Non-Parametric Inference of Relational Dependence

Abstract

Independence testing plays a central role in statistical and causal inference from observational data. Standard independence tests assume that the data samples are independent and identically distributed (i.i.d.) but that assumption is violated in many real-world datasets and applications centered on relational systems. This work examines the problem of estimating independence in data drawn from relational systems by defining sufficient representations for the sets of observations influencing individual instances. Specifically, we define marginal and conditional independence tests for relational data by considering the kernel mean embedding as a flexible aggregation function for relational variables. We propose a consistent, non-parametric, scalable kernel test to operationalize the relational independence test for non-i.i.d. observational data under a set of structural assumptions. We empirically evaluate our proposed method on a variety of synthetic and semi-synthetic networks and demonstrate its effectiveness compared to state-of-the-art kernel-based independence tests.

1 INTRODUCTION

Measuring dependence is a fundamental task in statistics. However, most existing independence tests assume that the observed data is independent and identically distributed (i.i.d.). This assumption makes them unsuitable for capturing statistical dependencies in real-world relational systems, from social networks to protein-protein interactions, in which data instances depend on each other.

Relational dependence refers to a statistical dependence, marginal or conditional, between random variables in which at least one of the variables is a relational variable [Lee and Honavar, 2017]. A relational variable is a set of random variables that belong to instances related to an instance of interest, such as friends of a person or proteins interacting with a target protein. Relational dependence testing is central to social influence studies and causal discovery in relational systems, yet there is no standard statistical tool for inferring different forms of dependence from observational relational data. In this paper, we present a practical tool for determining marginal and conditional independence between relational variables with consistency guarantees.

To understand the challenge of estimating relational dependence, let's consider the following example:

Example 1. *Sally doesn't smoke but she has some friends who smoke. Sally starts to smoke over time. Are the smoking habits of Sally's friends and Sally's decision to take up smoking independent?*

Part of the problem of detecting this dependence is that we need to know a priori the exact mechanism of dependence. Is it because all of Sally's friends share a certain behavior? Or is there a minimum threshold of friends necessary (e.g., half of her friends smoke) to activate the dependence? The nature of the relationships among her friends might play a role as well. Without prior knowledge of this mechanism of dependence, any existing statistical test is likely to fail. As such, the focus of our work is to develop a flexible non-parametric test that can capture multiple forms of dependence. Note that for the scope of this work, we do not focus on the causal aspects of this question (e.g., social influence) but the test that we propose is applicable to causal discovery as well.

Recent studies have proposed non-parametric independence tests designed for non-i.i.d data. [Flaxman et al., 2015, Lee and Honavar, 2017]. Flaxman et al. [2015] develop a test between propositional variables accounting for latent homophily in a grid network but do not consider relational variables. Lee and Honavar [2017] develop a test for conditional independence (KRCIT) between relational variables which is the current state-of-the-art test for relational dependence. They operationalize the test by flattening the relational data

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

into a single, propositional table. However, the current state-of-the-art test has three key limitations. First, it requires practitioners to make explicit assumptions about the data generating processes and to specify an aggregation function over the relational variable a priori. Second, existing tests rely on propositionalization, which refers to the process of projecting connected data to a single, propositional table, which raises statistical concerns Maier et al. [2013b]. Third, it is computationally expensive and inapplicable to large relational datasets.

In this work, we focus on developing a general definition of relational dependence and a statistical test, NIRD (Non-parametric inference of relational dependence), that is able to capture a family of aggregate functions for characterizing relational dependence. The contributions of our work are in providing 1) complete definitions of marginal and conditional relational dependence which extend the classical definition of Daudin [1980], 2) independence tests with consistency guarantees, 3) the explicit representation of neighboring sets through kernel mean embeddings, and 4) a scalable test through random Fourier feature formulation which makes the test practical for real-world applications. We compare our proposed test to KRCIT on a variety of synthetic networks and simulate several social network characteristics, such as structure, density, and size. We demonstrate the applicability of our method for detecting peer influence using both semi-synthetic and real-world social networks.

2 RELATED WORK

There are several independence tests from observational data. Partial correlation is used for the independence test on Gaussian variables with linear dependence [Baba et al., 2004]. A more advanced test, the Hilbert-Schmidt independence criterion (HSIC) [Gretton et al., 2005] is a statistical test that has been extended to marginal independence testing for structured data [Zhang et al., 2009] and random processes [Chwialkowski et al., 2014]. Flaxman et al. [2015] utilize HSIC to develop marginal and conditional independence tests for propositional variables in the presence of a latent confounder in a single-entity network with an additive noise generating function. Relational dependence exhibits itself through a pairwise edge in the network which is a noisy surrogate of latent homophily. Lee and Honavar [2017] extend these tests to explicitly include relational variables. They operationalize the tests by considering pairwise dependence between each member of the relational variable set in the flattened data. However, existing tests require prior domain knowledge in order to detect complex dependencies between relational variables such as entropy, variance or reaching a threshold of peers' actions by requiring the specification of aggregation functions. We lessen this requirement by employing kernel mean embeddings, which provide a non-parametric aggregation capable of capturing a large number of aspects of smooth functions over distributions belonging to the exponential family.

Relational independence testing is central to causal discovery and structure learning of relational models. However, existing relational causal discovery algorithms either rely on the existence of a relational independence oracle [Maier et al., 2010, 2013a, Lee and Honavar, 2016] or use conditional independence tests developed for i.i.d. data [Lee and Honavar, 2019], making them less practical for real-world applications. The relational dependence tests in this work together with existing tests for detecting causal direction of relational dependence [Arbour et al., 2016b] can make relational causal discovery achievable for real-world datasets with unknown dependence functions.

One of the applications of relational dependence testing is in characterizing social influence in observational relational data. Existing social influence studies measure influence through predictive models [Christakis and Fowler, 2007, Bakshy et al., 2011], or randomized controlled trials [Muchnik et al., 2013, Su et al., 2020]. A standard statistical tool to infer different forms of social influence from observational relational data does not exist. Our work can help develop such tools by treating social influence detection as an instance of a relational independence test.

3 PROBLEM DEFINITION

The central task of this paper is to develop a statistical test for determining marginal and conditional independence between relational variables. Before formalizing the problem, we introduce the necessary notations and definitions. We denote random variables and their realizations with uppercase and lowercase letters, respectively, and bold to denote sets. We use an Entity-Relationship model [Heckerman et al., 2007] to describe relational data following previous work [Maier et al., 2010, Lee and Honavar, 2017]. A relational schema $\mathcal{S} = \langle \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}} \rangle$ represents a relational domain where $\boldsymbol{\mathcal{E}}$, $\boldsymbol{\mathcal{R}}$ and $\boldsymbol{\mathcal{A}}$ refer to the set of entity, relationship and attribute classes respectively. We consider an undirected graph G to be the instantiation of the relational schema $\mathcal{S}$ where nodes and edges refer to the entity and relationship instances respectively. A given schema can entail numerous possible instantiations. For ease of exposition, we focus on a schema with a single entity type (e.g., Person) and a single relationship type (e.g., Friends) but discuss the extensions necessary to address multi-relational data in the Appendix. We refer to the set of vertices, edges and adjacency matrix of G as $\boldsymbol{V}$, $\boldsymbol{E}$ and A respectively. For example, nodes $v_1 = Ana$ and $v_2 = Bob$ with undirected edge between them can come from the $Person - Friends - Person$ relations. The attribute class $\boldsymbol{\mathcal{A}}$ represents the set of possible attributes for the specified entity of the given graph G. Let $v_i.X$ and $v_i.x$ refer to the attribute $X \in \boldsymbol{\mathcal{A}}$ and its realization respectively for node $v_i \in \boldsymbol{V}$ corresponding

to instance $i \in \boldsymbol{\mathcal{E}} \cup \boldsymbol{\mathcal{R}}$. Here, $v_i.X$ is considered to be a propositional random variable. Following prior work we define a relational variable as a set of propositional random variables [Maier et al., 2013a, Lee and Honavar, 2017].

Definition 1 (Relational Variable)**.** *Given a relational schema $\mathcal{S} = \langle \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}} \rangle$, its instantiation G 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 nodes $v_j \in \mathbf{V}$ reachable from $v_i \in \mathbf{V}$ such that $\boldsymbol{X} \subseteq \boldsymbol{\mathcal{A}}$, where the path predicate ρ is a function given by: $\rho(v_i, G) : \mathbf{V} \mapsto \mathcal{P}(\mathbf{V})$.*

Here, $\mathcal{P}(\mathbf{V})$ refers to the power set of $\mathbf{V}$. An example path predicate is $\rho(v_i, G) = \{v_j | v_j \in \hat{\mathcal{N}}(v_i)\}$ where $\hat{\mathcal{N}}(v_i)$ refers to the direct neighbors of v_i in G and the corresponding relational variable $\sigma(v_i, \boldsymbol{X}, G, \rho)$ refers to the set of attributes of the neighboring nodes. For simplicity and assuming only direct neighbors throughout the paper, we denote the relational variable corresponding to attribute X by $\sigma_X(v_i)$ and its value by $\sigma_x^{v_i}$. Note that $\sigma_X(v_i)$ can represent a *propositional* variable as a special case. For example, $\sigma_X(v_i) = \{v_i.X\}$ refers to the X attribute of v_i. We also make the following assumptions:

A 1. *Each node $v \in V$ has degree of at least 1.*

A 2. *The adjacency matrix of G is symmetric with edge weights bounded by some real constant.*

A 3. *Dependence between two instances i and j implies the existence of a path in the graph between v_i and v_j.*

Relational dependence refers to a statistical dependence, either marginal or conditional, between two variables where at least one of the variables is relational. The goal of a relational dependence test is to determine whether to reject the null hypothesis of independence between these variables or not. The representation of relational data for such a test is non-trivial because data instances are not i.i.d. A common practice to deal with relational data is *propositionalization* [Kramer et al., 2001], which refers to the process of projecting a set of connected data samples down to a single, propositional table. In the context of relational dependence testing, flattening has three main deficiencies. First, the entities in the flattened data are not i.i.d. Second, choosing the appropriate aggregation function is non-trivial as discussed in section 1. Failing to appropriately define the aggregate in this case could lead to increased type I errors in marginal tests, and both type I and II errors for conditional tests. Third, flattening raises statistical concerns for relational causal discovery, one of the application areas of relational conditional independence tests, by violating the causal Markov condition [Maier et al., 2013b]. Lee and Honavar [2017] address the first deficiency by proposing a solution framework based on graph kernels using an existing i.i.d. kernel-based CI test method. However, their approach does not directly address the other two concerns.

Let's look at the problem with a concrete example. We consider an entity Person which exhibits attributes such as smoking status before ($S0$) and after ($S1$) a given time threshold t and G represents the network of social ties. Detecting the dependence of peers on a person's smoking behavior can be formalized as an independence test. For example, detecting whether one's smoking behaviour is marginally independent of one's direct friends' smoking behaviour could be carried out by a marginal test of $v_i.S1 \perp\!\!\!\perp \sigma_{S0}(v_i)$. Similarly, a conditional test of $v_i.S1 \perp\!\!\!\perp \sigma_{S0}(v_i) | v_i.S0$ should detect whether one's current smoking behavior is independent of friends' old smoking behavior given one's old smoking behavior.

In this work, we propose a relational dependence test which captures complex dependencies between relational variables without relying on flattening or explicit aggregate representations. We extend the definition of conditional independence for non-parametric functions by Daudin [1980] and propose the following definitions of marginal and *relational* conditional independence:

Definition 2 (Relational Marginal Independence)**.** *Two relational variables, $\sigma_X(v_i)$ and $\sigma_Y(v_i)$ are said to be marginally independent of each other for any smooth square measurable functions $g_X(\cdot), g_Y(\cdot)$ if and only if, $\mathbb{E}\left[g_X(\sigma_X(v_i)) g_Y(\sigma_Y(v_i))\right] = \mathbb{E}\left[g_X(\sigma_X(v_i))\right] \mathbb{E}\left[g_Y(\sigma_Y(v_i))\right]$.*

Definition 3 (Relational Conditional Independence)**.** *Two relational variables, $\sigma_X(v_i)$ and $\sigma_Y(v_i)$ are said to be independent of each other given a third, $\sigma_Z(v_i)$ for any smooth square measurable functions $g_X(\cdot), g_Y(\cdot), g_Z(\cdot)$ if and only if, $\mathbb{E}\left[g_X(\sigma_X(v_i)) g_Y(\sigma_Y(v_i)) | g_Z(\sigma_Z(v_i))\right] = \mathbb{E}\left[g_X(\sigma_X(v_i)) | g_Z(\sigma_Z(v_i))\right] \mathbb{E}\left[g_Y(\sigma_Y(v_i)) | g_Z(\sigma_Z(v_i))\right]$.*

Here, $g_X(\cdot), g_Y(\cdot), g_Z(\cdot)$ are *aggregate* functions that map σ to a real-valued vector. They could be *sum*, *mean* or any other complex non-linear function. The rejection of the null hypothesis of marginal independence would mean that the variables are possibly dependent, either due to a directed path between them or due to a direct, causal relationship, or the presence of a confounding relationship. For a relational conditional independence (RCI) test, the rejection of the null hypothesis would imply that the two variables are not independent given the conditioning set. Note that because we are considering the dependence between sets of relational variables and their propositional counterparts we circumvent the three problems with flattening described earlier.

4 RELATIONAL INDEPENDENCE TESTS

In this section, we discuss the components which operationalize the definition of relational dependence into an empirical test. We first describe a non-parametric relational aggregate formed by local kernel means. Then we formulate marginal and conditional independence tests using the

kernel mean embedding. Then, we discuss the theoretical boundaries for the consistency of the proposed test. Finally, we introduce techniques for large-scale approximation of the proposed relational kernels that can speed up the independence test significantly.

4.1 NON-PARAMETRIC AGGREGATE REPRESENTATIONS

One of the central problems in estimating dependence in relational settings is defining a sufficient representation for the sets of observations for individual instances of a relational variable. Prior work [Maier et al., 2013a, Arbour et al., 2016a, Lee and Honavar, 2017] considered aggregation functions, usually one, which are specified *apriori* by the practitioner. However, in many scenarios, it is unreasonable to expect practitioners to reason over a very complex joint distribution or to know the exact parametric form of dependence. For example, the possible aggregation in effect for the spread of obesity in social networks [Christakis and Fowler, 2007] can be different from people's influence on the Twitter platform [Bakshy et al., 2011]. A generalized definition and associated operationalization of relational dependence can help the practitioner by directly measuring dependence without prior domain knowledge about aggregations on the given relational system.

The distance between the embedding of joint distribution and embedding of product of marginals can be used to infer independence according to definition 2, while avoiding explicit density estimation as an intermediate step. Note that the aggregate functions are represented implicitly through the kernel mean embedding (KME) [Smola et al., 2007, Muandet et al., 2017]. An appealing property of the kernel mean embedding is that if the kernel is universal then the kernel mean uniquely represents all moments for any member of the exponential family [Smola et al., 2007] [1].

Adopting the kernel mean as an aggregation function removes the burden of reasoning over parametric families and predefined aggregates. Specifically, the kernel mean embedding considers the mean of a variable after applying a projection $\phi(\cdot)$ into some RKHS, $\mu = \int \phi(x)p(x)dx$, with the corresponding empirical estimate of $\hat{\mu} = \frac{1}{N}\sum_i^N \phi(x_i)$ where N is the number of observations and $x_1, \ldots, x_N$ are observations from a random variable X [Smola et al., 2007].

We present the practical implementation of the kernel mean as a relational aggregate. For a given node v_i, we define the kernel mean aggregate of its neighbors with respect to the attribute X as $\mu(v_i) = \frac{1}{\deg(v_i)} \sum_{m \in \hat{\mathcal{N}}(v_i)} \phi(m.x)$

where $\hat{\mathcal{N}}(\cdot)$ refers to a path predicate which is restricted to immediate neighbors for ease of exposition. Because ϕ may map to an infinite dimension, it is impractical to explicitly represent this quantity. Fortunately, because our statistics of interest are concerned with the covariance, the kernel trick, i.e. considering the inner product rather than the feature representations directly, can be employed. Specifically, the inner product between relational kernel mean is given as

$$\langle \mu(v_i), \mu(v_j) \rangle = \frac{\sum_{m \in \hat{\mathcal{N}}(v_i)} \sum_{p \in \hat{\mathcal{N}}(v_j)} k(m.x, p.x)}{\deg(v_i)\deg(v_j)},$$

which can be written for an entire sample in terms of a matrix product between the network adjacency matrix, A, the inverse degree matrix D^{-1} where $D_{i,i} = \frac{1}{\deg(v_i)}$, and the kernel matrix K_X, by observing $(D^{-1}A\phi(\mathbf{x}))(D^{-1}A\phi(\mathbf{x}))^T = D^{-1}AK_XAD^{-1}$.

In contrast to the propositional kernel mean, the convergence of the relational to its population counterpart is not necessarily guaranteed because of sample dependence. We discuss convergence and consistency guarantees under the assumption of weak dependence after describing the relational independence tests.

4.2 RELATIONAL MARGINAL INDEPENDENCE TEST

With the relational kernel mean defined we now turn to the central task of this paper, non-parametric inference of relational dependence (NIRD). As a test statistic, we use the Hilbert-Schmidt independence criterion (HSIC) [Gretton et al., 2005]. HSIC measures the maximum distance between an embedding of the observed joint distribution, and the product of the marginals, i.e., $\|\mathbb{E}[\phi(x) \otimes \phi(y)] - \mathbb{E}[\phi(x)] \otimes \mathbb{E}[\phi(y)]\|^2$. We perform a hypothesis test using HSIC as the test statistic where the null hypothesis refers to independence. The test produces a p-value which is used to decide whether to reject the null or not. Testing relational independence using HSIC is straightforward with the relational kernel mean by using the kernel matrix defined earlier in the empirical HSIC estimator. Defining the centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, an empirical estimate of HSIC is given by $\frac{1}{n^2}\text{trace}\,(K_XHK_YH)$, where K_X and K_Y are kernel matrices corresponding to the random variables X and Y, respectively. Independence testing with HSIC can be performed by using the corresponding relational kernel in the test statistic.

4.3 RELATIONAL CONDITIONAL INDEPENDENCE TEST

A similar construction can be employed to test for relational conditional independence, defined in Definition 3. Following Strobl et al. [2019], we consider the following

[1] We refer readers to Szabó and Sriperumbudur [2017] for conditions for a kernel to be universal. Many popular kernels such as the RBF kernel are universal

L^2 spaces,

$$F_{XZ} \triangleq \left\{ \tilde{f} \in L^2_{XZ} \mid E(\tilde{f} \mid Z) = 0 \right\}$$
$$F_{YZ} \triangleq \left\{ \tilde{g} \in L^2_{YZ} \mid E(\tilde{g} \mid Z) = 0 \right\}$$
$$F_{Y \cdot Z} \triangleq \left\{ \tilde{h}' \mid \tilde{h}' = h'(Y) - E\left(h' \mid Z\right), h' \in L^2_Y \right\}$$

Each of these quantities can easily be constructed by considering regressions, e.g. $\tilde{f}$ can be obtained by taking the residuals after performing a regression. We consider a mean of the feature basis representation as an aggregation function whenever one of the variables is relational. Under the assumption that the direct sum of the reproducing kernel Hilbert spaces, $k_x k_y$ and k_z is dense in L_2, Strobl et al. [2019] (proposition 5) showed that conditional linear covariance of zero implies uncorrelatedness, i.e., $\mathbb{E}\left[\tilde{f}\tilde{g}\right] = 0 \implies X \perp\!\!\!\perp Y | Z \implies \Sigma_{XY|Z} = 0$. This motivates the use of a multiple output kernel ridge regression as an estimator of the conditional expectation, $\beta = (\phi(z)^T\phi(z) + \lambda I)^{-1}\phi(z)^t\phi(x)$. The test is then constructed by considering the residuals, $\widetilde{\phi(x)} = \phi(x) - \phi(z)\beta_{xz}, \widetilde{\phi(y)} = \phi(y) - \phi(z)\beta_{yz}$ and sum of the squared covariances between them.

There are two considerations in employing this procedure in a relational setting, namely how to handle relational variables in the conditioning set (Z) and the test set (X), respectively. When a member of the conditioning set is relational, the test procedure is identical after replacing $\phi(z)$ with its relational counterpart, $\frac{1}{|\hat{\mathcal{N}}(z)|}\sum_{m \in \hat{\mathcal{N}}(z)} \phi(m)$. When a member of the test set is relational, the problem is reduced to predicting each member of the set independently by considering the regression of the perspective of the relational variable, as described by Maier et al. [2013a]. After regressing individual members, the mean of residuals is then considered for the marginal tests, $\widetilde{\sigma(\phi(x))} = \frac{1}{|\hat{\mathcal{N}}(x)|}\sum_{m \in \hat{\mathcal{N}}(x)} \phi(m) - \phi(z)\beta_{mz}$.

4.4 CONSISTENCY OF RELATIONAL INDEPENDENCE TEST

In order to reason about the behavior of test statistics under non-i.i.d. samples and understand asymptotic behavior we need to characterize the behavior of dependence amongst instances as a function of some notion of distance between instances. There are a number of formalisms for reasoning about dependent data [Andrews and Pollard, 1994, Bickel and Bühlmann, 1999, Dedecker et al., 2007]. In this work we focus on weak dependence [Dedecker et al., 2007], which we describe next.

4.4.1 Weak Dependence

In order to accommodate dependent observations and maintain consistency of the testing procedure we will assume that observations are weakly dependent. Weak dependence provides a flexible notion of dependence that requires only the definition of distance between instances and the presence of a measurable probability space. Within this work we will make use of the notion of weak dependence, i.e. τ-dependence.

Definition 4. *[Dedecker et al., 2007] Let $\boldsymbol{\pi}$ be a filtration*[2] *over the set of nodes in a graph, G, defined by performing a breadth first search at an arbitrary node, $v \in G$. Further, define $\delta(i, j)$ to be the shortest path distance between nodes i and j and X to be a $\mathbb{L}^p$-integrable random variable. The* ***weak-dependence coefficient*** *is defined as $\tau_{p,r}(X) = \sup_{(i,j)} \| \sup_g Cov(g\left(X_{\boldsymbol{\pi}(i)}\right), g\left(X_{\boldsymbol{\pi}(j)}\right) \|_p$, where $i \leq j$ and $j - i \leq r$, and $g()$ is a Lipschitz function.*

Intuitively, the weak dependence coefficient, $\tau_{p,r}(X)$ measures the covariance between a vector, X_i and another random vector X_j drawn from the same process separated by at least distance of r. We call a process *weakly dependent* if τ tends to zero as the r tends to infinity. Note that this is a strictly weaker condition than alternative assumptions on dependence such as strong mixing and m-dependence which require independence at a a finite distance, whereas weak-dependence only requires it asymptotically.

4.4.2 Weak Dependence in Relational Domains

We provide a natural extension of weak dependence within the relational setting by replacing usual definition of distance to the shortest path distance between two nodes in a graph. The role of τ in this case can be interpreted as measuring the decay of dependence between instances as a function of shortest-path distance. We will assume from here out that as the distance between any two nodes in the network tends to infinity, the dependence between them converges to zero. More formally, we will employ the following assumption:

A 4. *$(X_t)_{t \in \pi}$ is a strictly stationary τ-dependent process with $\sum_{r=1}^{\infty} r^2 \sqrt{\tau_r(X)} \leq \infty$ for some filtration π, where r is shortest-path graph distance.*

The notion of weak dependence within the network setting is not novel to this work, Xiang and Neville [2011] make use of the τ-coefficient in the context of deriving asymptotic consistency for transductive learning with an assumption of linear dependence amongst instances. However, to our knowledge, our work is the first to consider weak dependence with arbitrary dependence for independence testing of relational data. Consistency of the relational independence testing is provided by the following theorem and corollary, after applying two additional assumptions.

[2] A filtration is an ordering of a set such that for any two subsets, $S_{1,\ldots,j}, S_{1,\ldots,k}, j \leq k \rightarrow S_{1,\ldots,j} \subseteq S_{1,\ldots,k}$.

A 5. *The maximum degree of any node in the network is bounded by a real constant.*

A 6. *The network structure is fixed and doesn't change during the generation of the observed random variables.*

Assumption 5 ensures that the average shortest path distance from any node to all other nodes in the graph tends to infinite as the number of nodes tends to infinite, which is necessary in order to have convergence of weakly dependent sequences. Assumption 6 ensures that the observed neighborhoods for nodes correspond to the structure which generated the data.

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., $|HSIC_n - HSIC_{population}| \underset{d}{\longrightarrow} 0$.*

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., $|HSIC_n - HSIC_{population}| \underset{d}{\longrightarrow} 0$.*

The proof of theorem 1, which is deferred to the supplement, follows by observing that the empirical estimate of HSIC is a degenerate V-statistic and then through a proof which shows consistency of degenerate V-statistics under weak-dependence in structured domains, which may be of independent interest. Similarly, the proof of the corollary, also deferred to the supplement follows from theorem 1 and showing that the weak dependence coefficient remains finite for relational variables.

Large Scale Approximations While the proposed model is theoretically appealing, the associated time and space complexity render it infeasible for most modern network settings. To address this, we appeal to an approximation of the kernels known as Random Fourier Features [Rahimi and Recht, 2008]. Random Fourier Features exploit Bochner's theorem, which states that a continuous, time-invariant kernel is positive definite if and only if the kernel is the Fourier transform of some non-negative measure. For example the Gaussian kernel can be represented with the following Fourier transformation $\hat{k}(\omega) = \frac{1}{2\pi}\int e^{-j\omega^\top \delta} k(\delta) d\delta$. This property implies that a kernel can be approximated via the following procedure:

- Draw D samples, from some distribution (i.e Normal), to approximate the Gaussian kernel where the variance σ corresponds to the bandwidth of the kernel.
- Construct the Fourier basis explicitly as $z(x) = \sqrt{\frac{2}{d}}\left[\cos\left(w_1^T x\right), \sin\left(w_1^T x\right), \ldots,\right]$.
- Perform linear operations using z.

Following [Zhang et al., 2018, Strobl et al., 2019], we approximate HSIC using random Fourier features by considering $\widehat{\text{HSIC}}(X,Y) = \left\|\frac{1}{n}\boldsymbol{Z}_X^T \boldsymbol{H} \boldsymbol{Z}_Y\right\|^2$ where Z is a $n \times d$ dimensional matrix with each row consisting of the random Fourier features for an observation. We can represent the relational kernel mean as $D^{-1}AZ$, and the corresponding test statistic as $\|\frac{1}{n} Z_X^T A D^{-1} H Z_Y\|$ where D and A are the diagonal degree and adjacency matrix as before. In several experiments we show that using approximate statistic leads to significant performance improvements with minimal effect on the efficacy of the test, even with only a few random features.

5 EXPERIMENTS

We run experiments with multiple network datasets, relational dependence cases, and synthetic attribute generators to evaluate the effectiveness of the proposed test.

5.1 NETWORK DATASETS

We consider networks from two synthetic graph generators and three non-PII real-world networks. First, for the Barabási-Albert (BA) model, we vary the parameter that controls the number of nodes a new node can attach to. For the Erdős-Rényi (ER) model, we vary the probability of edge creation between each pair of nodes. For each set of parameters, we generate 100 networks with size 100. The small size of the synthetic networks is driven by the baseline method which does not scale well, as shown in Figure 3a. We also demonstrate the applicability of our approach through a Facebook ego-network with $4,039$ nodes and $88,234$ edges [Leskovec and Mcauley, 2012]. The other two real-world datasets (*Twitter*, *50 Women*) and corresponding experimental results are described in the Appendix.

5.2 THREE CASES OF RELATIONAL DEPENDENCE

We choose three representative relational dependence cases to cover a range of possible tests. We consider attributes $Z, X, Y \in \boldsymbol{\mathcal{A}}$ which measure characteristics in time steps $t-1, t, t+1$ respectively. The three cases are represented with arrows showing the direction of dependence:

1. **Case 0:** $\sigma_X(v_i) \rightarrow v_i.Y$
2. **Case 1:** $\sigma_X(v_i) \leftarrow v_i.Z \rightarrow v_i.Y \leftarrow \sigma_X(v_i)$
3. **Case 2:** $v_i.X \leftarrow \sigma_Z(v_i) \rightarrow v_i.Y \leftarrow v_i.X$

where $\sigma_X(v_i)$ and $\sigma_Z(v_i)$ are relational variables on the attributes X and Z of the direct neighbors of v_i. Case 0 refers to marginal independence between a relational and a propositional variable ($\sigma_X(v_i) \perp\!\!\!\perp v_i.Y$). Cases 1

and 2 introduce conditional independence given a confounder. Case 1 refers to a propositional confounder ($\sigma_X(v_i) \perp\!\!\!\perp v_i.Y | v_i.Z$) whereas case 2 refers to a relational confounder ($v_i.X \perp\!\!\!\perp v_i.Y | \sigma_Z(v_i)$). A test should be able to reject the null hypothesis of no dependence in all cases. Note that direction is ignored in the test.

5.3 SYNTHETIC ATTRIBUTE GENERATION

The attribute values for the independent variables are drawn from the uniform distribution ($\sim U(0,1)$) while the noise is standard normal ($\sim \mathcal{N}(0,1)$). We consider polynomial dependence function with a mean aggregator for relational dependence in all cases except for the diffusion experiment where we consider linear dependence. The attribute generation is described in the Appendix.

5.4 EXPERIMENTAL SETUP

We empirically evaluate the proposed approach, NIRD, to the state-of-the-art RCI test method, KRCIT [Lee and Honavar, 2017]. We report the average Type I and Type II errors with significance level 0.05 over 100 trials for each set of parameters. We use *Radial Basis Function kernel* (RBF) as the base kernel. KRCIT is implemented with HSIC as the kernel-based marginal independence test method and KCIT [Zhang et al., 2011] as the kernel-based conditional independence (CI) test. We use the approximate method of NIRD in all experimental evaluation with 20 and 50 random Fourier features for marginal and conditional test respectively. We use a permuted bootstrap method for estimating the null distribution for both kinds of tests. We also compare both RCI methods (NIRD, KRCIT) to a recent i.i.d. CI test method, Sobolov Independence Criterion (SIC) [Mroueh et al., 2019] (see Appendix). We study NIRD's strengths and weaknesses in five experimental setups:

Relational dependence sensitivity: We evaluate the sensitivity of the dependence tests to different relational dependence strengths. We report results on polynomial models while varying the dependence coefficient in range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ for the alternate hypothesis and report both Type I, II errors. Edge connectivity is 3 for BA and edge probability is 0.02 for ER model.

Network sensitivity: We examine performance over a variety of network structures. We vary edge connectivity of BA in range $\{1, 2, 3\}$ and edge probability of ER in range $\{0.005, 0.01, 0.015, 0.02, 0.025\}$. We use a fixed dependence coefficient value of 0.5.

Diffusion: We apply NIRD to test for contagion by simulating a linear threshold model [Granovetter, 1978] with initial treatment probability of 0.1 on the semi-synthetic Facebook network. We reassign treatment values in each diffusion step and generate outcomes (Y) based on treatments (X) generated in the last diffusion step. The attribute generation process is described in the Appendix. We expect the distribution of $\sigma_X(v_i)$ to change with increasing diffusion steps and investigate at what step it is possible to detect relational dependence. We vary the number of steps, sample size and measure Type II error. We also investigate the impact of activation probability on the Type-II error on the Twitter ego-network with 10,000 nodes (results in the Appendix).

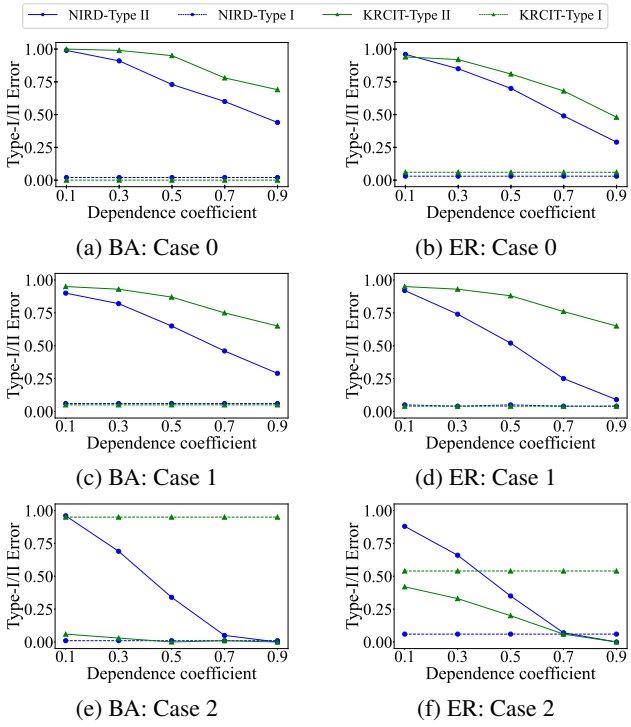
Scalability: To compare the scalability of NIRD against the baseline KRCIT, we generate Erdős-Rényi synthetic networks (edge probability 0.02) with varying number of nodes and report their execution time for marginal and conditional independence testing. We vary the network size (x-axis) in the range $\{100, 200, 300, 400, 500\}$. The choice of small networks was driven by the fact that KRCIT scales exponentially as shown in Figure 3a and it is impossible to run for larger networks. In the Appendix, we also demonstrate the scalability of NIRD through the diffusion experiment which is run on networks of size up to 10k nodes. The experiment is run on a 2.4GHz 8-core machine with 50GB memory.

Real world demonstration: In the Appendix, we demonstrate the applicability of our test for detecting peer influence in a well-studied real-world social network (*50 Women*) where our test discovers smoking-, drug- and sport-related peer dependencies that concur with previous research.

5.5 RESULTS

Relational dependence sensitivity. Figure 1 shows Type I and Type II errors for the polynomial dependency model on synthetic data. The rows correspond to the relational dependence cases and the columns to network models. The solid and dashed lines correspond to Type II and Type I errors respectively. The test is most challenging when the dependence coefficient, β_d (x-axis) is low. The figure shows that both RCI methods are well calibrated with low Type I error (max 0.06 by KRCIT in Erdős-Rényi) for the first two cases. In these cases, NIRD consistently produces lower Type II errors compared to KRCIT. It is most visible in Erdős-Rényi model (1d) with 86% reduction in Type II error for $\beta_d = 0.9$. The performance gain of NIRD increases slightly from case 0 to case 1 as the difficulty increases. In case 2, KRCIT is poorly calibrated and exhibits an unusually high Type I error. Across cases, NIRD shows desired behavior: it is consistently well-calibrated and its Type II error decreases with the increase of relational dependence.

Network sensitivity. Figure 2 shows Type I and Type II errors for two network models. The x-axis represents the corresponding parameter values for each model. We observe that increased parameter values exhibit higher Type II errors in general for Barabási-Albert model but not for Erdős-Rényi. A possible reason is that Barabási-Albert exhibits a more skewed degree distribution compared to Erdős-Rényi. Note

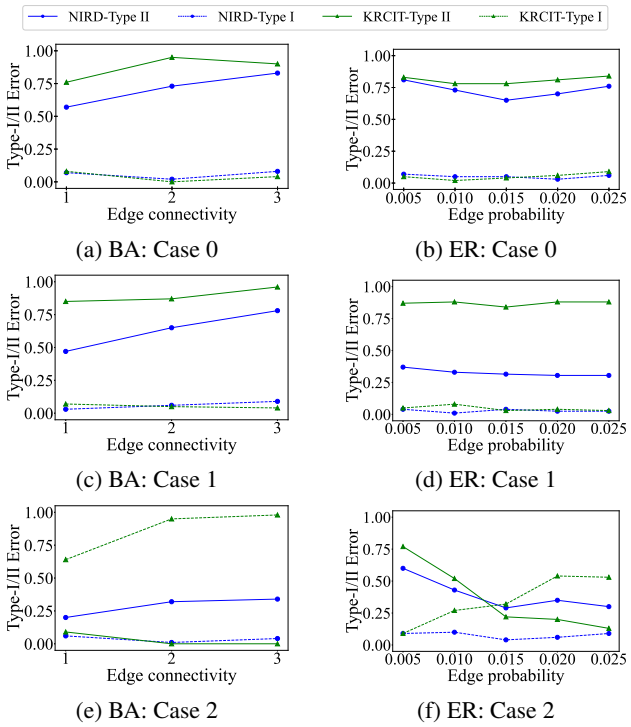


(a) BA: Case 0 (b) ER: Case 0

(c) BA: Case 1 (d) ER: Case 1

(e) BA: Case 2 (f) ER: Case 2

Figure 1: Relational dependence impact on Type I/II errors.



(a) BA: Case 0 (b) ER: Case 0

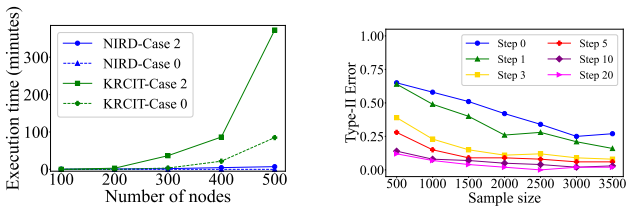
(c) BA: Case 1 (d) ER: Case 1

(e) BA: Case 2 (f) ER: Case 2

Figure 2: Impact of network parameters on Type I/II errors.

that the increased parameter values indicate higher density of the network. We expect Erdős-Rényi to show a similar trend if the edge probability is further increased. NIRD outperforms KRCIT in terms of Type II error (except in Figures 2e and 2f which is due to poor calibration of the baseline method) irrespective of network density. Type II error is reduced as high as 65% for Erdős-Rényi model with edge probability 0.025 (Figure 2d). Moreover, Type I error for NIRD is consistent whereas KRCIT suffers in case 2 (Figures 2e, 2f).

Diffusion. Figure 3b shows the impact of the number of diffusion steps (lines) and sample size (x-axis) on the effectiveness of NIRD. At initial activation (1 diffusion step) there is a high Type II error across sample sizes which decreases with higher number of steps. We see a significant decrease in error with just 5 diffusion steps. Further steps drastically lower the Type II error and at 20 steps and larger samples it can reject the null hypothesis consistently. This suggests that relational dependence is easier to detect after several diffusion steps rather than at early activation. It also demonstrates the effectiveness and scalability of NIRD in terms of detecting social phenomena in real world networks.



(a) Test scalability (b) Linear threshold model

Figure 3: Impact of sample size on exec. time and Type II error.

Note that it is computationally infeasible to run the baseline method on such size of samples.

Scalability. Figure 3a shows execution time in minutes (y-axis) for both marginal (case 0) and conditional (case 2) independence test for different network sizes in terms of number of nodes (x-axis). The solid and dashed lines represent the conditional and marginal test result respectively. KRCIT exhibits an exponential complexity whereas NIRD shows much less sensitivity to network size. This is expected given the complexity of the corresponding algorithms.

6 CONCLUSION

In this work we examine the problem of defining and measuring statistical dependence in relational data. We propose NIRD, a consistent, non-parametric test for detecting relational dependence that improves state-of-the-art relational dependence testing by capturing a wide range of possible relational dependencies. Moreover, we introduce an approximate method that makes NIRD scalable to larger networks. We evaluate the effectiveness of our method across diverse relational settings and find that our proposed test exhibits significantly less sensitivity to network properties and dependence types. Our work paves the way for a number of promising future research directions, from testing for social influence to causal structure learning from relational data.

References

Donald WK Andrews and David Pollard. An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review/Revue Internationale de Statistique*, pages 119–132, 1994.

D. Arbour, D. Garant, and D. Jensen. Inferring network effects in observational data. In *KDD*, 2016a.

D. Arbour, K. Marazopoulou, and D. Jensen. Inferring causal direction from relational data. In *UAI*, 2016b.

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, 2011.

Peter J Bickel and Peter Bühlmann. A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. *Bernoulli*, pages 413–446, 1999.

N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *NEJM*, 357(4): 370–379, 2007.

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

JJ Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980.

Jerome Dedecker, Paul Doukhan, and Gabriel Lang. *Weak Dependence: With Examples and Applications*. Springer, 2007.

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. R. Flaxman, D. B. Neill, and A. J. Smola. Gaussian processes for independence tests with non-iid data in causal inference. *TIST*, 7(2):1–23, 2015.

M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, 2005.

D. Heckerman, C. Meek, and D. Koller. Probabilistic entity-relationship models, prms, and plate models. *Introduction to statistical relational learning*, pages 201–238, 2007.

S. Kramer, N. Lavrač, and P. Flach. Propositionalization approaches to relational data mining. In *Relational data mining*, pages 262–291. Springer, 2001.

S. Lee and V. Honavar. On learning causal models from relational data. In *AAAI*, 2016.

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

S. Lee and V. Honavar. Towards robust relational causal discovery. In *UAI*, 2019.

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, B. Taylor, H. Oktay, and D. Jensen. Learning causal models of relational domains. In *AAAI*, 2010.

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

M. Maier, K. Marazopoulou, D. Arbour, and D. Jensen. Flattening network data for causal discovery: What could go wrong? In *WIN*, 2013b.

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.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10:1–144, 2017.

L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146), 2013.

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.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, 2008.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *ALT*, 2007.

E. V. Strobl, K. Zhang, and S. Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *JCI*, 7(1), 2019.

J. Su, K. Kamath, A. Sharma, J. Ugander, and S. Goel. An experimental study of structural diversity in social networks. In *ICWSM*, 2020.

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

Z. Szabó and B. K. Sriperumbudur. Characteristic and universal tensor product kernels. *JMLR*, 18(1):8724–8752, 2017.

Rongjing Xiang and Jennifer Neville. Relational learning with one network: An asymptotic analysis. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 779–788. JMLR Workshop and Conference Proceedings, 2011.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2011.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 113–130, 2018.

X. Zhang, L. Song, A. Gretton, and A. J. Smola. Kernel measures of independence for non-iid data. In *NIPS*, 2009.

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*[3].

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 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

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

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

Leucht and Neumann [2013] provide the prerequisites necessary for equation A.1 converges absolutely and uniform on compact subsets of 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

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

Where Δ_r is the set of nodes whose shortest path distance from X_0 is r, $\max(\text{deg})$ 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., $|HSIC_n - HSIC_{population}| \underset{d}{\longrightarrow} 0$.*

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.,* $|HSIC_n - HSIC_{population}| \xrightarrow[d]{} 0$.

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., 2013a]. 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 5 (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 0 and $v_i.Z$ for cases 1,2 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 0):

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

Conditional dependence (case 1) 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 2 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 0:

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

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 ACTIVATION PROBABILITY ON DIFFUSION

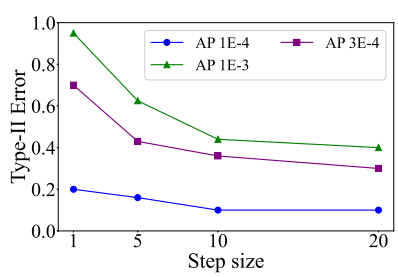


Figure 4: Type II error for the Linear Threshold Model on Twitter ego-network.

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 4 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.

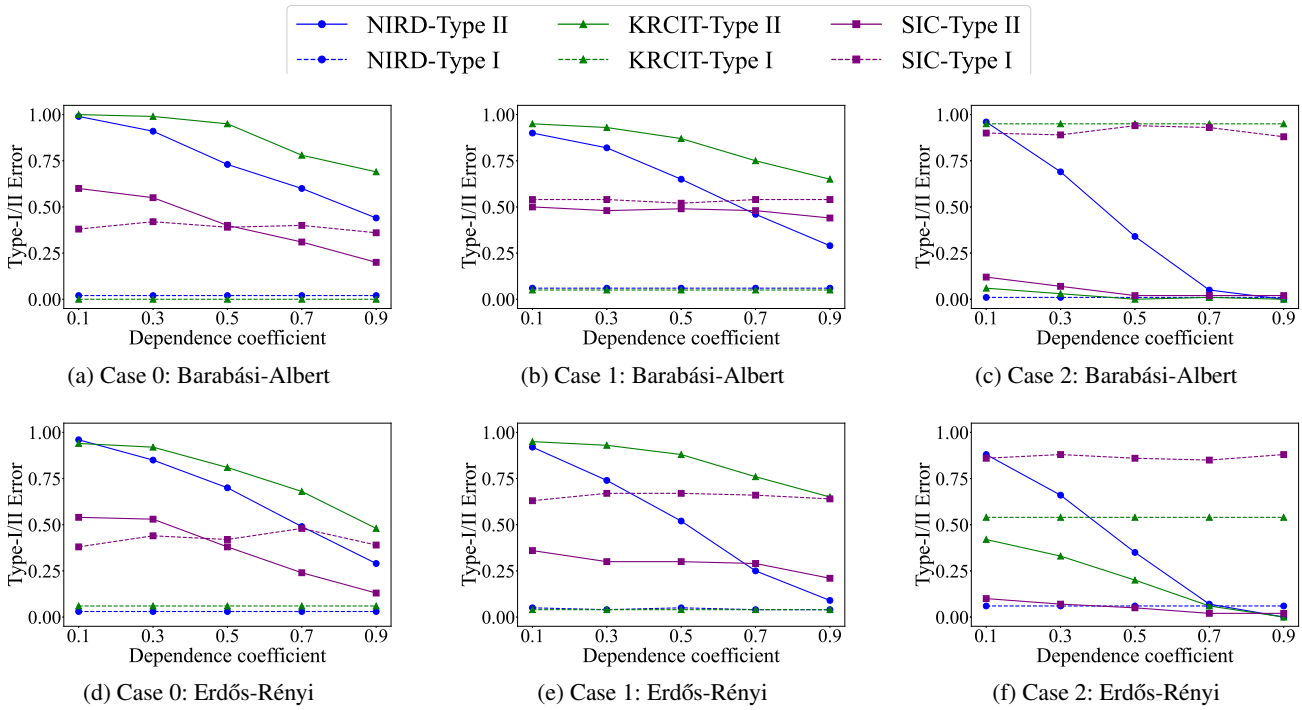


Figure 5: Type I/II errors with polynomial dependency model on synthetic networks for all three cases.

C.3 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 5 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.

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 $NIDR_t0$)

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

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 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.