Analyzing Spread of Influence in Social Networks for Transportation Applications

Pritesh Pimpale   Anand Panangadan
Department of Computer Science
California State University, Fullerton
Fullerton, California 92831, USA
Telephone: +1-657-278-3998
Fax: +1-657-278-7168
pritesh.pimpale@csu.fullerton.edu
apanangadan@fullerton.edu

Lourdes V. Abellera
Civil Engineering Department
California State Polytechnic University, Pomona
Pomona, California 91768, USA
Telephone: +1-909-869-4863
Fax: +1-909-869-4342
lvabellera@cpp.edu

***Abstract*—The spread of influence on social networks has been extensively studied but has been primarily demonstrated on large networks such as author collaboration networks. It is not clear how well these approaches translate to real-world social networks comprised of users discussing region-specific topics related to transportation. This work attempts to utilize the concept of influence in social media to identify the individuals who exert the most influence on transportation-related topics. The work describes a web-accessible system that enables a user to select transportation-related topics and then retrieve in real-time the most influential potential influential individuals and/or organizations on Twitter with respect to that specific topic. Influence is quantified using two simple measures of influence, the number of Twitter mentions and the number of retweets of a message. The locations of these influential entities are then indicated on a web-accessible front-end using Google Maps. Such a tool can eventually be used for many purposes including limiting misinformation and encouraging acceptance of a new transportation products or service.**

***Index Terms*—public transportation, machine learning, classification, natural language processing, sentiment analysis, crowdsourcing.**

I. Introduction

In southern California, public transit is an unpopular mode of transportation [1]. There are many reasons for this sentiment. Communities and business districts are so spread out that one person might live in Los Angeles and work in Claremont, a suburb separated by over 50 km. It is generally impractical to ride the train or the buses in this situation, unless, for example, the MetroLink can be a convenient alternative. People in southern California generally believe that public transportation is slow, not reliable, not safe, and not clean. Some of these beliefs have an actual basis, and some do not. For example, reliability depends on a particular route, and safety depends on the area surrounding a station. Most people who do not usually ride public transportation will never be able to experience its benefit if they hear negative experiences and views of other riders. Consider the following situation. An unpleasant experience happens to a rider, for instance, if she missed her plane due to the Los Angeles Airport FlyAway shuttle being late. The disgruntled passenger, especially if young, might tell her story in Twitter or Facebook, or write about the unpleasant experience in her blog. Consequently, her followers will take note of this experience, and will affect their decisions to take or not take the LAX FlyAway shuttle for their next trip to the airport depending on the amount of influence the dissatisfied passenger has exerted on them.

978-1-5386-4649-6/18/$31.00 ©2018 IEEE

Meanwhile, the use of online social media by the general public has been rapidly increasing, with an increasing number of people receiving news primarily from social media and forming opinions based on social media messages [2]. The goal of our work is to exploit online social media to enhance the public's perception of transportation services in the hope of increasing the number of people using shuttles, buses, and trains and others who use other sustainable transportation alternatives such as biking and walking. If more people participate in sustainable transportation, then there will be fewer cars that contribute to the production of greenhouse gases (GHGs) [3]. However, these benefits are not fully realized because of the continued under-utilization of transit capacity [4], [5].

The spread of influence on social networks has been extensively studied but has been primarily demonstrated on large networks such as author collaboration networks [6], [7]. It is not clear how well these approaches translate to real-world social networks comprised of users discussing region-specific topics related to transportation. In this project, we therefore attempt to analyze the concept of influence in social media, in particular, the Twitter social media site, and identify the individuals who exert the most influence on transportation-related topics. Utilizing spread of influence as a method to advertise the benefits of public transit, carpooling, biking, and walking challenges the status quo because the method is extremely different from the traditional ways of disseminating information.

In our preliminary work, we quantify influence using two simple measures of influence, the number of Twitter mentions and the number of retweets of a message. Currently, we have developed a web-accessible system that enables a user to select transportation-related topics and then retrieve in real-time the most influential members on Twitter with respect to that specific topic. We have completed a qualitative analysis of the results returned by the system. We eventually plan to incorporate more sophisticated measures of influence for this application.

The rest of the paper is organized as follows. Section II lists related work in use of social media for promoting public transportation. Section III describes the major parts of the software system. In Section IV, we discuss our current results. We give our conclusions and plans for future work in Section V.

II. Related Work

Merriam-Webster defines influence as “the power to change or affect someone or something”. Keller and Berry [8] identified the characteristics of 10% of the Americans who tell the rest what to buy, which political figures to vote for, and where to travel. With social media platforms, the influence exerted by one user becomes even more significant as the user has numerous venues to express his/her content in the form of tweets, blogs, videos, and images.

Facebook, Twitter, YouTube, Instagram, LinkedIn, and ResearchGate are popular examples of social media platforms. In social media, users share aspects of their personal life, pass on or report news and other information, voice their opinions on a current topic, and ask for or offer an advice about a product or service. A user can also scout for a potential employee or request for an academic paper. Participants of social media form virtual communities and networks of people with similar interests and goals. The Pew Research Center has conducted a study of social media usage from 2005 to 2015 [9] resulting in numerous statistics that show social media’s ubiquity. Therefore, social media is an immense venue for entities (individuals, groups, large organizations) to exert their influence, whether intentional or not, upon other users. There already exist services to compute for an entity’s influence such as Klout and PeerIndex.

There are several studies that use social media to assess public perception and sentiment regarding public transit. Bertrand et al. [10], using geospatial tools, discovered through Twitter that public sentiment is generally positive in public parks and negative at transportation hubs. Collins et al. [11] use sentiment analysis methods on Twitter messages to identify specific problems (e.g., fires) on public transportation in real-time.

Twitter is one of the most prominent micro-blogging services available. On this platform, users can send and read “tweets” which are short messages with a maximum limit of 140 characters (recently raised to 280 characters). Every day, approximately 500 million tweets are sent [12]. Advertisers of new products, presidential campaigners, and the like want to reach Twitter audience to push their content and influence the beliefs and actions of users. As such, various researchers have investigated the concept and spread of influence in Twitter using actual data [13] or models [14]. Many researchers define how to quantify influence. Twitter collects data from users and one of the common ones is the “follower”. Twitter applies a social-networking model known as “following”. Here, a user can select who he/she wants to “follow” to receive tweets from. The consent of the user to be followed is not required. One obvious way of defining influence is the number of followers. However, Cha et al. [15] have shown that in-degree, the number of people who follow a user, is not proof of influence. Weng et al. [16] found that users follow those users who follow them back out of politeness and not influence; this reciprocity is called homophily. This phenomenon has also been observed by Aral et al. [17]. Retweets, the number of times others “forward” a user’s tweet, has also been considered a measure of influence. Another parameter is mentions, which is the number of times others mention a user’s name. Anger and Kittl[18] also listed other proposed measures of influence such as Follower/Following Ratio, Retweet and Mention Ratio, and Interactor Ratio which is the number of individual users who retweet content or mention user X divided by the total number of followers of user X. Only mention and retweet prove to be consistent measures of influence [15], [13]. Hence, we used these two parameters to define our influential individuals, groups, or other entities in developing our tool. A retweet is associated with the value of the content, while a mention is related to the importance of the user’s name [18].

III. Methodology

Initial approach: We initially applied topic modeling and stochastic cascade modeling techniques to harvested tweets. We attempted to find the source A of a transportation-related tweet and how that tweet is propagating through the social network by finding the user B that retweeted the original tweet of A to users C and D. However, after the third level, it became difficult to determine how the tweet is propagating because some relevant information is not revealed by Twitter due to privacy issues. The propagation of a message in a social network has been studied to maximize the spread of influence [6], [15] and to limit the spread of misinformation [7]; however, these efforts assume a theoretical model of the network still in the theoretical stage.

As discussed in the Introduction section, the definition and measurement of influence has been studied and applied to Twitter datasets. None of these involved a transportation-related topic. Again, we define influence as mentions and retweets, and below is how we counted these parameters in our code and deployed the web application.

A. Software framework

The software system is divided in 3 parts:

1) Twitter stream capture - Python and MongoDB
2) Java REST API - Java
3) Web app front end - HTML and JavaScript (AngularJS)

Real time tweets are captured using the Twitter API with the help of the python library Tweepy[1]. The tweepy library helps in maintaining a continuous http connection with the server. The data obtained from Twitter is in JSON format which can be stored in a NoSQL database. MongoDB is used for the storing the tweets. It provides a document-oriented data structure which helps in directly storing the twitter data without the need for any change in format. It also provides drivers for Python and Java which we have used for our application.

The Server side REST API for the computation of the influential twitter handles is implemented in Java. Java provides the HashMap data structures which we have used for computing the influential twitter handles using the number of re-tweets and mentions. JAX-RS is used for developing the REST API for the web app in Java. AngularJS[2] is chosen as the client side script for the client web application. It provides a rich framework for developing single page applications with communication with server using http calls. It enables interfacing with the UI (User Interface) framework like Angular Material which is used for creating the UI of the web application[3].

B. Visualization

Visualization using Geographic Information Systems (GIS) has traditionalle been implemented with software such as ArcGIS . However, we realized ArcGIS might be unavailable for smaller transit agencies. Hence, we use Google Maps for this purpose. By using the free Google Maps API, we embedded the Google Maps site into our web application. The locations of the influential individual or organization and who they have influenced are shown in Google Maps. As Google Maps is a popular service (more than 1,000,000 web sites [19] use the Google Maps API), users are expected to be already familiar with its symbols (e.g., Google Map pin, map view, satellite view, terrain view) and functions (e.g., zoom in, zoom out).

IV. Results and Discussion

A. California High Speed Rail

The web application data is dynamic since it continually accumulates tweets. The results presented below are a snapshot of the retweets and mentions at a particular instant in time. To demonstrate how the web application works, we use the topic "High speed rail" as searched on July 26, 2016. Figure 1 presents the screen-shot of the web application showing the three available topics that the user can choose. These topics can be modified or removed, and more topics can be added. (The Google Maps pins, the inverted-drop-shaped icons that mark locations in Google Maps, show the locations of a previous search.) Note that this screen-shot is just a portion of the complete result of the search.

The user selects the topic "High speed rail", then clicks "Mentions", then clicks "FIND INFLUENCERS". Figure 2 shows the results of these steps. We observe that the Washington Examiner has the most number of mentions (i.e., 19 mentions). The Washington Examiner is an American political journalism website and weekly magazine based in Washington, D.C. Clicking on the icon for this Twitter user, Washington Examiner, results in the view shown in Figure 3. This view shows the tweets mentioning "California's bullet train on the track to extinction" of Washington Examiner. Aside from this mention, we can see that some tweeters also add their own sentiment.

Next, clicking on "Retweets", then "FIND INFLUENCERS", results in the viwe shown in Figure 4. This view shows that Scott Walker has the most number of retweets at 68, followed by DownsizeTheFeds (21 tweets) which ties with Paul Rogers also at 21 tweets. The applications lists only the first 10 entities. Click on user Scott Walker displays only the one tweet that was retweeted 68 times – "This is why I called the 'high speed rail' line between Milwaukee and Madison a boondoggle" (Figure 5). It is interesting to note that the Google Maps pin points to Wisconsin. This is because Scott Walker is the governor of Wisconsin and his tweet is about a recent news item about the California high speed rail. Table I shows the complete results of finding influencers using the number of mentions; Table II shows the corresponding influencers using the number of retweets. In these two tables, we observe that there are more retweets than mentions suggesting that content is more important than the name of the user in determining influence. While 2 to 3 mentions or 6 to 9 retweets are hardly indicative of influence, these results indicate the potential of this approach. Our tool lists the top 10 potential influencers; this number can be easily modified in the software.

It is interesting to note that Scott Walker, the governor of Wisconsin, is more influential (with 68 retweets) than Jerry Brown, the governor of California (with 3 mentions). It is reasonable to assume that the California governor is much more familiar with California issues than the Wisconsin governor. However, the Wisconsin governor is probably more well-known for the topic of bullet trains. For example, it is possible that Jerry Brown will be more influential for the keyword "traffic."

Finally, we observe that the tool finds relevant tweets with high accuracy with the current list of keywords. The user "CA Water 4 All" hardly qualifies as an interested entity in high speed rail. However, upon examining the individual tweets, we find that this user writes about California projects, which according to his/her opinion, are a waste of money. An example tweet is "#California's biggest boondoggles: The Delta tunnel projects and high-speed rail". Another interesting user is the Google Play Music. His/her inclusion in the resulting list is not an error. An example tweet is "I bet it'll be 2020 B4 @GooglePlayMusic has a sleep timer and alarm clock feature... California will have bullet trains 1st at this pace!". This tweet is not actually about the California

[1] http://www.tweepy.org/

[2] https://angularjs.org/

[3] https://material.angularjs.org/latest/

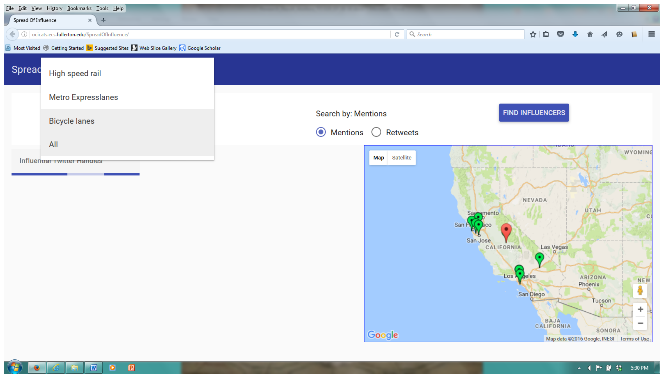
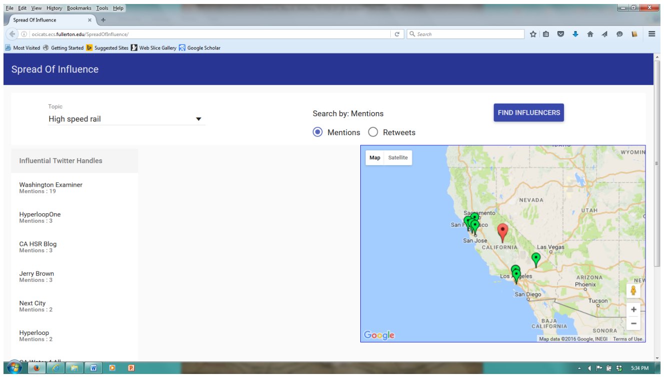


Fig. 1. Available topics in the web application.



Influential Twitter Handles

Washington Examiner
Mentions : 19

HyperloopOne
Mentions : 3

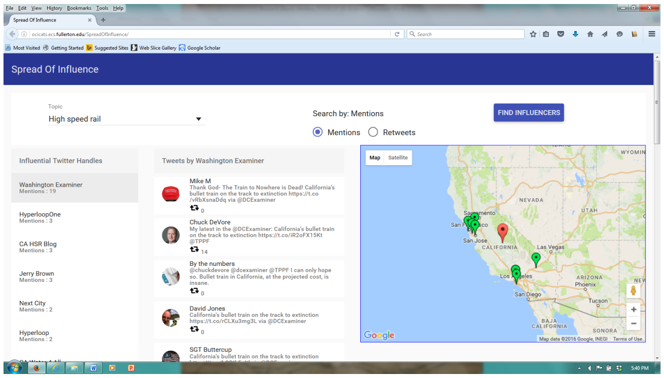
CA HSR Blog
Mentions : 3

Jerry Brown
Mentions : 3

Next City
Mentions : 2

Hyperloop
Mentions : 2

Fig. 2. Result of searching for mentions for the topic “High speed rail”



Influential Twitter Handles

Washington Examiner
Mentions : 19

HyperloopOne
Mentions : 3

CA HSR Blog
Mentions : 3

Jerry Brown
Mentions : 3

Next City
Mentions : 2

Hyperloop
Mentions : 2

Tweets by Washington Examiner

Mike M
Thank God- The Train to Nowhere is Dead! California's bullet train on the track to extinction https://t.co/vRbXsnaDdq via @DCExaminer
0

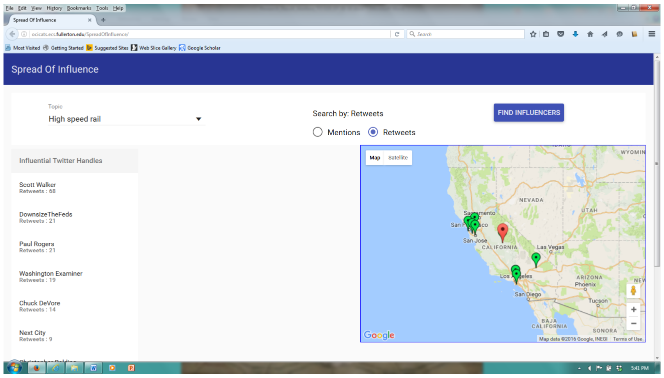
Chuck DeVore
My latest in the @DCExaminer: California's bullet train on the track to extinction https://t.co/iR2oFX15Kt @TPPF
14

By the numbers
@chuckdevore @dcexaminer @TPPF I can only hope so. Bullet train in California, at the projected cost, is insane.
0

David Jones
California's bullet train on the track to extinction https://t.co/rCLXu3mg3L via @DCExaminer
0

SGT Buttercup
California's bullet train on the track to extinction

Fig. 3. The tweets mentioning the tweets by Washington Examiner, the entity with the highest number of mentions



Influential Twitter Handles

Scott Walker
Retweets : 68

DownsizeTheFeds
Retweets : 21

Paul Rogers
Retweets : 21

Washington Examiner
Retweets : 19

Chuck DeVore
Retweets : 14

Next City
Retweets : 9

Fig. 4. Result of searching for retweets for the topic “High speed rail”

high speed rail, but it shows the popularity of this topic. The location of this particular tweet is Newcastle, Australia. The only irrelevant tweet included in the results is one tweet by Paul Rogers with the tweet “California imposing $45 fee on rail cars carrying toxic chemicals to fund emergency response”.

B. Strengths and Limitations

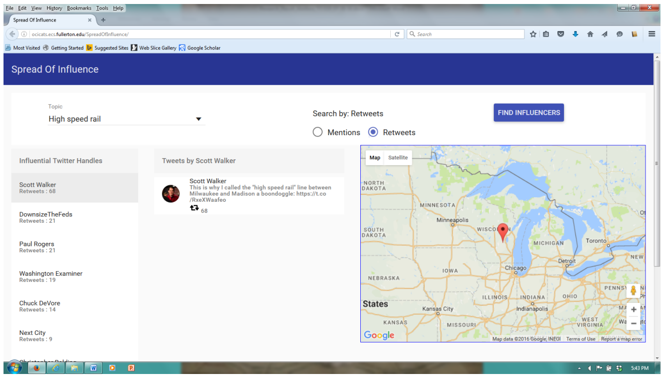
One of the strengths of the code we developed is that it gathers tweets free of charge. Commercial providers sell Twitter data. The search for the period from 1/1/2015 to 3/15/2016 retrieved approximately 35,000 tweets over the 435-day period and which would incur cost to purchase this dataset. Because the primary target for the proposed tool is transit agencies and not for-profit businesses, the system was designed to not require any expensive source of data. The Twitter API harvests data only from the last seven days. However, with the code we developed, it is continuously harvesting tweets from the date it was deployed (Jun 9, 2016). (The parameters of the code can easily be modified.) The standard APIs used in this application are all free and instructions can be easily accessed online. Our method made a connection between the Twitter API and Google Maps API, which are standard, accessible APIs that are already proven

TABLE I
SEARCH RESULTS OF USING MENTION AS THE MEASURE OF INFLUENCE

User	Number of mentions
Washington Examiner	19
HyperloopOne	3
CA HSR Blog	3
Jerry Brown	3
Next City	2
Hyperloop	2
CA Water 4 All	2
Texas Public Policy	2
Kira Fucker @home	2
Google Play Music	2

TABLE II
SEARCH RESULTS OF USING RETWEETS AS THE MEASURE OF INFLUENCE

User	Number of retweets
Scott Walker	68
DownsizeTheFeds	21
Paul Rogers	21
Washington Examiner	19
Chuck DeVore	14
Next City	9
Christopher Balding	9
janewells	9
Dagwood Bumstead	7
TED	6



Spread Of Influence

Topic
High speed rail

Search by: Retweets
○ Mentions ● Retweets

Influential Twitter Handles

Scott Walker
Retweets : 68

DownsizeTheFeds
Retweets : 21

Paul Rogers
Retweets : 21

Washington Examiner
Retweets : 19

Chuck DeVore
Retweets : 14

Next City
Retweets : 9

Tweets by Scott Walker

Scott Walker
This is why I called the "high speed rail" line between Milwaukee and Madison a boondoggle: https://t.co/RxeXWaafeo
68

Fig. 5. The tweet by Scott Walker that was retweeted 68 times

TABLE III
KEYWORDS USED TO CAPTURE THE LIVE TWEETS FROM TWITTER

#caltrans
caltrans
California transportation
California traffic
California cars
California rail
#expresslanes California
fastrak California
bicycle lanes California
California trains
California freight
high speed rail California
#hsr California
bullet train California
expresslanes California
transponder California
#bicyclelanes California

successes. Our methodology is also able to produce results relevant to the chosen topic with high accuracy.

One of the limitations of the tool is that we do not have a comprehensive list of transportation topics. Table III is a list of the keywords used to capture the live tweets from Twitter in the current version of the code. The three topics currently in the web application are just a subset of the topics we assigned. Transportation planners may be looking for very specific topics such as “Metro annual budget 2016” or “Caltrans 710 houses”. In this case, we can add these new keywords in the code. The system will the harvest the tweets for these additional keywords starting from the date those keywords were added.

The second limitation is that tweets with no user-specified location are not included in the results. This will limit the number of tweets processed. The third limitation is that it shows results from users that have provided their location outside of California. It is possible that (1) the user has moved recently to California and has not updated his/her Twitter account, or (2) the user does not live in California, but tweeting about a topic in California. There is no way to determine which of the two situations is correct. This situation can be handled in different ways. First, the transportation planner using the web application can disregard the tweet. As the locations of the Twitter users are shown in Google Maps, it is easy to see the user’s location. However, we can also argue that we can use this to influence California residents. For example, New York City is a public transit-friendly city, and those who tweet from New York may see the California High Speed Rail positively and tweet accordingly. As Southern California is spread out, many residents are averse to taking public transit. However, tweets from a New Yorker may influence them to take public transit. The last limitation is that it may be difficult to actually find these influential individuals if we choose to contact them. Twitter has control over privacy so the actual names and addresses of these influential people will be difficult to obtain. However, for organizations such as non-profits or newspapers, this may not be an issue.

V. CONCLUSIONS AND FUTURE WORK

Using Twitter data, we developed a tool for generating a list of potential influential individuals and/or organizations for particular transportation-related topics by counting the number of mentions of a specific Twitter user and retweets of a particular tweet. Their locations are indicated in Google Maps. We believe our tool will advance the state of the practice. Our tool can be used for many purposes including limiting misinformation and encouraging acceptance of a new transportation product or service.

In future work, we plan to use more sophisticated measures of influence on social networks that have been previously demonstrated on author-collaboration networks. We will also evaluate the approach on larger numbers of messages downloaded from the Twitter platform and that cover a wider variety of transportation-related topics. Finally, we will perform a detailed quantitative analysis of the results to judge the usefulness of this approach.

ACKNOWLEDGMENT

This material is based upon work supported by the University of California Center on Economic Competitiveness in Transportation (UCCONNECT). Students Nanwarin Chantarutai and Diana Lin contributed to this work.

REFERENCES

[1] L. J. Nelson and D. Weikel, “Billions spent, but fewer people are using public transportation in southern california,” *Los Angeles Times*, pp. 12–44, 2016.

[2] E. SHEARER, “News use across social media platforms 2017,” *Pew Res Cent*, 2017.

[3] C. A. Kennedy, “A comparison of the sustainability of public and private transportation systems: Study of the greater toronto area,” *Transportation*, vol. 29, no. 4, pp. 459–493, 2002.

[4] H. Min, *Assessing the Comparative Efficiency of Urban Mass Transit Systems in Ohio: Longitudinal Analysis*. Mineta National Transit Research Consortium, College of Business, San Jose State University, 2013.

[5] A. T. Murray, “Strategic analysis of public transport coverage,” *Socio-Economic Planning Sciences*, vol. 35, no. 3, pp. 175–188, 2001.

[6] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[7] C. Budak, D. Agrawal, and A. El Abbadi, “Limiting the spread of misinformation in social networks,” in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 665–674.

[8] E. Keller and J. Berry, *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy*. Simon and Schuster, 2003.

[9] A. Perrin, “Social media usage,” *Pew Research Center*, 2015.

[10] K. Z. Bertrand, M. Bialik, K. Virdee, A. Gros, and Y. Bar-Yam, "Sentiment in new york city: A high resolution spatial and temporal view," *arXiv preprint arXiv:1308.5010*, 2013.

[11] C. Collins, S. Hasan, and S. V. Ukkusuri, "A novel transit rider satisfaction metric: Rider sentiments measured from online social media data," *Journal of Public Transportation*, vol. 16, no. 2, p. 2, 2013.

[12] "Twitter usage statistics," http://www.internetlivestats.com/twitter-statistics, 2016.

[13] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert, "The influentials: New approaches for analyzing influence on twitter," *Web Ecology Project*, vol. 4, no. 2, pp. 1–18, 2009.

[14] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 65–74.

[15] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy." *Icwsm*, vol. 10, no. 10-17, p. 30, 2010.

[16] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[17] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 544–21 549, 2009.

[18] I. Anger and C. Kittl, "Measuring influence on twitter," in *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. ACM, 2011, p. 31.

[19] K. Hoetmer and M. Marks, "Google maps: into the future," in *presentation at Google I/O conference, Google Inc*, vol. 15, 2013.