Leveraging D-Separation for Relational Data Sets

Matthew J. H. Rattigan, David Jensen
Department of Computer Science
University of Massachusetts
Amherst, Masschusetts 01003

***Abstract*—Testing for marginal and conditional independence is a common task in machine learning and knowledge discovery applications. Prior work has demonstrated that conventional independence tests suffer from dramatically increased rates of Type I errors when naively applied to relational data. We use graphical models to specify the conditions under which these errors occur, and use those models to devise novel and accurate conditional independence tests.**

I. Introduction

Procedures for testing marginal and conditional independence are central to many algorithms for machine learning. For example, algorithms for learning the structure of Bayesian networks search over possible conditioning sets to identify pairs of variables that are conditionally independent [1], [2]. Algorithms for feature selection test whether a new feature is correlated with a dependent variable conditioned on the existing features. Algorithms for learning association rules evaluate whether new items are unexpectedly correlated with a target item conditioned on the existing items in the rule [3]. In each of these cases, assertions of marginal and conditional independence are one of the key statistical inferences made by the algorithm.

Unsurprisingly, inaccurate independence tests can cause serious errors in these algorithms. When tests incorrectly indicate independence, the algorithms disregard important predictive features, reducing the accuracy of learned models. When tests incorrectly infer dependence, algorithms add unnecessary structure to models, increasing the computational complexity of storing and employing those models. Finally, absent or superfluous statistical dependencies can cause a cascade of incorrect inferences in algorithms for learning model structure, particularly causal structure.

Prior research has identified several cases in which conventional tests of independence are seriously inaccurate, particularly when the underlying generative process for the data contains relational dependencies—statistical influences that cross the boundaries of individual entities such that the variables of related entities are correlated [4], [5]. Common domains that exhibit relational dependencies include social networks (the attributes of one person can affect the attributes of their friends), organizational networks (the attributes of an organization can affect the attributes of its members), and web pages.

In this paper, we formally specify causal generative models that can explain an observed dependence between a pair of variables in a relational data set. We show how to translate these models into non-relational generative models by introducing additional variables that capture key aspects of the relational structure. We show how the models directly produce computationally efficient tests of conditional independence. These tests allow algorithms to draw correct inferences despite conditions that mislead conventional tests. Using the principles of *d-separation* [6], we show how several classes of generative models can produce the same observed correlations, and thus cause errors in algorithms that assume a specific generative structure from these correlations.

II. Relational Data, Propositionalization and Type I Errors

The errors we describe have their origins in the mismatch between two data representations: the relational representation of the original data and the propositional representation required by a conventional test of independence. Propositional representations assume that each data instance can be represented solely by a vector of values; they cannot represent relationships explicitly and they typically represent only a single entity type. Relational representations often represent multiple entity types and explicitly represent relationships among instances. Unique identifiers denote instances of a specific type and indicate which instances participate in a relationship.

Propositionalization, sometimes called flattening, transforms data from a relational representation into a propositional one. Many relational learning algorithms incorporate propositionalization either as a pre-processing step or as an integral part of their search algorithms [7].

A. Propositionalization

Two key operations for propositionalization are replication and aggregation. Propositionalizing simple data sets requires only one of these operations, while propositionalizing more complicated data may require several replication and aggregation steps.

Figure 1a depicts a relational database representation of bipartite data in which every entity of type A is linked to

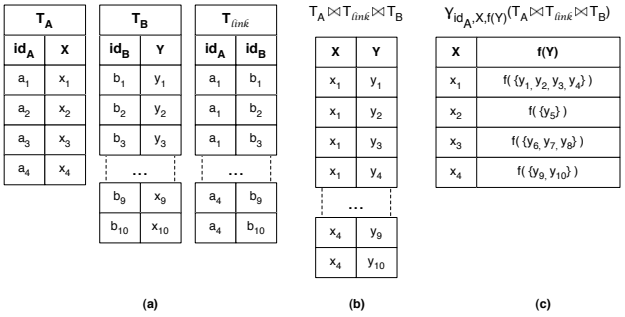

Figure 1. Relational database tables illustrating propositionalization operations (a). Replication (b) is the result of a three-way INNER JOIN of T_A, T_B and T_{link}. Aggregation (c) is the result of a GROUP BY applied to the same join used in conjunction with an aggregation function.

several entities of type B, where each entity type has a single associated categorical variable (X and Y, respectively).

Propositionalizing with replication can be illustrated with a three-way inner join between T_A, T_B, and T_{link}. The two-column projection of this join can be seen in Figure 1b. Each link in the data set produces a tuple in the resulting table. Since nodes with degree greater than one (e.g., A_1) participate in several tuples, their attribute values (in this case, x_1) are replicated in several rows.

Propositionalizing with aggregation can be illustrated with the same three-way inner join between T_A, T_B, and T_{link}. However, in this case, multiple values of Y corresponding to a single entity A are aggregated (Figure 1). The query uses an aggregation function $f()$ (e.g. SUM, AVG, MIN, or MAX) to operate over sets of values and produce a single value for the tuple. In database terminology, a GROUP BY operator with a specified aggregation function or functions is applied to the join. In our example, the X values of the group of B entities associated with each A entity produce a tuple in the target table, as seen in Figure 1.

Information about the relational structure of the data is lost during propositionalization. Relational data sets with different structures can produce the same propositionalized data tables, and statistics calculated on such tables will have the same values. However, the validity of statistical inferences based on these values depends partially on this lost information. In the following two sections, we discuss two specific examples of ways in which propositionalization can introduce spurious correlations.

III. Explaining Pathology with Graphical Models for Propositionalization

A. Errors with replication

Prior work has demonstrated that propositionalization with replication can lead to large increases in Type I errors (falsely inferring statistical dependence). Varieties of this effect have been known for more than a century [8], although the consequences of this effect for relational learning algorithms were first identified by Jensen & Neville [4]. Relational linkage and autocorrelation effectively reduce the sample size of a data set, increasing the variance of scores estimated using that set. Increased variability of the estimated value of any test statistic [9], [4] results in Type I error rates much higher than those expected from independent instances.

Figure 2 depicts the observed distribution of the chi-square statistic along with its Type I error rate for synthetic data containing two types of entities, A and B, each of which contains a single variable, X and Y, respectively. We generate 200 A entities and link each to between 1 and 20 B entities. The level of autocorrelation is expressed as the probability that any two "sibling" B entities will share the same Y value, calculated from the class distribution of Y and a parameter governing the strength of effect (for a data set with no autocorrelation effect, this quantity is equal to $p^2 + (1-p)^2$ for a binary variable with class probabilities p and $1-p$). Here, for a simulation with an autocorrelation level of 0.8 (moderate effect, given an even class split), 38% of the data sets generated had a chi-square value that was statistically significant at the $\alpha = 0.01$ level, substantially larger than the expected Type I error rate of 1%. As seen in the figure, the higher the level of autocorrelation among Y values, the more severe the bias.

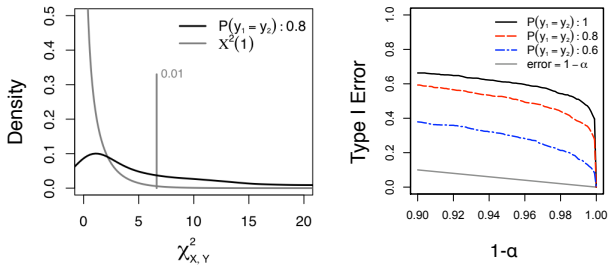

Figure 2. Values of the chi-square statistic are biased for autocorrelated data. Left: The empirical distribution has much higher variance than the theoretical χ^2 with one degree of freedom. Right: The Type I error rate greatly exceeds the expectation based on alpha; the bias becomes more severe for higher levels of autocorrelation.

While only recently explored in the statistical relational learning (SRL) community [4], [10], the effects of autocorrelation have been identified in the social sciences since the nineteenth century [8]. "Galton's problem" denotes the phenomenon of "group effects" causing instance dependence and elevating Type I error [11].

B. Errors with aggregation and degree disparity bias

An alternative to replication is propositionalization through aggregation. While aggregation avoids the types of errors describe above, prior work has shown that aggregation can also lead to mistaken judgments of dependence. Jensen, Neville, & Hay [5] show that aggregation can make uncorrelated variables appear correlated when those data are

propositionalized in the the presence of *degree disparity*. Degree disparity occurs when an attribute on an entity is correlated with the number of links to or from that entity. For instance, chronologically older researchers tend to have authored more research papers and persons from certain religious or ethnic backgrounds tend to have larger numbers of siblings.

Degree disparity combines with some common aggregation functions to produce systematically higher or lower aggregated values when the cardinality of the input values is high. For example, SUM, MAX, and COUNT all return systematically higher values given high cardinality; MIN will produce lower values; and MODE and AVG will produce less extreme values. When data are propositionalized using these aggregation functions, statistical dependencies between attributes and the aggregated value can be erroneously interpreted as dependence between the original attributes. Figure 3 depicts the distribution of Z-scores for relational data that exhibit degree disparity. Each of the different aggregators exhibits a different amount of bias, though all will clearly cause Type I errors for a two-tailed hypothesis test. Even AVG, which is centered, has increased variance when compared to the reference distribution.

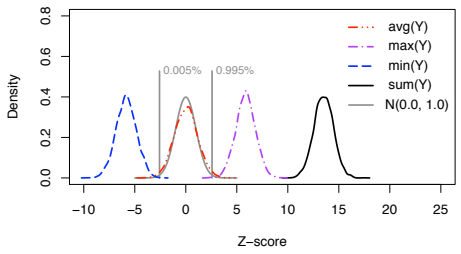

Figure 3. Distribution of z-score values for AVG, MAX, MIN, and SUM in a relational data set with moderate degree disparity. The sampling distributions indicate dependence even in the absence of dependence in the original data. Here, even though X and Y are marginally independent, X appears significantly correlated with aggregations of Y.

Figure 4 depicts Type I error curves for data with degree disparity using the SUM and MAX aggregations. As in the case with autocorrelation, error rates are much higher than those expected at the $\alpha = 1\%$ level. For data with a moderate level of degree disparity, the MAX aggregator has an error rate of 15% while SUM is greater than 70%.

C. Graphical models for propositionalization

The descriptions provided by prior work on independence tests for relational data provides an informal explanation for the existence and strength of the effects described above. However, they provide relatively little formal machinery to reason about these effects. In this section, we provide that machinery.

The situation we described informally in Section III-A can be described more formally by the directed acyclic

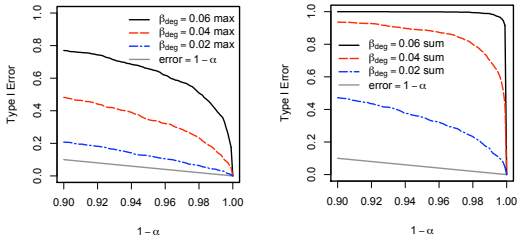

Figure 4. Type I error as a function of alpha for MAX (left) and SUM (right) aggregations under degree disparity. The value of X varies linearly with degree (parameterized by coefficient β_{deg}). At the $\alpha = 0.01$ level, the Type I error rates are 15% and 70% for MAX and SUM, respectively.

probabilistic entity relationship (DAPER) models [12] in Figure 5. Model H_0 corresponds to the null hypothesis that X and Y are marginally independent. Model H_1 indicates that X causes Y. Model H_2 indicates that Y is caused by a latent variable Z on the same entity as X, but otherwise is marginally independent of X. The values of Y on different entities B connected to the same A will be autocorrelated in either of the models H_1 or H_2.[1]

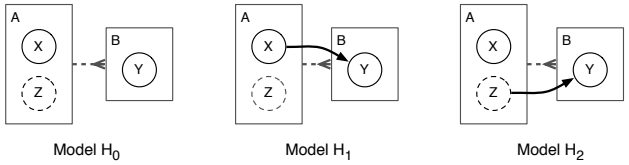

Figure 5. Three possible generative models for one-to-many data. In Model H_0, variables X and Y are independent. In Model H_1, X influences Y, while in H_2, Y is independent of X but related to a latent variable Z. Data generated by H_1 and H_2 will exhibit autocorrelation.

Model H_2 uses a common convention in graphical models to produce autocorrelation among related entities. The relational structure of the data indicates that a single entity A will be connected to several entities B. As a result, the dependence between a variable Z on A and several different instances of a variable Y on B will induce dependence among the values of Y on related entities B. This approach is often used in the social sciences to represent a "group effect" [13]. Models in machine learning frequently use this approach to model autocorrelation among members of latent groups [10] or among topics of related text documents [14], [15].

According to prior work [4], independence tests will frequently indicate that X and Y are marginally dependent

[1]The models in Figure 5 clearly do not exhaust the possible models that could relate these variables, but are meant to demonstrate that multiple generative models are consistent with the observed correlations. Examining the entire space of possible models that relates these variables allows a greater range of causal inferences, but that space is too large to discuss here.

when those data are generated using either model H_1 or model H_2. In general, given a significant value of a statistic alone, it is impossible to determine whether model H_1 or H_2 generated the data. Whether this distinction is important depends on the domain. However, if gaining a causal understanding is important, the distinction is crucial to determining whether manipulating X will change Y [16].

The graphical structure of model H_1 provides a clear indication of why X and Y are dependent in data drawn from this model, but model H_2 does not provide any correspondingly clear indication. One reason for this is that the DAPER model represents data in its relational state, and the results discussed in Section III-A derive from propositionalized data. Propositionalization may introduce additional dependencies not explicit in the DAPER model.

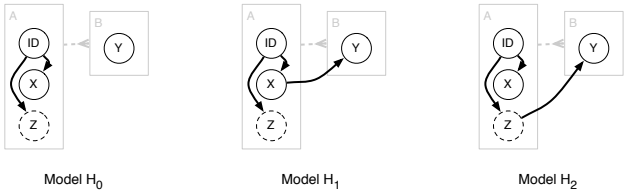

Figure 6. Propositionalized versions of generative models for autocorrelated data. The plate structure is included here for clarity only, and is not part of the graphical model.

Figure 6 shows propositionalized models corresponding to DAPER models H_1 and H_2. The entity-relationship structure is shown in gray for reference only and is not part of the model. The propositionalized models introduce a new variable: ID. The ID variable models the replication of the values of the X and Z variables during propositionalization. In the same way that Z models the autocorrelation among values of Y, ID models the autocorrelation among replicated values of X and Z variables.

The ID variable corresponds to the ID_A and ID_B columns in the relational database tables depicted in Figure 1. The value itself is arbitrary and has no intrinsic meaning; although frequently represented as a numeric value it is a categorical attribute with unbounded cardinality. Furthermore, it carries the constraint that no two entities in the relational data share the same value, although multiple data instances in propositional data can (and often do) have the same value of ID.

The ID variable deterministically causes every other variable whose values are replicated during propositionalization since information about an entity's ID completely determines the value of any variable associated with that entity. Given this, the ID attribute is an example of an infinite latent variable as proposed by Xu et al.[17] (only having perfect predictive ability), or a cluster identifier in the sense used by Kemp et al.[18].

Given the propositional models in Figure 6, the semantics of d-separation provides a formal explanation for the results from Section III-A [6]. In both models, the existence of an undirected collider-free path from X to Y corresponds to the observed correlations between the variables. In Model H_1, the path is direct; in Model H_2, the path flows from $X \leftarrow ID \rightarrow Z \rightarrow Y$. We can block the causal path by conditioning on any of the variables along that path. Conditioning on ID will d-separate X and Y under Model H_2 (but not Model H_1), allowing us to differentiate between the two. However, this fact does not provide a feasible test.

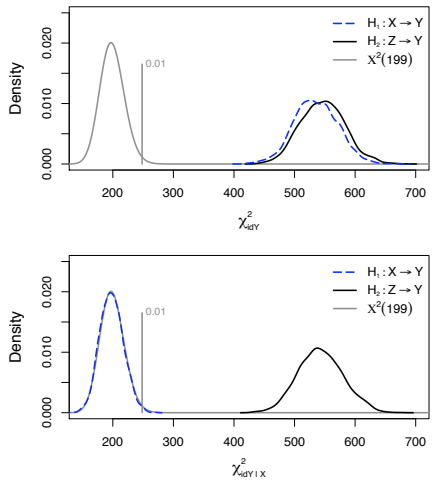

Figure 7. Empirical chi-square distributions for ID, Y. Top: Data generated under Model H_1 is indistinguishable from data generated by Model H_2 as both models create autocorrelation among Y values (captured here as an association between ID and Y). Bottom: The effect of conditioning on X, allowing clear discrimination between models.

Fortunately, the propositional model suggests another conditional independence test to differentiate Model H_1 from Model H_2. If the data were generated by Model H_1, we would expect that $ID \perp Y|X$. Figure 7 shows the empirical distributions of χ^2_{ID-Y} when conditioned on X. The association between ID and Y disappears when we condition for Model H_1, allowing us to retain the null hypothesis. For data from Model H_2, conditioning on X does not diminish the value of χ^2, allowing us to reject Model H_1 in favor of Model H_2. Thus, even with a graphical model that relies on a latent variable (Z), we have a test that allows us to differentiate between the two models.

We can use similar reasoning to understand and correct the bias introduced by degree disparity. Figure 8 shows three DAPER models representing alternative generative structures for the situations discussed in Section III-B. The variable E on the relationship between the A and B entities represents the existence of the relationship itself. We assume that degree disparity stems from a direct causal dependence between the variable X and the probability that one or more relationships exist. Thus, model H_2 indicates that the degree

of entities A depends on the value of X.

Model H_0 corresponds to the null hypothesis under which X and $f(Y)$ are marginally independent. Models H_1 and H_2 represent data in which X and $f(Y)$ are correlated. Once again, knowledge of marginal dependence between X and $f(Y)$ can be used to reject H_0, but it cannot differentiate between H_1 and H_2.

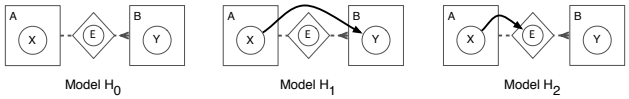

Figure 8. DAPER models for one-to-many data with degree disparity. Model H_0 represents the null hypothesis that X is marginally independent of both Y and E, while Model H_1 specifies that X has influence over Y. Model H_2 represents data that exhibit degree disparity.

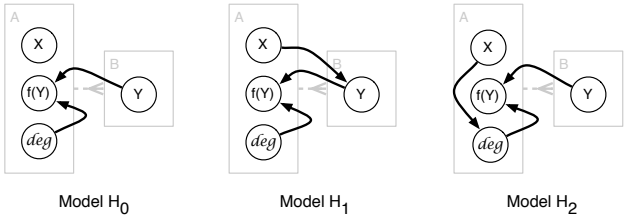

Figure 9. Propositionalized models of aggregated data corresponding to the DAPER models in Figure 8. The effects of degree disparity are represented by the dependence of the deg on X, coupled with an aggregation ($f(Y)$) that is sensitive to degree (and therefore dependent on deg).

Propositionalizing the data produces the corresponding models in Figure 9. The variable $f(Y)$ represents the variable produced by aggregating Y values, and the variable deg represents the number of related entities B (the degree of A). In contrast to the DAPER models in Figure 8, the propositionalized models make clear why both models H_1 and H_2 would exhibit dependence between X and $f(Y)$. In both cases, a collider-free undirected path exists between the variables. However, the models differ with respect to a direct causal dependence between X and Y.

The propositional models also suggest a simple test of conditional independence: conditioning on degree will d-separate X and $f(Y)$. Figure 10 depicts the empirical distributions of the conditional test for data generated under both models. The data generated under Model H_2 indicate no significant dependence, while the data under H_1 do show significant dependence. Conditioning on degree successfully differentiates between the two models.

D. Empirical results on Stack Overflow data

Stack Overflow (*http://stackoverflow.com*) is a website that allows users to post questions and answers concerning problems in computer programming. The Stack Overflow

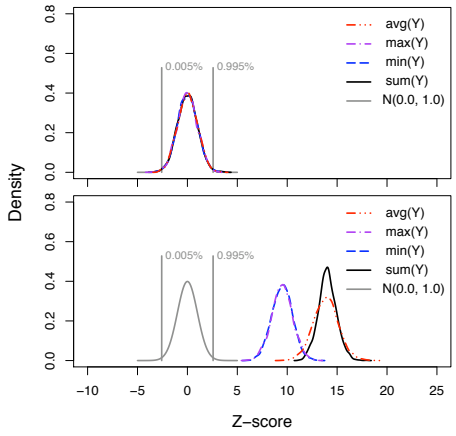

Figure 10. Conditioning on degree removes bias for statistics based on data with degree disparity, allowing differentiation from data containing actual association between X and Y. Top: Empirical distribution of Z-score for data generated under Model H_2. Bottom: Z-score distribution for data from model H_1.

data consists of *users*, *questions*, *answers*. Users may post new questions or provide answers to existing ones, and may *vote* (up or down) on the quality of both questions and answers posted by others. Furthermore, as users use the system, they are awarded *badges* designating some accomplishment. For example, the "Fanatic" badge is awarded to users who visit the site for a hundred days in a row, while the "Guru" badge is given to users who provide an answer that receives forty or more votes.

We examined the relationship between badge acquisition and answer *score* (up-votes minus down-votes). The dataset was drawn from February 1 and April 1, 2010. During this time period, there were 237,505 answers provided by 61,625 distinct users. For each of the 43 badge types, we generated a binary attribute on each user designating whether or not that badge had been awarded *before* April 1. Since users and answers are related in a one-to-many manner, we are able to propositionalize the data using both replication and aggregation. Using conditional independence tests, we can differentiate between models using the procedures outlined in Section III-C.

In the Stack Overflow data set, answer scores are heavily autocorrelated through user; that is, users are fairly consistent in the quality of the posts they provide (the Pearson corrected contingency coefficient is 0.75. Thus, in the replication case, every single badge attribute appeared to be correlated with a discretized answer score when naively tested. However, as discussed above, the marginal dependence between badges and score can be explained by different causal mechanisms as depicted in Figure 6 (for the Stack Overflow data, the badges and scores correspond to X, and Y respectively). Using the ID of each user, we can differentiate between model H_0 and H_1 by performing a

hypothesis test on $User.ID$ and $Answer.score$ conditioned on $User.badge$. In 22 of the 43 cases, the value of chi-square in the conditional test is not significant, allowing us to concluded that the relationship between that badge and answer score is not causal.

In the aggregation case, measured the correlation between the existence of a badge and an aggregated answer score for each user, using the models from 9 (again, $User.badge$ corresponds to X and $Answer.score$ corresponds to Y). By conditioning on degree, we can differentiate the cases where the marginal dependence between badges and scores are due to degree disparity vs. a direct causal mechanism. For SUM, MAX, and AVG, all 43 badge types have a marginal dependence with $Answer.score$; conditioning on degree removes this dependence for 39, 40, and 41 of these. Curiously, $score$ is marginally dependent on only 3 badges, and conditioning on degree *induces* a dependence.

Note that in the cases presented above, we considered each badge in isolation in terms of its causal effect on answer score. We leave a more thorough examination of the causal interactions between badge attributes for future work.

IV. Conclusion

We have used the framework of d-separation to provide the first formal explanation for two previously observed classes of statistical dependencies in relational data. This explanation applies to continuous and discrete variables and essentially any test of conditional independence.

Finally, it is worth noting that many data sets are created in propositional form, even when their underlying generative processes could more accurately be described by a relational representation. Thus, the propositional data sets initially provided to many learning algorithms are "born" without the information needed to draw correct inferences about the underlying generative processes that produced them. Disconcertingly, the effects discussed here apply equally to propositional learning algorithms when the data they analyze were originally drawn from relational domains.

Acknowledgment

This material is based on research sponsored by the Air Force Research Laboratory under agreement number FA8750-09-2-0187. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

References

[1] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. The MIT Press, 2001.

[2] I. Tsamardinos, L. Brown, and C. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006.

[3] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable techniques for mining causal structures," *Data Mining and Knowledge Discovery*, vol. 4, no. 2-3, pp. 163–192, 2000.

[4] D. Jensen and J. Neville, "Linkage and autocorrelation cause feature selection bias in relational learning," in *ICML 2002*. Morgan Kaufmann, 2002, pp. 259–266.

[5] D. Jensen, J. Neville, and M. Hay, "Avoiding bias when aggregating relational data with degree disparity," in *ICML 2003*. AAAI Press, 2003, pp. 274–281.

[6] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, 2000.

[7] S. Kramer, N. Lavrač, and P. Flach, "Propositionalization approaches to relational data mining," in *Relational data mining*. Springer-Verlag New York, Inc., 2001, pp. 262–286.

[8] S. F. Galton, "Discussion on 'On a method of investigating the development of institutions applied to laws of marriage and descent', E. Tylor," *Journal of the Anthropological Institute*, vol. 18, no. 270, 1889.

[9] D. Kenny and C. Judd, "Consequences of violating the independence assumption in analysis of variance," *Psychological Bulletin*, vol. 99, no. 3, pp. 422–431, 1986.

[10] J. Neville and D. Jensen, "Leveraging relational autocorrelation with latent group models," in *4th Int'l Workshop on Multi-Relational Data Mining*, 2005, p. 55.

[11] M. Dow, M. Burton, D. White, and K. Reitz, "Galton's problem as network autocorrelation," *American Ethnologist*, pp. 754–770, 1984.

[12] D. Heckerman, C. Meek, and D. Koller, "Probabilistic entity-relationship models, PRMs, and plate models," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2007, p. 201.

[13] D. Kenny and L. La Voie, "Separating individual and group effects," *Journal of Personality and Social Psychology*, vol. 48, no. 2, pp. 339–348, 1985.

[14] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *UAI 2004*, 2004, pp. 487–494.

[15] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on Enron and academic email," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 249–272, 2007.

[16] R. Scheines, "An introduction to causal inference," *Causality in Crisis*, pp. 185–99, 1997.

[17] Z. Xu, V. Tresp, K. Yu, and H. Kriegel, "Infinite hidden relational models," in *UAI 2006*, 2006, pp. 544–551.

[18] C. Kemp, J. Tenenbaum, T. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI 2006*, 2006, p. 381.