

The multiple facets of influence: Identifying political influentials and opinion leaders on Twitter.

April 30, 2013

250 word abstract

Keywords: Twitter, Online Communities, Influence, Social Media, Social Network Analysis

INTRODUCTION

The debate about whether or not the Internet is empowering individuals has been central to studies of political communication and political discussion communities, with some arguing significant increases in citizen power are likely (Blumler & Gurevitch, 2001; Dahlgren, 2000, 2005), and others cautioning against such hopes (Hindman, 2009). However, the importance of examining changes in political systems given the increasing pervasiveness of digital technologies is largely accepted (Agre, 2002). Digital technologies are making it possible for new players to become involved in political decisions (Dubois & Dutton, Forthcoming 2013). Some have described the political environment as consisting of an “ecology of actors” (Dutton & Peltu, 2007) which includes politicians, journalists, activists, and average citizens, interacting in ways that lead to varying political outcomes.

The ability to influence, that is to convince another individual to change their opinion, attitude, and/or behaviour is a potentially powerful skill. Broadly there are two ways an individual’s ability to influence others on political issues have been theorized: the “opinion leader” who uses social support and social pressure to influence their personal network (Katz & Lazarsfeld, 2006); and the influential who uses their visible position in a large network to spread messages widely (Rogers, 2010). The former speaks to individuals who occupy an important role in political talk and discussion communities which is thought to be integral to democratic integrity (J. P. Dillard, 1989; S. C. Dillard J. P. & Harden, 1989; Mutz, 2006). The later has been tied to work on diffusion of innovations and information cascades, for example: (Rogers, 2010; Bakshy, Hofman, Watts, & Mason, 2011; Lerman & Ghosh, 2010).

Our study relies on both interpretations of who is influential in political networks in order to answer the question: Which political players are the most influential within the two largest Canadian political communities on Twitter?

In a review of literature we trace the theoretical history of opinion leadership and influentials before identifying the most common ways these influential individuals have been identified in past Twitter research. Using online social network analysis and content analysis of tweets, measures of network centrality, interaction, knowledge and local embeddedness in the network present as important measures. We examine how each metric for identification compares to each other in terms of the way all individuals are ranked. We then take a more in-depth look at the top 20 most influential users identified according to each metric and examine what kinds of political players are considered most influential. Last, we consider a specific subset of each original network which excludes a selection of the most popular accounts. This allows us to identify users who are embedded in close-knit social groups but who also pay attention to popular accounts like media outlets, journalists and politicians. We call these users locally influential and suggest they are possible “opinion leaders.”

We have selected the case of Canadian political communities on Twitter for a number of reasons. The Canadian political system is well established and has similarities to many other countries like the UK and Australia in which the digital political environment has been more extensively investigated (to name only a few: (L. W. Gibson R. K. & Ward, 2005; Coleman, 2004; R. K. Gibson & Ward, 2002)). Further, Canadians are among the highest Internet users in the world (Duong, 2012) making this an instructive case for understanding political communities in a digital environment despite the fact that few have chosen to focus on the Canadian case.

Twitter has been selected as it is a popular site among Canadians, provides a clear set of boundaries for data collection, and has been studied by many others in the past which provides an instructive base from which to build our understanding of influence in political communities.

Finally, the Canadian political Twittersphere is large enough to provide sufficient data but small enough to conduct meaningful qualitative analysis which serves as a baseline for comparison. The use of two distinct hashtags, #CPC which stands for the Conservative Party of Canada (Government) and #NDP which stands for the New Democratic Party of Canada (Official Opposition), provides opportunity to compare two different communities within the Canadian Twittersphere at the same time ultimately lending reliability to our study (Yin, 2008).

Our findings suggest that when considering network-wide placement, those who traditionally hold political power—media and politicians—maintain such power. When indicators of expertise and interaction are considered, bloggers and political commentators gain prominence in the list of highly ranked users. Guided by the Two-Step Flow Hypothesis, we propose and test a way to identify those who are influential at a local level. These results are of course specific to this case, but the broader notion that various measures of influence, i.e. differing operational definitions of influence, in fact identify very different kinds of influencers, is an important point researchers can benefit from when defining influence and selecting measures for future studies. Our discussion of local level measures of influence as they relate to the notion of opinion leaders provide direction for future studies of political discussion communities.

Defining Influence

In early work Katz and Lazarsfeld conceived of the opinion leader who is able to influence their close personal ties. “Opinion leaders” influence those in their social circle by exerting social pressure and social support. These opinion leaders are important political players because they transmit messages to a wider public who are not tuned into messages from political elite. They are knowledgeable and trusted on a specific topic (Katz & Lazarsfeld, 2006). The hypothesis suggests four core facets of being influential: having a following, seen as an expert, knowledgeable/have expertise, and are in a position within their local community to exert social pressure and social support/social embeddedness.

The hypothesis was developed in response to theories of media effects which suggested the mass media directly influenced the opinions, attitudes, and behaviours of the general public. Their suggestion was that the mass media influenced a small segment of the population, the opinion leaders, who then influenced the wider public. The hypothesis has been tested in multiple settings (Katz, 1957), modified to include multiple steps (Weimann, 1982; Robinson, 1976), combined with other theories of media effects like agenda-setting (Brosius & Weimann, 1996), and applied recently in a digital media environment (2008).

Later work took these ideas in two divergent directions. The first is tied to political discussion communities where the notion of strong personal ties has remained crucial but the focus on identifying influential individuals has diminished (Eveland, Morey, & Hutchens, 2011; Huckfeldt, Sprague, Kuklinski, Wyer, & Feldman, 1995). Recent work tends to examine the quality of political discussion in terms of whether or not it is democratic (Huckfeldt et al., 1995; Mutz, 2006). That said, some have called for renewed focus on who discussants are and how they relate to one another (Eveland et al., 2011).

The second area which builds from the foundation of the Two-Step Flow is concerned with how messages flow through social networks. The main idea brought forward from early research is that the effects of a message can be traced from its sender, originally mass media but now whichever message creator is isolated, through multiple individuals within a chosen network. This has been an area of great interest in multiple areas, for example, market researchers interested in finding out who is best placed in a network to reach a very wide following CITE (Bakshy et al., 2011; D. Watts & Dodds, 2007; Hill, Provost, & Volinsky, 2006), and social movement scholars interested in how large networks self-organize without formal structures (González-Bailón, Borge-Holthoefer, & Moreno, 2012).

Despite their differences in approach, these two broad theoretical backdrops for identifying influentials remain important in the study of the various players in online political discussion communities. That is because political discussion is happening in a hybrid news and media environment in which various players all have access to a variety of tools with which to communicate with each other (Chadwick, 2011). Politicians, journalists, activists, bloggers and average users may use multiple strategies, like broadcasting, narrowcasting, asking and responding to questions, posting informative links, etc. to interact and discuss political issues. The context in which one is influential is much less certain in an online hybrid media environment than it was when theories of influence and political communication were first articulated.

This means that assigning someone the descriptor “influential” or “opinion leader” can be prob-

lematic because it is harder to identify traceable practices, specific tools or strategies or even structures of social connections which are necessarily unique to influencers. To address this problem researchers use a range of different metrics to identify influentials. Most advocate for a certain combination of metrics (Quercia, Ellis, Capra, & Crowcroft, 2011; González-Bailón et al., 2012; Lee, Kwak, Park, & Moon, 2010; Chang, Wang, Mei, Liu, & Park, 2013). Regardless, the end result is multiple ways to operationalize the notion of being influential based on the inclusion or exclusion of various facets of influence.

Measuring Influence

The most labor intensive way to identify influentials is to ask people who they are influenced by and if they feel they are influential. This is the method used in most studies of the Two-Step Flow of communication and related work (Katz, 1957). Done via interview or survey, the process is time consuming and limited in scale. But the presence of social media and online social networking sites has meant that much trace data exists from which networks of influence can be constructed (Welser, 2008).

The majority of recent studies making use of the opportunities afforded by this enhanced access to large scale social data have assumed a more easily quantifiable definition of influence. That is they tend to take the view of influence described by (Rogers, 2010) and others who look at number of followers and/or how far a message from a given users travels as indicators of influence (Bakshy et al., 2011; Wu, Hofman, Mason, & Watts, 2011; Ye & Wu, 2010; Rattnaritnont, Toyoda, & Kitsuregawa, 2012).

Among studies concerned with identifying those capable of influence on Twitter two dominant approaches exist. The first is concerned with users' placement within the network. These studies borrow tools from social network analysis to compute scores for degree, in-degree, out-degree, eigenvector centrality and the related PageRank, and betweenness centrality. These studies assume that when a node is placed in a network where they could be heard by many others in that network, that node is likely to be the most influential. Some examples of this are: (Bakshy et al., 2011; Brown & Feng, 2010; Anger & Kittl, 2011; Ye & Wu, 2010; Wu et al., 2011; Bigonha, Cardoso, Moro, Gonçalves, & Almeida, 2011; Weitzel, Quaresma, & Oliveira, 2012; Subbian & Melville, 2011). In essence, the facet of influence these studies rely on in order to provide an operational definition is having a following.

The second approach is to consider interaction between members of the network. This can be done by composing a network in which links connecting nodes are indicative of interaction, for example a re-tweet network, and applying measures mentioned above. Alternatively in a follower-follower network, or mutuals network, by assigning a value to each node based on how many times they were mentioned/their content was reposted by anyone else in that network, for example: (Sousa, n.d.; Overbey, Paribello, & Jackson, 2013). In recent studies examining cascades it is common to track how far a given user's content is diffused through the network as a measure of that users influence. The main facet of influence these studies are concerned with is being treated like an expert.

Content analysis provides an alternate option for identifying influentials. Though some use complex methods, like ranking quality of language or tracking urls over time, for assigning levels of influence to individuals in a given network based on content (Bakshy et al., 2011; Al-sharawneh, Sinnappan, & Williams, 2013; Ye & Wu, 2010), the simplest base form is the counting of keywords used in an individual's body of tweets. This method of identifying influentials is less common and also less transferable given that content produced by members of different networks is likely to vary quite widely. That said, it is a potentially useful measure for identifying opinion leaders since it is possible to assess a user's knowledge/expertise on a subject based on the content of their tweets. This approach theoretically addresses a similar facet of influence to the interaction measures noted above but instead of considering how a user is treated by their audience it considers the quality of messages a user sends, ie. knowledge and expertise.

Finally, while a small number of studies have used some of the measures mentioned above in order to identify influentials within smaller networks, for example a single personal network (Gruzd, Wellman, & Takhteyev, 2011), identification of influentials based on social embeddedness in their local community has been largely absent. This is problematic for those wishing to identify those political players most likely to influence individuals among the general public who are less tuned into messages coming from political elite – the opinion leaders.

The local clustering coefficient is one potential social network analysis measure which could provide insight. The measure scores nodes in terms of the degree of completeness of the graph between their immediate neighbors – a fully connected network receives a score of 1 and is a complete sub-graph, while a graph where none of these neighbors are connected (beyond the connection they have with the node in question) receives a score of zero (D. J. Watts & Strogatz, 1998). Despite only being used in a handful of studies, for example (Bigonha et al., 2011; Sousa, n.d.; Hill et al., n.d.), this information could be particularly useful to those interested in identifying opinion leaders and political discussants who are locally influential within their immediate group.

Evidently, we have a range of different theory driven ways in which we define and operationalize influence but little sense of how they empirically play out or how they compare to one another. As such, in addition to our research question concerning who is most influential within our chosen network we ask: Do different operational definitions of influence actually measure the same general trends when considering a Twitter network constructed in this way, and if not, how do they differ and why?

Accessing political discussion communities

While previous studies opt for very large data sets wherein as few restraints as possible are placed whilst sampling (Bakshy et al., 2011; Cha & Gummadi, 2010; Ye & Wu, 2010), this work instead chooses to focus in on a smaller case study in order to assess the networks of interest more exhaustively. In doing so, this work aims to highlight different facets of influence in a given context so as to provide a more thorough sense of the types of users who are influential, and how the different facets relate to one another.

What is particularly interesting to us is how influencers within specific political discussion communities are identified. Some have looked at specific communities within wider sets (eg. (Cha & Gummadi, 2010) selected three topics). But this approach only works if the wider set is the complete list of those in the smaller community, which is rare given technical constraints. Further, based on the theory laid out above, we are interested in users within a very specific group. Finally, identifying everyone in a given community helps to contextualize findings by allowing us to specify issue and time period and helps us understand the strengths and limitations of various methodological approaches to identifying influentials on social media platforms.

Given the importance of context to identifying influence we set a specific topic, Canadian politics, and collected tweets from a specific time period, March 12th and March 26th, 2013. Within Canadian politics we chose two distinct online communities as denoted by the use of the #CPC and #NDP hashtags. These are the prominent hashtags that the Government (CPC) and Official Opposition (NDP), respectively, organize around. The Liberal Party of Canada (LPC) which uses the hashtag #LPC, while still a major political party in Canada, was not included because we felt adding a less popular third party would not serve to better answer our research questions. Further, at the time of data collection the LPC were going through a leadership race which meant that their hashtag was being used for different purposes making it less comparable to the other two parties' hashtags.

Between March 12th and March 26th, 2013, a Twitter Streaming API ¹ connection was established, selecting tweets that matched the hashtags #CPC and #NDP in order to create a base index of users of interest. After the period of collection was finished, a network graph was generated where connections between users who followed one another were used as the edges between the nodes, or the individual users in the dataset. Additionally, up to a maximum of 200 recent tweets were collected for every user in the dataset in order to assess their interaction with the rest of the community compared to their general behavior. Table 2 provides a summary of the number of nodes, edges, and tweets in the two datasets, as well as aggregate network statistics.

A given Twitter hashtag is sometimes used by more than one community for very different reasons (Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer, 2011). A review of outlying users pointed to an instance where #CPC was being used by Spanish speaking Twitter users for categorizing content about the Centro de Predicción Climática, a climate change center².

On Twitter when one follows the conversation(s) of a given hashtag community they are subject to all messages with that specific tag. Additionally, the portion of total tweets which include references to non-political issues is estimated to be less than 10% in each network ³ and no top influencers across any metric included users who used the #CPC tag outside of the Canadian political context. As such, we have left all tweets in. There were no equivalent problems detected in the #NDP dataset.

¹For more information about the particulars about data access on Twitter for researchers, please consult (Gaffney, Forthcoming 2013)

²These outliers were detected by reviewing the followers_count field of each user's profile – this field records the total number of people following the account on Twitter.

³this is based on manual content analysis of a random sample of 10% of tweets in each network, see APPENDIX NUMBER for inter-coder reliability measures which followed guidelines outlined by (Lombard, Snyder-Duch, & Bracken, 2002)

First, a comparative analysis of popularly cited network metrics was used to assess possible facets of influence, and those metrics similarities in defining different facets of influence. Several popular network metrics, along with two bespoke metrics employed in this study, were selected for this process. Then, by using a rank correlation coefficient, a quantitative sense can be gained for the degree to which different metrics agree, and in turn, what different facets of influence may exist. Table 1 provides a list of the various network metrics employed in this study, outlines a brief description of their provenance and intended utility, and occasionally, notes about the metrics.

Of the metrics employed, the majority are standard network analysis metrics employed across various fields, one, “Interaction,” is commonly used in studies of influence on Twitter, and one, “Knowledge,” was generated for this research specifically. The interaction metric is a count of how many times a user was mentioned by other users during the two week sampling period. The knowledge metric was designed for this study with the aim to provide further insight about influence based on the content produced by the actual user in question.

To create the knowledge metric, we crafted a basic coding schedule and automated content analysis of the past tweets of each individual in the network in order to rank them based on the language use. The assumption we embedded in this content analysis was that those who use specific technical terms were likely to have higher levels of expertise and be portrayed as more knowledgeable. We chose to conduct content analysis in this way because it ensured the end result would be clearly intelligible given a succinct set of terms specific to discussion of politics in Canada. Operationally speaking, this number is the number of times a user posted a tweet in which one or more keywords⁴ across the tweets they posted within the dataset.

All metrics were run against the two separate datasets, and two standard rank correlation coefficients, Kendall’s τ and Spearman’s ρ , were employed in analysing their relationships. The observations made by this study relies primarily on Kendall’s τ , as Spearman’s ρ is less directly interpretable and more sensitive to few cases of extreme divergence in rankings, though Spearman’s ρ is reported for researchers more familiar with this metric. By employing a non-parametric ranking statistic such as Kendall’s τ , we aim to draw out pairwise comparisons of all metrics employed, and analyse the relative degrees to which these metrics agree (when a high τ is found), disagree (when a low τ is found), and when they diverge across the entire set (when a τ close to zero is found). Table 4 provides these scores for the #CPC dataset, while table 5 provides these scores for #NDP.

Next, of the six measures considered in the quantitative analysis, we selected five for more in-depth qualitative analysis⁵. In order to do this, we conducted a content analysis relying primarily on categorization (Miles & Huberman, 1994) of the profiles of all accounts found among the top 20 influentials of any of our five chosen measures. We chose the top 20 cut off as it represents a group large enough to provide variety but small enough to conduct meaningful qualitative analysis about what kinds of users are found to be most influential. The aim is not generalizability but rather to map the terrain of top influencers in these two hashtag communities. Other similar studies have

⁴The keywords used were “HOC”, “GOC”, “SenCA”, “byelxn”, “roft”, “cdnleft”, “p2ca”, and “QP”

⁵The “followers count” measure was omitted given that indegree provides similar information which is specific to the particular networks we are interested in

consistently chosen to look at the top 20 or fewer cases when comparing measures in this way (Cha & Gummadi, 2010; Wu et al., 2011).

In order to conduct the content analysis a coder and one of the researchers, both familiar with Canadian politics, classified each account based on whether they were media, partisan, activist, commentator/blogger, other notable, opinion leader, or average. Having multiple coders increases reliability (Potter & Levine-Donnerstein, 1999), please see APPENDIX NUMBER for a listing of Cohen's Kappa and per cent agreement levels which were all within an acceptable range.

Comparing Rankings

During the quantitative portion of analysis, we were most interested in whether or not different metrics, all attempting to score users based on their capacity to influence, ranked users in similar ways. By identifying these cases, we hope to further illustrate the distinct facets of influence, the metrics that are identifying those facets, and theoretical backing for why these metrics behave in such a way.

The ranking statistics bore out similar trends in both the #CPC and #NDP networks – eigenvector centrality and indegree tended to rank highly together, while followers count and interaction count fell into values that conferred only minor agreement with other rankings, and our knowledge metric and clustering coefficient scores seemed slightly more independent to many other metrics. One general point to note is that these metrics cannot be relied on to make strong arguments about what the distinctions are, or whether they are substantive, but instead provides us with a way to compare how each metric ranks users at an aggregate level.

That said, the clearest point to be made is that there are three general groups of metric pairings: pairs that generally agree (Indegree/Eigenvector Centrality), pairs that show lesser degrees of agreement (eg. Indegree/Interaction Count, Indegree/Followers Count, Eigenvector Centrality/Followers Count), and pairs that do not agree (Followers Count/Clustering Coefficient, Clustering Coefficient/Knowledge). What follows is a more specific review of the most meaningful pairs.

Metrics that ranked highly together suggest that they are either measuring the same facet of influence or they are measuring different facets, which ultimately tell the same story about who is influential. This makes sense when you consider the pair with the highest Kendall's τ – eigenvector centrality and indegree are both metrics which indicate how central a node is within a network in terms of the inbound connections towards it. If the knowledge metric were to rank highly with these measures, for example, it would suggest that the kind of influence it measures overlaps with eigenvector centrality and indegree. This is not the case, and as such qualitative analysis will be required to understand the differences between the measures further.

Of all relationships for #CPC data, followers count and clustering coefficient disagreed in ranks of the nodes strongest (τ -0.1658), and at that level is still much closer to independence rather than disagreement – for #NDP, the value expressed extremely high independence (τ -0.0043). For an exhaustive table of these metrics for the full #CPC and #NDP networks, alongside secondary rankings of a derivative network excluding top nodes by their eigenvector centrality value that will be examined further below, please consult tables 4 and 5.

In order to get a sense of how extreme differences in rankings provided by each measure are, we consider their absolute range of divergence across all other metrics. These values are available in Tables 6 and 7. By simplifying the more exhaustive tables, we can glean interesting trends about the overall distinctiveness of different metrics. For metrics to be comparably more distinct than other metrics, they should express ranges of Kendall's τ that are narrower and closer to zero, which is perfect disagreement (and is indicative of a high degree of independence, which may infer that the metric is possibly measuring something distinct). Measures that have ranges more towards the extremes are indicative of cases where metrics are measuring very similar phenomena (when the maximum of the range approaches one) or complete opposite phenomena (when the maximum of the range approaches zero).

Reviewing this data, eigenvector centrality and indegree have the highest maximums (as was seen with the entire set of pairs above) for both datasets. Followers Count and Interaction count were both generally in the range of 0.0 to 0.5, expressing some similarity to other metrics for a few cases, while to some degree knowledge and certainly clustering coefficient had considerably smaller ranges that were much closer to zero. The general interpretation of this trend is that eigenvector centrality and indegree rank nodes dissimilarly with respect to some other metrics, and those disagreements are stronger than any other ranking relationships across both datasets. For other metrics, smaller ranges are indicative of less agreement, while smaller maxima and minima for those ranges express increasing distinction of those metrics against others (and potentially, vastly different facets of influence).

Identifying Influential Political Players

We have discussed quantitatively the relationship between these various measures noting that they do not identify influentials in the same order but that some tend to agree more than others. The question is, are these differences ones that actually make a difference? Our qualitative analysis suggests yes, that different metrics identify different kinds of political players. Metrics concerned with the following of a user tended to identify traditionally important and highly visible political players like media outlets, journalists, and politicians. Metrics which embedded assumptions about the importance of being seen as and/or acting like an expert identified political commentators and bloggers as particularly influential. Finally, our metric which prioritizes local social embeddedness identified primarily average users which we discuss in detail in the section "Local Context."

(Cha & Gummadi, 2010) note in-degree has been the most basic measure of influence used across scholarship, however, they also explain it may be a better measure of popularity than actual influence. The core assumption is that the most important facet of the influence process is having a large following. Eigenvector centrality takes this concept a step further and says it is a matter of having a large following of those who also have a large following (Shamma, Kennedy, & Churchill, 2009).

Not surprisingly then, the highest ranked 20 in each list in both cases was populated almost entirely by media outlets, journalists, and politicians. The only exceptions were @iancapstick, a former communications director for the NDP who appeared in both #NDP indegree and eigenvector centrality lists, @stephen_taylor who is a well known CPC blogger who appeared in the #CPC

indegree list, and @leadnowca which is an activist group appearing in the indegree #CPC list.

Thinking about the #CPC and #NDP Twitter communities as opportunities for getting news (Kwak, Lee, Park, & Moon, 2010), it is not surprising the most central accounts are those of journalists and politicians, as they are typically the ones with the most access to political happenings. These accounts are most likely to send out first hand and/or very reliable information and they are accounts with a professional reputation to incite trust. The general public are interested in who they are and what they say and so too are others who form the political and media elite.

Are these the individuals who are actually most able to convince others to change their minds? The political elite are important and powerful players in political discussion, to be sure, but are they the only ones? These elite are not necessarily the influencers most able to convince others to change opinions and to lead political discussions.

Content based rankings, like our Interaction metric, offer an alternative route to network structure metrics. Though three quarters of the top 20 most mentioned users were media and politicians, the overlap in terms of which media and which politicians was not perfect. For example, @NSNDP, a twitter feed run by the Nova Scotia New Democrats was number six most mentioned in the #NDP network but was not found to be among the 20 most influentials by any other metric. Similarly, Liberal Leadership candidates @JoyceMurray and @MHallFindaly were both among the most mentioned in each network, yet neither appeared influential by any other measure. It is not surprising that, in terms of follower ties these two politicians are not found to be central as they are both members of the third party in Canada which uses the hashtag #LPC. They were likely being talked about because they were running for leadership of a major political party and they likely used the #CPC and #NDP hashtags because they wanted to access other politically inclined Canadians.

Next, we consider keyword ranking, what we have called Knowledge. Just under half of the most highly ranked accounts were deemed to be average users or opinion leaders with a mix of politicians, parties, journalists, and bloggers making up the rest of the list. Interestingly, it is in this list that we see the prominence of political staffers for the first time. The staffers in both the CPC and NDP list are all affiliated with the #NDP dataset. This is likely a result of the NDP's communication strategy where those within the party are prompted to provide the public with consistent language, language our coding scheme has picked up.

In terms of measuring influence, this example highlights a weakness and a strength. On the one hand, the difficulties related to reliability and validity researchers face using content analysis remain relevant. Though we increased reliability and validity by testing our coding schedule multiple times (Rourke & Anderson, 2004) and using a random sample of actual data to develop a hierarchical coding schedule (Richards & Richards, 1995), it remains possible results were skewed in favor of certain groups of individuals, in this case possibly NDP staff.

If, for example, the CPC strategy was to avoid discussion of the Senate while the NDP were heavily pushing for Senate discussion, then the term we identified as a non-partisan indicator of political awareness could become a political hot topic. Considering the #CPC network alone, it is possible that an NDP staffer using the #CPC hashtag may not have many followers within the network but does meet our requirements for language use. Since our coding schedule could have

avored the NDP language over CPC language that staffer may be identified as more influential than someone else who may actually have more pull within the #CPC community. Consulting news coverage over our sampling period, there were no major indicators of such situations but an exhaustive analysis of non-Twitter based content is beyond the scope of this paper. Such analysis is not required for our basic comparison of measures particularly since qualitative analysis is able to contextualize our findings.

However, this methodological worry can also be used to our advantage. For example, (Gentzkow & Shapiro, 2010), compile a coding schedule based on the language of partisans in the US Senate and use it to classify journalists as left and right leaning. Creating measures of influence specific to certain communities offers the opportunity to engage with that community in a deeper way by considering not just who has the largest potential audience or even active audience, but what they are saying and how it may be received which are theoretically important facets of the influence process.

Looking to the clustering coefficient ranking, almost every highly ranked user was considered an “average account,” yet average users only appeared once or twice among the top 20 most highly ranked for most other metrics. It is notable that far more than 20 accounts have clustering coefficients of one which is the top score, our analysis has taken a random sample of the top users. That said, not one user ranked within the top 20 on any metric has a clustering coefficient of one, the highest score being found among the top ranked given Knowledge at approximately 0.6 in both the #CPC and #NDP networks. Thinking about what a clustering coefficient of one means, that every follower a user is connected to also follows every other of those followers, this is not that surprising. Unless you follow only very few other users, it is hard to achieve a perfectly connected neighbourhood.

What this meant within the #CPC and #NDP networks is that users with very small local communities who did not follow politicians, journalists, or other very visible influentials, made up the vast majority of users with a clustering coefficient of one. As soon as a user follows a very visible account, for example a journalist, their clustering coefficient diminishes unless that journalist happens to be connected to everyone else the user is connected to. This is unlikely because political elites, like journalists, do not tend to follow many non-elite and so either the user follows only journalists who also follow each other, or the clustering coefficient is not one. The alternative route to a fully connected neighbourhood, assuming an undirected graph, would be for all of the users connections to follow that same journalist. This, however, presents a theoretical challenge because it is assumed that influentials will have some form of access to information with which to develop opinion and then influence others who do not access that information (Katz & Lazarsfeld, 2006). Put simply, local influencers (like opinion leaders in Two-Step Flow work) are expected to follow elites. As such, the clustering coefficient of the broad network is not optimal for identifying influencers, whether they are the very visible political elite or the local influencers who are embedded in a community.

In our comparison of measures, we have noted that traditional measures of centrality tend to agree on how to rank influencers. These network wide measures have identified political elites like politicians, media outlets, and journalists. Measures of interaction and other content based metrics help identify political commentators and bloggers outside the traditional elite. Though the clustering coefficient as applied to the full network did not identify individuals who appear likely to

be influential, this does not mean position in one's local network or even clustering coefficient are inconsequential. Given the theoretical importance of one's close personal ties and placement within a community, the following section investigates the potential use of the clustering coefficient further.

The local context

Given the algorithm behind clustering coefficient, the general characteristics of users following very visible accounts, and that following visible accounts is crucial for potential local influencers, we reason as follows: Those most likely to be local influencers will have lower clustering coefficients within the wider network because they follow very visible influentials. In other words, their clustering coefficients will be artificially low. Should those very visible influentials be removed, the clustering coefficients of those likely to be local influencers will increase. Those users whose clustering coefficient increases the most are most likely to be locally influential because they are users who have both access to information and are well positioned to disseminate that information to their local network.

We tested this logic by removing very visible users and creating a derivative network. We chose to remove nodes based on whether their eigenvector centrality score was at or above the elbow of the distribution of all eigenvector centrality scores within the network ⁶. Eigenvector centrality was selected because while indegree may remove most popular users, eigenvector centrality will remove popular users within the network who are in turn followed by popular users. In effect, this preferentially removes users who are followed by likely candidates for opinion leadership, as the exact nodes who are following these users is taken into the calculation of eigenvector centrality. The elbow of the distribution of eigenvector centrality values was selected as the point at which to isolate these users as it provides a readily interpretable cutoff point, and is based on the dataset itself rather than an arbitrary figure.

A qualitative analysis of the 20 users whose clustering coefficient increased the most supports our reasoning. Expectedly, no users were political elite, though some bloggers and small organizations were included. The next step is to differentiate between average users who are likely to be influential and those less likely. Posting political content, mentioning political elite, and having political conversations were all considered positive indicators of influence and was consistently found among this top 20 list whereas it was not for the network wide clustering coefficient top 20 list.

While this is a promising finding, there are a few important points to consider. First, the clustering coefficient applied to the derivative network still favours users with small neighbourhoods. While they may be well positioned to influence locally, that locale may be quite small. Therefore it may make sense to set a minimum indegree level as other studies of influence on Twitter have done (Cha & Gummadi, 2010). Second, just because clustering coefficient increases it does not mean the new clustering coefficient is high. In our case this was not an issue as only three users of our top 20 in the #CPC and none in the #NDP derivative networks had clustering coefficients below one, the others all had scores above 0.83 which remains relatively high, but in another network it could be

⁶These scores, for #CPC and #NDP respectively, were 0.1288 and 0.1311

different. Third, those with already high clustering coefficients are less likely to be found in the list of users who increased the most even if they have a high or perfect score in the derivative network. Theoretically, and given evidence in this case, this is justified because those with the highest original scores tend to be those who are not connected to elite players and do not access political information. That said, depending on the specific network it is possible that users which occupy a middle ground could be ranked lower than is actually ideal.

In short, the utility of the clustering coefficient for identifying local influentials is contextualized. It is the case with all the measures of influence we have used in this study that specific assumptions about which facets of the influence process are most important are embedded in the operationalization of each measure. These assumptions have been justified based on theory and tested by considering users who are deemed influential qualitatively.

DISCUSSION

In most measures of influence, the most important facet of influence is assumed to be who follows a given user and how often they talk about that user/if the user is treated as an expert. Presence is important, one cannot lead if they have no followers. However, other facets of influence are routinely ignored. The the role of expertise and knowledge and the importance of interpersonal interaction and personal connection are factors deemed theoretically relevant to the process of influencing someone (Katz & Lazarsfeld, 2006). It is not that having a following and being trusted, knowledgeable, and socially connected are in opposition, rather it is a matter of placing more theoretical and operational importance on some facets over others. While decisions must be made, using measures of influence out of context could lead to confusing or inaccurate results.

For example, we might gather from their professions that the journalists and politicians who top most measures of influence, particularly the standard measures of centrality most common to studies in this area, are in fact trusted as experts. It is a far cry to also assume they fulfill the need for close personal ties to help interpret information and actually do the hard work of convincing someone to change their attitude, opinion, or behaviour.

By ignoring certain facets of what makes an individual influential on a given topic or by attempting to assign influence regardless of topic or social connections, those who actually hold the most influence in a certain community may be drowned out by others who are well placed in other communities or globally popular.

Another interesting question is raised when we consider the example of Liberal leadership candidates within the #CPC and #NDP networks. Two candidates ranked as highly influential according to the number of times they were mentioned in both the #CPC and #NDP networks without appearing in any of the other lists of most influential.

Is interaction with certain individuals or position within a wider community more telling of a user's capacity to influence? Put differently, when we talk about a given "political community" is it those who are most active only, or also those who may passively exist within its bounds we are discussing?

From the perspective of the Two-Step Flow Hypothesis, if we think only of those who actively engage we are already limiting ourselves to likely opinion leaders and public figures, both of which have been described as “influentials” (Katz & Lazarsfeld, 2006). We necessarily eliminate those most likely to be primarily followers. With this the theoretical context shifts. Since we rank people in terms of influence the top bracket retain their title as influential but with the bottom bracket eliminated, the middle group become the “followers” and their type of influence is lost. “Influential” becomes a simpler concept which can be useful, but it also means we then lack clarity concerning the complexity of the social process that is influence.

That said, it is not theoretically sufficient to take the list of users ranked by interaction and assign the top portion the title public influential, bottom portion opinion leader, and any user not on the list follower. The notion of opinion leadership was seminal in the field of media studies and political communication because it connected theories of community and group dynamics to theories of mass media and political messaging. Social support and social pressure, applied by the opinion leader on his or her “every-day associates” – core social group, were the main mechanisms by which influence happened. Opinions changed when someone in a close-knit social group payed attention to a mass message and then used their position within that small group to personally influence the other members.

By this description interaction within a network is indeed important, and having some following is a necessity, but structural position within one’s local neighbourhood is also important. It is this final aspect the majority of influence metrics overlook and the point which our analysis using the clustering coefficient speaks to most clearly.

Importantly, we are not advocating for the clustering coefficient as a stand-alone measure of influence. As (González-Bailón et al., 2012) note, it is a useful addition to the a repertoire of influence measures. Indeed, we believe no single measure we have assessed in this paper is sufficient for identifying the range of different kinds of influentials found within a political discussion network on Twitter. That is because influence is a contextualized phenomenon and measuring how influential a communicator is presupposes an ability to isolate the components of influence and weigh them accurately within that context. The reason some of our measures varied so greatly is that the components of influence isolated are very different. Clear understandings of what these measures qualitatively represent can be used to help guide theory development and influential identification.

In sum, our study has used multiple measures of influence in order to identify the most influential members of the #CPC and #NDP Twitter communities. We have found that in terms of network placement, political elites such as media outlets, journalists, and politicians are most influential in each network. When interaction and content is considered both at the network level and globally, political elite remain prominent but political commentators and bloggers are integrated into the lists of most influential. Finally, considering how socially embedded a user is within their local neighbourhood, we are able to identify likely opinion leaders.

Though specific to our case, we believe these results are instructive for future studies of influence on Twitter and potentially other online social networking sites. As the ease with which we can trace interactions among people increases, we need to remain aware of how operational definitions can

impact the theoretical context of our research.

TABLES

Metric	Description	Papers Using this metric
In-degree	The number of nodes with a directed edge pointing towards the given node This can be thought of as a contextualized count of followers, a metric typically understood to confer influence.	(Cha & Gummadi, 2010; Java, Song, Finin, & Tseng, 2007; Romero & Kleinberg, 2010)
Followers Count	The total number of users following this account on Twitter – the global in-degree of the node across Twitter.	(Mendoza, Poblete, & Castillo, 2010; Kwak et al., 2010; Wu et al., 2011)
Eigenvector Centrality	Inspiration for Google’s PageRank algorithm, the metric was developed by Bonacich (1972) to quantify influence in a network.	(Weitzel et al., 2012; Bigonha, Cardoso, Moro, Almeida, & Gonçalves, 2010)
Clustering Coefficient	The clustering coefficient is a metric conferring the degree to which a given node is embedded within a tightly bound set of other nodes. Algorithm Employed is (Latapy, 2008)	(Lerman & Ghosh, 2010; Java et al., 2007)
Knowledge	The number of tweets a user posts containing context-specific terms divided by the number of tweets in the sample Terms derived from random sample of tweets collected during sampling period from both networks.	Derived in this work
Interaction	The total number of times all other users mentioned the given user within the dataset	Derived in this work

Table 1: Metric Overview

Metric	#CPC	#NDP
Users (nodes)	3,860	3,536
Friendships (edges)	163,506	144,658
Statuses (tweets)	730,562	653,989
Average In-degree	42.359	40.91
Maximum In-degree	1,428	1,309
Global Clustering Coefficient	0.295	0.285

Table 2: Aggregate Statistics for #CPC and #NDP datasets

Metric	#CPC	#NDP
Users (nodes)	3,459	3,159
Friendships (edges)	36,951	34,501
Statuses (tweets)	647,908	578,000
Average In-degree	10.683	10.921
Maximum In-degree	157	162
Global Clustering Coefficient	0.169	0.166

Table 3: Aggregate Statistics for #CPC and #NDP dataset, excluding top by eigenvector centrality

		Full Network		Network excluding top elbow by eigenvector centrality		
First Metric	Paired Metric	Kendall's τ	Spearman's ρ	Kendall's τ	Spearman's ρ	$\Delta(\tau)$
Indegree	Eigenvector Centrality	0.856	0.975	0.8251	0.9655	-0.0309
Indegree	Interaction Count	0.4933	0.6798	0.4178	0.5947	-0.0755
Indegree	Followers Count	0.4391	0.6057	0.3553	0.5061	-0.0838
Eigenvector Centrality	Interaction Count	0.4345	0.6077	0.3466	0.5007	-0.0879
Followers Count	Eigenvector Centrality	0.3918	0.5442	0.2978	0.4273	-0.094
Followers Count	Interaction Count	0.3854	0.5411	0.3115	0.4467	-0.0739
Eigenvector Centrality	Knowledge	0.2431	0.4458	0.1989	0.388	-0.0442
Indegree	Knowledge	0.231	0.4244	0.1848	0.3609	-0.0462
Knowledge	Interaction Count	0.1365	0.2533	0.08	0.1565	-0.0565
Followers Count	Knowledge	0.0577	0.1064	-0.0048	-0.0096	-0.0625
Clustering Coefficient	Knowledge	0.0544	0.1043	0.0835	0.1639	0.0291
Eigenvector Centrality	Clustering Coefficient	0.0148	0.1082	0.1065	0.2146	0.0917
Indegree	Clustering Coefficient	-0.0156	0.074	0.072	0.1756	0.0876
Clustering Coefficient	Interaction Count	-0.0659	-0.0687	-0.0045	0.0085	0.0614
Followers Count	Clustering Coefficient	-0.1658	-0.2095	-0.1132	-0.1479	0.0526

Table 4: #CPC Kendall's τ and Spearman's ρ Ranks

		Full Network		Network excluding top elbow by eigenvector centrality		
First Metric	Paired Metric	Kendall's τ	Spearman's ρ	Kendall's τ	Spearman's ρ	$\Delta(\tau)$
Indegree	Eigenvector Centrality	0.7968	0.9605	0.8518	0.9763	-0.055
Indegree	Interaction Count	0.476	0.6665	0.5182	0.7087	-0.0422
Eigenvector Centrality	Interaction Count	0.4246	0.6029	0.4599	0.6428	-0.0353
Indegree	Followers Count	0.4171	0.5907	0.4771	0.6553	-0.06
Followers Count	Eigenvector Centrality	0.3758	0.5387	0.4402	0.6111	-0.0644
Followers Count	Interaction Count	0.3474	0.5002	0.4187	0.5873	-0.0713
Eigenvector Centrality	Knowledge	0.1502	0.3007	0.2355	0.4365	-0.0853
Eigenvector Centrality	Clustering Coefficient	0.1482	0.2396	0.0625	0.1548	0.0857
Indegree	Knowledge	0.1446	0.2894	0.2189	0.4067	-0.0743
Indegree	Clustering Coefficient	0.1219	0.2124	0.0191	0.1059	0.1028
Clustering Coefficient	Interaction Count	0.0738	0.1138	-0.0618	-0.0628	0.1356
Knowledge	Interaction Count	0.0678	0.1354	0.1219	0.2283	-0.0541
Clustering Coefficient	Knowledge	0.0373	0.075	0.0786	0.1478	-0.0413
Followers Count	Knowledge	0.0136	0.0286	0.0707	0.1322	-0.0571
Followers Count	Clustering Coefficient	-0.0043	0.005	-0.1487	-0.1707	0.1444

Table 5: #NDP Kendall's τ and Spearman's ρ Ranks

Metric	Maximum Rank Correlation	Minimum Rank Correlation	Range
Eigenvector Centrality	0.856	0.0148	0.8412
Indegree	0.856	-0.0156	0.8716
Followers Count	0.4391	-0.1658	0.6049
Interaction Count	0.4933	-0.0659	0.5592
Clustering Coefficient	0.0544	-0.1658	0.2202
Knowledge	0.2431	0.0544	0.1887

Table 6: Range of Kendall's τ for metrics in comparison to all others in #CPC Dataset

Metric	Maximum Rank Correlation	Minimum Rank Correlation	Range
Eigenvector Centrality	0.8518	0.0625	0.7893
Indegree	0.8518	0.0191	0.8327
Followers Count	0.4771	-0.1487	0.6258
Interaction Count	0.5182	-0.0618	0.58
Clustering Coefficient	0.0786	-0.1487	0.2273
Knowledge	0.2355	0.0707	0.1648

Table 7: Range of Kendall's τ for metrics in comparison to all others in #NDP Dataset

References

- Agre, P. (2002). Real-Time Politics: The Internet and the Political Process. *The Information Society*, 18(5), 311–331. Available from <http://www.tandfonline.com/doi/abs/10.1080/01972240290075174>
- Al-sharawneh, J., Sinnappan, S., & Williams, M.-a. (2013). Credibility-Based Twitter Social Network Analysis. In (pp. 323–331).
- Anger, I., & Kittl, C. (2011). Measuring influence on Twitter. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '11*, 1. Available from <http://dl.acm.org/citation.cfm?doid=2024288.2024326>
- Bakshy, E., Hofman, J. M., Watts, D. J., & Mason, W. A. (2011). Everyone s an Influencer: Quantifying Influence on Twitter Categories and Subject Descriptors. In *Wsdm '11 proceedings of the fourth acm international conference on web search and data mining* (pp. 65–74).
- Bigonha, C., Cardoso, T. N., Moro, M. M., Almeida, V. A., & Gonçalves, M. A. (2010). Detecting evangelists and detractors on twitter. In *18th brazilian symposium on multimedia and the web* (pp. 107–114).
- Bigonha, C., Cardoso, T. N. C., Moro, M. M., Gonçalves, M. a., & Almeida, V. a. F. (2011, December). Sentiment-based influence detection on Twitter. *Journal of the Brazilian Computer Society*, 18(3), 169–183. Available from <http://www.springerlink.com/index/10.1007/s13173-011-0051-5>
- Blumler, J. G., & Gurevitch, M. (2001, January). THE NEW MEDIA AND OUR POLITICAL COMMUNICATION DISCONTENTS: DEMOCRATIZING CYBERSPACE. *Information, Communication & Society*, 4(1), 1–13. Available from <http://www.tandfonline.com/doi/abs/10.1080/713768514>
- Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1), 113-120. Available from <http://www.tandfonline.com/doi/abs/10.1080/0022250X.1972.9989806>
- Brosius, H.-B., & Weimann, G. (1996, October). Who Sets the Agenda: Agenda-Setting as a Two-Step Flow. *Communication Research*, 23(5), 561–580. Available from <http://crx.sagepub.com/content/23/5/561.abstract>
- Brown, P. E., & Feng, J. (2010). Measuring User Influence on Twitter Using Modified K-Shell Decomposition. In (pp. 18–23).
- Cha, M., & Gummadi, K. P. (2010). Measuring User Influence in Twitter : The Million Follower Fallacy. , 10–17.
- Chadwick, A. (2011, January). The Political Information Cycle in a Hybrid News System: The British Prime Minister and the Bullygate Affair. *The International Journal of Press/Politics*, 16(1), 3–29. Available from <http://hij.sagepub.com/content/16/1/3.abstract>
- Chang, Y., Wang, X., Mei, Q., Liu, Y., & Park, M. (2013). Towards Twitter Context Summarization with User Influence Models. In *sixth acm international conference on web search and data mining* (pp. 527–536).
- Coleman, S. (2004). Connecting parliament to the public via the internet. *Information, Communica-*

- tion Society*, 7(1), 1-22. Available from <http://www.tandfonline.com/doi/abs/10.1080/1369118042000208870>
- Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the political alignment of twitter users. Available from http://cnets.indiana.edu/wp-content/uploads/conover_prediction_socialcom.pdfexpress_ok_version.pdf
- Dahlgren, P. (2000). The Internet and the Democratization of Civic Culture. *Political Communication*, 17(4), 335–340. Available from <http://www.tandfonline.com/doi/abs/10.1080/10584600050178933>
- Dahlgren, P. (2005). The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation. *Political Communication*, 22(2), 147–162. Available from <http://www.tandfonline.com/doi/abs/10.1080/10584600590933160>
- Dillard, J. P. (1989). Types of influence goals in personal relationships. *Journal of Social and Personal Relationships*, 6, 293–308.
- Dillard, S. C., J. P., & Harden, J. M. (1989). Primary and secondary goals in the production of interpersonal influence messages. *Communication Monographs*, 56, 19–38.
- Dubois, E., & Dutton, W. H. (Forthcoming 2013). The Fifth Estate in Internet Governance: Collective Accountability of a Canadian Policy Initiative. *Revue Française d'Etudes Américaines RFEA*, 131(1).
- Duong, K. (2012). Canada Digital Future in Focus 2012. *comScore*. Available from <http://www.comscore.com/2012CanadaDigitalFutureinFocus>
- Dutton, W. H., & Peltu, M. (2007). The Emerging Internet Governance Mosaic: Connecting the Pieces. *Information Polity*, 12, 63–81.
- Eveland, W. P., Morey, A. C., & Hutchens, M. J. (2011, December). Beyond Deliberation: New Directions for the Study of Informal Political Conversation from a Communication Perspective. *Journal of Communication*, 61(6), 1082–1103. Available from <http://doi.wiley.com/10.1111/j.1460-2466.2011.01598.x>
- Gaffney, . P. C., D. (Forthcoming 2013). Data collection on Twitter. In J. B. M. M. . C. P. A. Bruns K. Weller (Ed.), *Twitter and society*. New York, NY: Peter Lang.
- Gentzkow, M., & Shapiro, J. M. (2010, January). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1), 35–71. Available from <http://dx.doi.org/10.3982/ECTA7195>
- Gibson, L. W., R. K., & Ward, S. (2005). Online Participation in the UK: Testing a “Contextualised” Model of Internet Effects. *The British Journal of Politics & International Relations*, 7, 561–583.
- Gibson, R. K., & Ward, S. (2002). Virtual campaigning: Australian parties and the impact of the internet. *Australian Journal of Political Science*, 37(1), 99-129. Available from <http://www.tandfonline.com/doi/abs/10.1080/13603100220119047>
- González-Bailón, S., Borge-Holthoefer, J., & Moreno, Y. (2012, March 8). Broadcasters and Hidden Influentials in Online Protest Diffusion. *American Behavioral Scientist*. Available from <http://dx.doi.org/10.2139/ssrn.2017808>

- Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an Imagined Community. *American Behavioral Scientist*, 55(10), 1294–1318. Available from <http://abs.sagepub.com/content/55/10/1294.abstract>
- Hill, S., Benton, A., Ungar, L., Macskassy, S., Chung, A., & Holmes, J. H. (n.d.). A Cluster-based Method for Isolating Influence on Twitter.
- Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2), 256–276.
- Hindman, M. S. (2009). *The Myth of Digital Democracy*. Princeton: Princeton University Press.
- Huckfeldt, R. R., Sprague, J., Kuklinski, J. H., Wyer, R. S., & Feldman, S. (1995). *Citizens, Politics and Social Communication: Information and Influence in an Election Campaign*. Cambridge University Press.
- Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th webkdd and 1st sna-kdd 2007 workshop on web mining and social network analysis* (pp. 56–65). New York, NY, USA: ACM. Available from <http://doi.acm.org/10.1145/1348549.1348556>
- Katz, E. (1957, March). The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *Public Opinion Quarterly*, 21(1), 61–78. Available from <http://poq.oxfordjournals.org/content/21/1/61.abstract>
- Katz, E., & Lazarsfeld, P. F. (2006). *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers.
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600). New York, NY, USA: ACM. Available from <http://doi.acm.org/10.1145/1772690.1772751>
- Latapy, M. (2008, November). Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theor. Comput. Sci.*, 407(1-3), 458–473. Available from <http://dx.doi.org/10.1016/j.tcs.2008.07.017>
- Lee, C., Kwak, H., Park, H., & Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. *Proceedings of the 19th international conference on World wide web - WWW '10*, 1137. Available from <http://portal.acm.org/citation.cfm?doid=1772690.1772842>
- Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th international conference on weblogs and social media (icwsm)*.
- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, 28(4), 587–604. Available from <http://dx.doi.org/10.1111/j.1468-2958.2002.tb00826.x>
- Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: can we trust what we rt? In *Proceedings of the first workshop on social media analytics* (pp. 71–79). New York, NY, USA: ACM. Available from <http://doi.acm.org/10.1145/1964858.1964869>
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*.

Thousand Oaks: Sage Publications.

- Mutz, D. C. (2006). *Hearing the other side: Deliberative versus participatory democracy*. New York, NY: Cambridge University Press.
- Norris, P., & Curtice, J. (2008, May). Getting the Message Out: A two-step model of the role of the Internet in campaign communication flows during the 2005 British General Election. *Journal of Information Technology & Politics*, 4(4), 3–13. Available from <http://dx.doi.org/10.1080/19331680801975359>
- Overbey, L. A., Paribello, C., & Jackson, T. (2013). Identifying Influential Twitter Users in the 2011 Egyptian Revolution. , 377–385.
- Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3), 258–284.
- Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2011). *In the Mood for Being Influential on Twitter*.
- Rattananaritnont, G., Toyoda, M., & Kitsuregawa, M. (2012). Characterizing Topic-Specific Hashtag Cascade in Twitter Based on Distributions. , 735–742.
- Richards, T., & Richards, L. (1995). Using hierarchical categories in qualitative data analysis. In U. Kelle (Ed.), *Computer-aided qualitative data analysis*. SAGE.
- Robinson, J. P. (1976, September). Interpersonal Influence in Election Campaigns: Two Step-flow Hypotheses. *Public Opinion Quarterly*, 40(3), 304–319. Available from <http://poq.oxfordjournals.org/content/40/3/304.abstract>
- Rogers, E. M. (2010). *Diffusion of Innovations, 4th Edition*. Free Press.
- Romero, D. M., & Kleinberg, J. (2010). The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *In icwsm*.
- Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development*, 52(1), 5–18. Available from <http://dx.doi.org/10.1007/BF02504769>
- Shamma, D. A., Kennedy, L., & Churchill, E. F. (2009). Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first sigmm workshop on social media* (pp. 3–10). New York, NY, USA: ACM. Available from <http://doi.acm.org/10.1145/1631144.1631148>
- Sousa, D. (n.d.). Characterization of the Twitter @ replies Network : Are User Ties Social or Topical ? In (pp. 63–70).
- Subbian, K., & Melville, P. (2011). Supervised Rank Aggregation for Predicting Influencers in Twitter. In (pp. 661–665).
- Watts, D., & Dodds, P. (2007, December). Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34(4), 441–458. Available from <http://www.jstor.org/stable/10.1086/518527>
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 409–10.
- Weimann, G. (1982, December). On the Importance of Marginality: One More Step into the Two-

- Step Flow of Communication. *American Sociological Review*, 47(6), 764–773. Available from <http://www.jstor.org/stable/2095212>
- Weitzel, L., Quaresma, P., & Oliveira, J. P. M. de. (2012). Measuring node importance on twitter microblogging. In *Proceedings of the 2nd international conference on web intelligence, mining and semantics* (pp. 11:1–11:7). New York, NY, USA: ACM. Available from <http://doi.acm.org/10.1145/2254129.2254145>
- Welser, S. M. A. G. E. . F. D., H. T. (2008). Distilling digital traces: Computational social science approaches to studying the internet. In L. R. Blank Grant & N. Fielding (Eds.), *Sage handbook of online research methods* (pp. 116–140). SAGE.
- Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 705–714). New York, NY, USA: ACM. Available from <http://doi.acm.org/10.1145/1963405.1963504>
- Ye, S., & Wu, S. F. (2010). Measuring Message Propagation and Social Influence on Twitter . com. *Social Informatics*, 6430(Section 3), 216–231.
- Yin, R. K. (2008). *Case study research: Design and methods (applied social research methods)* (Fourth Edition. ed.). Sage Publications.

Concerns Canadian politics?	Cohen's Kappa	Per cent agreement
CPC random sample (10%)	0.8731	97.67%
NDP random sample (10%)		

Table 8: Caption To Come

User Categorization Code	Cohen's Kappa	Agreement
Media outlet		
Journalist		
Party/partisan group		
Politician		
Activist/Advocacy group		
Commentator/Blogger		
Spam/bot/advertiser		
Average user		
Opinion leader		

Table 9: Caption To Come