Is Obesity Socially Contagious?

A Senior Project

Presented to

The Faculty of the Statistics Department

California Polytechnic State University- San Luis Obispo

A Partial Fulfillment

Of the Requirements for the Degree

Bachelor of Science

by

Ciani Jean Sparks

March 2013

Table of Contents

I. Overview

"Overweight and obesity are the fifth leading risk for global deaths. At least 2.8 million adults die each year as a result of being overweight or obese"(World Heath Organization, 2013). In the last decade, obesity has become extremely prevalent in our society due to food consumption, food prices, food production, and social behaviors. This paper will investigate the possibility of obesity being passed through social networks due to selections and choices between friends. We will critique three different articles, "*The spread of Obesity in a Large Social Network over 32 Years,"* by Nicholas A. Christakis and James H. Fowler, "*Is Obesity Contagious? Social Networks vs. environmental factors in the obesity epidemic*," by Ethan Cohen-Cole and Jason M. Fletcher, and "*Homophily and Contagion are Generically Confounded in Observational Social Network Studies*," by Cosma Rohilla Shalizi and Andrew C. Thomas. Each of these articles presents interesting information and statistical analysis that provide evidence for and against the idea of obesity being socially contagious.

The Christakis and Fowler (CF) article observed clustering of obesity in a large-scale social network: that is, people tend to be friends with others of similar obesity status. This could have been caused by three mechanisms: homophily, contagion, or confounding. Homophily is when people choose friends with similar traits, contagion is when a person can influence a change of BMI in their friends, and confounding is when two people gain weight simultaneously due to an external factor. After observing the network over 32 years, they saw that individuals' weight changes tend to correlate with those of their friends, suggesting that obesity could be socially contagious. CF claimed the clustering in the network was caused by contagion. They presented a statistical model, which they claim separates influence from selection and confounding variables. The other two articles critique CF's methodology and results. By fitting the CF model to a different data set and adjusting for confounding variables, Cohen-Cole and Fletcher (CoF) claim that confounding variables explain the relationship in a social network instead of contagion. The Shalizi and Thomas (ST) article simulated a toy network with no contagion. When they fit the CF model to the network, it showed contagion was present, which proved the CF model can provide false evidence for contagion. The CoF

and ST articles compare and contrast the results they received with the CF results to observe if an individual's BMI can be affected by social factors.

The Christakis and Fowler article was one of the first publications addressing the idea of social factors contributing to the obesity epidemic. Before explaining the study, I will define terminology that will be used throughout the paper. The following definitions are given in the CF article:

- **Ego** is the person whose behavior is being analyzed.
- **Alter** is the person connected to the ego who may influence the behavior of the ego.
- **Node** is an object that may or may not be connected to other objects in a network.
- **Tie** is the connection between two nodes that can either be one-way (directed) or two-way (mutual).
- **Homophily** is the tendency for people to choose relationships with people who have similar attributes.
- **Contagion** is the spread of behavior or trait from one person to another.
- **Cluster** is a group of nodes, each of which is connected to at least one other node in the group.

Christakis and Fowler looked at three mechanisms that could explain the clustering of obese persons in the network. First is homophily, which is the idea that egos become friends with alters that have the same characteristics as themselves. Second is confounding, which is when an ego and an alter simultaneously gain weight due to a common attribute or external factor. Lastly is contagion, where alters may influence egos, increasing the ego's likelihood to gain weight.

II. How Obesity Could Become Clustered in a Network

The data used in the CF article was from the Framingham Heart study, which started in 1971, enrolling a random sample of 2.3 of the residents of Framingham, MA. The original cohort included 5,209 people in the same location who gave repeated measurements of their height, weight, and written questionnaires. Over 32 years the offspring of members were consistently enrolled in the study, and by 2002, the third generation of offspring was involved. CF analyzed the second cohort of 5,124 subjects

and included anyone linked to these respondents in any way, resulting in a network with 12,067 people. All of the subjects were given tracking sheets, which provided complete information about their parents, spouses, siblings, children, and at least one "close friend". Since the study included a large percentage of Framingham, MA residents, many nominated friends were themselves respondents in the survey. Two friendships were possible, mutual or directed. Mutual friendships are when the ego and alter both nominate each other as close friends. A directed friendship is when an ego nominates an alter as a friend and the alter does not nominate the ego as a friend, or vice versa. This information allowed CF to examine different ways obesity could be spread including: the presence of clusters of obese people in a network, the relationship between one person's weight gain and their friends/family weight gain, the dependence of this relationship on the type of social tie, and the influence of gender, smoking behavior, and geographic distance between people in the network.

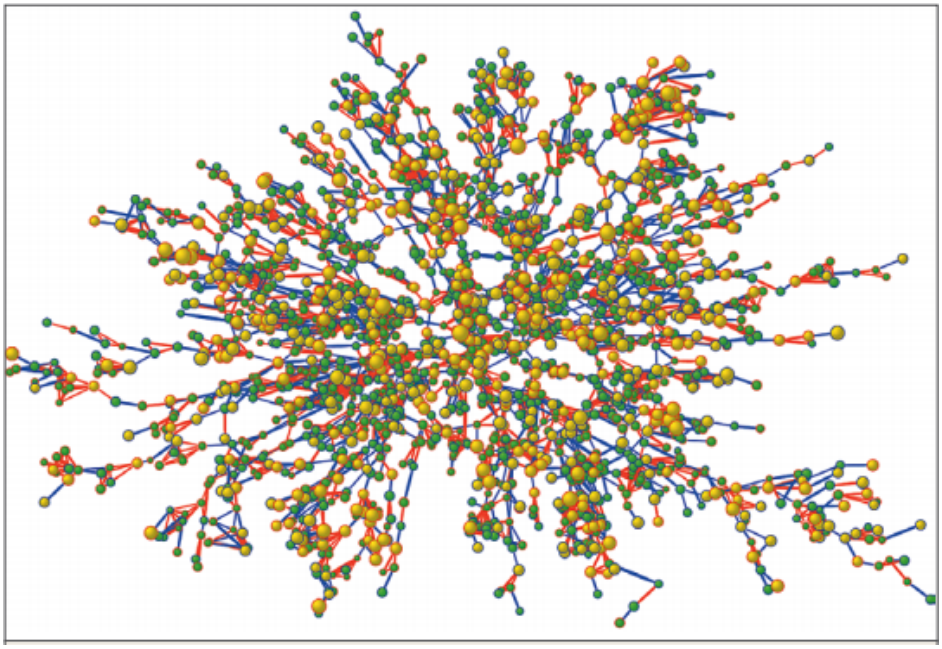

Figure 1: This is the largest connected subcomponent of the social network in the Framingham Heart Study in the year 2000. Each circle (node) represents one person in the network, women have red borders and men have blue. The size of the node is proportional to each person's body mass index with yellow signifying obese and green not obese. This figure was given in the Christakis and Fowler article.

The authors created a logistic regression model to examine the possible relationship between obesity and social networks, where the response variable was the

ego's obesity status, defined as BMI greater than 30 at the present time, t+1. BMI stands for Body Mass Index, which is calculated by dividing the weight of the ego in kilograms by the square of the height in meters. The model is a function based on the ego's BMI at the previous time, t, the ego's age, sex, and education, and the alter's BMI at time t and t+1.

$$BMI^{ego,t+1} = \beta_0 + \beta_1 * age^{ego} + \beta_2 * sex^{ego} + \beta_3 * educ^{ego} + \beta_4 * BMI^{ego,t} + \beta_5 * BMI^{alter,t} + \beta_6 * BMI^{alter,t+1}$$

Christakis and Fowler evaluated different aspects to see if the clustering of obesity was related to various potential confounding variables. Smoking has always been negatively linked to obesity, therefore the smoking status of egos and alters at times t and t+1 were added to the model to see if the spread in smoking-cessation behavior contributed to the spread of obesity. A variable representing geographical distance was also included in the model to observe whether living near an obese alter played a role in the ego becoming obese. After fitting the models, the authors drew multiple conclusions about how obesity can be clustered throughout a network. First, social distance was more important than geographical distance within the networks. This suggested the spread of behaviors associated with obesity depend on the nature of social ties. In other words, an increase in social distance decreases the effect an alter has on an ego, where as an increase in geographical distance did not change the effect of an alter on an ego. CF also concluded any spread of smoking behavior in a network was not significant in the clustering of obesity. Some confounding variables that were not included in the data were genetics, economic status, and environment, which may have produced false conclusions about how obesity is clustered in a network.

Christakis and Fowler then fit a linear regression model where the outcome was the obesity status of the ego, 1 signifying obese, and 0 not obese. They fit the model with two different types of friendship networks, ego-perceived (Ego sees Alter as friend only) and alter-perceived (Alter sees Ego as friend only) to see if the direction of the relationship affected ego's obesity status. If obesity is not socially contagious, then the coefficients for the ego-perceived and alter-perceived networks should be equal. This is because the confounders will affect the ego and alter simultaneously and equally, so the trend should not depend on the direction of the friendship. Therefore, the estimates of the

coefficients from the two networks should be similar. After fitting the model, Christakis and Fowler concluded the ego's chances for becoming obese increased by 57% if the alter became obese in an ego-perceived friendship, but obesity of alter-perceived friendships was not significantly related to the ego's risk of obesity (Christakis, 2007). Christakis and Fowler argue that a difference in coefficients gives evidence for contagion. Their premise is that people look up to their friends and want to be like them, so obesity in alters can cause the ego's perception about obesity to be accepted. This could directly influence the ego's behaviors. They expect that friends nominated by the ego are more likely to be esteemed, so the direction of the friendship is inversely related to the direction in transmission of obesity. This is referred to as the asymmetry argument.

Although this is a clever argument, Christakis and Fowler left out some key details to solidify their conclusion. They argued that the difference in coefficients is evidence for contagion, but failed to point out the overlap in confidence intervals. This overlap signifies the difference in coefficients is not statistically significant. This weakened their argument that obesity is contagious. With CF assuming the direction of transmission, equal coefficients would suggest alters didn't influence egos to become obese for ego-perceived friendships and vice versa. Hence, an alter who became obese would have no influence on whether the ego became obese. If friends did become obese at the same time, then it could have been caused by an external factor. This suggests there may be a different mechanism besides contagion that can cause obesity to become clustered in a social network. Christakis and Fowler presented an interesting theory for the transmission of obesity, but lacked support in their conclusion. We will highlight additional weaknesses of their analyses in our discussion of articles that built upon the CF methods.

III. Environmental Variables Could Partially Explain Clustering

The Cohen-Cole and Fletcher (CoF) article expanded CF's model by including environmental factors and also fit the model to a different data set to see if the coefficients changed. The authors used the Add Health data set, which has several advantages over the Framingham study. First, it was a nation wide sample of 7^{th}-12^{th} graders in 1994/95, which is advantageous because it is a larger population with a smaller

age range instead of a smaller population with a wider age range. The smaller age range decreased the variability, so if there was a relationship between the variables, it would be easier to detect. There were about 5,000 adolescents involved, with almost 2,000 whom were followed over time along with one same sex friend. The second advantage was that the setting was restricted to high schools, which may be more social than the lives of people in the Framingham study. This is advantageous because since the study was limited to high schools, and high school specific variables were collected, researchers were able to account for changes in social context experienced by all. The last advantage is CoF's results are more generalizable because the sample is representative of the entire country rather than limited to Framingham, Mass. The two similarities between the data sets include the time between interviews (about three years) and the desired information from each participant. However, CoF's data set includes potential confounders not available in the Framingham study.

The CoF article addresses more confounding variables in their model that were not included in the CF model. First, the authors discussed how genetic variability could not be the reason for the rapid increase in obesity because it happened in such a short period of time. This suggests there is a different reason for the spike in obesity across the nation. Second, CoF included a set of variables representing the environmental factors, $c_{c,t+1}$, which could affect the coefficients that CF obtained. Below is the CF model.

$$BMI^{ego,t+1} = \beta_0 + \beta_1 * age^{ego} + \beta_2 * sex^{ego} + \beta_3 * educ^{ego} + \beta_4 * BMI^{ego,t} + \beta_5 * BMI^{alter,t} + \beta_6 * BMI^{alter,t+1} + \varepsilon^{ego,t+1},$$

Cohen-Cole and Fletcher point out three features to the model that could impact results. First, if the environmental variables, denoted $c_{c,t+1}$, are positively related to an individual's BMI, then the unadjusted estimates for network effects will be too large, thus show more of a network effect than is indeed present. Excluding these environmental confounders can lead to contagion appearing to cause the transmission of obesity through a social network when it actually does not. Second, incorrect conclusions can be drawn if an individual's error term is positively correlated with their friend's BMI. In other words, if two people become friends because of a common unobserved trait and that trait can influence both BMI's, then the change in BMI can appear as contagion. Christakis and Fowler claim to account for this by using an independent variable for the alter's weight

status at two sequential time points, which controlled for homophily. Cohen-Cole and Fletcher propose this statement is false unless homophily is assumed only on this variable. Lastly, CoF claim the use of a lagged independent variable in a social network can lead to biases in estimation of the coefficients.

After acknowledging the flaws in the CF model, Cohen-Cole and Fletcher replicated the model using the Add Health data set. CoF obtained very similar results to Christakis and Fowler's. They estimated that the odds of an ego becoming obese increased by 80% if the alter was obese for an ego-perceived friendship (Cohen-Cole, 2008). After CoF added a set of environmental confounders, $c_{c,t+1}$, to the model, the odds decreased to 50%. Thus, the environmental variables partially explained the association found by CF.

The model was extended further by modeling BMI as a continuous variable. The CF model was first fit to the Add Health data set to produce a statistically significant coefficient of .054. This means that a one-unit increase in the alter's BMI is associated with a .054 unit increase in the ego's BMI. A school trend variable was then added to the model, which accounted for environmental factors shared by the students at the same school. When including this variable, the coefficient for alter's BMI was reduced to .037 and was no longer statistically significant. This meant the alter's BMI was not a significant variable in predicting the ego's BMI. Finally, individual fixed effects were included, which accounted for each person's own variability. The coefficient was decreased even more to .033 and was still not statistically significant. This led CoF to conclude that the clustering of obesity was related to the environment where an individual lived instead of contagion. Also, CoF concluded that omitted group-level characteristics could cause correlated body weight in peer groups. In the CF article, homophily was adjusted for by using time-lagged dependent variables with the alter being obese at the previous time point. This made it difficult to decipher if weight gain was caused by homophily or contagion. CoF controlled for homophily by measuring the ego's BMI at the start of the friendship and observed how it changed at the successive weight measurement. This allowed two people becoming friends based on their weight similarity to be distinguished from the friendship effect on weight gain.

An important point Cohen-Cole and Fletcher mention is the distinction between endogenous effects and contextual effects. Endogenous effects represent the probability of becoming obese because of the direct interaction with another individual, while contextual effects represent shared surroundings of a group that lead to similar weight outcomes. They stated that unless you know the individual's characteristics, preferences, choices, and environment, it is difficult to distinguish if weight gain is due to behavioral influences by friends, or due to changes in environmental factors. By analyzing the Add Health data, CoF included many contextual effects, measuring environmental influences in middle schools and high schools throughout the nation. This helped identify how obesity could become clustered in a social network, because all students in the study will be in the same environment.

After reading both of these articles, there are two critiques I would make about the analysis of the Cohen-Cole and Fletcher article. Although it is necessary to account for environmental factors, 7-12th graders are more likely to fluctuate in weight than adults because of puberty, peer pressure, school cafeteria food etc. This could have created more variability in the data, making it difficult to find an association between ego's BMI at the current time point and alter's BMI at the previous time point due to social factors. Also, one critique Cohen-Cole and Fletcher made about the Christakis and Fowler article was the use of lagged variables, even though they used lagged variables themselves. CoF claim it can lead to dynamic models producing coefficients with large degrees of bias, which suggests the coefficients they estimated may be biased as well. The ideas presented so far have been interesting, however the last article will show why this topic is controversial and why it is difficult to draw cause and effect conclusions from observational data.

IV. Confounded Mechanisms in Observational Data

The third article I reviewed, written by Shalizi and Thomas (ST), provides evidence to claim that homophily and contagion cannot be distinguished in analysis of observational data. The authors identify six different mechanisms that can appear like homophily or contagion in a network: biological influence, manifest homophily, latent homophily, social contagion, secondary homophily, and common external causation.

Secondary homophily is when two people become friends due to an unobserved trait, and that trait causes the friends to simultaneously gain weight. External causation is the same outer entity causing two people to gain weight at the same time. The main point of this article is to prove that it cannot be determined which one of these mechanisms causes clustering of obesity statuses in an observational data set. ST focus on two of these mechanisms: social contagion and latent homophily. An example of these two mechanisms would be; a gym membership is not in the study, therefore it is unobserved. This trait sparks a friendship between the two subjects of interest and also affects their BMI. Social contagion would be when the change in BMI of one friend affects the BMI of the other friend by directly influencing their habits. Latent homophily would be when belonging to the same gym causes both of their BMI's to change simultaneously. ST claim it is impossible to distinguish between these two mechanisms in observational data because the trait is unobserved, and the two variables of interest are linked by a pathway that includes unobserved traits. ST conducted two different statistical simulations to prove their claim. In the model, the unobserved trait is not time dependent, but they state it does not require much work to see that it applies to time dependent variables as well.

Shalizi and Thomas thought the Christakis and Fowler theory for friendship and obesity being inversely related (asymmetry argument) was clever, but they showed that is can fail under certain conditions. In this article, X represents an unobserved characteristic, Z represents an observed characteristic, Y represents obesity, and i/j represent individuals. ST showed that the CF argument breaks if two conditions are met. If people who influence obesity have different values about the unobserved trait, X, than people who are influenced then contagion will be falsely present in the network, and if there is a non-linear relationship between X and Y, then contagion will be falsely present in the network. To illustrate this claim, ST simulated a toy network with no contagion and demonstrated that applying the asymmetry argument to the network will show contagion exists when there is none. When Shalizi and Thomas fit CF's model to their network it showed evidence of contagion even though there was none. They replicated the network 5,000 times at multiple time points to estimate the coefficients. There were two coefficients of interest in the model, $ß_2$ and $ß_3$. If there are two individuals A and B, $ß_2$ would represent the effect of individual B's BMI at the previous time point on

individual A's BMI at the present time point when A nominates B as a friend. $ß_3$ would represent the opposite or "reverse" effect, i.e. the effect of individual B's BMI on individual A's BMI when B nominates A as a friend. Since there was no contagion included in the network, $ß_2$ and $ß_3$ should be equal to each other because the direction of the friendship doesn't matter. After performing the trials, $ß_2$'s mean was .0257 and $ß_3$'s mean was .00192. Also, the difference in the means was .0237. This difference is greater than zero in about 77% of the simulations, signifying the CF model was flawed, and the asymmetry argument can give false evidence of contagion being present in a data set. The results from the simulation are below.

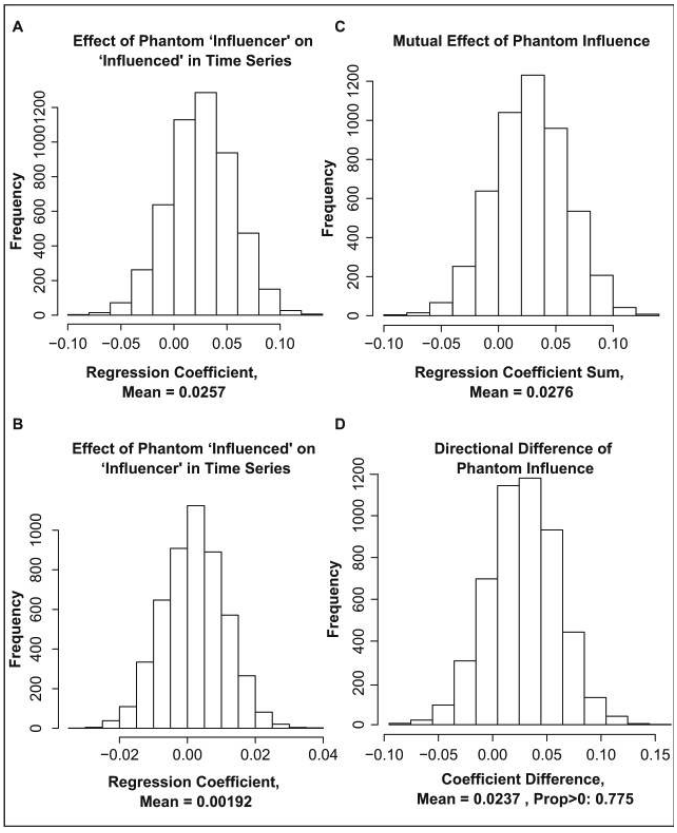

Figure 2: Results for the toy model where contagion is falsely present in a network. Note: Graph A is the estimate for $ß_2$, Graph B is the estimate for $ß_3$, Graph C is the estimate for the mutual effect, $ß_2 + ß_3$, and Graph D is the estimate for the difference, $ß_2 - ß_3$. This figure was given in the Shalizi and Thomas paper.

Next Shalizi and Thomas demonstrated that homophily and contagion combined can give a false appearance of causation in a social network. They created a different simulation model where an observed trait, unrelated to BMI, influences a friendship formation and the friends influence each other's change in BMI. Their claim is that a

dynamic network with these mechanisms can give the appearance of the trait causing a change in BMI, leading analysts to draw false conclusions from the data. First, ST simulated values for a trait, Z, for all nodes in a network. Next, they created homophily on Z by generating edges with a preference for others with similar values of Z. Then they simulated values for a cultural choice, Y, being independent of the trait, Z. The homophilous network had an association between one of the choices, Y, and the social trait, Z. They contrast this with a neutral network, which was generated by simulating edges without regard to Z values. This created a network with no homophily on Z. They performed the simulation 3,000 times in the homophilous and neutral networks, with increasing time points to observe how the trait would diffuse through the cluster over time.

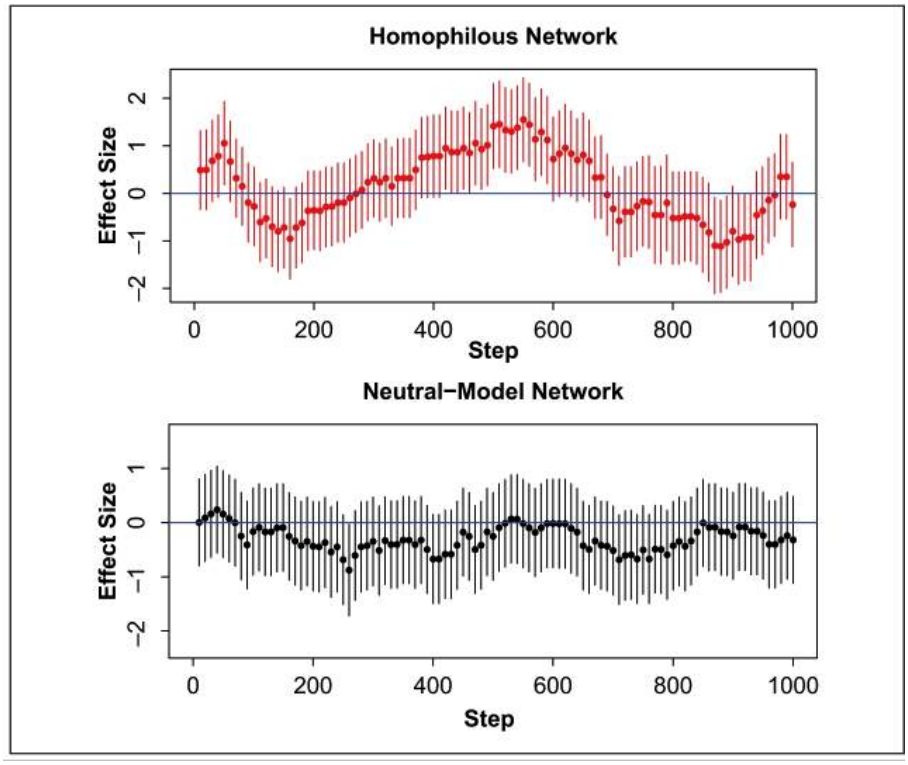

Figure 3: Results for the coefficieint estimates for logistic regression of choice on trait over time with 95% confident interval error bars. This figure was given in the Shalizi and Thomas article.

As you can see in Figure 3, the homophilous network coefficient becomes negative and statistically significant below zero and then returns to being positive and statistically significant. This signified an association between the trait, Z, and choice, Y.

It appears that the trait caused the change in Y since the coefficient is significantly different than zero, but it didn't. In the neutral network, the coefficients never become significantly different from zero, meaning there is no association between the trait and Y. This simulation study demonstrated that homophily and contagion can manifest similarity in a longitudinal social network, hence they can't be distinguished in observational data. Shalizi and Thomas conclude the only way to decipher homophily from contagion is either strong parametric assumptions about the individual's choices and characteristics, or including all possible confounding variables that influence BMI and the friendship between two people.

Shalizi and Thomas disproved the results found in the Christakis and Fowler article by simulating two of the three problems that arise in observational data. First, latent homophily appears to show contagion of obesity. For example, if two people become friends because they work at the same office building and the office influences their BMI, then it could appear that one person is affecting the other's BMI, when that is not the case. This was shown in Figure 2, where the network effect coefficient and reverse network effect coefficient were different in the simulation when contagion was excluded. Second, homophily on an observed characteristic and contagion of obesity can give the appearance of the characteristic causing obesity. Using the same scenario, if one individual started influencing the other to eat fast food every day, then it could appear that working at the same office was causing one person to gain weight, when the true cause is social contagion. This was shown in Figure 3, where the homophilous network had statistically significant coefficients and neutral network did not. Lastly, contagion of obesity and an observed characteristic causing obesity can appear to look like homophily. In other words, if two friends influence obesity, and they are friends because they work in the same office building, then this could look like a preference for friendship because of the similar obesity statuses. The simulation to show this claim was not demonstrated in the Shalizi and Thomas paper.

All three of these articles studied a health problem that has been tremendously increasing in our society in the last decade. The cause of the dramatic increase in numbers is still unclear, and the studies described here highlight why. Obesity is a difficult health risk to study because we rely on observational studies to gather data,

which makes it challenging to infer causality. Christakis and Fowler presented a model on how to observe the clustering of obesity in a social network. Their claim that contagion caused the clustering of obesity in a network sparked a debate that is ongoing today. Cohen-Cole and Fletcher adjusted CF's model, as well as included more confounding variables. Expanding the model lead to the conclusion that environmental factors explain some of the clustering of obesity in a network as opposed to contagion. Shalizi and Thomas presented simulations that showed contagion and homophily manifest similarly in social networks, meaning they are indistinguishable in observational data. These articles encourage this topic to continue to be studied, because it can create a variety of different dangerous health risks. The next section of my paper reproduces the statistical simulation performed in the Shalizi and Thomas article as well as changes the parameter values to try and quantify the bias produced in the Christakis and Fowler article.

V. Shalizi and Thomas Simulation Study

The Shalizi and Thomas article provided computer code to show how they created the toy networks used in their simulations. The first network created illustrated the ST claim that when the asymmetry argument is applied to a model, it provides false evidence of contagion in a network. First, ST created an undirected network consisting of 400 nodes. Then, a random number between 0 and 1 was generated for each node in the network to represent a latent attribute, X_i. They computed the difference between the X_i's for each pair of nodes in the network. They generated a friendship network with homophily on the latent trait as follows. They set the probability that individuals i and j are friends equal to inverse logit$(-3|X_i - X_j|)$, so people with similar traits will be more likely to become friends. They used these probabilities to generate random binomial numbers (0 or 1) to create the friendship matrix. This is a square matrix, A, with the number of rows equal to the number of people in the network. A_{ij} equals 1 if i and j are friends and A_{ij} equals 0 if not. A is symmetric since friendship is undirected. Each individual, i, then nominates one friend, j, with the probability equal to the inverse logit $(-|X_j – 0.5|)$. This replicates the sampling process in the Framingham Study analyzed by CF, in which each respondent nominated a single friend in the survey. ST picked 0.5 to

implement a preference for friends that are closer to the median value of that trait. Next, ST established time trends of the observable outcomes based on the latent variable at three time points. The equations are defined in the ST article as follows:

- $t=0,\ Y_i(0) = (X_i - 0.5)^3 + N(0,(0.02)^2)$, a nonlinear relationship between X and Y.
- $t=1,\ Y_i(1) = (X_i - 0.5)^3 + 0.4X_i + N(0,(0.02)^2)$
- $t=2,\ Y_i(2) = Y_i(1) + 0.4X_i + N(0,(0.02)^2)$

Shalizi and Thomas include the $0.4X_i$ term to model a greater increase in BMI for individuals with higher latent attribute values. For example, if X is the number of times a person eats fast food every week, then the BMI of the individual will increase at a faster rate when the number of times fast food was eaten increases. ST then simulated an undirected network with 400 nodes, and estimated the following linear model:

$$Y_i(2) = \alpha + \beta_1 Y_i(1) + \beta_2 \Sigma A_{ij} Y_j(1) + \beta_3 \Sigma A_{ji} Y_j(1) + \beta_4 \Sigma A_{ij} Y_j(0) + \beta_5 \Sigma A_{ji} Y_j(0) + \varepsilon_j,$$

This model was slightly different than what was used in the CF article. The CF article used a logistic regression model where the response variable is the odds of obesity for the ego at time t+1 based on age, sex, education, and the alter's obesity status at time t and t+1. The ST model is a linear regression equation that predicts the ego's obesity status at time 2 based on the directed friendship network effects. β_2 and β_4 estimate the network effect that the alter nominated has on the ego nominee at time 1 and time 0, while β_3 and β_5 estimate the network effect that the alter nominee has on the ego nominated at time 1 and time 0. ST claim that false evidence of contagion will be present when the influencers have higher values for the latent variable than the influenced and there is a non-linear relationship between the latent variable and obesity status. CF claim asymmetric friendships caused the clustering of obese people in the network. The asymmetry may be due to contagion being transmitted along the asymmetric edge. Therefore, the asymmetry argument would still hold if CF used a linear model as opposed to a logistic regression model. Assuming the direction of friendship is inversely related to the direction of transmission in obesity, the type of model used to show this relationship should not matter. Hence, it is okay that ST fit a different model than CF because it is just one example showing the asymmetry argument provides false evidence of contagion.

They repeated the simulation 5,000 times: each time simulating a new network, fitting this model, and saving the estimated coefficients. This simulation was used to

show that when ST applied CF's asymmetry argument to a network that contained no contagion, there was evidence of contagion. The coefficients of interest were β_2 and β_3, whose means were greater than 0. β_2 represents the effect of the alter's obesity status at time 1 on the ego's obesity status at time 2 when ego nominates alter as a friend. β_3 is the "reverse," or effect of the alter's obesity status at time 1 on ego's obesity status at time 2 when alter nominates ego as a friend. After the simulations, the average β_2 was 0.0237, meaning a one unit increase in the alter's BMI is associated with a 0.0237 increase in the ego's BMI when the ego nominates the alter as a friend. The average β_3 was 0.00192, meaning a one unit increase in the alter's BMI is associated with a 0.00192 increase in the ego's BMI when the alter nominates the ego as a friend. Since both of the means are greater than zero this signifies that the ego's BMI will increase as the alter's BMI increases for both types of friendships. Of key scientific interest is the difference between the two coefficients. In a network with no contagion, we expect the difference to be zero about 50 percent of the time, since the relationship between alter BMI and ego BMI is the same regardless of the direction of friendship. This would signify equal probability for the β's to be different, meaning there is no influence in the network. However, the mean difference over 5,000 simulations was greater than 0 and occurred in almost 77 percent of the simulations. This was an interesting result and brought up the question of how ST obtained results where β_2 was more likely to be greater than β_3. This proved the CF results were flawed and their claim that contagion caused clustering in a network was less reliable.

VI. Critiques to the Shalizi and Thomas Simulation

There were some interesting aspects about the ST code that I investigated in depth. One of the first things I noticed about the equations was the error term. The low standard deviation in the error term causes the latent variable to be highly correlated with obesity. The graph below shows the high correlation between the latent variable and the outcome variable at time point 0, $Y_i(0)$, and time point 1, $Y_i(1)$.

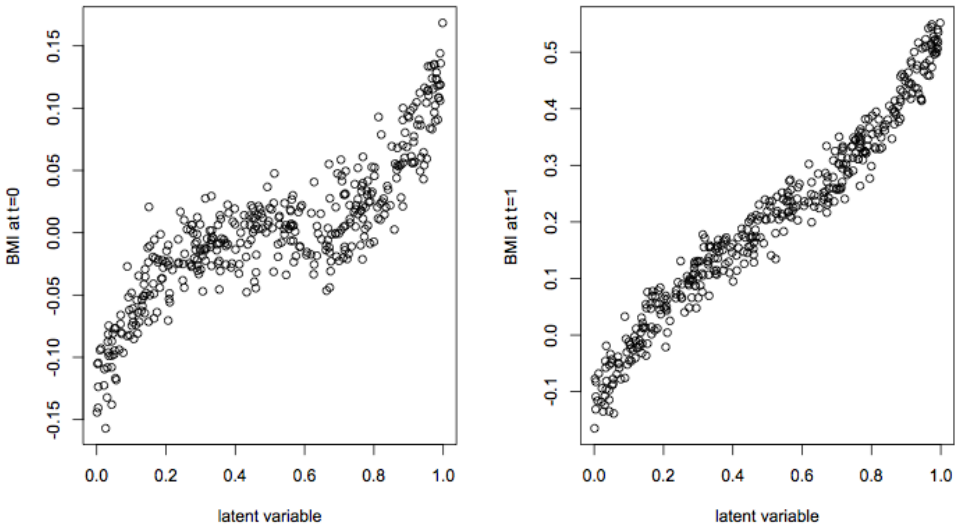

Figure 4: Simulated values of the latent variable and BMI at time points 0 and 1. At time 0 the correlation is 0.857 and at time 1 the correlation is 0.979.

Since weight is influenced by a large number of factors, we question whether it is likely that a single latent variable has such a strong correlation with BMI. This correlation between variables could be one of the reasons why the average difference in β_2 and β_3 is greater than zero in more than 50 percent of the Shalizi and Thomas simulations. To explore this, I created a negative correlation between the latent variable and BMI and observed that the coefficients were estimated the same distance below zero, and the difference between the coefficients was negative in about 60% of the simulations. This signified that the Shalizi and Thomas coefficients were greater than, rather than less than zero because the correlation between the latent variable and BMI was positive.

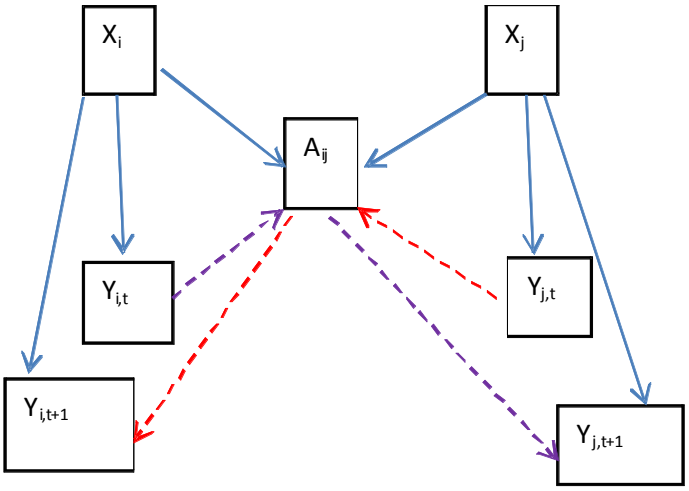

Figure 5: This causal diagram represents the key aspects of Shalizi and Thomas' toy model, required to demonstrate why this particular model shows contagion when no contagion is present. The blue arrows represent causal pathways from the latent variable to the friendship, and the latent variable to obesity status at time t and t+1. The red dashed lines represent β_2 and the purple dashed lines represent β_3.

The diagram above helps explain why the difference in coefficients was greater than zero in the ST simulation. Suppose the "influencers" are those on the right hand side. The β_2 coefficient will always be greater than β_3 because ST's model assumes the "influencers" have higher values for X, resulting in higher coefficients for $X_j \rightarrow Y_{j,t}$ and $X_j \rightarrow Y_{j,t+1}$ than for $X_i \rightarrow Y_{i,t}$ and $X_i \rightarrow Y_{i,t+1}$. The same is true if the correlation is negative, which is what I discovered when I simulated a network with a negative correlation between the latent variable and Y. It is important to notice that the coefficients, β_2 and β_3, were both greater than zero in ST's results. This shows that there is evidence of contagion in the network, however the coefficients should be equal because the direction of friendship shouldn't matter. Since the coefficients are greater than zero, a positive correlation between an unobserved variable and BMI could indicate that obesity may be socially contagious. ST disproved the results in the CF article because the difference in coefficients was greater than zero, but the fact that both of the coefficients were greater than zero signifies contagion may cause clustering of obesity in a network.

I also explored whether a similar result could be found in a network with a smaller correlation between the latent variable and the response variable. By increasing

the error term standard deviation to 0.20 instead of 0.02, the correlation was decreased at time point 0 and time point 1.

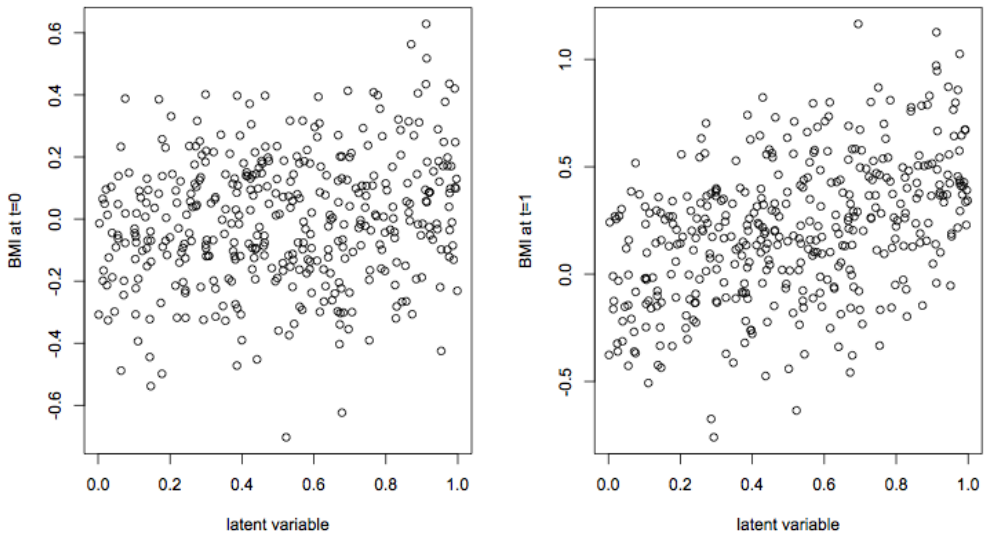

Figure 6: Simulated values of the latent variable and BMI after error standard deviation was changed to 0.20. For time 0 the correlation is 0.178 and for time 1 the correlation is 0.441.

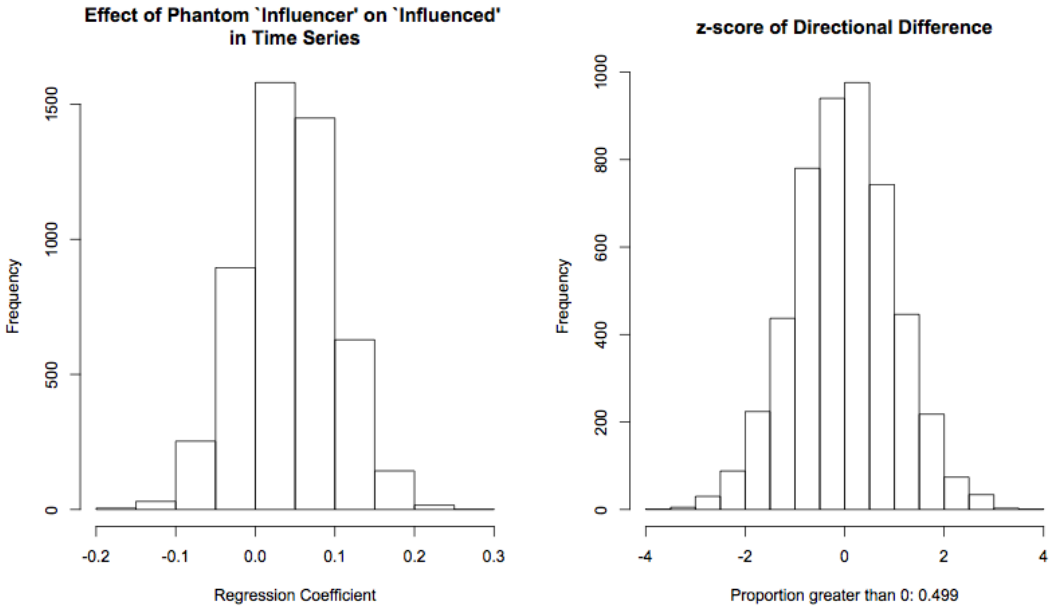

Figure 7: When the correlation between the latent variable and BMI are decreased, the average β_2 is still greater than zero (left), however the difference between the coefficients is centered at 0.50 (right). This signifies the asymmetry argument may not fail in a network with this correlation.

As you can see above, after performing the simulation 5,000 times with the decreased correlation, the mean difference in coefficients was greater than zero for 50% of the simulations. This means that Christakis and Fowler's asymmetry argument does not fail for a network with this correlation between the latent variable and obesity and also, an ego's BMI increases as the alter's BMI increases. Since the coefficients are still both greater than zero, contagion is still present in the network when the latent variable has a lower correlation with BMI. Again, this signifies obesity may be socially contagious.

I would like to further explore the conditions under which the asymmetry argument breaks down and whether real networks are likely to satisfy these conditions. An unobserved trait that I think would differ between people who influence others vs. people who are influenced would be stress. People who influence others are more likely to hold power or be in a leadership position, which will create more stress to perform well and maintain a good image. They will also be highly esteemed by others, so more likely to "transmit" their habits or values affecting BMI via social imitation. People who are influenced will not have as much pressure, therefore, could be less stressed. Stress could be measured by the ranking of your profession. I think a higher occupational status could involve higher stress levels because there is more responsibility and competition. I could create a network that substituted those values for the randomly generated values created in the Shalizi and Thomas code. Then fitting the model to the network and generating coefficients would give us more concrete evidence as to if the asymmetry argument produces false indication of contagion in a realistic network. As future work, I propose collecting this data and performing the simulation. This project could be illuminating because Shalizi and Thomas don't go into great detail about whether a real network would satisfy the conditions of their toy network. This is just one example that didn't meet the ST conditions about when the asymmetry argument would break down. There may be other latent variables that do satisfy the conditions stated in the ST article. Potential future work could include exploration of all other potential latent variables, stress just being one example, and investigating their correlations with BMI.

The transmission of obesity in social networks is a very controversial topic, and it is difficult to draw cause and effect conclusions because the data is observational, with a large number of potential confounding variables. Shalizi and Thomas did present interesting evidence to prove the Christakis and Fowler model was flawed, but their simulation results are based on a network with specific parameters, which may be unrealistic.

VII. Conclusion

In conclusion, these three articles bring up good points regarding statistical methods to give evidence for contagion in a social network. The CF article used the Framingham data set to find clustering of obesity in a network and presented a statistical model, which they claim demonstrates the clustering was caused by contagion. One premise for their conclusion was the idea that people look up to their friends, and esteem to be like them, which could change the person's behaviors that affect their BMI. This is one premise to the asymmetry argument, which sparked much controversy. The CoF article used a different data set, Add Health, and fit the CF model to the network. CoF obtained similar results to CF, but when they added environmental variables and individual effects into the model, the coefficient for the alter's effect on the ego's obesity status became insignificant. This led CoF to conclude that environmental confounders partially explained the clustering found in the CF article. I think CoF did have a good argument for the CF article. Since the environment was common for all subjects in the data across the nation, the changes within high schools and middle schools were accounted for. This showed that the CF findings could only be applied to people who lived in Framingham, MA. The CoF results indicated that the environment you live in could cause obesity to become clustered in a network. I believe this is true, because if you live in a lower income community, people are less focused on staying healthy and more focused on providing enough for their family to survive. The image of obesity would be much different in a community like this compared to a community where wealth is more prevalent. Therefore, I think the environment does have an affect on whether clustering of obesity occurs.

Shalizi and Thomas performed two simulations to show that homophily and contagion appear to have the same effect on a network, so cannot be distinguished in observational data sets. One simulation showed that homophily and contagion combined can give false appearance of causation in a network. ST simulated a toy network and showed that in a homophilous network the coefficient becomes significantly different from zero over time, indicating causation of obesity. The other simulation disproved the asymmetry argument. ST simulated a network that contained no contagion, and when they fit the CF model to the network it showed that contagion existed. When I looked at the simulated network they had created, I found that the correlation between the latent variable and BMI was very strong and positive. This caused the coefficient estimates to be greater than, rather than less than zero, and could have also caused their results to be unreliable. The high correlation also caused the difference in coefficients to be greater than zero in more than 50 percent of the simulations. In both of the simulated networks with higher and lower correlation between the latent variable and BMI, the results showed coefficients that were greater than zero. This signified that contagion was present, and shows obesity may be socially contagious. Since I found parameters where the difference in coefficients was centered at 0, the asymmetry argument could hold for a network satisfying these conditions.

Shalizi and Thomas's simulations showed that homophily and contagion appear to have the same effect on a network. Since the data are observational, it is impossible to determine which one caused the clustering of obesity. I think that the transmission of obesity could be inversely related to the direction of friendship, so if a real network was used instead of a simulated network, then there would be concrete evidence as to whether the asymmetry truly holds or not. Different latent variables will have different correlations with BMI, so some latent variables could have a stronger effect on BMI than others, which is what needs to be researched further in order to have a better understanding as to why obesity has become so prevalent in our society today. This disease remains controversial and is difficult to study because we rely on observational data sets. However, obesity has dramatically increased in our society and deserves attention because it can lead to other dangerous conditions.

VIII. Bibliography

"10 Facts on Obesity." *WHO*. World Health Organization, Mar. 2013. Web. 20 Dec. 2012. <http://www.who.int/features/factfiles/obesity/en/>.

Christakis, Nicholas A., and James H. Fowler. "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine* 357.4 (2007): 370-79. Print.

Cohen-Cole, Ethan, and Jason M. Fletcher. "Is Obesity Contagious? Social Networks vs. Environmental Factors in the Obesity Epidemic." *Journal of Health Economics* 27.5 (2008): 1382-387. Print.

Shalizi, Cosma R., and Andrew C. Thomas. "Homophily and Contagion Are Generically Confounded in Observational Social Network Studies." *Sociological Methods & Research* 40.2 (2011): 211-39. Print.

IX. Reference Codes

1) Asymmetry Argument

```
asymmetry.sim <- function (num.nodes=400, scale=3, offset=0,y.noise=0.02,
friend.noms=1,response="linear", nominate.by="distance",time.trend=0.4)

 invlogit <- function(cc) exp(cc)/(1+exp(cc))
 xx <- runif(num.nodes)
 distances <- as.matrix(dist(xx, diag=TRUE, upper=TRUE))

 adj.probs <- array(rbinom(length(distances), 1, invlogit(offset-scale*distances)), dim(distances))
 diag(adj.probs) <- 0
 adj.probs[lower.tri(adj.probs)] <- t(adj.probs)[lower.tri(adj.probs)]

 fixed.x.dist <- distances
 diag(fixed.x.dist) <- Inf
 nominees <- t(rbind(sapply(1:num.nodes, FUN=function(ii) {
  possibles <- which(adj.probs[ii,] == 1)
  return(ifelse(rep(nominate.by=="distance",friend.noms),
            possibles[which(order(fixed.x.dist[ii,possibles])<=friend.noms)],sample(possibles,
size=friend.noms, prob=invlogit(-abs((xx[possibles] - 0.5))))
            ))
 })))

 aa.mat <- array(0, dim(distances))
 for (ii in 1:num.nodes) aa.mat[ii, nominees[ii,]] <- 1
 rev.aa.mat <- t(aa.mat)
 #reciprocated matrix.
 #recip.aa.mat <- aa.mat*rev.aa.mat

 y1 = (xx-0.5)^3+rnorm(num.nodes,0,y.noise)
 y2 = y1+rnorm(num.nodes,time.trend*xx,y.noise)

 infl.y1 <- aa.mat%*%y1
 back.y1 <- rev.aa.mat%*%y1
 #recip.y1 <- recip.aa.mat%*%y1

 XX <- cbind(1, y1, infl.y1, back.y1)

 trial <- lm(y2 ~ y1 + infl.y1 + back.y1)
 coef.table <- summary(trial)$coefficients[,1]
 v.plus.c <- c(diag(summary(trial)$cov.unscaled),
summary(trial)$cov.unscaled[3,4])*summary(trial)$sigma^2
z.stat <- (coef.table[3]-coef.table[4])/sqrt(v.plus.c[3]+v.plus.c[4]+2*v.plus.c[5])

 out <- cbind(c(coef.table, v.plus.c, z.stat))
 rownames(out) <- c("int.b", "auto.b", "infl.b", "back.b",
                "int.s", "auto.s", "infl.s", "back.s",
                "cov.infl.back",
                "z.stat")
 return(out)
}
result <- replicate(5000, asymmetry.sim(friend=1, time.trend=0.4))
```

```
dim(result)

pdf("timeseriesmodel-act-5000.pdf", width=12, height=6)

par(mfrow=c(1,2))
hist(result[3,], main="Effect of Phantom `Influencer' on `Influenced' in Time Series", xlab="Regression
Coefficient"); hist(result[10,], main="z-score of Directional Difference", xlab=paste("Proportion greater
than 0:", mean(result[10,,]>0)))
dev.off()
```

2) Network Simulation

```
library(network)
num.nodes = 400

scale=3; offset=0; y.noise=0.02; response="linear"; nominate.by="distance";friend.noms=1; time.trend=0.4

  invlogit <- function(cc) exp(cc)/(1+exp(cc))
  xx <- runif(num.nodes)
  distances <- as.matrix(dist(xx, diag=TRUE, upper=TRUE))

adj.probs <- array(rbinom(length(distances), 1, invlogit(offset-scale*distances)), dim(distances))

  diag(adj.probs) <- 0
  adj.probs[lower.tri(adj.probs)] <- t(adj.probs)[lower.tri(adj.probs)]

net=network(adj.probs, matrix.type="adjacency", directed=F)
summary(net)
net %v% "latent.var" = xx

  y1 = (xx-0.5)^3+rnorm(num.nodes, mean=0, sd=y.noise)
 y2 = y1+rnorm(num.nodes, mean = time.trend*xx, sd = y.noise)

net %v% "y1"=y1
net %v% "y2"=y2

cor(cbind(xx,y1,y2))
plot(xx,y1)
plot(xx,y2)
plot(y1,y2)

hist(apply(adj.probs,1,sum))

plot(net)
```