Statistical Tests for Contagion in Observational Social Network Studies

Greg Ver Steeg and **Aram Galstyan**
USC Information Sciences Institute
Marina del Rey, CA 90292, USA
{gregv,galstyan}@isi.edu

Abstract

Current tests for contagion in social network studies are vulnerable to the confounding effects of latent homophily (i.e., ties form preferentially between individuals with similar hidden traits). We demonstrate a general method to lower bound the strength of causal effects in observational social network studies, even in the presence of arbitrary, unobserved individual traits. Our tests require no parametric assumptions and each test is associated with an algebraic proof. We demonstrate the effectiveness of our approach by correctly deducing the causal effects for examples previously shown to expose defects in existing methodology. Finally, we discuss preliminary results on data taken from the Framingham Heart Study.

Christakis and Fowler's paper suggesting that obesity may spread along social ties [3] has sparked years of discussion about what constitutes evidence of contagion in observational social network studies (see, e.g., this recent review [2]). The most general result from the causal modeling perspective shows that latent homophily acts as a confounder for contagion so that uniquely pinpointing the strength of contagion is impossible without additional assumptions [14]. In other words, contagion is non-parametrically unidentifiable. However, if the true goal is to test for the presence of contagion, a lower bound on the strength of contagion is all that is necessary. We present a general method to obtain such bounds in this paper.

Identifying causal effects in social networks through intervention is often impractical or even unethical. Mea-

suring all the human traits that affect link formation and observed actions is unrealistic. A method to measure the strength of causal effects without recourse to these alternatives is of central importance for studying social networks. Our method produces a sequence of bounds on the strength of contagion which converge in some limit to the best possible bounds. In this sense, our method is the best solution to the problem of measuring the strength of contagion that does not involve invoking additional (parametric) assumptions.

1 Model and Method

We have two actors Alice(A) and Bob(B) whose actions or attributes we observe at discrete time steps, $t = 1, \ldots, T$. In Fig. 1, we depict a Bayesian network that incorporates both contagion (also known as social influence) and homophily, following Shalizi and Thomas [14]. The only difference is that we are more pessimistic in that we will consider all the attributes of Alice(R_A) and Bob(R_B) to be hidden. In this figure, we condition on E, the presence of a directed edge from Alice to Bob. The formation of such an edge depends in some arbitrary way on the hidden attributes of Alice and Bob. We often refer to this process as *latent homophily*, even though the edge formation process is unrestricted and could, e.g., be heterophilous instead (i.e. edges are more likely to form between actors with different attributes).

We observe some sequence of actions $(A_1, \ldots, A_T)$ (sometimes abbreviated $A_{1:T}$ or simply A) and $B_{1:T}$. Given E, what correlations are possible between A and B? Below we use standard results about graphical models [12] for the network in Fig. 1 along with some simple manipulations using Bayes' rule. We also employ the common shorthand that capital letters represent random variables and we suppress their instantiation when no ambiguity arises, i.e., $P(A_1) \equiv P(A_1 = a_1)$. We additionally require that the transitions are stationary, i.e., $\forall t, t', \quad P(A_t|A_{t-1}, R_A) =$
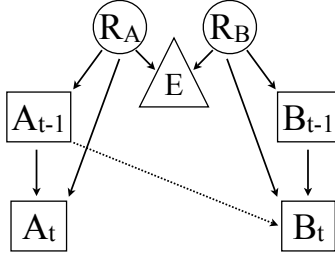
Figure 1: A slice of a Bayesian network representing both latent homophily and contagion (dotted line). We observe a sequence of actions for A and B that depend on previous actions and on hidden attributes R_A, R_B. This graph is conditioned on the presence of a directed edge in the social network, E, from A to B, whose formation depends on R_A, R_B.

$P(A_{t'}|A_{t'-1}, R_A)$, and similarly for B. In principle, we could allow the actions for Alice and Bob to come from any finite discrete set, but for simplicity we will consider $A_t, B_t \in \{0, 1\}$ from here on. Surprisingly, we can allow the size of the hidden attribute space to be infinite. The probability distribution over A, B, conditioned on the presence of an edge, E, takes the following form.

$$P(A_{1:T}, B_{1:T}|E) = \qquad\qquad (1.1)$$
$$\sum_{R_A, R_B} P(R_A, R_B|E)P(A_1|R_A)P(B_1|R_B) \times$$
$$\prod_{t=2}^{T} P(A_t|A_{t-1}, R_A)P(B_t|B_{t-1}, A_{t-1}, R_B)$$

The Bayesian network in Fig. 1 represents a class of models that can be specified by setting the conditional probability distributions for each node, $P(X|\text{parents}(X))$. In this case, unidentifiability means that a particular (non-experimental) probability distribution over observed variables does not pick out a unique model. Therefore, the strength of the dependence of B_t on A_{t-1} is also not uniquely determined. In technical terms the presence of a back-door path from A_{t-1} to B_t is a confounder preventing a unique identification of the causal effect of A_{t-1} on B_t (see [14, 12]). Our goal is not a unique identification of the strength of the causal effect, but rather to establish a lower bound on the strength. We begin by slicing up the space of all possible models according to the strength of the causal effect.

Definition 1.1. *We define Δ-causal models for $\Delta \equiv [\delta_l, \delta_u]$ as the set $\mathcal{P}_\Delta$, consisting of probability distributions, $P(A, B|E)$, s.t. there exists conditional probability distributions that satisfy Eq. 1.1 and that addi-*

tionally satisfy:

$$\delta_l \leq P(B_t = 1|B_{t-1}, A_{t-1} = 1, R_B) \qquad (1.2)$$
$$-P(B_t = 1|B_{t-1}, A_{t-1} = 0, R_B) \leq \delta_u,$$

for all possible values of B_{t-1}, R_B.

The class of models specified by $\delta_l = \delta_u = 0$ we denote by $\mathcal{P}_0$ and refer to as non-causal models.

The quantity in Eq. 1.2 is conventionally referred to as the *average causal effect of treatment*, or just average treatment effect, where the treatment in this case refers to Alice's action and the effect is measured on Bob [10]. We are really bounding the average treatment effect for every sub-population defined by R_B. Identifying whether a distribution is in the set of non-causal models is of special interest and because $\delta_l = \delta_u = 0$, this implies that $P(B_t|B_{t-1}, A_{t-1}, R_B)$ simplifies to $P(B_t|B_{t-1}, R_B)$ in that case.

1.1 Simple Example

Consider a simple function of the observed variables $c(A_{1:T}, B_{1:T})$, or $c(A, B)$. The expectation value of this function is

$$\langle c(A, B) \rangle_P \equiv \sum_{A, B \in \{0,1\}^T} P(A, B|E)c(A, B).$$

Set $T = 4$ and consider a specific observable,

$$c^{(1)}(A_{1:T}, B_{1:T}) = \qquad\qquad (1.3)$$
$$(\mathbf{1}\{A_2 = B_2 \neq A_3 = B_3\} - \mathbf{1}\{A_2 = B_3 \neq A_3 = B_2\})$$
$$\times (1 - \mathbf{1}\{A_1 \neq A_4\}\mathbf{1}\{B_1 \neq B_4\})$$

This operator can only take values 0 or ± 1, so its average must lie in this range. Using Def. 1.1 and simple but tedious algebra verifies that

$$\forall P \in \mathcal{P}_0, \quad \langle c^{(1)}(A, B) \rangle_P = 0. \qquad (1.4)$$

While a fact like this is straightforward to verify, it offers little understanding. In the rest of the paper, we develop methods to find equalities (and inequalities) of this form. Moreover, we focus the search by looking for useful tests so that, e.g., we find conditions that are satisfied $\forall P \in \mathcal{P}_0$, but are violated by models which contain contagion.

For instance, if we define a simple model of influence, $P_\delta(A, B)$ in which $P(A_t = 0) = P(A_t = 1) = P(B_1 = 0) = P(B_1 = 1) = 1/2$ and $B_t = A_{t-1}$ with probability δ otherwise B_t randomly becomes 0 or 1. The "average treatment effect" in this case is just δ. We can easily see that

$$\langle c^{(1)}(A, B) \rangle_{\hat{P}_\delta} = -\frac{3}{16}\delta.$$

Even for a tiny amount of influence Eq. 1.4 is violated, demonstrating that the distribution $P_\delta(A, B)$ cannot be explained by a non-causal model, even with an infinite number of hidden attributes.

1.2 Finding Useful Tests

Determining if a (non-experimental) probability distribution is compatible with a class of models defined by Def. 1.1 seems hopeless because the number of parameters depends on the size of the hidden attribute space which can be infinite. Luckily, as we have just seen, we can find simple conditions which all distributions in $\mathcal{P}_\Delta$ satisfy. A distribution that violates one of these conditions is incompatible with the associated class of models.

We begin by considering a candidate probability distribution, $\hat{P}(A, B)$, a class of models (specified by a set of distributions) $\mathcal{P}$, and some observable, $c(A, B)$. We are looking for the following condition to be satisfied.

$$\langle c(A, B)\rangle_{\hat{P}} - \langle c(A, B)\rangle_P \geq \gamma > 0, \quad \forall P \in \mathcal{P}$$

If this condition is satisfied then $c(A, B)$ constitutes a statistical test that is bounded $\forall P \in \mathcal{P}$, but is violated for the distribution $\hat{P}$. Looking for an observable $c(A, B)$ with associated bound γ leads us to an optimization problem.

$$\underset{\gamma, c(A,B)}{\text{maximize }} \gamma, \text{ s.t.} \tag{1.5}$$

$$-\gamma + \langle c(A, B)\rangle_{\hat{P}} - \langle c(A, B)\rangle_P \geq 0, \quad \forall P \in \mathcal{P}$$

Our goal is to transform this optimization problem into a sequence of linear programs (LP), so that the lower bound, γ becomes successively tighter as we increase the size of the LP. To that end, we will first represent the expectation values in terms of polynomials. This allows us to represent the condition in the second line using a result about representations of non-negative polynomials which we include here.

Theorem 1.1. *(Handelman's representation [7]) Any polynomial, $\gamma - h(x)$, that is positive on a compact domain $\mathcal{K} = \{x : g_1(x) \geq 0, \ldots, g_s(x) \geq 0\}$, where the $g_i(x)$ are linear, can be written in this form,*

$$\gamma - h(x) = \sum_{k \in \mathbb{N}^s} \lambda_k \prod_{i=1}^s g_i(x)^{k_i},$$

using non-negative λ's, with k representing a vector of non-negative integers.

Because the RHS consists of a sum of products of non-negative quantities on $\mathcal{K}$, we can see the LHS should be non-negative. The theorem ensures any positive polynomial can be written in this form. The main drawback, however, is that it does not say how many terms

are required. Although this shortcoming is remedied in [4], those bounds are often impractical. Instead, we can bound the number of terms so that $\sum_i k_i \leq d_{max}$, and therefore γ is an upper bound for $h(x)$ on $\mathcal{K}$, that becomes progressively tighter as we increase d_{max}. As a bonus, providing a concrete representation in terms of λ's provides a certificate, or an algebraic proof, that γ is an upper bound for $h(x)$ on $\mathcal{K}$ (see Sec. 1.4 for an explicit example).

Now we can proceed to re-write the optimization problem in Eq. 1.5 as an LP using Handelman's representation. First, looking at Eq. 1.1 and using convexity, we see that, for $P \in \mathcal{P}_\Delta$

$$\langle c(A, B)\rangle_{P \in \mathcal{P}_\Delta} \leq \max_{P' \in \mathcal{P}_\Delta} \langle c(A, B)\rangle_{P'} = \tag{1.6}$$

$$\max_{P(\cdot|\cdot)} \sum_{A, B \in \{0,1\}^T} c(A, B) P(A_1|R_A) P(B_1|R_B) \times$$

$$\prod_{t=2}^T P(A_t|A_{t-1}, R_A) P(B_t|A_{t-1}, B_{t-1}, R_B)$$

The maximization is over conditional probability distributions that satisfy normalization, positivity, and the condition in Eq. 1.2. We think of the conditional probability distributions as variables, e.g. $x_1 \equiv P(A_1 = 0|R_A)$, normalization is ensured by writing $P(A_1 = 1|R_A) = 1 - x_1$, positivity corresponds to conditions like $g_1(x) = x_1 \geq 0, g_2(x) = 1 - x_1 \geq 0, \ldots$, and Eq. 1.2 corresponds to more complicated linear inequalities involving these variables that depend on δ_l, δ_u. We represent all these linear inequalities with the set $\mathcal{K}_\Delta$. We will give a more concrete demonstration of this mapping in the next section.

To complete the transformation of Eq. 1.6, we also consider the vector of variables, $\boldsymbol{c}$, whose elements we will sometimes index $c_{AB} \equiv c(A, B)$, where the concatenated binary sequences A and B should be interpreted as an integer in $[0, 2^{2T} - 1]$. We can do the same to represent $\hat{P}(A, B)$ as a vector $\hat{\boldsymbol{p}}$, and then expectation values are just dot products. Putting this together, we re-write this equation as

$$\langle c(A, B)\rangle_{P \in \mathcal{P}_\Delta} \leq \max_{\boldsymbol{x} \in \mathcal{K}_\Delta} \boldsymbol{c} \cdot \boldsymbol{f}(\boldsymbol{x})$$

We can ensure $-\gamma + \langle c(A, B)\rangle_{\hat{P}}$ is an upper bound for the RHS (which is the condition written in the second line of Eq. 1.5) by ensuring that $-\gamma + \boldsymbol{c} \cdot \hat{\boldsymbol{p}} - \boldsymbol{c} \cdot \boldsymbol{f}(\boldsymbol{x})$ has a Handelman representation (and is therefore non-negative).

That leads us to the following form of the optimization

in Eq. 1.5.

$$\underset{\gamma, \boldsymbol{c}, \boldsymbol{\lambda}}{\text{maximize}} \; \gamma, \; \text{s.t.} \qquad (1.7)$$

$$-\gamma + \boldsymbol{c} \cdot \hat{\boldsymbol{p}} - \boldsymbol{c} \cdot \boldsymbol{f}(\boldsymbol{x}) = \sum_{k \in \mathbb{N}^s} \lambda_k \prod_{i=1}^{s} g_i(x)^{k_i}$$

$$\gamma, \boldsymbol{\lambda} \geq 0, c_i \in [-1, 1], \sum_i k_i \leq d_{max}$$

Equating the terms of the polynomials on both sides of the second line results in linear equalities among the variables. We restrict the c_i's to some fixed range so that γ cannot be made arbitrarily large by scaling all the c_i's. This optimization turns out to be a linear program (LP), and, hence can be efficiently solved in polynomial time. The feasibility of this LP proves that there is a linear equality that is obeyed by all distributions in $\mathcal{P}_\Delta$ but is violated by the distribution $\hat{P}$. Namely, we have shown that

$$\forall P \in \mathcal{P}_\Delta, \quad \langle c(A, B) \rangle_P \leq \langle c(A, B) \rangle_{\hat{P}} - \gamma.$$

So obviously if $\hat{P} \in \mathcal{P}_\Delta$ this would lead to a contradiction (assuming γ is positive).

Not only does this LP provide us with a concrete bound and the size of the violation by $\hat{P}$, the λ's can be interpreted as an algebraic proof of the upper bound. The main factor determining the size of the LP is the number of variables, λ_i, which is determined by the number of terms we use in our Handelman representation. Mathematica can solve LPs with hundreds of thousands of variables and our code is available [1]. In the next section, we provide a more concrete formulation of this optimization.

1.3 Non-Causal Models

We give a more explicit formulation of Eq. 1.7 for the special case of non-causal models. In this case, each variable sequence, $A_{1:T}$ is a mixture of Markov chains with associated transition probabilities that depend on the unknown value of R_A. We denote by $\alpha_+(\alpha_-)$ the probability that A flips from 0(1) to 1(0) at some time step and $\alpha_0 = P(A_1 = 0)$. We have similar parameters for $B : \beta_{+,-,0}$. We use just α or β when possible to avoid writing out all three.

$$q_A(\alpha) \equiv P(A_{1:T}|R_A) = \alpha_0^{1-A_1}(1-\alpha_0)^{A_1} \qquad (1.8)$$

$$\alpha_+^{F_{01}(A)} \alpha_-^{F_{10}(A)} (1-\alpha_+)^{F_{00}(A)} (1-\alpha_-)^{F_{11}(A)}$$

The same equations hold replacing A with B and α with β. $F_{ij}(A)$ counts the number of transitions from state i to j in string A. If A, A' have the same initial state and the same transition counts(e.g. $(0, 0, 1, 0)$ and $(0, 1, 0, 0)$), they are said to be *partially exchangeable* because they clearly have the same probability of

occurring. This observation alone, especially extended to joint strings on A and B, imposes serious constraints on possible observed probabilities and explains the existence of equalities like Eq. 1.3. We discuss tests based on this idea and their relationship to de Finetti theorems in Appendix A.

In this case, the bounds imposed by Eq. 1.2 are trivial and have already been taken into account by eliminating the dependence of B_t on A_{t-1} in defining the variable above. This leaves us with only 12 inequality constraints to enforce positivity, two for each variable: $g_1(\alpha, \beta) = \alpha_0 \geq 0, g_2(\alpha, \beta) = 1 - \alpha_0 \geq 0, \ldots, g_{11}(\alpha, \beta) = \beta_+ \geq 0, g_{12}(\alpha, \beta) = 1 - \beta_+ \geq 0$. Using $f_{AB}(\alpha, \beta) = q_A(\alpha)q_B(\beta)$ and these definitions for $g_i(\alpha, \beta)$, we can plug these into Eq. 1.7 to search for bounds that are satisfied by probability distributions explained by non-causal models, $\mathcal{P}_0$.

1.4 A Sample Bound

As a simple example of how we can use LPs to give bounds, we begin by bounding $P(A = (0, 0, 1))$ for $P \in \mathcal{P}_0$. We see that

$$P(A = (0, 0, 1)) \leq \max_{\alpha_0, \alpha_+ \in [0,1]} \alpha_0(1 - \alpha_+)\alpha_+$$

from Eq. 1.8 and Eq. 1.6. We are looking for an upper bound γ so that

$$\gamma - \alpha_0(1 - \alpha_+)\alpha_+ \geq 0 \text{ on } \mathcal{K},$$

with $\mathcal{K} = \{\alpha_0, \alpha_+ : \alpha_0, \alpha_+, (1 - \alpha_+), (1 - \alpha_0) \geq 0\}$. Casting the problem as in Eq. 1.7 with $d_{max} = 3$ (with $\boldsymbol{c}$ a constant in this case), we get an LP with 36 variables (counting γ) whose solution results in the following representation.

$$\frac{1}{3} - \alpha_0(1 - \alpha_+)\alpha_+ =$$

$$\frac{1}{3}(1 - \alpha_+)^3 + \frac{1}{3}\alpha_+^3 + (1 - \alpha_0)(1 - \alpha_+)\alpha_+$$

The RHS constitutes an algebraic proof that $\gamma = 1/3$ is an upper bound for $P(A = (0, 0, 1))$ on this domain. Although in principle we can generate algebraic proofs for all bounds presented in the paper, they are unwieldy for all but the simplest examples. Instead we provide code to generate bounds [1].

1.5 Equality Constraints

A special case of Eq. 1.7 occurs when we set $d_{max} = 0$. Essentially, we are looking for $\boldsymbol{c}$, so that $\boldsymbol{c} \cdot \boldsymbol{f}(\boldsymbol{x}) = 0$, and we do this by ensuring that the coefficient of each monomial order $x_1^{k_1} x_2^{k_2} \ldots$ equals zero. Putting the constraints from all these coefficients together leads to some matrix M so that $M\boldsymbol{c} = \boldsymbol{0}$. If T=3, the null

space of M has dimension 0, but if T=4 and we consider $\mathcal{P}_0$, (so that $\boldsymbol{f}(\boldsymbol{x})$ is set by Eq. 1.8), the null space of M has a dimension of 60. This implies that there are 60 linearly independent equalities like the one in Eq. 1.3 that are satisfied by each probability distribution $P(A, B|E) \in \mathcal{P}_0$. A 28-dimensional subspace of these equalities are satisfied for any $\mathcal{P}_\Delta$. For $T = 4$, all of these equalities can be understood in terms of partial exchangeability, see Appendix A. Alternately, we can look for particular useful equalities like Eq. 1.3, or another considered in Appendix B.

In the next section, we focus on inequalities satisfied by all distributions in $\mathcal{P}_0$, so that violation of these inequalities can rule out a non-causal model.

2 Results

Shalizi and Thomas (ST) illustrate how the confounding effects of latent homophily cause standard tests for contagion to fail with two examples [14]. In the first, they demonstrate a non-causal model that looks like contagion. In the second, they consider a simple copying model and show how the observed results appear to be explained by homophily. Our tests correctly identify the underlying mechanism in both cases.

2.1 Homophily Looks Like Contagion

A now popular test for contagion considers unreciprocated, directed edges so that A can influence B, but not vice versa [3]. Then, if we regress B's action based on A's history, versus regressing A's action based on B's history, we should see an asymmetry in the size of the regression coefficient if A influences B. ST's example shows that this asymmetry can be reproduced by latent homophily as long as there is an asymmetry in the edge formation mechanism. E.g., consider all nodes to take some static hidden attribute in the range $[0, 1]$. We say that nodes are more likely to form links with someone who has a similar attribute (homophily), but they also tend to prefer people whose attribute is closer to the median, 0.5, leading to an asymmetry in preference of edge formation. If each node's state at each time step only depends on their hidden attribute and their previous state, we have a model which clearly has no influence. However, ST show that this model does reproduce asymmetries in regression coefficients which would be interpreted as a sign of influence.

We ran the code that ST provided in their paper [14], making only one change so that the state of each node at each time step is a binary variable. For a given graph, we consider all pairs of nodes, A, B so that there is a directed edge from A to B. Then we look

at the frequency of observing a given joint sequence of states for $A_{1:4}, B_{1:4}$, and we use this to construct the empirical probability distribution, $\hat{P}_{LH}(A, B|E)$. We estimated this distribution based on $M = 400,000$ samples. As a first test, we can consider the equality constraints that should be satisfied for any non-causal model, given by Eq. 1.3, Sec. B.

$$\langle c^{(1)} \rangle_{\hat{P}_{LH}} = \qquad (+1499 - 1493)/400000$$
$$\langle c^{(2)} \rangle_{\hat{P}_{LH}} = \qquad (+20006 - 19871)/400000$$

For a non-causal model we expect exactly 0, but for an empirical distribution the results are not exact due to error in our sampled distribution. In this case, we can calculate a simple confidence bound. Because $c^{(1)}, c^{(2)}$ take only the values $0, \pm 1$, and we are trying to determine if the mean value is nonzero, we can use the binomial distribution to give the exact probability of getting an excess of $+1$(heads) over -1(tails) for a fair coin. The p-values we get for the statistics above are 0.54 and 0.25, respectively, which is not extreme enough to rule out the null hypothesis that $\hat{P}_{LH} \in \mathcal{P}_0$.

The previous test is only one of many conditions we expect non-causal models to satisfy. A more comprehensive test is to take the empirical distribution, $\hat{P}_{LH}$ and plug it in to Eq. 1.7 (we used $d_{max} = 9$). The result is an observable $c_{LH}(A, B)$, $\gamma_{LH} = 0.0024$, so that $\forall P \in \mathcal{P}_0, \langle c_{LH} \rangle_{\hat{P}_{LH}} - \langle c_{LH} \rangle_P \geq 0.0024$. How can we interpret this result? Because we have optimized our test based on the data, we cannot apply a straightforward confidence bound like Hoeffding's inequality. Instead, we want to check that the Euclidean distance between the empirical distribution and any distribution in $\mathcal{P}_0$ (which is at least $\gamma/|\boldsymbol{c}_{LH}|$) is much larger than we would expect from sampling error. Based on the central limit theorem (and verified by numerical experiments), we expect the average Euclidean distance between a probability distribution and an empirical distribution estimated based on M samples to be $1/\sqrt{M}$. In Fig. 2, we see that the lower bound on the Euclidean distance between $\mathcal{P}_0$ and $\hat{P}_{LH}$, which is $\gamma/|\boldsymbol{c}_{LH}| = 0.00022$ is much less than we would expect from error in sampling the empirical distribution, ≈ 0.0016, so once again we fail to reject the null hypothesis that $\hat{P}_{LH} \in \mathcal{P}_0$.

2.2 Contagion Looks Like Homophily

In this case, we construct a simple network (see Fig. 3), where each node is defined by a static trait, square or circle. Links are more likely between nodes of the same type. We start each node randomly in the state red or green. We then evolve the state of the graph by repeatedly picking an edge and then copying the state of one node to its neighbor. After many iterations,
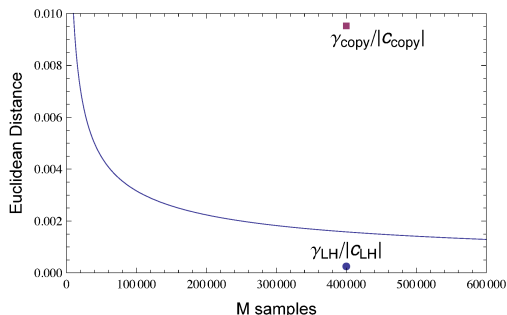
Figure 2: Average Euclidean distance between the vector of empirical frequencies estimated with M samples and the vector of true probabilities from which the samples were drawn. Compared is the Euclidean deviation from $\mathcal{P}_0$ by two empirical distributions.
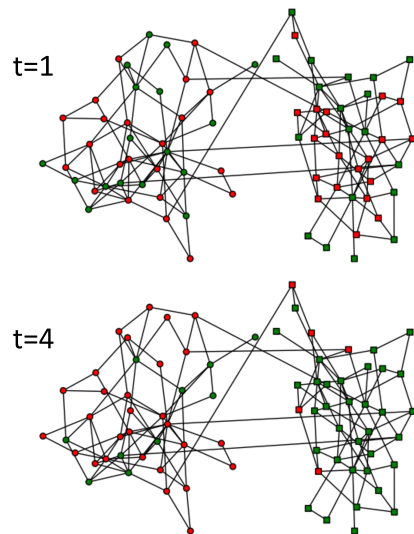

Figure 3: An example of simple copying dynamics on a network in which color (red/green) spuriously appears to evolve according to the static node type (circle/square).

we observe the new state of the graph. ST point out that by looking at the dynamics it can appear that a tendency to become green, e.g., is explained by the static attribute of being a square, while circles tend to become red. However, this is just transient behavior caused by the network structure, the node type has nothing to do with the copying mechanism.

Again, we generate $M = 400,000$ samples using this model (details and code in [14]), generating an empirical distribution, $\hat{P}_{copy}$. Our test easily identifies contagion in this case. For instance, $\langle c^{(1)} \rangle_{\hat{P}_{copy}} = 0.062$, for which the p-value under the null hypothesis is $\sim 10^{-3000}$. If, on the other hand, we solve Eq. 1.7 using $\hat{P}_{copy}$, we get $\gamma_{copy}/|c_{copy}| = 0.0095$, which is much larger than we would expect from error in sampling the empirical distribution (0.0016, see Fig. 2).

2.3 Experimental Results

We have done a preliminary analysis of the Framingham Heart Study (FHS) data, a longitudinal social network study which includes many covariates (e.g. obesity, marriage, depression, smoking, alcohol) and link types (friend, neighbor, co-worker, sibling, spouse). Detailed analysis of this dataset appears in several works [3, 2]. Methodologically, we proceed in an identical fashion to the previous sections.

We consider a particular example from the Framingham Heart Study regarding BMI (a BMI greater than 30 is defined as "obese"). We considered waves 4,5,6, and 7 of the original and offspring cohort, and considered pairs A, B where B nominated A as a friend, and A and B are not related. Because our goal is to rule out the non-causal model, the timing of the edge creation is not a factor. We defined the binary variable $A_t = 0(1)$ to indicate that A's BMI did not (did) increase by more than the median amount since

the last survey. This definition was intended to reduce the effect of dynamic factors influencing all actors in the same way. The median change in BMI for these 4 waves was $(0.55, 0.57, 0.42, 0.20)$. As we mention in the conclusion and Appendix A, we can also test whether the data is consistent with a model including latent homophily and unbounded contagion as a way of identifying the presence of other unmodeled causal effects. We were not able to rule out that $\hat{P}_{BMI} \in \mathcal{P}_\Delta$ (which would suggest the presence of unmodeled causal effects), but we could rule out $\hat{P}_{BMI} \in \mathcal{P}_0$ with high confidence. For instance, using the test in Eq. 1.3 we can rule out a non-causal model with 99% confidence. Since even this claim is controversial, we have included the relevant statistics for observing joint sequences of actions of friends in the supplementary material [1], so that the violation of Eq. 1.3 can be verified directly.

As a contrast, let $A_t = 0(1)$ indicate that Alice smoked at least one cigarette per day. In that case, we could rule out $\hat{P}_{smoking} \in \mathcal{P}_\Delta$ with high confidence for any Δ. Even latent homophily combined with contagion do not explain the correlations in smoking. A complete analysis of FHS data will appear in future work.

3 Related work

Christakis and Fowler were not the first to look for contagion in observational social network studies, but their study on obesity [3] marks the beginning of an eruption of methodological introspection and critique which CF have recently summarized and ad-

dressed [2]. While most of the responses to that work center around the robustness of various parametric modeling assumptions (e.g., sensitivity analysis [15]), our central concern is with the broader, so far unanswered, critique leveled by ST that latent homophily poses a significant barrier to identifying contagion [14].

The main difference to ST's paper is that instead of identifying the strength of causal effects, we put lower bounds on these effects (which would seem to be the primary aim for most practitioners). In this sense, we are actually in the realm of "partial identifiability" [9], an approach ST themselves suggest as an open problem in the final section of their paper [14].

The approach we have taken here considers probability distributions as points in some high-dimensional vector space. In this context we can apply tools from algebraic geometry to answer a variety of questions about set membership. While the possibility of this approach was recognized long ago [6, 8], those approaches relied on computationally infeasible, exact methods. Recent advances in convex relaxations for algebraic geometry problems [11] make this approach feasible. This perspective was considered in a general context [16], but the result here differs in important ways. Casting the optimization problem here as an LP instead of a semi-definite program eliminates some ambiguities in defining an optimal test, leads to a more tractable optimization, and allows us to address more complex problems. Ultimately, this added power allows us to solve a concrete, open problem in causal analysis by bounding the strength of contagion even in the presence of confounders.

4 Discussion

When it comes to human behavior, measuring and controlling for every variable that might be relevant is unrealistic. For that reason, it is important to obtain the best possible bounds on causal strength with the fewest possible assumptions. In the context of a specific graphical model like Fig. 1, we can unambiguously state our result as a lower bound on the strength of contagion.

In the real world, contagion is only one of many effects that could cause deviation from the non-causal model. For example, common external causation could be a factor. Even in that situation our test can provide valuable insight. For instance, we can test whether a distribution belongs to the set of models which include latent homophily and unbounded contagion strength. If not, we can conclude that other factors must be involved. In our analysis of FHS data, we found this to be the case for smoking, for example. This is not surprising if changing laws and restrictions on public smoking act as a common external cause of correlations.

The methods outlined in this paper provide a powerful way to statistically test for causal effects, even in the presence of confounding variables, while invoking minimal assumptions. While this solves the open problem of distinguishing latent homophily and contagion, our analysis highlights another widespread shortcoming of social network studies. Assuming that no external dynamics influence correlations, while a common assumption, is often unjustified. The methods we present here can also be used to test the validity of this assumption. Future work will apply these tests to an in-depth analysis of causal effects in the Framingham Heart Study data.

A Joint Partial Exchangeability

Consider the set of pairs $z, z' \in \{0, 1\}^T$ that obey the relation $z \equiv_e z'$. Here $\equiv_e$ is an equivalence relation that denotes partial exchangeability (PE). Two sequences are partially exchangeable if and only if they have the same initial state and the same count for each transition (which we called $F_{ij}(Z)$ in the text). Clearly, a stationary Markov chains produces partially exchangeable sequences with equal probability, because the probabilities depend only on the initial state and number of transition. Even if we average over Markov chains with unknown, arbitrary transition matrices, this property is preserved. A famous result known as the de Finetti theorem for Markov chains is that in the limit of long sequences, PE (along with a small technical requirement that there is always a probability of visiting some state) is also a sufficient condition for data to be described as a mixture of Markov chains [5]. Essentially, we have a theorem proving that a symmetry property justifies a particular form for a latent variable graphical model.

In the context of the null model, $\mathcal{P}_0$ (model 1 in Fig. 4), where we ask whether Alice and Bob's sequences are both described by mixtures of Markov chains this suggests a simple test. For each joint sequence A, B we should see that $P(A, B) = P(A', B')$ if $A \equiv_e A'$ and $B \equiv_e B'$. We call this notion joint partial exchangeability (JPE), which differs from PE, because statistics for Alice and Bob's sequences could both be PE while failing to be JPE (i.e. $P(A) = P(A'), P(B) = P(B'), P(A, B) \neq P(A', B')$). As an example of the different possibilities consider the sequences $w = (1, 0, 0, 0), x = (0, 0, 1, 0), y = (0, 1, 0, 0), z = (0, 0, 0, 1)$. Clearly, $x \equiv_e y$. Consider also the alternative models in Fig. 4. We can view model (3), for instance, as a mixture of stationary Markov chains on the joint variable (A_t, B_t),

but this implies different exchangeability properties. Going back to our example, we see the difference.

Equality	Models Satisfying
$P(A = x, B = x) = P(A = y, B = y)$	(1),(2),(3),(4)
$P(A = w, B = x) = P(A = w, B = y)$	(1),(4)
$P(A = x, B = x) = P(A = x, B = y)$	(1)
$P(A = z, B = x) = P(A = z, B = y)$	(1),(2)

For longer sequences of observations, this type of test would be computationally much easier than solving the optimization suggested in Eq. 1.7, which is exponential in T. De Finetti's theorem suggests that in the long T limit, this test may even be sufficient.

This test has another nice interpretation in terms of a symmetry condition. If there is influence from A to B some (but not all) of the symmetries are eliminated (see Sec. 1.5). If there is influence in both directions, an even smaller (but nonzero) set of symmetries remain. Essentially, model checking then becomes a question of (approximately) matching symmetries in the data to the appropriate models. Interestingly, these symmetries differ from typical conditional independence relations. Because we have a mixture of Markov processes, none of the observed variables are independent. A similar perspective has been applied for testing the order of a Markov process [13].
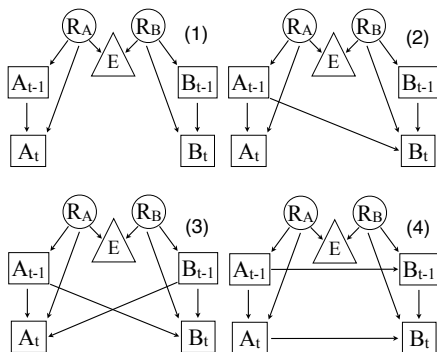

Figure 4: Four model variations with different exchangeability properties.

B Another Equality

We can look for equalities that are useful for identifying certain types of influence by solving Eq. 1.7 using some influence model for $\hat{P}$ and setting $d_{max} = 0$. We already saw an example in Eq. 1.3 which was obtained by setting $T = 4$ and defining $P_\delta^{(i)}(A, B)$ so that $P(A_t = 0) = P(A_t = 1) = 1/2$ and $B_t = A_t$ with probability δ otherwise B_t randomly becomes 0 or 1. This can be thought of as an "instant" influence model, where Alice's choice at time t influences Bob's choice at time t. In contrast, we had previously defined

$$
\begin{array}{cccccccccccccccc}
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
-1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \\
0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\
-1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
\end{array}
$$

Table 1: Specifying an observable for which $\forall P \in \mathcal{P}_0, \langle c^{(2)}(A, B) \rangle_P = 0$.

a delayed influence model, $P_\delta(A, B)$ where $B_t = A_{t-1}$ with probability δ. Solving Eq. 1.7 using P_δ leads to another equality constraint, $\forall P \in \mathcal{P}_0, \langle c^{(2)} \rangle_P = 0$, listed in Table 1. For models with influence, on the other hand, we have the following.

$$
\langle c^{(2)} \rangle_{P_\delta} = \frac{7}{16}\delta \qquad \langle c^{(1)} \rangle_{P_\delta^{(i)}} = \frac{1}{8}(3 - \delta^2)
$$

The value of $c^{(2)}(A, B)$, for a particular $A \in \{0,1\}^4, B \in \{0,1\}^4$ can be read from Table 1. Simply consider the sequence A or B as a binary number and pick the A-th row and the B-th column. This equality is the maximally violated one for a particular model of delayed influence, but it provided significant violations for synthetic and real-world data in Sec. 2.

Acknowledgements

References

[1] Code and data is available in the associated supplementary material or at `http://www.isi.edu/~gregv/contagion.nb`.

[2] N. Christakis and J. Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, 2012.

[3] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, 357(4):370–379, July 2007.

[4] E. De Klerk and M. Laurent. Error bounds for some semidefinite programming approaches to polynomial minimization on the hypercube.

SIAM Journal on Optimization, 20(6):3104–3120, 2010.

[5] P. Diaconis and D. Freedman. De Finetti's theorem for Markov chains. *The Annals of Probability*, 8(1):115–130, 1980.

[6] D. Geiger and C. Meek. Quantifier elimination for statistical problems. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.

[7] D. Handelman. Representing polynomials by positive linear functions on compact convex polyhedra. *Pacific Journal of Mathematics*, 132(1):35–62, 1988.

[8] C. Kang. Inequality constraints in causal models with hidden variables. In *In Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, pages 233–240, 2006.

[9] C. Manski. *Identification for prediction and decision*. Harvard University Press, 2007.

[10] S. Morgan and C. Winship. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press, 2007.

[11] P. A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Math. Program.*, 96(2, Ser. B):293–320, 2003.

[12] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2009.

[13] F. Quintana and P. Müller. Nonparametric bayesian assessment of the order of dependence for binary sequences. *Journal of Computational and Graphical Statistics*, 13(1):213–231, 2004.

[14] C. Shalizi and A. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

[15] T. VanderWeele. Sensitivity analysis for contagion effects in social networks. *Sociological Methods & Research*, 40(2):240–255, 2011.

[16] G. Ver Steeg and A. Galstyan. A sequence of relaxations constraining hidden variable models. In *Proc. of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.