Non-Parametric Inference of Relational Dependence (Supplementary File)

Ragib Ahsan[1] **Zahra Fatemi**[1] **David Arbour**[2] **Elena Zheleva**[1]

[1]Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
[2]Adobe Research, USA

A PROOFS

In this section we present the proofs of consistency for HSIC and relational HSIC under weak dependence. The approach here is to extend the results of Chwialkowski et al. [2014] and Leucht and Neumann [2013], who analyze degenerate U and V-statistics (which includes HSIC as a specific instantiation) under weak dependence in spaces that admit euclidean distances to the more general setting of graph structured spaces. Much of the results carry through after modifications to accommodate the fact that the number of reachable instances at a specific distance is irregular. We present a modification of the relevant proof which shows the convergence of the distribution of degenerate V-statistics, which may be of independent interest, and then describe the application to our setting and the extension to relational variables.

A.1 V-STATISTICS UNDER RELATIONAL WEAK DEPENDENCE

Let $X = \{X_1, \dots X_n\}$ be the set of given observations. Define h to be a symmetric function, taking m arguments. A V-statistic is a function defined with respect to h taking the form

$$V(h, X)_n = \frac{1}{n^m} \sum_{i \in i_1 \dots i_m \in N^m} h(X_{i_1}, \dots X_{i_m})$$

where N^m is defined as the Cartesian product of the set $1, \dots, n$ and n is the total number of observations. In the sequel, we will write $V(h, X)$ as $V(X)$ to reduce notational clutter. We will refer to h as the *core*[1].

We say that a core h is *j-degenerate* if for every $x_1, \dots, x_j$,

$$E[h(X_1, \dots, X_j, X^*_{j+1}, \dots, X^*_m)] = 0$$

where $X^*_{j+1}, \dots, X^*_m$ are independent samples drawn from the same distribution as X_1. A core is called canonical if for all $j \leq m - 1$ it is j-degenerate. Finally, we call a V-statistic with a 1-degenerate core a *degenerate V-statistic*.

We now provide a proof of consistency of degenerate V-statistics for relational data under weak dependence. The strategy of this proof is to first approximate the V-statistic with weighted sums of squares, and then apply the central limit theorem to this approximation. The approximation used is the spectral decomposition of the core

$$h(x, y) = \sum_k \lambda_k \Phi(x) \Phi(y)$$

where λ_k are the nonzero eigenvalues of $E[h(x, X_0)\Phi(X_0)] = \lambda\Phi(x)$, and $\Phi(x)$ are the associated eigenvectors. This strategy largely mirrors what is found by Leucht and Neumann [2013]. However, in that case, the approximations are

[1]In order to prevent confusion, we follow Chwialkowski et al. [2014] and do not follow the canonical convention of calling h the kernel.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

constructed as a function of distance in time. Our contribution is a generalization of the approximations to network domains that follow the aforementioned assumptions. This is done by considering *sets* of instances separated by shortest path distance of k, rather than assuming that there is always a single instance at distance k, and adapting results accordingly.

Theorem 2. *Let $(Z_k)_k$ be centered, jointly normal random variables with $\mathrm{Cov}(Z_j, Z_k) = \sum_{r=-\infty}^{\infty} \mathrm{Cov}(\Phi_j(X_0), \Phi_k(X_r))$, and $(\lambda_k)_k, (\Phi_k)_k$ be the sequence of non-zero eigenvalues and corresponding eigenfunctions of $E\left[h(x, X_0)\Phi(X_0)\right] = \lambda\Phi(x)$. Under the aforementioned assumptions, $V_n \xrightarrow{d} Z := \sum_k \lambda_k Z_k^2$, as $n \to \infty$, and $EZ = \sum_{r\in\mathbb{Z}} Eh(X_0, X_r) < \infty$ i.e., the infinite series that defines Z converges in L_1.*

Proof. Let $(\lambda_k)_k$ be an enumeration of the positive eigenvalues of $Eh(x, X_0)\Phi(X_0) = \lambda\Phi(x)$ sorted in decreasing order, and $(\Phi_k)_k$ be the corresponding eigenfunctions. Following Leucht and Neumann [2013], we set $\lambda_k := 0, \Phi_k \equiv 0, \forall k > L$, when the number L of non-zeros eigenvalues is finite. We are given from a version of Mercer's theorem (given by Theorem 2 of Sun Sun [2005]) that

$$h^{(K)}(x, y) = \sum_{k=1}^{K} \lambda_k \Phi_k(x)\Phi_k(y) \underset{K\to\infty}{\longrightarrow} h(x, y), \forall x, h \in \mathrm{supp}(P^{X_0})$$

Leucht and Neumann [2013] provide the prerequisites necessary for the equation to converge absolutely and uniform on compact subsets of $\mathrm{supp}(P^{X_0})$, which apply directly in our setting as well. We will consider an approximation of V_n by a V-statistic with a kernel with finite spectral decomposition given by $V_n^{(K)} = \frac{1}{n}\sum_{s,t}^{n} h^{(K)}(X_s, X_t)$. Because h is positive semi-definite by definition, all eigenvalues are non-negative, implying $V_n - V_n^{(K)} \geq 0$. This implies

$$\begin{aligned}
&E\left|V_n - V_n^{(K)}\right| = E\left[V_n - V_n^{(K)}\right] \\
&= E\left[h(X_0, X_0) - h^{(K)}(X_0, X_0)\right] + \sum_{r=1}^{n-1} 2(1 - r/n)E\left[h(X_0, X_r) - h^{(K)}(X_0, X_r)\right]
\end{aligned}$$

By majorized convergence the first term converges to zero as $K \to \infty$. For the second term, repeated application of Cauchy-Schwarz gives

$$\begin{aligned}
&\sum_{r=1}^{n-1} 2(1 - r/n)E\left[h(X_0, X_r) - h^{(K)}(X_0, X_r)\right] \\
&\leq 2\sum_{r=1}^{\infty}\left|\sum_{j\in\Delta_r} E\left[\sum_{k=K+1}^{\infty} \lambda_k\Phi_k(X_0)\Phi_k(X_j)\right]\right| \\
&= 2\sum_{r=1}^{\infty}\left|E\left[\sum_{j\in\Delta_r}\sum_{k=K+1}^{\infty} \lambda_k\Phi_k(X_0)(\Phi_k(X_j) - \Phi_k(\widetilde{X}_j))\right]\right| \\
&\leq 2\sum_{r=1}^{\infty}\sqrt{E\left[\sum_{j\in\Delta_r}\sum_{k=K+1}^{\infty} \lambda_k\Phi_k^2(X_0)\right]}\sqrt{E\left[\sum_{j\in\Delta_r}\sum_{k=K+1}^{\infty} \lambda_k\left(\Phi_k(X_r) - \Phi_k(\widetilde{X}_j)\right)^2\right]} \\
&\leq 2\sqrt{\sum_{r=1}^{\infty}\lambda_k}\sum_{r=1}^{\infty}\sqrt{E\left[\sum_{j\in\Delta_r}\sum_{k=1}^{\infty} \lambda_k\left(\Phi_k(X_j) - \Phi_k(\widetilde{X}_j)\right)^2\right]} \\
&\leq 2\sqrt{\sum_{k=K+1}^{\infty}\lambda_k}\sum_{r=1}^{\infty}\sqrt{\sum_{j\in\Delta_r} E\left[h(X_j, X_j) - h(X_j, \widetilde{X}_j) - h(\widetilde{X}_j, X_j) + h(\widetilde{X}_j, \widetilde{X}_j)\right]} \\
&\leq 2\sqrt{\sum_{k=K+1}^{\infty}\lambda_k}\sum_{r=1}^{\infty}\sqrt{2\max(\mathrm{deg})^r\mathrm{Lip(h)}}\sqrt{\tau(r)}
\end{aligned}$$

Where Δ_r is the set of nodes whose shortest path distance from X_0 is r, $\max(\deg)$ is the largest degree in the network, and $\widetilde{X}_r$ denotes a copy of X_r that is independent of X_0 and satisfies $E\|X_r - \widetilde{X}_r\|_1 \leq \tau(r)$. Because $\sum_{k=1}^{\infty} \lambda_K = Eh(X_0, X_0) < \infty)$, thus $\sum_{k=K+1}^{\infty} \lambda_k \to 0$ as $K \to \infty$ we arrive at $\sup_n E\left|V_n - V_n^{(K)}\right| \underset{K\to\infty}{\longrightarrow} 0$.

The proof of the central limit theorem for partial sums, i.e., for $K \leq L$

$$V_n^{(K)} = \sum_{k=1}^{K} \lambda_k \left(n^{-1/2} \sum_{t=1}^{n} \Phi_k(X_t) \right)^2 \xrightarrow{d} \sum_{k=1}^{K} \lambda_k Z_k^2 \tag{1}$$

follows a direct application of Leucht and Neumann [2013] Theorem 2.1 proof part (*ii*). Combining these two results, to satisfy the requirements of Theorem 2 of Dehling et al. [2009] we arrive at $V_n \xrightarrow{d} Z := \sum_k \lambda_k Z_k^2$. The only item remaining to be shown is $EX < \infty$, which follows from a direct application of part (*iv*) of the proof of Theorem 2 provided by Leucht and Neumann [2013]. □

We now turn our attention to the Hilbert-Schmidt independence criterion. Note that both follow almost immediately from implications of theorem 2.

Theorem 1. *Under the aforementioned assumptions the Hilbert-Schmidt independence criterion of two weakly dependent propositional variables converges in L_1 to its population counterpart, i.e., $\left|\overline{HSIC_n} - HSIC_{population}\right| \underset{d}{\longrightarrow} 0$.*

Proof. Recall that the Hilbert-Schmidt independence criterion (HSIC) is a test of dependence, i.e. a hypothesis test of paired samples where the null hypothesis is that the two samples are generated independently, $\mathbb{P}_{x,y} = \mathbb{P}_x \mathbb{P}_y$. Our focus is on the empirical estimator of HSIC, which can be written as degree-four V-statistic with a core defined by:

$$h(x_1, x_2, x_3, x_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)}) k(y_{\pi(1)}, y_{\pi(2)}) + k(y_{\pi(3)}, y_{\pi(4)}) - 2k(y_{\pi(2)}, y_{\pi(3)}) \tag{2}$$

where S_n is the set of permutations over a set of n elements. Convergence then follows as a direct application of theorem 2 and the weak law of large numbers. Note that under independence Z is a zero mean, jointly Gaussian variable and the resulting sequence $\sum_i \lambda_i Z_i^2$ is mean zero. □

Corollary 1. *Under the aforementioned assumptions the Hilbert-Schmidt independence criterion between a weakly relational and a weakly dependent propositional variable converges in L_1 to its population counterpart, i.e., $|HSIC_n - HSIC_{population}| \underset{d}{\longrightarrow} 0$.*

Proof. The central items to be shown in order to apply the results of theorem 2 to apply are (1) relational kernels define a valid V-statistic, and (2) the relational variable remains weakly-dependent. Item (1) follows directly by denoting one of the variables in equation 2 to be a set of instances return by the path predicate and k to be the relational kernel defined in the main text. Item (2) follows as a consequence of assumption 5 which bounds the degree of each node by a finite constant, c. As a result, any path predicate which defines a finite length path will return a set no larger than $c < c' < \infty$. As a result, so long as the initial random variable is weakly dependent, the relational variable constructed from the initial random variable will also be weakly-dependent, albeit with a slower rate of convergence since the coefficient τ_r (the weak dependence coefficient) will necessarily decay more slowly. □

B EXTENSION TO MULTI-RELATIONAL SYSTEMS

In our problem definition we assumed a single-entity, single relationship relational schema for ease of exposition. Here, we discuss necessary extensions for a multiple entity, multi-relational system. We consider a set of item classes $\boldsymbol{\mathcal{I}}$ to be the union of entities and relationship classes, $\boldsymbol{\mathcal{I}} = \boldsymbol{\mathcal{E}} \cup \boldsymbol{\mathcal{R}}$, following prior work [Lee and Honavar, 2017, Maier et al., 2013]. We refer to the attribute class of an item class $I \in \boldsymbol{\mathcal{I}}$ as $\boldsymbol{\mathcal{A}}(I)$. Moreover, let $G(I)$ denote a set of items of an item class $I \in \boldsymbol{\mathcal{I}}$.

Here, we point out two major differences in a multi-relational system:

1. The relational dependence is specifically defined between two item classes $I \in \boldsymbol{\mathcal{I}}$ and $I \in \boldsymbol{\mathcal{J}}$.

2. The path predicate ρ is likely to be defined with relational queries rather than random walks over a neighborhood.

Now, we revisit definition 1 from the main text with the new notation as follows:

Definition 1 (Relational Variable)**.** *Given a relational schema $\mathcal{S} = \langle \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}} \rangle$, its instantiation G, two item classes $I, J \in \boldsymbol{\mathcal{I}}$ and a path predicate ρ, a relational variable $\sigma(v_i, \boldsymbol{X}, G, \rho)$ is the set of attributes $v_j.\boldsymbol{X}$ selected by ρ of items $v_j \in G(J)$ reachable from items $v_i \in G(I)$ such that $\boldsymbol{X} \subset \boldsymbol{\mathcal{A}}(J)$, where the path predicate ρ is a function given by:*

$$\rho(v_i, G) : G(I) \mapsto \mathcal{P}(G(J))$$

The necessary assumptions and relational dependence definitions still hold. The major difference arises in the compact representation of the relational kernel. Equation 1 stays valid with an updated notion of path predicate. However, the compact representation in equation 2 is no longer trivial since the adjacency matrix A is no longer directly applicable. There are two potential workarounds. First, since the compact representation is not mandatory for our method to work, we can still work with equation 1 for multi-relational systems. Second, we can essentially consider the bipartite graph between sets of items between item classes $I, J \in \boldsymbol{\mathcal{I}}$ and use the adjacency matrix A_{IJ} of this bipartite graph instead of A. Similarly a corresponding degree matrix D_{IJ} can be constructed from A_{IJ}.

C EXPERIMENTS

C.1 SYNTHETIC ATTRIBUTE GENERATION

Here, we describe the synthetic attribute generation procedure for the three cases mentioned in the main text. Note that, only the generation of $v_i.Y$ differs in null and alternate hypothesis while others stay the same. We consider polynomial dependency model for most of our experiments. $v_i.X$ for case 1 and $v_i.Z$ for cases 2,3 is drawn from a uniform distribution $U(0, 1)$ while $v_i.X$ is always *binarized* to resemble the effect of treatment assignment. The outcome $v_i.Y$ is generated according to the following equation for marginal dependence (case 1):

$$v_i.Y \sim \begin{cases} U(0,1) & null \\ \beta_d \cdot (g(\sigma_x(v_i)))^2 + \epsilon & alternate \end{cases} \tag{3}$$

Conditional dependence (case 2) is reflected by the following equation:

$$\begin{aligned} v_i.X &\sim \beta_c \cdot (v_i.Z)^2 + \epsilon \\ v_i.Y &\sim \begin{cases} \beta_c \cdot (v_i.Z)^2 + \epsilon & null \\ \beta_d \cdot (g(\sigma_X(v_i)))^2 + \beta_c \cdot (v_i.Z)^2 + \epsilon & alternate \end{cases} \end{aligned} \tag{4}$$

Here, β_d and β_c are dependence and confounding coefficients respectively. β_c is considered 1.0 in our experiments. ϵ is noise drawn from standard normal ($N(0, 1)$) distribution. g refers to the *mean* aggregate function. We can get the generating function for case 3 by replacing $g(\sigma_X(v_i))$ and $v_i.Z$ with $v_i.X$ and $g(\sigma_Z(v_i))$ respectively in equation 4. Next, we consider the following procedure to simulate linear threshold model for the diffusion experiment which falls under case 1:

$$\begin{aligned} T_i &\sim U(0,1) \\ v_i.x_{t+1} &= \mathbb{1}(mean(\sigma_{x_t}(v_i)) > T_i) \\ v_i.y_{t+1} &= \mathbb{1}(g(\sigma_{x_t}(v_i)) > T_i) \end{aligned} \tag{5}$$

where we reassign $v_i.x$ values to simulate each diffusion step based on its value in previous step. The $v_i.y$ values are assigned based on $v_i.x$ values in the last diffusion step.

C.2 IMPACT OF VARIED NOISE VARIANCE

We conducted an experiment where we draw noise variance from a normal distribution $\sigma^2 \sim N(1, 0.2)$ over different trials. From figure 5 we can see a slight change of type-II errors compared to Figure 1 in the main paper. However, the trend seems to be very similar.

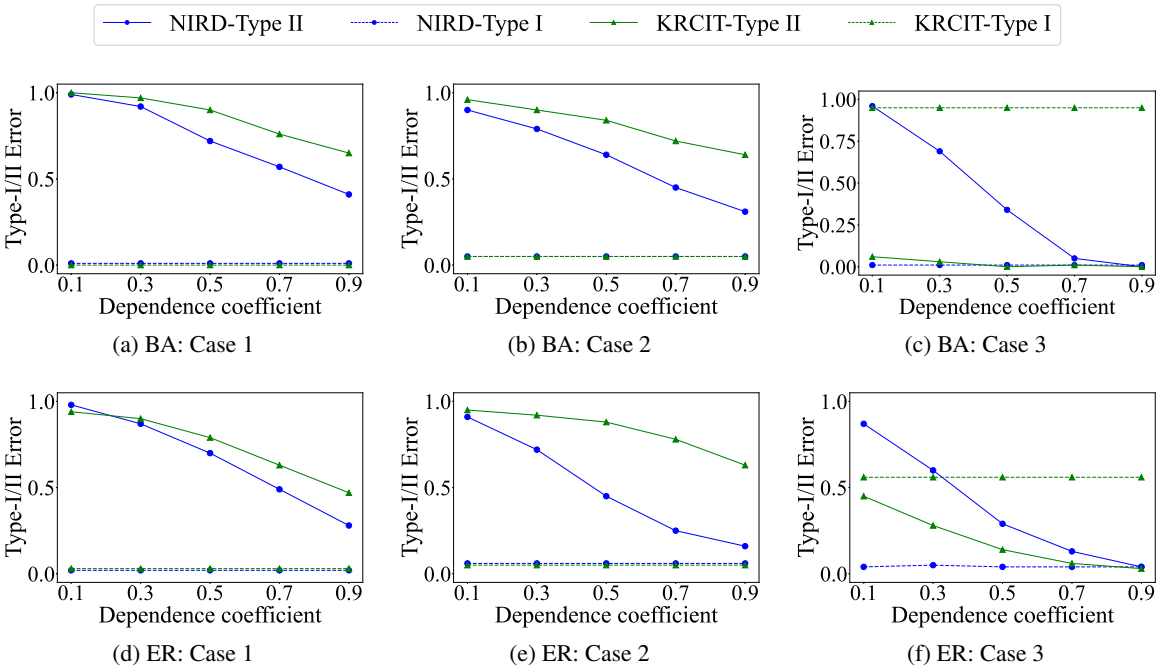


Figure 5: Relational dependence impact on Type I/II errors while variance of noise varied $\sim \mathcal{N}(1, 0.2)$ over multiple trials.

C.3 IMPACT OF ACTIVATION PROBABILITY ON DIFFUSION

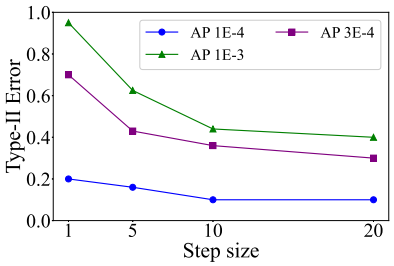
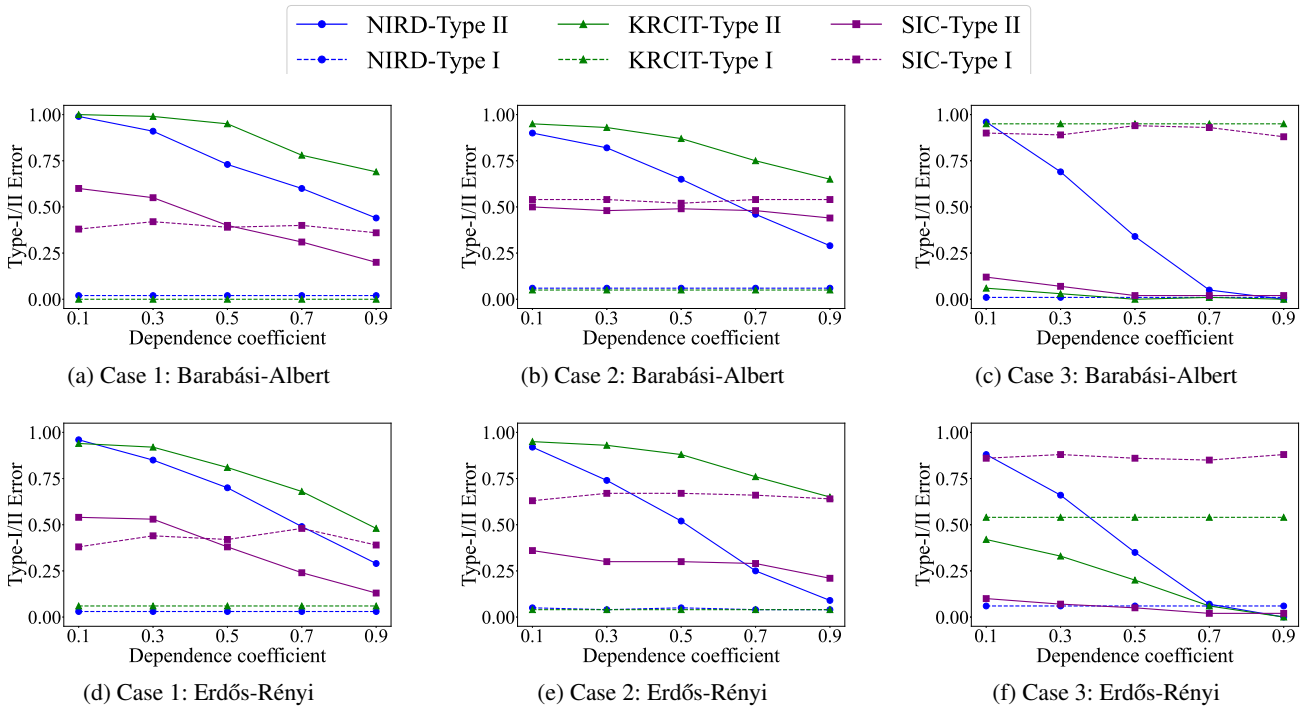


Figure 6: Type II error for the Linear Threshold Model on Twitter ego-network.

In order to showcase the applicability of the proposed method on large scale real world relational data, we show an extended version of the diffusion experiment from the main paper. We consider a similar semi-synthetic setup with Twitter ego-network which is a larger real world network consisting 11,176 nodes and 1,44,653 edges [Leskovec and Mcauley, 2012]. We consider a sample of 10,000 nodes and vary the initial activation probabilities. Figure 6 shows the Type-II errors (y-axis) for different diffusion step sizes (x-axis). The lines correspond to the initial activation probabilities (AP) for the diffusion process. We see the general trend of decreasing Type-II error with higher step sizes. It seems to be almost saturated with step 10. Moreover, the result indicates that the test is sensitive to activation probabilities and with higher activation probability, it shows higher type II error.

C.4 COMPARISON TO SOBOLEV INDEPENDENCE CRITERION (SIC)

To show the effectiveness of relational CI methods vs. CI methods developed for i.i.d. data, we compare both RCI methods (NIRD, KRCIT) to a recent i.i.d. CI test, the Sobolov Independence Criterion [Mroueh et al., 2019]. SIC is an interpretable dependency measure between multivariate random variables characterized by integral probability metric between the joint distribution and the product of the marginals. We perform the SIC test on the flattened representation of the relational data, similar to KRCIT. Figure 7 extends the results shown in Figure 1 in the main text. In all three cases, the i.i.d. baseline SIC exhibits high Type I error which shows its poor calibration to reasoning over the relational data.



(a) Case 1: Barabási-Albert (b) Case 2: Barabási-Albert (c) Case 3: Barabási-Albert

(d) Case 1: Erdős-Rényi (e) Case 2: Erdős-Rényi (f) Case 3: Erdős-Rényi

Figure 7: Type I/II errors with polynomial dependency model on synthetic networks for all three cases.

D REAL-WORLD DEMONSTRATION

One of the main challenges in social studies is to identify the effect of friends on their peers and the strength of such effects in different domains, e.g. health and violence. Studies show that patterns of interactions among adolescents can reveal possible reasons for changes in their behavior over time. The central question in such studies is how to identify and measure the existence of such effects. The proposed independence test can facilitate reasoning over the existence of dependence between peers' behaviors in social networks by providing a mechanism for falsifying statistical hypotheses.

As a demonstration, we examine the 50 Women dataset [Michell and Amos, 1997]. This dataset has the smoking, sport, drug, alcohol consumption habits of 50 female students, along with their friendship information, over the course of three years. Each of the behavioral variables are coded as categorical variables indicating how regularly women engage in each of the behaviors. Assuming independence between the behavior peers as the null hypothesis, the goal of this analysis is to explore whether the habits of a student's friends are associated with her habits in subsequent years.

Table 1 shows p-values estimated by our kernel test method considering four attributes in 50 Women dataset. We use column $Period$ to indicate the years we consider for the test, e.g., in $Period$ $1 \rightarrow 2, 3$, we explore students' behavior change from first year to the second and third year. We consider both the original categorical coding and a binarization of the categorical attributes, which is 1 if the student uses a substance at least once during the year and 0 otherwise. The number of students who did not engage in the behavior during the first time point is shown in column $t0$, e.g, in the first row of the table, 4 students did not drink alcohol in the first year. We exclude $t0$ for categorical data (indicated by NA) because the frequency of the habit is intrinsic to the hypothesis of interest in these cases. The last two columns ($NIDR\_all$ and $NIDR\_t0$) show p-values measured by NIRD. In $NIDR\_all$ and $NIDR\_t0$ we consider all women (whether they have the habit in a year or not), and women who do not have the habit in the first time point, respectively. Overall we find:

- Sports activity of peers is not associated with whether a student plays a sport or not. High values of $NIDR\_t0$ and $NIDR\_all$ are enough evidences to accept the null hypothesis of independence.
- Peer smoking habits are associated with students' frequency of smoking: $NIDR\_all = 0$ and $NIDR\_t0 < 0.022$ for all time periods, except period $2 \rightarrow 3$ where $NIDR\_t0 \approx 0.28$.
- Peer drug use is not associated with subsequent drug use in previously non-drug using students ($NIDR\_t0 > 0.05$). However, when we consider the effect of drug users on non-drug users and vice versa, it becomes associated in both the

Table 1: Real-world demonstration: exploration of the dependence between the habits of students and their first-hop neighbors in 50 Women dataset

period	attribute	attribute type	t0	NIRD_all	NIRD_t0
$1 \rightarrow 2$	alcohol	binary	4	0.425532	0.000000
$1 \rightarrow 2$	alcohol	categorical	NA	0.000000	NA
$1 \rightarrow 2$	drug	binary	35	0.000000	0.138298
$1 \rightarrow 2$	drug	categorical	NA	0.000000	NA
$1 \rightarrow 2$	smoke	binary	35	0.000000	0.021277
$1 \rightarrow 2$	smoke	categorical	NA	0.000000	NA
$1 \rightarrow 2$	sport	binary	12	0.925532	0.978723
$1 \rightarrow 2$	sport	categorical	NA	0.925532	NA
$1 \rightarrow 2,3$	alcohol	binary	5	0.114583	0.197917
$1 \rightarrow 2,3$	alcohol	categorical	NA	0.000000	NA
$1 \rightarrow 2,3$	drug	binary	35	0.000000	0.583333
$1 \rightarrow 2,3$	drug	categorical	NA	0.000000	NA
$1 \rightarrow 2,3$	smoke	binary	36	0.000000	0.000000
$1 \rightarrow 2,3$	smoke	categorical	NA	0.000000	NA
$1 \rightarrow 2,3$	sport	binary	12	1.000000	0.166667
$1 \rightarrow 2,3$	sport	categorical	NA	1.000000	NA
$2 \rightarrow 3$	alcohol	binary	3	0.125000	0.666667
$2 \rightarrow 3$	alcohol	categorical	NA	0.281250	NA
$2 \rightarrow 3$	drug	binary	32	0.000000	0.125000
$2 \rightarrow 3$	drug	categorical	NA	0.000000	NA
$2 \rightarrow 3$	smoke	binary	31	0.000000	0.281250
$2 \rightarrow 3$	smoke	categorical	NA	0.000000	NA
$2 \rightarrow 3$	sport	binary	20	0.864583	0.479167
$2 \rightarrow 3$	sport	categorical	NA	0.864583	NA

use and rate of consumption ($NIDR\_all = 0$).

- Peer alcohol consumption is associated with the level subsequent alcohol use ($NIDR\_all = 0$, except in period $2 \rightarrow 3$ where $NIDR\_all > 0.1$), but not with the decision for a non-drinking student tot begin drinking.

Different studies [Michell and Amos, 1997, Pearson and Michell, 2000] deploy 50 women data to explore the association between gender, risk-taking or social position and smoking or drug usage in groups of youngsters. In particular our results comport with Pearson et al. [Pearson and Michell, 2000] who show that drug usage and smoking are contagious among group of friends who are highly connected and people who are loosely connected to a friendship group.

References

K. P. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *NIPS*, 2014.

Herold Dehling, Olivier Durieu, and Dalibor Volny. New techniques for empirical processes of dependent data. *Stochastic Processes and their Applications*, 119(10):3699–3718, 2009.

S. Lee and V. Honavar. A kernel conditional independence test for relational data. 2017. UAI.

J. Leskovec and J. J. Mcauley. Learning to discover social circles in ego networks. In *NIPS*. 2012.

Anne Leucht and Michael H Neumann. Dependent wild bootstrap for degenerate u-and v-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.

M. Maier, K. Marazopoulou, D. Arbour, and D. Jensen. A sound and complete algorithm for learning causal models from relational data. In *UAI*, 2013.

L. Michell and A. Amos. Girls, pecking order and smoking. *Social Science & Medicine*, 44(12):1861 – 1869, 1997. ISSN 0277-9536. doi: https://doi.org/10.1016/S0277-9536(96)00295-X. URL http://www.sciencedirect.com/science/article/pii/S027795369600295X.

Y. Mroueh, T. Sercu, M. Rigotti, I. Padhi, and C. Nogueira dos Santos. Sobolev independence criterion. *NeurIPS*, 2019.

M. Pearson and L. Michell. Smoke rings: social network analysis of friendship groups, smoking and drug-taking. *Drugs: Education, Prevention and Policy*, 7(1):21–37, 2000. doi: 10.1080/dep.7.1.21.37. URL https://doi.org/10.1080/dep.7.1.21.37.

Hongwei Sun. Mercer theorem for rkhs on noncompact sets. *Journal of Complexity*, 21(3):337–349, 2005.