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 a more direct 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 metrics of topic centrality – one 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 their implications for political polarization are discussed.

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

Contents

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)*

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 op-

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

posed 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.

1.2 Content as a "First-Class Citizen"

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). While there exists studies on user engagement in social networks, online or otherwise, such research generally focuses on connections between individuals (*User A* is connected to *User B*) and rarely acknowledges the context with which engagement occurs (Zhang *et al.*, 2017; Kavanaugh, 2002). In an online environment, certain users may engage with certain content producers when they produce certain forms of content, but choose not to engage with that same producer if the subject matter differs. As such, more nuanced graph models are needed and developed that acknowledge the importance of the producer of the content *as well as the content itself* in shaping engagement. Content holds an especially important role in online mediums, where users have a more granular ability to control the content they expose themselves to. This theoretical underpinning allows for a more nuanced analysis and in doing so makes content a "first-class citizen."

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

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

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.

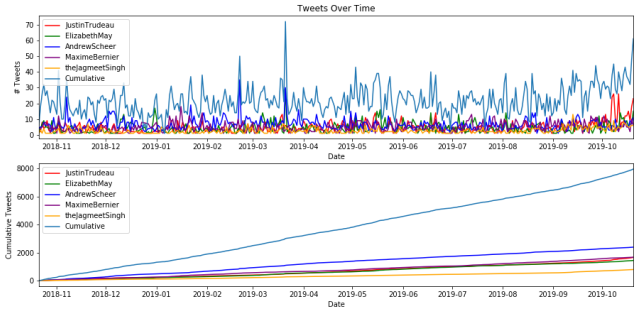


Fig. 1. Daily and Cumulative Tweets over Time

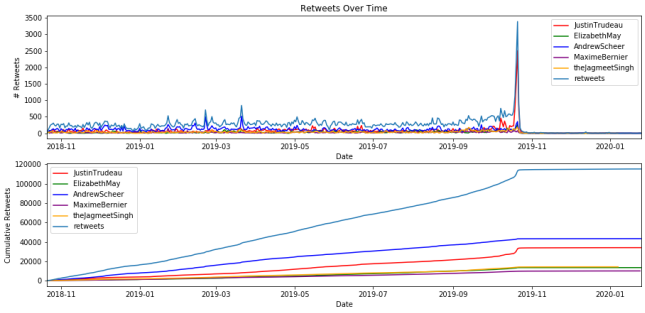


Fig. 2. Daily and Cumulative Retweets over Time

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

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.

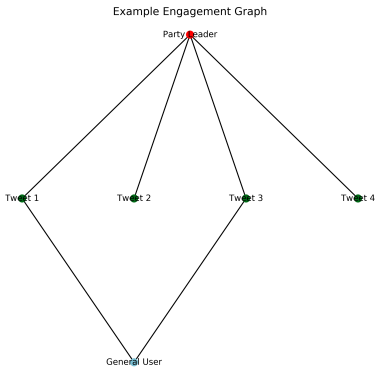


Fig. 4. Example Engagement Graph

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.

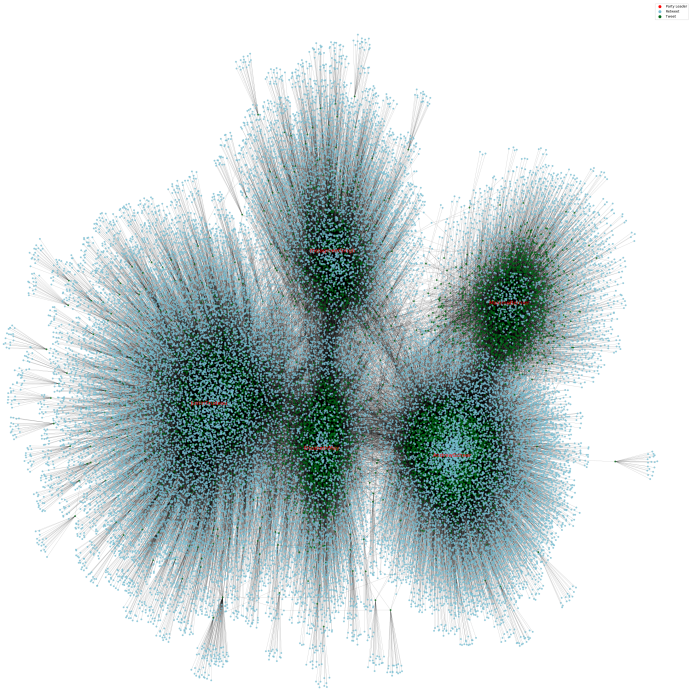


Fig. 5. Full Engagement Graph

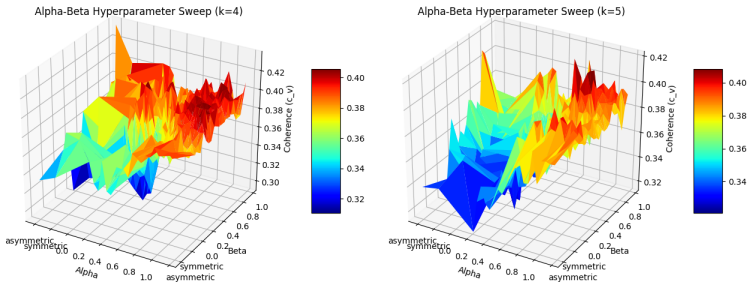
2.2 Topic Modeling

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. 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

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

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).



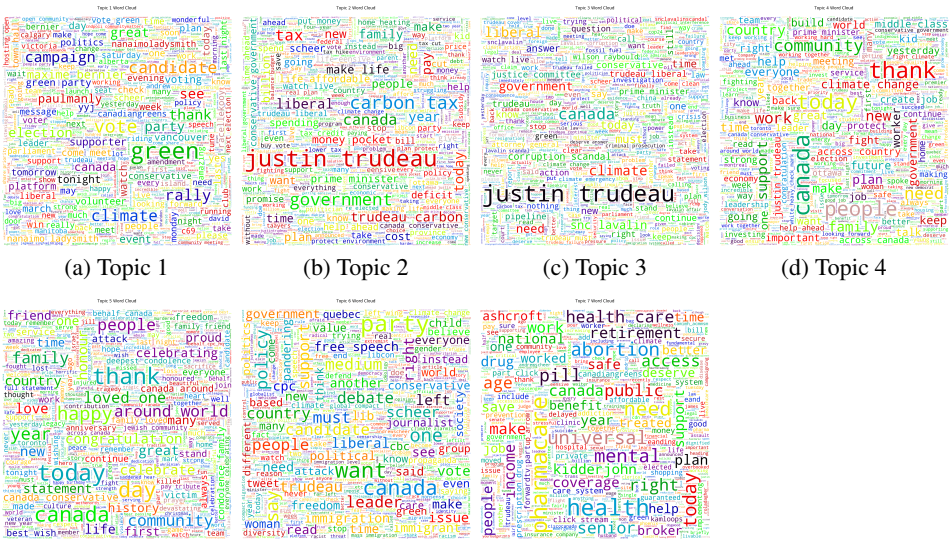
(a) $k = 4$ (b) $k = 5$

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

Fig. 6. LDA Parameter Sweep Results

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, 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.

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

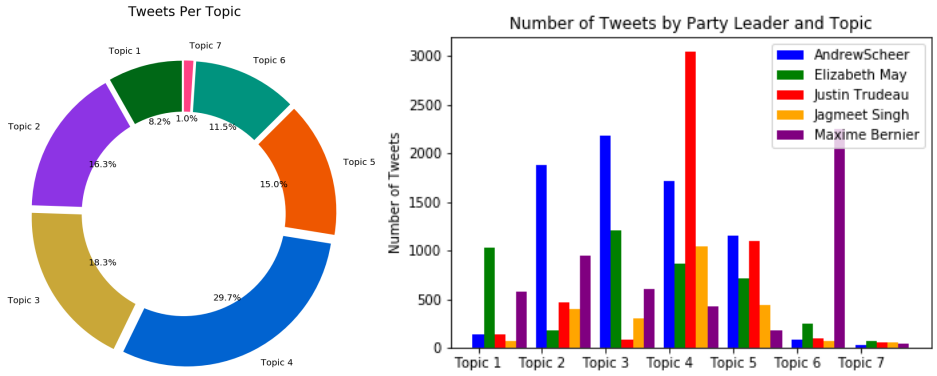


(a) Topic 1 (b) Topic 2 (c) Topic 3 (d) Topic 4

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

Fig. 7. LDA Topic Word Clouds

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's 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.

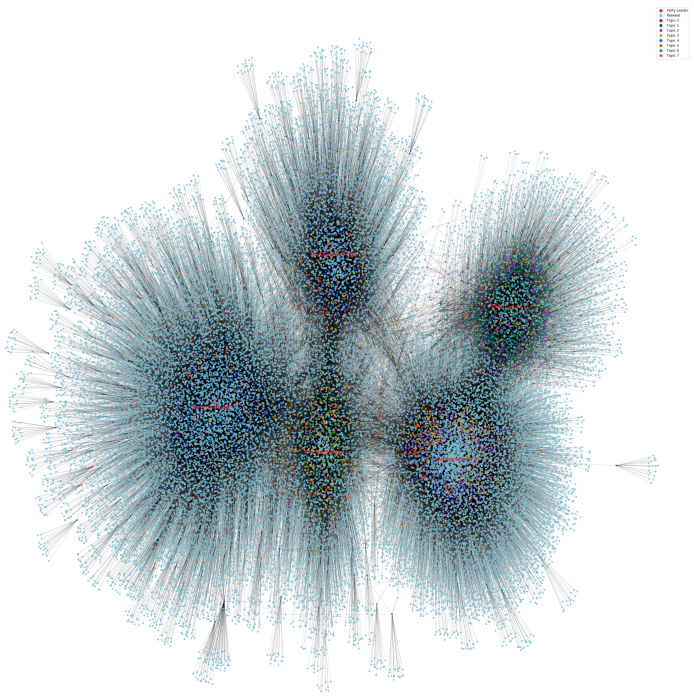


Fig. 9. Full Engagement Graph With Tweet Vertices Colored By Topic

2.3 Topic Centrality

Descriptive statistics in Figure 8 show the frequency with which Tweets of different topics were promoted by Canadian party leaders leading up to a major election. However, these Figures say little about how different policies rallied groups of existing supporters, or bridged different voting blocs. In order to answer such questions – a measure of vertex centrality, eigenvector centrality, is expanded upon. The result are two variations of topic centrality: total network topic centrality and party leader topic centrality. By juxtaposing the two, various insights about how different topics influenced discourse can be derived.

2.3.1 Eigenvector Centrality

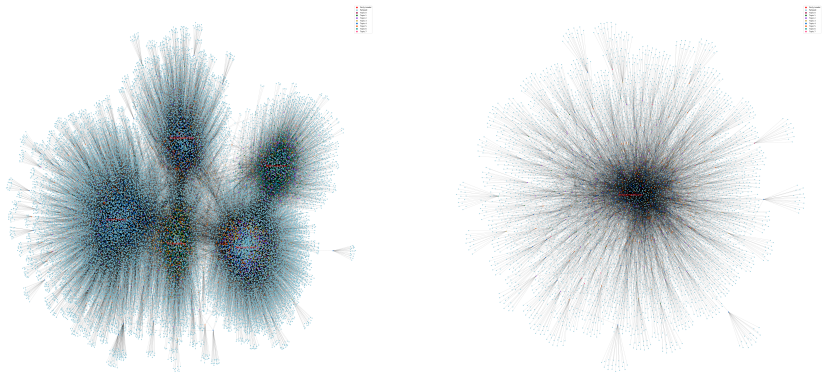
Research concerning the centrality of vertices in a graph has successfully been applied to problems in marketing, economics and epidemiology; Stephenson and Zelen explored the utility of centrality measures in studying the social dynamics of Gelada baboons (Stephenson & Zelen, 1989). Common centrality measures include measures of degree and betweenness. This article 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.

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." (Newman, 2008) The eigencentrality of vertex i given an adjacency matrix A is defined as x_i where x_i is proportional to the average eigenvector centrality of i's neighbors:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j \tag{1}$$

where λ is some constant. By defining the vector of centralities as $\vec{x} = (x_1, x_2, \ldots)$ this equation can be rewritten as $\lambda \vec{x} = A \cdot \vec{x}$, and it is evident that $\vec{x}$ is an eigenvector of the adjacency matrix with eigenvalue λ (Newman, 2008). By Perron-Frobenius theorem, picking the largest eigenvalue of A will result in all elements of $\vec{x}$ being non-negative (Newman, 2008).

2.3.2 Total Network Topic Centrality



(a) Full Engagement Graph (b) Political Engagement Subgraph for Jagmeet Singh

Fig. 10. Example Engagement Graph and Subgraph

Total network topic centrality is an aggregate of the eigenvector centrality of all tweets of a certain topic in the entire engagement graph. For reference, this graph has been reproduced in part (a) of Figure 10. More formally, and with slight abuse of notation,

the total network topic centrality of topic t, T_t, can be defined as the set of all eigenvector centrality measures for tweet vertices of topic t in a graph G.

$$T_t := \{x_i, \forall i \in G \mid type(i) = tweet, topic(i) = t\} \tag{2}$$

When taking the aggregate (sum, mean, z-score relative to other topics), a single number can be assigned to the relative importance of topic t. And in keeping with the strengths of eigenvector centrality, T_t will be larger if: those tweets are retweeted a lot, and if they're retweeted by *highly engaged* users. This provides a nuanced metric for how central different issues were to the entire online discourse.

2.3.3 Party Leader Topic Centrality

In order to measure how central a topic is to a party leader's base, party leader topic centrality was developed. This assumes a world in which party leader j is the only actor with which generic users can engage with. This is done by taking a subgraph of $G - G_j$ – which only contains party leader j, all of j's tweets, I, and all edges from $i \in I$ to the generic users who retweeted one of j's Tweets. An example of this is in part (b) of Figure 10, which demonstrates the subgraph for Jagmeet Singh.

After the subgraph G_j is constructed, the party leader topic centrality P_{tj} is defined as the set of all eigenvector centrality measures for tweet vertices of topic t in a graph G_j.

$$P_{tj} := \{x_i, \forall i \in G_j \mid type(i) = tweet, topic(i) = t\} \tag{3}$$

Similarly, by taking the aggregate of P_{tj} a single centrality score can be assigned that measures how important topic t was to those who engaged with that party leader. This will be higher if it is: has a high number of Retweets and retweeted by j's *most engaged followers*. Highly engaged users who rarely engage with that specific party leader are therefore discounted in this metric relative to someone who is less engaged, but concentrates their attention on that party leader's content.

3 Results

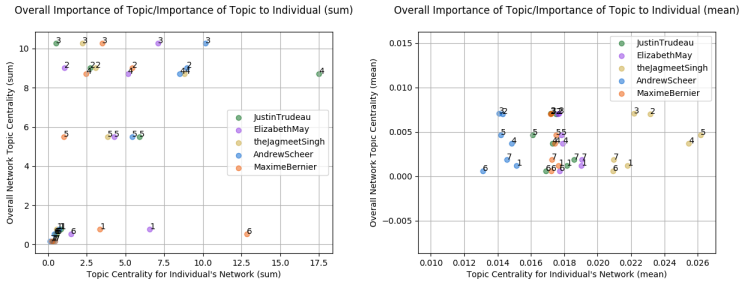
After calculating $\mathbf{T} = \{T_t, \forall i \in topics\}$[5] and $\mathbf{P} = \{P_{tj}, \forall i \in topics and \forall j \in party\_leaders\}$[6], each individual centrality measure, $T_t \in \mathbf{T}$ and $P_{tj} \in \mathbf{P}$, can be summated ($\hat{T}_t/\hat{P}_{tj}$), have it's mean calculated ($\overline{T}_t/\overline{P}_{tj}$), or have it's z-score taken relative to other topics in its set[7] ($\overset{z}{T}_t/\overset{z}{P}_{tj}$).

Figure 11 shows plots of $\mathbf{T}$ as a function of $\mathbf{P}$; each point is colored according to its party leader and is annotated with the topic that it pertains to. The x-axis is the party leader

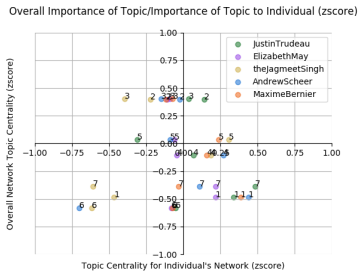
[5] *topics* refers to the seven topics that tweets could be assigned to.

[6] *party_leaders* refers to the five Canadian Federal party leaders whose Tweets were collected.

[7] For the party leader topic centrality, the z-score of topic centrality is based off of other tweet topics for that party leader.



(a) $\hat{T}$ by $\hat{P}$ (b) $\overline{T}$ by $\overline{P}$



(c) $\overset{z}{T}$ by $\overset{z}{P}$

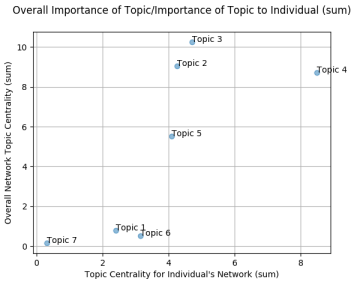
Fig. 11. **T** aggregates in relation to **P**

topic centrality score for all party leader and topic combinations – and the y-axis is the total network topic centrality score for each topic[8].

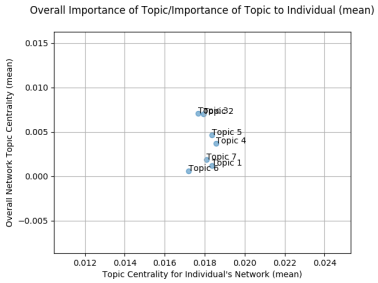
Some initial insights from comparing $\hat{T}$ by $\hat{P}$ with $\overline{T}$ by $\overline{P}$; while Maxime Bernier promoted a large number of tweets regarding immigration, free speech and diversity as indicated by a high $\hat{P}_{6M}$ – these tweets gained little traction overall in the entire network ($\overset{z}{T}_6$) as well as compared to other topics Bernier tweeted about $\overset{z}{P}_{6M}$. Additionally, despite the relatively few tweets pertaining to health care and pharmacare (topic 7), and the low total network centrality score ($\overset{z}{T}_7$), they were disproportionately important to Justin Trudeau's network ($\overset{z}{P}_{7T}$). Discrepancies in a topic that gives the highest $\hat{P}_{tj}$ for a party leader (total topic engagement) and $\overline{P}_{tj}$ (mean topic engagement) can illustrate inefficiencies in connecting with their target demographic.

As well, for each topic – the party leader topic centrality score can be averaged across all the different party leaders to give a more general analysis of how central topics were to the entire network, or to any individual party leader. This is shown in Figure 12. Here it is most insightful to look at $\overset{z}{T}$ by $\overset{z}{P}$, specifically in the lower right-hand quadrant and upper left-hand quadrant. The former indicates tweet topics that are more important to a party leader's base than to the entire network (topics 7 and 1). Topic 1 – campaign dynamics – intuitively makes sense in this category; it is not surprising that messages about the campaign, where rallies are, etc... would appeal more to a party leader's base than to the

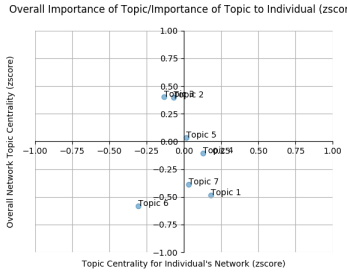
[8] This explains why all party leader topic centrality scores for the same topic lie on the same point on the y-axis.



(a) $\hat{T}$ by $\hat{P}$



(b) $\overline{T}$ by $\overline{P}$



(c) $\overset{z}{T}$ by $\overset{z}{P}$

Fig. 12. **T** aggregates in relation to **P** (party leader average)

entire network. Conversely, tweets in the upper left-hand quadrant indicates tweet topics that are more important to the overall network than to the individual network – which may be an indication of those topics spanning partisan divides (topics 2 and 3).

Topic	$\hat{T}$	$\overline{T}$	$\overset{z}{T}$	$\hat{P}$	$\overline{P}$	$\overset{z}{P}$	$\hat{P}$	$\overline{P}$	$\overset{z}{P}$
1	0.779	0.001	-0.486	0.726	**0.015**	**0.438**	**6.566**	**0.019**	0.216
2	9.034	**0.007**	0.395	8.947	0.014	-0.024	1.074	0.017	-0.110
3	**10.263**	**0.007**	**0.402**	**10.176**	0.014	-0.148	7.097	0.018	-0.090
4	8.719	0.004	-0.106	8.500	**0.015**	0.269	5.156	0.018	-0.043
5	5.507	0.005	0.033	5.425	0.014	-0.088	4.278	0.018	-0.061
6	0.519	0.001	-0.580	0.380	0.013	-0.703	1.490	0.018	-0.082
7	0.153	0.002	-0.387	0.145	**0.015**	0.107	0.438	**0.019**	**0.218**

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.

Kavanaugh, Andrea L. (2002). Community networks and civic engagement: A social network approach. *The good society*, **11**(3), 17–24.

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

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

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.

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

Zhang, Fan, Zhang, Ying, Qin, Lu, Zhang, Wenjie, & Lin, Xuemin. (2017). Finding critical users for social network engagement: The collapsed k-core problem. *Aaai*.