Ruling out latent homophily in social networks

Greg Ver Steeg
Information Science Institute
University of Southern California
Marina del Rey, CA 90292
gregv@isi.edu

Aram Galstyan
Information Science Institute
University of Southern California
Marina del Rey, CA 90292
galstyan@isi.edu

Abstract

Despite recent high profile studies identifying counter-intuitive behaviors (e.g. obesity [3]) as being socially contagious, Shalizi and Thomas [12] have demonstrated that homophily on *latent* attributes is indistinguishable from influence. For sociologists to unequivocally identify influence effects in networks they must rule out the possibility of latent homophily as an explanation. This requires either undertaking the Sisyphean task of measuring every hidden attribute that might influence the formation of links in social networks or, our goal, determine the conditions for distinguishing influence from homophily even in the presence of unobserved attributes. Our test is inspired by the Bell inequalities: a simple inequality involving observed probability distributions which is obeyed by classical physics, but violated by quantum physics. We show any model producing correlations between actors through *static* latent homophily alone will obey certain constraints, and we develop and test a technique to detect violation of these constraints.

1 Introduction

Sociologists often observe that individuals who are connected in a social network exhibit behaviors that are highly correlated. This correlation is usually explained via two effects: homophily and influence. Influence, or contagion, supposes that actors change to become more similar to their neighbors in the network. Whereas, homophily posits that individuals form connections in the network precisely because they are already similar. Distinguishing true sources of influence is very important in situations where we might like to affect the influencer to promote a desired change as in, e.g., social policy or viral marketing.

An example illustrates the difficulty of distinguishing the two. Suppose Alice is friends with Bob, a smoker, and some time later Alice begins smoking. If Alice would not have begun smoking if she had not known Bob, we would certainly say she was influenced by Bob. Unfortunately, this counterfactual is impossible to test. An alternate explanation is that Alice and Bob both suffer from depression, and that is why they became friends. Alice is already predisposed to start smoking, and would have begun even if she had never met Bob. A typical sociological study would attempt to control for this covariate – either by measuring Alice and Bob's depressive tendencies or some substitute that indicates those tendencies. In this case, the difficulty comes from trying to measure all possibly relevant covariates; this is the approach taken in [2, 3].

Shalizi and Thomas [12] show that under a general non-parametric model of homophily, influence and homophily on unobserved attributes cannot be distinguished. Their results rely only on basic facts about the

structure of graphical models (see, e.g., [9] for a review). They conclude that either one needs to measure all relevant covariates or strong parametric constraints on the correlation model are necessary. Previous work distinguishing homophily and influence either assumed such parametric constraints explicitly [13], or only considered homophily on observed attributes [5], or assumed a symmetry between the hidden attributes of influencers and those influenced [1].

We develop a test for latent homophily that makes none of the previous assumptions. Our contributions are as follows:

- In Section 2, we define homophily models as in [12], adding a natural restriction that, effectively, hidden attributes are fixed over time. We call these static latent homophily (SLH) models, see Fig. 1.
- From the perspective of algebraic geometry, we show in Section 3 that the space of probability distributions over observed behaviors is restricted for SLH models. Section 4 introduces an efficient method to find a nonlinear convex relaxation of this space.
- We use this relaxation in Section 5 to demonstrate that a simple influence model acting on a real social network taken from the online news community "digg.com" produces correlations outside of the bounds of those allowed by SLH.

2 Latent homophily models

In Fig. 1, we start with the most general picture of latent homophily. We have two actors Alice(A) and Bob(B) whose actions we observe at various time steps, $t = 1, \ldots, T$. We consider some hidden attributes of Alice(R_A) and Bob(R_B) and E depends somehow on both hidden attributes and represents information about edges between them (e.g., a time-dependent sequence of edges, possibly directed or weighted, possibly including edges of various kinds). Unlike previous works [1, 5], we do not assume that E depends symmetrically on R_A and R_B; an important consideration in the case of asymmetric (directed) links. Although we do not explicitly include some observed attributes in this model as in [12], there presence makes no difference to the results.

Given E, what correlations are possible between A and B? Below we use the definition of the graphical model [9] in Fig. 1 along with some simple manipulations using Bayes' rule.

$$\begin{aligned} p(A^{1:T}, B^{1:T}|E) &= \sum_{R_A, R_B} p(A^{1:T}|R_A)p(B^{1:T}|R_B)p(E|R_A, R_B)p(R_A)p(R_B)/p(E) \\ &= \sum_{R_A, R_B} p(A^{1:T}|R_A)p(B^{1:T}|R_B)p(R_A, R_B|E) \end{aligned} \tag{1}$$

We also take into account that A^t may depend on A^{t-1} in addition to R_A.

$$p(A^{1:T},B^{1:T}|E) = \sum_R p(R_A, R_B|E) \prod_t p(A^t|A^{t-1}R_A)p(B^t|B^{t-1}R_B) \tag{2}$$

Since the hidden variables R_A, R_B can in principle be arbitrarily correlated and the dependence on the hidden variable is also arbitrary, we shorten the notation from R_A, R_B to just R from here on.

It is easy to see that an arbitrary marginal distribution $\bar{p}(A^{1:T}, B^{1:T})$ can be written in a manner consistent with this graphical model,[1] given appropriate definition of $p(R)$. Let R be a vector of 2^{2T} values so that

[1] In fact, in a future longer version of this work we will demonstrate that $|\text{dom}(R)| \sim 2^T$ is a necessary and sufficient condition to reproduce an arbitrary correlation between A and B.

each is associated with a pair of sequences for Alice and Bob, $R = (A, B)$. Now take

$$\begin{aligned} p(R = (A, B)|E) &= \bar{p}(A, B) \\ p(A^t|A^{t-1}, R = (A', B')) &= \delta_{A^t, A'^t} \end{aligned} \tag{3}$$

and similarly for B. Plugging these values into Eq. 2, we reproduce the arbitrary distribution $\bar{p}$.

Intuitively, we have reproduced arbitrary correlations by making the space of our hidden attribute large and then allowing the dependence of A^t on R to change at each time step. Effectively, this is the same as allowing the hidden attribute to fluctuate with time. Generally, when we say that actions depend on hidden attributes, this is not what we mean. We assume that the attributes (e.g. gender, IQ, etc.) are not changing (or at least not quickly) with respect to the observed actions (smoking, posting on a social network, etc.), and we wish to explain the latter in terms of the former. Therefore, below, we restrict ourselves to models with this property.

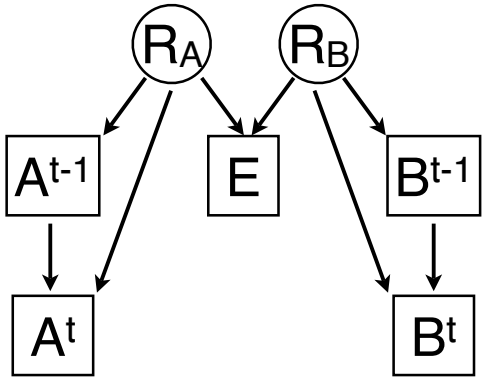

Figure 1: A slice of a latent homophily model. We observe a sequence of actions for $A^1, \ldots, A^{t-1}, A^t, \ldots, A^T$ (sometimes abbreviated A) and B that depend on some hidden attributes R_A, R_B. Presence and properties of edges between them, E, depend in some arbitrary way on R_A, R_B.

Static latent homophily model We now define static latent homophily models (SLH) by demanding the crucial addition of stationarity: the transition probability does not change over time.

$$\forall t, t', r: \; p(A^t = a|A^{t-1} = b, R = r) = p(A^{t'} = a|A^{t'-1} = b, R = r). \tag{4}$$

The homophily model of [12] also looks like Fig. 1, but without the stationary assumption in Eq. 4. Note that the demonstration in Eq. 3 that latent homophily reproduces arbitrary correlations only holds without stationarity.

The stationary Markov assumption restricts the probability of observing certain sequences. If Bob's state is a sequence of coin flips, it is highly unlikely that Alice independently produces the same sequence without seeing (or being influenced by) Bob's coin flips. We will make this intuition more precise in the next section.

3 Algebraic geometry of SLH

Looking at Eq. 2, we can see that we have just defined a polynomial mapping from the small space of conditional probabilities to the larger space of Alice and Bob's observed joint probability distribution. The structure of Eq. 2 is a convex combination over the (possibly infinite) factorizable joint distributions.

For simplicity, we now consider a SLH where we restrict ourselves to variables $A^t, B^t \in \{\pm 1\}$, and we have conditioned on some arbitrary measurement of E (e.g., E is a directional link from A to B, or E is an edge of a certain weight, etc.). Each variable sequence, $A^{1:T}$ is a Markov chain with associated transition probabilities that depend on the unknown value of R. We denote by $\alpha_+(\alpha_-)$ the probability that A flips from $+(-)$ to $-(+)$ at some time step and $\alpha_0 = p(A^1 = +1)$. We have similar parameters for $B : \beta_{+,-,0}$. For legibility below, we suppress the functional dependence of these probabilities on R and we take A to represent the sequence $A^{1:T}$.

$$p(A^{1:T}|R) = \alpha_+^{F_+(A)} \alpha_-^{F_-(A)} (1-\alpha_-)^{S_-(A)} (1-\alpha_+)^{S_+(A)} \alpha_0^{1/2(1+A^1)} (1-\alpha_0)^{1/2(1-A^1)} \tag{5}$$

$$F_\pm(A) = \sum_{t=1}^{T-1} \frac{1}{4}(1 \pm A^t)(1 - A^{t+1}A^t)$$

$$S_\pm(A) = \sum_{t=1}^{T-1} \frac{1}{4}(1 \pm A^t)(1 + A^{t+1}A^t)$$

The same equations hold replacing A with B and α with β.

We can define the vector,

$$x = (x_1, \ldots, x_6) \equiv (\alpha_+, \alpha_-, \alpha_0, \beta_+, \beta_-, \beta_0).$$

Now consider the expected outcomes from some arbitrary set of measurements

$$y_j \equiv \langle \mathcal{O}_j(A, B) \rangle, j = 1, \ldots, n.$$

In principle, the set of $\mathcal{O}_j(A, B)$ could consist of the indicator functions for each possible outcome (in which case $n = 2^{2T}$), but we would like to reserve the ability to pick a smaller set of measurements for computational reasons later on. Then

$$\begin{aligned} f_j(x_R) \equiv & \sum_{AB} p(A^{1:T}|R) p(B^{1:T}|R) \mathcal{O}_j(A, B) \\ y_j = & \sum_R p(R|E) f_j(x_R) \end{aligned} \tag{6}$$

This represents a polynomial mapping from $\mathbb{R}^6 \to \mathbb{R}^n$ where the domain is the region

$$K = \{x \in \mathbb{R}^6 : g_i(x) = x_i(1 - x_i) \geq 0, i = 1, \ldots, 6\}$$

because each x_i represents a different transition (or prior) probability. The set of all y is just the convex hull of $f(x)$ where $x \in K$.

Consider the following representations of this convex set,

$$\begin{aligned} SLH = & \quad \text{conv}(\{y \in \mathbb{R}^n : \exists x \in K, y = f(x)\}) \\ B = & \quad \{b \in \mathbb{R}^n : \forall x \in K, y = f(x), 1 - b \cdot y \geq 0\}. \end{aligned} \tag{7}$$

The set function "conv" represents the convex hull of a set. The second line gives a representation of the convex hull in terms of an intersection of half-spaces [11].[2] Clearly, if $\forall x \in K, b \cdot f(x) \leq 1$, this is also true for any convex combination, $\forall x_R \in K, b \cdot (\sum_R p(R) f(x_R)) \leq 1$.

[2]Note that our formulation implicitly presupposes that the origin is on the interior of our convex set. This condition can be insured with a simple translation of the vector y. E.g. $y \to y - \int_K dx f(x) / \int_K dx$

Given a representation of the convex hull, B, one can determine that a point $\hat{y}$ is outside SLH by finding a $b_0 \in B$ such that $b_0 \cdot \hat{y} > 1$. If we consider a subset $RB \subset B$, this set amounts to a convex relaxation on the original set SLH. That is,

$$b_0 \in RB \wedge b_0 \cdot \hat{y} > 1 \rightarrow \hat{y} \notin SLH,$$

but for this relaxation, the converse is not true.

4 SOS Relaxation

We now consider a subset of B which can be efficiently described and optimized over. We are trying to describe the set of b so that $1 - b \cdot f(x) \geq 0$, for $x \in K$. When dealing with positive polynomials, a simple, effective relaxation is to consider bounded degree sums-of-squares polynomials

$$SOS_d = \{s(x) : \exists q_i(x) \in \mathbb{R}[x], \deg(q_i(x)) \leq d/2, s(x) = \sum_i q_i(x)^2\}.$$

That is, if we can write $1 - b \cdot f(x) = s_0(x)$, for $s_0 \in SOS_d$, we are guaranteed that it is positive. In general, it is not true that every positive polynomial can be written as an SOS (see [8] for a review of the large body of work about SOS and positive polynomials). SOS polynomials have the desirable property that they can be written in the form

$$s(x) = z^\intercal A z,$$

where $z = (1, x_1, x_1 x_2, x_1 x_2^2 ...)$ is a vector of monomials in the variables and $A \succeq 0$ indicates a positive semidefinite matrix. Then our condition for $1 - b \cdot f(x) = s_0(x)$ amounts to linear relationships between coefficients along with a linear matrix inequality, $A \succeq 0$. These types of problems are called semidefinite programs(SDP) and many powerful techniques exist to solve them. We use SOSTools [10] in MATLAB to convert SOS programs to SDP which are then solved by, e.g., SeDuMi [14].

In our case, because we only demand positivity on a bounded region K, defined by polynomials $g_i(x) \geq 0$, we make things a little easier. Not only do we consider SOS polynomials $s(x)$, we also consider polynomials in the "positive cone" of $g_i(x)$. Roughly, that is just the set of polynomials that are formed as sums of products of the $g_i(x)$ and $s(x) \in SOS$. Clearly, if $g_i(x) \geq 0, \forall x \in K \rightarrow s(x) g_i(x) \geq 0, \forall x \in K$ and $s(x) g_1(x) g_2(x) \geq 0, \forall x \in K$, and so on.

Therefore, we define the set, $RB_1 \subseteq \ldots \subseteq RB_d \subseteq B$.

$$RB_d = \{b \in \mathbb{R}^n : \forall x \in K, s_i(x) \in SOS_d, \quad 1 - b \cdot f(x) = s_0(x) + \sum_i s_i(x) g_i(x)\}$$

By construction, $s_0(x) + \sum_i s_i(x) g_i(x) \geq 0$ for all $x \in K$. For any $b \in RB_d$, this proves $1 - b \cdot f(x) \geq 0$, and for any y in the convex hull of $f(x)$, this will also be true. This amounts to a sequence of convex relaxations of the set SLH. Note that as defined, this sequence does not necessarily converge to SLH, though straightforward generalizations of this technique can provide theoretical, if computationally impractical, guarantees of convergence [4].

For a specific observed distribution $\hat{y}$, we search for a hyperplane $b \in RB_d$ so that $b \cdot \hat{y}$ is maximized.

$$\begin{aligned} &\max_{b, s_i(x)} \; b \cdot \hat{y} \\ 1 - b \cdot f(x) - \sum_i s_i(x) g_i(x) = \quad & s_0(x) \\ s_i(x) \in \quad & SOS_d \end{aligned} \tag{8}$$

This format corresponds to a sum-of-squares (SOS) program and it can be efficiently translated into a semidefinite program and solved numerically [8].

Solving this SOS is constructive in that it provides specific SOS polynomials proving that $1 - b \cdot y \geq 0$, for any $y \in SLH$. Furthermore, if we find a solution b such that $b \cdot \hat{y} > 1$, this constitutes proof that the statistics $\hat{y}$ could not have been generated by a SLH.

5 Results

We tested our results using a real world social network from the online news portal "digg.com" [6]. This network had $M = 1,731,659$ edges and $N = 279,634$ nodes. In principle, we cannot know for sure whether information spread on this network is due to influence or homophily, so we begin, as in [5], by doing a semi-synthetic analysis by simulating a known influence model on the real graph of the social network.

For our influence model we started all the nodes in a random state ± 1. At each of M steps, we picked a random (directed) edge from $A \to B$ and had B copy A's state. Then we considered three time slices from this evolution to construct the statistics $\hat{p}(A^{1:3}, B^{1:3}|E = 1)$, where $E = 1$ means there exists a directed edge from A to B. Using Eq. 8, we construct an observable $\mathcal{O}$, such that $\langle \mathcal{O} \rangle_{SLH} \leq 1$, while the mean value of this observable on our data is maximized. We can use Hoeffding's inequality to give the confidence that $\langle \mathcal{O} \rangle_{data} > 1$. Because this confidence goes like $\sim 1 - e^{-M}$, our confidence to rule out SLH in this case is indistinguishable from 1.

5.1 Conclusion

Intuitively, our test to rule out SLH is motivated by Bell inequalities in quantum physics (see [7] for a computer scientist friendly introduction). In that case, you have two particles that are spatially separated and you want to check whether there exists a "local hidden variable" theory that describes the measured correlations. Bell showed that all correlations produced by local hidden variable theories would satisfy a simple linear equality. Subsequent experiments violated this inequality confirming the existence of what Einstein referred to as "spooky action at a distance." In some sense, we have constructed a Bell inequality for social networks where we are detecting "spooky" correlations between friends.

We have constructed an efficient test that can rule out static latent homophily as an explanation for correlations in social networks. Tests on a simulated influence model on a real world social network show that our test can rule out SLH in such cases. Future work will test the technique on real processes occurring on large online social networks. We would also like to extend the nonlinear convex relaxation technique for other graphical models with hidden variables.

Acknowledgments

Thanks to Cosma Shalizi and Jennifer Neville for useful discussions while visiting ISI. This research was partially supported by the National Science Foundation Grant No. 0916534.

References

[1] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15, New York, NY, USA, 2008. ACM.

[2] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, December 2009.

[3] Nicholas A. Christakis and James 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] Etienne De Klerk and Monique Laurent. Error bounds for some semidefinite programming approaches to polynomial minimization on the hypercube. *SIAM Journal on Optimization*, 20(6):3104–3120, 2010.

[5] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 601–610, New York, NY, USA, 2010. ACM.

[6] Kristina Lerman and Rumi Ghosh. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, May 2010. Data available, http://www.isi.edu/integration/people/lerman/downloads.html.

[7] Nielsen, Michael A. and Chuang, Isaac L. *Quantum Computation and Quantum Information*. Cambridge University Press, 1 edition, October 2000.

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

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

[10] Stephen Prajna, Antonis Papachristodoulou, Peter Seiler, and Pablo A. Parrilo. Sostools and its control applications. In Didier Henrion and Andrea Garulli, editors, *Positive Polynomials in Control*, volume 312 of *Lecture Notes in Control and Information Sciences*, pages 273–292. Springer Berlin / Heidelberg, 2005.

[11] Ralph T. Rockafellar. *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*. Princeton University Press, December 1996.

[12] Cosma R. Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *arxiv:1004.4704*, Oct 2010.

[13] Tom A. B. Snijders, Christian E. G. Steglich, and Michael Schweinberger. Modeling the co-evolution of networks and behavior. In *In*, 2006.

[14] Sturm, J. F. Using sedumi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999.