Identifying Influential Users in Social Media: A Study of U.S. Immigration Reform

Wingyan Chung
School of Business Administration
Stetson University
wchung@stetson.edu

Daniel Zeng
Eller College of Management
The University of Arizona
zeng@email.arizona.edu

Nathan O’Hanlon
School of Business Administration
Stetson University
nohanlon@stetson.edu

Abstract

As opinion leaders and key participants emerge in social media, identifying influential users can help decision makers to effectively target the source of influence and hence bring about change in the communities. In this research, we developed an approach to identifying influential users in online social networks of interest to policy makers and the general public. We present findings from an empirical study of the U.S. immigration reform discussion, in which more than 300,000 users posted 909,035 tweets during May-November 2013. We present findings of our analysis, provide the lists of influential users identified, and discuss the implication on predictive analytics and social media analytics. This research should contribute to providing a new case and new empirical findings of applying influence analytics to analyzing social media networks, and has strong implications on predictive analytics, business intelligence, and social media analytics.

Keywords

Influence, influential users, social media analytics, social network analysis, text mining, predictive analytics, U.S. immigration reform, business intelligence.

Introduction

Social media has been widely used in building online communities. As opinion leaders and key participants emerge in these communities, identifying influential users can help decision makers to effectively target the source of influence and hence bring about change in the communities.

With the rapid growth of social media content and usage, policy makers and ordinary citizens are able to efficiently obtain voluminous data about public sentiment. Every day, more than 100 million users post over 230 million “tweets” (text messages with up to 140 characters) on the social media website Twitter.com (Miller 2011). Opinion leaders, authorities, and activists who share their ideas on Twitter are often followed closely by thousands of users. These leaders provide valuable content as well as linkage information that can offer insights for policy decision-making.

To gauge public opinion, policy makers, social scientists, and public servants rely extensively on networks of human contacts and political groups to obtain information. Traditionally, this information is obtained through various media, news agencies, and social channels. However, the large volumes and variety of expressions on social media have challenged traditional policy analysis and public sentiment assessment. Focusing on the agendas and activities of influential leaders in a community, policy makers can quickly identify current trends and movements of the community. Thus the identification of influential leaders in a community can save significant time and resources of polling and gauging opinion from the larger community. Existing works on identifying influential users from online social networks often incorporate simple metrics such as counts of replies, friends, and followers, which need to be augmented by advanced techniques and methods.
In this paper, we proposed the use of both content-based metrics and social network metrics to identify influential users in a large social network of interest to policy makers, analysts, and the general public. We define an influential user as a person who has high strategic importance and participation in connecting other users in a social network, often by channeling messages and resources in the community. We conducted an empirical study using data collected on the U.S. immigration reform discussion posted on Twitter. The data contain 909,035 tweets posted by over 300,000 users during May-November 2013. We present findings of our analysis, provide the lists of influential users identified, and discuss the implication on predictive analytics and social media analytics.

**Literature Review**

Previous studies have explored various aspects of influence across social networks. One aspect is the spread of ideas in a social network, for which researchers have developed models to maximize the spread (and hence influence) of ideas in the network (Chen et al. 2010). These studies use heuristic algorithmic approaches based on greedy algorithms to replicate efficient spread of influence in a network. Building on academic work of influence spread, several companies have developed proprietary algorithms of measuring influence across social networks, notable of which are services such as Klout, Twitalyzer, and BehaviorMatrix. Klout, one of the most well-known of these services, states that it uses a combination of statistics across multiple social networks and internet databases to compute their influence metric called “Klout score” (Klout 2014). Twitalyzer similarly uses a wide range of statistics across aspects of influence, however it focuses on Twitter analytics only (Twitalyzer 2012). BehaviorMatrix computes an influence score based on emotional cues picked up in the content tweets to analyze and rank users (BehaviorMatrix 2013). Apart from proprietary influence analytics developed as trade secrets or patents, other research was published in publicly-available outlets.

Articles have been written on the subject of measuring influence using Twitter data, however the methods are varied and focus on different aspects of the topic; some look towards novel definitions of influence that are comparable only to the original data source (Cha et al. 2010), while others focus on replicating industry measures such as the Klout score (Anger et al. 2011). A study by Hewlett-Packard uses a count of the number of IP hits on twitter posts to determine influence, finding that popularity does not imply influence (Romero et al. 2011). Others explored the traits of tweets that are commonly retweeted in an effort to discover the basis behind the spread of an individual message (Kucman 2012). Along the same path of content-based information, a study has also found that diverse messages and focused messengers have the greatest impact in a social network (Weng et al. 2014). Each of these approaches has some amount of variance in yielding valid results, as was made clear by critical analyses (Probst et al. 2013). Due to the inherent variability present in opinion- and influence-based research, steps should be taken to quantify and evaluate the results.

A study in radicalization and political violence shows that influence is highly concentrated among the top 1 percent of users in the set; high scores in both influence and exposure showed a strong correlation to engagement with the seed ideology (Berger et al. 2013). The researchers measured influence by counting the number of replies and retweets in Twitter, and measured exposure by observing the tendency of a user to respond to another user. Researchers also have explored the propagation of messages in a Twitter media network (Ye et al. 2010), using simple metrics such as follower count and reply count to measure influence. Another research project studied social unrest through social media, highlighting the need to explore the underlying dynamic processes and characteristics of political entrepreneurship (Hua et al. 2013) and security informatics (Chung 2012).

**Proposed Approach**

While existing works have examined a wide range of methods, approaches, and applications, there is little work done to analyze the influence of users in a large-scale online social network of importance to international security, the U.S. government, and policy makers and legislators. We focus our data collection and analysis on the U.S. immigration reform, which has been undergoing significant changes in the year of 2013. Since the beginning of 2013, the U.S. immigration policy has experienced the most far-reaching reform in the history of the country that will affect the economy, national security, and foreign policies (Bush et al. 2009). As nearly twelve million undocumented immigrants reside in the United States, the U.S. Government and Congress are under tremendous pressure of developing timely
legislations for reforming immigration policies. For example, a mass protest occurred on October 8, 2013 in Washington D.C. (see Figure 1), where hundreds of demonstrators rallied in support of immigration reform, resulting in arrest of at least 100 people, several of whom were U.S. legislators including Reps. Luis V. Gutierrez (D-Ill.), Charles B. Rangel (D-N.Y.), John Lewis (D-Ga.) and Jan Schakowsky (D-Ill.) (Constable et al. 2013). This immigration reform promises to enhance the security of the U.S. border, because terrorists, drug traffickers, and illegal gang members constantly seek to infiltrate the U.S. homeland.

Figure 1. A mass protest in Washington D.C. in support of immigration reform

Data Collection

We developed a tweet crawler to automatically collect tweets related to a set of queries about the U.S. immigration reform. These queries were formulated by reviewing a wide collection of documents, books, papers, government publications, and press releases (Bush et al. 2009; Gans et al. 2012; LeMay 2004; U.S. Commission on Immigration Reform 1994; United States. Bureau of Immigration and Customs Enforcement. 2003; United States. Dept. of Homeland Security. Office of Inspector General. 2011) and by testing their usage empirically on Twitter. Key terms that return mainly relevant results are adopted. Terms that return too few results (e.g., fewer than 50 per day) or a significant portion of non-relevant results (e.g., greater than 20%) are removed. Because of these requirements, some terms are too broad to be used, such as “US border security” that also produces results about Russia’s border security and Brussel’s border security, whereas some terms are too narrow, such as “immigrant assimilation.” Between May 21, 2013 and November 15, 2013, our automatic crawler collected 909,035 tweets posted by over 300,000 users. On average, each tweet is associated with 1.1 query terms.

Analysis

In addition to collecting the textual content of the tweets, we analyzed the network relationship among users. We define a network link as a connection between two users who are associated in one of the following ways:

• (1) Targeted tweet (TT): When a user sends a tweet with a target recipient (with an @ symbol in front of a user name placed in the message), then a “targeted tweet” link is created. This type of tweet occurs when the user wants to connect another specific user, showing either a challenge or a desire to relate.

• (2) Retweet (RT): When a user re-sends a tweet without a specific target user as the recipient, the tweet is labeled as a retweet. Retweets often indicate endorsement of an opinion expressed in the tweet being retweeted.
(3) Modified tweet (MT): When a user modify the content of a tweet and sends out the modified tweet, the tweet is labeled as a modified-tweet, which often indicates the user's different opinion or his additional arguments. Modified tweets often represent disapproval or alternative opinion.

Examples of some of these tweets are as follows:

Retweet Example:

*From the @nytimes: "Thousands Rally Nationwide in Support of an Immigration Overhaul"* [http://t.co/3dpnoi4uqJ #TimeIsNow. October 8, 2013.]

Modified Tweet Example:

*SEE THE PATTERN? MT* @ByronYork:=>@charliespiering: Park Service OKs immigration reform rally on 'closed' Nat'l Mall. http://t.co/An81aBuDSE, October 8, 2013.

The 36-day time frame was a sliding window that we used to study evolution of network activities over time. The first network spans the 36 days beginning May 21 (the first day that we collected tweets) onwards. Using the network relationship, we computed the Betweenness Centrality to represent the influence score of each user. The *Betweenness Centrality* of a node *i* measures the proportion of the number of shortest paths that pass through node *i* to the number of all shortest paths between a pair of users (Jackson 2008; Scott 2000). Besides its intuitive application to identifying users serving as bridges in online social networks, betweenness centrality has been shown to provide a higher performance of ranking of prominence than other widely-used metrics (e.g., closeness, degree, in-degree) (Adali et al. 2013).

**Empirical Findings**

To enable visualization and summarization of the collected data, we developed an interface that presents to users daily counts of the collected tweets as well as top 10 leaders in different categories.

Figure 2 shows summaries of the tweet counts (overall, by weekday, by week, and by time of day).

Among the list of influential users who got the highest betweenness centrality, CAP Immigration is ranked top with an influence score more than double of the second most influential user (Alan Gomez, a journalist specializing on U.S. immigration reform). Table 1 provides a sample of the top 10 most influential users, identified by betweenness centrality scores.

**Figure 2. Summaries of Tweets Collected**
The leaders identified in Table 1 represent influential participants because of their strategic positions and their active participation in the Twitter network of U.S. immigration reform discussion. By studying the agendas and activities of these leaders who have high influence scores, policy makers can identify possible trends and movements of the community. Thus the identification can save significant time and resources of polling and gauging opinion from the larger community.

In the sampled results, the users whose tweets were retweeted most frequently by other users are Barack Obama, Charles Garcia, Senator Ted Cruz, The New York Times, and Marco Rubio (Table 2). These users are often major politicians or popular press.

The identification of users in Table 2 as well as those in Tables 3-7 used other metrics such as number of retweets, number of tweets posted, number of opinions posted, number of targeted tweets received, number of followers, and number of friends.

On the other hand, the users who posted most tweets include mjselker (who strongly opposes immigration reform), Ricardo Parra, and Miles Pierce (see Table 3). An interesting finding is the DRM Action Coalition, which supports undocumented youths in the U.S. While these users may not be widely known, but are eager to spend much time on social media to post messages.

A group similar to the most frequent tweeters is called “most opinions expressed” (see Table 4), which features users who post opinion-type tweets. An opinion-type tweet does not link to other tweets but only expresses one’s opinion on some issues. The most opinionated users include Miles Pierce, Benouis Law Office, Linda P, Immigration Law News, and US Senators. Many of these users are related to immigration laws.
Implications

The aforementioned findings demonstrate several research implications and trends in the U.S. immigration reform debate.

First, the number of tweets varied significantly with press release. Our results indicate a high tweet daily counts when approaching and shortly after the dates during which the Senate debated and eventually passed (on June 27, 2013) the comprehensive immigration reform bill. These counts drop significantly during July-September when momentum of passing the bill at the House stalled. These counts picked up again in October when mass protests for immigration reform occurred again in Washington D.C. and other locations.

Second, the lists of top leaders change drastically when new tweets were posted and new connections were made with other top leaders and influential users. For example, tweets posted by such users as Senator Ted Cruz and Barack Obama often spurred waves of new tweets, retweets, and modified tweets, changing the ranks of these users in the top leader lists. Fourth, the overall tweet count has increased significantly in June, followed by a gradual decrease in the beginning of July as fewer new developments of immigration reform were reported.

Third, the use of different influence metrics produced significant different ranked lists of users, prompting for different interpretation of the ranks. In general, the list produced by using betweenness centrality (Table 1) indicates the top users who connect the crowds through civil groups, mass media, and political affiliation. Other lists (Tables 2-7) provide different perspectives of leaderships, each identified by the way users behave on the social media community. These different perspectives allow analysts and decision makers to choose the suitable way to identify influential users.

Conclusion

In this paper, we developed an approach to identifying influential users in a large social network of interest to policy makers and the general public. We present findings from an empirical study using data collected on the U.S. immigration reform discussion posted on Twitter. Influential users were identified using various metrics and techniques, demonstrating different ways that policy makers can use to find key opinion leaders in an evolving online social community. This research should contribute to providing a new case and new empirical findings of applying influence analytics to analyzing social media networks, and has strong implications on predictive analytics, business intelligence (Chung 2014; Chung et al.
2009), BI visualization (Chung et al. 2005), and social media analytics (Zeng et al. 2010). Our ongoing works include comparing different statistics over different periods of immigration reform developments, developing and validating influence computation, evaluating different influence metrics, analyzing the spreading of viral messages, incorporating sentiment and other analyses to understand mass social movements using the current study as a case.

Acknowledgments

This paper is based upon work supported partially by funding from the U.S. Department of Homeland Security (through the 2013 Summer Research Team Program for MSI), by the National Center for Border Security and Immigration at the University of Arizona, by the Stetson University’s School of Business Administration (SBA) and the SBA Board of Trustees, and by the Center for Business Intelligence and Analytics at Stetson University (http://cbia.stetson.edu/). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DHS. We thank Mr. Daniel Ballard, the AMCIS track chairs, reviewers, and coordinators for their assistance and valuable suggestions.

References


