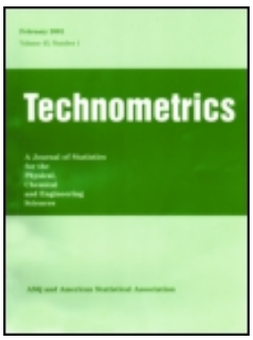

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Measuring Influence of Users in Twitter Ecosystems Using a Counting Process Modeling Framework

Donggeng Xia, Shawn Mankad & George Michailidis

To cite this article: Donggeng Xia, Shawn Mankad & George Michailidis (2016) Measuring Influence of Users in Twitter Ecosystems Using a Counting Process Modeling Framework, Technometrics, 58:3, 360-370, DOI: 10.1080/00401706.2016.1142906

To link to this article: https://doi.org/10.1080/00401706.2016.1142906

Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH

Measuring Influence of Users in Twitter Ecosystems Using a Counting Process Modeling Framework

Donggeng XIA*

University of Michigan
Ann Arbor, MI 48109
(*donggeng@umich.edu*)

Shawn MANKAD*

Department of Operations,
Technology, and Information Management
Cornell University
Ithaca, NY 14853
(*spm263@cornell.edu*)

George MICHAILIDIS*

University of Michigan
Ann Arbor, MI 48109
(*gmichail@umich.edu*)

Data extracted from social media platforms are both large in scale and complex in nature, since they contain both unstructured text, as well as structured data, such as time stamps and interactions between users. A key question for such platforms is to determine influential users, in the sense that they generate interactions between members of the platform. Common measures used both in the academic literature and by companies that provide analytics services are variants of the popular web-search PageRank algorithm applied to networks that capture connections between users. In this work, we develop a modeling framework using multivariate interacting counting processes to capture the detailed actions that users undertake on such platforms, namely posting original content, reposting and/or mentioning other users' postings. Based on the proposed model, we also derive a novel influence measure. We discuss estimation of the model parameters through maximum likelihood and establish their asymptotic properties. The proposed model and the accompanying influence measure are illustrated on a dataset covering a five-year period of the Twitter actions of the members of the U.S. Senate, as well as mainstream news organizations and media personalities. Supplementary material is available online including computer code, data, and derivation details.

KEY WORDS: Algorithms; Massive datasets; Multivariate analysis.

1. INTRODUCTION

Leading business and nonprofit organizations are integrating growing volumes of increasingly complex *structured* and *unstructured* data to create big data ecosystems for content distribution, as well as to gain insights for decision making. A recent, substantial area of growth has been online review and social media platforms, which have fundamentally altered the public discourse by providing easy to use forums for the distribution and exchange of news, ideas and opinions. The focus in diverse areas, including marketing, business analytics and social network analysis, is to identify trends and extract patterns in the vast amount of data produced by these platforms, so that more careful targeting of content distribution, propagation of ideas, opinions and products, as well as resource optimization is achieved (Probst, Grosswiele, and Pfleger 2013; Dave 2015).

One platform that has become of central importance to both business and nonprofit enterprises is Twitter. According to its fourth quarter 2014 financial results announcement, Twitter had more than half a billion users, out of which more than 288 million were active ones (Twitter Inc. 2014). Although Twitter lags behind in terms of active users to Facebook, it is nevertheless perceived by most businesses and nonprofit organizations as an integral part of their digital presence (Bulearca and Bulearca 2010).

The mechanics of Twitter are as follows: the basic communication unit is the account. The platform allows account users to post messages of at most 140 characters, and thus has been described as the Short Message Service (SMS) of the Internet. As of mid-2014, over half a billion messages were posted on a daily basis. Further, Twitter allows accounts to "follow" other accounts, which means the follower receives notification whenever the followed account posts a new message. Thus, the follow-follower relations serve as a primary channel for content to spread within the social networking platform. Accounts tend to interact with each other over these channels in two directed ways. First, an account can *copy* or *rebroadcast* another account's tweet, which is referred to as a "retweeting." Second, an account can *mention* another account within a tweet by referring

*All authors contributed equally to this work.

TECHNOMETRICS, AUGUST 2016, VOL. 58, NO. 3
DOI: 10.1080/00401706.2016.1142906
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/tech*.

to their account name with the @ symbol as a prefix. These two actions, retweeting and mentioning, are directed responses from one account to another and thus, provide the mechanisms for online conversation.

The mechanics of Twitter, together with the original messages generated by users, give rise to rich Big Data. Specifically, the content of the message, together with easily searchable key terms or topics that use the "hashtag" symbol # as a prefix, constitute a large corpus of unstructured text. The hashtag function enables searches to identify emerging themes and topics of discussion. In 2014, more than 2.1 billion search queries were generated (Twitter Inc. 2014). Further, the following built-in capability, creates a network for *potential information flow and dissemination*, while the retweeting and mentioning actions create subnetworks of *actual interactions* between user accounts.

A key problem in all social networking platforms is that of identifying *user influence*, since such users are capable of driving action (e.g., steer discussions to particular themes and topics) or provoking interactions among other users and thus, are also potentially more valuable to businesses (Trusov, Bodapati, and Bucklin 2010). Thus, the ranking of Twitter users based on their influence constitutes both an active research topic and a business opportunity, as manifested by services that market and sell to businesses and other organizations influence scoring metrics. The most standard metric employed is the number of followers an account has. However, a number of studies (Cha et al. 2010; Weng et al. 2010) have concluded that the count of followers is not a good indicator, since most followers fail to engage with the messages that have been broadcast. For that reason, the number of retweets an account receives (Kwak et al. 2010) is a better measure of influence. Since we are interested in ranking of users, more sophisticated influence measures based on the popular PageRank (Page et al. 1999) and HITS (Kleinberg 1999) ranking algorithms, widely used for ranking search results on the Web, have been used (Haveliwala 2003; Kwak et al. 2010; Weng et al. 2010; Gayo-Avello, Metaxas, and Mustafaraj 2011). However, these algorithms have been developed for and applied to the followers network, which clearly captures the general popularity of users, but not necessarily of their influence. For example, Cha et al. (2010) found that the 20 most globally followed accounts are composed mostly of athletes and celebrities, with the exceptions being President Obama and two news sources (CNN and the New York *Times*).

In this article, we propose to measure an account/user's influence on the Twitter social media platform, by taking into consideration both their ability to produce new content by posting messages, but also to generate interactions from other accounts through retweeting and mentioning. To that end, we build a statistical model for an account's actions and interactions with other accounts. It uses a counting process framework to capture the posting, retweeting and mentioning actions. In addition, based on this model we introduce a novel *influence measure* that leverages both the follower network (that captures the potential for posted messages to generate interactions with other users) and the *intensity* over time of the basic actions involved (posting, retweeting and mentioning). Hence, underlying the model in this article is the idea that conversations, and in particular the rate of directed activity, between accounts reveal their real-world position and influence.

We illustrate the modeling framework on a closely knit community, namely that of the members of the United States Senate, the upper legislative house in the bicameral legislative body for the United States. Two senators are democratically elected to represent each state for six-year terms. We further augment the set of Twitter accounts analyzed by including selected prominent news organizations (e.g., *Financial Times*, *Washington Post*, CNN), as well as popular bloggers (e.g., Nate Silver, Ezra Klein), the accounts of President Obama and the White House, and two influential federal agencies (the U.S. Army and the Federal Reserve Board); for details refer to Section 7. Thus, we examine an ecosystem of key participants that influence the political conversation and discourse of the country.

We identify particular senators and news agencies that tend to elicit interactions from other accounts, thus revealing their influence on Twitter. Our results in Section 7 further indicate that the proposed approach produces influence measures for the U.S. Senators that correspond more closely with their legislative importance than purely network-based solutions based on the PageRank algorithm.

The remainder of the article is organized as follows: in Section 2, we review recent literature on measuring influence in online social networks. In Section 3, we introduce the modeling framework and the proposed influence measure. Section 4 presents the algorithm to obtain the model parameter estimates, as well as establish their statistical properties and those of the influence measure in Section 5. The performance of the model is evaluated on synthetic datasets in Section 6, while the U.S. Senate application is presented in Section 7. Finally, some concluding remarks are drawn in Section 8.

2. BACKGROUND AND LITERATURE REVIEW

There has been a great deal of work on ranking nodes in online social networks by their influence motivated by fundamental questions in marketing, such as how to identify the best set of users to create cascades or viral campaigns. Probst, Grosswiele, and Pfleger (2013), in an extensive survey article, found that the most common measures to quantify the influence of a certain node are completely based on network topology and fail to account for "further characteristics of influential users" or the actual dynamics on the social network. They identify several papers that propose variations to the core idea of measuring influence with network metrics of the followers network. An illustration of the standard methodology with similar data is Dubois and Gaffney (2014), where Canadian political communities on Twitter are explored using degree, clustering coefficient, and other network metrics calculated from the followers network to identify "opinion leaders," that is, accounts that steer online conversations.

To create a more nuanced influence measure that addresses the challenges highlighted by Probst, Grosswiele, and Pfleger (2013) and references therein, researchers have begun to use the content of the communication like the underlying topic or theme of conversation, which allows for more realistic models, since some individuals are authoritative or receptive to others only along certain topical dimensions. As such, a number of recent works have extended the classical network topology measures to account for topic of conversation. Haveliwala (2003) and Weng

et al. (2010) take into account topic similarity of the actual messages and the social link (followers network) structure via modified PageRank algorithms that are applied to the followers network. Barbieri, Bonchi, and Manco (2013) proposed a similar idea for the related problem of identifying the optimal choice of initial users for inducing cascades. The model we propose relates to these previous works by also separating behavior according to the topic of conversation. Our contribution lies in measuring influence with actual conversation dynamics by combining the mentions and retweets along different topics with the followers link structure.

Our approach extends recent work in the statistics community, which uses counting processes to combine conversation dynamics (mentions and retweets) with the followers network structure. In this stream of literature, the hazard rate represents a measure of influence and typically quantifies the effect of a message from one node on each of its followers (Du et al. 2012; Gomez-Rodriguez, Leskovec, and Schölkopf 2013). Thus, as in Perry and Wolfe (2013), Hunter et al. (2011), and Butts (2008), the interactions between nodes are modeled as *independent* counting processes. The model posited in this work exhibits certain key differences, as illustrated in Figure 1, because the hazard rate of a node to retweet or mention is a function of the cumulative effect of tweets from its followers. The use of *interacting* counting processes is an important modeling nuance, since it allows for more realistic account behavior. For instance, accounts that are very popular and receive many tweets on the same topic within a short period of time usually respond once both out of convenience and to avoid spamming their followers. Thus, the model we posit should result in more accurate influence measures for Twitter ecosystems like the U.S. Senate that we investigate in Section 7.

Another closely related work, Aral and Walker (2012), uses an exponential proportional hazard rate model to understand how covariates, like age, gender, marital status, and so on, effect peer adoption of a software application for mobile phones on Facebook. The nuanced differences in underlying data and context lead to different modeling decisions. For instance, Aral and Walker (2012) used a single-failure model, whereas we use a multiple failure framework. Since we are studying conversations, we think of the hazard rate for an account as varying between topics of conversation and related to the number of actions seen on this topic. Aral and Walker (2012) take another approach and model the hazard rate as a function of covariates typically captured in meta-data, like gender, age, etc. We assign each account two parameters that describe its ability to generate responses and its susceptibility to respond to others. These two parameters capture information closely related to leadership skills, authority levels and communication ability, which may be difficult to quantify in typical meta-data drawn from social media platforms.

3. THE MODEL AND THE INFLUENCE MEASURE

We start our presentation by defining some key quantities for future developments. Let $G = (V, L)$ denote the followers network, where V corresponds to the set of nodes of all the Twitter accounts under consideration and $L = \{L_{i,j}, 1 \leq i \neq j \leq n\}$ the edge set between them and captures whether an account follows another account. Note that the network is bidirectional in nature and not necessarily symmetric, since account i may follow account j, but not vice versa. In principle, L can be dynamically evolving, but in this work we consider L to be static and not changing over time. As explained in the introductory section, in the Twitter platform, accounts (nodes) can undertake the following three actions: post a new message, retweet a message posted by another account that they follow and finally mention another account that they follow. Further, the vast majority of messages posted, retweeted or mentioned have key terms (with a # prefix) that identify the topic(s) that are discussed.

Next, we define the following two key counting processes. Let $N_j(t, l)$ denote the total number of retweets and mentions that account j generates on topic l by time t and let $A_j(t, l)$ denote the total number of posted messages by account j on topic l by time t. Define α_j to be a parameter that captures the long-term capability of account j to generate responses by other accounts from the content posted, and β_j a parameter that captures the long term susceptibility of account j to respond (retweet/mention) to the postings of the accounts it follows. In this article, we mainly focus on $N_j(t, l)$, since it reflects the interactions between accounts, while $A_j(t, l)$ is related to general posting habits. We model $\{N_j(t, l)\}_{i=1}^n$ as a set of counting processes through their hazard rates, using a version of Cox (Cox 1972) proportional hazard model; specifically, the hazard rate $\lambda_{j,l}(t)$ of process $N_j(t, l)$ is given by

$$\lambda_{j,l}(t) = \lambda_{0,l}(t) \exp\left(\sum_{i \neq j} L_{ij}(\alpha_i + \beta_j) \log(M_i(t, l) + 1)\right), \tag{1}$$

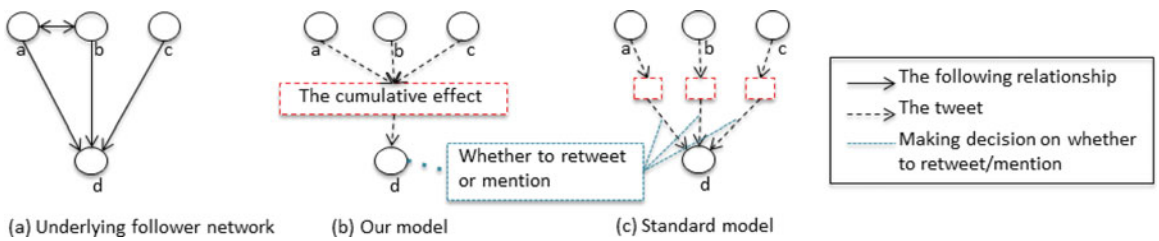

Figure 1. Solid lines in panel (a) represent edges in the follower's network. Panel (b) illustrates the proposed model, where node d decides to retweet or mention by the cumulative effect of the three tweets from nodes a, b, and c. Panel (c) illustrates the standard counting process model on interactions between nodes. Instead of considering the cumulative effect of the three tweets, node d makes a decision on whether to respond (retweet or mention) three separate times.

where

$$M_i(t,l) = (N_i(t,l) + A_i(t,l))I(N_i(t,l) + A_i(t,l) \le F) + F \cdot I(N_i(t,l) + A_i(t,l) > F).$$

$A_j(t,l) + N_j(t,l)$ is the total number of posting, retweets and mentions for account j on topic l by time t. And we consider the effect of seeing actions from account i can get saturated when the total number of actions reaches the constraint, F. We assume that the parameters $\alpha_i, \beta_i \in (-\infty, \infty)$, since accounts and their users may be positively or negatively inclined toward other accounts, as well as being more keen in joining specific conversations or passively retweeting messages from favorite accounts. The nonparametric baseline component $\lambda_{0,l}(t)$ is time varying. In general, we would expect this baseline to be small for large times t, since topics in social media platforms have a high churn rate; they become "hot" and generate a lot of action over short time scales and after awhile it stops being discussed (Kwak et al. 2010). The model posits that account j interacts with other accounts at a baseline level $\lambda_{0,l}(t)$, modulated by its ability to generate responses by accounts in its followers network, as well as its own susceptibility to respond to accounts it follows postings and rebroadcasting of messages. Note that we model the retweet-mention process $N_j(t,l)$, since it reflects interactions between nodes and use the total activity process $M_j(t,l)$ as a covariate.

To complete the modeling framework, denote the set of topics in the data as $\{1,\dots,\Gamma\}$. Further, let $\mathcal{T}_j^l = \{T_{j,1}^l, \dots, T_{j,n_j^l}^l\}$, $t = 1,\dots,n_j^l$, denote the set of time points that account j took action (post, retweet, mention) on topic l, until our end of observation time point t_0. Finally, for identification purposes, we require one member of the parameter vector $\Omega = (\alpha_1, \alpha_2, \dots, \alpha_n, \beta_1, \dots, \beta_n)$ to be set to a fixed value, and without loss of generality we set $\alpha_1 = 0$. Following, Andersen and Gill (1982) and Cox (1972), we employ a partial-likelihood function to obtain estimates of Ω. Specifically, we treat the baseline $\lambda_{0,l}(t)$ as a nuisance parameter and decomposing the full-likelihood to obtain

$$PL(t) = \prod_{1\le l\le\Gamma}\left(\prod_{1\le j\le n}\prod_{1\le k\le n_j^l, T_{j,k}^l\le t}\frac{\lambda_j(T_{j,k}^l)}{\sum_{1\le i\le n}\lambda_i(T_{j,k}^l)}\right).$$

Plugging the exact form of the hazard rate from (1) into the partial-likelihood function (PL), we get:

$$PL(t) = \prod_{1\le l\le\Gamma}\left(\prod_{1\le j\le n}\prod_{1\le k\le n_j, T_{j,k}^l\le t} \times \frac{\exp\left(\sum_{i\ne j} L_{ij}(\alpha_i+\beta_j)\log(M_i(T_{j,k}^l,l)+1)\right)}{\sum_{1\le i\le n}\exp\left(\sum_{u\ne i} L_{ui}(\alpha_u+\beta_i)\log(M_u(T_{j,k}^l,l)+1)\right)}\right). \quad (2)$$

3.1 The Influence Measure

Leveraging the structure of the model, we propose to measure an account's (node) influence as the total hazard rate change it will bring to its followers. Specifically, for an account j its relative hazard rate (ignoring the baseline) at time t is given by: $H_j = \exp(\sum_{k\ne j}\log(M_k(t,l)+1)L_{kj}(\alpha_k+\beta_j))$. Further, the contribution of node i is $H_j^{(i)} = \exp(\log(M_i(t,l)+1)L_{ij}(\alpha_i+\beta_j))$. Then, after some algebra we obtain that the total hazard rate change i brings to its followers can be written as

$$TH^{(i)} = \sum_{j\ne i} L_{ij}\cdot\exp\left(\log(M_i(t,l)+1)(\alpha_i+\beta_j)\right). \quad (3)$$

Since $M_i(t,l)$ is a random value, we approximate it by its observed average value, $\bar{M}_i$, calculated from the data over all time points and topics. Hence, the influence measure becomes

$$\tilde{TH}^{(i)} = \sum_{j\ne i} L_{ij}\cdot\exp\left(\log(\bar{M}_i+1)(\alpha_i+\beta_j)\right). \quad (4)$$

Finally, we express it in a log-scale, so as to linearize the scale and make it compatible with the range of values of the response and susceptibility parameters α and β:

$$\Xi^{(i)} = \log\left[\sum_{j\ne i} L_{ij}\cdot\exp\left(\log(\bar{M}_i+1)(\alpha_i+\beta_j)\right)\right]. \quad (5)$$

In our real data analysis, we work with an estimate of $\Xi^{(i)}$ by plugging in estimated $\hat{\alpha}_i$ and $\hat{\beta}_j$ values.

4. COMPUTATION AND INFERENCE

Next, we present a Newton-type algorithm for computing the parameter estimates Ω. The logarithm of the partial likelihood function (2) is given by

$$LL(t) = \log(PL(t)) = \sum_{1\le l\le\Gamma}\left\{\sum_{1\le j\le n}\sum_{1\le k\le n_j^l, T_{j,k}^l\le t}\sum_{i\ne j} L_{ij}(\alpha_i+\beta_j) \times \log(M_i(T_{j,k}^l,l)+1) - \sum_{1\le j\le n}\sum_{1\le k\le n_j^l, T_{j,k}^l\le t} \times \log\left[\sum_{1\le i\le n}\exp\left(\sum_{u\ne i} L_{ui}(\alpha_u+\beta_i)\log(M_u(T_{j,k}^l,l)+1)\right)\right]. \quad (6)$$

The objective function corresponds to $LL(t_0)$, which considers all events k in its Equation (6). For the sake of notational simplicity, we will use LL to represent $LL(t_0)$ in the rest of the article. Due to its smoothness, we employ Newton's algorithm that uses the gradient and the Hessian of LL. The detailed expressions for the gradient vector $G \equiv \nabla_\Omega LL$ and the Hessian $H \equiv \nabla_\Omega\nabla_\Omega(LL)$ are given in Section S1 of the supplementary material.

To speed up calculations, we take advantage of the structure of the problem by precomputing and storing several quantities for repeated use, thus saving on computational time in practice. Further details are explained in the supplementary material.

The steps of the optimization are given in Algorithm 1. As stated in the algorithm, s is a positive constant to judge the convergence of the the Newton's algorithm. The computational complexity of this algorithm is dominated by the computation of H. Denote by $m_n = \max_{1\le j\le n}\{n_j\}$. Based on

Algorithm 1 Estimating the parameters by Newton's algorithm

1: Initialize the vector Ω value by $\alpha_1 = \cdots = \alpha_n = \beta_1 \cdots = \beta_n = 0$
2: Define s as a positive thresholding constant for the minimum step size
3: **while** $\tau > s$ **do**
4: Calculate G by using (1) and (2) in the supplementary material.
5: Calculate H by using (3) to (8) in the supplementary material.
6: Find the optimum positive τ value such that $\Omega - \tau \cdot H^{-1}G$ will maximize the log-partial-likelihood (6)
7: Update $\Omega \leftarrow \Omega - \tau \cdot H^{-1}G$.
8: In the updated Ω, set $\alpha_1 = 0$.
9: **end while**
10: **return** Ω

Equations (1) and (2) in the supplementary material, it costs $O(\Gamma n m_n)$ operations to calculate an entry of G. Further, since G is of dimension $2n$, it takes $O(\Gamma n^2 m_n)$ to obtain the entire G vector. Analogously, based on Equations (3) to (8) in the supplementary material, it costs $O(\Gamma n m_n)$ operations to calculate an entry of H, if proper book-keeping is used on the results obtained for the gradient G. Further, since H is of dimension n^2, it takes $O(\Gamma n^3 m_n)$ to obtain the entire H matrix. Hence, the overall time complexity for each iteration of the algorithm is of the order $O(\max\{\Gamma n^3 m_n\})$. The time complexity for the whole algorithm is then $O(\max\{\Gamma n^3 m_n R\})$, where R is the number of repetitions needed for the algorithm to converge, which depends on the threshold s. When setting $s = 10^{-3}$ in our simulations in Section 6 and real data analysis in Section 7, we find the algorithm generally converges within 10 repetitions.

5. PROPERTIES OF THE $\hat{\Omega}$ ESTIMATES

Next, we establish that the estimator $\hat{\Omega}$ which maximizes (6) will converge to the true parameter Ω in probability under certain mild regularity conditions.

Conditions

A. (Bounded hazard rate) $C_0 \le \lambda_{0,l}(t) \le C_1$ for $0 \le t \le t_0$ $1 \le l \le \Gamma$,
B. (Bounded parameters) $\max_{1\le i,j\le n}\{|\alpha_i|, |\beta_j|\} \le C_2$,
C. (Limited posting frequencies)

$$P\left(A_j(t+h,l) - A_j(t,l) \ge 1\right) \le C_3 \cdot h,$$
$$P\left(N_j(t+h,l) - N_j(t,l) \ge 1\right) \le C_3 \cdot h, \quad (7)$$

when $t, h \ge 0, t + h \le t_0$.
D. (Positive definite limit of Hessian) Let Ω' be any choosable parameter vector satisfying (B). For large enough Γ and some C_4, we have the holding condition to hold on the smallest eigenvalue of $-\nabla_{\Omega'}\nabla_{\Omega'} LT(\Omega', t)$, at $\Omega' = \Omega, t = t_0$,

$$P\left(\lambda_{\min}\left(-\nabla_{\Omega'}\nabla_{\Omega'} LT(\Omega', t)\right)|_{\Omega'=\Omega, t=t_0} > C_4\right) \to 1, \text{ as } \Gamma \to \infty,$$

where

$$LT(\Omega, t) \equiv \Gamma^{-1}\left\{-\sum_{l=1}^{\Gamma}\lambda_{0,l}(u)\log\left\{\sum_j \exp\right.\right.$$
$$\left.\times\left(\sum_{i\ne j} L_{ij}(\alpha_i' + \beta_j')\log(M_i(t,l)+1)\right)\right\}$$
$$\left.\cdot\left(\sum_{j=1}^{n}\exp\left(\sum_{1\le i\le n, i\ne j} L_{ij}(\alpha_i+\beta_j)\log(M_i(t,l)+1)\right)\right)\right\} \quad (8)$$

Theorem 1. Consider the four Conditions A, B, C, and D above, with C_0, C_1, C_2, C_3, and C_4 are all positive constants. Under these conditions, we have:

$$\hat{\Omega} \to_P \Omega \text{ as } \Gamma \to \infty.$$

The detailed proof is given in the supplementary material.

Conditions A and B of Theorem 1 are fairly mild and should be satisfied in most applications. Condition C is more involved, but Lemma 1 gives an example of a counting process for which it naturally holds. On the other hand, it is challenging to impose conditions on the basic processes under which Condition D will hold, and consequently, as shown in Section 6, we propose to verify it empirically.

Lemma 1. When both $A_j(t,l)$ and $N_j(t,l)$ are both Poisson processes and the hazard rate of $A_j(t,l)$ is smaller than a constant K, Condition C in Theorem 1 is satisfied.

The detailed proof is also presented in the supplementary material.

Based on Theorem 1, by leveraging the properties of continuous functions, we can establish the consistency of the proposed influence measure.

Proposition 1. Let $\Xi(t) = (\Xi^1(t), \ldots, \Xi^n(t))$ denote the n-dimensional vector of influence measures at time t. Further, denote by $\hat{\Xi}(t) = (\hat{\Xi}^1(t), \ldots, \hat{\Xi}^n(t))$ their empirical estimates. Under the conditions of Theorem 1, we have that

$$\left\|\hat{\Xi}(t) - \Xi(t)\right\| \to_P 0 \quad (9)$$

for any $t \ge 0$.

Based on Theorem 1, the proof of the proposition is straightforward, since each element of the vector $\hat{\Xi}$ is a continuous function of $\hat{\Omega}$.

6. PERFORMANCE EVALUATION

In this section, we evaluate the proposed model and influence measure on synthetic data. We start by outlining the data generation mechanism.

Step 1: Building the followers network L.

The tasks employed for Step 1 are presented next.

- First, for each node i, generate $K_1(i)$ from a uniform distribution on the integers $\{1, \ldots, K\}$, where $K = \lfloor *n/2 \rfloor$ and $\lfloor *\cdot \rfloor$ is the floor function that returns the maximum integer not larger than the value inside.
- Generate $F_1(i)$ for node i by randomly sampling $K_1(i)$ users from $\{1, \ldots, n\}\backslash\{i\}$. If $k \in F_1(i)$, let $L_{ik} = 1$, $1 \le i \le n$.

- For each node j, sample $K_2(j)$ uniformly from the set $\{1, \ldots, K\}$. Generate $F_2(i)$ for node j by randomly sampling $K_2(j)$ users from $\{1, \ldots, n\}\backslash\{j\}$. If $k \in F_2(j)$, let $L_{kj} = 1, 1 \le j \le n$.

At the end of this procedure, every node in the network has at least one follower and at least an account that it follows.

Step 2. Generate the post, retweets and mentions sequences.

Since the baseline hazard rate $\lambda_{0,l}(t)$ always gets canceled out within the partial-likelihood function (2), we select $\lambda_{0,l}(t)$ as $\lambda_{0,l}(t) = a$, whenever $0 \le t \le t_0$ and $\lambda_{0,l}(t) = 0$ when $t > t_0$, where a is a small positive constant.

We then generate actions on this network with Algorithm 2 below for each topic $l \in \Gamma_1$ or Γ_2. In this algorithm, we first let each node send out a number of tweets with distribution Binomial(J, p) at $t = 0$. Then we generate the retweets and mentions in the standard survial analysis way, by using the hazard rate (1), as in the algorithm below.

Algorithm 2 Generate Group A actions

```
Initialize Indicator which is the sequence to record the nodes that have mentioned or retweeted as an empty sequence.
Initial t = 0. Let each node has a tweet with probability p.
Let each node send out tweets from Binomial(J, p).
while t < t_0 (stopping time for all topics) do
    Generate survival time for each node with its hazard rate (1)
    Find node i with the shortest time t_s.
    if t + t_s < t_0 then
        Update t to be t + t_s. Record the node that has done this retweet or mention.
    end if
    if t + t_s > t_0 then
        Break
    end if
end while
return Indicator
```

We first illustrate the performance of the Newton estimation algorithm, on a random network of varying size. We set the parameter $a = 0.5$ for the baseline hazard rate and choose a time horizon of $t_0 = 7$, to emulate a week's worth of data. We also select the parameters Ω uniformly at random in the interval $[-0.3, 0.3]$.

Due to the bounded baseline hazard rate and simulated parameters, and since the retweets and mentions are generated as Poisson, Condition A, B, C of Theorem 1 have been satisfied. Then we empirically "check" Condition D. With a large $\Gamma = 1000$, network size $n = 10, 50$, we repeated Step 1 and 2 for 20 times to simulate the network and actions. In each repetition, the square root of the smallest eigenvalue of $\lambda_{\min}(-\Gamma\nabla_{\Omega'}\nabla_{\Omega'}LT(\Omega', t))|_{\Omega'=\Omega, t=t_0}$ is computed. The results are plotted in Figure 2, where it can be seen that smallest eigenvalues of $[\lambda_{\min}(-\Gamma\nabla_{\Omega'}\nabla_{\Omega'}LT(\Omega', t))|_{\Omega'=\Omega, t=t_0}]^{1/2}$ are generally large and greater than 0.5.

Then as we have verified all conditions are satisfied with network sizes $n = 10, 50$ and $\Gamma = 1000$, we plot in Figure 3 the mean squared error of the parameter and influence estimates $\frac{\|\hat{\Omega}-\Omega\|}{\sqrt{2n-1}}$ and $\frac{\|\hat{\Xi}-\Xi\|}{\sqrt{n}}$ to check the performance of our estimation algorithm, where $\|\cdot\|$ corresponds to the ℓ_2 norm of a vector. The results are based on 20 replicates of the underlying followers networks, as well as the actions (postings, retweets and mentions) data.

It can be seen that the quality of the estimates improves as a function of the number Γ of topics discussed, while it deteriorates as a function of the number of nodes in the followers network L. Another way to look at the quality of the estimates, is to examine the relative error of the parameter and influence estimates, given by $\frac{\|\hat{\Omega}-\Omega\|}{\|\Omega\|}$ and $\frac{\|\hat{\Xi}-\Xi\|}{\|\Xi\|}$.

It can be seen in Figure 4 that especially the influence measure which is of prime interest in applications, exhibits a small (less than 10%) relative error rate.

Next, we use a size 10 network, specially constructed to gain insight into the workings of the proposed influence measure. The settings for the data generation are as follows: $\Gamma = 500$, $\alpha_1 = 0, \alpha_2 = -2, \alpha_3 = \cdots = \alpha_{10} = 0.2$ and $\beta_1 = \cdots = \beta_{10} = 0$. Finally, the topology of the followers networks is given in Figure 5.

Since $\alpha_2 = -2$, node 2 is an "unpopular" one and hence can hardly generate any retweets and mentions of its postings. On the other hand, all nodes have approximately an equal number of followers, which suggests that their ranking according to the PageRank metric (or many other popular ones based on that network like Haveliwala (2003) and Weng et al. (2010)) will be approximately similar. The results based on a single realization of the user actions data generation process is shown in Figure 6. It can be seen that relying on the followers network structure

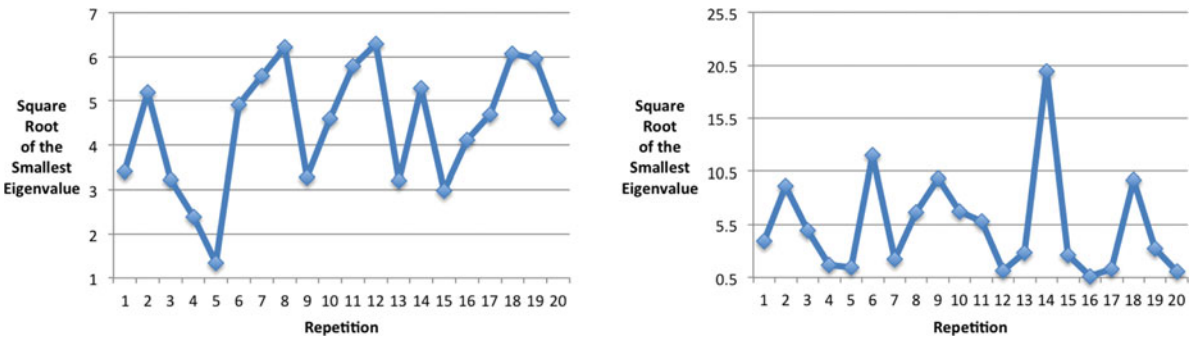

Figure 2. Diagnostics for Condition D with the simulated data: $[\lambda_{\min}(-\Gamma\nabla_{\Omega'}\nabla_{\Omega'}LT(\Omega', t))|_{\Omega'=\Omega, t=t_0}]^{1/2}$ at $\Gamma = 1000$, $n = 10$ (left) and $n = 50$ (right). Due to large variations, the square root of the smallest eigenvalues is shown for better visualization.

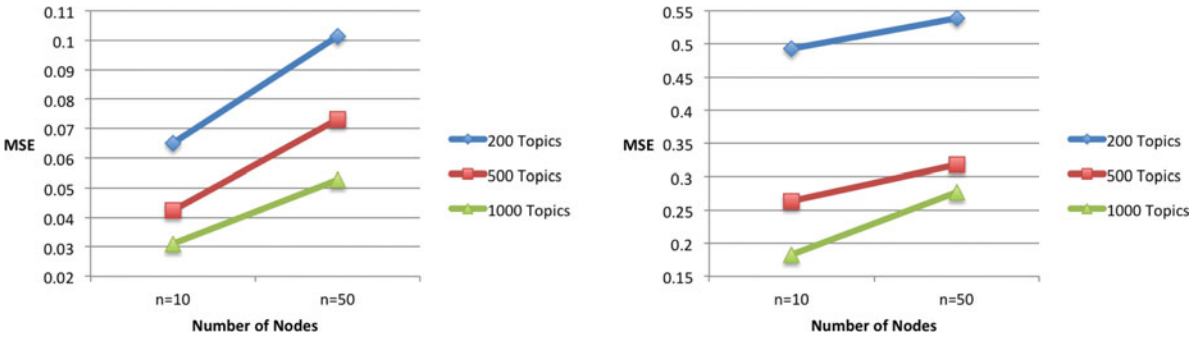

Figure 3. Mean squared error of the model parameter estimates Ω (left) and Ξ (right).

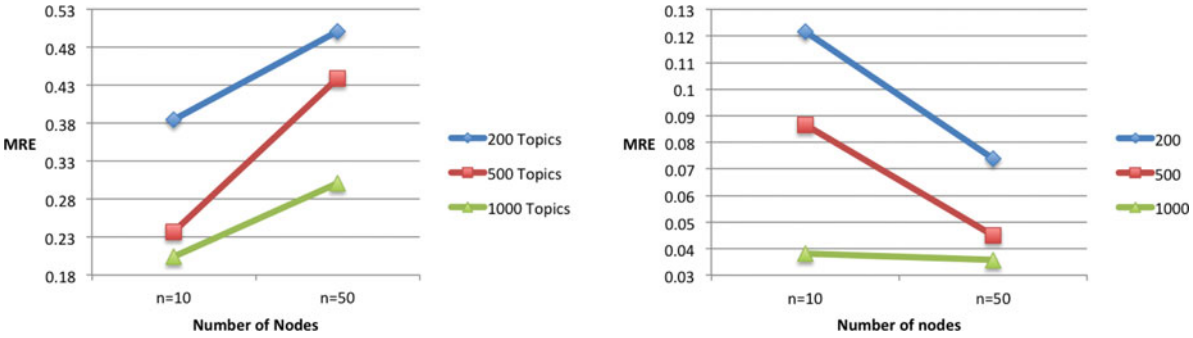

Figure 4. Mean relative error of the model parameter estimates Ω (left) and Ξ (right).

gives a false impression, while the proposed influence measure that incorporates the actions of the accounts provides a more insightful picture.

7. IDENTIFYING INFLUENTIAL SENATORS

Tweets and follower lists are collected using Twitter's API and consist of approximately 200,000 tweets and 4671 follower links within the set of 120 accounts from April 2009 to July 2014. Accounts are registered to 55 Democratic politicians (U.S. Senators and the President of the U.S.), 46 Republican Senators, 2 government organizations (U.S. Army and the Federal Reserve Board), and 16 media outlets, including newspapers (Financial Times, *Washington Post*, *New York Times*, Huffington Post), television networks (MSNBC, Fox News, CNN, CSPAN), reporters (Nate Silver (538), Ezra Klein) and television hosts (Bill O'Reilly, Sean Hannity). The top panel of the figure shows some periods of increased activity, as in the months surrounding the inauguration of President Obama (January 2013), the debate on raising the debt ceiling of the U.S. government and its temporary suspension around April 2013 and the summer of 2014 (soccer World Cup). Note that the sudden increase during the summer of 2014 may be an artifact of rate limiting data acquisition. Specifically, Twitter's API allows access to only the past 3000 tweets for any account. As a consequence, for extremely high volume

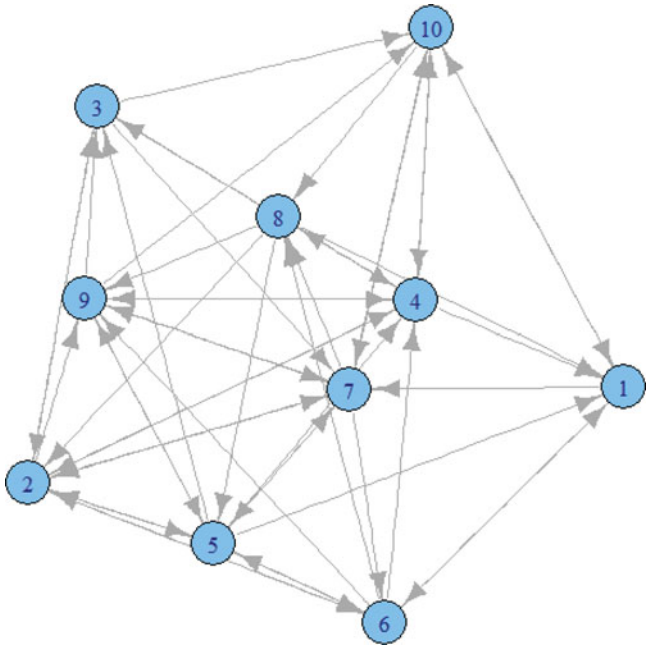

Figure 5. Artificial topology of a plot with "unpopular" node.

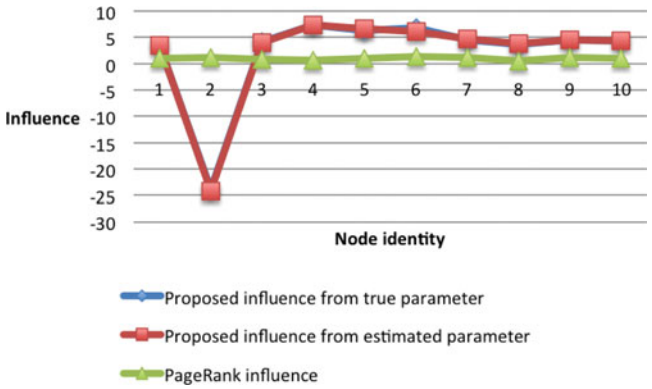

Figure 6. Proposed influence VS PageRank Influence, in a plot with "unpopular" node.

Table 1. Actual tweets mentioning or retweeting the most influential accounts over from May 15, 2014 to July 3, 2014

Date	Account	Tweet
05/19/2014	Menendez	"@SenBlumenthal & in #NJ the avg student loan debt is over $29K. It's unacceptable! #GameofLoans http://t.co/hUJMSeJbfd"
05/23/2014	Cornyn	"RT @nytimes: Former Defense Secretary Gates Is Elected President of the Boy Scouts http://t.co/C7STUSVIP3"
05/27/2104	Blumenthal	"RT @msnbc: @SenBlumenthal calls for reviving gun reform debate after mass shooting near Santa Barbara: http://t.co/7sqtf1IAFy"
06/02/2014	Markey	"RT @washingtonpost: A huge majority of Americans support regulating carbon from power plants http://t.co/lj6ieL5D1Y http://t.co/2CA63hTqmm"
06/17/2014	Markey	"Proud to intro new bill w @SenBlumenthal 2 protect domestic violence victims from #gunviolence http://t.co/MsgK40oLiT http://t.co/ynEHrEbh2x"
06/20/2014	Blumenthal	"Proud to stand w/ @CoryBooker & others on enhancing rules to reduce truck driver fatigue. Their safety & safety of others is paramount. -RB"
06/20/2014	Markey	"Proud to support our workers and this commonsense bill w @SenatorHarkin Keeping Track: Overtime Pay, via @nytimes http://t.co/TnAS96Hro5"
06/25/2014	Durbin	"Watch now: @OfficialCBC @HispanicCaucus @CAPAC @USProgressives @SenatorCardin on racial profiling #MoreThanAProfile http://t.co/ZX0Eu65dgi"
06/25/2014	Cardin	"RT @TheTRCP: Thank you @SenatorCardin for standing with sportsmen today for #CleanWater #protectcleanwater"
06/27/2014	Markey	"Thanks @alfranken @CoryBooker @amyklobuchar @SenBlumenthal for joining me in support of community #broadband http://t.co/O8Px2MzrCg"
06/27/2014	Menendez	"Took my first #selfie at #NJs @ALJBS! Hope @CoryBooker is proud of his NJ Sen colleague. http://t.co/FrEJonUy9d"
06/28/2014	Booker	"Thanks Adam RT @AIsaacs7 Props to @CoryBooker and @SenRandPaul for their bipartisanship in introducing their amendment #MedicalMarijuana"

users, like newspapers and television networks, our data traces their Twitter usage for months. For the least active users in our data, 3000 tweets dates back multiple years.

An inspection of actual tweets in Table 1 shows, consistent with Golbeck, Grimes, and Rogers (2010), that senators tend to retweet and mention as a means of self or legislative promotion. In fact, we see a number of references to legislative activity, such as calls for gun reform, carbon emissions, and references to actual bills on overtime pay, domestic violence protections, among others. Senators often cite news coverage by retweeting or mentioning news media accounts that support their political agenda, which would suggest that the media outlets collectively have enormous influence. This also suggests that Twitter is used by senators as part of a larger strategy to build and coalesce public support to pass bills through congress.

To test these hypotheses and also rigorously compare the proposed influence measure to PageRank applied to the followers networks (which constitutes the backbone of many ranking algorithms of Twitter accounts), we perform a regression analysis to assess how well each measure explains *legislative leadership* in Congress. Our response variable is the leadership score, published by Govtrack (GovTrack.us 2014). GovTrack creates the leadership score by applying the PageRank algorithm to the adjacency matrix of bill cosponsorship data. Thus, the leadership score for each senator is a number between 0 and 1, where higher values denote greater ability to find successful cosponsors (collaborators) within the senate. The regression model we are interested in is

$$\text{Leadership} = \beta\text{Influence} + \Theta\text{Controls}, \tag{10}$$

where Influence contains the proposed measure and/or PageRank, and Controls includes party affiliation, gender, age, and number of years in the senate. Seniority endows a number of benefits including preferential assignment to committees. Thus, these control variables likely associate strongly with legislative leadership.

To estimate the proposed influence measure, the data is organized into weekly intervals after using the follow-follower relations to construct the adjacency matrix L. As mentioned in Section 1, hashtags can be used as an indicator of different conversations. However, we find that senators do not utilize hashtags often. To overcome this challenge, we follow previous works on Twitter (Hong and Davison 2010; Ramage, Dumais, and Liebling 2010) by applying probabilistic topic modeling, which was first introduced in Blei, Ng, and Jordan (2003). Due to space constraints, for statistical and algorithmic details on the topic model, see Blei (2012) and Blei, Ng, and Jordan (2003) and references therein.

Topic modeling results in a probabilistic decomposition of tweets by topic (the underlying content) $P(\text{topic}|\text{tweet})$, which is appropriate since a single tweet could touch on multiple issues. Cross-validation is applied to the entire dataset of tweets (ignoring time) to select 10 topics as optimal. Tweets are then assigned to topics that had $P(\text{topic}|\text{tweet}) \geq 0.25$. Finally, topics from each week are treated as unique, since the micro content tends to churn rapidly, for example, "the military" may be a consistent discussion topic, but from week to week the specific conversation changes from military action in Iraq and Afghanistan, to appropriations, to potential actions in Ukraine, etc. Given the length of the entire data (over 5 years), we arrive at 2770 topics overall. After preprocessing, we apply Algorithm 1 to estimate the α and β parameters for every account using all data. The final influence measure is constructed by computing the influence measure vector $\hat{\Xi}$ over different time intervals to

Table 2. Top 10 rankings according to the proposed model and PageRank from May 15, 2014–July 3, 2014

Rank	Proposed measure	PageRank
1	Financial Times	Financial Times
2	Washington Post	U.S. Army
3	NYTimes	CNN
4	MSNBC	Barack Obama
5	Ezra Klein	CSPAN
6	Fox News	New York Times
7	Cory Booker	Washington Post
8	Ben Cardin	Cory Booker
9	Nate Silver (538)	MSNBC
10	Richard Blumenthal	Wall Street Journal

study how influence evolved; that is, $\hat{\Xi}$ was computed by using the average of $M_i(\mathcal{T}_m, l)$ over all time points in $\mathcal{T}_m$ and topics, where $\mathcal{T}_m$ denotes the mth time interval of interest.

The first time interval $\mathcal{T}_1$ we investigate is May 15, 2014–July 3, 2014, which captures the most active period in our data and also represents a period when rate limiting is not a concern, that is, the data for even high volume users extends this far. During this time several major events occurred worldwide, including the soccer World Cup, debate on immigration reform, and the Islamic State in Iraq and the Levant (also known as the ISIS or ISIL) began an offensive in northern Iraq. Table 2 shows the top 10 most influential accounts under the proposed method and PageRank (Page et al. 1999) calculated from the followers network. Both methods estimate that the Financial Times is the most influential Twitter account, and in general find that the media has an enormous influence that facilitates online conversation between politicians. We see from Figure 7 that these top accounts were actively retweeted and mentioned throughout this period.

Next we empirically verify that Conditions A-D for Theorem 1 are satisfied in this data application. Condition A of bounded hazard rates is easily satisfied, since, based on topic division, we can compute that $M_i(t, l) \leq A_i(t, l) + N_i(t, l) \leq 647$ for all topics l and nodes i. After applying our proposed Algorithm 1, we find that all estimated parameters $\alpha_i, \beta_j \in [-3.1, 1.6]$. Thus, if we select $C_2 = 3.2$, Condition B of bounded parameters is satisfied. Checking Condition C is also straightforward. In our observed data, the smallest unit in time is the second. Thus, $h \geq 1/3600$, and Condition C is satisfied by letting $C_3 = 3600$. Finally to check Condition D, for the sake of simplicity, we assume $\lambda_{0,l}(u) = 1$. Then using the estimated α and β values, we plot in Figure 8 the smallest eigenvalue of the Hessian matrix $-LT(\Omega', t)$ at $\Omega' = \Omega$ and $t = t_0$. The

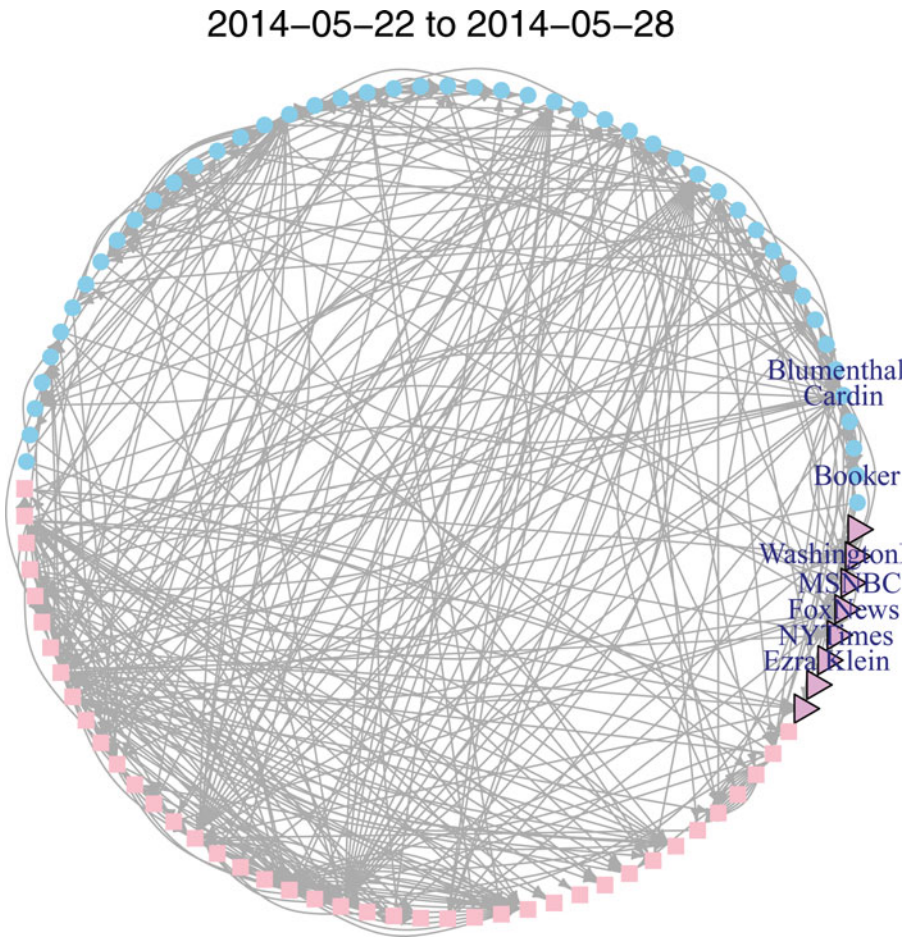

Figure 7. A representative Twitter retweet and mention network drawing from the 2014 summer. Nodes that were active that week and in the top ten most influential accounts are labeled. The nodes (Twitter accounts) contain Democratic senators (circles), Republican senators (squares), media (triangles), and government agencies (stars).

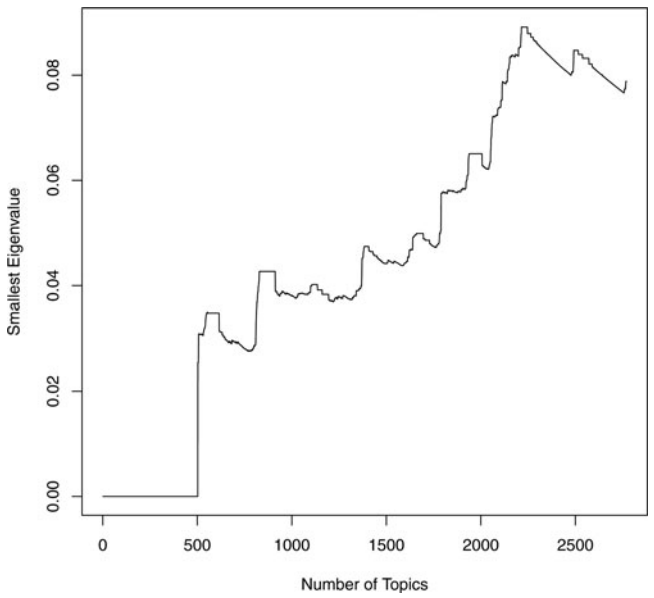

Figure 8. The smallest eigenvalue of the Hessian matrix $-LT(\Omega', t)$ at $\Omega' = \Omega$ and $t = t_0$ is shown for different number of topics (Γ) using the U.S. Senator Twitter data. For large enough Γ, Condition D is satisfied.

figure shows an increasing trend and that Condition D is satisfied when $C_4 = 0.04$ and $\Gamma > 1500$.

Since the assumptions for the model are satisfied, we next estimate the regression model in Equation (10). We note that Senators Baucus, Kerry, Cowan, Lautenberg, and Chiesa are scored by Govtrack, but are not in our analysis. Max Baucus and John Kerry are left out, because they vacated their Senate seats to become, respectively, Ambassador to China and U.S. Secretary of State. Mo Cowan succeeded Kerry and was senator from February 1, 2013, to July 16, 2013, until a special election could be held. Cowan chose not to run in the election. Likewise, due to the death of Senator Frank Lautenberg, Jeffrey Chiesa was appointed by Governor Chris Christie to be the junior senator from New Jersey from June 6, 2013, to October 31, 2013. He declined to run in the special election and thus, is also not included in the analysis.

Since the leadership score provided by GovTrack are between 0 and 1, we estimate two models. One model uses the raw leadership scores, and another uses $\log(\frac{\text{leadership}}{1-\text{leadership}})$ for the response variable. In both cases, as shown in Table 3, we consistently find that the proposed influence measure explains more variation in leadership and when both the proposed and PageRank influence measures are included as independent variables, PageRank does not provide additional explanatory power. Tables 4 and 5 show a significant positive coefficient for the proposed influence measure, meaning that senators who are more influential in Twitter by successfully steering conversation of their colleagues onto particular topics, tend to be more influential in real life in crafting legislation. These results are consistent across different time intervals. For instance, in Section S2 of the supplementary material, we find similar results, where influence is calculated from January 1, 2013, to March 1, 2013, corresponding to sequestration and also from November 1, 2012, to January 31, 2013, corresponding to the President's reelection and subsequent inauguration.

Table 3. Estimated R-squared values for different regression models, where the proposed measure and/or PageRank is included in the set of independent variables and the influence is computed for the entire data sample. We consistently find that the proposed measure is a better indicator of legislative importance

Response	Proposed Measure	PageRank	R^2
leadership	√		0.311
		√	0.276
	√	√	0.311
$\log(\frac{\text{leadership}}{1-\text{leadership}})$	√		0.114
		√	0.098
	√	√	0.114

Table 4. Regression estimates, where the response variable is the raw leadership scores from Govtrack and influence is computed for the entire data sample. $R^2 = 0.311$; $F = 8.228$ on 5 and 92 DF (p-value: 0.000)

Variable	Estimate	Std. Error	t value	$p(>\|t\|)$
Intercept	−0.086	0.232	−0.368	0.714
Proposed Influence	0.062	0.028	2.241	0.027
Republican	−0.154	0.039	−3.945	0.000
Age	0.002	0.003	0.923	0.359
Years in Senate	0.007	0.003	2.518	0.014
Male	0.020	0.050	0.395	0.694

8. DISCUSSION

The goal in this article was to characterize the influence of users in a large scale social media platform when given information about the detailed *actions* users take on it. Our comprehensive analysis of the ecosystem comprising of U.S. Senators and influential government agency and media related accounts demonstrated that conversations, and in particular the rate of directed activity, between accounts are correlated with their real-world position and influence. We expect similar conclusions to hold broadly for other types of directed interaction data when the nodes form a clearly defined ecosystem or closely knit social group/community. For example, one could apply our model and analysis framework to study whether corporate leaders with high external (e.g., Twitter) influence also have higher internal influence for tasks, like corporate planning—a key issue in organizational behavior and management science.

Table 5. Regression estimates, where the response variable is $\log(\frac{\text{leadership}}{1-\text{leadership}})$, where leaderhip is from Govtrack and influence is computed for the entire data sample. $R^2 = 0.114$; $F = 2.334$ on 5 and 92 DF (p-value: 0.048)

Variable	Estimate	Std. error	t value	$p(>\|t\|)$
Intercept	−3.590	2.604	−1.379	0.171
Proposed Influence	0.437	0.308	1.416	0.160
Republican	−1.112	0.438	−2.538	0.013
Age	0.009	0.029	0.323	0.747
Years in Senate	0.034	0.032	1.063	0.290
Male	0.470	0.563	0.834	0.407

The proposed approach only utilizes network information (e.g., followers network), plus time stamps of actions (e.g., retweets and mentions), thus allowing to process a large volume of data. However, it does not consider the tone of the message (positive, negative or neutral), a topic addressed in Taddy (2013), where the goal is to understand how messages related to a specific topic are perceived by other users. Since in that approach the message content needs to be analyzed—a computationally demanding task—Taddy (2013) develops efficient sampling designs for that task. It is of interest though to combine such sampling ideas with the current approach to be able to address user influence issues in very large ecosystems comprising of millions of users.

The modeling and statistical inference issues, associated with large scale data obtained from these social media platforms are different from those in the related literature on network community detection (Kolaczyk 2009; Fienberg 2012; Salter-Townshend et al. 2012), where the goal is to identify relatively dense groups of nodes (users), even though the underlying data (observed adjacency matrices) are the same. Relative to other recent work on modeling directed networks, as in Perry and Wolfe (2013), our study has important modeling differences motivated by the online social media platform domain. For instance, our approach incorporates the fundamental differences between actions like retweeting, mentioning, and posting. As a consequence, our final influence measure, which sums all possible influences from the social network, is able to outperform traditional topology driven approaches like PageRank (Page et al. 1999). Perhaps most importantly, given the massive volumes of data generated by platforms like Twitter, we presented a fast estimation algorithm and established statistical properties for the model estimates and those of the final influence measure.

SUPPLEMENTARY MATERIALS

R code and datasets: The R codes and data files can be downloaded as a zip file.

Derivation details and additional empirical results: Full details are given for the algorithmic derivation in addition to regression results.

ACKNOWLEDGMENTS

The authors thank Professor Moulinath Banerjee for many insightful comments on the theoretical results.

[Received November 2014. Revised December 2015.]

REFERENCES

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 4, 1100–1120. [363]

Aral, S., and Walker, D. (2012), "Identifying Influential and Susceptible Members of Social Networks," *Science*, 337, 337–341. [362]

Barbieri, N., Bonchi, F., and Manco, G. (2013), "Topic-Aware Social Influence Propagation Models," *Knowledge and Information Systems*, 37, 555–584. [362]

Blei, D. M. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84. [367]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [367]

Bulearca, M., and Bulearca, S. (2010), "Twitter: A Viable Marketing Tool for SMEs?," *Global Business and Management Research: An International Journal*, 2, 296–309. [360]

Butts, C. T. (2008), "A Relational Event Framework for Social Action," *Sociological Methodology*, 38, 155–200. [362]

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010), "Measuring User Influence in Twitter: The Million Follower Fallacy," *ICWSM*, 10, 10–17. [361]

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society,* Series B, 34, 187–220. [362,363]

Dave, P. (2015), "Firms Turn to Online 'Influencers' to Spread the Word on Social Media," *The Los Angeles Times*, Accessed: 2015-01-05. [360]

Du, N., Song, L., Yuan, M., and Smola, A. J. (2012), "Learning Networks of Heterogeneous Influence," in *Advances in Neural Information Processing Systems 25*, eds. F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Curran Associates, Inc., pp. 2780–2788. [362]

Dubois, E., and Gaffney, D. (2014), "The Multiple Facets of Influence: Identifying Political Influentials and Opinion Leaders on Twitter," *American Behavioral Scientist*. doi: 10.1177/0002764214527088. [361]

Fienberg, S. E. (2012), "A Brief History of Statistical Models for Network Analysis and Open Challenges," *Journal of Computational and Graphical Statistics*, 21, 825–839. [370]

Gayo-Avello, D., Metaxas, P. T., and Mustafaraj, E. (2011), "Limits of Electoral Predictions Using Twitter," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*. [361]

Golbeck, J., Grimes, J. M., and Rogers, A. (2010), "Twitter Use by the U.S. Congress," *Journal of the American Society for Information Science and Technology*, 61, 1612–1621. [367]

Gomez-Rodriguez, M., Leskovec, J., and Schölkopf, B. (2013), "Modeling Information Propagation with Survival Theory," *CoRR*, abs/1305.3616. [362]

GovTrack.us (2014), GovTrack.us: Tracking the United States Congress. Available at *https://www.govtrack.us/about/analysis#leadership*. [367]

Haveliwala, T. H. (2003), "Topic-Sensitive Pagerank: A Context-Sensitive Ranking Algorithm for Web Search," *Knowledge and Data Engineering, IEEE Transactions on*, 15, 784–796. [361,365]

Hong, L., and Davison, B. D. (2010), "Empirical Study of Topic Modeling in Twitter," in *Proceedings of the First Workshop on Social Media Analytics*, ACM, pp. 80–88. [367]

Hunter, D., Smyth, P., Vu, D. Q., and Asuncion, A. U. (2011), "Dynamic Egocentric Models for Citation Networks," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 857–864. [362]

Kleinberg, J. M. (1999), "Authoritative Sources in a Hyperlinked Environment," *Journal of ACM*, 46, 604–632. [361]

Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, Springer Series in Statistics, New York, NY: Springer. [370]

Kwak, H., Lee, C., Park, H., and Moon, S. (2010), "What is Twitter, a Social Network or a News Media?," in *Proceedings of the 19th International Conference on World Wide Web*, ACM, pp. 591–600. [361,363]

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999), The PageRank Citation Ranking: Bringing Order to the Web. [361,368,370]

Perry, P. O., and Wolfe, P. J. (2013), "Point Process Modelling for Directed Interaction Networks," *Journal of the Royal Statistical Society,* Series B, 75, 821–849. [362,370]

Probst, F., Grosswiele, L., and Pfleger, R. (2013), "Who Will Lead and Who Will Follow: Identifying Influential Users in Online Social Networks," *Business & Information Systems Engineering*, 5, 179–193. [360,361]

Ramage, D., Dumais, S., and Liebling, D. (2010), "Characterizing Microblogs With Topic Models," in *Proc. ICWSM 2010*, American Association for Artificial Intelligence. [367]

Salter-Townshend, M., White, A., Gollini, I., and Murphy, T. B. (2012), "Review of Statistical Network Analysis: Models, Algorithms, and Software," *Statistical Analysis and Data Mining*, 5, 243–264. [370]

Taddy, M. (2013), "Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression," *Technometrics*, 55, 415–425. [370]

Trusov, M., Bodapati, A. V., and Bucklin, R. E. (2010), "Determining Influential Users in Internet Social Networks," *Journal of Marketing Research*, 47, 643–658. [361]

Twitter Inc. (2014), 10-K Report, available at *http://www.sec.gov/Archives/edgar/data/1418091/000156459015001159/twtr-10k_20141231.htm*. [360,361]

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010), "Twitterrank: Finding Topic-Sensitive Influential Twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ACM, pp. 261–270. [361,365]