Discovering Causal Knowledge by Design

Marc E. Maier Matthew J. Rattigan David D. Jensen

Knowledge Discovery Laboratory
Department of Computer Science
University of Massachusetts Amherst
{maier, rattigan, jensen}@cs.umass.edu

Abstract

Causal knowledge is frequently pursued by researchers in many fields, such as medicine, economics, and social science, yet very little research in knowledge discovery focuses on discovering causal knowledge. Those researchers rely on a set of methods, called experimental and quasi-experimental designs, that exploit the ontological structure of the world to limit the set of possible statistical models that can produce observed correlations among variables. As a result, designs are powerful techniques for drawing conclusions about cause-and-effect relationships. However, designs are almost never used explicitly by knowledge discovery algorithms. In this work, we provide explicit evidence that designs have the potential to be highly useful as part of algorithms to discover causal knowledge. We first formalize the basic elements of experimental and quasi-experimental designs to characterize a design search space. We then quantify the range and diversity of designs that can be applied to examine the central questions associated with a large and complex domain (Wikipedia). Finally, we show that explicit consideration of designs can substantially improve the accuracy of causal inference and increase the statistical power of algorithms for learning the structure of graphical models.

1. Introduction

The strongest type of knowledge we can discover is causal. Causal knowledge is actionable, as opposed to associational knowledge, which is only predictive. The existence of a causal dependence between two variables, x and y, implies that manipulating x will result in a change in y. While a necessary precondition for causality, statistical association between x and y merely implies that knowledge of x can help predict the value of y. Thus, causal knowledge is commonly pursued in fields such as medicine, economics, biology, and social science where true understanding of behavior and knowledge of how to effect changes are necessary.

Causal discovery is strictly a more difficult task than non-causal discovery because it requires inferring a superset of the conditions of non-causal discovery. The most challenging of these conditions is to eliminate the effects of all potential common causes, whether observed or latent. As a result, very little research in knowledge discovery focuses on causal discovery, and discovering causal knowledge is left as a manual activity for researchers in fields that require explicit determination of causal knowledge. Those researchers rarely use the algorithms developed within the knowledge discovery community; instead, they much more often employ a set of formal frameworks and guidelines for gathering and analyzing data that has been developed over the past eighty years. Some of these approaches—collectively called *experimental designs*—help structure the planning and conduct of prospective experiments. The rest of these approaches—called *quasi-experimental designs*—help structure the analysis of retrospective studies of observational data. Such designs can alleviate the challenges of causal knowledge discovery by increasing statistical power and eliminating common causes.

In work published last year, Jensen conjectured that incorporating specific consideration of designs could improve algorithms for machine learning and knowledge discovery [10], and Jensen et al. later showed that quasi-experimental designs could be identified automatically [11]. In this paper, we formalize designs with respect to relational databases and adapt them to knowledge discovery in order to characterize the search space of designs. The formalization of a search space of designs presents an unexploited opportunity to develop algorithms that automate the identification and application of designs for causal discovery. We show that the design space is large for relational domains, and that relational data enables the expression of designs and facilitates their applicability to causal discovery. We also

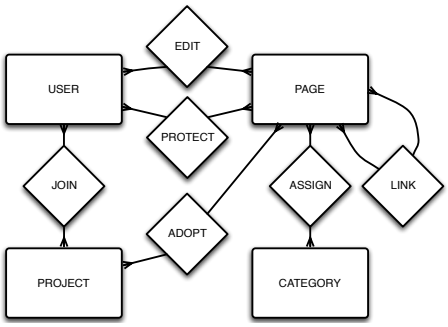

Figure 1. Wikipedia is a complex domain with various entities, relationships, and attributes.

demonstrate how designs can improve causal knowledge discovery by eliminating common causes (both latent and observed), reducing the space of alternative causal models, and increasing statistical power.

1.1. Example

Consider the problem of understanding the operation of Wikipedia, a peer-produced general knowledge encyclopedia [21]. 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). See Figure 1 for a more complete relational data schema describing the major entities and relations that make up Wikipedia.

Wikipedia exemplifies many common aspects of modern data sets. It is made up of heterogeneous entities connected by relations. Many data elements record temporally-varying characteristics; Wikipedia records the precise moment of all user actions, page edits, etc. so any associations can easily be examined with respect to temporal order. Finally, the data exhibit sufficient complexity that they allow examination of a remarkably wide variety of questions.

For example, one of the most persistent claims about Wikipedia is that its reputability 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 claim were actually causal, then we could theoretically increase the quality of an article by asking more users to participate in revisions. However, to actually determine that there exists a causal dependence between the number of users editing an article and its quality, we must eliminate other plausible alternative models that could explain the observed correlation. 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. In fact, the data allow the use of a number of different designs, each eliminating different potential threats to a valid causal conclusion.

A naive approach to this question would be to simply 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 2, and given this design, the variables are highly correlated. A chi-square test yields χ^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 knowledge discovery and data mining—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.

We could remove this potential common cause by using an alternative design. Since projects govern pages that are thematically similar, we can use page-project relations to factor out the influence of subject matter. This more complex design helps to differentiate between the graphical model shown in Figure 2 and the model in Figure 3a. When we use project links to arrange pages into groups (called "blocks" in the language of experimental and quasi-experimental design), we find that the average correlation between editor count and page quality has decreased. A Cochran-Mantel-Haenszel test [1] yields M^2=82.33 (n=189; DOF=12; p=1.48×10^{-12}). Although lower, this value is still highly significant, and roughly 53.4% 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

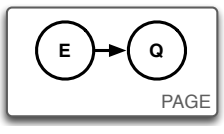

Figure 2. A simple graphical model can describe the dependence between the number of editors and quality of an article, but it does not account for common causes.

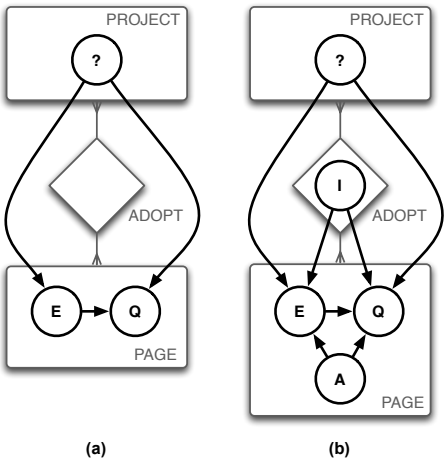

Figure 3. (a) A more complex graphical model incorporates common causes (potentially latent) due to project; (b) An even more complex graphical model adds measured variables of article importance and age to account for all plausible common causes of number of editors and quality.

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 this type of design. This design is easily found by exploiting the relational structure of the data, yet it is unexploited in current knowledge discovery algorithms.

There are still other potential common causes that could account for the observed correlation between editors and quality. For example, the importance or age of a page could influence both the number of edits and the page quality. Fortunately, we can build on the design above to factor out the potentially confounding influence of importance and age. Both of these effects can be mathematically modeled, allowing us to statistically control for their effects. When we do so, the correlation between editors and quality drops to M^2=29.13 (n=189; DOF=12; p=0.00377), and the effect size also drops to 18.9% of the variance. 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 indeed causal, and that the "many-eyes hypothesis" is valid.

1.2. Central concepts

The example in the previous section illustrates many important concepts. First, learning causal knowledge is very useful and can have greater utility than associational knowledge. If many eyes cause page quality, then Wikipedia administrators seeking to improve article quality could try encouraging a diversity of editors within single pages. However, if we only know that the number of editors and page quality covary, then various alternative causal models could explain the correlation, each implying a different action. Thus, causal knowledge is highly useful in domains we would like to control, and, in settings that are not amenable to change, true understanding of behavior can be derived only under causal interpretations.

Second, causal knowledge discovery is a difficult task. To infer a causal dependence between two variables x and y, three conditions must be met:

1. *Association* — x and y must be statistically correlated.
2. *Direction* — The direction of causality is known.
3. *No common causes* — All possible common causes of x and y have been accounted for.

Clearly, causal discovery requires inferring a superset of the conditions of standard knowledge discovery. Finding statistical associations is the focus of most current knowledge discovery algorithms. The second condition is typically achieved with temporal precedence and is straightforward given some notion of time within the data. The third condition of eliminating the effects of all possible common causes is challenging.

One approach is to statistically model all possible common cause variables. In fact, structure learning algorithms that learn probabilistic models of a set of variables, including propositional algorithms (e.g., Bayesian networks [8]) and relational algorithms (e.g., Probabilistic Relational Models [6]), follow this approach. They determine structure by finding dependencies among the variables through statistical control of restricted sets of parent variables. However, even with a highly accurate model, this approach succumbs to various problems, such as the existence of latent, unmeasured variables and low statistical power.

Third, designs are useful for causal knowledge discovery. Designs are mostly employed by researchers in other fields to alleviate the challenges of discoverying causal knowledge, and the techniques used in the previous section emphasize their utility. The first design used in the example is similar to those implicitly encoded in many knowledge discovery algorithms. It can certainly identify a statistical association between two variables, but it does not address common causes. It is important to note that causal conclusions are only valid up to the assumptions, or causal constraints, that we make. If we assume that there are no common causes between the number of editors and page quality, then this design, upon detecting a significant association, could conclude causation. The other designs in the example incorporate additional elements of blocking and statisti-

cal control that help to account for more potential common causes.

Finally, the space of designs is large. The example highlights that, for a given analytical task, several potential designs apply, and we are interested in identifying and selecting among those designs. Additionally, it is the relational structure of the data that enables many designs. For example, the project-page relations in Wikipedia allow the blocking design that controls for all variables attributed to projects. By utilizing the complex structure of our data, different designs are able to rule out different sets of competing hypotheses that can explain an observed correlation.

In the sections that follow, we formalize the concept of designs for knowledge discovery and relate them to experimental and quasi-experimental designs found in other empirical sciences. We demonstrate that, for typical data sets such as Wikipedia, there exist many valid designs for each given task, and that selecting among these designs provides higher analytical power than consistently selecting a single simple design.

2. Designs

The full range of experimental and quasi-experimental designs is unfamiliar to many researchers in knowledge discovery. As a result, almost no knowledge discovery algorithms incorporate explicit consideration of experimental and quasi-experimental designs. A system might consistently make use of one hard-coded design, and human investigators sometimes explicitly select a design before submitting data to a knowledge discovery algorithm, but no system known to us automatically and dynamically selects among multiple potential designs in an effort to maximize the statistical power or identify causal relationships. The algorithm developed by Jensen et al. was the first system to explicitly identify one type of quasi-experimental design, but this work was primarily proof-of-concept and very simplistic [11]. Given the widespread use of designs in data analyses directed by human analysts and the absence of designs from current knowledge discovery algorithms, developing algorithms that automate the identification and application of designs appears to be a great opportunity for causal knowledge discovery.

Our goal in this section is to adapt and translate the literature of experimental and quasi-experimental design into a body of ideas that complement existing concepts familiar to researchers and practitioners in the knowledge discovery community. We formally define designs in terms of relational databases and the relational algebra. Given the natural composition of the relational algebra, this formalization of designs clearly characterizes a large search space for designs and stresses the potential, and need, for automating the process of identifying applicable designs.

2.1. Designs defined

Given a relational database D characterized by a schema $S = \{R_1, \ldots, R_k\}$ with attributes $\mathcal{A}(R_i)$ defined over each relation R_i, a *design* is a function from S to a result table R, such that R specifies the data necessary to conduct a hypothesis test that determines the dependence between some treatment variable $x \in \mathcal{A}(R_i)$ and some outcome variable $y \in \mathcal{A}(R_j)$. For example, the naive design that tests the many-eyes hypothesis for Wikipedia specifies both the treatment and outcome variables as attributes of the Page relation. The aim of a design is to construct a situation in which the outcome of a single hypothesis test will determine, with high probability, whether a causal dependence holds between two specified variables. That is, designs are formulated in such a way that a valid conclusion about statistical association will correctly determine causal dependence.

The design itself does not specify the hypothesis test, only the necessary data to conduct such a test. An analyst (or algorithm) must choose both a design and an accompanying statistical test. The statistical test depends on the levels of measurement for the variables (i.e., nominal, ordinal, or continuous) and the type of design. For example, the design that blocks for associated projects must group the instances and use a hypothesis test that can handle stratified data (e.g., the Cochran-Mantel-Haenszel test). As we are primarily concerned with formally defining designs, we leave the choice of hypothesis tests as a separate issue.

As defined above, a design can be formulated given knowledge about the entities, relationships, and attributes in a domain such as those in the schema for Wikipedia shown in Figure 1. A design tests the dependence between two variables, the treatment and outcome, of a particular class of units. *Units* denote some experimental subject (e.g., entity, group of entities, relationship) during a specific period of time and corresponds to a single row in the result table. *Treatments* are potential causes and can be any measurable item that affects a unit, including an attribute value (of a specific entity or an aggregate) or a link to another entity. Formally, treatments can be defined in the relational algebra as

$$\pi_{treatment}(R_{Base} \bowtie \cdots \bowtie R_{Treatment})$$

where R_{Base} is the base table of the unit, and the treatment variable is projected following a series of joins (i.e., a path in the relational schema) to the underlying table that defines the treatment variable. *Outcomes* are potential effects and can be defined in the same manner as treatments.

2.2. Design elements

We discuss different designs in terms of three main elements: (1) sampling; (2) blocking; and (3) statistical con-

trol. Each of these elements can be formally defined within the relational algebra, and they can be naturally composed to form a vast design space. We believe this space of designs has great potential as a search space that can be automatically explored, and this is an avenue of research we are actively pursuing. In this work, we focus on defining the search space and, in later sections, demonstrating the wide applicability of designs to relational domains and the large benefit associated with using designs to discover causal knowledge.

Sampling is the process of refining units, treatments, or outcomes, and it is formally defined as

$$\sigma_{\varphi}(R_{Base} \bowtie \cdots \bowtie R_{Sample})$$

where φ is a propositional formula specifying conditions necessary to be included in the sample.

In the KD community, data sampling is usually only employed as a computational necessity. We tend to sample data to speed up algorithms or calculations with unacceptable runtimes, or to clean data in hopes of reducing noise. Otherwise, KD practitioners are hesitant to "throw away" data by sampling; by analyzing all the data that are available, we hope to maximize the sample sizes involved in our calculations and increase power. However, selecting subsets of the data to analyze can improve the accuracy of causal inference, yield an increase in statistical power, or help identify context-sensitive dependencies. For example, twin studies are able to draw conclusions about the population at large by looking at a small fraction of all available individuals [3]. Of course, the accuracy gained through sampling may come at the expense of generalizability. If twins are not representative of the population at large, the conclusions we draw from their study will not be widely applicable.

Blocking is a data grouping strategy, a way of organizing units to improve models and increase power. Blocking serves similar purposes to statistical control: reducing variability and pruning the space of alternative causal models. However, blocking simultaneously controls for the influence of entire classes of variables as opposed to that of individual ones. The goal of blocking is to organize experimental units into groups such that variability within each group, or block, is reduced. Blocks contain units with varying treatments and outcomes while homogenizing confounding factors that make detecting the relationship between treatment and outcome more difficult. These blocks are then modeled individually, and the results are combined into a hypothesis test to make a determination about the population at large.

Relational data sets lend themselves to the creation of blocks, as link structure is often correlated with attribute values. One main type of blocking is entity blocking, which can be defined as

$$\pi_{blockId}(R_{Base} \bowtie \cdots \bowtie R_{Block})$$

in which we identify entities that are related to a particular base unit through a series of joins in the schema. Units with common blocking entity identifiers can then be grouped together. In the example above, we blocked Wikipedia articles according to their related projects. By blocking in this manner, we control for any possible confounding aspect of projects on page quality. This allows us to control for all project attributes simultaneously, whether they are measured or unmeasured. This ability to factor out the effects of latent variables is one of the key strengths of blocking over more traditional approaches to statistical control of individual variables.

In addition to blocking by entities, we can block instances of the same experimental unit over time.

$$\pi_{id}(\sigma_{\tau_1}(R_{Base})) \cup \sigma_{\tau_2}(R_{Base})))$$

The same base unit, at different times τ_i, is unioned with itself. This allows analysts to control for all aspects of an entity, except those that are influenced by the passage of time itself, so-called maturation effects [4, 17]. In a sense, a unit at an instant in time t_1 is a "twin" of the same unit at some time t_2. Temporal blocking is often implicit in relational learning; any time we examine multiple measurements of the same attribute on an entity we are blocking their associated instances temporally.

Finally, in addition to sampling and blocking, we can account for common causes using *statistical control*.

$$\pi_{control}(R_{Base} \bowtie \cdots \bowtie R_{Control})$$

The design can specify additional variables to control for, retrieving them in a similar manner as the treatment and outcome variables. This class of methods models the effects of potential common causes so that their effects can be removed mathematically. If there can be any claim that concepts from quasi-experimental design are used in existing knowledge discovery algorithms, then they would be based on the use of statistical control. For example, work by Pearl [14], Spirtes, Scheines, and Glymour [18], and others has explored this approach within the context of graphical models.

The design testing the many-eyes hypothesis that combines blocking on projects and controlling for importance and age can be expressed as follows:

$$\begin{aligned}
&\pi_{editorcount}(\text{PAGE}) \bowtie_{pageid} \pi_{quality}(\text{PAGE}) \\
&\quad \bowtie_{pageid} \pi_{projectid}(\text{PAGE} \bowtie \text{ADOPT} \bowtie \text{PROJECT}) \\
&\quad \bowtie_{pageid} \pi_{age}(\text{PAGE}) \\
&\quad \bowtie_{pageid} \pi_{importance}(\text{PAGE} \bowtie \text{ADOPT})
\end{aligned}$$

While the precise details of the definitions of design in the literature vary [16, 17, 19], the description above is largely compatible with most other definitions of design.

Entity and temporal blocking may appear to be fundamentally different techniques, but they are actually quite similar when we consider that both operations group units. When we include multiple observations of the same entities within an experiment, we're including multiple instances of the same entity and blocking them together. Furthermore, if we consider these instances to be connected by "temporal" links, then relational and temporal blocking amount to the same technique: using link structure to group instances together to reduce variability and eliminate common causes.

One distinction between our categorization of designs and those found in the social science literature involves how we define treatment. Traditional experimental descriptions separate units into distinct treatment and control groups. In our context, treatment can take on many possible values; they are not merely binary designations. Also, most discussions of experimental design in the social science literature include detailed discussions of different "threats to validity" that accompany designs or design elements (e.g., measurement error, maturation, resentful demoralization among subjects). While we do not discuss specific threats here, all internal validity concerns can be viewed as possible alternative models that explain the data and invalidate the favored hypothesis.

3. Applicability of designs

In order to determine whether designs are a profitable direction for enhancing knowledge discovery algorithms, we must assess how often designs actually apply to typical domains targeted by knowledge discovery researchers. To examine this question, we take an in-depth look at Wikipedia. Although the Wikipedia schema (as described in Section 1.1) is not overly complex, it can spawn dozens of causal questions with hundreds of applicable designs. There are four main entity types (articles, categories, projects, and users) and six types of relationships. Wikipedia is, however, quite large, with over 2 million pages, 8 million users, and close to 300 million edits. Furthermore, although Wikipedia has been the subject of several recent studies [9, 20, 22], we know very little about how it functions, especially from a causal standpoint. These aspects make Wikipedia an ideal candidate for studying the applicability and utility of designs.

We wanted to produce a representative list of tasks of real concern that could then be examined in the context of design. We surveyed a group of ten people, each with a bachelors or masters degree in Computer Science, to obtain a sample of interesting causal questions in the Wikipedia domain. Respondents were given a simple list of attributes (see Table 1) and asked to indicate ten pairs of treatments and outcomes they found compelling for study. Treatments and outcomes were presented in one of five random orders,

Table 1. Complete list of survey variables.

Entity	Attributes
Page	Adopted by Project, Age, Assessment, Editors, Edits, Featured, Importance, Length, Notice, Number of Links, Protected, Quality, Views
User	Role, Edits, Membership in Project
Edit	Size, Vandalism, Minor, Reverted

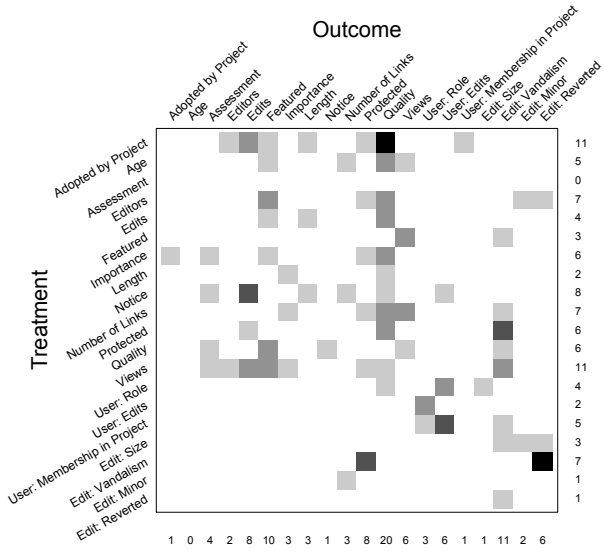

Figure 4. For each treatment and outcome variable pair, the degree of shade indicates how frequently that particular task was selected by survey respondents. The row and column marginals indicate how often a particular treatment or outcome variable was selected, respectively.

to eliminate biases associated with presentation order. The group generated a list of 99 causal discovery tasks (one respondent provided only 9 tasks), 71 of which were unique. A graphical representation of the survey results can be seen in Figure 4 (unless otherwise noted, variables correspond to page entities). For each pair of treatment and outcome variables, the shading in the corresponding square indicates the number of respondents who highlighted that particular relationship in his/her response. Page quality and vandalism were the most popular outcome variables selected, while page views and project inclusion were deemed interesting treatments.

In Section 2.2, we discussed the advantages of using blocking designs to eliminate common causes and reduce variability. Our ability to employ a blocking design does not rely on the particular treatment or outcome variables;

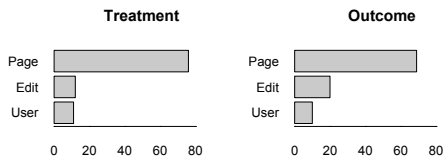

Figure 5. The frequency with which selected treatment and outcome variables appear on particular entity types.

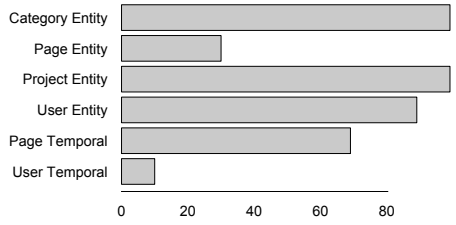

Figure 6. The number of tasks that apply to a given method of entity or temporal blocking.

rather, it is based on the structure of the schema surrounding the entity type associated with the outcome variable. Figure 5 shows the frequency that different entity types (page, user, edit) occur as the underlying entity for either the treatment or outcome variables. Page was the most frequently selected underlying entity for both treatment (76 out of 99 tasks) and outcome (69 out of 99 tasks) variables. This is probably due to the disproportionately large number of variables associated with page as opposed to users or edits. We manually identified the different design choices, and Figure 6 depicts the applicability of the different relational and temporal blocking schemes for the tasks identified in our survey. The overwhelming majority of tasks identified were amenable to at least one blocking design.

Below, we consider the design options for three example tasks generated from the survey in detail. Our goal here is to demonstrate the breadth of applicable designs for each task rather than focus on selecting the "best" design to evaluate the relationship in question. The first two tasks correspond to the most cited treatment-outcome pairs in the group (each was identified by four respondents). The third example is one of several that were chosen by three respondents. These three tasks are representative of the entire list of tasks, as each employs a different outcome entity type.

Example 1: Page adoption by a project → Page quality

This task concerns the dependence between articles being included in Wikipedia projects and their quality. In Wikipedia, quality is assessed as an ordinal variable with 7 levels, ranging from *Stub* to *Featured Article*. The simplest design would involve only the treatment and outcome variables, thereby ignoring all other related entities, but this design would not be robust in terms of factoring out potential common causes of treatment and outcome. For instance, certain categories of pages might be highly likely to be adopted by projects as well as of uncommon quality; blocking instances by category could eliminate this interpretation of the data from the hypothesis space. Similarly, common cause threats from users, linked pages, and additional projects can be addressed via blocking, yielding 32 different designs. In addition, a temporal design can be applied to the pages themselves, examining the page quality longitudinally before and after it is added to a project. Finally, any subset of the several page entity attributes (e.g., importance, age, length) could be statistically controlled. All together, there are dozens of designs that apply to this task.

Example 2: Edit vandalism → Edit reversion

This task involves the relationship between two variables on edits; specifically, whether edits categorized as vandalism are reverted. While the association between these variables may appear trivial, establishing the strength of a generalizable causal relationship would be of great interest to Wikipedia administrators. In terms of blocking choices, edits are made by users to particular articles, so designs blocking on pages and users apply. In addition, the influence of page category could also be a confounding factor, as could project inclusion. Unlike the previous example, however, not all combinations of blocking choices are compatible within the same design—for example, it is meaningless to block on page as well as project, since page blocks would subsume project blocks. In total there are ten different combinations of factors to use to group edit instances into blocks via relations. Since edits are instantaneous, they cannot be blocked temporally.

Example 3: User project membership → User edit count

This task considers whether user membership in a Wikipedia project has an effect on the number of edits they make on various articles. Possible alternative hypotheses that could explain a correlation include the influence of page content or project theme. Both of these can be ruled out by blocking user instances on those entities. Furthermore, temporal blocking can be employed to control for the inherent variability of the edit behavior of the users themselves.

As we demonstrate by these example tasks, the applicable design choices for even simple causal questions are numerous and potentially overwhelming to investigate manually. A knowledge discovery algorithm that could automat-

ically search the space of possible designs would be beneficial, which is the primary motivation for performing this assessment.

4. Utility of designs

In Section 2, we described the different elements of experimental and quasi-experimental design in a knowledge discovery context. In Section 3, we demonstrated how various designs can be used to examine interesting causal questions in a typical rich data set. Now we examine the key advantages of sampling, blocking, and statistical control through the use of two examples taken from the Wikipedia domain. Where possible, we quantify the effects of each design element, and we support these findings through additional simulations.

The two example tasks from Wikipedia can be summarized with the following questions:

1. *Do "many eyes" cause page quality?* That is, does the number of unique contributors to an article affect its quality? Recall this is the example used in Section 1.1. For this task, the units are Wikipedia articles, the treatment variable is the number of distinct users that have edited the article, and the outcome is its quality. To analyze this question, we randomly sampled 189 Wikipedia articles from within ten different projects.

2. *Does exposure of an article cause a change in editing behavior?* For this question, we examine the effect of being featured on Wikipedia's main page as "Today's featured article" on the number of edits made to that article. In this setting, the units are also Wikipedia articles, the treatment is being featured on Wikipedia's front page (which changes daily), and the outcome is the number of edits measured over different time periods. For this question, we randomly selected 97 Wikipedia articles that were featured in 2008.

4.1. Sampling

Sampling, the first element of design we discussed, can be utilized in designs for both examples. Sampling has two primary benefits: enabling the creation of blocks and reducing variability. To analyze the many-eyes hypothesis, we sample pages that have associated projects and have been assessed for quality. Not only does this ensure that the outcome variable (quality) is measured, but it also enables a more complex design to block on projects. In this case, the sampling procedure reduces the size of the data by 72% because only a fraction of articles have been assessed. In the second example, we sample only articles that have "featured article" quality. Only pages designated with this status may be chosen as "Today's featured article," and only 1 out of every 1,130 articles actually have featured article quality. This is equivalent to more than a 99.9% reduction in the number of articles studied.

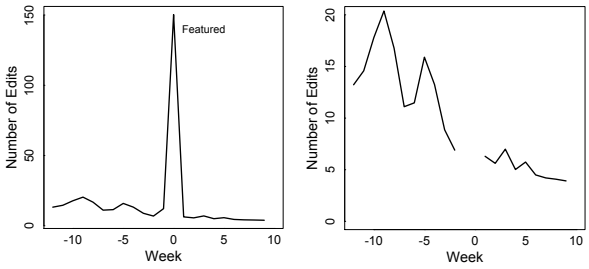

Figure 7. Featured article have an immediate and vast increase in the number of edits when they appear on Wikipedia's front page (left). If we remove this edit spike, a clear downward trend in the number of edits is visible (right).

Additionally, sampling can be used to reduce variability, which may change the detected effect size or increase power. Suppose we hypothesize that users prepare an article for an upcoming request to be "Today's featured article" by frequently editing the page in the weeks leading up to the event. However, after the article's exposure, the number of edits decreases since it no longer requires preparation for widespread exposure. We can test this by performing a one-tailed t-test on the number of edits in the month before and month after being featured, but we will incorrectly reject this hypothesis regardless of doing a paired or unpaired t-test. On closer inspection, it is clear that being featured leads to an immediate increase in edits (see Figure 7). Therefore, if we sample the number of edits in the month before and month after, but remove the edits during the week surrounding being featured, then the effect size completely changes. In fact, the sign of the effect changes from an average increase of 128 edits due to being featured to an average decrease of 21 edits due to being featured.

4.2. Blocking

Blocking designs can relax the causal sufficiency assumption by eliminating entire classes of potential common causes, including both measured and latent variables. In the many-eyes task, we block pages by project in order to control for potential common causes, as discussed in Section 1.1. Without blocking, there is a strong correlation between the number of editors and page quality. Using a simple Chi-square test, we obtain a value of χ^2=101.83 (n=189; DOF=12; p=2.44×10^{-16}). After blocking on

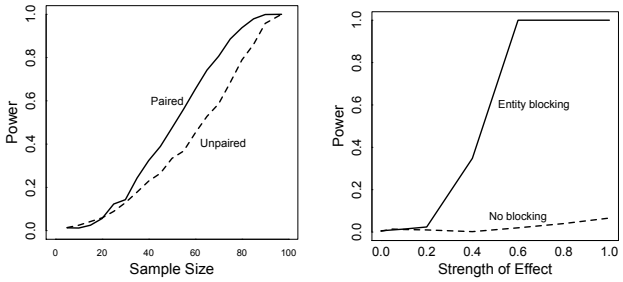

Figure 8. (a) Looking at article edits in the month before and after an article becomes featured, a design that employs blocking can increase statistical power by reducing variance. (b) As the strength of effect increases between two variables x and y, a design blocking on common entities can detect correlation with increasingly higher power as opposed to a design without blocking.

project, the Cochran-Mantel-Haenszel test achieves a value of M^2=82.33 (n=189; DOF=12; p=1.48×10^{-12}), indicating that blocking on project explains some of the originally observed correlation between the number of editors and page quality. By blocking, we have also reduced the hypothesis space by eliminating any potential alternative causal models in which variables on project solely control for correlation between number of editors and quality. As a result, we can make a stronger claim about the source of the remaining association.

In the featured article example, we block temporally on page by comparing the number of edits in the month before and after being featured (and ignoring the surrounding weeks of the featured event). In this case, not only do we factor out potential common causes residing on the page itself, but we also increase the power since the within-block variability is less than the between-block variability. This is the second main benefit of blocking: Blocking can be used to decrease variability and increase the power to detect correlation. In the featured article example, we use a paired t-test which increases the power of the test. In Figure 8a, we see that blocking provides for the equivalent of approximately 10 additional data instances for the same level of power (e.g., 80% power). However, beyond power, we have again reduced the space of alternative models by eliminating potential common causes attributed to the page itself.

We can use simulation to examine the phenomenon of blocking to increase power in more detail. We generate data from a bipartite model that mirrors the structure of articles and projects in Wikipedia. Each page entity has two discrete variables x and y, and there is a latent variable z on the project entity; x and y covary by a given strength, but z also contributes to the value of y. In Figure 8b, we illustrate the power gained from blocking by project as a function of the strength of the dependence between x and y. As we increase the strength of effect between x and y, the power of the blocking design increases dramatically; without blocking, even a strong dependence goes undetected.

4.3. Statistical control

Statistical control is an element of design applicable to all tasks. Like blocking, it can be used to eliminate potential common causes of measured variables and reduce variability. In the many-eyes task, we assume that age and importance of an article are common causes of the number of distinct editors and page quality. As a result, it is necessary to factor out the additional source of variation due to these two variables. When we block on project and additionally condition on age and importance, the Cochran-Mantel-Haenszel test still indicates a highly significant correlation between number of distinct editors and page quality (M^2=29.1286; n=189; DOF=12; p=0.00377). Again, by controlling for additional factors we reduce the space of alternative models. Given the assumption that there are no remaining common causes, we can conclude that the many-eyes hypothesis holds true for Wikipedia: To achieve high quality articles, we should encourage more volunteers to contribute to articles.

As shown in Section 3, large numbers of different designs can be formulated for most causal questions of interest. In this section, we have shown that the benefits of applying any individual design can be large, and that those benefits vary substantially depending on which design is employed. This suggests an unexploited opportunity for knowledge discovery algorithms to identify, evaluate, and apply designs to facilitate causal discovery.

5. Related work

The ideas behind experimental and quasi-experimental design date back decades [4, 5], but they continue to be an active area of research to date [16, 17, 19] as new types of designs and their associated analysis are formulated and classified. Most traditional quasi-experimental designs fit nicely into the framework of sampling, blocking, and statistical control. For example, the examination of editor count and page quality was an expanded "posttest only with nonequivalent control group" design, while the temporally-blocked design examining page exposure and edit frequency in Wikipedia was an example of an "untreated control group with dependent pretest and posttest samples using switching replications."

The concept of blocking as a means of controlling variance has been a widely used technique in statistical analysis [19]. The utilization of relational structure to block by entire entities rather than attributes is a generalization 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 [3].

The idea of using designs to maximize the utility of data collection is also studied extensively in the combinatorics community [2]. Many of these ideas are not as directly applicable to our work, as we work with "found" observational data rather than designing studies to be tabulated. Furthermore, they focus almost exclusively on binary (or at the very least, discrete) treatments and outcomes and are unsuitable for the types of continuous data sets studied in machine learning.

A small but active research community in computer science, statistics, and philosophy has focused on moving from correlational studies to causal ones, utilizing Bayes nets with causal semantics [14, 18]. This research is providing a solid theoretical foundation for causal analysis, but it focuses almost entirely on propositional data sets that don't explicitly represent relations or time. Extending these ideas to relational-temporal domains is a promising area of future investigation.

Several branches of the social sciences utilize statistical models based on the hierarchical nature of many data sets to infer causality; these include hierarchical linear models [15], structural equation modeling [13], and multi-level modeling [7]. However, each of these approaches consistently, and implicitly, relies on one specific design or set of designs. They do not define a space of designs or choose among the applicable designs.

6. Conclusions

In this paper, we have formally defined designs with respect to relational databases and the relational algebra, which naturally compose to form a searchable space of designs. Given a formally defined search space, it is clear that we can develop algorithms to identify and evaluate designs. We have shown that large numbers of design alternatives exist for a representative sample of causal questions, which suggests that algorithms could significantly aid investigators engaged in causal discovery. Additionally, we have presented quantitative evidence that designs employing sampling, blocking, and statistical control can limit the set of possible causal models and improve the statistical power of causal discovery. The utility of designs for causal discovery indicates that algorithms identifying such designs would be very beneficial. We believe this is a major area of future research, and we are actively pursuing algorithms to search the design space and automatically apply designs to discover causal knowledge.

7. Acknowledgments

We thank Cynthia Loiselle for her assistance in preparing this manuscript. This material is based on research sponsored by the Air Force Research Laboratory and the Intelligence Advanced Research Projects Activity (IARPA), under agreement number FA8750-07- 2-0158. 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.

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley New York, 1990.

[2] R. Bailey. *Association Schemes: Designed Experiments, Algebra and Combinatorics*. Cambridge University Press, 2004.

[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. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1300–1309, August 1999.

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

[8] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[9] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pages 243–252, 2007.

[10] D. Jensen. Beyond prediction: Directions for probabilistic and relational learning. In *Inductive Logic Programming, Seventeenth International Conference, Revised Selected Papers*, number 4894, pages 4–21. Springer, Berlin, Corvallis, OR, 2008.

[11] D. Jensen, A. Fast, B. Taylor, and M. Maier. Automatic identification of quasi-experimental designs for discovering causal knowledge. In *Proceedings of the Fourteenth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 372–380, 2008.

[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] R. Kline. *Principles and Practice of Structural Equation Modeling*. The Guilford Press, 2005.

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

[15] S. Raudenbush and A. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Thousand Oaks, CA, 2002.

[16] P. Rosenbaum. *Observational Studies*. Springer, 2002.

[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] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *Proceedings of the First International Conference on Web Search and Web Data Mining*, pages 171–182, 2008.

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

[22] D. Wilkinson and B. Huberman. Assessing the value of cooperation in Wikipedia. *First Monday*, 12(4), April 2 2007.