0} \tag{4.2}$$

Since the heat trace signatures of two graphs lie in the same dimensional vector space, taking the L_2 distance of the difference between the two heat trace vectors provides a suitable distance metric.

4.2 Random Graphs

Real world interactions are not deterministic. That is – who your friends are, whether you contract a disease, or whether you choose to retweet a tweet are the result of random processes, patterned by your connections with others. Therefore, modelling how social networks form requires investigating the underlying of the processes that generate these random ties. As Robins' discusses in his article, *a tutorial on methods for the modeling and analysis of social network data*, graphs where edges are generated in some stochastic process can help illustrate how relationships drive behaviour [17].

4.2.1 Stochastic Blockmodels

Initially, random graph models were primarily developed to "model variations in local connectivity of each network actor, the degree of local clustering among actors, and the general distribution of connectivity." [17] However, intrinsic characteristics of actors in a network often patterns the relationships found; inter vs intra-group dynamics drive Shakespeare's Romeo and Juliet, and how the characters behaved was largely determined by which family they were a part of [7]. Blocks therefore group all nodes of the same characteristics. The proportion of all possibles edges within a group – and the proportion of all possible edges that span to all other blocks – formulate the probabilities of forming edges in the generative blockmodel. Figure 4.3 visualizes a friendship network with each node (representing a person) coloured according to their sex, and edges between nodes denoting friendship between them.

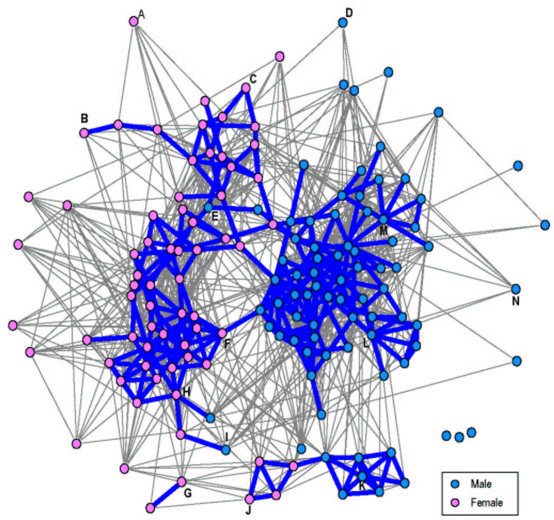


Figure 4.3: Friendship Blockmodel With Nodes Coloured by Sex

Of all the possible edges for a network of that size, the ones observed are overwhelmingly between members of the same sex, with fewer edges spanning that characteristic. These are attributes that shape behaviour that blockmodels attempt to capture.

Stochastic blockmodelling, first proposed by Holland et al. in 1983, is a subset of blockmodels that takes classes of nodes to be latent [11]. The goal, therefore, of stochastic blockmodels are to determine the latent classes necessary to define the blockmodel.

For the purposes of this thesis, stochastic blockmodels will be adapted and extended to capture how users on Twitter show preferences for party leaders and/or policies when deciding whether to engage with tweets. The model will have to realistically model the spectrum of users; from those who only engage with one party leader, to those who retweet multiple party leaders, to those whose engagement driven by specific issues.

4.3 Measures of Centrality

Centrality is a measure of prominence for vertices within a graph. For the purpose of this thesis, it will be used to measure the relative importance of different topics tweeted about in the lead up to Canada's 2019 federal election.

4.3.1 Background

There are various different ways of measuring vertex centrality that have successfully been applied to problems in marketing, economics and epidemiology; Stephenson and Zelen explored the utility of centrality measures in studying the social dynamics of Gelada baboons. [19]. Common centrality measures include measures of degree and betweenness. This thesis will focus on the notion that central vertices are close to other central vertices, which is one of the founding intuitions behind Google's "page-rank" algorithm and eigenvector centrality.

4.3.2 Eigenvector Centrality

As Newman lays out in his 2016, Mathematics of Networks: "the eigenvector centrality [...] accords each vertex a centrality that depends both on the number and the quality of its connections: having a large number of connections still counts for something, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts." [15]

If we define the eigencentrality of vertex x as $C_E(x)$, where $C_E(x)$ is proportional to the average eigenvector centrality of x's neighbours, multiplied by some constant

λ:

$$C_E(x) = \frac{1}{\lambda}\sum_{j=1}^{n} A_{xj} C_E(x) \tag{4.3}$$

By defining the vector of centralities as $C_E(X) = (C_E(x_1), C_E(x_2), ...)$ this equation can be rewritten as $\lambda C_E(X) = \lambda A$, and it is evident that CE(V) is an eigenvector of the adjacency matrix with eigenvalue, λ [15]. By Perron-Frobenius theorem, picking the largest eigenvalue of A will result in all elements of $C_E(X)$ being non-negative [15].

Thesis Contribution

5.1 Topic Modelling

In order to evaluate the relative importance of policy and party leaders in driving political engagement on Twitter, all the tweets collected must first be organized by topic. In order to do so, a latent Dirichlet allocation (LDA) was trained on the English tweets of Canada's five major, english speaking party leaders: Andrew Scheer, Elizabeth May, Jagmeet Singh, Justin Trudeau, and Maxime Bernier. The timeframe of collection ranges from October 21, 2018 to October 21, 2019 - the eve of Canada's federal election. While the tweets from each Federal party's official accounts were also collected, they predominantly acted as logistical tools – informing party affiliates of events and rallies. The personal accounts for party leaders were generally more pertinent to their beliefs, platforms and style of rhetoric, and thus are better suited in this context. In this spirit, only tweets of the party leader were collected, excluding retweets. Figure 5.1 visualizes the daily and cumulative number of tweets over time, in aggregate and by party leader, resulting in 7978 total tweets.

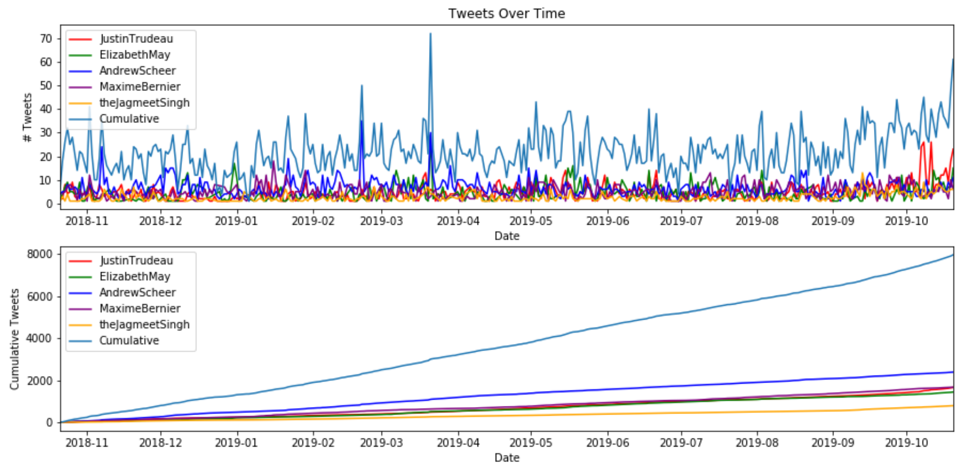


Figure 5.1: Daily and Cumulative Tweets over Time

Text Cleaning

Given the inherent noise and extraneous info in text data, it is standard and necessary to preprocess text before modelling [18]. The text cleaning pipeline removes punctuation marks, stop words, words with fewer than three characters, and URLS, as well as common twitter symbols like "RT:", "@" and "#". Emojis were converted to text using the python package `emoji`. After this process, all text was converted to lower-case and lemmatized to get rid of common suffixes. Therefore the tweet in figure 5.2 after preprocessing reads: *wherever maple leaf fly represents rich history bright future value hold dear happy flag day canada.*

Figure 5.2: Example Tweet

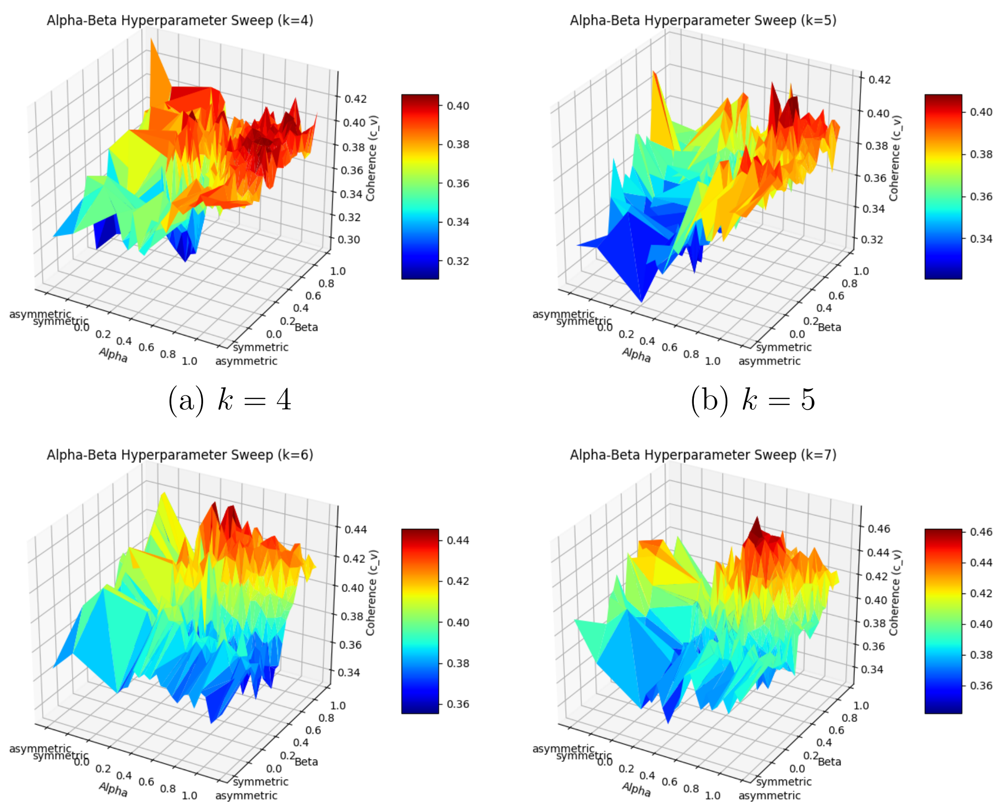
5.1.1 Hyper-Parameter Tuning

As discussed in section 3, the LDA takes in three parameters: α - which acts as a concentration parameter for how documents are modelled as topics; β - which acts as a concentration parameter for how topics are modelled as words; and k which is the number of topics to be modelled. By performing a parameter sweep, where α and β lie on the interval $[0, 1]$ with increments of 0.05, and k ranges between 4 and 7, the LDA was exposed to the entire corpus and then evaluated using c_v coherence. Figure 5.3 shows, for each k value, the c_v coherence as a function of different combinations of α and β.

5.1.2 Results

After performing the parameter sweep described in section 5.1.1, the most performant model had a k value of 7, α of 0.31 and β of 0.81 and a c_v coherence score of 0.48. By labelling each tweet as the maximum probability value in its topic mixture, each tweet was assigned a single topic. The word clouds for each topic are described in figure 5.4.

Topic 1 pertained to campaign messages, rallies and logistics – and makes up

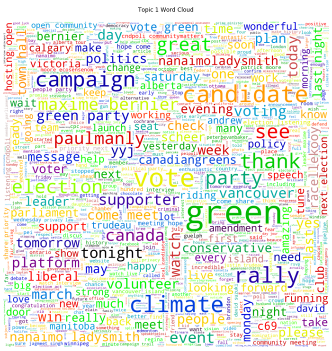


(a) $k = 4$ (b) $k = 5$

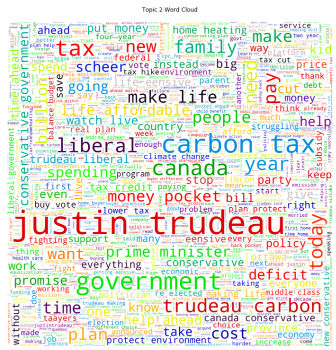
(c) $k = 6$ (d) $k = 7$

Figure 5.3: LDA Parameter Sweep Results

8.2% of all tweets. Topic 2 contains tweets regarding a carbon tax, pipelines and the economy – and makes up 16.3% of all tweets. Topic 3 contains tweets about the SNC Lavalin affair, a scandal that plagued Justin Trudeau, and tweets about corruption – making up 18% of all tweets. Topic 4 is predominantly tweets appealing to the middle-class and economy – and is 29.7% of all tweets; topic 5 contains celebratory messages about the campaign, as well as tweets regarding national holidays and days of remembrance – and make up 15% of all tweets. Topic 6 is made up of tweest about immigration, diversity and free speech – and makes up 11.5% of all tweets. Finally, topic 7 contains tweets regarding healthcare, abortion and pharmacare – and makes



(a) Topic 1



(b) Topic 2

(c) Topic 3

(d) Topic 4

(e) Topic 5

(f) Topic 6

(g) Topic 7

Figure 5.4: LDA Topic Word Clouds

up 1% of all tweets. The magnitude of how many tweets were assigned to each topic is shown in figure 5.5. The vertices representing tweets of different topics in figure 4.1 are assigned different colours.

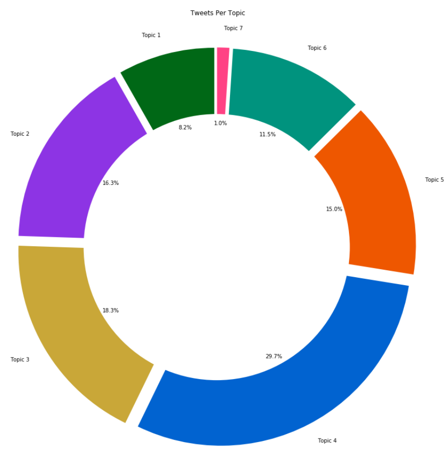


Figure 5.5: LDA Topic Distribution

5.2 Topic Centrality

5.2.1 Total Network Topic Centrality

Define what it is, show the full graph again.

5.2.2 Party Leader Topic Centrality

Now, consider a fact stating that relation `r` is total, i.e.,

5.2.3 Results

5.3 Stochastic Blockmodels

As described in section 4.2.1, stochastic blockmodels use edge densities across and within blocks to generate random graph models [11]. However, while users may belong to a certain "party leader block" or "topic block"[1] to some degree, these cannot be known a priori and users can be members of infinite combinations and mixtures of different blocks. As such, a novel method for generating stochastic blockmodels was developed that approximated edge densities and is specific to reproducing engagement graphs described in section 4, where certain users produce objects (tweets, songs, goods, etc...) and other users choose which ones to engage with. The algorithm takes in 6 parameters: n, $tweet\_dist$, k, m, $retweet\_histogram$, ϵ – which are described in table 5.1.

[1] In these cases, the blocks would be users who would only engage with a certain party leader or topic, or both.

Table 5.1: Adapted Stochastic Blockmodel Parameters

Parameter	**Type**	**Description**
n	*Integer*	The number of party leader vertices
tweet_dist	*Tuple*	A tuple representing a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ that is sampled n times to determine the number of tweet vertices per party leader
k	*Integer*	The number of topics that a tweet vertex can take on
m	*Integer*	The number of generic user vertices
retweet_histogram	*Histogram*[a]	The distribution of retweets that is sampled m times to get the degree for each generic user
ϵ	$Float \in [0-1]$	The proportion of the time a generic user will choose the greedy tweet-type[b] rather than choosing based off of the edge probabilities

[a] As of March, 2020 – this took the form as a Numpy histogram: a tuple containing the bin boundaries and corresponding densities.

[b] For the models generated in this thesis a tweet-type could be its topic, who tweeted it, or its topic *and* who tweeted it.

The algorithm generates the stochastic block model in three phases. First, all the party leader, tweet, and generic user vertices are generated based off of n, *tweet_distribution* and m, along with edges between the the party leader and tweet vertices. Second, topics are assigned to tweets randomly and m samples of the *retweet_histogram* are generated to determine the retweet degree for each user, D. Finally, the algorithm loops through each generic user and retweets D_i tweets. For each retweet a user makes, edge probabilities for each type of tweet are calculated based on the user's prior retweet history. The probability of forming an edge between user i and a tweet of type t, given edge probabilities e_i, is given by the policy[2] π in equation 5.1. This is a variation of the ϵ-*greedy* algorithm, where the tweet-type with the highest probability is chosen ϵ percent of the time, and $(1-\epsilon)$ percent of the time tweets are chosen based off of e_i. After each edge is formed user i's retweet history is updated.

[2] Policy in this context refers to the probability of choosing an action from some set of possible actions and is denoted by π; all other mentions to policy in this thesis refer to categories of messages.

$$\pi(t|e_i)\begin{cases}\epsilon & \text{if } t = argmax_t e_i \\ \\ (1-\epsilon)e_{it} & \text{otherwise}\end{cases}$$

While initially the edge probabilities for a user i would be proportional to their retweet history[3], section 5.3.1 demonstrates how more nuanced edge probabilities can be developed with deep stochastic blockmodels. The full algorithm is given in algorithm 1.

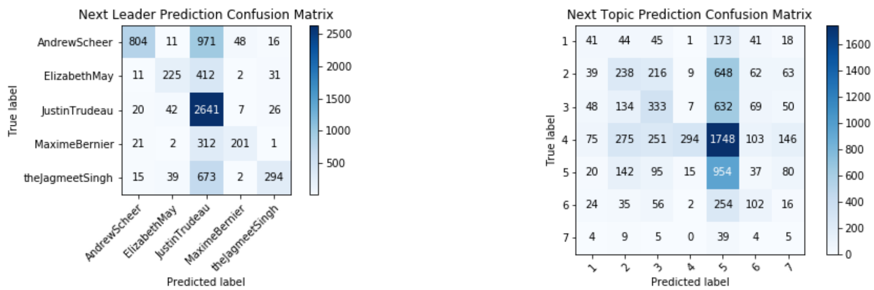
Algorithm 1: Stochastic blockmodel for modelling political engagement

---

```
Result: A stochastically generated political engagment graph
initialize G as an empty graph;
initialize T as an array of all different types of tweets;
initialize n party leader vertices in G;
initialize m generic user vertices in G;
for i ← 1 to n do
    initialize X ~ N(μ, σ²) tweet vertices in G;
    assign tweet vertices random topics from [1, ..., k];
    add edges from tweet vertices to party leader i;
end
generate retweet degree array D of size m by sampling retweet_histogram;
for i ← 1 to m do
    initialize user_history array of size |T| with all 0s;
    for j ← 1 to D_i do
        generate edge probabilities e_i;
        add an edge from user i to a tweet of type t based on π(t|e_i);
        user_history_t += 1
    end
end
return G
```

---

[3]This can be done efficiently by taking the $softmax$ of the retweet history.



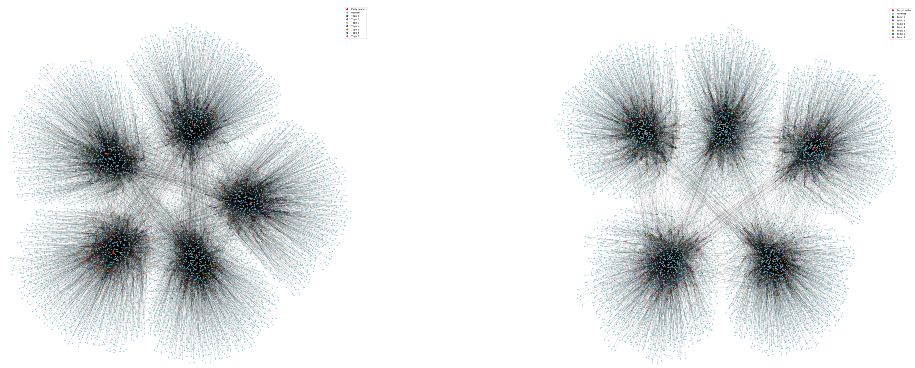
(a) "Next Leader" Confusion Matrix (b) "Next Topic" Confusion Matrix

Figure 5.6: ANN Confusion Matrices

5.3.1 Deep Stochastic Blockmodels for Modelling Political Engagement

As discussed in section 5.3, the edge probabilities for a user retweeting a tweet-type can be calculated in various ways. A further nuance that can be added is incorporating real world data into the edge probability calculation. Data used to generate the complete political engagement graph in figure 4.1 was used to train two feed-forward artificial neural networks (ANN) to aid in the edge probability calculation: one that predicts the next party leader a user would tweet given the prior party leaders they had retweeted, and one that predicts which topic a user would retweet given the previous topics of tweets that user had retweeted. Figure 5.6 shows the confusion matrices for these two ANNs. Both ANNs attained between 85-87% accuracy.

Sections 5.3.1, 5.3.1 and 5.3.1 will demonstrate both the utility of these ANNs in generating the stochastic blockmodels in comparison to the standard edge probability calculation.

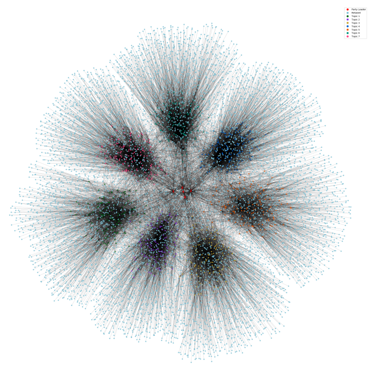


(a) Simple Stochastic Party Leader Blockmodel (b) ANN Adaption

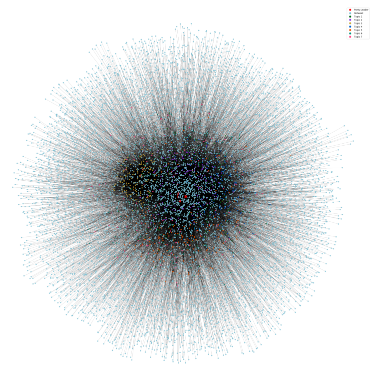
Figure 5.7: Stochastic Party Leader Blockmodels

Stochastic Party Leader Blockmodel

The stochastic party leader model generates user behaviour only taking into account the previous party leaders each user had engaged with prior. The tweet-types therefore are all the different party leaders that could be retweeted and e_i represents the weights of retweeting each party leader. For each user, when deciding which tweet they are to retweet, their retweet history is converted into a probability distribution (ex. $user\_history_i = [JT = 0, AS = 1, JS = 3, EM = 2, MB = 0]$ generates $e_i = [JT = 0.09, AS = 0.19, JS = 0.36, EM = 0.27, MB = 0.09]$). In this sense, it models a world in which politically engaged Twitter users only engage along the axis of party leaders, with a complete disregard for the topics tweeted about. Figure 5.7 shows two examples of stochastic party leader blockmodels: one in which edge probabilities are proportional to the number of times that user's retweeted each party leader, and one in which edge probabilities are determined with the ANN described in section 5.3.1.



(a) Simple Stochastic Topic Blockmodel



(b) ANN Adaption

Figure 5.8: Stochastic Topic Blockmodels

Stochastic Topic Blockmodel

Conversely, the stochastic topic blockmodel models a world in which politically engaged Twitter users only engage along the axis of topics. In this case, the tweet-types are the various different topics that a user can engage with. Here, the tweet-type history of a user is converted into a probability distribution, and with a probability of ϵ, that user will retweet the topic with the highest activation – regardless of which party leader tweeted it. Figure 5.8 shows two examples of stochastic topic blockmodels: in which edge probabilities are proportional to a user's topic retweet history, and one in which edge probabilities are determined with the ANN described in section 5.3.1.

Stochastic Hybrid Blockmodel

The final model developed is a hybrid of the stochastic party leader block model, and the stochastic topic blockmodel. Here, two history vectors for each user are captured – the n dimensional party leader history vector, and the k dimensional topic history vector. After each respective vector is converted into a probability distribution, the edge probability of topic i by party leader j is determined by some constant α[4] and the function:

$$edge\_probability(partyleader = i, topic = j) = \alpha P(i) + (1 - \alpha)P(j) \tag{5.1}$$

Where $P(i)$ is index i of that user's *party leader* probability distribution, $P(j)$ is index j of that user's *topic* probability distribution, and α is some constant that determines the relative weighting of the two. As α approaches 1, the hybrid model becomes equivalent to the stochastic party leader blockmodel – and as α approaches 0 the model approaches the stochastic topic blockmodel. This model then generates different "worlds" in which users' political engagement falls on the spectrum from only caring about *party leaders* to only caring about *topics*.

5.3.2 NetLSD for Describing Political Engagement

The final objective of comparing the relative importance of topics and party leaders in driving political engagement requires comparing the structure of target graph shown in 4, and various hybrid models generated with different values of α. Using the Network Laplacian Spectral Descriptor described in section 4.1.1 by Tsitsulin et al. the optimal alpha value can be determined in a scale-adaptive, size-invariant, and permutation-invariant manner [20].

Given the $O(n^3)$ complexity involved in calculating the eigenvalues of a graphs normalized Laplacian matrix it is advantageous to generate sufficiently small hybrid models when fitting them to the original engagement graph (see figure 4.1). To

[4]This has no relation to the parameter α referred to in section 3.2

determine how large a graph of this nature needs to be to capture its underlying structure – tweets of the original graph were sampled, keeping all the retweet edges, and then compared to the heat trace signature of the original graph. This is shown graphically in figure 5.9, the x-axis represents how many tweets *per party leader* were sampled from the original graph and the y-axis represents the L_2 distance from the original engagement graph's heat trace signature.

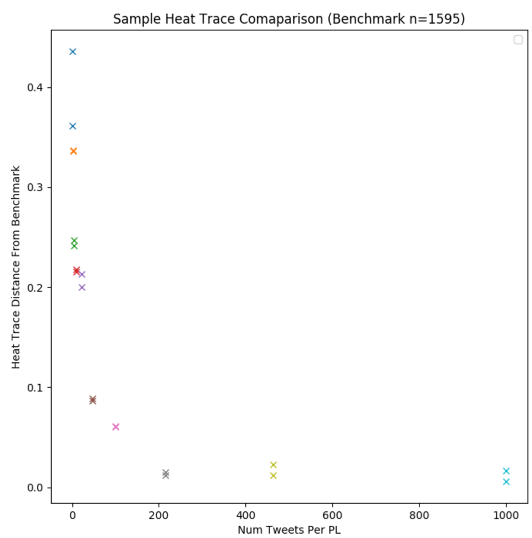


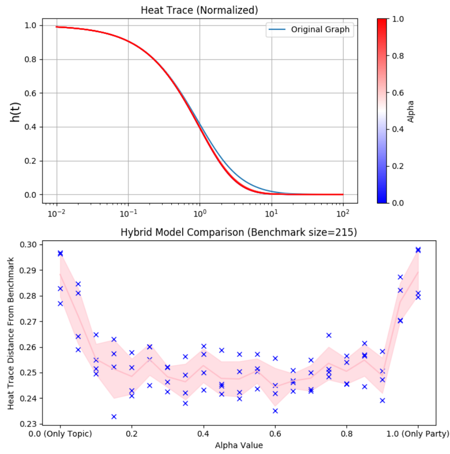
Figure 5.9: Heat Trace Signature Distance as a Function of Graph Sample Size

As can be seen from figure 5.9, there are diminishing returns for graphs with more than 215 tweets per party leader. Therefore, for each of the party leader vertices generated with the hybrid models – each one generated a number of tweets by a random variable X that is distributed normally where $X \sim \mathcal{N}(\mu = 215,\, \sigma^2 = 70)$.

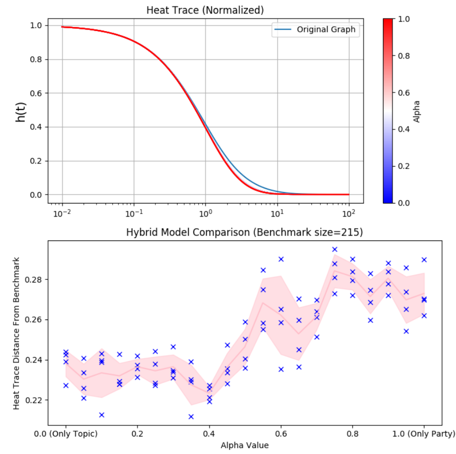
5.3.3 Results

The stochastic hybrid blockmodel (SHBm) gives the ability to define different "worlds" in which policy and party leaders drive engagement to different degrees. NetLSD gives the ability to compare the structural similarity of graphs in a scale-adaptive, size-invariant, and permutation-invariant manner. By performing a sweep of different α values for the hybrid models, and measuring which one produces the heat trace signature with the smallest L_2 distance to the original graph, the final objective of putting Ezra Klein's hypothesis to task can be achieved. All graphs were generated with $n = 5$, $k = 7$, $tweet\_dist = (\mu = 215, \sigma^2 = 70)$, $m = 8826$, $retweet\_histogram$ derived from the original engagement graph, and $\epsilon = 0.95$. α values ranging between 0 and 1, with increments of 0.05 were used to generate the different hybrid models. Four models were generated per α and the mean distance was used to determine the final distance.

The results of this experiment are striking: in both the deep SHBm, which uses the ANN to calculate edge probabilities, and the standard SHBm, which makes edge probabilities proportional to the user's retweet history, it is clear that both *what* the message is and *who* tweeted the message drive political engagement. The standard SHBm shows a clear trend where α values that privilege policy over party leader, or vice versa, have a further distance from original engagement graph than ones that blend the two. These hybrid models had the best fit when $\alpha = 0.6$. The deep SHBm had an optimal fit when $\alpha = 0.4$. The full tabular data is shown in table 5.2, and the results are visualized in figure 5.10.



(a) Standard SHBm L_2 Plots



(b) Deep SHBm L_2 Plots

Figure 5.10: Stochastic Topic Blockmodels

Table 5.2: Results of NetLSD Comparison of SHBm Models

	Standard SHBm			Deep SHBm		
α	Min L_2	Mean L_2	σ^2	Min L_2	Mean L_2	σ^2
0.00	0.2648	0.2756	0.0083	0.2173	0.2276	0.0062
0.05	0.2475	0.2603	0.0104	0.2111	0.2201	0.0073
0.10	0.2385	0.2440	0.0057	0.2030	0.2231	0.0117
0.15	**0.2225**	0.2403	0.0109	0.2176	0.2217	0.0061
0.20	0.2303	0.2375	0.0066	0.2210	0.2260	0.0037
0.25	0.2343	0.2439	0.0059	0.2172	0.2241	0.0067
0.30	0.2319	0.2374	0.0039	0.2207	0.2260	0.0058
0.35	0.2274	0.2355	0.0067	**0.2022**	0.2174	0.0095
0.40	0.2328	0.2418	0.0063	0.2096	**0.2134**	0.0031
0.45	0.2313	0.2370	0.0063	0.2181	0.2257	0.0067
0.50	0.2296	0.2368	0.0066	0.2255	0.2355	0.0085
0.55	0.2329	0.2396	0.0046	0.2440	0.2565	0.0116
0.60	0.2246	**0.2336**	0.0071	0.2249	0.2508	0.0187
0.65	0.2320	0.2358	0.0028	0.2260	0.2418	0.0125
0.70	0.2320	0.2369	0.0048	0.2260	0.2418	0.0125
0.75	0.2374	0.2425	0.0061	0.2610	0.2718	0.0079
0.80	0.2347	0.2394	0.0047	0.2603	0.2692	0.0062
0.85	0.2338	0.2436	0.0060	0.2484	0.2597	0.0081
0.90	0.2286	0.2378	0.0066	0.2604	0.2684	0.0058
0.95	0.2584	0.2654	0.0071	0.2432	0.2581	0.0111
1.00	0.2672	0.2765	0.0085	0.2506	0.2611	0.0098

Summary and Conclusions

6.1 Summary

6.2 Other Work

As dictated by the Queen's School of Computing, 25% of the evaluation for CISC 500 is left to be arranged between the student and their supervisor. Given the nature of this research, lab work or developing software appeared out of scope. I believe that one of the benefits of applied research is its ability to take abstract concepts and frameworks and show their utility in solving complicated problems. This took the form of performing a lecture on the project's research and the application of graphs in Dr. Robin Dawes' CISC 235: Data Structures course.

6.2.1 International Network for Social Network Analysis

The findings and methodologies proposed in this thesis will be presented at the International Network for Social Network Analysis' annual Sunbelt conference in Paris, France.

6.3 Future Work

* better ANNs * different contexts (UK, US, Chile, France) * more variables? * more efficient way of fitting the model

6.4 Conclusion

Bibliography

[1] *Annual report on Government of Canada advertising activities.* Public Works and Government Services Canada., 2018.

[2] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] R. Kenneth Carty and William Cross. Political parties and the practice of brokerage politics. *The Oxford Handbook of Canadian Politics*, pages 191–207, 2010.

[5] Fan RK Chung and Fan Chung Graham. *Spectral graph theory.* Number 92. American Mathematical Soc., 1997.

[6] Derrick L. Cogburn and Fatima K. Espinoza-Vasquez. From networked nominee to networked nation: Examining the impact of web 2.0 and social media on political participation and civic engagement in the 2008 obama campaign. *Journal of political marketing*, 10(1-2):189–213, 2011.

[7] Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj. *Generalized blockmodeling*, volume 25. Cambridge university press, 2005.

[8] Hewayda Elghawalby and Edwin R Hancock. Measuring graph similarity using spectral geometry. In *International Conference Image Analysis and Recognition*, pages 517–526. Springer, 2008.

[9] Linton Freeman. The development of social network analysis. *A Study in the Sociology of Science*, 1:687, 2004.

[10] Nathan Gardels and Nicolas Berggruen. *Renovating democracy: Governing in the age of globalization and digital capitalism*, volume 1. Univ of California Press, 2019.

[11] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[12] Steven Levitsky and Daniel Ziblatt. *How democracies die*. Broadway Books, 2018.

[13] Brian McNair. *An introduction to political communication*. Taylor & Francis, 2017.

[14] Yascha Mounk. *The people vs. democracy: Why our freedom is in danger and how to save it*. Harvard University Press, 2018.

[15] Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.

[16] David W. Nickerson and Todd Rogers. Political campaigns and big data. *Journal of Economic Perspectives*, 28(2):51–74, 2014.

[17] Garry Robins. A tutorial on methods for the modeling and analysis of social network data. *Journal of Mathematical Psychology*, 57(6):261–274, 2013.

[18] Ma Shiela C Sapul, Than Htike Aung, and Rachsuda Jiamthapthaksin. Trending topic discovery of twitter tweets using clustering and topic modeling algorithms. In *2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6. IEEE, 2017.

[19] Karen Stephenson and Marvin Zelen. Rethinking centrality: Methods and examples. *Social networks*, 11(1):1–37, 1989.

[20] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. Netlsd: hearing the shape of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2347–2356, 2018.