Published in final edited form as:

Stat Politics Policy. 2012 February 1; 3(1): 2151–1050. doi:10.1515/2151-7509.1050.

Why and When “Flawed” Social Network Analyses Still Yield Valid Tests of no Contagion

Tyler J. VanderWeele,
Harvard University

Elizabeth L. Ogburn, and
Harvard University

Eric J. Tchetgen Tchetgen
Harvard University

Abstract

Lyons (2011) offered several critiques of the social network analyses of Christakis and Fowler, including issues of confounding, model inconsistency, and statistical dependence in networks. Here we show that in some settings, social network analyses of the type employed by Christakis and Fowler will still yield valid tests of the null of no social contagion, even though estimates and confidence intervals may not be valid. In particular, we show that if the alter’s state is lagged by an additional period, then under the null of no contagion, the problems of model inconsistency and statistical dependence effectively disappear which allow for testing for contagion. Our results clarify the setting in which even “flawed” social network analyses are still useful for assessing social contagion and social influence.

Keywords

confounding; contagion; dependence; social influence; social networks

In his paper, Lyons (2011) offers a number of criticisms of social network analyses that attempt to estimate contagion effects, such as those of Christakis and Fowler (2007, 2008). A number of his criticisms, and a number of other critiques (Cohen-Cole and Fletcher, 2008; Shalizi and Thomas, 2011; Noel and Nyhan, 2011), are important and need to be taken seriously in the conduct and interpretation of such studies. Some progress has been made in addressing or working around some of these critiques (Fowler and Christakis, 2008; Ver Steeg and Galstyan, 2010; Christakis and Fowler, 2011; VanderWeele, 2011). However, many of the issues raised have arguably not yet been dealt with adequately. In this paper, we offer further discussion on several points raised by Lyons (2011), focusing specifically on model consistency and inference. We argue that, although the issues raised by Lyons (2011) can lead to biased estimates and invalid inference, social network analyses like those of Christakis and Fowler (2007, 2008) will, in some circumstances, still suffice as a valid test of the null hypothesis of no contagion (no social influence) in the social network.

On p. 13 of his paper, Lyons (2011) considers a model like that used in Christakis and Fowler (2007, 2008) in which the log odds of the state of a "focal participant" or "ego" at time t, $Y_{i,t}$, is modeled as a linear function of the state of the "linked participant" or "alter" at time t, $Y_{j,t}$, and at time $t - 1$, $Y_{j,t-1}$, of the ego's state at time $t - 1$, $Y_{i,t-1}$, and of the covariates for the ego, X_i. In this model, β_1 is the coefficient for $Y_{j,t}$ (the alter's state at time t) and β_2 is the coefficient for $Y_{j,t-1}$ (the alter's state at time $t - 1$):

$$\log\left(\frac{P(Y_{i,t}=1|Y_{j,t},Y_{j,t-1},Y_{i,t-1},X_i)}{P(Y_{i,t}=0|Y_{j,t},Y_{j,t-1},Y_{i,t-1},X_i)}\right)=\beta_0+\beta_1Y_{j,t}+\beta_2Y_{j,t-1}+\beta_3Y_{i,t-1}+\beta_4^{'}X_i$$

Christakis and Fowler (2007, 2008) interpret the estimate of β_1 as their "contagion effect" or causal estimate of social influence. Lyons (2011) argues that, if, in the network, there is a person i with a tie to person k and that person k has a tie to person $m \neq i$ then the models themselves imply that $\beta_1 = 0$. The models themselves effectively contradict the existence of the very effect Christakis and Fowler want to assess. Lyons further argues that when the state is continuous and linear regression is used as in the loneliness social network analyses of Cacioppo et al. (2009) then if person i has a tie with person j and person j with person i, and if likewise person j has a tie with person k and person k with person j with $k \neq i$ then it follows from the models that $\beta_1 = \beta_2 = 0$.

This issue raised by Lyons is essentially that there are more equations than unknowns. This arises because the state of the ego at time t is regressed on the current state of the alter at the same time t, rather than only on the lagged state of the alter. When there is reciprocation between persons with regard to their ties, this creates modeling problems. Intuitively, the problem develops because the same variable at the same time period, e.g. the ego's state at time t, is the dependent variable in one regression and the independent variable in another regression.

As noted by Lyons, the models themselves then effectively contradict the conjecture of social influence that Christakis and Fowler want to assess. An important exception, however, arises when the null hypothesis of no contagion is true. In this case, provided that homophily and environmental confounding have been properly controlled for, β_1 does indeed equal 0 (cf. Shalizi and Thomas, 2011). And, if $\beta_1 = 0$, then the models may be correctly specified, provided e.g. the log odds of the ego's state is indeed linear in the covariates. Under the null hypothesis of no contagion, the problem of model inconsistency effectively vanishes[1]. Thus, under the null hypothesis of no contagion, a statistical test for $\beta_1 = 0$ would provide a joint test of (i) no contagion, (ii) no homophily or environmental confounding conditional on the covariates and (iii) correct model specification with regard to the covariates. The estimate and confidence interval for β_1 would *not* constitute a valid estimate of the contagion effect, even if there is no homophily or environmental confounding conditional on the covariates. However, whether the confidence interval for β_1 contained 0 would constitute a valid test of the null hypothesis of no contagion, again

[1] The problem of model consistency may still arise if multiple alters are used for a single ego, but such issues would not arise with the Framingham Heart Study data with the analyses of mutual friends, ego-nominated friendships, or spouses since, in these cases, there will only be one alter per ego. In the Framingham Heart Study, each ego nominates only one friend.

provided the assumptions of no homophily and no environmental confounding conditional on the covariates and that of correct model specification with respect to the covariates held. Under these assumptions, we can in theory do testing, but not estimation.

This brings us to yet another critique offered by Lyons (2011), that of statistical modeling under the dependence structures that are generated by a social network. Christakis and Fowler (2007, 2008) use a method referred to as generalized estimating equations, clustering on the ego, to take into account the use of multiple time points for the ego. Unfortunately, as Lyons (2011) notes, this is not the only source of dependence in the data. If there is social influence (contagion) then the clusters defined by the ego will not be independent of one another. Moreover, even under the null of no contagion, when contemporaneous ego-alter data is used, the generalized estimating equations standard error is not always valid. Christakis and Fowler (2007, 2008) consider social influence for different types of relationships including ego-nominated friends, alter-nominated friends, mutual friends, spouses, neighbors and siblings. We show in the Appendix that because Christakis and Fowler (2007, 2008) use contemporaneous data for the ego and the alter, and because one person's state at time t is thus both an outcome in one regression and an independent variable in another, the standard errors for β_1 obtained by Christakis and Fowler (2007, 2008) are anti-conservative whenever relationships are reciprocal e.g. for mutual friends, spouses, siblings and neighbors. In these cases, even under the null hypothesis of no contagion, the standard errors will be invalid and the confidence intervals will be too narrow. One could derive a valid estimator of the standard error under the null but unfortunately the generalized estimating equation standard error used by Christakis and Fowler (2007, 2008) is not valid. However, we also show that for relationships which are not reciprocal, e.g. ego-nominated friendships or alter-nominated friendships (that are not mutual friendships), the generalized estimating equation standard error used by Christakis and Fowler (2007, 2008) is valid under the null hypothesis of no contagion and thus whether their confidence interval includes 0 does constitute a valid test for the null of no contagion, provided control has been made for homophily and environmental confounding.[2]

For the purposes of testing, both the problem of model inconsistency and the problems of statistical dependence and standard error estimation can be easily addressed if the alter's state is lagged by an additional period in the regressions. The argument used by Lyons (2011) to show that the models are inconsistent in the presence of contagion is no longer applicable. Moreover, under the null of no contagion, and provided adequate control has been made for homophily and environmental confounding, the clusters defined by the ego are independent of one another, avoiding statistical dependence throughout the network. Finally, by lagging the alter's state by an additional period, the same variable is no longer an outcome in one regression and a dependent variable in another regression at the same period time t, circumventing the issue of obtaining, under the null, valid standard errors when using generalized estimating equations. The generalized estimating equation standard error will be valid, under the null of no contagion. Thus, if a researcher lags the alter's state by an extra time period so that the log odds of the ego's state at time t, $Y_{i,t}$, is modeled as a linear

[2] Of course, such a test statistic will only be useful if it has non-trivial power; however, in the analyses of Christakis and Fowler (2007, 2008), they were able to reject the null at least for ego-nominated ties.

function of the alter's state at time $t - 1$, $Y_{j,t-1}$, and at time $t - 2$, $Y_{j,t-2}$, the ego's state at time $t - 1$, $Y_{i,t-1}$, and the covariates for the ego:

$$\log\left(\frac{P(Y_{i,t}=1|Y_{j,t-1},Y_{j,t-2},Y_{i,t-1},X_i)}{P(Y_{i,t}=0|Y_{j,t-1},Y_{j,t-2},Y_{i,t-1},X_i)}\right)=\beta_0+\beta_1Y_{j,t-1}+\beta_2Y_{j,t-2}+\beta_3Y_{i,t-1}+\beta_4^{'}X_i$$

then, whether the generalized estimating equation confidence interval for the coefficient of $Y_{j,t-1}$ contains 0 will constitute a valid test of the null of no contagion. We can at least still do testing using the same approach of Christakis and Fowler (2007, 2008) but simply lagging the alter's state by an additional period.

All of our discussion thus far has assumed that adequate control has been made for homophily and environmental confounding. As noted by Lyons (2011) and by Shalizi and Thomas (2011), this is, of course, a very strong assumption. VanderWeele (2011) proposed a sensitivity analysis technique to assess the extent to which an unmeasured factor responsible for homophily or environmental confounding would have to be related to both the ego's and the alter's state in order to substantially alter qualitative and quantitative conclusions. The technique itself made simplifying parametric assumptions but a more general approach could alternatively be used (VanderWeele, 2011; VanderWeele and Arah, 2011). Unfortunately, however, it is not clear that this technique would apply in the context of inconsistent models when contemporaneous data for the ego and the alter are used. This is because the sensitivity analysis parameters in VanderWeele (2011) related the observed expectation for the ego's state, controlling for observed covariates, to the expectation that would have been obtained had control also been possible for an unobserved covariate; however, when the models are inconsistent then it is no longer clear that the estimates, e.g. in Christakis and Fowler (2007, 2008), using the observed data, provide a consistent estimate of the expectation conditional on the observed covariates, for the very reasons raised by Lyons. The sensitivity analysis technique could, however, be applied to estimates obtained by lagging the alter's state by an additional period because, once again, the problem of model inconsistency then no longer arises.

We have given numerous arguments for lagging the alter's state by an additional period: (1) the problem of model inconsistency raised by Lyons (2011) does not arise, (2) the analyses using generalized estimating equations clustering by ego as in Christakis and Fowler (2007, 2008) will give valid tests of the null of no contagion, and (3) the sensitivity analysis technique of VanderWeele (2011) can be applied to the estimates obtained from such analyses.

In fact, Christakis and Fowler (2007, 2008) report, in the online supplement to their papers, that they ran such analyses in which the alter's state was lagged by an additional period and that the results of such analyses were similar to those of their main analyses using contemporaneous data for the ego and alter, i.e. they once again find evidence of significant contagion effects for smoking and obesity. Moreover, with these lagged social network analyses, the sensitivity analysis techniques to assess that the extent to which latent homophily and unmeasured environmental confounding could explain away the estimates are again applicable and suggest the contagion effect for smoking cessation between spouses and obesity between mutual friends are quite robust to potential latent homophily and unmeasured environmental confounding (VanderWeele, 2011).

A few further caveats are, however, in order. First, the sensitivity analysis techniques, in their present form, are not applicable to dynamic forms of homophily, such as “unfriending”, as considered by Noel and Nyhan (2011). However, this unfriending problem in the Framingham Heart Study data used by Christakis and Fowler (2007, 2008) does not seem, by Noel and Nyhan’s own simulations, sufficiently common to result in substantial biases. Second, even with the alter’s state lagged an additional period, the sensitivity analysis technique of VanderWeele (2011) is applicable to the estimates, but may not be to the limits of the confidence interval obtained by using generalized estimating equations, because, under the alternative hypothesis that contagion effects are present, the standard error for the supposed contagion effect still may not be valid because of statistical dependence in outcomes across the network. If valid confidence intervals were obtained the sensitivity analysis technique of VanderWeele (2011) would be applicable to the limits of the confidence interval as well. Finally, although some progress can be made with testing the null of no contagion, ultimately, we would also like to be able to obtain valid inferences and confidence intervals, not just tests and estimates. Doing so will require the development of statistical theory to handle the sorts of dependence structures that arise on social networks. In our view, this should be one of the central priorities in subsequent work that aims to provide a more rigorous foundation for the types of social network analyses for contagion effects exemplified by the studies of Christakis and Fowler (2007, 2008).

Acknowledgments

This research was support by NIH grants ES017876 and HD060696.

References

Cacioppo JT, Fowler JH, Christakis NA. Alone in the crowd: the structure and spread of loneliness in a large social network. Journal of Personality and Social Psychology. 2009; 97(6):977–991. [PubMed: 19968414]

Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. New England Journal of Medicine. 2007; 357:370–379. [PubMed: 17652652]

Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. New England Journal of Medicine. 2008; 358:2249–2258. [PubMed: 18499567]

Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. Statistics in Medicine. 2011 to appear.

Cohen-Cole E, Fletcher JM. Is obesity contagious? Social network vs. environmental factors in the obesity epidemic. Journal of Health Economics. 2008; 27(5):1382–1387. [PubMed: 18571258]

Fowler JH, Christakis NA. Estimating peer effects on health in social networks. Journal of Health Economics. 2008; 27(5):1386–1391.

Lyons R. The spread of evidence-poor medicine via flawed social-network analyses. Statistics, Politics and Policy Article. 2011; 2(1):Article 2, 1–26.

Noel H, Nyhan B. The ’unfriending’ problem: the consequences of homophily in friendship retention for causal estimates of social influence. Social Networks. 2011; 33:211–218.

Shalizi CR, Thomas AC. Homophily and contagion are generically confounded in observational social network studies. Sociological Methods and Research. 2011; 40:211–239. [PubMed: 22523436]

VanderWeele TJ. Sensitivity analysis for contagion effects in social networks. Sociological Methods and Research. 2011; 40:240–255.

VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. Epidemiology. 2011; 22:42–52. [PubMed: 21052008]

Ver Steeg, G.; Galstyan, A. Ruling out latent homophily in social networks. NIPS Worksop on Social Computing; 2010. URL: http://mlg.cs.purdue.edu/lib/exe/fetch.php?id=schedule&cache=cache&media=machine_learning_group:projects:paper19.pdf

Appendix

Consider outcome data $Y_{i,t}$, $t = 1, \dots, T; i = 1, \dots n$; let $X_{i,t}$ denote p covariates for person i observed up to time t. If $X_{i,t}$ is time-invariant then we could also write $X_{i,t} = X_i$. Let $\mathscr{R}_{i,t}$ denote the set of all $Y_{j,t}$ with a specific type of tie to person i at time t, $i \neq j$, which can be of Type A for an “alter-nominated tie”, of type E for an “ego-nominated tie”, or of type M for a “mutually-nominated tie” (or similarly for any other type of tie which is reciprocal e.g. spouse, neighbor, sibling). To test for contagion we may test the null hypothesis H_0 that $Y_{i,t}$ is jointly independent of $\{Y_{j,t}: Y_{j,t} \in \mathscr{R}_{i,t}\}$ given $(X_{i,t}, Y_{i,t-1}, Y_{j,t-1})$. We make the following assumptions:

1. The cardinality of $\mathscr{R}_{i,t}$ remains bounded as n goes to infinity.
2. The support of $X_{i,t}$ is bounded.
3. $Y_{i,t}$ is the independent of $(\{X_{j,t'} : t' \leq t\}, \{Y_{j,t'} : t' < t - 1\}, \{X_{j,t'} : t' < t\})$ given $(X_{i,t}, Y_{i,t-1}, Y_{j,t}, Y_{j,t-1})$.

$$\begin{gathered}\text{logit}\, P_{i,j,t}(\beta) := \text{logit}\Pr(Y_{i,t}=1 \mid Y_{j,t}, X_{i,t}, Y_{i,t-1}; \beta)\\ = \beta_0 + \beta_1 Y_{j,t} + \beta_2 Y_{j,t-1} + \beta_3 Y_{i,t-1} + \beta_4' X_{i,t},\\ \text{for all } Y_{j,t} \in \mathscr{R}_{i,t}, \text{where } \beta_1 = 0 \text{ encodes the null effect of } Y_{j,t},\\ \text{and } \beta = [\begin{matrix}\beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4^T\end{matrix}]^T = [\begin{matrix}\beta_0 & 0 & \beta_2 & \beta_3 & \beta_4^T\end{matrix}]^T.\end{gathered} \qquad (1)$$

Christakis and Fowler (2007, 2008) estimate β by maximizing the objective function

$$\log \prod_i \prod_t \prod_{j:Y_{j,t} \in \mathscr{R}_{i,t}}^{Y_{i,t}} P_{i,j,t}(\beta^*)^{Y_{i,t}\,(1-Y_{i,t})} (1 - P_{i,j,t}(\beta^*))^{(1-Y_{i,t})}$$

with respect to β^* which produces β, $\hat{}$ the solution to the estimating equation:

$$\begin{gathered}\sum_i \sum_t \sum_{j \neq i: Y_{j,t} \in \mathscr{R}_{i,t}} U_{i,j,t}\left(\hat{\beta}\right) = 0\\ \text{where } U_{i,j,t}\left(\hat{\beta}\right) = W_{i,j,t}\varepsilon_{it}\\ W_{i,j,t} = [\begin{matrix}1 & Y_{j,t} & Y_{j,t-1} & Y_{i,t-1} & X_{i,t}^T\end{matrix}]^T\\ \varepsilon_{it} = \left\{Y_{i,t} - P_{i,j,t}\left(\hat{\beta}\right)\right\}\end{gathered} \qquad (2)$$

this set of equations define a standard GEE for correlated outcomes with a logit link, and the independence working correlation matrix. In this setting, the large sample variance of $\hat{\beta}$ is typically approximated by Σ_{emp} =

$$\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1} \times n^{-1}\left\{\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)\right]^{\otimes 2}\right\} \times\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1}$$

We now state the main result.

Result

Suppose assumptions 1-4 hold, then, under H_0, the following hold:

i. if for all *i* and for all *t*, $R_{i,t}$ is strictly of type E or A only, then Σ_{emp} is, when *n* is large, approximately equal to the large sample variance-covariance of $\hat{\beta}$.

ii. if for all *i* and for all *t*, $R_{i,t}$ is strictly of type M only, then Σ_{emp} is guaranteed, when *n* is large, to be anti-conservative; that is Σ_{emp} is generally smaller (in the semipositive definite sense) than the variance-covariance of $\hat{\beta}$.

Proof

Under H_0 it can be verified that under Assumption 4, $\beta=\left[\begin{array}{ccccc}\beta_0 & 0 & \beta_2 & \beta_3 & \beta_4^T\end{array}\right]^T$ solves the equation

$$\mathbb{E}\left\{U_{i,j,t}\left(\beta\right)\right\}=0$$

which in turn implies that under mild regularity conditions, $\hat{\beta}$ is consistent for β. Furthermore, a Taylor series expansion of equation (2) can be used to establish that in large samples, under Assumptions 1-4:

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \approx \left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1} \frac{1}{\sqrt{n}}\sum_i\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)$$

This further implies that in large samples, the variance of $\sqrt{n}\left(\hat{\beta}-\beta\right)$ is approximately equal to

$$\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1} \times n^{-1}\mathbb{E}\left\{\sum_i\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)\right\}^{\otimes 2}$$
$$\times\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1}$$

where $A^{\otimes 2} = AA^T$. The middle factor reduces to

$$\begin{aligned} &n^{-1}\mathbb{E}\left\{\sum_i\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)\right\}^{\otimes 2} \\ &=n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)\right]^{\otimes 2} \\ &+n^{-1}\sum_{i\neq j} \mathbb{E}\left[\left\{\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)\right\}\left\{\sum_t \sum_{k:Y_{k,t}\in\mathscr{R}_{j,t}} U^T_{j,k,t}\left(\beta\right)\right\}\right] \end{aligned}$$

and

$$\begin{aligned} &n^{-1}\sum_{i\neq s} \mathbb{E}\left[\left\{\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)\right\}\left\{\sum_{t'} \sum_{k:Y_{k,t'}\in\mathscr{R}_{s,t'}} U^T_{s,k,t'}\left(\beta\right)\right\}\right] \\ &=n^{-1}\sum_{i\neq s} \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \sum_{k:Y_{k,t}\in\mathscr{R}_{s,t}} U_{i,j,t}\left(\beta\right)U^T_{s,k,t}\left(\beta\right)\right] \end{aligned}$$

since under Assumption 3, for $t < t'$, $Y_{j,t} \in \mathcal{R}_{i,t}$, $Y_{k,t'} \in \mathcal{R}_{i,t}$ and $i \neq s$

$$\mathbb{E}\left\{U_{i,j,t}\left(\beta\right)U^T_{s,k,t'}\left(\beta\right)\right\}=0.$$

Furthermore, Assumption 3 implies that:

$$\begin{aligned} n^{-1}\sum_{i\neq s} \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \sum_{k:Y_{k,t}\in\mathscr{R}_{s,t}} U_{i,j,t}\left(\beta\right)U^T_{s,k,t}\left(\beta\right)\right] &=n^{-1}\sum_{i\neq s}\sum_t 1(Y_{s,t}\in\mathscr{R}_{i,t})1(Y_{i,t}\in\mathscr{R}_{s,t})\mathbb{E}\left\{U_{i,j,t}\left(\beta\right)U^T_{s,k,t}\left(\beta\right)\right\} \\ &=n^{-1}\sum_{i<s}\sum_t 1(Y_{s,t}\in\mathscr{R}_{i,t})1(Y_{i,t}\in\mathscr{R}_{s,t})\left[\mathbb{E}\left\{U_{i,s,t}\left(\beta\right)U^T_{s,i,t}\left(\beta\right)\right\}+\mathbb{E}\left\{U_{s,i,t}\left(\beta\right)U^T_{i,s,t}\left(\beta\right)\right\}\right] \\ &=n^{-1}\sum_{i<s}\sum_t 1(Y_{s,t}\in\mathscr{R}_{i,t})1(Y_{i,t}\in\mathscr{R}_{s,t})\left[\mathbb{E}\left\{W_{i,s,t}W^T_{s,i,t}\varepsilon_{it}\varepsilon_{st}\right\}+\mathbb{E}\left\{W_{s,i,t}W^T_{i,s,t}\varepsilon_{it}\varepsilon_{st}\right\}\right] \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}\left\{W_{i,s,t}W^T_{s,i,t}\varepsilon_{it}\varepsilon_{st}\right\}=&\left[\begin{array}{cc} 0_{(p+2)\times(p+2)} & 0_{(p+2)\times 1} \\ 0_{1\times(p+2)} & P_{i,s,t}\left(\beta\right)\left(1-P_{i,s,t}\left(\beta\right)\right)P_{s,i,t}\left(\beta\right)\left(1-P_{s,i,t}\left(\beta\right)\right) \end{array}\right] \\ &:=\Gamma_{i,s,t} \end{aligned}$$

Therefore the variance of $\sqrt{n}\left(\hat{\beta}-\beta\right)$ is approximately equal to

$$\begin{array}{c}\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1} \\ \times n^{-1}\mathbb{E}\left\{\begin{array}{c} \sum_i \mathbb{E}\left[\sum_t\sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} U_{i,j,t}\left(\beta\right)\right]^{\otimes 2} \\ +\sum_{i<s}\sum_t 1(Y_{s,t}\in\mathscr{R}_{i,t})1(Y_{i,t}\in\mathscr{R}_{s,t})\Gamma_{i,s,t}\end{array}\right\} \\ \times\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1}.\end{array}$$

The standard sandwich estimator implemented by Christakis and Fowler (2007, 2008) is under H_0 approximately equal to Σ_{emp} and therefore the bias of their estimator is approximately equal to

$$\begin{array}{c}\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1} \\ \times n^{-1}\sum_{i<s}\sum_t 1(Y_{s,t}\in\mathscr{R}_{i,t})1(Y_{i,t}\in\mathscr{R}_{s,t})\Gamma_{i,s,t} \\ \times\left\{n^{-1}\sum_i \mathbb{E}\left[\sum_t \sum_{j:Y_{j,t}\in\mathscr{R}_{i,t}} \frac{\partial U_{i,j,t}\left(\beta^*\right)}{\partial\beta^{*T}}|_\beta\right]\right\}^{-1}\end{array}$$

which can be verified to be semipositive definite. Under the null hypothesis H_0, the variance estimator of $\hat{\beta}$ used by Christakis and Fowler (2007, 2008) is valid in case (i) of the result, since then

$$1(Y_{s,t}\in\mathscr{R}_{i,t})1(Y_{i,t}\in\mathscr{R}_{s,t})=0\,\text{of all}\,t,\ \text{and}\ (s,i),\, s\neq i$$

and thus the bias term is equal to 0. However, in case (ii), with mutual/reciprocal ties, this is not the case and thus their variance estimator may be anti-conservative in large samples.