Scalable Techniques for Mining Causal Structures

Craig Silverstein * Sergey Brin Rajeev Motwani † Jeff Ullman ‡

Department of Computer Science
Stanford University
Stanford, CA 94305
E-mail: {csilvers,brin,rajeev,ullman}@cs.stanford.edu

Contact Author:	Craig Silverstein
Address:	(as above)
Phone:	(650) 723-1787
Fax:	(650) 725-4671
Email:	csilvers@cs.stanford.edu

Abstract

Mining for association rules in market basket data has proved a fruitful area of research. Measures such as conditional probability (confidence) and correlation have been used to infer rules of the form "the existence of item A implies the existence of item B." However, such rules indicate only a statistical relationship between A and B. They do not specify the nature of the relationship: whether the presence of A causes the presence of B, or the converse, or some other attribute or phenomenon causes both to appear together. In applications, knowing such causal relationships is extremely useful for enhancing understanding and effecting change. While distinguishing causality from correlation is a truly difficult problem, recent work in statistics and Bayesian learning provide some avenues of attack. In these fields, the goal has generally been to learn complete causal models, which are essentially impossible to learn in large-scale data mining applications with a large number of variables.

In this paper, we consider the problem of determining *casual* relationships, instead of mere associations, when mining market basket data. We identify some problems with the direct application of Bayesian learning ideas to mining large databases, concerning both the scalability of algorithms and the appropriateness of the statistical techniques, and introduce some initial ideas for dealing with these problems. We present experimental results from applying our algorithms on several large, real-world data sets. The results indicate that the approach proposed here is both computationally feasible and successful in identifying interesting causal structures. An interesting outcome is that it is perhaps easier to infer the *lack of causality* than to infer causality, information that is useful in preventing erroneous decision making.

*Supported by an ARCS Fellowship and NSF Award CCR-9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

†Supported by an Alfred P. Sloan Research Fellowship, an IBM Faculty Partnership Award, an ARO MURI Grant DAAH04–96–1–0007, and NSF Young Investigator Award CCR–9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

‡This work was partially supported by the Community Management Staff's Massive Digital Data Systems Program, NSF grant IRI–96–31952, ARO grant DAAH04–95–1–0192, and grants of IBM and Hitachi Corp.

1 Introduction

In this paper we consider the problem of determining *casual* relationships, instead of mere associations, when mining market basket data. We discuss ongoing research in Bayesian learning where techniques are being developed to infer casual relationships from observational data, and we identify one line of research in that community which appears to hold promise for large-scale data mining. We identify some problems with the direct application of Bayesian learning ideas to mining large databases, concerning both the issue of scalability of algorithms and the appropriateness of the statistical techniques, and introduce some ideas for dealing with these problems. We present experimental results from applying our algorithms on several large, real-world data sets. The results indicate that the approach proposed here is both feasible and successful in identifying interesting causal structures. A significant outcome is that it appears easier to infer the *lack of causality*, information that is useful in preventing erroneous decision making. We conclude that the notion of causal data mining is likely to be a fruitful area of research for the database community at large, and we discuss some possibilities for future work.

Let us begin by briefly reviewing the past work involving the *market basket problem*, which involves a number of *baskets*, each of which contains a subset of some universe of *items*. An alternative interpretation is that each item has a *boolean variable* representing the presence or absence of that item. In this view, a basket is simply a boolean vector of values assigned to these variables. The *market basket problem* is to find "interesting" patterns in this data. The bulk of past research has concentrated on patterns that are called *association rules*, of the type: "item Y is very likely to be present in baskets containing items $X_1, \ldots, X_t$."

A major issue in mining association rules has been that of finding appropriate definitions of "interest" for specific applications. An early approach, due to Agrawal, Imielinski, and Swami [2], was to find a pair of items that occur together often (that is, have high *support*), and also have the property that one item often occurs in baskets containing the other item (that is, have high *confidence*).[1] In effect, this framework chooses conditional probability as the measure of interest. Many variants of this interest measure have been considered in the literature, but they all have a flavor similar to conditional probability. These measures were critiqued by Brin, Motwani, and Silverstein [9], who proposed statistical correlation as being a more appropriate interest measure for capturing the intuition behind association rules.

In all previous work in association rules and market basket mining, the rules being inferred, such as "the existence of item A in a basket implies that item B is also likely to be present in that basket," often denoted as $A \Rightarrow B$, indicate only the *existence* of a statistical relationship between items A and B. They do not, however, specify the *nature* of the relationship: whether the presence of A causes the presence of B, or the converse, or some other phenomenon causes both to appear together. The knowledge of such causal relationships is likely to be useful for enhancing understanding and effecting change; in fact, even the knowledge of the lack of a casual relationship will aid decision making based on data mining. We illustrate these points in the

[1] The support-confidence definition easily extends beyond pairs of items to incorporate sets of arbitrary size.

following hypothetical example.

Example 1 [2]

Consider a supermarket manager who notes that his meat-buying customers have the following purchasing pattern: buy hamburgers 33% of the time, buy hot dogs 33% of the time, and buy both hamburgers and hot dogs 33% of the time; moreover, they buy barbecue sauce if and only if they buy hamburgers. Under these assumptions, 66% of the baskets contain hot dogs and 50% of the baskets with hot dogs also contain barbecue sauce. The manager will find that the association rule HOT-DOGS ⇒ BARBECUE-SAUCE *has both high support and confidence. (Of course, the rule* HAMBURGER ⇒ BARBECUE-SAUCE *has even better confidence, but that is an obvious association.)*

A manager who has a deal on hot dogs may choose to sell them at a large discount, hoping to increase profit by simultaneously raising the price of barbecue sauce. However, the correct causal model (that the purchase of hamburgers causes *the purchase of barbecue sauce) tells us that this approach is not going to work. In fact, the sales of both hamburgers and barbecue sauce are likely to plummet in this scenario, as the customers buy more hot dogs and fewer hamburgers, leading to a reduction in sales of barbecue sauce. If the manager could infer the correct causal model, or even infer that "*HOT-DOGS *causes* BARBECUE-SAUCE*" is not part of any possible causal model, he could avoid a pricing fiasco.* ▮

A basic tenet of classical statistics ([6], [20]) is that correlation does not imply causation. Thus, it appears impossible to infer causal relationships from mere observational data available for data mining, since we can only infer correlations from such data. In fact, it would seem that to infer causal relationships it is essential to collect experimental data, in which some of the variables are controlled explicitly. This experimental method is neither desirable nor possible in most applications of data mining.

Fortunately, recent research in statistics and Bayesian learning communities provide some avenues of attack. Two classes of technique have arisen: *Bayesian causal discovery*, which focuses on learning complete causal models for small data sets [8, 12, 14, 15, 16, 17, 21, 22, 23, 25, 26, 27], and an offshoot of the Bayesian learning method called *constraint-based causal discovery*, which use the data to limit — sometimes severely — the possible causal models [11, 26, 24]. While techniques in the first class are still not practical on very large data sets, a limited version of the constraint-based approach is linear in the database size and thus practical on even gigabytes of data. We present a more flexible constraint-based algorithm, which is linear in the number of records (baskets) in the database, though it is cubic in the number of items in each record. Despite the cubic time bound, the algorithm proves to be practical for databases with thousands of items.

In this paper, we explore the applicability of a constraint-based causal discovery to discovering causal relationships in market basket data. Particularly, we build on ideas presented by Cooper [11], using local tests to find a subset of the causal relationships. In the rest of this section, we discuss causality for data mining in the context of research into causal learning. We begin, in Section 2 with a particular constraint-based algorithm, due to Cooper [11], upon which we build the algorithms

[2] This example is borrowed from a talk given by Heckerman.

presented in this paper. We then enhance the algorithm so that for the first time causality can be inferred in large-scale market-basket problems.

- Section 2 introduces "CCU" inferences, a form of causal structure not used by [11].
- Section 3 discusses weaknesses of the Cooper algorithm, notably a susceptibility to statistical error, and how power statistics such as correlation can be used to mitigate these problems.
- In Section 4 we describe in detail the algorithms we developed for discovering causal relationships, and we also discuss discovery of noncausal relationships, an important technique that filters many statistically unlikely inferences of causality.
- For the first time, we are able to run causality tests on real, large-scale data. In Section 5 we test this algorithm on a variety of real-world data sets, including census data and text data. In the former data sets we discover causal relationships (and nonrelationships) between census categories such as gender and income. In the text data set we discover relationships between words.
- Finally, in Section 6, we discuss possible directions for future research.

1.1 Previous Research in Causality

As we have mentioned, there has been significant work in discovering causal relationships using Bayesian analysis. A *Bayesian network* is a combination of a probability distribution and a structural model that is a directed acyclic graph in which the nodes represent the variables (attributes) and the arcs represent probabilistic dependence. In effect, a Bayesian network is a specification of a joint probability distribution that is believed to have generated the observed data. A *causal Bayesian network* is a Bayesian network in which the predecessors of a node are interpreted as directly causing the variable associated with that node.

In Bayesian learning techniques, the user typically specifies a prior probability distribution over the space of possible Bayesian networks. These algorithms then search for that network which maximizes the posterior probability of the data provided. In general, they try to balance the complexity of the network with its fit to the data.

The possible number of causal networks is severely exponential in the number of variables, so practical algorithms must use heuristics to limit the space of networks. This process is helped by having a quality prior distribution, but often the prior distribution is unknown or tedious to specify, particularly if the number of variables (i.e., items) is large. In this case, an uninformative prior is used. Even when informative priors are available, the goal of finding a full causal model is aggressive, and Bayesian algorithms can be computationally expensive. While improved heuristics, and the use of sampling, may make Bayesian algorithms practical, this has yet to be demonstrated for data sets with many variables.

In our view, inferring *complete* causal models (i.e., causal Bayesian networks) is essentially impossible in large-scale data mining applications with thousands of variables. For our class of applications, the so-called *constraint-based causal discovery method* [24, 26] appears to be more useful. The basic insight here, as articulated by Cooper [11], is that *information about statistical independence and dependence relationships among a set of variables can be used to constrain (sometimes significantly) the possible causal relationships among a* subset *of the variables.* A simple example of such a constraint is that if attributes A and B are independent, then it is clear that there is no causal relationship between them. It has been shown that, under some reasonable set of assumptions about the data (to be discussed later), a whole array of valid constraints can be derived on the causal relationships between the variables.

Constraint-based methods provide an alternative to Bayesian methods. The PC and FCI algorithms [26] use observational data to constrain the possible causal relationships between variables. They allow claims to be made such as "X causes Y," "X is not caused by Y," "X and Y have a common cause," and so on. For some pairs, they may not be able to state the causal relationship.

These constraint-based algorithms, like the Bayesian algorithms, attempt to form a complete causal model and therefore can take exponential time. (Due to the complexity of their causal tests, they may also be less reliable than simpler algorithms.) Cooper [11] has described an algorithm called LCD that is a special case of the PC and FCI algorithms and runs in polynomial time. Since our algorithm is based on Cooper's, we discuss it in some detail in Section 2.

1.2 Causality for Market Basket Analysis

Finding causality in the context of data mining is particularly difficult because data sets tend to be large: on the order of megabytes of data and thousands of variables. While this large size poses a challenge, it also allows for some specializations and optimizations of causal algorithms, holding out promise that algorithms crafted particularly for data mining applications may yield useful results despite the large amount of data to be processed. This promise holds particularly true for market basket data, which is an even more specialized application. Below, we note some issues involved with finding causality in the market basket case.

- Market basket data is boolean. This may allow for added efficiency over algorithms that work with discrete or continuous data. As we shall see in Section 3.1, some statistical tests that are essential to constraint-based causal discovery have pleasant properties in the boolean case.

- The traditional market basket problem assumes there is no missing data: that is, for a given item and basket it is known whether or not the item is in the basket. This assumption obviates the need for complex algorithms to estimate missing data values.

- Because market basket data is usually voluminous, algorithms that need a large amount of data to develop a causal theory are well suited to this application.

- With thousands of items, there are likely to be hundreds of thousands of causal relationships. While an optimal algorithm might find all these relationships and output only the "interesting" ones (undoubtedly using another data mining algorithm as a subroutine!), it is acceptable to find and output only a small number of causal relationships. Selection may occur either due to pruning or due to an algorithm that finds only causal relationships of a certain form. Algorithms that output a small number (possibly arbitrarily decided) of causal relationships may not be useful outside a data mining context. Data mining, however, is used for exploratory analysis, not hypothesis testing. Obviously, a technique for finding all causal relationships and somehow picking "interesting" ones is superior to one that chooses relationships arbitrarily, but even techniques in the latter category are valuable in a data mining setting.

- Again, for market basket applications with thousands of items, finding a complete causal model is not only expensive, it is also difficult to interpret. We believe isolated causal relationships, involving only pairs or small sets of items, are easier to interpret.

- For many market basket problems, discovering that two items are *not* causally related, or at least not directly causally related (that is, one may cause the other but only through the influence of a third factor), may be as useful as finding out that two items are causally related. While complete causal models illuminate lack of causality as easily as they illuminate causality, algorithms that produce partial models are more useful in the market basket setting if they can discover some noncausal relationships as well as causal relationships.

These aspects of discovering causality for market basket data drove the development of the algorithms we present in Section 4. Many of these issues point to constraint-based methods as being well suited to market basket analysis, while others indicate that tailoring constraint-based methods — for instance by providing error analysis predicated on boolean data and discovering lack of causality — can yield sizable advantages over using generic constraint-based techniques.

2 The LCD Algorithm

The LCD algorithm [11] is a polynomial time, constraint-based algorithm. It uses tests of variable dependence, independence, and conditional independence to restrict the possible causal relationships between variables. The crux of this technique is the Markov condition [26].

Definition 1 (Markov Condition) *Let A be a node in a causal Bayesian network, and let B be any node that is not a descendant of A in the causal network. Then the* Markov condition *holds if A and B are independent, conditioned on the parents of A.*

The intuition of this condition is as follows: If A and B are dependent, then B must either be a (possibly indirect) cause of A or (possibly indirectly) caused by A. In the second case, B is a descendant of A, while in the first B is an ancestor of A and has no effect on A once A's immediate parents are fixed.

Assuming the Markov condition, we can make causal claims based on independence data. For instance, suppose we know, possibly through a priori knowledge, that A has no causes. Then if B

is dependent on A, B must be caused by A, though possibly indirectly. If we have a third variable C dependent on both A and B, then the three variables lie along a causal chain. Variable A, since it has no causes, is at the head of the chain, but we don't know whether B causes C or *vice versa*. If, however, A and C become independent conditioned on B, then we conclude by the Markov condition that B causes C.

In the discussion that follows, we denote by $B \to C$ the claim that B causes C. Note that, contrary to normal use in the Bayesian network literature, we do not use this to mean that B is a *direct* cause of C. Because we have restricted our attention to only three variables, we cannot in fact say with assurance that B is a direct cause of C; there may be a *confounding* variable D, or some *hidden* variable, that mediates between B and C. A confounding variable is a variable that interacts causally with the items tested, but was not discovered because it was not included in the tests performed. A hidden variable represents an effect that interacts causally with the items tested, but is not captured by any variable in the data set.

Even with hidden and confounding variables, we can say with assurance that A is a cause of B and B a cause of C. We can also say with assurance that A is *not* a direct cause of C, since its causality is mediated by B, at the minimum.

If we drop the assumption that A has no causes, then other models besides $A \to B \to C$ become consistent with the data. In particular, it may be that $A \leftarrow B \to C$, or $A \leftarrow B \leftarrow C$. In this case, it is impossible, without other knowledge, to make causality judgments, but we can still say that A is not a direct cause of C, though we do not know if it is an indirect cause, or even caused by C instead of causing C.

We summarize these observations in the *CCC rule*, so named since it holds when A, B, and C are all pairwise correlated.[3]

Rule 1 (*CCC* causality) *Suppose that A, B, and C are three variables that are pairwise dependent, and that A and C become independent when conditioned on B. Then we may infer that one of the following causal relations exists between A, B, and C:*

$$A \leftarrow B \to C \qquad\qquad A \to B \to C \qquad\qquad A \leftarrow B \leftarrow C$$

Now suppose two variables B and C are independent, but each is correlated with A. Then B and C are not on a causal path, but A is on a causal path with both of them, implying either both are ancestors of A or both are descendants of A. If B and C become dependent when conditioned on A, then by the Markov condition they cannot be descendants of A, so we can conclude that B and C are causes of A. This observation gives rise to the *CCU rule*, so named since two variable pairs are correlated and one is uncorrelated.

Rule 2 (*CCU* causality) *Suppose A, B, and C are three variables such that A and B are correlated, A and C are correlated, and B and C are uncorrelated, and that B and C become correlated when conditioned on A. Then we may infer that B and C cause A.*

Again we cannot say whether hidden or confounding variables mediate this causality.

[3] Here correlation indicates dependence and not a specific value of the correlation coefficient. While use of the term "correlation" we intend should be clear from context.

Algorithm: LCD
Input: A set V of variables. w, a variable known to have no causes. A data set D. Tests for dependence, independence, and conditional independence (D, I, and CI respectively).
Output: A list of causal relationships supported by the data.

```
For all variables x ≠ w
    If D(x, y)
        For all variables y ∉ {x, w}
            If D(x, y) and D(y, w) and CI(x, y, w)
                output 'x causes y'
```

Table 1: The LCD algorithm.

The LCD algorithm uses the CCC rule (but not the CCU rule) to determine causal relationships. It looks at triples of items, where one item is known a priori to have no cause. In this way it can disambiguate the possible causal models. The algorithm assumes tests for dependence and conditional independence. It is shown in Table 1.

The LCD algorithm depends on the correctness of the statistical tests given as input. If one test wrongly indicates dependence or conditional independence, the results will be invalid, with both false positives and false negatives. An additional assumption, as has already been stated, is in the applicability of the Markov condition. We list some other assumptions, as described by Cooper [11], and their validity for market basket data.

Database Completeness The value of every variable is known for every database record. This is commonly assumed for market basket applications.

Discrete Variables Every variable has a finite number of possible values. The market basket problem has boolean variables.

Causal Faithfulness If two variables are causally related, they are not independent. This is a reasonable assumption except for extraordinary data sets, where, for instance, positive and negative correlations exactly cancel.

Markov Condition This condition is reasonable if the data can actually be represented by a Bayesian network, which in turn is reasonable if there is no feedback between variables.

No Selection Bias The probability distribution over the data set is equal to the probability distribution over the underlying causal network. The reasonableness of this assumption depends on the specific problem: if we only collect supermarket data for customers who use a special discount card, there is likely to be selection bias. If we collect data for all customers, or random customers, selection bias is unlikely to be a problem.

Valid Statistical Testing If two variables are independent, then the test of independence will say so. If they are dependent, the test of dependence will say so. This assumption is unreasonable, since all tests have a probability of error. When many tests are done, as is the case for the LCD algorithm, this error is an even bigger concern (see Section 3.1).

A criticism of the LCD algorithm is that it finds only causal relationships that are embedded in CCC triples, presumably a small subset of all possible causal relationships. Furthermore, this pruning is not performed on the basis of a goodness function, but rather because of the exigencies of the algorithm: these are the causal relationships that can be discovered quickly. While this trait of the LCD algorithm is limiting in general, it is not as problematic in the context of data mining. As we mentioned in Section 1.2, data mining is used for exploratory analysis, in which case it is not necessary to find all, or even a small number of specified, causal relationships. While not ideal, finding only a small number of causal relationships is acceptable for data mining.

3 Determining Dependence and Independence

Cooper [11] uses tests for dependence and independence as primitives in the LCD algorithm, and also proposes Bayesian statistics for these tests. In our approach, we use instead the much simpler χ^2 statistic; refer to [9] for a discussion on using the chi-squared tests in market basket applications. The necessary fact is that if two boolean variables are independent, the χ^2 value is likely to exceed the threshold value χ^2_α with probability at most α. There are tables holding χ^2_α for various values of α.[4] We say that if the χ^2 value is greater than χ^2_α, then the variables are correlated with probability $1-\alpha$. We extend this definition to the market basket problem by adding the concept of support, which is the proportion of baskets that a set of items occurs in.

Definition 2 (Correlation) *Let $s \in (0,1)$ be a* support threshold *and $c \in (0,1)$ be a* confidence threshold. *An itemset $S \subseteq I$ is said to be (s,c)-*correlated *(hereafter, merely* correlated*) if the following two conditions are met:*

1. *The value of support(S) exceeds s.*
2. *The χ^2 value for the set of items S exceeds the χ^2 value at significance level c.*

Typical values for the two threshold parameters are $s = 1\%$ and $c = 5\%$. If $c = 5\%$, then we would expect that, for 5% of the pairs that are actually uncorrelated, we would claim (incorrectly) they are correlated.

Support is not strictly necessary; we use it both to increase the effectiveness of the chi-squared test and to eliminate rules involving infrequent items.

Intimately tied to the notion of correlation is that of *uncorrelation*, or independence. Typically, uncorrelation is defined as the opposite of correlation: an itemset with adequate support is uncorrelated if the χ^2 value does not support correlation. In effect, the chi-squared test is being applied as a *one-tailed* test.

[4] In the boolean case the appropriate row of the table is the one for 1 degree of freedom.

This definition is clearly problematic: with $c = 5\%$, item sets with a χ^2 value just below the cutoff will be judged uncorrelated, even though we judge there is almost a 95% chance the items are actually correlated. We propose, instead, a two-tailed test, which says there is evidence of dependence if $\chi^2 > \chi^2_\alpha$ and evidence of independence if $\chi^2 < \chi^2_{\alpha'}$. The following definition is based on this revised test.

Definition 3 (Uncorrelation) *Let* $s \in (0, 1)$ *be a* support threshold *and* $c \in (0, 1)$ *be a* confidence threshold. *An itemset* $S \subseteq I$ *is said to be* (s, c)-uncorrelated *(hereafter, merely* uncorrelated*) if the following two conditions are met:*

1. *The value of support*(S) *exceeds* s.
2. *The* χ^2 *value for the set of items* S *does not exceed the* χ^2 *value at significance level* c.

If $c = 95\%$, then we would expect that, for 5% of the pairs that are actually uncorrelated, we would fail to say they are uncorrelated. Note that we would *not* necessarily say they are correlated: a pair of items may be neither correlated nor uncorrelated. Such a pair cannot be part of either CCC causality or CCU causality.

We can use the chi-squared test not only for dependence and independence, but also for conditional dependence and conditional independence. Variables A and B are independent conditioned on C if $p(AB \mid C) = p(A \mid C)p(B \mid C)$. The chi-squared test for conditional independence looks at the statistic $\chi^2(AB \mid C = 0) + \chi^2(AB \mid C = 1)$, where $\chi^2(AB \mid C = i)$ is the chi-squared value for the pair A, B limited to data where $C = i$. As with standard correlation, we use a two-tailed chi-squared test, using different thresholds when testing for conditional dependence as opposed to conditional independence.

For example, suppose everybody over 18 both drives and votes, but nobody under 18 does either. Then driving and voting are dependent — quite powerfully, since driving is an excellent predictor of voting, and vice versa — but they are independent conditioned on age. To see this, note that if we know somebody's age, knowing whether they drive yields no extra insight in predicting whether they vote.

Note that both the correlation and the uncorrelation tests bound the probability of incorrectly labeling uncorrelated data, but do not estimate the probability of incorrectly labeling *correlated* pairs. This is a basic problem in statistical analysis of correlation: while rejecting the null hypothesis of independence requires only one test, namely that the correlation is unlikely to actually be 0, rejecting the null hypothesis of dependence requires an infinite number of tests: that the correlation is not 0.5, that the correlation is not 0.3, and so on. Obviously, if the observed correlation is 0.1, it is likelier that the actual correlation is 0.3 than that it is 0.5, giving two different probabilities. It is unclear what number would capture the concept that the pair is "correlated." One solution to this problem is to define correlation as "the correlation coefficient is higher than a cutoff value." For boolean data, this is equivalent to testing the chi-squared value as we do above; see Section 3.1 and Appendix A for details.

3.1 Coefficient of Correlation

The LCD algorithm can perform the tests for dependence and independence tens of thousands of times on data sets with many items. Though the individual tests may have only a small probability of error, their frequent use means there will be hundreds of errors in the final result. This problem is exacerbated by the fact a single erroneous dependence judgment could form the basis of up to m^2 causal rules.

This problem is usually handled in the statistical community by lowering the tolerance value for each individual test, so that the total error rate is low. In general, with thousands of tests the error rate will have to be set intolerably low. However, for boolean data even a very low tolerance value is acceptable because of a connection between the probability of error and the strength of correlation, presented in the following theorem. This proof of this theorem, along with the concept of correlation coefficient at its heart, can be found in Appendix A.

Theorem 1 *Let X and Y be boolean variables in a data set of size n, with correlation coefficient ρ. Then $\chi^2(X,Y) = n\rho^2$. Thus, X and Y will fail to be judged correlated only if the confidence level for the correlation test is below that for $\chi^2_\alpha = n\rho^2$.*

Because of this relationship, by discarding rules that are more likely to be erroneous, we are at the same time discarding rules with only a weak correlation. Weak rules are less likely to be interesting in a data mining context, so we are at the same time reducing the probability of error and improving the quality of results.

4 Algorithms for Causal Discovery

In the following discussion we shall use the following terminology: A pair of items constitutes a C-edge if they are correlated according to the correlation test, and they constitute a U-edge if they are uncorrelated according to the uncorrelation test. (Note that an item pair may be neither a C-edge nor a U-edge.) We denote the number of items by m, the number of baskets by n, and the degree of node A — that is, the number of C- and U- edges involving item A — by Δ_A. When necessary, we shall also refer to Δ^C_A and Δ^U_A, which are the degree of A when restricted to C- and U-edges, respectively. Let Δ be the maximum degree, that is, $\max_A\{\Delta_A\}$; Δ^C and Δ^U are defined similarly.

We consider the performance of algorithms with respect to three factors: memory use, running time, and number of passes required over the database. Since our techniques look at triples of items, $O(m^3)$ memory is enough to store all the count information needed for our algorithms. Because of this, machines with $O(m^3)$ memory require only one pass over the database in all cases. The algorithms below assume that m is on the order of thousands of items, so caching the required database information in this way is not feasible. However, we consider $O(m^2)$ memory to be available. Situations where less memory is available will have to rely on the naive algorithm, which requires only $O(1)$ memory.

4.1 The Naive Algorithm

Consider first the brute force search algorithm for determining all valid causal relations from market basket data. Effectively, we iterate over all triples of items, checking if the given triple satisfies the conditions for either CCC or CCU causality. This requires a conditional independence test, which requires the count n_{ABC}. Thus, the brute force algorithm requires $O(m^3)$ passes over the database, and this alone takes time $O(nm^3)$. However, the algorithm requires only $O(1)$ memory. If M words of memory are available, we can bundle count requests to reduce the number of database passes to $O(m^3/M)$.

4.2 The CC-path Algorithm

The naive algorithm can be speeded up easily if $O((\Delta^C)^2)$ memory is available: Consider each item A in turn, determine all items connected to A via C-edges, and for each pair B and C of these C-neighbors check if either causality rule applies to ABC. This approach requires that we examine $O(m(\Delta^C)^2)$ triples, instead of $O(m^3)$. More importantly, it requires only n passes over the database, since in the pass for item A, we can use the $O((\Delta^C)^2)$ space to store counts for all ABC in which B and C are connected to A by a C-edge. The running time of the resulting algorithm is $O(nm(\Delta^C)^2)$. (More accurately, the running time is $O(n\sum_A(\Delta^C_A)^2)$. More accurately still, it is $O(n\sum_A((\Delta^C_A)^2 + m))$, since it takes $O(nm)$ time just to calculate which neighbors of A are correlated. In the future, we ignore such fine distinctions.) This algorithm has the same worst-case running time as the naive algorithm, but unless Δ^C_A is very large it is faster than performing the naive algorithm and bundling n^2 count requests.

4.3 The CU-path Algorithm

The CC-path algorithm is so named because it looks at $C-C$ paths (with A as the joint) and then checks for the existence of the third, uncorrelated, edge. Another approach, appropriate only for finding CCU causality, is to look for $C-U$ paths and check if the third edge is correlated. This algorithm is superior when $\Delta^U_A < \Delta^C_A$ for most A. Let $\Delta^{CU} = \max_A\{\Delta^C_A\Delta^U_A\}$ (in general, this will be less than $\Delta^C\Delta^U$). The CU-path algorithm requires $O(\Delta^{CU})$ memory, $O(nm\Delta^{CU})$ time, and $O(m)$ passes over the database. Again, there is the tighter but more complicated time bound of $O(n\sum_A \Delta^C_A\Delta^U_A)$.

4.4 The CU-path Algorithm with Heuristic

The CU-path algorithm allows for a heuristic that is not available for the CC-path algorithm. It follows from the fact every CCU triple has two $C-U$ paths but only one $C-C$ path. Therefore, for every U edge there is a choice, when looking for $C-U$ paths, of whether to look at one endpoint or the other. From a computational point of view it makes sense to pick the endpoint that abuts fewer C-edges. As a result there will be fewer $C-U$ paths to process. (One way to think of it is this: there are $C-U$ paths that are part of CCU triples, and those that are not. The former

Algorithm	Theoretical Space	Theoretical Time	`clari` Time	Theoretical DB passes	`clari` DB passes
naive	$O(1)$	$O(nm^3)$	–	$O(m^3)$	–
CC-path	$O((\Delta^C)^2)$	$O(nm(\Delta^C)^2)$	3684 sec	$O(m)$	12606
CU-path	$O(\Delta^{CU})$	$O(nm\Delta^{CU})$	1203 sec	$O(m)$	9718
CU-path with heuristic	$O(\Delta^{CU})$	$O(nm\Delta^{CU})$	631 sec	$O(m)$	9718

Table 2: Summary of running time and space for finding CCU causal relationships, in both theory and practice on the `clari` data set (Section 5.4). This data set has $m = 6303$ and $n = 27803$. Time is in seconds of user time. To improve running time, the algorithm grouped items together to use the maximum memory available on the machine. Thus, a comparison of memory use is not helpful. The naive algorithm was not run on this data set.

we must always look at, but the latter we can try to avoid.) This heuristic has proven extremely successful, particularly when the number of C-edges is large. For the `clari` data set (Section 5.4), the CU-path heuristic cut the running time in half. The improvement in the smaller `clari.world` data set (Section 5.2 was 30 percent.

Optimizations are possible. For instance, the algorithms described above check twice whether a pair of items share an edge, once for each item in the pair. It would be faster, memory permitting, to determine all correlated and uncorrelated edges once as a pre-processing step and store them in a hash table. Even better would be to store edges in an adjacency list as well as in the hash table, to serve as a ready-made list of all C- and U-edges abutting the "joint" item. In experiments, this improvement in data structures halved the running time. Caching as many triple counts as will fit in main memory will also improve the running time by a constant factor.

4.5 Comparison of Performance

Table 2 holds a summary of the algorithms we consider and their efficiencies. Note that the number of database passes is not the same for all algorithms. This is because if an item lacks a correlated (or uncorrelated) neighbor, we need not perform a database pass for that item. For the `clari` data set (Section 5.4), there are 316295 C-edges but only 5417 U-edges. This explains the superior performance of the CU-path algorithm, both in terms of time and database passes.

When the data is very large, we expect the I/O cost — the cost of moving data between main and secondary memory — to dominate, rather than processing cost. However, there is really no difference, for these algorithms, between processor and I/O costs. In justification, for a fixed amount of main memory, the number of passes over the data that we need is proportional to the number of DB passes times the space required in each pass. The reason is that we may process in parallel the passes for just as many items as can have their needed space in main memory simultaneously. If we look at Table 2, we see that the time required for each algorithm is n times the product of the space and DB passes. Thus, "time" is proportional to I/O cost for the algorithms we consider,

A	B	C	causality	ρ_{AB}	ρ_{AC}	ρ_{BC}
MOVED-LAST-5YRS	MALE	SUPPORT	$A \leftarrow B \rightarrow C$	0.0261	−0.0060	−0.2390
NEVER-MARRIED	EMPLOYED	CAR/CAB	?	−0.0497	−0.0138	0.2672
HOUSEHOLDER	20-40K	NATIVE-AMER	$A \leftarrow B \leftarrow C$	0.2205	−0.0111	−0.0537
MILITARY	PAY	GOVT	?	0.1350	−0.0795	−0.5892

Table 3: Some of the 25 causal CCC relationships found in census data. The causal relationship is given when it can be disambiguated using a priori information. ρ is the coefficient of correlation between the pair of items, and is positive when the two items are found often together, and negative when they are rarely found together.

A	and	B	each cause	C	ρ_{AC}	ρ_{BC}
BLACK	and	NOGRAD-HS	each cause	CAR/CAB	−0.0207	−0.1563
ASIAN	and	LABORER	each cause	<20K	0.0294	−0.0259
ASIAN	and	LABORER	each cause	20-40K	−0.0188	0.0641
EMPLOYED	and	MILITARY	each cause	UNDER-43	−0.0393	−0.2104
EMPLOYED	and	MILITARY	each cause	NEVER-MARRIED	−0.0497	−0.0711
SALES	and	HOUSEHOLDER	each cause	NOGRAD-HS	−0.0470	−0.0334

Table 4: Some of the 36 causal CCU relationships found in census data. The causal relationship is uniquely determined. ρ is the coefficient of correlation between the pair of items. It is not given for the AB pair because this pair is uncorrelated.

and henceforth we shall only consider processor time in our comparisons.

5 Experimental Results

We use two data sets for our analysis, similar to the ones used by Brin, Motwani, and Silverstein [9]. One holds boolean census data (Section 5.1). The other data set is a collection of text data from UPI and Reuters newswires (Section 5.2). We actually study two newsgroup corpora, one of which is significantly larger than the other.

In the experiments below, we used a chi-squared cutoff $c = 5\%$ for C-edges and $c = 95\%$ for U-edges. We use the definition of support given by Brin, Motwani, and Silverstein [9]. All experiments were performed on a Pentium Pro with a 166 MHz processor running Solaris x86 2.5.1, with 96 Meg. of main memory. All algorithms were written in C and compiled using `gcc 2.7.2.2` with the `-O6` compilation option.

5.1 Census Data

The census data set consists of $n = 126229$ baskets and $m = 63$ binary items; it is a 5% random sample of the data collected in Washington state in the 1990 census. The items are listed in Appendix B. To simplify the pinterpretation of many items, we discarded responses of those under

25 and over 60 years old.[5] Census data has categorical data that we divided into a number of boolean variables. Thus, for marital status we have several items: MARRIED, DIVORCED, SEPARATED, WIDOWED, NEVER-MARRIED. Every individual has TRUE for one of these variables and FALSE for the rest.

The test for CCU causality took 3 seconds of user CPU time to complete, while the test for CCC causality took 35 seconds of user CPU time. This indicates the census data has many more C edges than U edges, which is not surprising since all variables derived from the same census question are of necessity correlated.

In Table 3 we show some of the results of finding CCC causality. Since several variables (such as MALE and UNDER-43) cannot have causes, census data fits well into Cooper's LCD framework, and it is often possible to determine the direction of causation. Because of the possibility of confounding and hidden variables however, we cannot determine direct causality. The CCC test, however, allows us to rule out direct causality, and this in itself yields interesting results.

Intuitively, an apparent relationship between two variables is explained by a third variable. For example, being in a support position is correlated with having moved in the past five years. This may lead one to believe that support personnel are unusually unlikely to move around. However, when we condition on being white, the apparent relationship goes away. From this, we can guess that being white causes one to move frequently, and also causes one not to be in support. Notice that in any case the correlation between support and moving is very weak, indicating this rule is not powerful.

Another CCC rule shows that if people who are never married are less likely to drive to work, it is only because they are less likely to be employed. (The conditional is used here because we cannot be sure of the causal relationship: are the unmarried less likely to have jobs, or are the unemployed less likely to get married?)

Table 4 shows some of the CCU causal relationships discovered on census data. While the causal relationship is uniquely determined, confounding and hidden variables keep us from determining if causality is direct. For instance, in the first row of Table 4, we can say not having graduated high school causes one not to drive to work, but we do not know if this is mediated by the fact high school dropouts are less likely to have jobs. A causal rule such as NOGRAD-HS $\rightarrow$ EMPLOYED may exist, but if so neither the CCC nor CCU causality tests found it. As we see, these algorithms are better at exploratory analysis than hypothesis testing.

Note that both CCC and CCU causality tests discovered a causal relationship between being employed and never having been married. The CCU result can be used to disambiguate among the possible causal relationships found from the CCC test. A danger of this if that if the CCU result is inaccurate, due to statistical error, using it to disambiguate the CCC result propagates the error.

As it is, an improper uncorrelation judgment can cause many erroneous causal inferences. For instance, the uncorrelated edge SALES–HOUSEHOLDER is the base of 10 CCU judgments, causing

[5] For instance, one choice for marital status is, "Never married or under 15 years old," while the question regarding means of transportation to work conflates being unemployed and being under 16.

20 causal inferences. If this edge was marked uncorrelated incorrectly, all 20 causal inferences are unjustified. In fact, our a priori knowledge would lead us to believe there is a correlation between being in sales and being the head of a household. Causal inferences based on this U-edge, such as the last entry in Table 4 are clearly false — dropping out of high school is temporally prior to getting a job or a house, and thus cannot be caused by them — leading us to question all causal inferences involving the SALES–HOUSEHOLDER edge.

5.2 Text Data

We analyzed 3056 news articles from the `clari.world` news hierarchy, gathered on 13 September 1996, comprising 18 megabytes of text. For the text experiments, we considered each article to be a basket, and each word[6] to be an item. These transformations result in a data set that looks remarkably different from the census data: there are many more items than baskets, and each basket is sparse. To keep the number of items to a reasonable level, we considered only words that occurred in at least 10 articles. We also removed commonly occurring "stop words" such as "the," "you," and "much." We were left with 6723 distinct words.

Since we have no a priori knowledge to distinguish between the possible causal models returned by the CCC algorithm, we ran only the CCU algorithm on the text data. The algorithm returned 73074 causal relationships. To study these, we sorted them by (the absolute value of) their correlation coefficient. We would expect the very top pairs to be obvious causal relationships, and indeed we see from Figure 5 that this is the case. To explore more interesting causal relationships, we also show some results from 5% down the list of correlations.

Even the first set of causal relationships, along with its obvious relationships such as "united" causing "states," has some surprises. One is in the relationships "quoted" causes "saying," probably part of the set phrase, "...was quoted as saying ...." Though this may not illuminate the content of the corpus, it does lend insight into the writing style of the news agency.

Another interesting property is the frequency of causal relationships along with their converse. For instance, "prime" causes "minister" and "minister" causes "prime." The probable reason is that these words are usually found in a phrase and there is therefore a deterministic relationship between the words; that is, one is unlikely to occur in an article without the other. When words are strongly correlated but not part of a phrase — "iraqi" and "iraq" are an example — then we only see the causal relationship in one direction. This observation suggests a somewhat surprising use of causality for phrase detection. If words that always occur together do so only as part of a phrase, then we can detect phrases even without using word location information, just by looking for two-way causality. Presumably, incorporating this strategy along with conventional methods of phrase detection would only improve the quality of phrase identification.

The causal relationships at the 5% level are as intriguing as those at the top of the list. The causal relationship "infiltration" $\rightarrow$ "iraqi" points to an issue that may bear further study. Other

[6] A word was defined to be a series of alphanumeric characters. However, we allowed a word to have a single internal punctuation mark, which we removed. We also lower-cased all words. Thus the sentence "It'll cost you one—no, two—pence for that pince-nez" has the words `itll`, `cost`, `you`, `one`, `no`, `two`, `pence`, `for`, `that`, and `pincenez`.

Causal relationships	χ^2 value	ρ	Causal relationships	χ^2 value	ρ
upi → reuter	2467.2895	−0.8985	state → officials	70.5726	0.1520
reuter → upi	2467.2895	−0.8985	saturday → state	70.6340	0.1520
iraqi → iraq	2362.6179	0.8793	infiltration → iraqi	70.5719	0.1520
united → states	1691.0389	0.7439	forces → company	70.5456	−0.1519
states → united	1691.0389	0.7439	company → forces	70.5456	−0.1519
prime → minister	1288.8601	0.6494	win → party	70.3964	0.1518
minister → prime	1288.8601	0.6494	commitment → peace	70.2756	0.1516
quoted → saying	866.6014	0.5325	british → perry	70.2082	0.1516
news → agency	718.1454	0.4848	support → states	70.1291	0.1515
agency → news	718.1454	0.4848	states → support	70.1291	0.1515

Table 5: Causal relationships from the top and 5% mark of the list of causal relationships for words in the `clari.world` news hierarchy. The list is sorted by (the absolute value of the) correlation coefficient (last column). The χ^2 value measures the confidence that there is a causal relationship; all these χ^2 values indicate a probability of error of less than 0.0001. The ρ value measures the power of the causality.

relationships, such as "saturday" → "state," seem merely bizarre. Note also how quickly the correlation coefficients drop for this data set; only 2.3% of the causal relationships have $\rho > 0.2$.

A final note about the causal relationships is they seem to disproportionately concern Iraq. While Iraq was much in the news during September 1996, the news hierarchy includes articles concerning all parts of the world, from Latin America to Africa to Oceania. However, there were a diffuse set of issues relating to these areas, each with its own vocabulary. Only Iraq seemed to have the large number of cohesive articles necessary to pass the support and confidence thresholds. Whether this is a feature of the causal algorithm[7] or a bug depends on the individual application.

5.3 Comparing Causality with Correlation

A question naturally arises: what is the advantage of causal discovery over merely ranking correlated item pairs? In Figure 6 we show the top 10 correlations, as measured by the correlation coefficient. These results are directly comparable to the top portion of Figure 5. Two difference are immediately noticeable. One is the new item pairs. Some of them, like "iraq" and "warplanes," seem like significant additions. Others, like "hussein" and "northern," have plausible explanations (the U.S. flight zone is restricted to the Northern and Southern bands of the Iraq) but do not seem to belong so high on the list; other, more perspicuous, causal relationships have a lower correlation coefficient. The other noticeable difference is that, since correlation is symmetric, there is no case of a pair and its converse both occurring. Insofar as asymmetric causalities yield extra understanding of the data set, identifying causal relationships yields an advantage over identifying correlations.

A third difference, not noticeable in the figure, is that there are many more correlation rules

[7]Actually, since support and confidence are concepts used to identify C- and U-edges, this is a concern not only for causal discovery but for any methodology based on a similar definition of correlation.

Corr. relationships	χ^2 value	ρ
reuter – upi	2467.2895	−0.8985
iraq – iraqi	2362.6179	0.8793
states – united	1691.0389	0.7439
minister – prime	1288.8601	0.6494
quoted – saying	866.6014	0.5325
democratic – party	777.7790	0.5045
agency – news	718.1454	0.4848
iraqs – northern	705.4615	0.4805
hussein – northern	678.5580	0.4712
iraq – warplanes	655.5450	0.4632

Table 6: Correlations from the top of the list of correlations, sorted by (the absolute value of the) correlation coefficient (last column). This list is a superset of the list of top causal relationships (Figure 5). The χ^2 value measures the confidence that there is a correlation; all these χ^2 values indicate a probability of error of less than 0.0001. The ρ value measures the power of the causality.

than causal rules. While there are around 70 thousand causal relationships in this data set, there are 200 thousand correlated pairs.

5.4 Performance on a Large Text Data Set

While the `clari.world` hierarchy is somewhat homogeneous, which as we have seen is an advantage when pruning, at 18 megabytes it is a small dataset. We therefore repeated the text experiments on the entire `clari` hierarchy, a larger, more heterogeneous news hierarchy that covers sports, business, and technology along with regional, national, and international news. This data set was gathered on 5 September 1997. While `clari.world` is logically a subtree of the `clari` hierarchy, the `clari.world` database is not a subset of the `clari` database since the articles were collected on different days. The `clari` data set consists of 27803 articles and 186 megabytes of text, and is thus ten times larger than the `clari.world` data set. However, the number of items was kept about the same — the larger data set, at 6303 items, actually has fewer items than the `clari.world` data set — due to the pruning of infrequent words. In both cases, words were pruned that occurred in fewer than 0.3% of all documents; for the `clari` data set this worked out to an 84 document minimum.

As in the smaller data set, the terms with the highest correlation come from a coherent subset of documents from the collection. Unfortunately, the coherent subset of the `clari` collection is a large mass of Government postings soliciting bids for automotive supplies, and most of the words used in these posts are abbreviations of technical terms. Thus, the top causalities are "recd" causes "solnbr" and "desc" causes "solnbr"; we must go down 25 lines before we find causality involving two English words. Even then, the relationship — "office" causes "contact" — involves words from the procurement articles.

The causal relationships found 5% down the list are a little more interesting. Some are shown

Causal relationships	χ^2 value	ρ
cause $\rightarrow$ company	558.2142	0.1417
15 $\rightarrow$ 90	557.9750	0.1417
constitutes $\rightarrow$ number	557.8370	0.1416
modification $\rightarrow$ poc	557.2716	0.1416
email $\rightarrow$ 1997	557.1664	0.1416
today $\rightarrow$ time	557.0560	0.1415
time $\rightarrow$ today	557.0560	0.1415
rise $\rightarrow$ market	556.6937	0.1415
people $\rightarrow$ update	556.6686	0.1415
134 $\rightarrow$ 28	556.1250	0.1414

Table 7: Causal relationships from the list of causal relationships for words in the `clari` news hierarchy, starting from 5% down the list when the list is sorted by (the absolute value of the) correlation coefficient (last column). The χ^2 value measures the confidence that there is a causal relationship; all these χ^2 values indicate a probability of error of less than 0.0001. The ρ value measures the power of the causality.

in Figure 7. Note that, as in the smaller text data set, the correlation coefficient rapidly becomes rather low.

6 Conclusion and Further Research

In data mining context, constraint-based approaches promise to find causal relationships with the efficiency needed for the large data sets involved. The size of the data mining data sets mitigate some of the weaknesses of constraint-based approaches, namely that they sometimes need large amounts of data in order to make causal judgments, and instead of finding all causal relationships, they only find a subset of these relationships. For data mining, which seeks to explore data rather than to test a hypothesis, finding only a portion of the causal relationships is acceptable. Furthermore, with large data sets, the number of actual causal relationships is likely to be so large that algorithms that return only a small portion of the relationships are likely to yield interesting results. Another weakness of constraint-based algorithms, the error inherent in repeated use of statistical tests, is mitigated in boolean data by using a power statistic to reduce the probability of error without discarding powerful causal relationships.

We developed a series of algorithms, based on techniques used in Cooper's LCD algorithm, that run in time linear in the size of the database and cubic in the number of variables. For large data sets with thousands of variables, these algorithms proved feasible and returned a large number of causal relationships and, equally interesting, not-directly-causal relationships. This feasibility came from heuristics and algorithmic choices that improved on both the time and memory requirements of the naive cubic-time algorithm.

Finding causal relationships is useful for a variety of reasons. One is that it can help in visualizing relationships among variables. Another is that, unlike correlation, causation is an asymmetric concept. In contexts where it is possible to intervene on the variables (for instance, when a manager can choose not to stock certain food items) causality can help predict the effect of the intervention, whereas a correlation analysis cannot. In the context of text analysis, causation can help identify phrases.

There are still a variety of unresolved and unexplored issues in the area of mining for causal relationships. We briefly list some of them below.

Choosing Thresholds Is there a way to determine optimal values for the correlation and uncorrelation cutoffs for a given data set? Better yet, is it possible to replace cutoff values with estimates of the probability of correlation efficiently?

Disambiguation We mentioned in Section 5.1 that, at the risk of propagating error, we can use known causal rules to disambiguate the CCC causality rule. However, we have seen that both $A \rightarrow B$ and $B \rightarrow A$ may occur in data. Is there a principled way to resolve bidirectional causality for disambiguation? Can we then devise efficient algorithms for using incremental causal information to perform disambiguation?

Hidden Variables Bidirectional causality may indicate deterministic relationships (as in text phrases), error in statistical tests, or the presence of hidden variables. Under what situations can we be confident hidden variables are the cause? What other techniques can we use to discover hidden variables?

Heuristics for Efficiency How can we make the above algorithm even more efficient? The largest speedups could be obtained by avoiding the need to check all triples. Can we determine when the conditional independence test will fail on triples without explicitly testing them? Can we reduce the number of items, perhaps by collapsing items with similar distributions?

Acknowledgments

We would like to thank the members of the Stanford Data Mining research group, particularly Lise Getoor, for their useful comments and suggestions. We would also like to thank Greg Cooper and David Heckerman for fruitful discussions.

References

[1] S. Abiteboul, R. Motwani, and S. Nestorov. Inferring structure in semistructured data. *PODS/SIGMOD Workshop on Management of Semistructured Data*, pp. 42–47, 1997.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on the Management of Data*, pages 207–216, May 1993.

[3] R. Agrawal, T. Imielinski, and A. Swami. Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5(1993): 914–925.

[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In Fayyad et al [13], pages 307–328, 1996.

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, September 1994.

[6] A. Agresti. *Categorical Data Analysis.* John Wiley & Sons, New York, 1990.

[7] A. Agresti. A Survey of exact inference for contingency tables. *Statistical Science* 7(1992): 131–177.

[8] A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, Seattle, WA, pages 46–54.

[9] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD Conference on Management of Data,* pages 265–276, Tucson, AZ, May 1997.

[10] S. Brin, R. Motwani, S. Tsur, and J.D. Ullman. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD Conference on Management of Data,* pages 255–264, Tucson, AZ, May 1997.

[11] G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery,* 2(1997).

[12] G. Cooper and E. Herskovits. A Bayesian method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9, pages 309–347.

[13] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthrusamy. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, CA, 1996.

[14] D. Heckerman. A Bayesian approach to learning causal networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 285–295, 1995.

[15] D. Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery,* 1(1997): 79–119.

[16] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 293–301, July 1994.

[17] D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. Technical Report MSR-TR-97-05, Microsoft Research, February 1997.

[18] D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3(1995), pages 405–430.

[19] D. Heckerman and R. Shachter. A definition and graphical representation of causality. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1995, pages 262–273.

[20] W. Mendenhall, R. Scheaffer, and D. Wackerly. *Mathematical Statistics with Applications.* Duxbury Press, 3rd edition, 1986.

[21] J. Pearl. From Bayesian networks to causal networks. In *Proceedings of the Adaptive Computing and Information Processing Seminar*, pages 25–27, January 1993.

[22] J. Pearl. Causal diagrams for empirical research. *Biometrika,* 82(1995): 669-709.

[23] J. Pearl. Graphical models for probabilistic and causal reasoning. *The Computer Science and Engineering Handbook*, 1997, pages 697–714,

[24] J. Pearl and T.S. Verma. A theory of inferred causation. In *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning,* 1991, pages 441–452.

[25] P. Spirtes, C. Glymour, and R. Scheiner. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review,* 9(1991): 62–72.

[26] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

[27] P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 499–506, 1995.

[28] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of the 21st International Conference on Very Large Data Bases*, pages 407–419, September 1995.

[29] H. Toivonen. Sampling large databases for finding association rules. In *Proceedings of the 22nd International Conference on Very Large Data Bases*, pages 134–145, September 1996.

[30] S. Wright. Correlation and causation. *Journal of Agricultural Research,* 20(1921): 557–585.

A Relationship between χ^2 and ρ

We provide the manipulations to show that $\chi^2(X,Y) = n \cdot \rho(X,Y)^2$. The χ^2 statistic, as we have mentioned, measures the probability that two variables would yield an observed count distribution if they were independent. The correlation coefficient $\rho(X,Y)$, on the other hand, measures the *strength* of the dependence observed in the data. It does this by summing over the joint deviation from the mean. Formally, the two concepts are defined as follows:

$$\begin{aligned}
\chi^2(X,Y) &= \sum_{x\in\{0,1\},y\in\{0,1\}} \frac{(O(X=x,Y=y) - O(X=x)O(Y=y)/n)^2}{O(X=x)O(Y=y)/n} \\
\rho(X,Y)^2 &= \frac{\left(\sum_i (X_i-\mu_X)(Y_i-\mu_Y)\right)^2}{\sigma_X^2\sigma_Y^2}
\end{aligned}$$

In this case, $O(X = x, Y = y)$ is the observed count of records where $X = x$ and $Y = y$. $O(X = x)O(Y = y)/n$ is the expected value of $O(X = x, Y = y)$ under the hypothesis that X and

Y are independent. μ_X is the observed mean of X, while σ_X^2 is the observed variance. To simplify the notation, we will use $x = O(X = 1)$, $y = O(Y = 1)$, and $b = O(X = 1, Y = 1)$. Let n be the total number of records (i.e., baskets) in the database. Then $O(X = 0)$ is just $n - x$, while $O(X = 1, Y = 0) = O(X = 1) - O(X = 1, Y = 1) = x - b$.

The χ^2 and ρ^2 statistics can easily be simplified in the boolean case. For instance, $\mu_X = O(X = 1)/n = x/n$, and $\mu_Y = y/n$. The variance can be determined from first principles.

$$\begin{aligned}
\sigma_X^2 &= \sum_i (X_i - \mu_X)^2 \\
&= x \cdot (1 - \mu_X)^2 + (n - x) \cdot (0 - \mu_X)^2 \\
&= x(1 - 2x/n + x^2/n^2) + (n - x)(x^2/n^2) \\
&= x - x^2/n \\
&= x/n \cdot (n - x).
\end{aligned}$$

$\sigma_Y^2 = y/n \cdot (n - y)$. Thus the denominator of ρ^2 is $xy(n - x)(n - y)/n^2$. The numerator of ρ is $(\sum X_i Y_i - \mu_X \sum Y_i - \mu_Y \sum X_i + n\mu_X\mu_Y)^2 = (b - xy/n - yx/n + yx/n)^2 = (b - xy/n)^2$.

If we perform the substitutions above for χ^2 and ρ^2, we obtain the following formulas:

$$\begin{aligned}
\chi^2(X,Y) &= \frac{(b - xy/n)^2}{xy/n} + \frac{(x - b - x(n - y)/n)^2}{x(n - y)/n} + \frac{(y - b - y(n - x)/n)^2}{(n - x)y/n} \\
&\quad + \frac{(n - x - y + b - (n - x)(n - y)/n)^2}{(n - x)(n - y)/n} \\
\rho(X,Y)^2 &= \frac{(bn - xy)^2}{xy(n - x)(n - y)}
\end{aligned}$$

χ^2 can be simplified further; the key is to note that all the numerators are actually equal to $(b - xy/n)^2$. For instance

$$(x - b - x(n - y)/n)^2 = (x - b - xn/n + xy/n)^2 = (-b + xy/n)^2.$$

Simplifying in this way, and putting all four terms over a common denominator, yields

$$\chi^2(X,Y) = \frac{n \cdot (b - xy/n)^2 \cdot [(n - x)(n - y) + (n - x)y + (n - y)x + xy]}{xy(n - x)(n - y)}$$

Further simplifying the terms in brackets, we obtain

$$\chi^2(X,Y) = \frac{n \cdot (b - xy/n)^2 \cdot n^2}{xy(n - x)(n - y)}$$

or $\chi^2 = n \cdot (bn - xy)^2/xy(n - x)(n - y)$. It is easy to see that this is $n \cdot \rho^2$.

B Table of Census Variables

The list of boolean census variables used in Section 5.1, and the census categories from which they were derived, is shown below. We formed the set of items by collapsing a number of census questions into binary form.

census question	boolean variables (each is a single item)
Head of house	HOUSEHOLDER
Gender	MALE
Race	WHITE BLACK ASIAN NATIVE-AMER
Age	UNDER-43
Marital status	MARRIED WIDOWED DIVORCED SEPARATED NEVER-MARRIED
Hispanic	HISPANIC
Where born	BORN-US
Citizenship	US-CITIZEN
Education level	NOGRAD-HS NOGRAD-COLLEGE GRAD-COLLEGE POSTGRAD
Mobility	MOVED-LAST-5YRS
Language	ENGLISH-ONLY
English proficiency	ENGLISH-GOOD
Military service	MILITARY
Children	NO-KIDS 1-KID 2-3-KIDS 3+-KIDS
Employment	EMPLOYED
Get to work	CAR/CAB BIKE/WALK PUBLIC-TRANS AT-HOME
Leave for work	BEFORE-8AM
Commute time	>1HOUR
Profession	MANAGER ENGINEER/SCIENTIST HEALTH-MD HEALTH-TREATMENT
	TEACHER LAWYER/JUDGE ENTERTAINER TECHNICIAN MILITARY
	SALES SUPPORT SERVICE AGRICULTURE PRODUCTION LABORER
Employment	PAY GOVT SELF
Income	<20K 20-40K 40-60K 60-80K 80-100K >100K
Dividends	NO-DIV-$
Public Assistance	PUB-ASSIST