Automatic Identification of Quasi-Experimental Designs for Discovering Causal Knowledge

David D. Jensen, Andrew S. Fast, Brian J. Taylor, Marc E. Maier
Knowledge Discovery Laboratory
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
413-545-9677
{jensen,afast,btaylor,maier}@cs.umass.edu

ABSTRACT

Researchers in the social and behavioral sciences routinely rely on quasi-experimental designs to discover knowledge from large databases. Quasi-experimental designs (QEDs) exploit fortuitous circumstances in non-experimental data to identify situations (sometimes called "natural experiments") that provide the equivalent of experimental control and randomization. QEDs allow researchers in domains as diverse as sociology, medicine, and marketing to draw reliable inferences about causal dependencies from non-experimental data. Unfortunately, identifying and exploiting QEDs has remained a painstaking manual activity, requiring researchers to scour available databases and apply substantial knowledge of statistics. However, recent advances in the expressiveness of databases, and increases in their size and complexity, provide the necessary conditions to automatically identify QEDs. In this paper, we describe the first system to discover knowledge by applying quasi-experimental designs that were identified automatically. We demonstrate that QEDs can be identified in a traditional database schema and that such identification requires only a small number of extensions to that schema, knowledge about quasi-experimental design encoded in first-order logic, and a theorem-proving engine. We describe several key innovations necessary to enable this system, including methods for automatically constructing appropriate experimental units and for creating aggregate variables on those units. We show that applying the resulting designs can identify important causal dependencies in real domains, and we provide examples from academic publishing, movie making and marketing, and peer-production systems. Finally, we discuss the integration of QEDs with other approaches to causal discovery, including joint modeling and directed experimentation.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

SIGKDD 2008, August 24–27, 2008, Las Vegas, NV, USA.
Copyright 2008 ACM 1-58113-000-0/00/0008…$5.00.

General Terms

Algorithms, Design, Experimentation, Languages, Theory.

Keywords

Quasi-Experimental Design, Causal Discovery.

1. INTRODUCTION

Quasi-experimental designs are a staple of research in the social and behavioral sciences, economics, and medicine. A quasi-experimental design (QED) is an approach to data analysis that exploits fortuitous characteristics of the data that allow the equivalent of experimental control or randomization [4][5][19]. QEDs are sometimes called "natural experiments" because they emulate the conditions that allow investigators to infer causal dependencies from small amounts of data using laboratory experiments.

QEDs can be a powerful tool for inferring causal knowledge, but applying these designs is a painstaking manual affair, requiring extensive knowledge of both the data and the conditions under which a QED can be applied. As a result, the opportunity to apply QEDs can be missed by investigators, even though such application requires no additional data collection. Investigators must cull through their data schema with great care to identify situations in which QEDs can be applied, and opportunities for causal inference using these designs can go unrecognized.

In addition, the opportunities to apply QEDs in more general settings have increased dramatically in recent years [11]. First, the expanding complexity of databases is increasing the number of situations that match the conditions necessary to apply a QED. Second, the expanding size of databases is increasing the probability that the subsets of data used by QEDs will provide the necessary statistical power to identify subtle causal dependencies. Finally, the availability of new data and knowledge representations, particularly relational and temporal representations, and their associated inference methods, is making it possible to reason automatically about the preconditions for applying QEDs.

In this paper, we report the first instance of a fundamentally new approach to knowledge discovery in databases. We describe and evaluate a proof-of-concept system that shows how QEDs can be identified automatically. We show that QEDs can be found using only a relational database schema, additional information about the temporal durations of specific events, and limited prior knowledge about potential causes. We report several representational innovations that facilitate automated discovery of

QEDs, including automatic construction of event streams and aggregated variables. Finally, we apply the approach to several data sets and discover non-trivial and useful causal dependencies based on the identified QEDs.

1.1 Example

Many of the ideas that led to this work are illustrated by recent studies published in the sociology literature. In late 2007, sociologists at The Ohio State University published a paper in the *Journal of Youth and Adolescence* [1]. The paper reported that early sexual activity among adolescents increases their risk of delinquency.[1] The study not only concluded that the two types of behavior are statistically correlated, but that early sexual activity actually *causes* an increase in delinquency. Their findings indicated that, to reduce delinquency, public programs should focus, at least in part, on efforts to reduce early sexual activity.

The study was based on a mostly manual analysis of a large and complex data set — the National Longitudinal Study of Adolescent Health — commonly known to researchers as "Add Health". The data set resulted from a nationally representative study that explored health-related behaviors of more than 15,000 adolescents in grades 7 through 12 and their outcomes in young adulthood [21]. The Add Health data allow researchers to examine how adolescents' experiences and social contexts (families, friends, peers, schools, neighborhoods, and communities) influence their health and risk behaviors. The data have been used in more than 1,000 published reports and journal articles.

The Ohio State researchers were trying to do something very difficult: draw causal conclusions from non-experimental data. One primary challenge of such work is that, to conclude that a statistical association between two variables indicates causal dependence, the analysis must account for the effects of all common causes of the variables. To address this concern, the researchers attempted to control for potential common causes of both sexual activity and delinquency by modeling those effects and mathematically removing them, an approach often called "statistical control." The potential common causes they modeled included gender, race, parental education, receipt of public assistance, family structure, prior delinquency, depression, school grades, parental support, illegal substance use, relative physical development, dating experience, and virginity pledge status.

Unfortunately, it appears that these controls were inadequate. Another study, completed by researchers at the University of Virginia at Charlottesville [8] and first released just a few months later, came to a very different conclusion. The study found that genetic and environmental differences between families explained the statistical association between early sexual experience and delinquency. Indeed, after controlling for these genetic and environmental causes, early sexual experience predicted slightly *lower* levels of delinquency in early adulthood rather than the higher levels that the Ohio State researchers had found. The findings suggest that some other factor, perhaps genetics, increases risk-taking behaviors, including both sexual activity and delinquency. The study was so convincing that the authors of the first study later agreed to this reinterpretation of their findings [22].

Why was the later study so convincing? Rather than apply statistical controls, the second group of researchers applied a quasi-experimental design. Specifically, they identified 534 same-sex twin pairs in the Add Health data. Twins share similar or identical genetics (depending on whether they are fraternal or identical twins) and similar fetal and early childhood environments. As a result, studying twins provides a way to control for genetic and environmental factors without the need to explicitly identify and model the effects of these factors. In the case of the Add Health data, focusing on twins allowed the Virginia team to control for additional factors that were not successfully measured or controlled in the initial study.

This example points to one of the key insights about QEDs for knowledge discovery: analyzing only a *subset* of all available data can increase the validity of the resulting conclusions, provided that subset meets some highly specific conditions. Indeed, such subsets can provide evidence for stronger conclusions about causality than can an uninformed analysis of the entire data set.

2. CAUSAL INFERENCE AND QEDS

QEDs make it possible to discover *causal knowledge* from observational data. Here, causality means the assertion that dependence exists between *A* (the cause) and *B* (the effect) such that manipulation of the cause will result in the manipulation of the effect. Causation implies that varying *A* will make *B* vary.

Causal knowledge differs substantially from the type of knowledge identified by most knowledge discovery algorithms, which captures only *statistical associations*. Classification trees, association rules, support vector machines, Bayesian classifiers, and nearly all other types of statistical models constructed by knowledge discovery algorithms make no commitments about causality. They only attempt to represent statistical associations. Knowing the value of *A* will help you *predict* the value of *B*, but *changing* the value of *A* may or may not affect the value of *B*.

Causal knowledge has unique advantages over knowledge that identifies only statistical association. Causal knowledge is actionable in ways that statistical associations are not. A statistical association between two variables *A* and *B* could indicate that *A* causes *B*, that *B* causes *A*, or that some third variable C causes both *A* and *B*. If the goal is to affect the value of *B*, each of these situations implies different actions. For additional details, see a recent discussion with examples of causal knowledge discovery [11].

In addition, causal knowledge provides a more compact representation of knowledge about the associations among a set of variables. Rather than showing a complex pattern of statistical associations among a set of variables, causal models show a much smaller set of causal dependencies from which the larger set of statistical associations can be derived.

2.1 Causal Inference

Inferring causal dependencies from data is strictly more difficult than identifying statistical associations. Classically, the inference that *A* causes *B* relies on three conditions:

- *Association* — The values of *A* and *B* are statistically associated.

[1] "Delinquency" refers generally to illegal acts by minors, including those applicable only to minors, such as truancy and alcohol use. In these studies, researchers assessed delinquency by scoring self-reported incidents of graffiti, property damage, shoplifting, other theft, and drug dealing.

- *Direction* — The direction of causality is known (e.g., based on temporal criteria).
- *No common causes* — The effects of all common causes of *A* and *B* have been eliminated.

The challenge of eliminating common causes is particularly daunting, and different methods for causal inference approach this challenge in different ways.

One approach is to employ a classical experiment in which researchers can explicitly affect the conditions under which data are gathered. In classical experiments, researchers use both control and randomization to eliminate the effects of potential common causes [7]. Control holds potential confounding variables constant so that they cannot affect the experimental outcome, and randomization assigns experimental subjects to treatments randomly so that potentially confounding variables cannot systematically affect outcomes. Randomization is particularly powerful, because it can eliminate or implicate entire classes of variables as potential common causes, even if an investigator has never defined or measured those variables. For example, an experimenter need not know which characteristics of an experimental subject (e.g., a medical patient) might be a common cause of *A* and *B* outside of the experimental context, as long as subjects are randomly assigned to groups receiving different treatments (values of *A*) within the experiment. The random assignment of *A* removes the potential for *any* variable to be a common cause of both *A* and *B*.

However, investigators often wish to infer causality in situations that are not amenable to classical experiments, either for logistical or ethical reasons. An alternative to an experiment is to use observational (non-experimental) data and to identify, measure, and model potential common causes of *A* and *B*. With an accurate model, an investigator can mathematically remove the effects of common causes and then ascribe any unexplained association to the causal effect of *A*. This approach has been pioneered over the past several decades by researchers in several fields, including statistics [9][10][18], computer science [16], and philosophy [20]. Successfully applying this approach requires identifying and measuring all potential common causes, an assumption referred to as "causal sufficiency." The Ohio State team took one version of this approach in their study (although their analysis appears to have violated the causal sufficiency assumption).

Another approach to analyzing observational data is to apply quasi-experimental designs. QEDs identify configurations of the data that provide the equivalent of control or randomization (sometimes called "pseudo-control" or "pseudo-randomization"). These designs were pioneered by sociologist Donald T. Campbell and his colleagues, beginning in the 1960s [4][5][19], and they have since been used in thousands of published papers in the social sciences, economics, and medicine. QEDs employ a variety of methods to emulate control and randomization. For example, one design (the non-equivalent control group design) attempts to identify two sets of data instances that have similar responses to temporal events but that differ in whether they experience a given treatment event. Another (the regression-discontinuity design) models the combined effect of both a discrete treatment and another variable that determines which units receive treatment. Other types of quasi-experimental designs that have been devised include the proxy pretest design, double pretest design, non-equivalent dependent variables design, pattern matching design, and the regression point displacement design [5].

Nearly all QEDs can be thought of as exploiting the temporal or relational structure of the world to provide quasi-control or quasi-randomization. For example, the twin design employed by Harden et al. exploits the fact that two different individuals share (are related to) a common genotype. As a result, systematically examining behavioral differences between twins can control for the vast array of genetic factors that could affect behavior. Similarly, the several QEDs exploit the fact that the characteristics of a single entity (say, a company) are likely to remain relatively stable over short time-periods. This fact facilitates the inference that an external event causes company behavior if that behavior changes substantially just after the event.

2.2 Advantages and Disadvantages of QEDs

QEDs have a number of advantages over statistical control or classical experiments. First, QEDs can surpass the validity of attempts at statistical control because they can control for entire classes of variables, even though those variables are not identified, measured, or modeled. Statistical control requires all three of these things, while some QEDs identify subsets of data for which the relational structure assures that entire classes of variables will be controlled. For example, the Virginia team was able to use a QED to factor out variables characterizing genetics and early family environment, even though they did not specifically identify, measure, or model any variables in these sets. Rather, they relied on the fact that twins had identical genetics and early family environment, regardless of which aspects of these factors might have influenced sexual activity or delinquency.

Second, QEDs can surpass the validity of controlled randomized experiments because they apply to data collected "in place" rather than in an artificial laboratory setting. While laboratory experiments often allow exquisite levels of control and randomization, these advantages are often purchased at a high price by introducing artificiality into the study. Thus, while experiments have higher "internal validity" (they are internally consistent), they can sacrifice "external validity" (ability to generalize to the real world) [4]. Conclusions reached by QEDs typically have higher external validity than the corresponding experimental study, though they may sacrifice some degree of internal validity.

Third, QEDs do not require the collection of additional data. Instead, investigators can apply them to an existing data set and draw strong causal conclusions. Indeed, as we show below, the *identification* of a QED does not require *any* data collection, only a specification of a data schema. This means that designs can be identified in advance and then used to guide data collection.

Finally, QEDs do not preclude alternative methods for causal inference. Indeed, they can serve as a valuable adjunct to statistical control (by eliminating or identifying potential causal relationships) and to experiments (by limiting the number of potential dependencies that must be experimentally evaluated).

That said, QEDs also suffer from several limitations. First, the designs are only applicable in a very limited number of situations. The increasing size and complexity of relational databases offer expanding opportunities for applying QEDs, but still only a small fraction of causal dependencies will be amenable to examination by these designs.

Second, because many QEDs use only a subset of the data to infer causal dependencies, the validity of their conclusions relies on the representativeness of that subset. For example, twin studies rely on the assumption that twins do not differ substantially from non-twins with respect to the characteristics under study. This assumption has been largely valid in the past, because twins occurred relatively randomly within the population of all births. However, in the past two decades, fraternal twins have become far more common due to the use of in vitro fertilization. These children tend to be born to older mothers, among other differences, and thus may differ systematically from non-IVF children.

3. AUTOMATED DISCOVERY OF QEDS

QEDs are widely used because of their advantages and despite their limitations. As one rough indicator, Google Scholar lists over 4,500 citations to each of two classic texts on the subject [4][5] and lists over 20,000 papers that use the terms "quasi-experimental design" or "quasi-experiment". However, these uses are entirely manual — investigators identify the potential to apply a QED based on their own knowledge of the data and of QEDs, and they do so without help from an automated system. We are unaware of any prior work on automated or semi-automated systems for identifying applicable QEDs based on information about a database.

That said, the potential benefits of an automated system are large. First, such a system would allow automated checking of large and complex schemas for applicable QEDs. As the example in the introduction shows, it is easy for even experienced investigators to overlook important opportunities to apply QEDs. An automated system could alert researchers to applicable designs with relatively little work on their part.

Second, an automated system would allow easy rechecking when changes occur to a database's schema or the knowledge of potential causal dependencies. When one study confirms or disproves a given causal dependence, it is not currently easy to assess the wider implications of this finding for the applicability of QEDs for other potential dependencies. An automated system could continuously evaluate the impact of new findings and alternative sets of assumptions made by an investigator on applicable QEDs. This, in turn, could significantly aid the process of collaboratively constructing knowledge bases (e.g., [17]).

Finally, an automated system for identifying QEDs would allow these methods to be integrated with other automated methods for causal modeling [16][18][20]. These methods learn the structure and parameters of a joint probability model of a large collection of variables. While research on these methods continues, they face a large number of challenges both in terms of accuracy and computational complexity. These challenges could be partially addressed by applying QEDs whenever possible to identify key dependencies, reduce the size of the search space, and reduce the sample complexity of learning. This is particularly true as the complexity of models increases, as it has with the relatively recent advent of relational and temporal models.

3.1 The AIQ Algorithm

To evaluate whether automated discovery is possible and realistic, we have developed AIQ (for "Automated Identification of Quasi-experiments," pronounced "a-eye-que"), a system for reasoning about the applicability of QEDs to specific data sets. AIQ (v. 1.0) is implemented in SWI Prolog.[2] Source code and Prolog-encoded inputs for several public databases are available from our website.[3]

The algorithm takes as input a relational database schema augmented with information about the temporal extent of specific types of entities and relationships, along with information about the potential causes of specific variables on those entities and relationships. From this information, AIQ constructs several types of intermediate representations, including temporal streams of events, aggregated variables on those streams, and units that identify the data entities that will be used to test specific causal dependencies. There are a large number of combinations of these intermediate representations, and each combination applies to only a handful of potential QEDs. The algorithm checks which combinations of these intermediate representations correspond to valid QEDs, and outputs all such designs. The investigator can then evaluate the validity of each design and run appropriate statistical tests based on them. Alternatively, designs output by the algorithm may indicate flaws in the database schema or ancillary information, in which case the investigator can modify the database-specific information and iterate.

3.2 Relational Database Schemas

In section 4, we provide results of running AIQ on several data sets, for which we encoded their database schema and ancillary information in first-order logic. For convenience, we illustrate those schemas through augmented Entity-Relationship (ER) diagrams [6]. Specifically, we represent the ER diagram using a slightly modified Barker Notation [2] where entities are rectangular boxes and the relationship between two entities is a solid connecting line with the cardinality of the relationship represented as symbols on both ends of the connection. Entities can be related through one-to-one, one-to-many, or a many-to-many cardinality. For example, given the two entities paper and conference and a relationship where a paper appears in one conference but a conference has many papers would be considered as many-to-one cardinality. In Barker Notation, open circles are used to identify zero, a vertical bar or dash is used to identify one, and a symbol where three lines intersect represents many. For simplicity, we do not use cardinality of zero, and cardinality of one is represented by an absence of a symbol.

Many QEDs rely on knowledge about the existence of events and the temporal extent of entities. ER diagrams do not specify such temporal characteristics of data, so we must augment the schema with this information. Temporal extent identifies the average lifetime of an object and is used in causal modeling to identify if a stream associated with an entity can be a valid cause for some observed effect. Temporal frequency identifies how often one object changes in relation to another object.

Our temporal ER diagrams identify the extent duration of entities by adding a time label inside of the entity box (see section 4.1). To introduce temporal frequency, we have extended the ER diagram to include a frequency label that annotates one-to-many relationship between entities. Where a many-to-many relationship exists, two directed temporal frequencies would be provided, one on either side of the many-to-many relation. For example a Store may have many Customers, and the temporal frequency from the Store to Customers would represent how often the Store receives a new Customer. In cases where the frequency is varied across

[2] http://www.swi-prolog.org

[3] http://kdl.cs.umass.edu/causality/

different objects of the same type, an average frequency is used to represent the set. For all temporal annotations in this paper, we measured frequencies in days.

3.3 Potential Causes

In addition to temporal annotations on entities and relations, individual variables in AIQ are annotated with information about their potential causes. By default, every other variable in the database is a potential cause, but prior knowledge of investigators and existing analysis approaches (e.g., joint modeling) can be used to prune the set of potential causes. In the most extreme case, an investigator might know that a variable has no causes; in the language of QEDs, this variable would be considered *quasi-random*. For example, some government programs that provide services to individuals (e.g., job training) are allocated by lottery among all qualified applicants. Clearly, such situations represent highly useful knowledge for identifying potential QEDs.

3.4 Streams and Aggregated Variables

From the database schema and temporal annotations, AIQ automatically constructs *streams* that represent series of events occurring over a period of time with respect to a given static entity. For example, scientific researchers typically produce multiple papers at irregular intervals each year; from the perspective of the researcher, there exists a stream of papers over their entire career. AIQ automatically identifies potential streams based on the augmented database schema. Streams are defined as pairs of entities and an associated connecting path. Valid streams must match three conditions: (1) The two entities must be connected by a path of relationships; (2) The first entity must have a one-to-many relationship with the second entity; and (3) The first entity must have an extent longer than the average frequency along the stream.

These conditions define the structure of a stream — a base entity, followed by a sequence of zero or more items on a relationship path, followed by an entity that is "dynamic" with respect to the base entity. It also gives rise to the algorithm in which streams can be recursively defined by enumerating all candidate paths. Additionally, in a particular direction, a "many" relationship will be propagated for the remainder of the path. For example, if item A is connected to B through a one-to-many relationship, and item B is connected to item C through a many-to-one relationship, then item A has a one-to-many relationship with item C.

Given a stream, AIQ also constructs a set of variables on those streams that can be used in one or more designs. Stream variables include aggregations of variables on the dynamic entity, such as the average box office receipts of movies released by a given studio (the dynamic and base entities, respectively) or the sum of the page counts of articles published by a given author. Stream variables also include existence variables, which allow QEDs to identify causal influences on the occurrence of an entity. Any variable defined on an entity that becomes the dynamic entity of a stream can be aggregated into a stream variable.

3.5 Identification of Units

The notion of an experimental unit is crucial for developing effective quasi-experimental designs [19]. Essentially, a unit defines the boundaries for possible causes in an experiment. For example, in a clinical trial, a unit would be a person receiving a drug treatment. However, in the relational setting a unit can be any collection of related entities. We automatically define our units based on generated streams. The base item common to the two streams in a proposed design is the core of the unit, and the dynamic items, as well as the items on the paths in the streams, compose units in our QEDs. Additionally, a unit is defined with respect to a specific period of time, which is also automatically derived from the temporal frequencies provided through the schema description.

3.6 Identification of Designs

AIQ combines input information (the database schema and potential causes) with information it constructs (potential streams and units) to identify and evaluate potential QEDs. In version 1.0, we focus on a single class of QEDs — *the non-equivalent control group design*. This design combines relative simplicity, wide applicability, and intuitively understandable results. Future versions of AIQ will implement a much wider array of classes of QEDs.

The non-equivalent control group design, also called the "non-equivalent comparison group design," is similar to one of the most widely used experimental designs (the "pretest-posttest control group design"). The experimental version of this design randomly assigns units to either a control group or a treatment group, measures a set of variables on all units (the pretest), then administers treatment to only the treatment group, and measures the variables again (the posttest). In the quasi-experimental version of the design, investigators cannot assign treatment randomly. Instead, in the non-equivalent control group design, treatments are assigned non-randomly and causal inferences are based on the relative rates of change of the two groups (e.g., the threats to validity might be judged low if the treatment group's observed response increases sharply after treatment and the control group's response remains unchanged) [19]. That said, AIQ currently also requires that all potential common causes be ruled out (a condition most frequently met by treatments being designated as quasi-random), to provide additional confidence that treatment and control groups are similar.

AIQ identifies an instance of the non-equivalent control group design whenever: (1) two distinct stream variables — potential cause A and potential effect B — can be defined with respect to the same unit; (2) A and B have no common causes (e.g., A is quasi-random); and (3) the entities defining A and B match certain temporal constraints (e.g., causes occur less frequently than effects). Such cases correspond to the canonical non-equivalent control group design, in which a potential causal event happens rarely enough that measurements of effects can be made both before (pretest) and after (posttest) the potential causal event. When identified, instances of this QED are output for inspection and evaluation by the investigator.

3.7 Hypothesis Tests

The QED specifications identified by AIQ provide sufficient information that formulating a statistical test for causal dependence is fairly simple, given that data are readily available. In theory, the test could be done automatically by AIQ, but there were a sufficient number of implementation decisions that could crucially affect the validity of the hypothesis test that we chose to leave such tests outside the boundaries of automation (see section 4.3 for an example). These include questions of sampling, test statistic, and aggregation method for stream variables. Instead, AIQ 1.0 identifies the QED in sufficient detail that statistical tests can be run fairly easily by the investigator.

4. RESULTS

To evaluate the utility of the AIQ algorithm, we applied it to three public databases that are widely used in the KD research community and that have reasonably complex relational schemas. While it is not a simple matter to objectively assess the performance of the algorithm, we report both subjective and quantitative results on these data sets. We provide several detailed examples, including one discovered causal dependence and one case of iterative refinement of the database schema to identify more interesting and useful QEDs.

4.1 Data Sets

HEP-Th — HEP-Th is a bibliographic database of papers from the ArXiv.org repository. Originally published as part of the 2003 KDD Cup competition,[4] the data set contains preprints from 1992 until 2003, with over 30,000 papers, 13,000 authors, and 500,000 links among them. Figure 1 shows the relationship between authors, their submissions and papers, journals, and citations. We have represented the submissions and specific author credits as separate entities in the data set.

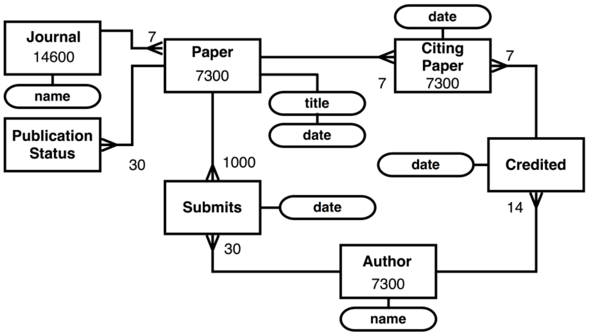

Figure 1: Entity-Relationship diagram with temporal frequencies and extents for the HEP-Th database. Authors make submissions to ArXiv that may or may not become papers for a journal and papers can refer to other papers giving authors credits. We assume that authors make submissions to ArXiv about once a month, that it takes a submission about 3 years to undergo journal review and that the status of the paper while in review can change monthly. The remaining frequencies are estimates used to reflect the dynamic nature of citations. We assume that authors and papers are a part of the repository for at least 20 years and that journals last even longer.

IMDb/Netflix — The Internet Movie Database[5] contains information on movies released worldwide, including release dates, directors, producers and actors, as well as the nominees and recipients of Academy Awards. We selected a subset of these awards covering films released in the years 1997 to 2007. We included information on the nominees and winners of Best Picture, Best Director, Best Actor, and Best Actress. We augmented the IMDb data with the Netflix Prize data set,[6] which contains the title and year of release for 17,770 movies released on DVD and ratings of those movies from more than 400,000 customers. The date range for ratings is from November 11, 1999 to December 31, 2005. The schema shown in figure 2 represents the combination of the two data sets.

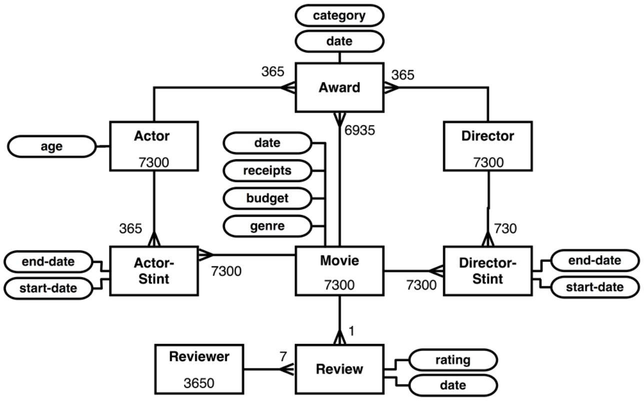

Figure 2: Entity-Relationship diagram with temporal frequencies and extents for the IMDb+Netflix database. Each movie has a series of actor and director stints as well as a review by a user of the Netflix Prize database. Awards are presented to actors, directors, and movies. We assume that movies, actors, and directors last as long as the database itself. Reviews can occur daily and award ceremonies occur once a year. Actors work on two films a year, directors make one movie per year, and once an actor or director works on a film, that information never changes.

Wikipedia — Wikipedia[7] is a collaborative peer-production system with the ultimate goal of providing free encyclopedia content to everyone. The database consists of millions of articles maintained by thousands of users. Anyone that registers can become a user and edit any page. Consequently, vandalism is inevitable, and occasional misinformation is provided. Thus, actions on users and pages frequently occur (e.g., users can have privileges revoked, pages can be restricted), and these events are stored in logging tables. The schema in Figure 3 presents the relationships among users editing articles, as well as the specific logging events that may take place. The main temporal assumption for this data set is that logging and editing events are frequent enough to assume they occur daily.

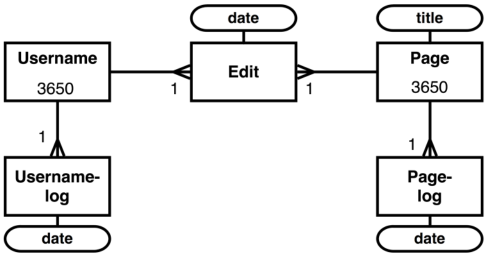

Figure 3: Entity-Relationship diagram with temporal frequencies and extents for the Wikipedia database. This simple view shows how users can edit pages, and logs can be created to record the activity in the database. The assumption that editing and logging events can occur daily is reflected in the temporal frequency labels. We assume that users and pages last for the duration of the Wikipedia database.

[4] http://www.cs.cornell.edu/projects/kddcup/

[5] http://www.imdb.com/

[6] http://www.netflixprize.com/

[7] http://www.wikipedia.com/

4.2 Focusing on Plausible QEDs

From a large number of possible designs, determined by the cardinality of entities, relations, and variables, AIQ identifies a relatively small number of plausible designs. To evaluate the degree to which the algorithm can focus the attention of investigators on plausible designs, we compared the total number of available designs to those selected by AIQ. Table 1 shows data on the total available QEDs, the total number actually identified, and the breakdown of those designs by base entity. The larger the number of possible QEDs, the greater the percentage reduction in the QEDs actually identified by AIQ as plausible.

Table 1: Identified and Possible QEDs

Database Name	Base Entities	Number QEDs	Unique QEDs	Available QEDs	Percent
IMDb + Netflix	Total	102	79	2028	3.9%
	Movie	30	13		
	Director	30	27		
	Actor	42	39		
HEP-Th Original	Total	60	60	560	10.7%
	Author	38	38		
	Paper	22	22		
HEP-Th Extended	Total	97	97	830	11.7%
	Author	49	49		
	Paper	48	48		
Wikipedia	Total	44	44	192	22.9%
	User	24	24		
	Page	20	20		

4.3 Identifying Useful QEDs

The next reasonable question is the degree to which AIQ's identified designs are actually useful for identifying previously unknown causal dependencies. We selected and evaluated several of the QEDs identified by the algorithm. The majority showed neither statistically significant associations nor sufficient statistical power to conclude independence between the variables. However, several were statistically significant.

For example, one instance of a QED identified by AIQ on the IMDb/Netflix data involves the variables of award existence and an aggregate of user ratings on a base item of movies. This design implies, rather intuitively, that the granting of an Academy Award to a movie may cause changes in user ratings of that movie. This design was made possible because whether an award entity exists was designated as pseudo-random among all nominated movies (i.e., all nominated movies are equally likely to win an award). This is clearly an assumption, but a plausible one.

We test this design by computing the average rating a movie receives in the two months prior to and the two months after Academy Awards are granted. For each movie, we computed the difference in the average ratings. Then we compared the mean difference for movies that won an award with the mean difference for those who were nominated but did not win.

This general configuration of a hypothesis test is directly implied by the matching QED (the non-equivalent control group design), though the details still require prior knowledge of the movie domain not encoded within the domain-specific knowledge base. For example, we restrict the ratings to a window of two months since movies are not generally released on DVD until shortly before the Academy Award ceremony. Additionally, since the major nominations for the Academy Awards consist of best picture, best director, best actor, and best actress, we only consider these four categories as potential influences on user ratings.

We perform a two-sample t-test on the differences in average ratings for these two populations of movies. The average rating decreases by 0.066 for movies that win an award compared to a drop of 0.247 for those that do not. This difference is weakly significant ($p=0.07$; $N_1=14$, $N_2=47$) indicating a causal relationship between winning an award and user ratings. The relative differences in the two populations could be due to a variety of underlying mechanisms, including anchoring (the Academy Award confers a high initial rating that raters are loath to change). Note that both populations see an overall decrease in average ratings, which could be due to regression toward the mean (early raters of movies tend to give higher ratings), unreasonably high expectations (award-winning movies are expected to very good), or seasonal effects.

4.4 Interactive Refinement of Schemas

The potential utility of AIQ goes beyond a simple one-shot analysis of an input schema. The algorithm can be used in an iterative manner to refine a schema to become more useful for causal discovery.

For example, the initial HEP-Th schema consisted of entities for journals, authors, papers, citations, credits, and submissions and the relations between them (see figure 1). For this schema, AIQ generated a set of 60 possible QEDs. Many of the designs suggested that variables on submission caused variables on citation. While this might be plausible, another potential cause seemed more likely.

We added an entity to the schema called *publication status*. These entities represent events in the life of a paper, including being in review, accepted, rejected, or published, and each status entity occurs at a specific point in time. The addition of the publication status entity increased the number of possible QEDs to 97. We again reviewed the set of designs, and several of them now indicated that changes in paper status could causally influence the existence of citations, an entirely plausible and interesting design.

To evaluate this design, we selected a list of papers from HEP-Th that were published at least one year after they were first submitted. For each of these papers, we counted the number of citations the paper received during the first year it waited for publication and for the two years after publication. Then we selected the set of papers that were submitted but were not published. For each of these, we counted their number of citations during the three years after they were first submitted.

We computed the difference in the means of the citation rate for the two time periods for each group of papers. For published papers, the difference in the means indicated that, in the period after publication, the paper's citation rate improved by 40%. For the unpublished papers, the difference also indicated an improvement in their citation rate, but only by 14%. We applied a two-sample t-test to analyze the difference between the two sub-populations. The test indicated a highly significant difference between the citation rates of the published and unpublished papers (p=2.2e-16; N_1=17394, N_2=4559).

However, is this strong evidence for a causal dependence between publication and citation rate? Unfortunately not. Upon reflection, the augmented schema leaves out any measure of paper quality, a potential common cause of both publication and citation. Good papers are both more likely to be accepted for publication and more likely to be cited by other authors. On the one hand, this is precisely the type of situation that AIQ was designed to avoid; on the other hand, AIQ was unable to exclude such a QED because it lacked the information that would have allowed it to identify this possibility. The schema should be changed. While we have no data measuring this variable, we can include it in the schema and eliminate this QED from consideration.

This example only contains two iterations of schema redesign, but it illustrates how, through many such iterations, an investigator could refine the schema to both expand and trim the list of QEDs.

5. PRIOR WORK

The most obvious body of prior work concerns the manual application of experimental and quasi-experimental designs [4][5][7][19]. This covers a long tradition of philosophical writing stretching back to the origins of modern science, as well as work on experimental design since the 1920s and formal quasi-experimental design beginning with the work of Campbell and Stanley [4] and continuing to the present day [19]. We build on this work to produce algorithms to identify QEDs automatically.

Work in cognitive psychology and artificial intelligence has investigated a related area — the processes by which scientific experiments are designed and analyzed. Heuristic search has been used in systems that are capable of rediscovering laws and inventing new ones in fields of science such as physics and chemistry [3][15]. For example, the KEKADA program has been used to model the process by which Hans Krebs developed and executed the experiments necessary to discover the urea cycle [14].

Similar advances have been exploited to automatically plan and conduct actual experiments. For example, researchers have recently automated the nearly the entire process needed to discover gene functions in yeast [13]. This “robot scientist” automatically generates hypotheses from the available data, designs and runs experiments, and analyzes the results. The algorithm's experiment selection has been shown to have equivalent or better performance than humans and vastly improves upon random selection of experiments.

To our knowledge, however, no prior work exists on automatic identification of QEDs for the analysis of non-experimental data. With the increasing use of large-scale systems for data collection, the number and size of observational data sets is growing at unprecedented rates. These data sets provide a rich resource that should be automatically exploited to infer causal knowledge. AIQ offers a first step in that direction.

A second large body of relevant research concerns causal discovery through joint modeling [9][10][16][20]. As mentioned in the introduction, this work differs substantially from the topic of this paper. It uses the entire data set to jointly model the probabilistic dependencies among all variables, rather than selecting subsets of data to control or randomize the effects of large classes of variables and thus allow individual dependencies to be tested with high statistical power.

In addition, nearly all work on joint modeling for causal discovery assumes the data consist of independent and identically distributed (i.i.d.) instances. In contrast, our work (along with much of the work on manual identification of QEDs) assumes that data instances are joined by relations that represent temporal, genetic, organizational, or institutional linkages. These relations imply dependencies between variables on related entities, and they are essential background knowledge to identifying QEDs, whether manually or automatically. Little work on causal discovery exploits these relations (though a notable exception is Karimi & Hamilton [12], who use temporal information to identify causal sequences).

That said, both work in QEDs and joint modeling use similar underlying notions of causality, control, randomization, and statistical inference. In addition, some of the more complex quasi-experimental designs (e.g., the regression point displacement and regression discontinuity designs) rely on some degree of statistical modeling to achieve their effect. There is substantial scope to combine these methods in complementary ways, and perhaps even to unify them into a common framework for causal discovery.

6. DISCUSSION AND FUTURE WORK

The results and examples in this paper demonstrate the potential for automatic identification of quasi-experimental designs. For the first time, an automated program can identify QEDs within large and complex databases. AIQ 1.0 is only the first step toward a more complete and useful tool. While our implemented designs are not perfect, AIQ makes it possible to quickly see the implications of different assumptions, and to evaluate and improve those assumptions. The system provides a “what if” capability for investigators, and facilitates rapid improvement and exploration.

Still, a wide variety of improvements remain. Future versions of AIQ should search for a much wider array of QED types and allow specification of unobserved entities and variables. More extensive changes would involve automated hypothesis testing of the potential causal dependencies and integrating QEDs with joint modeling algorithms that are currently used for causal discovery.

Finally, a much more extensive evaluation is necessary, examining questions such as the breadth of application of QEDs, the proportion of all causal dependencies that are discoverable by QEDs, and the extent to which the use of QEDs facilitates joint modeling (and vice versa).

7. ACKNOWLEDGMENTS

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.

8. REFERENCES

[1] Armour, S. and Haynie, D. 2007. Adolescent sexual debut and later delinquency. *Journal of Youth and Adolescence.* 36, 2, 141-152.

[2] Barker, R. 1990. *CASE*Method: Entity Relationship Modelling.* Addison-Wesley, Boston, MA.

[3] Bradshaw, G., Langley, P., and Simon, H. 1983. Studying scientific discovery by computer simulation. *Science*, 222, 4627, 971-975.

[4] Campbell, D. and Stanley, J. 1963. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally.

[5] Cook, T. and Campbell, T. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Rand McNally.

[6] Chen, P. 1976. The entity-relationship model - Toward a unified view of data. *ACM Transactions on Database Systems* 1, 1, 9-36.

[7] Cochran, W. and Cox, G. 1954. *Experimental Designs*. Wiley, New York.

[8] Harden, K., Mendle, J., Hill, J., Turkheimer, E., and Emery, R. 2008. Rethinking timing of first sex and delinquency. *Journal of Youth and Adolescence* 37, 4, 373-385.

[9] Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association*. 81, 396, 945-960.

[10] Holland, P. and Rubin, D. 1988. Causal inference in retrospective studies. *Evaluation Review* 12, 203–231.

[11] Jensen, D. 2008. Beyond prediction: Directions for probabilistic and relational learning. *Lecture Notes in Computer Science 4894*, 4-21. Springer, Berlin.

[12] Karimi, K. and Hamilton, H. 2003. Distinguishing causal and acausal temporal relations. *The Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2003).* Seoul, South Korea, 234-240.

[13] King, R., Whelan, K., Jones, F., Reiser, P., Bryant, C., Muggleton, S., Kell, D., and Oliver, S. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427, 6971, 247-252.

[14] Kulkarni, D. and Simon, H. 1988. The processes of scientific discovery: The strategy of experimentation. *Cognitive Science* 12, 139-176.

[15] Langley, P. 1981. Data-driven discovery of physical laws. *Cognitive Science* 5, 1, 31-54

[16] Pearl, J. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge.

[17] Richardson, M. and Domingos, P. 2003. Building large knowledge bases by mass collaboration. *Proceedings of the 2nd international conference on Knowledge capture.* 129-137.

[18] Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 66, 5, 689.

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

[20] Spirtes, P., Glymour, C., and Scheines, R. 2000. *Causation, Prediction, and Search*. MIT Press, Cambridge.

[21] UNC Carolina Population Center. 2008. Add Health Home Page. http://www.cpc.unc.edu/addhealth. Accessed on February 27, 2008.

[22] Weiss, R. 2007. Study debunks theory on teen sex, delinquency. *Washington Post*. November 11, 2007, A03.