Under consideration for publication in Network Science

Measures of Topic Centrality for Online Political Engagement

Cameron Raymond
School of Computing, Queen's University, Kingston K7L 3N6, CA
(*e-mail:* c.raymond@queensu.ca)

Abstract

The advent of social media has enabled political parties to engage with the broader populous in new and unforeseen ways – and the ability to bypass the traditional mediating forces of mass media allows for an unfiltered promotion of policy, ideology and party stances. Drawing on Twitter data leading up to the 2019 Canadian Federal Election, this paper develops two novel, graph-based methods that capture how different categories of messages drive different patterns of political engagement. Through the two proposed variations of topic centrality – on which measures how central a topic was to the general discourse, and one which measures how central a topic was to a particular voting bloc – statistically significant variations in topic centrality are then shown and discussed.

Keywords: *centrality, political communication, social media, topic modeling*

Contents

1 Introduction 1
1.1 Social Media in the Canadian Context 2
2 Methods 3
2.1 Data 3
2.2 Topic Modeling 5
2.3 Topic Centrality 9
3 Results 9
3.1 Topic Saliency 9
3.2 Total Network Topic Centrality 9
3.3 Party Leader Topic Centrality 9
4 Discussion 9
References 9

1 Introduction

The way information is distributed and received has changed significantly over the past decade. As Cogburn and Espinoza-Vasquez argue, 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 (Cogburn & Espinoza-Vasquez, 2011). The same holds true for Canada, 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 (ann, 2018). Over the past 12 years, political elites have subverted traditional models of political communication by using social media to directly promote various policies, topics and issues to the electorate[1](McNair, 2017).

Additionally, it is important to note that not all messages promoted by political elites are likely to serve the same purpose. Some topics may be logistical in nature, informing party affiliates of campaign events; other topics may be promoted in an attempt to rally that party's core voting bloc; others, finally, may be an attempt to attract engagement from new, untapped demographics. The latter two categories are in many ways analogous to Robert Putnam's conception of social capital (Putnam, 2001). Here, Putnam draws the distinction between two forms of social capital: bonding social capital, which occurs within a group – and bridging social capital, which unites different demographics (Putnam, 2001). Therefore, the research question being proposed is: are their data to support the notion that some political messages are bonding in nature, rallying members within a group, while other political messages are bridging in nature? This question will be answered within the context of the 2019 Canadian Federal Election with the Tweets of Canada's five major, english speaking party leaders: Andrew Scheer, Elizabeth May, Jagmeet Singh, Justin Trudeau, and Maxime Bernier.

In order to answer this question, a justification of Canadian politics and social media data in this context will be given. Then an overview of the data collected and a formal definition of the political engagement graph used will allow for the exploration of two measures of topic centrality: total network topic centrality, and party leader topic centrality. Finally, results from this process and a discussion of their implications will highlight possibilities for future research.

1.1 Social Media in the Canadian Context

While it is clear that technology is changing how information is received, and thus also 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, it 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.(Carty & Cross, 2010)*

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

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 (Carty & Cross, 2010). 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" (Carty & Cross, 2010).

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 (Carty & Cross, 2010). The need to capture pluralities in a diverse range of electoral districts means that most parties have to take stances on most issues, and thus when a user engages with a specific issue, it doesn't necessarily invoke a specific party or vice versa.

Given the utility of Canadian politics in answering questions about different axes of political engagement, the question then is: how do we observe these phenomena? Social media data, culled from platforms like Twitter, are inherently relational – and thus lend themselves well to being represented as graphs. An empirical analysis that observes and measures how users behave and engage with political parties online privileges this 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" (Freeman, 2004).

2 Methods

2.1 Data

The novel dataset used was collected via Twitter's historical search application programming interface (API), which allows user's to programmatically access any publicly available Tweet. The API was used to collect all of the English[2] tweets from Canada's five, 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 Twitter 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 do analyze the bridging versus bonding nature of various topics. In this spirit, only Tweets of the party leader were used, excluding Retweets. Figure 1 visualizes the daily and cumulative number of Tweets over time, in aggregate and by party leader, resulting in 7,978 total Tweets. Additionally, for each tweet collected from a party leader, all of the available Retweets by general users[3] were collected for a total of 113,293 Retweets by 36,450 general users. This is, again, visualized in aggregate and by party leader in Figure 2.

[2] Denoted by a language marker in the historical search API.

[3] The term general users will denote those active on Twitter who are not party leaders.

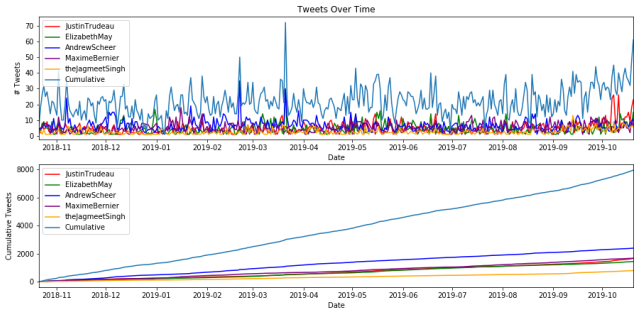


Fig. 1. Daily and Cumulative Tweets over Time

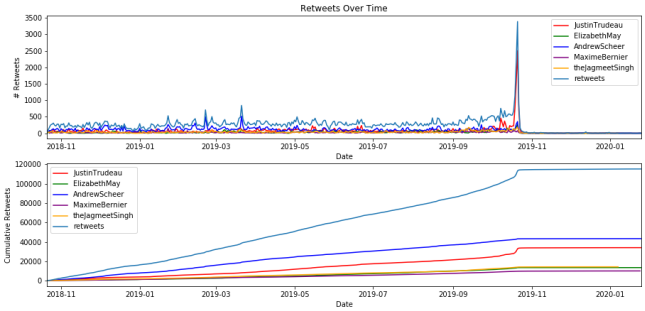


Fig. 2. Daily and Cumulative Retweets over Time

Additionally, given the inherent noise and extraneous info in text data, it is standard and necessary to preprocess text before modeling (Sapul *et al.*, 2017). 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 3 after preprocessing reads: *wherever maple leaf fly represents rich history bright future value hold dear happy flag day canada.*

Fig. 3. Daily and Cumulative Retweets over Time

2.1.1 Engagement Graph

The networks considered in this paper are assumed to be connected, unweighted, and undirected. Let (V,E) be a network, where V is the set of vertices and E is the set of edges. If vertex v_1 is connected to vertex v_2, it is denoted by $(v_1,v_2) \in E$. Engagement graph's are defined with additional constraints that denote three categories of vertices: those that *produce objects* (Tweets, songs, goods, services, etc…), those that represent the objects produced, and a third set of vertices that chooses to engage with the various produced objects in the network. In this context the engagement graph represents the Tweets that party leaders produce, and the general users who choose which Tweets to Retweet. An example of this type of political engagement graph is shown in figure 4. More formally the political engagement graph is defined 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; and let $V_3 = \{v_1, v_2, ..., v_k\}$ be the set of "general users" who Retweet Tweets. Let the total set of vertices $V = V_1 \cup V_2 \cup V_3$.
- *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 generic user vertex to a tweet vertex.

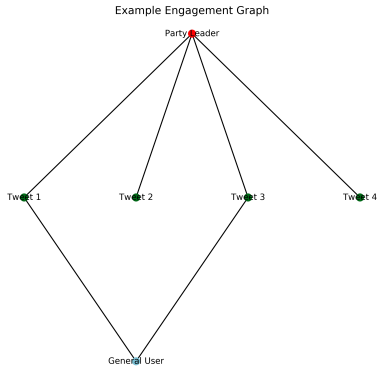


Fig. 4. Example Engagement Graph

Further nuances can be added to distinguish between different types of objects, which in this context would refer to Tweets of different topics. Figure 5 visualizes the full political engagement graph collected with all 5 party leaders, 7,978 Tweets, 36,450 general users, and 113,293 Retweets[4] built with these constraints.

2.2 Topic Modeling

In order to evaluate the relative importance of various political messages in driving political engagement on Twitter, all the tweets collected must first be organized by topic.

[4] The number of Retweets is equivalent to the number of edges from Tweet vertices to general user vertices.

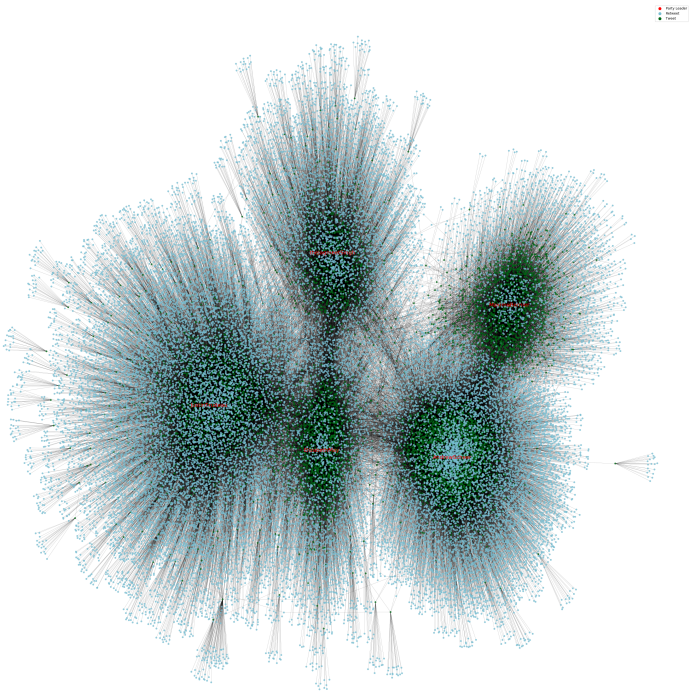
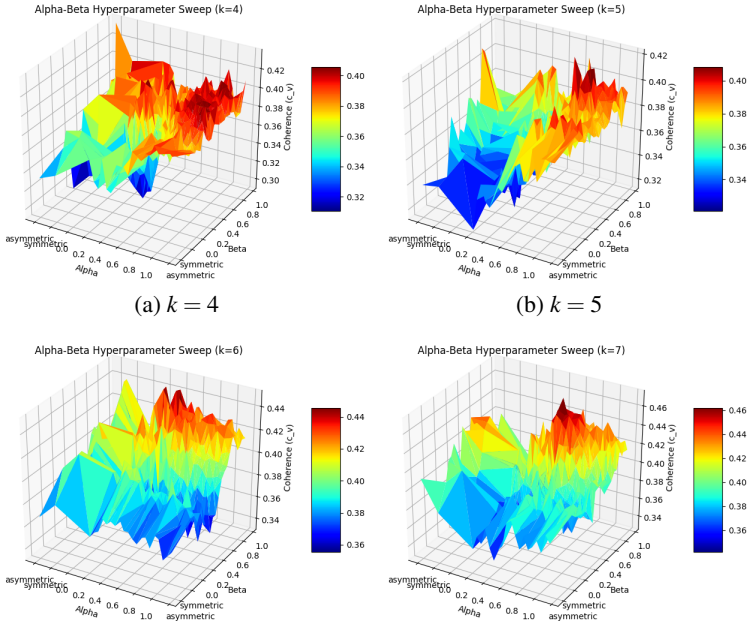


Fig. 5. Full Engagement Graph

Given the size of the data, manually coding each label is inefficient as well as subject to individual biases. As a result, there is a need for techniques that autonomously organize big, unclassified corpuses of text.

Topic modeling methods finds clusters of words that frequently occur together (topics), connects words with similar meanings, and distinguishes different uses of words with multiple meanings (Alghamdi & Alfalqi, 2015). This is based on the underlying assumption that a document is concerned with a fixed set of topics, and that the frequency of words used is indicative of this latent structure (Blei *et al.*, 2003). Given that topic extraction approaches based on keywords are brittle, context specific and are unable to capture emergent topics – unsupervised machine learning techniques, like the latent Dirichlet allocation (LDA) developed by Blei et al., are especially useful for autonomous topic extraction. An LDA model defines two distributions – one which models topics as "a distribution over a fixed vocabulary of terms," and another which models documents as a distribution of topics based on the occurrences of words in each document, the topic distribution per word and the importance of words to each topic (Blei *et al.*, 2003). The LDA requires four inputs: the text corpus – which in this context are the cleaned tweets from all 5 party leaders; α – which acts as a concentration parameter for how documents are modeled as topics; β – which acts as a concentration parameter for how topics are modeled as words; and k – which is the number of topics to be modeled (Blei *et al.*, 2003).

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, various LDAs were exposed to the entire corpus of Tweets and then evaluated using c_v coherence. Figure 6 shows, for each k value,

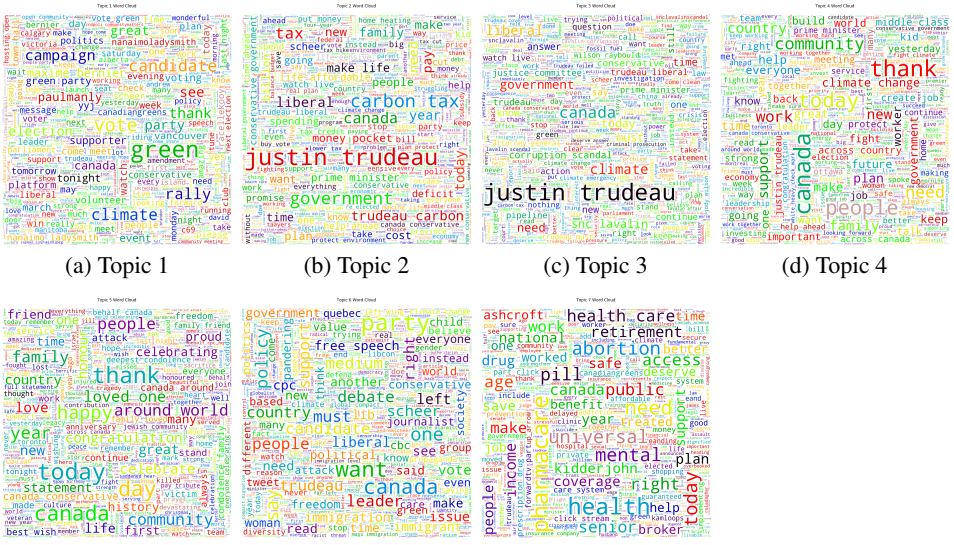


(a) $k = 4$ (b) $k = 5$

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

Fig. 6. LDA Parameter Sweep Results

the c_v coherence as a function of different combinations of α and β. 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 7.

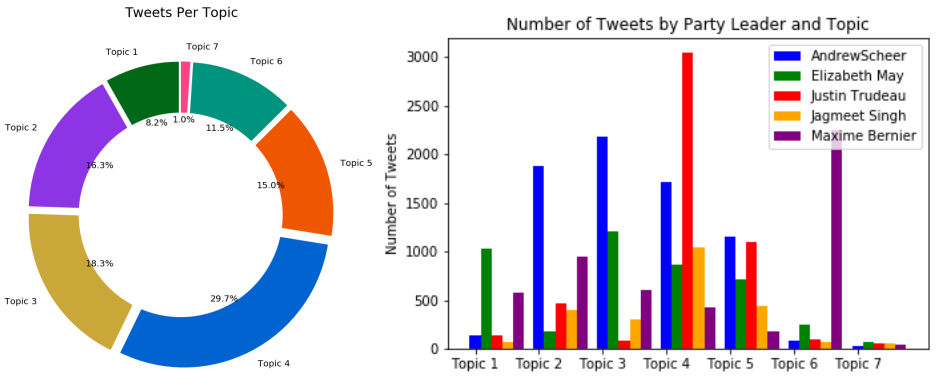


(a) Topic 1 (b) Topic 2 (c) Topic 3 (d) Topic 4

(e) Topic 5 (f) Topic 6 (g) Topic 7

Fig. 7. LDA Topic Word Clouds

Topic 1 pertained to campaign messages, rallies and logistics – and makes up 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 tweets 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 up 1% of all tweets. The magnitude of how many tweets were assigned to each topic is shown in Figure 8.



(a) Topic Breakdown (b) Number of Tweets per Topic by Party Leader

Fig. 8. LDA Topic Distribution

As can be seen by part (b) of Figure 8 – while Topic 6, Tweets pertaining to diversity and immigration, only made up 11.5% of all Tweets they made up the majority of Maxime Bernier – the leader of the newly founded, populist People's Party of Canada – Tweets. This signals an attempt on his part to bring these topics to the forefront of the debate despite few other party leader's broaching the topic. Additionally, Andrew Scheer and Elizabeth May's campaigns relied heavily on Topic 3, the SNC Lavalin affair, with a disproportionate number of Tweets aimed at Justin Trudeau's handling of the affair. Finally, it can be noted that Maxime Bernier and Andrew Scheer also had a disproportionate number of Tweets pertaining to Topic 2 – rebutting Justin Trudeau's plan to implement a carbon tax, and arguing for greater access to Alberta's Oil Sands. While an initial assumption may be that each party tailors such messages to maximize engagement, that remains to be seen without a more rigorous analysis of how different segments or the electorate as a whole engaged with each category of message. By assigning each Tweet vertex to a topic and then coloring the Tweet vertices accordingly, the full engagement graph (Figure 9) for online political content is complete and ready for analysis.

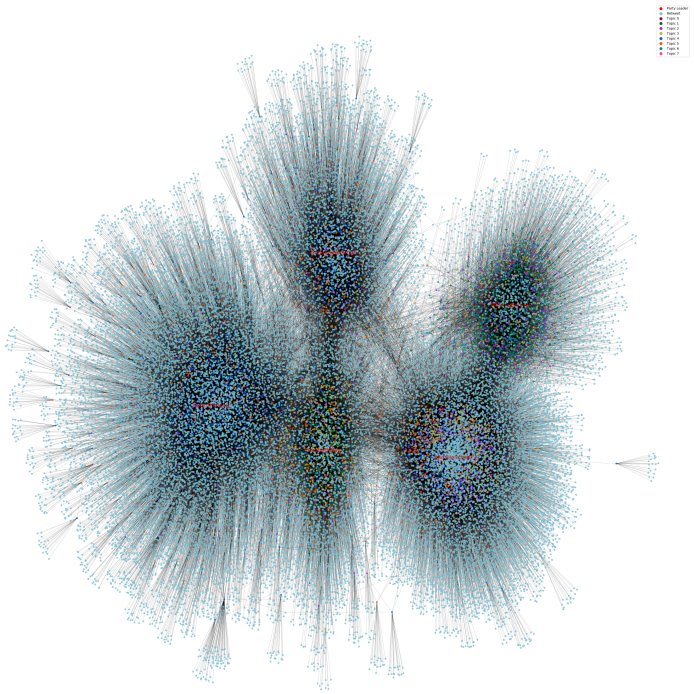


Fig. 9. Full Engagement Graph With Tweet Vertices Colored By Topic

2.3 Topic Centrality

2.3.1 Eigenvector Centrality

2.3.2 Total Network Topic Centrality

2.3.3 Party Leader Topic Centrality

3 Results

3.1 Topic Saliency

3.2 Total Network Topic Centrality

3.3 Party Leader Topic Centrality

4 Discussion

References

(2018). *Annual report on government of canada advertising activities*. Public Works and Government Services Canada.

Alghamdi, Rubayyi, & Alfalqi, Khalid. (2015). A survey of topic modeling in text mining. *Int. j. adv. comput. sci. appl.(ijacsa)*, **6**(1).

Blei, David M, Ng, Andrew Y, & Jordan, Michael I. (2003). Latent dirichlet allocation. *Journal of machine learning research*, **3**(Jan), 993–1022.

Carty, R. Kenneth, & Cross, William. (2010). Political parties and the practice of brokerage politics. *The oxford handbook of canadian politics*, 191–207.

Cogburn, Derrick L., & Espinoza-Vasquez, Fatima K. (2011). 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.

Freeman, Linton. (2004). The development of social network analysis. *A study in the sociology of science*, **1**, 687.

McNair, Brian. (2017). *An introduction to political communication*. Taylor & Francis.

Putnam, Robert. (2001). Social capital: Measurement and consequences. *Canadian journal of policy research*, **2**(1), 41–51.

Sapul, Ma Shiela C, Aung, Than Htike, & Jiamthapthaksin, Rachsuda. (2017). Trending topic discovery of twitter tweets using clustering and topic modeling algorithms. *Pages 1–6 of: 2017 14th international joint conference on computer science and software engineering (jcsse)*. IEEE.