2010 Conference Proceedings

2010

The "Unfriending" Problem: The Consequences of Homophily in Friendship Retention for Causal Estimates of Social Influence

Hans Noel
University of Michigan - Ann Arbor, hansnoel@umich.edu

Brendan Nyhan
University of Michigan - Ann Arbor, bnyhan@umich.edu

The "Unfriending" Problem
The Consequences of Homophily in Friendship Retention for Causal Estimates of Social Influence

Hans Noel
RWJ Scholar in Health Policy Research
University of Michigan
hansnoel@umich.edu

Brendan Nyhan
RWJ Scholar in Health Policy Research
University of Michigan
bnyhan@umich.edu

May 14, 2010

Abstract

Christakis, Fowler, and their colleagues have recently published numerous articles estimating "contagion" effects in social networks (Christakis and Fowler 2007, 2008; Fowler and Christakis 2008*a*; Cacioppo, Fowler and Christakis 2009; Rosenquist et al. 2010; Rosenquist, Fowler and Christakis 2010; Mednick, Christakis and Fowler 2010). In response to concerns that their results are driven by homophily, Christakis and Fowler describe Monte Carlo results showing no evidence of homophily-induced bias in their statistical model's estimates of peer effects (Fowler and Christakis 2008*b*, N.d.; Fowler et al. N.d.). However, their simulations do not address the role of homophily in friendship *retention*, which may cause significant problems in longitudinal social network data. We investigate the effects of this mechanism using Monte Carlo simulations and demonstrate that homophily in friendship retention induces significant upward bias and decreased coverage levels in the Christakis and Fowler model.

We thank James Fowler for sharing the simulation code adapted in this paper, and Megan Andrew, Michael Heaney, Jonathan Ladd, and Fabio Rojas for helpful suggestions.

1 Introduction

Until recently, social science and medical research has given relatively little attention to the influence of friends and family on human behavior. Instead, people have been studied largely as atomistic individuals ripped from their social context. Thankfully, this impoverished approach has started to give way to an interdisciplinary movement seeking to understand the influence of social networks in domains ranging from health to politics. While much progress has been made, scholars have struggled for years with the difficulties in obtaining valid causal estimates of peer effects in observational data, especially as the problems raised in Manski's seminal article (1993) have become more widely appreciated.

In this context, recent studies by Christakis, Fowler, and their colleagues estimating "contagion" effects in social networks have attracted a great deal of attention and criticism. These studies, which leverage rich longitudinal social network data from the Framingham Heart Study and National Longitudinal Study of Adolescent Health (Add Health), make strong claims about the effects of one's friends[1] on a wide range of dependent variables: obesity (Christakis and Fowler 2007), smoking (Christakis and Fowler 2008), happiness (Fowler and Christakis 2008*a*), loneliness (Cacioppo, Fowler and Christakis 2009), depression (Rosenquist et al. 2010), alcohol consumption (Rosenquist, Fowler and Christakis 2010), and sleep loss (Mednick, Christakis and Fowler 2010). Each paper uses the same source data and statistical model.

In response to concerns that their findings are biased due to homophily, Christakis and Fowler present Monte Carlo simulations indicating that homophily in friendship formation does not result in increased bias under their model (2008*b*; N.d.). However, these simulations test the CF model under low levels of homophily in friendship formation and do not account for homophily in friendship *retention*, a key concern in analysis of longitudinal social network data.

[1] The studies typically also estimate social influence effects among family members; we do not consider the validity of those estimates here.

This paper evaluates the robustness of the CF model to homophily in friendship formation *and* retention. Our Monte Carlo simulations, which are adapted from those of Christakis and Fowler, show that the CF model's estimates of peer effects are unbiased and have accurate coverage levels when homophily in friendship retention is not present, but show substantial upward bias and decreased coverage levels for peer effects as homophily in friendship retention increases. As such, their estimates of peer effects are likely to be too optimistic.

2 The Christakis-Fowler model: A solution?

It is now widely accepted that peer effects are an important phenomenon in human behavior. Results from cases in which peers were randomly or quasi-randomly assigned such as college roommates have provided credible evidence of such effects (e.g., Sacerdote 2001; Zimmerman 2003; for a review, see Kremer and Levy 2008). However, random assignment of peers is typically not feasible for many populations and phenomena of interest. In these cases, researchers typically must use observational data, which creates difficult inferential problems.

In particular, Manski (1993) identified several key difficulties in estimating peer effects. The most important for our purposes is that some "correlated effect" may induce correlation in behavior among friends that is not the result of the behavior or characteristics of one's peers. This effect could be a common environmental shock (e.g., the opening of a new McDonald's in a study of obesity) or the result of a shared characteristic (e.g., a high degree of athleticism). While the incidence of environmental shocks may vary, a vast literature demonstrates that peers are likely to be similar on a range of characteristics due to homophily—the tendency of humans to associate with people who are like themselves (for a review, see McPherson, Smith-Lovin and Cook 2001). Distinguishing between environmental shocks, homophily, and peer effects is thus a very difficult challenge.

A number of authors have proposed strategies for resolving these problem and accurately estimating peer effects (recent examples include Anagnostopoulos and Mahdian 2008; Bramoullé, Djebbaria and Fortin 2009; and Arala, Muchnika and Sundararajana 2009; for a review of the previous literature, see Soetevent 2006). However, none has been as widely employed or published as that of Christakis, Fowler, and their colleagues (hereafter CF), who estimate versions of the following model for ego *i* and alter *j* observed at time t_0 and time t_1:

$$\mathrm{Y_{i,t_1}} = f(\mathrm{Y_{i,t_0}}, \mathrm{Y_{j,t_0}}, \mathrm{Y_{j,t_1}}, \text{controls}) \tag{1}$$

CF's models are typically estimated using generalized estimating equations (GEE) with an independent correlation structure to account for repeated observations of the same ego (specifically, those who name more than one friend) and dyad (those who name each other and are thus included twice, one each as the ego and once as the alter). The functional form of the model varies depending on the distribution of the dependent variable (logistic regression if the dependent variable is binary; ordinary least squares regression otherwise). CF argue that this specification controls for initial homophily (i.e., the likely similarity between Y_{i,t_0} and Y_{j,t_0}), allowing us to identify the effect of *changes* in the alter's trait from t_0 to t_1 by estimating the effect of Y_{j,t_1} controlling for Y_{j,t_0}. In Christakis and Fowler (2007), they write that including alters' lagged obesity as a covariate "controlled for homophily" (373). In later work, the language is somewhat more hedged—for instance, they write in Christakis and Fowler (2008, 2251) that a lagged measure of alter smoking "*helped* to account for homophily" (our emphasis)—but the suggestion that the coefficient for Y_{j,t_1} is a causal estimate of peer effects remains.

Cohen-Cole and Fletcher (2008, 1385) question whether CF's model adequately controls for homophily (see also Lyons N.d.; Shalizi and Thomas 2010). In response, CF describe Monte Carlo simulation results "documenting that homophily (ranging from no homophily to complete homophily) does not result in bias in the estimates of induction in this model specification" (Fowler and Christakis 2008*b*, 1404). These results, which are presented in an unpublished version of the paper on Fowler's website (Fowler and Christakis N.d.) and in a very similar form in Fowler et al. (N.d.), are derived from a stylized model in which a population of individuals with a randomly chosen value on some characteristic of interest form friendships and then influence each other or not (we discuss the details of the procedure in more detail below). CF find that estimates of this influence coefficient are approximately unbiased across varying levels of homophily when the true peer effect is 0 and have a slight downward bias when the true peer effect is 0.1. On this basis, they conclude that "This simulation evidence suggests that the [Cohen-Cole and Fletcher] assertion that homophily causes us to overestimate the size of the induction effect is false." However, as we discuss below, their simulation does not incorporate friendship *attrition* and thus fails to fully account for the effects of homophily.

3 The "unfriending" problem in longitudinal data

Due to the prevalence of cross-sectional data and interest in fixed characteristics such as race and gender, scholars of social networks have tended to think about homophily in relatively static terms and to analyze it as a propensity to form ties with others who share similar characteristics. However, social networks are actually the result of a dynamic process of friendship *formation* and *dissolution*.

Homophily influences social relationships through both of these mechanisms. Just as people who are similar are more likely to be friends, friends who are less similar are more likely to *stop* being friends. Most of us have had friends from whom we have grown apart in this way. As we have less in common with those people, we stop spending time with them and eventually fall out of touch. In some cases, one person may deliberately end the relationship as a result of differences in political views, religious beliefs, mood, alcohol consumption, or a range of other behaviors and characteristics. We call this the "unfriending" problem in honor of the Facebook practice of removing a person from one's list of friends on the online social network site.

Several cases of this pattern have been documented in the sociology literature on friendships—in particular, a two-wave study of adolescent friendships by Kandel (1978). She describes homophily in friendship retention based on both initial characteristics and subsequent behavior (430):

> At time 1, prior to any subsequent change, pairs that will remain stable over time are much more similar in their behaviors and values [marijuana use, educational goals, political views, and delinquency] than the subsequently unstable pairs... At time 2, homophily among former friends is lower than among new friendship pairs or stable pairs.

She interprets these results as a combination of selection (choosing to become and stay friends with those who are like you) and socialization (acting more like your friends in those relationships you maintain) (433–435):

> The results support the general conclusion that adolescents coordinate their choices of friends and their behaviors, in particular the use of marijuana, so as to maximize congruency within the friendship pair. If there is a state of unbalance such that the friend's attitude or behavior is incongruent with the adolescent's, the adolescent will either break off the friendship and seek another friend or will keep the friend and modify his own drug behavior.

Similarly, Newcomb, Bukowski and Bagwell (1999, 72) find that sixth grade students "are more likely to maintain a friendship if they choose a friend who is similar to them on these dimensions [aggression and class competence at the beginning of the study] than if they choose a friend who is less similar" and Degirmencioglu et al. (1998) finds that same-gender friendships among adolescents are more stable than cross-gender ones.

The unfriending problem is a significant concern for the CF approach, which relies on longitudinal network data. First, the specification of their generalized estimating equation models requires an ego to name an alter as a friend for two or more consecutive waves. What happens when some of the dyads at t_0 are no longer friends at t_1? Fowler and Christakis argue that including such friendship pairs in their data will bias the results against finding an effect because "it essentially adds 'random' non-friend relationships (i.e., people who are no longer friends) to the pool of friends" (Fowler and Christakis 2008*b*, 1401). This is a legitimate issue; non-friends presumably can no longer influence the person in question.

However, the friendships that have been terminated may not have be "random." Relationships often end for a reason. If the process of friendship termination is linked to changes between t_0 and t_1 in the underlying trait we are examining, it will induce an association between Y_{i,t_1} and Y_{j,t_1} that is not captured by the lagged value of the variable in question. In the CF model, the coefficient for Y_{j,t_1} is interpreted as a causal effect. As such, the association induced by homophily in the unfriending process could create the appearance of an influence effect even if none exists.[2]

4 Monte Carlo simulation procedure

To determine the extent to which homophily in friendship retention biases the estimates produced by CF's model, we conduct Monte Carlo simulations in which we know the true value of the parameter in question (in this case, 0).[3] Our procedure is adapted from code used in Fowler and Christakis (N.d.) and Fowler et al. (N.d.).

The simulation proceeds as follows:

1. A normally distributed trait Y_{t_0} is randomly generated at time t_0 for a population of n actors where $Y_{t_0} \sim \mathrm{N}(50, 10)$.

[2] In principle, one might attempt to model the selection process by which friendships are maintained in order to recover the true value of the influence coefficient. However, it seems impossible to obtain data that is granular enough to separate stochastic changes in the trait of interest from t_0 to t_1 from subsequent peer effects. In the absence of such data, accurately modeling the friendship retention process requires knowing the value of Y_{i,t_1} that would have been observed if no influence had taken place—an unobserved counterfactual.

[3] A broader question that we do not engage here is whether statistical models of such effects are formally identified. Shalizi and Thomas (2010) presents a graphical causal model arguing that such effects are generically unidentified in observational data for a person i when some latent trait "$X(i)$ directly influences $Y(i,t)$ for all t." In such cases, even controlling for $Y(i,t-1)$ and $Y(j,t-1)$ is not sufficient to identify the causal effect of a network tie $A(i,j)$ on $Y(i,t)$.

2. Differences in Y are computed for all dyads of actors i and j:

$$d_{i,j} = -|Y_{i,t_0} - Y_{j,t_0}|$$

   The difference term is negatively valued so that dyads with similar traits have high values.

3. Ties $A_{i,j}$ are created as a function of $d_{i,j}$ using a probit model based on a latent variable $A^*_{i,j}$. These ties are directed (i.e., $A_{i,j}$ does not imply $A_{j,i}$).

$$A_{i,j} = \begin{cases} 1 & \text{if} \quad A^*_{i,j} > \epsilon_{i,j} \sim N(0,1) \\ 0 & \text{if} \quad A^*_{i,j} \leq \epsilon_{i,j} \sim N(0,1) \end{cases}$$

   where

$$A^*_{i,j} = \beta_0 + \beta_1 d_{i,j}$$

   β_0 represents the baseline propensity to form ties and β_1 represents the coefficient for homophily (positive values indicate higher levels of homophily).

4. All actors receive a normally distributed, independent shock u to their trait Y_{t_0} where $u \sim \mathrm{N}(0,5)$.

5. All egos' values of Y are updated as a weighted average of their own current value of $Y_{t_0} + u$ and the average value of $Y_{t_0} + u$ for their alters:

$$Y_{i,t_1} = (1 - b_1)(Y_{i,t_0} + u_i) + b_1 \left(\frac{\sum_j A_{ij}(Y_{j,t_0} + u_j)}{\sum_j A_{ij}} \right)$$

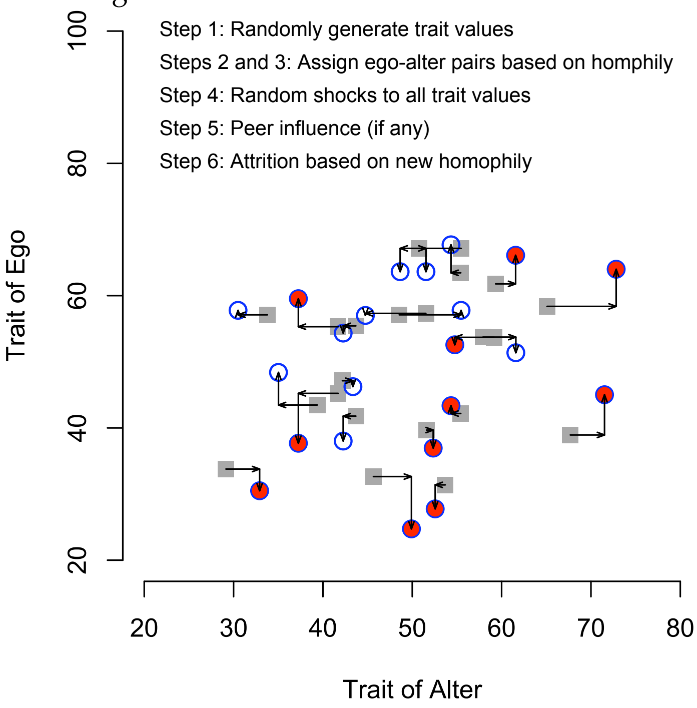
   where b_1 is a measure of the influence of alters on egos.

6. All actors update their friendship ties as in step 3.

7. We estimate b_1 for the egos i and alters j who remain friends using a generalized estimating equation specified in the same manner as CF (see Equation 1 above).

The simulation process is illustrated in Figure 1. Gray squares represent

Figure 1: Illustration of one simulation



values of Y for ego-alter pairs after initial friendships have been formed. There is some correlation due to homophily. The actors then both experience shocks to their values of the trait, represented by arrows. The new values of the trait are indicated by the open blue circles at the end of each path. The GEE should estimate the effect of the alter's shock on the ego controlling for the alter's previous Y value. However, at the friendship retention stage, some of the pairs cease to be friends. Only those pairs indicated by the solid red circles remain in the data.

This procedure modifies CF's original approach in two key ways. First, we model friendship using a latent variable probit model where $A^*_{i,j} = \beta_0 + \beta_1 d_{i,j} + \epsilon$ with a single error term $\epsilon \sim \mathrm{N}(0,1)$. By contrast, CF generate a probability of a tie that is a weighted average of Y_0 and a random component and then conduct a random draw with that probability to determine whether a tie exists.

The CF process combines two sources of random noise. The first is meant to model factors other than homophily in friendship choice, while the second models the inherently stochastic component of friendship formation. In practice, however, this partition is not readily interpretable—any unobservable influence on the outcome variable in a statistical model can reasonably be included in the stochastic component if it is not also systematically related to the independent variables.

More importantly, however, the result of this double-randomness is that CF's simulations do not generate high correlations in the outcome variable among friends even when friendships are formed on the basis of "complete homophily" (Fowler and Christakis 2008*b*, 1405). Our replications of CF's simulation model found correlations between ego and alter on outcome variable of 0.20 or lower. In practice, however, social network datasets often display higher levels of correlation among friends. For instance, the 2000 American National Election Study asks respondents to name up to four people to whom they regularly speak about politics and their best guess of the candidate for whom that person voted. The correlation in vote choice between respondents and their friends ranged from 0.43 for the last person named to 0.57 for the first person named (results available upon request). Even accounting for the effects of projection (i.e., falsely perceiving that your friends agree with your views), these results suggest that simulations should consider higher levels of homophily. Similarly, the "Faux Magnolia High" friendship network of high school students shows a correlation of gender across all ties of approximately 0.39 (results available upon request).[4] By changing the coefficient for β_1, we can vary homophily up to

[4] The "Faux Magnolia High" dataset is a synthetic version of high school friendships based on the Add Health dataset for a number of large high schools in the American South.

realistically high levels.

The second difference we introduce is step 6, which repeats the friendship model from step 3, allowing for friendships to end based on homophily after a shock to Y. This step is crucial in longitudinal network data as discussed above. The shock u to Y_0 is assumed to be randomly distributed. However, if some people cease to be friends due to the values of the random shock u that they received, it will induce a correlation in outcomes Y_{i,t_1}, Y_{j,t_1} for the remaining subjects that will appear to be a causal influence effect. Controlling for lagged values of the trait for egos and alters will not resolve this problem.

5 Monte Carlo results

Following standard procedure in Monte Carlo evaluations of statistical models, we set the true contagion effect is 0 and estimate mean bias and coverage levels for the CF model. In our simulations, we set β_0 to -2.75 at the friendship formation stage in order to generate realistic numbers of friendships at t_0.[5] We then vary β_1 at both stages to generate realistic levels of homophily in both friendship formation and retention and also vary β_0 at the friendship retention stage to consider different friendship attrition rates:[6]

- Homophily in friendship formation: We vary the homophily parameter β_1 in step 3, considering no effect (0.0), a moderate effect (0.025), and a larger effect (0.05).
- Homophily in friendship retention: We vary the homophily parameter β_1 in step 6, considering a value representing no effect (0) up to a

Since the actual friendship ties in Add Health are protected, the Faux data provides a synthetic version for public use. It is designed to have the same basic characteristics as the real data (Resnick et al. 1997; Handcock et al. 2003).

[5] CF's Framingham subjects typically only name one friend due to the nature of the instrument used. However, this structure is unusual and we do not mimic it here.

[6] Additional simulations in which we also vary the standard deviation of the shock u to Y_0 and the number of subjects in the data n generate similar results. They are thus omitted but available on request.

large effect (0.05) in five increments.

- Levels of friendship retention: We vary β_0 in step 6, considering values of -0.25, 0.25, and 0.75 to represent realistic variation in attrition rates between t_0 and t_1.

Complete results from the Monte Carlo simulations, which were performed 500 times for each unique combination of model parameters, are in Table 1 at the end of the document. Figures 2 and 3 plot how well the estimator performed for varying values of the homophily coefficients when the constant $\beta_0 = 0.25$ and $\beta_0 = 0.75$ at the friendship retention stage (those for $\beta_0 = -0.25$ are not plotted but are even worse; see Table 1).

First, Figure 2 presents the probability that that estimated 95% confidence interval covers the true contagion parameter of 0. When homophily in friendship retention is 0, the confidence intervals for the CF models accurately bracket the true value approximately 95 percent of the time. However, as homophily in friendship retention increases, coverage rates decline dramatically. In the most extreme cases, the confidence interval includes the true value of the influence coefficient less than 15 percent of the time.

As we might expect, coverage degrades because the model is overestimating the contagion effect, as is evident in Figure 3, which presents the mean value of the estimated peer effect (which has been set to 0 in the simulations). When unfriending is not affected by homophily, the estimator is unbiased. But as homophily in friendship retention increases, estimated bias levels increase substantially—up to 0.08 in the worst case. As described above, the reason for the bias is that homophily-based friendship attrition induces a correlation among ego and alter traits that is not controlled for by including lagged values for ego and alter as controls. In additional simulations, we find that coverage problems worsen further as sample size increases (which increases the likelihood that the CF model will falsely reject the null hypothesis that the influence effect is 0).

Is this level of bias meaningful? As a point of comparison, we note that estimated peer effect coefficients for continuous variables in the literature are often in the range described in Table 1 and Figure 3. For example, the co-

Figure 2: Estimated coverage rates for CF model in Monte Carlo simulations

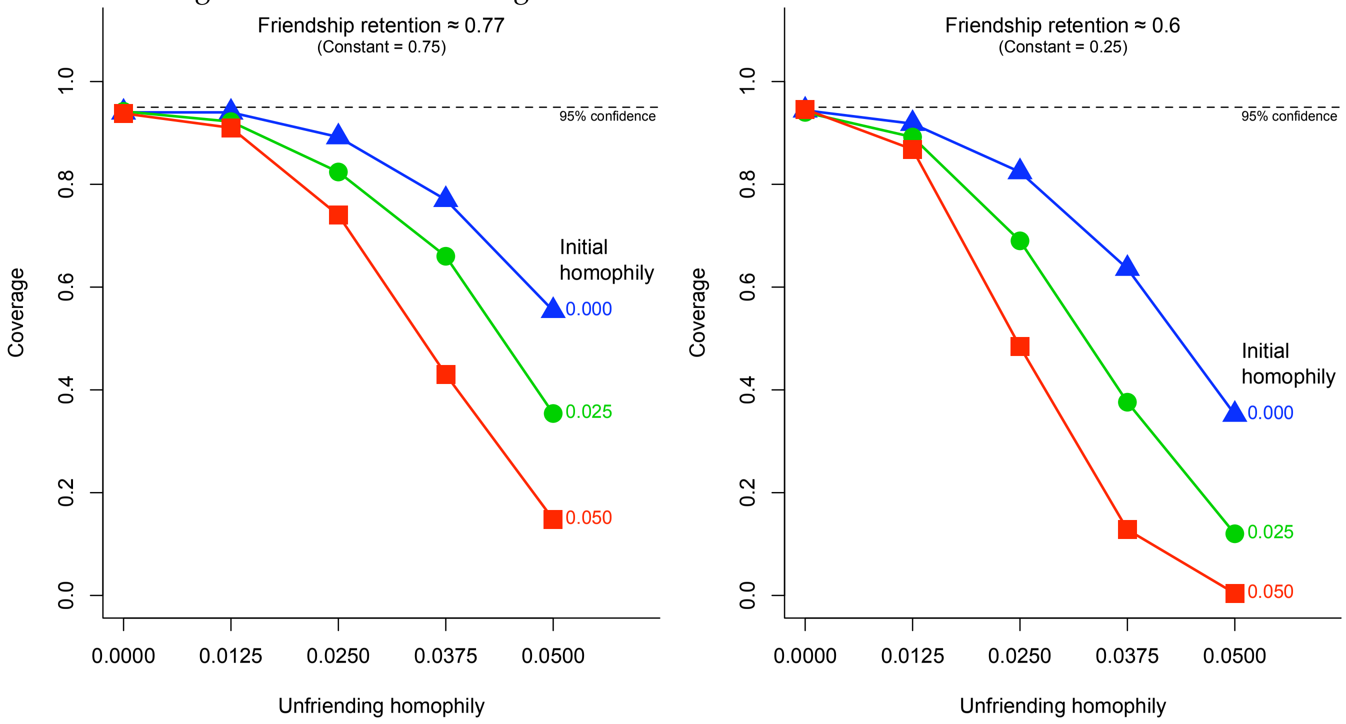
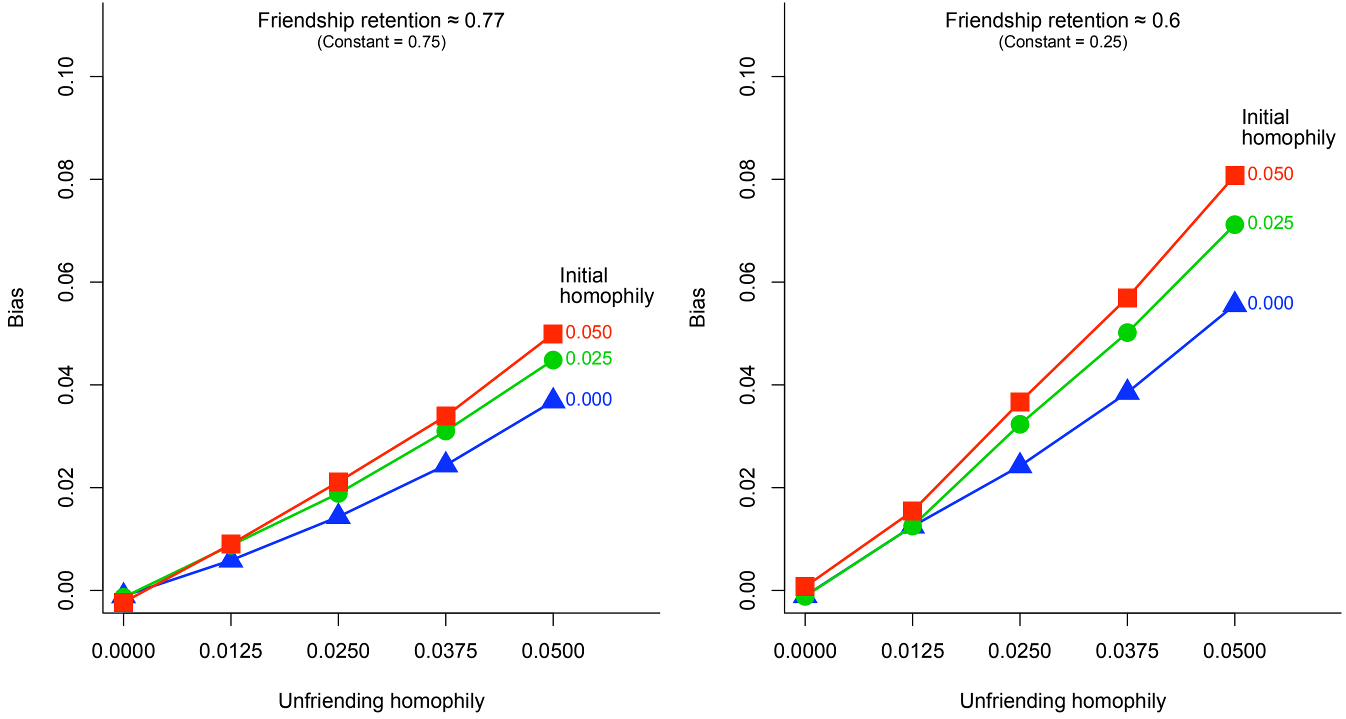


Figure 3: Estimated bias for CF model in Monte Carlo simulations



efficient on alter's current BMI is 0.053 in the Framingham Heart Study data (SE=0.018) and 0.033 in Add Health data (SE=0.014) (Fowler and Christakis 2008*b*).[7]

Table 1 also shows that ego-alter trait correlations in the simulation cover a plausible range of values both before and after friendship attrition. In particular, correlations at t_1 for the cases in which coverage was below 0.8 range from 0.11 to 0.60. Consider, for instance, the simulations summarized in line 9 of Table 1. Though homophily at the second stage was only half as strong ($\beta_1 = 0.025$) as initial levels ($\beta_1 = 0.05$), the coverage probability dropped all the way to 0.74 (ego-alter trait correlation at $t_1 = 0.48$).

Depending on the values selected for β_0, the current simulation yields friendship attrition rates of approximately 23% in the low attrition case (retention constant $\beta_0 = 0.75$), 40% in the moderate attrition case ($\beta_0 = 0.25$), and 60% in the high attrition case ($\beta_0 = -0.25$). These appear to be realistic. Observed levels of attrition in longitudinal social network studies depend on both the underlying attrition rate and the time elapsed between waves of the survey. For instance, Mollenhorst (2009) finds that about half of our friends are replaced every seven years. Framingham and Add Health reinterview respondents every three to four years, which implies attrition rates of approximately 30%. However, studies of social networks among children and adolescents have found higher rates of attrition. Schneider et al. (2006), for instance, found that 60% of third and fourth grade dyads in a Canadian sample and 71% in an Italian sample remained reciprocal friends over the course of a single school year.

6 Extensions

The framework we have developed here is potentially very flexible. In future work, we hope to extend it in several possible directions discussed below. (Our hope is that researchers concerned about other parameters can

[7] It would be worthwhile to repeat this exercise with a binary variable as in the CF studies of smoking or depression.

adapt it for their use in the future as well.)

6.1 Asymmetric friendship

CF claim in many of their articles to address the inferential threat posed by environmental shocks by demonstrating influence effects on egos who name alters but are not named reciprocally (this idea is also exploited in Anagnostopoulos and Mahdian 2008). They argue that, while we might expect most friendships to be reciprocal, it is frequently the case that person A names B as a friend, but person B does not name A. The implication is that person A views B as a close friend and so might be influenced by B, while person B does not view A in the same way and thus will not be influenced. This intuition suggests a way to rule out contextual effects due to environmental shocks. If the common context of the alter and ego is responsible for the effect, then it should not matter how the actors perceive their relationship. So if we observe differences in the effects between the two actors, CF argue, it must be due to their asymmetric perceptions.

However, the asymmetric friendship relation is not necessarily unrelated to the traits we are investigating. In a typical asymmetric relationship, both people know each other, but they perceive the relationship differently. Such perceptions are likely bound up with the effects we wish to isolate. First, the "revealed" network might include only those ties where that are highly valued by the namer. This selection process of naming may therefore exaggerate the magnitude of peer effects among friends and complicate the interpretation of those effects. Rather than your friends influencing you, we would conclude you are influenced by that subset of your friends whom you hold in high esteem. This argument suggests the need to model a status or valence dimension that is at least partially distinct from Y and influences selection of the friends in the naming process.

Along similar lines, Shalizi and Thomas (2010) present simulation results demonstrating that CF's claim about the direction of effects does not hold under plausible conditions:

> [T]he argument breaks down if two conditions are met: first,

> the influencers (the j in the pair) differ systematically in their values of X from the influenced (the i), and, second, different neighborhoods of X have different local relationships to Y .

In the case of the argument above, the values of X would represent a latent status trait on which influencers differ from the influenced.

6.2 Modeling latent traits

In our simulation, the trait of interest and the trait on which homophily is formed are exactly the same. Of course, in the real world, the characteristics on which homophily operates often have latent causes (e.g., a predisposition to become obese). Shalizi and Thomas (2010) have shown that such confounding cannot be accounted for with the CF model. The problem is that such a latent trait might have continuing independent effects on the outcome variable of interest even when we control for lagged values of the outcome variable. For instance, a latent interest in abstract problem solving might cause two scholars to become friends and it might also cause them to spend time at the computer instead of exercising. The result is that both scholars are more likely to gain weight as they age than comparable individuals. In this case, controlling for their initial weight would not account for their tendency to disproportionately gain weight during any subsequent study period. Formally, the problem is that X_{t_0} might influence friendship formation (A_{ij}) and also influence Y_{t_1} separately from Y_{t_0}. In that case, controlling for Y_{t_0} would not sufficient to block X_{t_0} in a graphical causal model of the sort that Shalizi and Thomas present.

Our current simulation does not include any such latent factors. Shalizi and Thomas have agued that homophily and contagion are generically confounded, but the magnitude of the confounding effect is not clear. For our purposes, it's also worth considering how such confounding would interact with the unfriending problem we identify. It would be straightforward to introduce a latent X_0 into our simulation to address these questions.

6.3 Realistic network features

Network formation in our simulation is random conditional on the outcome variable Y. This is not the case in real networks, which differ systematically from comparable random graphs. As a consequence, our networks (like those in CF's original model) differ from human social networks in several ways. In particular, there are few mutual friendships and clustering is very low (in real data, two people with a common friend are likely to be friends themselves).

We would like to bring the insights of the literature on network formation to our simulation. Ideally, we would generate the network using a well-known model, such as a Barabasi game, that generates graphs of the sort that we observe in empirical data. However, we are unaware of any such canonical graph model that incorporates homophily. Alternatively, we could model some of the expected features directly, adding parameters in our tie formation model for mutuality and the closing of triangles and observing how changes in those parameters affect our estimates.

7 Conclusion

In this paper, we have argued that homophily in friendship retention complicates efforts to estimate causal influence effects in longitudinal social network data. This "unfriending" problem can induce an association between random shocks to an outcome variable for those dyads that remain friends at both t_0 and t_1. Our simulations support this result, showing that bias increases and coverage decreases dramatically as unfriending homophily increases. As such, we conclude that estimates of peer effects based on CF's model are likely to confound homophily in friendship retention with true influence effects.

References

Anagnostopoulos, Aris, Ravi Kumar and Mohammad Mahdian. 2008. Influence and correlation in social networks. In *KDD08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. Bing Liu, Sunita Sarawagi and Ying Li. pp. 7–15.

Arala, Sinan, Lev Muchnika and Arun Sundararajana. 2009. "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks." *Proceedings of the National Academy of Sciences* 106(51):21544–21549.

Bramoullé, Yann, Habiba Djebbaria and Bernard Fortin. 2009. "Identification of peer effects through social networks." *Journal of Econometrics* 150(1):41–55.

Cacioppo, John T., James H. Fowler and Nicholas A. Christakis. 2009. "Alone in the crowd: The structure and spread of loneliness in a large social network." *Journal of Personality and Social Psychology* 97(6):977.

Christakis, Nicholas A. and James H. Fowler. 2007. "The spread of obesity in a large social network over 32 years." *New England Journal of Medicine* 357(4):370–379.

Christakis, Nicholas A. and James H. Fowler. 2008. "The collective dynamics of smoking in a large social network." *New England Journal of Medicine* 358(21):2249–2258.

Cohen-Cole, Ethan and Jason M. Fletcher. 2008. "Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic." *Journal of Health Economics* 27(5):1382–1387.

Degirmencioglu, Serdar M., Kathryn A. Urberg, Jerry M. Tolson and Protima Richard. 1998. "Adolescent friendship networks: Continuity and change over the school year." *Merrill-Palmer Quarterly* 44(3).

Fowler, James H., Michael T. Heaney, David W. Nickerson, John F. Padgett, and Betsy Sinclair. N.d. "Causality in Political Networks." Unpublished manuscript.

Fowler, James H. and Nicholas A. Christakis. 2008*a*. "Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study." *British Medical Journal* 337(a2338):doi:10.1136/bmj.a2338.

Fowler, James H. and Nicholas A. Christakis. 2008*b*. "Estimating peer effects on health in social networks: A response to Cohen-Cole and Fletcher and Trogdon, Nonnemaker, and Pais." *Journal of Health Economics* 27:1400–1405.

Fowler, James H. and Nicholas A. Christakis. N.d. "Estimating Peer Effects on Health in Social Networks." Unpublished version of *Journal of Health Economics* article. Downloaded from `http://jhfowler.ucsd.edu` on May 11, 2010.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau and Martina Morris. 2003. *statnet: Software tools for the Statistical Modeling of Network Data*. Seattle, WA: . Version 2.0.
URL: *http://statnetproject.org*

Kandel, Denise B. 1978. "Homophily, selection, and socialization in adolescent friendships." *American Journal of Sociology* 84(2):427–436.

Kremer, Michael and Dan Levy. 2008. "Peer effects and alcohol use among college students." *Journal of Economic Perspectives* 22(3):189–206.

Lyons, Russell. N.d. "The spread of evidence-poor medicine via flawed social-network analysis." Unpublished manuscript.

Manski, Charles F. 1993. "Identification of endogenous social effects: The reflection problem." *The Review of Economic Studies* 60(3):531–542.

McPherson, Miller, Lynn Smith-Lovin and James M Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual Review of Sociology* 27(1):415–444.

Mednick, Sara C., Nicholas A. Christakis and James H. Fowler. 2010. "The Spread of Sleep Loss Influences Drug Use in Adolescent Social Networks." *PLoS One* .

Mollenhorst, G.W. 2009. Networks in contexts : How meeting opportunities affect personal relationships PhD thesis University Utrecht.

Newcomb, Andrew F., William M. Bukowski and Catherine L. Bagwell. 1999. Knowing the Sounds: Friendship as a Developmental Context. In *Relationships as developmental contexts: The Minnesota Symposia on Child Psychology*, ed. W. Andrew Collins and Brett Laursen. Vol. 30 Lawrence Erlbaum Associates p. 72.

Resnick, Michael D., Peter S. Bearman, Robert Wm. Blum, Karl E. Bauman, Kathleen M. Harris, Jo Jones, joyce Tabor, Trish Beurhring, Renee E. Sieving, Marcia Shew, Marjorie Ireland, Linda H. Bearinger and J. Richard Udry. 1997. "Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health." *Journal of the American Medical Association* 278:823–32.

Rosenquist, J. Niels, James H. Fowler and Nicholas A. Christakis. 2010. "Social network determinants of depression." *Molecular Psychiatry* .

Rosenquist, J. Niels, Joanne Murabito, James H. Fowler and Nicholas A. Christakis. 2010. "The Spread of Alcohol Consumption Behavior in a Large Social Network." *Annals of Internal Medicine* 152(7):426–433.

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics* 116(2):681–704.

Schneider, Barry H., Ada Fonzi, Franca Tani and Giovanna Tomada. 2006. "A cross-cultural exploration of the stability of children's friendships and the predictors of their continuation." *Social Development* 6(3):322–339.

Shalizi, Cosma R. and Andrew C. Thomas. 2010. "Homophily and Contagion Are Generically Confounded in Observational Social Network Studies." Arxiv preprint arXiv:1004.4704.

Soetevent, Adriaan R. 2006. "Empirics of the identification of social interactions: An evaluation of the approaches and their results." *Journal of Economic Surveys* 20(2):193–228.

Zimmerman, David J. 2003. "Peer effects in academic outcomes: Evidence from a natural experiment." *Review of Economics and Statistics* 85(1):9–23.

Table 1: Monte Carlo simulation results

	Retention	Formation	Retention			Ego-alter	Ego-alter	Friends/	Friends/	Friendship
	constant	homophily	homophily		Coverage	correlation	correlation	person	person	retention
	β_0 (t_1)	β_1 (t_0)	β_1 (t_1)	Bias	probability	(t_0)	(t_1)	(t_0)	(t_1)	rate
1	0.750	0.000	0.0000	-0.00	0.94	-0.00	-0.00	3.0	2.3	0.77
2	0.750	0.025	0.0000	-0.00	0.94	0.38	0.28	3.5	2.7	0.77
3	0.750	0.050	0.0000	-0.00	0.94	0.62	0.44	4.9	3.8	0.77
4	0.750	0.000	0.0125	0.01	0.94	-0.00	0.03	3.0	2.3	0.77
5	0.750	0.025	0.0125	0.01	0.92	0.38	0.31	3.5	2.7	0.77
6	0.750	0.050	0.0125	0.01	0.91	0.62	0.46	4.9	3.8	0.77
7	0.750	0.000	0.0250	0.01	0.89	-0.00	0.07	3.0	2.3	0.77
8	0.750	0.025	0.0250	0.02	0.82	0.38	0.34	3.5	2.7	0.77
9	0.750	0.050	0.0250	0.02	0.74	0.62	0.48	4.9	3.8	0.77
10	0.750	0.000	0.0375	0.02	0.77	-0.00	0.11	3.0	2.3	0.76
11	0.750	0.025	0.0375	0.03	0.66	0.38	0.37	3.5	2.7	0.77
12	0.750	0.050	0.0375	0.03	0.43	0.62	0.50	4.9	3.8	0.77
13	0.750	0.000	0.0500	0.04	0.55	-0.00	0.16	3.0	2.2	0.75
14	0.750	0.025	0.0500	0.04	0.35	0.38	0.40	3.5	2.6	0.76
15	0.750	0.050	0.0500	0.05	0.15	0.62	0.52	4.9	3.8	0.76
16	0.250	0.000	0.0000	-0.00	0.94	-0.00	-0.00	3.0	1.8	0.60
17	0.250	0.025	0.0000	-0.00	0.94	0.38	0.28	3.5	2.1	0.60
18	0.250	0.050	0.0000	0.00	0.95	0.62	0.44	4.9	3.0	0.60
19	0.250	0.000	0.0125	0.01	0.92	-0.00	0.05	3.0	1.8	0.60
20	0.250	0.025	0.0125	0.01	0.89	0.38	0.32	3.5	2.1	0.60
21	0.250	0.050	0.0125	0.02	0.87	0.62	0.47	4.9	3.0	0.60
22	0.250	0.000	0.0250	0.02	0.82	-0.00	0.11	3.0	1.8	0.60
23	0.250	0.025	0.0250	0.03	0.69	0.38	0.36	3.5	2.1	0.60
24	0.250	0.050	0.0250	0.04	0.48	0.62	0.50	4.9	2.9	0.60
25	0.250	0.000	0.0375	0.04	0.64	-0.00	0.17	3.0	1.8	0.59
26	0.250	0.025	0.0375	0.05	0.38	0.38	0.41	3.5	2.1	0.60
27	0.250	0.050	0.0375	0.06	0.13	0.62	0.53	4.9	2.9	0.60
28	0.250	0.000	0.0500	0.06	0.35	-0.00	0.23	3.0	1.8	0.59
29	0.250	0.025	0.0500	0.07	0.12	0.38	0.45	3.5	2.1	0.60
30	0.250	0.050	0.0500	0.08	0.00	0.62	0.55	4.9	2.9	0.60
31	-0.250	0.000	0.0000	-0.00	0.94	-0.00	-0.00	3.0	1.2	0.40
32	-0.250	0.025	0.0000	0.00	0.95	0.38	0.29	3.5	1.4	0.40
33	-0.250	0.050	0.0000	-0.00	0.96	0.62	0.44	4.9	2.0	0.40
34	-0.250	0.000	0.0125	0.02	0.93	-0.00	0.08	3.0	1.2	0.40
35	-0.250	0.025	0.0125	0.02	0.87	0.38	0.34	3.5	1.4	0.40
36	-0.250	0.050	0.0125	0.03	0.81	0.62	0.48	4.9	2.0	0.40
37	-0.250	0.000	0.0250	0.04	0.76	-0.00	0.16	3.0	1.2	0.41
38	-0.250	0.025	0.0250	0.05	0.59	0.38	0.40	3.5	1.4	0.40
39	-0.250	0.050	0.0250	0.05	0.33	0.62	0.52	4.9	2.0	0.40
40	-0.250	0.000	0.0375	0.06	0.49	-0.00	0.24	3.0	1.2	0.41
41	-0.250	0.025	0.0375	0.08	0.19	0.38	0.45	3.5	1.4	0.41
42	-0.250	0.050	0.0375	0.08	0.03	0.62	0.56	4.9	2.0	0.40
43	-0.250	0.000	0.0500	0.08	0.20	-0.00	0.31	3.0	1.2	0.42
44	-0.250	0.025	0.0500	0.10	0.03	0.38	0.50	3.5	1.4	0.41
45	-0.250	0.050	0.0500	0.11	0.00	0.62	0.60	4.9	2.0	0.41

$\beta_0 = -2.75$ at t_0; true contagion effect $b_1 = 0$; standard deviation of $u = 5$, $n = 1000$