# Relational Blocking for Causal Discovery in Networks

Matthew J. H. Rattigan
rattigan@cs.umass.edu

Marc Maier
maier@cs.umass.edu

David Jensen
jensen@cs.umass.edu

Knowledge Discovery Laboratory
Department of Computer Science
University of Massachusetts Amherst

## 1. INTRODUCTION

Algorithms that learn models of relational data [7][13] rely on statistical conditioning as a mechanism for identifying dependencies. Conditioning allows the algorithm to isolate effects by controlling variability and factoring out the influence of covariates. In this work, we present *relational blocking* as a new algorithmic operator that can be used for learning model structure.

Block designs are commonly utilized in the social sciences to account for confounding variables [19]. Here, we describe blocking designs that can be applied to relational data sets, where blocks are determined by network structure. By blocking on entire entities rather than conditioning on categorical variables, relational blocking allows us to control for both measured and unobserved factors. These designs, while common in manual statistical analysis, are not currently used within automated learning algorithms.

In this paper, we demonstrate the effectiveness of relational blocking for use in causal discovery by showing how blocking reduces variability and increases statistical power; and by showing how blocking controls for entire classes of observed and latent confounders. We describe these advantages using recently introduced formalisms for describing graphical models of relational data. Finally, we show that blocking is distinct from simple conditioning, and thus represents a fundamentally new operator for causal discovery. Specifically, blocking is not susceptible to spurious correlations induced by conditioning on common effects. These benefits can provide strong evidence for drawing causal conclusions.

### 1.1 An Example

Consider the problem of understanding the operation of Wikipedia, a peer-produced encyclopedia of general knowledge [20]. Wikipedia articles, or *pages*, are produced collectively by thousands of volunteer users. Pages are created and modified by users, and users often organize themselves into groups called *projects*, each of which covers a general topic. Within a project, individual pages are assessed by editors for "importance" (how central the page is to the project theme) and "quality" (a project-independent objective evaluation of key criteria).

One of the most persistent claims about Wikipedia is that its high quality stems from the large number of users that collaborate to write each article [12]. We call this the "many-eyes hypothesis" — the more users that revise an article, the higher the quality of that article. If we knew that this association was causal, then we could increase the quality of an article by asking more users to participate in revisions. However, to determine that a causal dependence exists between the number of users editing an article and its quality, we must eliminate other plausible alternative models that could explain an observed dependence. In other words, we must account for all potential common causes, which can be very challenging. Fortunately, the data available on Wikipedia make it possible to evaluate this claim and eliminate some potential threats to a valid causal conclusion.

A naive approach to this question would examine a large number of pages at a given point in time and estimate the correlation between the number of editors and the quality of the page. This design tests the assumptions of the graphical model shown in Figure 1a; given this design, the variables are highly correlated. A chi-square test yields $\chi^2$=101.83 ($n$=189; DOF=12; $p$=2.44×10$^{-16}$), and approximately 66% of the variance of page quality would be attributed to the number of editors. This approach is quite similar to those conducted by many algorithms in machine learning — it identifies a statistical association between two variables, but it does little to identify cause and effect. The observed correlation could stem from a common cause, such as general topic. Pages on topics of high interest to Wikipedians may be edited by a disproportionately large number of users, and that interest could also drive editors to exert special care when editing, thereby improving quality.

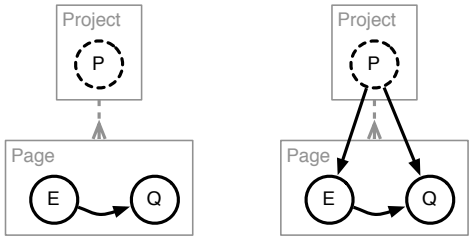


**Figure 1: (a) A simple graphical model can describe the dependence between the number of editors *E* and quality *Q* of an article, but it does not account for common causes. (b) A more complex graphical model incorporates latent common causes *P* associated with project.**

We can remove this potential common cause by using blocking. Since projects govern pages that are thematically similar, blocking on project can factor out the influence of subject matter. This more complex design helps to differentiate between the graphical model shown in Figure 1a and the model in Figure 1b. When we use project links to arrange pages into groups, we find that the average correlation between editor count and page quality has decreased. A Cochran-Mantel-Haenszel test yields $M^2$=82.33 ($n$=189; DOF=12; $p$=1.48×10$^{-12}$). Although lower, this value is still highly significant, and roughly 53% of the variance would now be attributed to the number of editors. The effect size has dropped, but it is still significant. Moreover, using this approach allows a stronger claim regarding the source of the association because we have plausibly factored out at least one potential (unmeasured) common cause. The ability to factor out multiple

variables, observed or latent, is a highly valuable benefit of blocking. After ruling out several plausible common causes of variation, we now have much stronger evidence that the relationship between editor count and page quality is causal, and that the "many-eyes hypothesis" is valid.

The example above highlights three concepts whose intersection forms the basis of this work. First, the Wikipedia data set is *relational*, made up of heterogeneous, interrelated data instances drawn from a relational network. Second, the question being investigated is *causal*. While there is a marginal association between editor count and quality, we are trying to establish a more powerful claim. Lastly, we were able to control for confounding factors (and draw a causal conclusion) through the use of a block design rather than simple conditioning. In the sections below, we discuss each of these three elements in greater detail.

## 2. GRAPHICAL MODELS, CONDITIONAL INDEPENDENCE, AND CAUSALITY

Most modern machine learning algorithms focus on identifying correlations in data. Whether inferring class labels, generating association rules, or clustering, correlational algorithms distinguish, classify, and generalize data by representing the statistical associations within a domain. In this work we are concerned with causal relationships between entities and their associated attributes in relational data.

In recent years, a small but growing effort in machine learning has focused on causal, rather than associational, learning. In addition to computer science, formal reasoning about causal structures has roots in several fields, including philosophy, economics, and statistics. There is much active debate over the proper way to define, detect, and analyze causal dependence. Shadish, Cook and Campbell present a definition of causality that is rooted in experimental design [17], while Rubin provides a framework based on counterfactual logic [15]. In this work, we approach causal reasoning in the manner of the graphical models community, including the work of Pearl [14] and Spirites, Glymour, and Scheines [18]. A brief review of this framework is provided below.

The graphical approach to causality has its roots in Bayesian network learning. At the core of this formulation is the representation of a causal system as a directed acyclic graph (DAG). Two vertices share an edge when those variables share a direct dependence, and edges are oriented to point from cause to effect. Any variable, whether measured or latent, can be considered both a cause and effect, and while it is not always made explicit, it is assumed that the direction of each edge respects the flow of time.

As with associational Bayesian networks, the causally interpreted DAG offers a compact way to represent conditional independence relationships within data. The mechanism for identifying these relationships is Pearl's notion of d-separation [14]. The d-separation criteria describe the graphical scenarios that entail conditional independence relationships in data. When nodes in a DAG are d-separated, they are conditionally independent; when they are d-connected, they are dependent.

A frequent assumption of techniques rooted in the graphical paradigm is that of *causal sufficiency*, stating that any and all common cause variables have been explicitly represented and modeled within the DAG [16]. Of course, doing so usually requires that any possibly confounding variables be observed (although some techniques, such as the FCI algorithm, can reason about unobserved variables [18]). In general, the less certain we are of causal sufficiency, the higher the risk of inferring a causal relationship when only an associational one exists.

The traditional approach in machine learning is to statistically model all possible common cause variables. Structure learning algorithms that learn probabilistic models of a set of variables, including propositional algorithms (e.g., Bayesian network learning [18]) and relational algorithms (e.g., RPC [13]), follow this approach. These techniques determine structure by finding dependencies among variables through statistical control of restricted sets of parent variables. However, even with a highly accurate model, algorithms that rely exclusively on conditioning can succumb to various problems related to the existence of latent, unmeasured variables and low statistical power.

In this work, we are concerned with relational data sets (here, we will focus on bipartite data sets, where entities are related in a one-to-many manner, and leave the analysis of alternative network structures for future exploration). Although the graphical model framework was originally developed for use with propositional data, recent work has focused on adapting the formalism to relational domains. For example, Directed Acyclic Probabilistic Entity Relationship (DAPER) models seek to marry the machinery of graphical models with Entity-Relationship diagrams, a schematic representation of relational domains [10]. In addition, the RPC algorithm extends Bayesian structure learning to a network setting.

## 3. BLOCK DESIGNS

At its core, *blocking*[1] is a data grouping strategy used to control variation and factor out common causes. The block design is traditionally used for causal discovery [19], originating in the agricultural experimental design work of Fisher [5]. In blocking designs, data instances are divided into disjoint groups, or *blocks*, according to the value of one or more *blocking variables*. In a network setting, units can be blocked using network structure in addition to variables. *Relational blocking* groups units that share links with a common entity, called the *blocking entity*. For example, papers written by common authors, or groups of movies produced by the same studio, may form blocks. Blocking in this manner can be used to facilitate causal discovery in network data sets consisting of entities (e.g., people, events, or places) that share some type of relationship or action among them.

Blocking is commonly used in experimental studies; for example, the Randomized Complete Block Design refers to a configuration where each possible value of the treatment (cause) variable is paired with each value of the blocking variable to form the blocks. Within each block, confounding factors (often called "nuisance factors") associated with the blocking variable are held constant, reducing any variability in the outcome (effect) variable that is due to these factors. For example, the analysis of a drug trial might block on the hospital where the treatment was administered,

[1] The term "blocking" is overloaded in the statistical sciences. In this work, blocking refers to instance grouping, and should not be confused with the concept of "path blocking" found in graphical modeling literature.

allowing experimenters to control for any environmental factors associated with the facility.

Block assignment should not be confused with the notion of experimental group assignment found in experimental design literature. Experimental groups are homogeneous with regard to treatment (or lack thereof). In contrast, experimental blocks contain instances with varying treatments and outcomes while homogenizing confounding factors that make detecting the relationship between treatment and outcome more difficult.

Blocking is used less commonly in observational, or quasi-experimental settings. In contrast to experimental domains, treatment is not explicitly assigned in non-experimental settings, so factors associated with each block may affect both treatment and outcome.

The benefit of blocking is twofold. First, by organizing experimental units into groups such that variability within each block is reduced, we improve statistical power. In this respect, blocking serves the same purpose as statistical control. However, blocking simultaneously controls for the influence of entire classes of variables at once rather than a single factor. When applied to hierarchical domains (such as the synthetic domains described in the following section), relational blocking serves a similar purpose to multilevel modeling, where the influence of factors associated with common group or entity is modeled within the appropriate regression equation associated with each level of the hierarchy [8].

The second benefit relates to causal reasoning. Factors that are held constant within each block can be eliminated as possible common causes of treatment and outcome, allowing for a stronger claim of causal sufficiency and pruning the space of alternative causal models of the system. This utilization of relational structure to block by entire entities rather than attributes can be thought of as an extension of the classic twin design. For more than a century, researchers have relied on twin data to control for whole classes of (often unmeasurable) attributes related to family environment and heredity [1].

The dual aims of a block design, increased statistical power and causal sufficiency, can both be served by relational blocking. Below, we consider each in turn.

Block designs increase statistical power by eliminating "nuisance" factors and decreasing the variability within each block. Previous work in relational learning provides strong evidence that blocking by network structure will have this effect. Relational autocorrelation, a commonly observed trait of network data sets, is indicative of an association between network structure and attributes such that entities sharing common neighbors often share similar attribute values as well [11]. This autocorrelation may be the result of differing causal mechanisms; when the existence of relationships stems from attributes, it is referred to as *homophily*; when the reverse is true, it is called *network influence* [1][6]. In either case, blocks constructed from using network structure will exhibit less variability than the population at large in terms of treatment, outcome, or both. Structural blocks will hold constant any attribute associated with the blocking entity, even if it is unobserved.

Furthermore, controlling for the variables associated with the blocking entity can bolster claims of causality. By eliminating entire classes of potential common causes, including both measured and latent variables, the causal sufficiency assumption is relaxed, in that confounding factors can be accounted for even if they are unobserved.

# 4. BLOCKING VS. CONDITIONING

It may be tempting to equate blocking with simple conditioning. While the two serve common purposes — reducing variability and eliminating common causes — they do not produce the same statistical results in relational settings. To illustrate this point, we generate synthetic bipartite data and compare the results of blocking and conditioning for different generative models of attribute structure. Each data set consists of entities of two types, *A* and *B*, connected in a one-to-many manner. In all cases, there are 10,000 *B* entities, with the number of *A* entities varying between different experiments. Each *A* entity carries two attributes, *Z* and *H*, with the former considered measured and the latter latent. The *B* entities also have two attributes, *X* and *Y*, both of which are observable.

In each experiment, the goal is to assess the relationship between *X* and *Y* while either blocking on *A* or conditioning on *Z*. Note that while *Z* is generated as a continuous variable, in each experiment it is discretized to a fixed number of levels in order to compare the results of blocking and conditioning using the same hypothesis test (we use Guo's weighted Pearson's *r* correlation [9]. While not presented here, we found that the results of experiments using partial correlation with an untransformed *Z* were qualitatively similar.

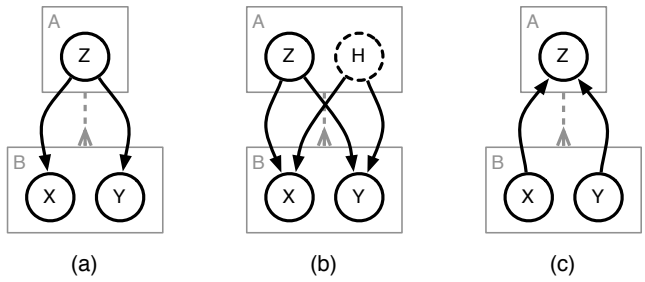


**Figure 2: Different generative models for bipartite one-to-many data. In case (a), blocking and conditioning will both render *X* and *Y* conditionally independent. In (b), blocking is able to control for the influence of latent confounder *H*, while conditioning is not. Case (c) depicts *Z* as a common effect of *X* and *Y*; here, *X* and *Y* are rendered dependent when conditioned on *Z* (Berkson's Paradox), yet remain independent when *Z* is controlled through blocking using entities of type *A*.**

For the first experiment, *X* and *Y* are each dependent on the value of *Z* on the related *A* entity, such that $X = \beta Z + \varepsilon$, $Y = \beta Z + \varepsilon$ where $Z, \varepsilon \sim N(0, 1)$. The generative model is illustrated in Figure 2a, and represents the simple case of a common cause creating marginal association between variables. Using the graphical model formalism, this marginal dependence is evident from the existence of a "collider-free" path connecting *X* and *Y*. Conditioning on *Z*, of course, interrupts this path and renders *X* and *Y* conditionally independent. Here, blocking by *A* entity has the same effect as conditioning.

In the presence of latent variables, however, conditioning and blocking do not perform equivalently. Figure 2b depicts the generative model for data with both a measured (*Z*) and latent (*H*)

variable exerting influence on $X$ and $Y$, such that $X, Y = \beta_Z Z + \beta_H H + \varepsilon$. The plot in Figure 3 depicts Type I error rate at the $\alpha$=0.05 level with $\beta_Z$ held constant at 0.5, and $\beta_H$ varying from 0 to 0.5. Since blocking controls for all confounders, it can be used to establish conditional independence in the presence of unmeasured factors. Thus, in cases where two variables are marginally dependent, conditioning alone is inadequate for ruling out alternative models such as that in Figure 1b.

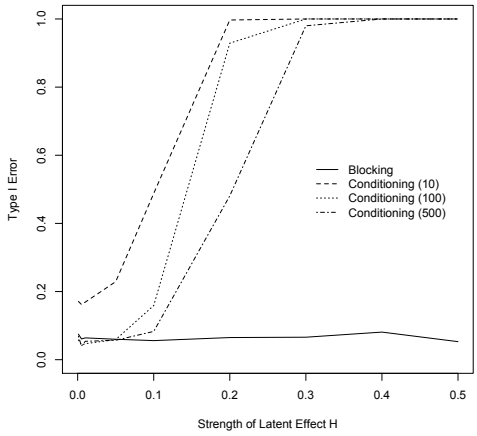


**Figure 3: The effects of blocking and conditioning differ for data generated under the model shown in Figure 2b. Since conditioning can only be applied to measured variables, as the strength of the latent effect ($\beta_H$) varies from 0.0 to 0.5, it is susceptible to high rates of Type I error. The higher the cardinality of discretized $Z$, the smaller the effect. Since blocking controls for both $H$ and $Z$, it is not affected by $\beta_H$.**

As depicted in Figure 3, the cardinality of the discretized $Z$ variable plays a role in modulating the Type I error rate associated with conditioning. While conditioning and blocking can be thought of as distinct operations for analysis, a suitably high cardinality of $Z$ can render the two techniques effectively (and, at the extreme case where $|Z| = |A|$, algebraically) identical.

An additional synthetic case is described by the model shown in Figure 2c. In this case, $X$ and $Y$ are marginally independent, while $Z$ is generated such that $Z = \beta X' + \beta Y' + \varepsilon$, where $X'$ and $Y'$ are the sums of the values of the $X$ and $Y$ values for each related $B$ entity. This case presents an example of "Berkson's paradox" [3], where conditioning on a common effect (or collider, in the language of graphical models) will induce dependence between marginally independent variables. Here, blocking and conditioning lead to opposite conclusions, even in the absence of latent factors. As expected, conditioning on $Z$ does indeed induce correlation between $X$ and $Y$; however, blocking on $A$ does not, even though doing so effectively controls for variable $Z$ as in the conditioning case.

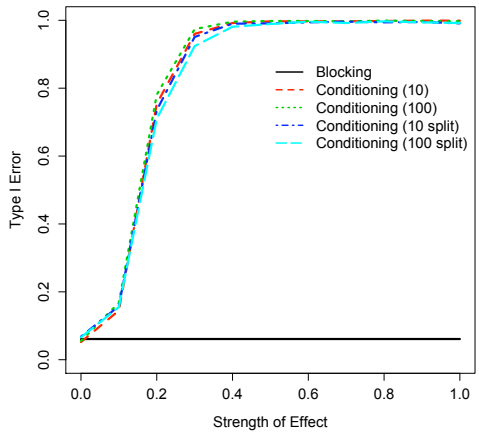


**Figure 4: In the common effect case described by the model in Figure 2c, blocking and conditioning behave differently even when there are no latent factors present. Conditioning on $Z$ will render $X$ and $Y$ conditionally independent. Blocking on $A$ entity, which implicitly controls for $Z$, does not produce the same effect.**

Figure 4 illustrates the difference between blocking and conditioning for the common effect case as a function of $\beta$. As we increase the strength of effect, conditioning induces a dependence between $X$ and $Y$ more frequently. Blocking, on the other hand, does not produce any of the conditional dependence described by Berkson's paradox. Furthermore, the differences between blocking and conditioning cannot be attributed to statistical power. For the case presented above, the block size (10 instances) is significantly smaller than the conditioning groups (100 and 1000). To compensate for this difference, we randomly split each conditioning group into subgroups of 10 instances (labeled as "split" in Figure 4). Even with conditioning groups of equal size to the blocks, the proportion of significant $p$-values is unchanged. This distinction reinforces the notion that blocking and conditioning are fundamentally different operations.

## 5. CONCLUSIONS AND FUTURE WORK

In this work, we have presented relational blocking as a technique to facilitate causal learning in relational data sets. Blocking is similar in function to simple conditioning in its ability to reduce variability and increase statistical power. Unlike conditioning, blocking is able to control for whole classes of observed and latent factors but does not induce dependence when controlling for common effects. By relaxing the causal sufficiency assumption, blocking allows for more robust causal discovery.

This preliminary investigation suggests several avenues of future inquiry. First, we would like to derive a formal understanding of blocking using the graphical model framework and d-separation. In addition, our empirical results are currently limited to network data sets with one-to-many relationships; future studies should include a more complex network structure. Finally, we have described blocking as a new algorithmic operator, and we would like to incorporate it into a constraint-based system for causal discovery in networks.

## 6. ACKNOWLEDGMENTS

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 and the Intelligence Advanced Research Projects Activity (IARPA), or the U.S. Government.

## 7. REFERENCES

[1] S. Aral, L. Muchnik and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences. 106 (51): 21544-21549, 2009.

[2] J. Berkson. Limitations of the application of fourfold table analysis to hospital data. Biometrics Bulletin 2(3):47--53, 1946.

[3] D. Boomsma, A. Busjahn, and L. Peltonen. Classical twin studies and beyond. Nature Reviews Genetics, 3:872–882, November 2002.

[4] D. Campbell and J. Stanley. Experimental and Quasi-Experimental Designs for Research. Rand McNally, Chicago, IL, 1966.

[5] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

[6] N. E. Friedkin. *A Structural Theory of Social Influence*. New York, NY: Cambridge University Press, 1998.

[7] L. Getoor, N. Friedman, D. Koller, A. Pfeffer, and B. Taskar. Probabilistic relational models. In Getoor, L., and Taskar, B., eds. *Introduction to Statistical Relational Learning*. Cambridge, MA: MIT Press. 129--174, 2007.

[8] H. Goldstein. *Multilevel Statistical Models*. Arnold London, 1995.

[9] J.-H. Guo. Four Correlation Coefficients with a Third Blocking Variable: Their Efficacy, Relative Efficiency, and Test Statistics. Communications in Statistics, Theory and Methods. 32(9):1835--1858, 2003.

[10] D. Heckerman, C. Meek, and D. Koller. Probabilistic entity-relationship models, PRMs, and plate models. In Getoor, L., and Taskar, B., eds., *Introduction to Statistical Relational Learning*. Cambridge, MA: MIT Press. 201--238, 2007.

[11] D. Jensen and J. Neville. Autocorrelation and linkage cause bias in evaluation of relational learners. In Proceedings of the Twelfth International Conference on Inductive Logic Programming, number 2583, pages 101–116, 2002.

[12] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work, pages 37–46, 2008.

[13] M. Maier, B. Taylor, H. Oktay, and D. Jensen. Learning causal models of relational domains. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.

[14] J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, 2000.

[15] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701, October 1974.

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

[17] W. R. Shadish, T. D. Cook, and D. T. Campbell. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA, 2002.

[18] P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction and Search. MIT Press, Cambridge, MA, 2nd edition, 2000.

[19] W. M. Trochim. The Research Methods Knowledge Base, 2nd Edition. www.socialresearchmethods.net/kb/, October 2006.

[20] Wikipedia. Wikipedia, The Free Encyclopedia. en.wikipedia.org/wiki/Main_Page, 2009.