Statistica Sinica **33** (2023), 499-518
doi:https://doi.org/10.5705/ss.202020.0310

INFERRING SOCIAL INFLUENCE IN DYNAMIC NETWORKS

Xiang Cui and Yuguo Chen

University of Illinois at Urbana-Champaign

Abstract: An interesting problem in social network analysis is whether individuals' behaviors or opinions spread from one to another, which is known as social influence. The degrees of influence describes how far the influence passes through individuals. Here, we explore the degrees of influence in dynamic networks. We build a longitudinal influence model to specify how people's behaviors are influenced by others in a dynamic network. In order to determine the degrees of influence, we propose a sequential hypothesis testing procedure and use generalized estimating equations to account for multiple observations of the same individual across different time points. In addition, we show that the power of our proposed test goes to one as the network size goes to infinity. We illustrate the performance of our proposed method using simulation studies and real-data analyses.

Key words and phrases: Degrees of influence, dynamic network, generalized estimating equations, longitudinal analysis, social influence.

1. Introduction

Social network analysis has become popular in many fields, including sociology, psychology, computer science, and statistics. A social network consists of individuals and the relationships between them, represented by nodes and edges, respectively, in a graph. Social networks can be static or dynamic. A static network is a snapshot of a network at a certain time point, and a dynamic network is a sequence of observations of networks at different time points.

An interesting problem in social network analysis is whether the behaviors or opinions of an individual can be influenced by others in the network, which is known as social influence or social contagion. Several methods have been developed to study the spread of individuals' behavior within a social network (Valente (1995); Centola (2010)).In addition, researchers have examined the spread of various individual health outcomes, including obesity (Christakis and Fowler (2007)), smoking (Christakis and Fowler (2008)), sleep loss and drug use (Mednick, Christakis and Fowler (2010)), alcohol consumption (Rosenquist et al.

Corresponding author: Yuguo Chen, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA. E-mail: yuguo@illinois.edu.

(2010)), and sexual orientation (Brakefield et al. (2014)). See Sun and Tang (2011) for a summary of the models and algorithms developed for social influence analysis. Kempe, Kleinberg and Tardos (2003) proposed methods for selecting the most influential nodes in a network to maximize the spread of the influence (i.e., social influence maximization). O'Malley (2013) used instrumental variables to account for the confounding effect when analyzing peer effects, and proposed a network influence model with multiple types of relationships.

The degrees of influence (DOI) describes how far an influence passes through individuals in a network. For static networks, Christakis and Fowler (2013) proposed a permutation test to identify the behavior association between individuals across a social network, using the Framingham Heart Study data. They claimed that the spread of influence in social networks obeys the three degrees of influence rule. VanderWeele (2013) discussed three distinct interpretations of this rule. However, O'Malley (2013) pointed out an issue with the choice of the null hypothesis in Christakis and Fowler (2013). Later, Su (2019) proposed a new sequential test procedure with more appropriate null hypotheses for determining the degrees of influence.

Existing works can only detect the degrees of influence for static networks (Christakis and Fowler (2013); Su (2019)), and it is not clear how to extend their methods to dynamic networks. Here, we introduce a longitudinal influence model and a sequential hypothesis testing procedure for determining the degrees of influence in dynamic networks. We also provide theoretical properties of the level and power of the proposed test. In particular, we show that the power of the proposed test goes to one as the network size goes to infinity.

The remainder of the paper is organized as follows. Section 2 provides the basic notation. Section 3 introduces the longitudinal influence model. Section 4 gives the proposed sequential hypothesis testing procedure. Section 5 provides the theoretical properties of the proposed method. Section 6 describes the simulation studies. Section 7 reports the results for the Higgs Twitter data set (De Domenico et al. (2013)) and the Digg data set (Hogg and Lerman (2012)). Section 8 concludes the paper.

2. Notation

Consider a dynamic social network consisting of n individuals (nodes) and a set of dyadic relationships (edges) between them at time $t = 1, 2, \ldots, T$. We are mainly concerned with directed networks with no loops (both ends of an edge connect to a single node) or multiple edges between a pair of nodes. Such a

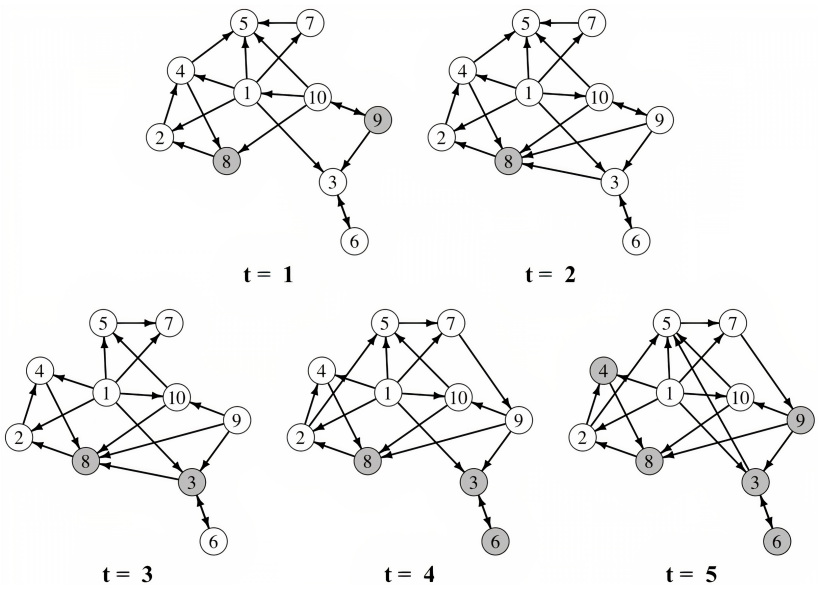

Figure 1. A toy example of a dynamic social network, with gray nodes denoting smokers ($y_{i,t} = 1$) and white nodes denoting nonsmokers ($y_{i,t} = 0$).

dynamic network can be represented by its adjacency matrix $A_t = (a_{ij,t})_{n\times n}$, for $t = 1, 2, \ldots, T$, where each A_t is an $n \times n$ binary square matrix, with $a_{ij,t} = 1$ if there is a directed edge from node i to j at time t (i.e., individual i is following individual j at time t), and $a_{ij,t} = 0$ otherwise. For a directed network, A_t does not need to be symmetric.

In addition, we observe whether each individual in the network possesses a specific trait, such as obesity, smoking, or happiness, at each time point t. This can be modeled by a binary random vector $Y_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})$, with $y_{i,t} = 1$ indicating the trait is present in individual i at time t, and zero indicating otherwise. A toy example of such a dynamic network is given in Figure 1.

In social networks, the individual we are focusing on is called the ego. If there is a directed path from the ego to an individual at time t, then that individual is called an alter. If the shortest directed path from the ego to an alter is d at time t, then this alter is referred to as a dth-degree alter, denoted by $\text{alter}_{d,t}$. Here is a simple illustration:

$$\text{ego} \to \text{alter}_{1,t} \to \text{alter}_{2,t} \to \cdots$$

Obviously the first-degree alters ($\text{alter}_{1,t}$) are directly connected to the ego. The second-degree alters ($\text{alter}_{2,t}$) have a length-two path from the ego, but

they are not directly connected to the ego. In other words, $a_{\text{ego, alter}_{1,t},t} = 1$, $a_{\text{alter}_{1,t},\ \text{alter}_{2,t},t} = 1$, and $a_{\text{ego, alter}_{2,t},t} = 0$. For example, at time $t = 2$ in Figure 1, individuals 6 and 8 are first-degree alters of ego 3, individual 2 is a second-degree alter of ego 3, and individual 4 is a third-degree alter of ego 3, and so on.

Let $d_{ij,t}$ be the length of the shortest directed path from i to j at time t. Then, $d_{\text{ego, alter}_{1,t},t} = 1$, $d_{\text{ego, alter}_{2,t},t} = 2$, and so on. We define the dth-degree alter set for each ego i at time t as

$$S_{i,t}^d = \{j :\ d_{ij,t} = d\}.$$

For an individual i at time t, we define the dth-degree influence factor $x_{i,t}^d$ as the average status of individual i's dth-degree alters at time t, that is,

$$x_{i,t}^d = \begin{cases} \frac{1}{|S_{i,t}^d|} \sum_{j \in S_{i,t}^d} y_{j,t}, & |S_{i,t}^d| > 0, \\ 0, & |S_{i,t}^d| = 0. \end{cases} \tag{2.1}$$

3. Longitudinal Influence Model

Social influence describes the process by which an individual's behavior or opinion is affected by others in the network. The influence may go beyond the people a person is directly linked to. We are interested in studying the degrees of influence, which describes how far an influence can pass through links between individuals. In this section, we specify a longitudinal influence model for different degrees of influence in dynamic networks.

We assume that

$$y_{i,t} \sim \text{Bernoulli}(p_{i,t}), \quad i = 1, 2, \ldots, n \text{ and } t = 1, 2, \ldots, T. \tag{3.1}$$

If the degrees of influence is zero, then the behavior of each individual is not affected by others in the network. Therefore $y_{i,t+1}$ depends only on individual i's status at time t. We propose the following longitudinal influence model:

$$\begin{aligned} y_{i,t+1} &\sim \text{Bernoulli}(p_{i,t+1}), \\ \text{logit}(p_{i,t+1}) &= \gamma + \beta_0 y_{i,t}, \quad i = 1, 2, \ldots, n \ \text{ and } \ t = 1, 2, \ldots, T-1, \end{aligned} \tag{3.2}$$

where γ is the intercept and β_0 is the coefficient for the time-lagged status $y_{i,t}$.

If the true degrees of influence is $D^* > 0$, then each individual may be influenced by anyone to whom the individual is connected by a path with length no more than D^*. Thus each ego i's binary status at time $t + 1$ depends on ego i's status at time t and the status of ego i's alters with degrees one to D^* at time

t. We propose the following longitudinal influence model:

$$\begin{aligned} y_{i,t+1} &\sim \text{Bernoulli}(p_{i,t+1}), \\ \text{logit}(p_{i,t+1}) &= \gamma + \beta_0 y_{i,t} + \sum_{d=1}^{D^*} \beta_d x_{i,t}^d, \quad i = 1, 2, \ldots, n \text{ and } t = 1, 2, \ldots, T-1, \end{aligned} \tag{3.3}$$

where γ is the intercept, β_0 is the coefficient for the time-lagged status $y_{i,t}$, $x_{i,t}^d$ is the dth-degree influence factor defined in (2.1), and β_d is its coefficient. In the above model with true degrees of influence $D^* > 0$, because we assume that each individual i can be influenced by individual i's alters with degrees one to D^*, we are essentially assuming $\beta_d \neq 0$, for $d = 1, 2, \ldots, D^*$.

4. Hypothesis Testing

In order to determine the degrees of influence in a dynamic network, we propose a sequential hypothesis testing procedure. This procedure is similar to forward variable selection in linear regression models, where we add one new predictor variable to the model at a time, and perform a goodness of fit test to compare it with the model without the new predictor variable (Hocking (1976); Everitt and Dunn (2001)). We propose sequentially testing the following hypothesis:

$$H_0 : \text{ DOI} = D - 1 \quad \text{vs.} \quad H_1 : \text{ DOI} \geq D. \tag{4.1}$$

We start with $D = 1$, and if the null hypothesis is rejected, then we test (4.1) again with D increased by 1 to $D = 2$. The procedure continues until the null hypothesis cannot be rejected for a certain value of D, and we then report $D - 1$ as the degrees of influence.

For the test in (4.1), the null model M_0 under H_0 is

$$\begin{aligned} y_{i,t+1} &\sim \text{Bernoulli}(p_{i,t+1}), \\ \text{logit}(p_{i,t+1}) &= \gamma + \beta_0 y_{i,t} + \sum_{d=1}^{D-1} \beta_d x_{i,t}^d, \quad i = 1, 2, \ldots, n \quad \text{and} \quad t = 1, 2, \ldots, T-1. \end{aligned} \tag{4.2}$$

If the alternative hypothesis is true, based on the discussion after (3.3), the coefficient for the Dth-degree influence factor is nonzero. Hence, hypothesis (4.1) is testing

$$H_0 : \ \beta_D = 0 \quad \text{vs.} \quad H_1 : \ \beta_D \neq 0. \tag{4.3}$$

Under H_1, the model closest to the null hypothesis is the following alternative candidate model M_1, with DOI $= D$:

$$\begin{aligned} y_{i,t+1} &\sim \text{Bernoulli}(p_{i,t+1}), \\ \text{logit}(p_{i,t+1}) &= \gamma + \beta_0 y_{i,t} + \sum_{d=1}^{D} \beta_d x_{i,t}^d, \quad i = 1, 2, \ldots, n \quad \text{and} \quad t = 1, 2, \ldots, T-1. \end{aligned} \tag{4.4}$$

The null model ($\beta_D = 0$) is nested within the alternative candidate model. In order to implement the test to compare the two models, we need to estimate the parameter β_D and the variance of the estimator in the alternative candidate model M_1.

To account for multiple observations of the same individual across different time periods, we use generalized estimating equations (GEEs) (Liang and Zeger (1986)) to estimate β_D and the variance of the estimator. We first establish the notation for the parameter estimation. For the ith individual, let $y_i = (y_{i,2}, y_{i,3}, \ldots, y_{i,T})^T$ be a vector of the outcome values and $X_i = (x_{i,1}, \ldots, x_{i,T-1})^T$ be a matrix of the covariate values, where $x_{i,t} = (1, y_{i,t}, x_{i,t}^1, \ldots, x_{i,t}^D)^T$, for $t = 1, \ldots, T-1$. In the alternative candidate model M_1, we have $\text{E}(y_{i,t+1}) = p_{i,t+1}$ and $\text{logit}(p_{i,t+1}) = x_{i,t}^T \beta$, where $\beta = (\gamma, \beta_0, \beta_1, \ldots, \beta_D)^T$. When using GEEs for parameter estimation, we assume an independence working correlation structure. Under certain conditions, this can yield a consistent estimator $\hat{\beta}_D$ for β_D, and the variance of the estimator $\hat{\beta}_D$ can be consistently estimated by a sandwich estimator $\hat{\Sigma}(\hat{\beta}_D)$ (Liang and Zeger (1986)). We used the R package geepack (Halekoh, Højsgaard and Yan (2006)) to solve the GEEs by providing the outcome values y_i and the matrix of the covariate values X_i. More details on the use of GEEs can be found in Liang and Zeger (1986) and Halekoh, Højsgaard and Yan (2006).

We use the Wald test with test statistic

$$W = \frac{\hat{\beta}_D^2}{\hat{\Sigma}(\hat{\beta}_D)}. \tag{4.5}$$

Under the null hypothesis and certain conditions, the test statistic W approximately follows a $\chi^2(1)$ distribution. For a given significance level α, the critical value for the test is $c^* = \chi_{1-\alpha}^2(1)$, and we reject H_0 if $W > c^*$. In the remainder of the paper, we choose $\alpha = 0.05$ for all simulation and real-data analyses.

4.1. Toy example

We use the toy example in Figure 1 to illustrate the sequential testing procedure. The dynamic network in Figure 1 was generated in the following way. At time $t = 1$, we generated a network from the Erdős–Rényi model $\mathrm{ER}(n, p_e)$ (Erdös and Rényi (1960)), where n is the number of nodes and p_e is the edge probability. We set $n = 10$ and $p_e = 0.2$. At the following time step t, for $2 \le t \le 5$, the network structure is allowed to change. In particular, we assume that for every pair of (i, j), $a_{ij,t}$ is equal to $a_{ij,t-1}$ with probability 0.95, and is equal to $1 - a_{ij,t-1}$ with probability 0.05. We assigned the smoking status for each node at time $t = 1$ based on Bernoulli(p_m) with $p_m = 0.2$. At time t, for $2 \le t \le 5$, each node's status was generated according to the longitudinal influence model in (3.3), where we set the degrees of influence $D^* = 1$, and the parameters $\gamma = -3$, $\beta_0 = 4$, and $\beta_1 = 4$. In Figure 1, gray nodes denote smokers and white nodes denote nonsmokers.

To explore the degrees of influence in the toy example, we set the significance level $\alpha = 0.05$ and started by testing: $H_0 : \mathrm{DOI} = 0$ vs. $H_1 : \mathrm{DOI} \ge 1$. The dth-degree influence factor $x^d_{i,t}$ for each individual can be calculated based on Equation (2.1). To obtain the estimates of the parameters in Equation (3.3) and the corresponding estimated variance, we solved the GEEs under an independence working correlation structure using the R package geepack (Halekoh, Højsgaard and Yan (2006)). Given $y_{i,t}$ and $x^1_{i,t}$, for $i = 1, 2, \ldots, 10$ and $t = 1, 2, \ldots, 5$, we obtained the estimate $\hat{\beta}_1 = 6.191$ and the estimated variance $\hat{\Sigma}(\hat{\beta}_1) = 2.401$. The test statistic is

$$W = \frac{\hat{\beta}_1^2}{\hat{\Sigma}(\hat{\beta}_1)} = 15.961,$$

which is larger than the critical value $\chi^2_{0.95}(1) = 3.841$. The null hypothesis $H_0 : \mathrm{DOI} = 0$ is rejected.

Then, we tested $H_0 : \mathrm{DOI} = 1$ vs. $H_1 : \mathrm{DOI} \ge 2$. Given $y_{i,t}$ and $x^d_{i,t}$, for $d = 1, 2$, $i = 1, 2, \ldots, 10$, and $t = 1, 2, \ldots, 5$, we obtained the estimate $\hat{\beta}_2 = 1.147$ and the estimated variance $\hat{\Sigma}(\hat{\beta}_2) = 1.659$. The test statistic is

$$W = \frac{\hat{\beta}_2^2}{\hat{\Sigma}(\hat{\beta}_2)} = 0.792,$$

which is smaller than the critical value 3.841. The null hypothesis $H_0 : \mathrm{DOI} = 1$ cannot be rejected, and the degrees of influence in the toy example is reported to be one based on the sequential test.

5. Theoretical Properties

In this section, we provide some theoretical properties of the proposed sequential testing procedure. The following theorem shows how the level and power of the test change as the network size increases.

Theorem 1. *Suppose we observe a dynamic network with n nodes at time $1, \ldots, T$ and binary vectors $Y_1, \ldots, Y_n$ indicating the presence or absence of a trait for each individual in the network. Let D^* be the true degrees of influence in the network. Let W in (4.5) be the proposed test statistic for testing $H_0 : DOI = D - 1$ vs. $H_1 : DOI \geq D$, with W estimated from observations that are independent across individuals. Let α be the significance level and c^* be the critical value of the test. We have the following results:*

(a) The level of the test $P(W > c^* \mid DOI = D - 1) \longrightarrow \alpha$ as $n \to \infty$, for $D - 1 = D^*$.

(b) The power of the test $P(W > c^* \mid DOI \geq D) \longrightarrow 1$ as $n \to \infty$, for all $1 \leq D \leq D^*$.

The proof of Theorem 1 is provided in Appendix A. The above theorem indicates that the level of the test goes to the significance level α and the power of the test goes to one as the network size $n \to \infty$. This shows that the test can always tell the difference between the null and the alternative hypotheses when the network size is large. Furthermore, for large networks, the true degrees of influence can be detected by our method with high probability.

The theorem requires that the test statistic $W = \hat{\beta}_D^2 / \hat{\Sigma}(\hat{\beta}_D)$ be estimated based on observations that are independent across individuals. This ensures that the theoretical properties of the estimates using GEEs are applicable here. To obtain independent data across individuals, a convenient assumption is that $y_{1,t}, \ldots, y_{n,t}$ are independent, conditional on all the observations at time $t - 1$. Under this assumption, there are different ways of obtaining independent data. For example, the observations $y_{i,t}$ at $t = 2$ are conditionally independent given $y_{i,t}$ at $t = 1$. Furthermore, $y_{i,t}$ at $t = 2k$ $(k = 1, 2, \ldots)$ are conditionally independent given $y_{i,t}$ at $t = 2k - 1$. Because these independent data do not make full use of the information in the observed data, in practice, using all observations $y_{i,t}$, as discussed in Section 4, tends to perform better. Therefore, we use the full data in our simulation studies and real-data analyses.

In the proof for the power of the test (part (b) of the theorem), the test statistic W is estimated based on the true model in the alternative hypothesis. In practice, the true model is not known in the middle of the sequential test,

Table 1. Results for detecting the degrees of influence for different parameter settings.

$(n, p_e, \beta_1, \beta_2)$	$< D^*$	$= D^*$	$> D^*$
(300, 0.02, 3, 3)	0	47	3
(500, 0.02, 3, 3)	0	49	1
(1000, 0.01, 2, 2)	4	44	2
(3000, 0.005, 2, 2)	1	47	2
(5000, 0.003, 2, 2)	0	47	3

so we estimate W based on the alternative candidate model M_1 in (4.4). This approach works well in the simulation studies and real-data analyses. This type of approach has also been suggested in sequential testing for forward variable selection in linear regression models (Hocking (1976); Everitt and Dunn (2001)). In fact, our simulation shows that the test based on M_1 is even more powerful than that based on the true model for testing (4.1) with $D < D^*$. An intuitive explanation might be that the estimates based on the alternative candidate model M_1 need to reflect the additional influence from distances larger than D, which makes it easier to reject the null hypothesis of DOI $= D - 1$.

6. Simulation Results

6.1. Detecting the degrees of influence

In this section, we show the performance of our proposed test procedure in detecting the degrees of influence in dynamic networks. We generated a network at time $t = 1$ from the Erdős–Rényi model ER(n, p_e), where n is the number of nodes and p_e is the edge probability. At the following time step t, for $2 \le t \le 5$, the network structure changes in the following way. For each pair of $\{i, j\}$, if $a_{ij,t-1} = 1$, then $a_{ij,t} = 1$ with probability 0.95, and $a_{ij,t} = 0$ with probability 0.05. If $a_{ij,t-1} = 0$, then $a_{ij,t} = 1$ with probability $p_{\text{change}} = 0.05p_e/(1 - p_e)$, and $a_{ij,t} = 0$ with probability $1 - p_{\text{change}}$. At time $t = 1$, we assigned the status $y_{i,1}$ for each individual from Bernoulli(p_m) with $p_m = 0.2$. For time $t = 2, 3, 4, 5$, each individual's status was generated according to the longitudinal influence model in (3.3), where we set the true degrees of influence $D^* = 2$, and the parameters $\gamma = -3$ and $\beta_0 = 4$. The values of β_1 and β_2, together with n and p_e, are presented in Table 1. For each set of parameter values, we generated data and applied our proposed method to detect the DOI. We ran 50 trials for each simulation; the results are presented in Table 1.

In Table 1, the first column gives the parameter settings for the network size n, the edge probability p_e, and the coefficients β_1, β_2 in the model in Equation

(3.3). The columns "$< D^*$," "$= D^*$," and "$> D^*$ " indicate that the number of trials the DOI detected using our proposed method is smaller than, equal to, and larger than, respectively, the true DOI. From Table 1, we can see that our proposed test procedure detects the true DOI in most cases.

We also considered detecting the DOI when the network structure is fixed for the whole time period $1 \le t \le 5$. Using the same parameter settings as in the above simulation, we obtained similar results to those shown in Table 1. This shows that our proposed test procedure works well for both fixed and varying network structures.

6.2. Level and power of the tests

In this section, we show the level and power of our proposed test for different parameter settings and different true degrees of influence. We generated a network at time $t = 1$ from the Erdős–Rényi model ER(n, p_e). At the following time step t, for $2 \le t \le 5$, the network structure changes in the following way. For each pair of $\{i, j\}$, if $a_{ij,t-1} = 1$, then $a_{ij,t} = 1$ with probability 0.95, and $a_{ij,t} = 0$ with probability 0.05. If $a_{ij,t-1} = 0$, then $a_{ij,t} = 1$ with probability $p_{\text{change}} = 0.05p_e/(1 - p_e)$, and $a_{ij,t} = 0$ with probability $1 - p_{\text{change}}$. At time $t = 1$, we assigned the status $y_{i,1}$ for each individual from Bernoulli(p_m) with $p_m = 0.2$. For time $t = 2, 3, 4, 5$, each individual's status was generated according to the longitudinal influence model in (3.3). For a given DOI D^*, we generated data with true DOI $= D^*$ and $\beta_1, \ldots, \beta_{D^*}$ set to prespecified values. We then estimated the power of the test $H_0 : \text{DOI} = D - 1$ vs. $H_1 : \text{DOI} \ge D$ for $1 \le D \le D^*$ and the level of the test $H_0 : \text{DOI} = D - 1$ vs. $H_1 : \text{DOI} \ge D$ for $D - 1 = D^*$. In practice, when the size of the network n increases, it may become sparse and the edge probability may decrease. Therefore, we assigned smaller values to p_e for larger networks. For the rest of this section, we set $\gamma = -3$ and $\beta_0 = 4$ in model (3.3). We ran 100 trials for each simulation setting to obtain the level and power.

6.2.1. Testing when the true DOI is zero

In this section, we assume the true DOI $D^* = 0$, and the data are generated from model (3.2). We test $H_0 : \text{DOI} = 0$ vs. $H_1 : \text{DOI} \ge 1$. Because H_0 represents the true DOI, we only examine the level of the test. The first column of Table 2 gives the parameter values for the network size n and the edge probability p_e at time $t = 1$. We ran 100 trials for each simulation to estimate the level. The results in Table 2 show that our test procedure achieves the level around the prespecified $\alpha = 0.05$.

Table 2. Levels for testing H_0 : DOI $= 0$ vs. H_1 : DOI ≥ 1 when the true DOI $D^* = 0$.

(n, p_e)	Level
(500, 0.02)	0.06
(1000, 0.01)	0.04
(3000, 0.005)	0.06
(5000, 0.003)	0.07

Table 3. Power for testing: H_0 : DOI $= 0$ vs. H_1 : DOI ≥ 1 and levels for testing: H_0 : DOI $= 1$ vs. H_1 : DOI ≥ 2 when the true DOI $D^* = 1$.

(n, p_e, β_1)	Power	Level
(500, 0.02, 3)	1	0.06
(1000, 0.01, 3)	1	0.05
(3000, 0.005, 3)	1	0.03
(5000, 0.003, 3)	1	0.05
(500, 0.02, 2)	0.96	0.03
(1000, 0.01, 2)	1	0.06
(3000, 0.005, 2)	1	0.04
(5000, 0.003, 2)	1	0.03

6.2.2. Testing when the true DOI is one

In this section, we assume the true DOI $D^* = 1$. The first column of Table 3 gives the parameter values for the network size n, the edge probability p_e at time $t = 1$, and the coefficient β_1 in model (3.3). The nonzero β_1 is used to generate data with true DOI $D^* = 1$. We first test H_0 : DOI $= 0$ vs. H_1 : DOI ≥ 1, and report the power of the test in the second column of Table 3. Then, we test H_0 : DOI $= 1$ vs. H_1 : DOI ≥ 2, and report the level of the test in the third column of Table 3. We ran 100 trials for each simulation to estimate the level and power.

The results in Table 3 show that our test procedure achieves the level around the prespecified $\alpha = 0.05$. The power of our test is close to or equal to one in all settings. Thus, our test is powerful in all of the above parameter settings when the true DOI is one.

6.2.3. Testing when the true DOI is two

In this section, we assume the true DOI $D^* = 2$. The first column of Table 4 gives the parameter values for the network size n, the edge probability p_e at time $t = 1$, and the coefficients β_1 and β_2 in model (3.3). We first test H_0 : DOI $= 0$ vs. H_1 : DOI ≥ 1, and report the power of the test in the second column of Table 4 (denoted by Power-1). Then, we test H_0 : DOI $= 1$ vs. H_1 : DOI ≥ 2, and report the power of the test in the third column of Table 4 (denoted by Power-2).

Table 4. Power for testing: $H_0 : \text{DOI} = 0$ vs. $H_1 : \text{DOI} \geq 1$ and $H_0 : \text{DOI} = 1$ vs. $H_1 : \text{DOI} \geq 2$, and levels for testing: $H_0 : \text{DOI} = 2$ vs. $H_1 : \text{DOI} \geq 3$ when the true DOI $D^* = 2$ and $\beta_2 = \beta_1$.

$(n, p_e, \beta_1, \beta_2)$	Power-1	Power-2	Level
(500, 0.02, 3, 3)	1	0.94	0.03
(1000, 0.01, 3, 3)	1	1	0.06
(3000, 0.005, 3, 3)	1	1	0.02
(5000, 0.003, 3, 3)	1	1	0.07
(500, 0.02, 2, 2)	1	0.55	0.05
(1000, 0.01, 2, 2)	1	0.91	0.03
(3000, 0.005, 2, 2)	1	1	0.04
(5000, 0.003, 2, 2)	1	1	0.06

Table 5. Power for testing: $H_0 : \text{DOI} = 0$ vs. $H_1 : \text{DOI} \geq 1$ and $H_0 : \text{DOI} = 1$ vs. $H_1 : \text{DOI} \geq 2$, and levels for testing: $H_0 : \text{DOI} = 2$ vs. $H_1 : \text{DOI} \geq 3$ when the true DOI $D^* = 2$ and $\beta_2 < \beta_1$.

$(n, p_e, \beta_1, \beta_2)$	Power-1	Power-2	Level
(500, 0.02, 3, 2.25)	1	0.93	0.04
(1000, 0.01, 3, 2.25)	1	1	0.06
(3000, 0.005, 3, 2.25)	1	1	0.04
(5000, 0.003, 3, 2.25)	1	1	0.07
(500, 0.02, 2, 1.5)	1	0.30	0.05
(1000, 0.01, 2, 1.5)	1	0.49	0.04
(3000, 0.005, 2, 1.5)	1	0.84	0.03
(5000, 0.003, 2, 1.5)	1	0.93	0.05

Finally, we test $H_0 : \text{DOI} = 2$ vs. $H_1 : \text{DOI} \geq 3$, and report the level of the test in the fourth column of Table 4. We ran 100 trials for each simulation to estimate the level and power.

In Table 4, the coefficients β_1 and β_2 were chosen to be the same. In some situations, the influence from the second-degree alters may be weaker than the influence from the first-degree alters, so we set $\beta_2 < \beta_1$ in Table 5 and re-ran the same tests. The results are presented in Table 5.

From Tables 4 and 5, we can see that our proposed method preserves the level of the test with a type I error close to the prespecified level $\alpha = 0.05$. The power of the test $H_0 : \text{DOI} = 0$ vs. $H_1 : \text{DOI} \geq 1$ is always one. The power of the test $H_0 : \text{DOI} = 1$ vs. $H_1 : \text{DOI} \geq 2$ increases to about one as the network size increases. For fixed network size n and edge probability p_e, the test $H_0 : \text{DOI} = 1$ vs. $H_1 : \text{DOI} \geq 2$ is more powerful for larger values of β_2. This is not surprising, because larger values of β_2 indicate a stronger influence from second-degree alters, which makes it easier to detect the misspecified null hypothesis of

Table 6. Power for testing: H_0 : DOI $= D - 1$ vs. H_1 : DOI $\geq D$ for $D = 1, 2, 3$, and levels for testing: H_0 : DOI $= 3$ vs. H_1 : DOI ≥ 4 when the true DOI $D^* = 3$ and $\beta_3 = \beta_2 = \beta_1$.

$(n, p_e, \beta_1, \beta_2, \beta_3)$	Power-1	Power-2	Power-3	Level
(500, 0.02, 3, 3, 3)	1	1	0.47	0.05
(1000, 0.01, 3, 3, 3)	1	1	0.69	0.03
(3000, 0.005, 3, 3, 3)	1	1	0.82	0.07
(5000, 0.003, 3, 3, 3)	1	1	0.95	0.06
(500, 0.02, 2, 2, 2)	1	1	0.35	0.04
(1000, 0.01, 2, 2, 2)	1	1	0.68	0.03
(3000, 0.005, 2, 2, 2)	1	1	0.78	0.06
(5000, 0.003, 2, 2, 2)	1	1	0.90	0.05

Table 7. Power for testing: H_0 : DOI $= D - 1$ vs. H_1 : DOI $\geq D$ for $D = 1, 2, 3$, and levels for testing: H_0 : DOI $= 3$ vs. H_1 : DOI ≥ 4 when the true DOI $D^* = 3$ and $\beta_3 < \beta_2 < \beta_1$.

$(n, p_e, \beta_1, \beta_2, \beta_3)$	Power-1	Power-2	Power-3	Level
(3000, 0.005, 3, 2.25, 1.5)	1	1	0.44	0.06
(5000, 0.003, 3, 2.25, 1.5)	1	1	0.61	0.05
(10000, 0.002, 3, 2.25, 1.5)	1	1	0.70	0.05
(3000, 0.005, 2, 1.5, 1)	1	0.99	0.24	0.03
(5000, 0.003, 2, 1.5, 1)	1	1	0.35	0.07
(10000, 0.002, 2, 1.5, 1)	1	1	0.47	0.03

no influence from second-degree alters. For fixed values of β_1 and β_2, the power of the test increases as the network size increases.

6.2.4. Testing when the true DOI is three

In this section, we assume the true DOI $D^* = 3$. The first column of Table 6 gives the parameter values for the network size n, the edge probability p_e at time $t = 1$, and the coefficients β_1, β_2, and β_3 in model (3.3). We test H_0 : DOI $= D-1$ vs. H_1 : DOI $\geq D$ for $D = 1, 2, 3$, and report the power of the test in the second, third, and fourth columns of Table 6 (denoted by Power-1, Power-2, Power-3, respectively). Then, we test H_0 : DOI $= 3$ vs. H_1 : DOI ≥ 4, and report the level of the test in the fifth column of Table 6. We ran 100 trials for each simulation to estimate the level and power.

In Table 6, the coefficients $\beta_1 = \beta_2 = \beta_3$ were chosen to be the same. In some situations, the influence from the Dth-degree alters may be weaker than the influence from the $(D - 1)$th-degree alters, so we set $\beta_3 < \beta_2 < \beta_1$ in Table 7 and re-ran the same tests. The results are presented in Table 7.

From Tables 6 and 7, we can see that the correct level is achieved in different

settings. The power of the tests $H_0 : \text{DOI} = 0$ vs. $H_1 : \text{DOI} \geq 1$ and $H_0 : \text{DOI} = 1$ vs. $H_1 : \text{DOI} \geq 2$ is always around one. With the same parameter settings for β_i $(i = 1, 2, 3)$, the power of the test $H_0 : \text{DOI} = 2$ vs. $H_1 : \text{DOI} \geq 3$ increases as the network size increases. For a fixed network size n and edge probability p_e, the test $H_0 : \text{DOI} = 2$ vs. $H_1 : \text{DOI} \geq 3$ is more powerful for larger values of β_3, which is consistent with our intuition.

We also re-ran all simulations related to the levels and power in Section 6.2 with a fixed network structure for the whole time period $1 \leq t \leq 5$. Using the same parameter settings as in the above simulations, we obtained similar results to those shown in Tables 2 to 7. Thus, our proposed test procedure also works well for dynamic networks with a fixed network structure.

7. Real-Data Analyses

7.1. Higgs Twitter data

Twitter is a popular American social networking site, with a microblogging system that allows users to post and interact with posts, called "tweets." In this section, we analyze the Higgs Twitter data set collected by De Domenico et al. (2013) and available at `https://snap.stanford.edu/data/higgs-twitter.html`. These data were built by keeping track of the spreading process on Twitter before, during, and after the announcement on July 4, 2012, of the discovery of a new particle with the elusive Higgs boson features.

The data set contains a network of Twitter users who posted messages about this discovery between July 1, 2012, and July 7, 2012. Nodes in the network correspond to users, and an edge from node i to node j means node i follows node j. The data set also contains interactions (including retweets, mentions, and replies) between users with a time stamp. Here, we focus on the mention behavior as a feature of interest, which indicates a user mentioned other users when he/she posted on Twitter about the Higgs boson discovery. This feature is represented by a binary random variable $y_{i,t}$, where $y_{i,t} = 1$ indicates that user i mentioned other users in tweets about the discovery before a certain time t, and $y_{i,t} = 0$ otherwise.

Because the network in the original data set has a large number of nodes, we consider a subset of the network by choosing the node that first showed the feature of interest, and then selecting those nodes that have a path with a length of no more than two to this node. This subnetwork has 1,757 nodes and is shown in Figure 2. The network structure does not change in this data.

Because the announcement of the discovery was on July 4, 2012, the spread

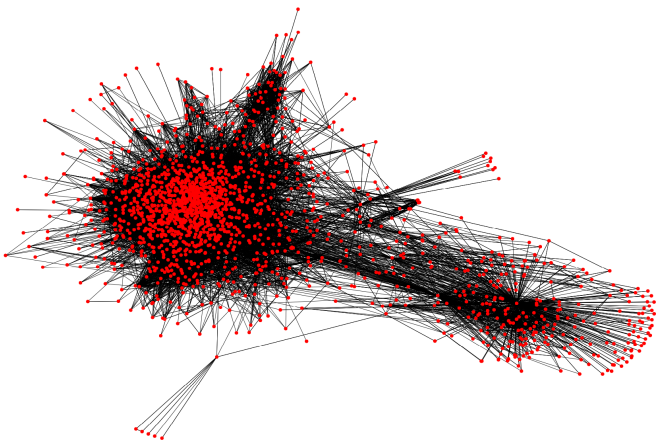

Figure 2. Higgs Twitter network.

of the feature of interest after the announcement might not come from social influence in the network. Therefore, the time interval of interest is from July 1, 2012, to July 3, 2012. We divided this interval into six time steps, with t_1 12 p.m. of 7/1/2012 and t_6 12 a.m. of 7/4/2012; the gap between two neighboring time steps is 12 hours.

We applied our sequential hypothesis test to determine the degrees of influence for the mention behavior. We started by testing $H_0 : \mathrm{DOI} = 0$ vs. $H_1 : \mathrm{DOI} \geq 1$. The test statistic W in (4.5) is 9.43, which is larger than the critical value $\chi^2_{0.95}(1) = 3.841$. Thus, we rejected the null hypothesis and continued to test $H_0 : \mathrm{DOI} = 1$ vs. $H_1 : \mathrm{DOI} \geq 2$. This time $W = 57.76$, which is still larger than the critical value, so the null hypothesis is again rejected. Then, we tested: $H_0 : \mathrm{DOI} = 2$ vs. $H_1 : \mathrm{DOI} \geq 3$, and the test statistic $W = 0.0019$ is smaller than the critical value. Therefore, the null hypothesis H_0 cannot be rejected, and we report the degrees of influence as two. This means that the mention behavior in Twitter can be influenced by an individual's followees, as well as by his/her followees' followees. Note that there are many length-three paths in the subnetwork, and every node in the subnetwork has some third-degree alters. Thus, accepting $H_0 : \mathrm{DOI} = 2$ is not because there are not enough length-three paths.

7.2. Digg data

Digg is a social news website with a curated front page that selects interesting stories related to viral Internet issues, science, and political news for an Internet

audience. Users can read and share the most popular and interesting stories on the Internet. The Digg2009 data set, collected by Hogg and Lerman (2012), contains data on the stories promoted to Digg's front page based on users' votes in one month in 2009. The data are available at `https://www.isi.edu/~lerman/downloads/digg2009.html`. This data set is anonymized and has the voting records for 3,553 different stories during that month. The voting record for each story contains the ID of the voter and the time stamp of the vote. In addition, the data set contains links of the voters and the time stamp of the formation of the link. Here, a link from node i to node j means user i is a fan/follower of user j.

We consider users' votes for the most voted story (which is story 714) as the feature of interest. This feature is represented by a binary random variable $y_{i,t}$, where $y_{i,t} = 1$ means user i voted for story 714 before time t, and $y_{i,t} = 0$ otherwise. Because the original network has a large number of nodes, we consider a subset of the network by choosing the node that made the first vote for story 714, and then selecting the nodes directly following this node at the time when the first vote for story 714 was made. This subnetwork has 1,408 nodes.

For the selected subnetwork, a total of 304 votes were made for story 714. After the first vote at 17:42:46 on June 25, 2009, there were 169 votes before 21:00:00 of the same day. This fast increase of votes was probably not due to social influence. Furthermore, there were only 12 votes after June 26, 2009, and we believe social influence was very weak by that time. Therefore, the time interval of interest is from 21:00:00 of June 25, 2009 (t_1), to 00:00:00 of June 27, 2009 (t_{10}), during which the number of votes increased from 169 to 292. We divided this interval into 10 time steps, and the gap between two neighboring time steps is three hours. The network structure also changed slightly during the selected time interval. The subnetwork at time t_1 is shown in Figure 3.

We applied our sequential hypothesis test to determine the degrees of influence for the voting behavior for story 714. We tested $H_0 : \mathrm{DOI} = D - 1$ vs. $H_1 : \mathrm{DOI} \geq D$ for $D = 1$, 2, and 3, yielding the test statistics W equal to 4.02, 28.80, and 56.40, respectively, which are all larger than the critical value $\chi^2_{0.95}(1) = 3.841$. Thus, the null hypothesis H_0 for all three tests is rejected. Then, we continued to test: $H_0 : \mathrm{DOI} = 3$ vs. $H_1 : \mathrm{DOI} \geq 4$, and the test statistic $W = 0.87$ is smaller than the critical value. Thus, the null hypothesis H_0 cannot be rejected, and we report the degrees of influence as three. This shows that users' voting behavior in Digg network can be influenced by their directly connected neighbors, their neighbors' neighbors, and their third-degree alters. Note that there are many length-four paths in the subnetwork. On average, each

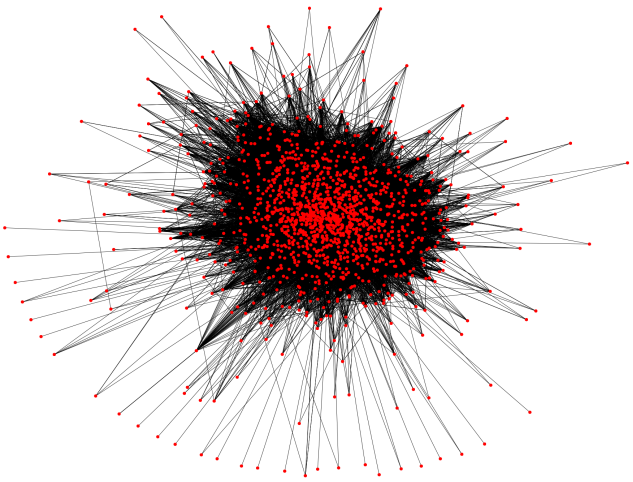

Figure 3. Digg network at time t_1.

individual has more than 50 fourth-degree alters that can potentially influence this individual. Thus, the fact that we cannot reject $H_0 : \mathrm{DOI} = 3$ is not because there are not enough length-four paths.

8. Conclusion

We have presented a longitudinal influence model for dynamic networks with various degrees of influence. We have proposed a sequential testing procedure for determining the degrees of influence in dynamic networks. We also provide a theoretical justification for our proposed test, and show that the power of the test goes to one as the network size goes to infinity. Our proposed test performs well in simulation studies and real-data analyses. The sequential testing procedure may involve multiple tests, but because the degrees of influence is usually small (often no more than three), we need only perform the test a few times in most cases. Therefore, we do not consider the issue with multiple tests in this study.

The proposed longitudinal model and sequential test for dynamic networks are quite different to the testing procedure for static networks (Christakis and Fowler (2013); Su (2019)). It would be of interest to consider extensions of the method proposed by Su (2019) to dynamic networks, and to compare the performance with the approach presented here. A related topic is predicting how a certain opinion/behavior spreads in a network, and how individuals' behavior changes in the future based on social influence. Missing values in individuals' status or missing edges between individuals is quite common in real data. Devel-

oping methods to deal with missing data is also very useful in practice. Another interesting problem is to test whether social influence decreases as social distance increases. This is beyond the scope of this study and is left to future work.

Acknowledgments

This work was supported in part by National Science Foundation grants CCF-1934986 and DMS-2015561. We thank the editor, associate editor, and two referees for their constructive comments.

Appendix

A. Proof of Thereom 1

Part (a):

For testing: H_0 : DOI $= D-1$ v.s. H_1 : DOI $\geq D$, the test statistic is $\hat{\beta}_D^2/\hat{\Sigma}(\hat{\beta}_D)$ in (4.5), where $\hat{\beta}_D$ is the estimate based on the generalized estimating equations and $\hat{\Sigma}(\hat{\beta}_D)$ is the sandwich estimator for the variance of $\hat{\beta}_D$. Under H_0, the longitudinal influence model is:

$$\begin{aligned} y_{i,t+1} &\sim \text{Bernoulli}(p_{i,t+1}), \\ \text{logit}(p_{i,t+1}) &= \gamma + \beta_0 y_{i,t} + \sum_{d=1}^{D} \beta_d x_{i,t}^d, \end{aligned}$$

where $\beta_D = 0$.

By the property of generalized estimating equations (Liang and Zeger (1986)), $\hat{\beta}_D$ is a consistent estimator for β_D and

$$\sqrt{n}\hat{\beta}_D \xrightarrow{d} N(0, V(\beta)_{D+2,D+2}),$$

where $V(\beta)_{D+2,D+2} = \lim_{n\to\infty} nV_n(\beta)_{D+2,D+2}$ and $V_n(\beta) = H_1(\beta)^{-1}H_2(\beta)H_1(\beta)^{-1}$ (Liang and Zeger (1986)). Thus,

$$\frac{\hat{\beta}_D^2}{V(\beta)_{D+2,D+2}/n} = \frac{n\hat{\beta}_D^2}{V(\beta)_{D+2,D+2}} \xrightarrow{d} \chi^2(1).$$

Given $n\hat{\Sigma}(\hat{\beta}_D)$ is a consistent estimator for $V(\beta)_{D+2,D+2}$ (Liang and Zeger (1986)) and by Slutsky's theorem, we have

$$W = \frac{\hat{\beta}_D^2}{\hat{\Sigma}(\hat{\beta}_D)} = \frac{n\hat{\beta}_D^2}{n\hat{\Sigma}(\hat{\beta}_D)} = \frac{n\hat{\beta}_D^2}{V(\beta)_{D+2,D+2}} \frac{V(\beta)_{D+2,D+2}}{n\hat{\Sigma}(\hat{\beta}_D)} \xrightarrow{d} \chi^2(1).$$

Since $c^* = \chi^2_{1-\alpha}(1)$, we have:

$$P(W > c^* \mid \text{DOI} = D - 1) \longrightarrow \alpha.$$

Part (b):

Under H_1, the longitudinal influence model is:

$$y_{i,t+1} \sim \text{Bernoulli}(p_{i,t+1}),$$
$$\text{logit}(p_{i,t+1}) = \gamma + \beta_0 y_{i,t} + \sum_{d=1}^{D^*} \beta_d x^d_{i,t},$$

where $\beta_D \neq 0$ for $1 \leq D \leq D^*$.

By the property of generalized estimating equations (Liang and Zeger (1986)), $\hat{\beta}_D$ is a consistent estimator for β_D and

$$\sqrt{n}(\hat{\beta}_D - \beta_D) \xrightarrow{d} N(0, V(\beta)_{D+2,D+2}),$$

where $V(\beta)_{D+2,D+2} = \lim_{n\to\infty} nV_n(\beta)_{D+2,D+2}$ and $V_n(\beta) = H_1(\beta)^{-1}H_2(\beta)$ $H_1(\beta)^{-1}$ (Liang and Zeger (1986)). Also $n\hat{\Sigma}(\hat{\beta}_D)$ is a consistent estimator for $V(\beta)_{D+2,D+2}$ (Liang and Zeger (1986)).

Since $\hat{\beta}_D$ is a consistent estimator for β_D, we have $\hat{\beta}_D \xrightarrow{p} \beta_D$. So $\hat{\beta}^2_D \xrightarrow{p} \beta^2_D > 0$. Note that $V(\beta)$ is a covariance matrix, so $V(\beta)_{D+2,D+2}$ is a finite positive number. Since $n\hat{\Sigma}(\hat{\beta}_D)$ is a consistent estimator of $V(\beta)_{D+2,D+2}$, we have

$$n\hat{\Sigma}(\hat{\beta}_D) \xrightarrow{p} V(\beta)_{D+2,D+2},$$

so $\hat{\Sigma}(\hat{\beta}_D) \xrightarrow{p} 0$. Since $\hat{\beta}^2_D \xrightarrow{p} \beta^2_D > 0$, we have $W = \hat{\beta}^2_D/\hat{\Sigma}(\hat{\beta}_D) \xrightarrow{p} \infty$. This shows $P(W > c^* \mid \text{DOI} \geq D) \xrightarrow{p} 1$, where $c^* = \chi^2_{1-\alpha}(1)$.

References

Brakefield, T. A., Mednick, S. C., Wilson, H. W., De Neve, J.-E., Christakis, N. A. and Fowler, J. H. (2014). Same-sex sexual attraction does not spread in adolescent social networks. *Archives of Sexual Behavior* **43**, 335–344.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* **357**, 370–379.

Christakis, N. A. and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine* **358**, 2249–2258.

Christakis, N. A. and Fowler, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine* **32**, 556–577.

De Domenico, M., Lima, A., Mougel, P. and Musolesi, M. (2013). The anatomy of a scientific rumor. *Scientific Reports* **3**, 2980.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–60.

Everitt, B. S. and Dunn, G. (2001). *Applied Multivariate Data Analysis*. Edward Arnold, London.

Halekoh, U., Højsgaard, S. and Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software* **15**, 1–11.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1–50.

Hogg, T. and Lerman, K. (2012). Social dynamics of Digg. *EPJ Data Science* **1**, 5.

Kempe, D., Kleinberg, J. and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Mednick, S. C., Christakis, N. A. and Fowler, J. H. (2010). The spread of sleep loss influences drug use in adolescent social networks. *PLoS One* **5**, e9775.

O'Malley, A. J. (2013). The analysis of social network data: An exciting frontier for statisticians. *Statistics in Medicine* **32**, 539–555.

Rosenquist, J. N., Murabito, J., Fowler, J. H. and Christakis, N. A. (2010). The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine* **152**, 426–433.

Su, X. (2019). *Variational Approximation for Importance Sampling and Statistical Inference on Social Influence*. PhD thesis. University of Illinois at Urbana-Champaign, Illinois, USA.

Sun, J. and Tang, J. (2011). A survey of models and algorithms for social influence analysis. In *Social Network Data Analytics* (Edited by C. C. Aggarwal), 177–214. Springer, Boston.

Valente, T. W. (1995). Network models of the diffusion of innovations. *Computational and Mathematical Organization Theory* **2**, 163–164.

VanderWeele, T. J. (2013). Inference for influence over multiple degrees of separation on a social network. *Statistics in Medicine* **32**, 591–596.

Xiang Cui

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, U.S.A.

E-mail: xiangc5@illinois.edu

Yuguo Chen

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, U.S.A.

E-mail: yuguo@illinois.edu

(Received August 2020; accepted June 2021)