Selecting the Right Correlation Measure for Binary Data

LIAN DUAN, New Jersey Institute of Technology
W. NICK STREET, University of Iowa
YANCHI LIU, SONGHUA XU, and BROOK WU, New Jersey Institute of Technology

Finding the most interesting correlations among items is essential for problems in many commercial, medical, and scientific domains. Although there are numerous measures available for evaluating correlations, different correlation measures provide drastically different results. Piatetsky-Shapiro provided three mandatory properties for any reasonable correlation measure, and Tan et al. proposed several properties to categorize correlation measures; however, it is still hard for users to choose the desirable correlation measures according to their needs. In order to solve this problem, we explore the effectiveness problem in three ways. First, we propose two desirable properties and two optional properties for correlation measure selection and study the property satisfaction for different correlation measures. Second, we study different techniques to adjust correlation measures and propose two new correlation measures: the Simplified χ^2 with Continuity Correction and the Simplified χ^2 with Support. Third, we analyze the upper and lower bounds of different measures and categorize them by the bound differences. Combining these three directions, we provide guidelines for users to choose the proper measure according to their needs.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Measurement

Additional Key Words and Phrases: Knowledge discovery, correlation, association rules

ACM Reference Format:
Lian Duan, W. Nick Street, Yanchi Liu, Songhua Xu, and Brook Wu. 2014. Selecting the right correlation measure for binary data. ACM Trans. Knowl. Discov. Data 9, 2, Article 13 (September 2014), 28 pages.
DOI: http://dx.doi.org/10.1145/2637484

1. INTRODUCTION

With the development of scanning devices, the Internet, and computer storage technologies, companies in diverse sectors such as retailing, banking, and telecom compile large databases on consumers' past transactions. Each record in a typical market basket transaction dataset corresponds to a transaction, which contains a unique identifier and a set of items bought by a given customer. The analysis of relationships between items is fundamental in many data mining problems. For example, the central task of association analysis [Agrawal et al. 1993] is to discover a set of items that co-occur frequently in a transaction database. Regardless of how the relationships are defined, such analysis requires a proper measure to evaluate the dependencies among items. The stronger the dependence relationship is, the more interesting the pattern.

In this article, we only study the correlation for binary data. Numeric data can be handled by canonical correlation analysis [Johnson and Wichern 2001], which won't be

Authors' addresses: L. Duan, Y. Liu, S. Xu, and B. Wu, Department of Information Systems, New Jersey Institute of Technology; emails: {lian.duan, yl473, songhua.xu, wu}@njit.edu; W. Nick Street, Department of Management Sciences, University of Iowa; email: lian.duan@njit.edu.

DOI: http://dx.doi.org/10.1145/2637484

discussed here. For binary data, although we are, in general, interested in correlated sets of arbitrary size, most of the published work with regard to correlation is related to finding correlated pairs [Tan et al. 2004; Geng and Hamilton 2006]. Related work with association rules [Brin et al. 1997a, 1997b; Omiecinski 2003] is a special case of correlation pairs since each rule has a left- and right-hand side. Given an association rule $X \Rightarrow Y$ where X and Y are itemsets, $Support = P(X \cap Y)$ and $Confidence = P(X \cap Y)/P(X)$ [Agrawal et al. 1993; Omiecinski 2003] are often used to represent its significance. However, these can produce misleading results because of the lack of comparison to the expected probability under the assumption of independence. In order to overcome the shortcoming, Lift [Brin et al. 1997a], Conviction [Brin et al. 1997b], and Leverage [Piatetsky-Shapiro 1991] are proposed. Dunning [1993] introduced a more statistically reliable measure, Likelihood Ratio, which outperforms other correlation measures. Jermaine [2005] extended Dunning's work and examined the computational issue of Probability Ratio and Likelihood Ratio. Bate et al. [1998] proposed a correlation measure called Bayesian Confidence Propagation Neural Network (BCPNN), which is good at searching for correlated patterns occurring rarely in the whole dataset. These correlation measures are intuitive; however, different correlation measures provide drastically different results. Although Tan et al. [2004] proposed several properties to categorize these correlation measures, there are no guidelines for users to choose the desirable correlation measures according to their needs. In order to solve this problem, we will propose several desirable properties for correlation measures and study the property satisfaction for different correlation measures in this article.

By studying the literature related to correlation, we notice that different correlation measures are favored in different domains. In the text mining area, people use Likelihood Ratio [Dunning 1993]. BCPNN is favored in the medical domain [Bate et al. 1998], while Leverage is used in the social network context [Clauset et al. 2004]. Our research will answer this question of why different areas favor different measures.

Evaluating the performance of different correlation measures requires a ground true ranking list matching human intuition and each measure will be evaluated by checking how similar the retrieved ranking list is to the ground true ranking list. However, when dealing with human intuition to get the ground true ranking list, different people have different opinions. Take the pairs $\{A, B\}$ and $\{C, D\}$ for example. When Event A happens, the probability of observing Event B will increase from 0.01% to 10%. When Event C happens, the probability of observing Event D will increase from 50% to 90%. Which correlation pattern is stronger between $\{A, B\}$ and $\{C, D\}$? Different people have different answers. Therefore, there is no ground true ranking list to test the performance of each measure. Instead, people all agree that a good correlation measure can at least tell correlated patterns from uncorrelated patterns, and it is much easier to identify ground true correlated patterns, especially in simulated datasets. Therefore, our evaluation emphasizes more the precision of each measure by telling correlated patterns from uncorrelated patterns. Still, when using precision to measure performance, two correlation measures can both perfectly tell correlated patterns from uncorrelated patterns to achieve 100% precision but rank correlated patterns differently. For example, one measure can rank $\{A, B\}$ higher, while the other can rank $\{C, D\}$ higher. As a complement to precision, the preference differences among measures will be studied when they achieve similar precision.

There are two very influential papers [Tan et al. 2004; Geng and Hamilton 2006] on the correlation study in the data mining area. They both recognize the performance differences among different measures. They categorized measures according to their different property satisfaction. By categorizing measures, users only need to check the performance of the typical measure in each category instead of all the possible measures. However, two measures can rank patterns differently even if they satisfy

the same set of properties. Instead, one recent paper [Tew et al. 2013] categorized measures directly according to their ranking list similarity. No matter how measures are categorized, two fundamental questions are still not answered. Which one can tell correlated patterns from uncorrelated patterns better? If there is a ranking difference between two measures, what is the difference? Some other studies focus on selecting the best measure that can tell correlated patterns from uncorrelated patterns in a specific application. But they cannot explain why this measure can do better than others. In addition, different measures are selected as the best in different domains [Dunning 1993; Bate et al. 1998]. Furthermore, different measures are selected as the best in the same application but different datasets [Liu et al. 2013]. Therefore, the goal of our research is trying to address these two fundamental questions better. In addition, different from most of the previous research, we study the performance not only on correlated pair search but also on correlated itemset search. Therefore, to the best of our knowledge, we maintain a comprehensive list of correlation measures that have an explicit comparison to the expectation from the assumption of independence and test them against our proposed guidelines. In addition, we only focus on the selection of correlation measures in this article. However, another related issue on correlation search is the efficiency problem. Even if we successfully select the right correlation measure, we might not be able to get the result for a large dataset if we cannot solve the efficiency problem. Xiong et al. [2006a, 2006b] made some significant progress on correlated pair search for ϕ Correlation Coefficient. Duan et al. [2013] proposed a Token-ring algorithm on correlated pair search for more general correlation measures. Zhang and Feigenbaum [2006] studied the distribution of the ϕ-coefficient and relaxed the upper bound in TAPER in order to speed up the search. Duan and Street [2009] proposed a fully correlated itemset framework to speed up the high-dimensional correlation search for any good correlation measure that satisfies the three mandatory properties in Section 2.1.

The remainder of this article is organized as follows. In Section 2, important correlation properties and different correlation measures will be discussed. The correlation properties are used to guide the choice of different correlation measures. Section 3 shows the experimental results. Finally, we draw a conclusion in Section 4.

2. CORRELATION MEASURES

Since some correlation measures can only be used for pairs, we categorize correlation measures into the general and pair-only types. Both types can evaluate correlation for pairs, but only the general type can evaluate correlation for itemsets. In this section, we discuss correlation properties and measures for both types.

2.1. Correlation Properties

There are some general correlation properties that both types need to satisfy. However, there are some additional properties for the pair-only measures.

2.1.1. General Correlation Properties. Given an itemset $S = \{I_1, I_2, \ldots, I_m\}$ with m items in a dataset with sample size n, the true probability is $tp = P(S)$, the expected probability under the assumption of independence is $ep = \prod_{i=1}^{m} P(I_i)$, and the occurrence is $k = P(S) \cdot n$. These parameters will be used in different correlation measures, and we use "Support" and "true probability" interchangeably in this article. Correlation measures can be generalized as $M = f(tp, ep)$. For example, Leverage [Piatetsky-Shapiro 1991] is $tp - ep$. The following seven properties provide guidelines for a good correlation measure M according to users' needs:

P1: M is equal to a certain constant number C when all the items in the itemset are statistically independent.

P2: M monotonically increases with the increase of $P(S)$ when all the $P(I_i)$ remain the same.
P3: M monotonically decreases with the increase of any $P(I_i)$ when the remaining $P(I_k)$ (where $k \neq i$) and $P(S)$ remain unchanged.
P4: The upper bound of M does not approach infinity when $P(S)$ is closer to 0.
P5: M gets closer to C (including negative correlation cases whose M is smaller than C) when an independent item is added to S.
P6: The lower bound of M gets closer to the lowest possible function value when $P(S)$ is closer to 0.
P7: M gets further away from C (including negative correlation cases) with increased sample size when all the $P(I_i)$ and $P(S)$ remain unchanged.

The first three properties are proposed by Piatetsky-Shapiro [1991]. They regulate the change between tp and ep and should be mandatory for any good correlation measure. The first property requires a constant C to indicate independency when the actual probability is the same as the expected probability. It is positive correlation when above C and negative correlation when below C. The second property requires the correlation value to increase when the expected probability stays the same while the actual probability goes up. In other words, the itemset deserves more credit when we have the same expectation but the actual performance is better. Similarly, the correlation value will decrease when the expected probability goes up while the actual probability stays the same.

These three mandatory properties screen out some bad correlation measures, but there are still some other poor correlation measures that satisfy all three mandatory properties. In order to solve the problem, we proposed another two desired properties, which are the fourth and fifth properties. The fourth property is that it is impossible to find any strong positive correlation from itemsets occurring rarely. In other words, it at least has to happen several times in order to be statistically validated. We want to find significant patterns rather than coincidences. This property helps to rule out measures that cannot tell correlated patterns from uncorrelated patterns well. For the fifth property, we give some penalty for adding independent items in order to make highly correlated itemsets stand out. In the extreme case, when a lot of independent items are added, the final itemset is dominated by the independence and the correlation value should be very close to the constant C. The fifth property is related to an agreement of how to rank differently.

In addition, we proposed two optional properties, which are the sixth and seventh properties. The sixth property expects the strongest negatively correlated itemsets coming from the low Support region. It is quite intuitive since the stronger negative correlation means a lower chance of them happening together. The fourth property checks whether correlation measures can correctly evaluate positive correlation, while the sixth property checks the correct evaluation for negative correlation. Therefore, if users are only interested in positive correlation, it doesn't matter if the correlation measure cannot satisfy the sixth property. Similarly, if users are only interested in negative correlation, it doesn't matter if the correlation measure cannot satisfy the fourth property. However, we treat the fourth property as desired and the sixth property as optional for the following reason. If we consider the absence of an item I as the presence of the absence, we can find a positive correlation like $\{A, \bar{B}, C\}$ and we understand how A, B, and C are correlated with each other. From the negative correlation $\{A, B, C\}$, we don't know how A, B, and C are correlated with each other. The seventh property indicates that the correlation measure should increase our confidence about the positive or negative correlation of the given itemset S when we get more sample data from the same population. However, we prefer it to be optional for two reasons. First, we

Table I. A Two-Way Contingency Table for Variables A and B

	B	$\bar{B}$	$\sum$ Row
A	f_{11}	f_{10}	f_{1+}
$\bar{A}$	f_{01}	f_{00}	f_{0+}
$\sum$ Column	f_{+1}	f_{+0}	N

Table II. The Grade–Gender Example from 1993

	Male	Female	$\sum$ Row
High	30	20	50
Low	40	10	50
$\sum$ Column	70	30	100

Table III. The Grade–Gender Example from 2004

	Male	Female	$\sum$ Row
High	60	60	120
Low	80	30	110
$\sum$ Column	140	90	230

can always make our correlation measures a function of the sample size isolated from other parameters. In that way, we can either keep the sample size parameter to satisfy the last property or drop the sample size parameter. Second, we might want to compare the correlation from different sources that have different sample sizes. In order to conduct a fair comparison, we might not want the correlation measure to satisfy the last property.

2.1.2. Additional Properties for Pair-Only Measures. In addition to the previous seven properties for both general and pair-only measures, Tan et al. [2005] proposed three additional properties for the pair-only-type measures based on operations for 2×2 contingency tables. Table I shows a typical 2×2 contingency table for a pair of binary variables, A and B. Each entry f_{ij} in this 2×2 tables denotes a frequency count.

The three proposed properties on a two contingency table are as follows:

AP1: M remains the same when exchanging the frequency count f_{11} with f_{00} and f_{10} with f_{01}.

AP2: M remains the same by only increasing f_{00}.

AP3: M remains the same under the row/column scaling operation from Table T to T', where T is a contingency table with frequency counts $[f_{11}; f_{10}; f_{01}; f_{00}]$, T' is a contingency table with scaled frequency counts $[k_1k_3\,f_{11}; k_2k_3\,f_{10}; k_1k_4\,f_{01}; k_2k_4\,f_{00}]$, and k_1, k_2, k_3, k_4 are positive constants.

The first property is the inversion property, which argues that the correlation of A and B should be the same as that of $\bar{A}$ and $\bar{B}$. The second property is the null addition property. Tan et al. argued that the correlation between two items should not be affected by adding unrelated data to a given dataset. For example, we are interested in analyzing the relationship between a pair of words, such as *data* and *mining*, in a set of computer science papers. The association between *data* and *mining* should not be affected by adding a collection of articles about ice fishing. The third property is the scaling property. Mosteller [1968] presented the following example to illustrate the scaling property. Tables II and III show the contingency tables for gender and the grades achieved by students enrolled in a particular course in 1993 and 2004, respectively. The data in these tables showed that the number of male students has doubled since 1993, while the number of female students has increased by a factor of

Table IV. The Original Table

	B	$\bar{B}$	$\sum$ Row
A	5	4	9
$\bar{A}$	11	80	91
$\sum$ Column	16	84	100

Table V. The Modified Table

	B	$\bar{B}$	$\sum$ Row
A	5	7	12
$\bar{A}$	7	81	88
$\sum$ Column	12	88	100

3. However, the male students in 2004 are not performing any better than those in 1993 because the ratio of male students who achieve a high grade to those who achieve a low grade is still the same, that is, 3:4. Similarly, the female students in 2004 are performing the same as those in 1993.

However, the purpose of these three properties proposed by Tan et al. is to categorize correlation measures. They don't provide any guideline for choosing the proper correlation measure according to users' situations. In fact, we argue that these three properties are not desirable correlation properties.

For the first inversion property, we argue that the correlation of A and B is determined by f_{11} or the other three occurrences f_{01}, f_{10}, and f_{00} instead of f_{00} alone. In Table IV, we can fix the true probability and the expected probability of $A \cap B$, but alternate the values of f_{01}, f_{10}, and f_{00} to generate another table, Table V. Since the true probability and the expected probability of A and B are the same in these two tables, the correlation of A and B in these two tables is the same. If the inversion property stands, we can conclude that the correlation of $\bar{A}$ and $\bar{B}$ in Table IV is equal to the correlation of $\bar{A}$ and $\bar{B}$ in Table V, which is controversial.

For the second null addition property, the correlation of A and B should be the same if only we fix the values of f_{11}, f_{01}, and f_{10}. Given the extreme example that we add a huge number of ice fishing documents into the set of computer science papers, the ice fishing documents start to dominate the corpus and the set of computer science papers becomes the background noise. Although the co-occurrence of the pair of words "data" and "mining" is not changed, intuitively, we don't want the correlation between "data" and "mining" to be as strong as that in the initial setting since it is the background noise now. We can also analyze this case in another way. The actual probability of A and B is $f_{11}/(f_{11} + f_{01} + f_{10} + f_{00})$ and the expected probability of A and B is $(f_{11} + f_{10})/(f_{11} + f_{01} + f_{10} + f_{00}) \cdot (f_{11} + f_{01})/(f_{11} + f_{01} + f_{10} + f_{00})$. For the pair A and B, both the actual probability and the expected probability decrease when f_{00} increases. The decrease of the actual probability lowers the correlation according to Property 2, and the decrease of the expected probability increases the correlation according to Property 3. The final change of correlation is due to the tradeoff between the effect from the actual probability and that from the expected probability, and it is unnecessary for these to be the same. When we add a huge number of ice fishing documents into the set of computer science papers, the Support of the pair of words "data" and "mining" is close to 0. The correlation of the pair of words "data" and "mining" should also be close to the constant number C according to Property 4, which also contradicts the null addition property.

For the third scaling property, let's reconsider the gender–grade example shown in Tables II and III. Though the ratio of male students who achieved a high grade to those who achieved a low grade in 1993 is still the same as that of 2004, the ratio of male

students who achieved a high grade to those who achieved a low grade is different from that of females. Since the portion of the male students has changed from 1993 to 2004, we are expecting that the high-grade students are less likely to be male in 2004. The correlation between grade and gender should be changed.

Though we doubt the three addition properties qualify as desirable properties and the experimental results in this article support our arguments, it is still up to users' choice.

2.2. Formulas and Property Satisfaction

In this section, we study the correlation measures for both the general and the pair-only types and their property satisfaction.

2.2.1. General Correlation Measures.

Support. Support of the itemset S is the proportion of transactions that contain S. Using Support, the level of correlation is simply the fraction of times that the items co-occur. As a metric, Support facilitates fast search, but it has drawbacks [Brin et al. 1997a, 1997b]. For example, the finding that A and B occur together in 81% of the transactions is not interesting if A and B both occur in 90% of the transactions. This would be expected since $P(A) = P(A|B)$ and $P(B) = P(B|A)$. A and B are not correlated, even though they together have very high Support. Among all the seven properties mentioned previously, Support only satisfies Properties 2 and 6.[1] Since even two mandatory properties are violated by Support, it is a poor correlation measure.

Any-confidence. Any-confidence [Omiecinski 2003] of the itemset S is the ratio of its probability to the probability of the item with the lowest probability in S: $AnyConfidence(S) = P(S)/min(P(I_1), P(I_2), \ldots, P(I_m))$. The value of Any-confidence is the upper bound of the confidence of all association rules that can be generated from the itemset S. It helps us to determine whether we can find at least one rule that has a confidence greater than the specified threshold. However, it is not designed as a correlation measure and does not have a downward closure property to facilitate search.

All-confidence. All-confidence [Omiecinski 2003] of the itemset S is the ratio of its probability to the probability of the item with the highest probability in S: $AllConfidence(S) = P(S)/max(P(I_1), P(I_2), \ldots, P(I_m))$. The value of All-confidence is the lower bound of the confidence of all association rules that can be generated from the itemset S. Although All-confidence itself possesses the downward closure property to facilitate search, it is not designed as a correlation measure and suffers the same problems as Support. Theoretically, All-confidence lacks comparison to the expected probability under the independence assumption. Like Support, it satisfies only the second and sixth of the seven desired correlation measure properties. Practically, All-confidence shares three problems with Support. First, it is biased toward itemsets with high Support items. If an itemset S consists of independent, high Support items, $Support(S)$ will be high (despite the independence), and $AllConfidence(S)$ will also be high. This problem is exaggerated if we extend our search to include the presence of some items and the absence of others, since absence of a rare item is itself a high Support item. This is typically not relevant in marketing but could be in, for example, genetic data. Second, intuitively, we want exact-length correlation patterns. However, All-confidence is biased to short itemsets as its value decreases monotonically with increasing itemset size. More maximal All-confidence sets are likely to be 2-itemsets like maximal frequent itemsets. Third, the antimonotone property makes it difficult to compare correlation among itemsets of different sizes.

[1]The property satisfaction proofs related to each correlation measure are in the appendix.

Bond / Jaccard. Bond [Omiecinski 2003] of the itemset S is the ratio of its probability to the probability of the union of all the items in S: $Bond(S) = P(S)/P(I_1 \cup I_2 \cup \cdots \cup I_m)$. Usually, Jaccard is used for pairs and Bond is used for itemsets, but they share the same idea. Bond is similar to Support but with respect to a related subset of the data rather than the entire dataset. Like Support and All-confidence, Bond possesses the downward closure property. Given a set of strongly related rare items, both Bond and All-confidence can assign a high score for this itemset, which can relieve the disadvantage of Support. However, worse than All-confidence, Bond satisfies only the sixth of the seven correlation measure properties and measures correlation in a suboptimal way.

The Simplified χ^2-statistic. The χ^2 is calculated as $\chi^2 = \sum_i \sum_j (r_{ij} - E(r_{ij}))^2/E(r_{ij})$. If an itemset contains n items, 2^n cells in the contingency table must be considered for the previous Pearson χ^2 statistic. The computation of the statistic itself is intractable for high-dimensional data. However, we can still use the basic idea behind the χ^2-statistic to create the Simplified χ^2-statistic: $\chi'^2 = (r - E(r))^2/E(r)$, that is, $n \cdot (tp - ep)^2/ep$, where the cell r corresponds to the exact itemset S. Since the Simplified χ^2-statistic is more computationally desirable, in the rest of the article we only discuss the properties and experimental results of the Simplified χ^2-statistic. The value of the Simplified χ^2-statistic is always larger than 0 and cannot differentiate positive from negative correlation. Therefore, we take advantage of the comparison between tp and ep. If $tp > ep$, it is a positive correlation. Then the Simplified χ^2-statistic is equal to $n \cdot (tp - ep)^2/ep$. If $tp < ep$, it is a negative correlation. Then the Simplified χ^2-statistic is equal to $-n \cdot (tp - ep)^2/ep$. This transformed Simplified χ^2-statistic is mathematically favorable. Larger positive numbers indicate stronger positive correlation, 0 indicates no correlation, and larger (in magnitude) negative numbers indicate stronger negative correlation.

Probability Ratio / Lift / Interest Factor. Probability Ratio [Brin et al. 1997b] is the ratio of its actual probability to its expected probability under the assumption of independence. It is calculated as follows: $ProbabilityRatio(S) = tp/ep$. This measure is straightforward and means how many times the itemset S happens more than expected. In some cases, we also use the log value of Probability Ratio. In that way, we make the constant number C in the first mandatory property 0, which is consistent with other measures. However, this measure might not still be a reasonable correlation measure to use. The problem is that it favors the itemsets containing a large number of items rather than significant trends in the data. For example, given a common transaction containing 30 items and each item in this transaction has a 50% chance to be bought individually, the expected probability for this transaction is $9.31 * 10^{-10}$ if all the items are independent. Even if this transaction coincidentally happened once out of 1 million transactions, its Probability Ratio is 1,073, which is still very high. However, a single transaction is hardly something that we are interested in.

Leverage. An itemset S with higher Support and low Probability Ratio may be more interesting than an alternative itemset S' with low Support and high Probability Ratio. Introduced by Piatesky-Shapiro [1991], $Leverage(S) = tp - ep$. It measures the difference between the actual probability of an itemset S and its expected probability if all the items in S are independent from each other. Since $\prod_{i=1}^{m} P(I_i)$ is always no less than 0, $Leverage(S)$ can never be bigger than $P(S)$. Therefore, Leverage is biased to high Support itemsets.

Likelihood Ratio. Likelihood Ratio is similar to a statistical test based on the log-likelihood ratio described by Dunning [1993]. The concept of a likelihood measure can

be used to statistically test a given hypothesis by applying the Likelihood Ratio test. Essentially, we take the ratio of the highest likelihood possible given our hypothesis to the likelihood of the best "explanation" overall. The greater the value of the ratio is, the stronger our hypothesis will be.

To apply the Likelihood Ratio test as a correlation measure, it is useful to consider the binomial distribution. This is a function of three variables: $Pr(p, k, n) \rightarrow [0 : 1]$. Given our assumption of independence of all items in the itemset S, we predict that each trial has a probability of success ep. Then the binomial likelihood of observing k out of n transactions containing S is $Pr(ep, k, n)$. However, the best possible explanation of the probability of containing S is tp instead of ep. Therefore, we perform the Likelihood Ratio test, comparing the binomial likelihood of observing k transactions under the assumption of independence with the best possible binomial explanation. Formally, the Likelihood Ratio in this case is $LikelihoodRatio(S) = Pr(tp, k, n)/Pr(ep, k, n)$.

In the rest of the article, we use a transformed Likelihood Ratio to measure correlation for two reasons. First, since the actual Likelihood Ratio could be extremely large, we use the ln value instead of its original value. Second, the numerator of the Likelihood Ratio is the maximal likelihood of the real situation, so the Likelihood Ratio is always larger than 1 and cannot differentiate positive from negative correlation. When calculating the transformed Likelihood Ratio, we take advantage of the comparison between tp and ep. If $tp > ep$, it is a positive correlation. Then the transformed Likelihood Ratio is equal to $ln(LikelihoodRatio(S))$. If $tp < ep$, it is a negative correlation. Then the transformed Likelihood Ratio is equal to $-ln(LikelihoodRatio(S))$. This transformed Likelihood Ratio is mathematically favorable. Larger positive numbers indicate stronger positive correlation, 0 indicates no correlation, and larger (in magnitude) negative numbers indicate stronger negative correlation.

BCPNN. Probability Ratio is straightforward and means how many times the combination happens more than expected. However, the Probability Ratio is very volatile when the expected value is small, which makes it favor coincidences rather than significant trends in the data. In order to solve the problem, we use shrinkage [Bate et al. 1998; Dumouchel 1999; Norén et al. 2008] to regularize and reduce the volatility of a measure by trading a bias to no correlation for decreased variance. For an itemset S, the calculated tp is 0 if S is not observed in the dataset. However, we might get some transactions containing S if we get more samples. In order to make the conservative estimation to the ground tp and ep, we add a continuity correction number here. Suppose the continuity correction is cc; the formula of BCPNN is $BCPNN = ln(tp + cc)/(ep + cc)$. Normally, we set $cc = 0.5/n$ when the dataset is relatively clean; however, it could be any positive number theoretically. Especially when the dataset contains a lot of noisy data, we might use a larger number to make a more conservative estimate. This shrinkage strength has been successfully applied to pattern discovery in the analysis of large collections of individual case safety reports. Norén et al. [2008] claimed that it precludes highlighting any pattern based on less than three events but is still able to find strongly correlated rare patterns. In general, the strength and direction of the shrinkage can be adjusted by altering the magnitude and ratio of the constants added to the nominator and denominator, which will be fully discussed in Section 2.3. From a frequency perspective, BCPNN is a conservative version of Probability Ratio, tending toward 0 for rare events and with better variance properties. As tp and ep increase, the impact of the shrinkage diminishes.

LEMMA 1. *Given the ground true probability p for an itemset S in the dataset with n transactions, the variance of Probability Ratio approaches to ∞ and the variance of BCPNN approaches to 0 when $p \rightarrow 0$.*

PROOF. Since the occurrence of S, denoted by X, follows the binomial distribution, we get $E(X) = n \cdot p$ and $Var(X) = n \cdot p \cdot (1 - p)$.

(1) $ProbabilityRatio = X/(n \cdot p)$. Therefore, $Var(ProbabilityRatio) = Var(X)/(n^2 \cdot p^2) = (1 - p)/(n \cdot p)$. According to the formula, we can see that $Var(ProbabilityRatio) \to \infty$ when $p \to 0$.

(2) $BCPNN = (X+cc)/(n \cdot p+cc)$. Therefore, $Var(BCPNN) = Var(X + cc)/(n \cdot p + cc)^2 = (n \cdot p - n \cdot p^2)/(n^2 \cdot p^2 + 2 \cdot cc \cdot n \cdot p + cc^2)$. As $p \to 0$, $Var(BCPNN) \to 0/cc^2 = 0$. □

Simplified χ^2 with Continuity Correction Inspired by the shrinkage technique applied to BCPNN, we propose a new correlation measure, Simplified χ^2 with Continuity Correction (SCWCC). Suppose the continuity correction is cc; we add cc additional occurrence to both the actual events and the expected events. The formula of SCWCC is $SCWCC = n \cdot (tp - ep)^2/(ep + cc)$. As tp gets closer to 0, the upper bound of Likelihood Ratio, Leverage, BCPNN, and SCWCC gets closer to the constant number C. However, different measures have different biases toward different Support regions, which will be discussed in Section 2.3.

IS measure. Since interest factor (Probability Ratio) favors rare combinations rather than significant trends in the data and Support favors frequent combinations rather than strong correlation, Tan et al. [2000] proposed Interest factor with Support (IS), which is the square root of the product of Interest factor and Support, that is, $IS(S) = \sqrt{ProbabilityRatio(S) \cdot Support(S)} = tp/\sqrt{ep}$. When measuring pairs for binary data, IS is exactly cosine similarity, which is $n \cdot P(A \cap B)/\sqrt{n \cdot P(A) \cdot n \cdot P(B)} = tp/\sqrt{ep}$. Therefore, we treat the cosine similarity as a special case of IS when measuring binary data. Intuitively, IS is large when both Probability Ratio and Support are large enough. In fact, the IS value of a rare large combination is still very large, which is not that much better than Probability Ratio. In addition, when all the items in the itemset S are independent of each other, that is, $tp = ep$, $IS = \sqrt{ep}$, which is not constant. It violates the first mandatory property.

Two-way Support / the Simplified Mutual Information. Sharing the same idea with IS, Zhong et al. [2001] proposed the Two-way Support measure, which is the product of Support and the log value of Probability Ratio, that is, $TwoWaySupport(S) = tp \cdot ln(tp/ep)$. It adopts the exact idea of Mutual Information [Everett 1957] to measure the correlation for the target cell. The relationship between Two-way Support and Mutual Information is very similar to that between Simplified χ^2-statistic and χ^2-statistic mentioned earlier. The computation of the Mutual Information itself is intractable for high-dimensional data. Therefore, Two-way Support, a more computational desirable version, is selected for evaluation. Better than IS, Two-way Support satisfies the first mandatory property and uses the log value of Probability Ratio to suppress the increase of Probability Ratio. As Support approaches 0, the decrease from Support dominates the increase from the log value of Probability Ratio; that is, its upper bound is close to 0. However, the side effect is that its lower bound also approaches 0 when Support is close to 0. In other words, there are no significant negatively correlated patterns for low Support itemsets, which is wrong.

Simplified χ^2 with Support. Both Simplified χ^2 and Probability Ratio favor rare combinations rather than significant trends in the data. Inspired by the IS measure, we propose a new correlation measure called Simplified χ^2 with Support (SCWS), which is the product of Simplified χ^2 and Support. The formula of SCWS is $SCWS(S) = tp \cdot (tp - ep)^2/ep$. Better than the IS measure, SCWS satisfies the first mandatory property. However, the same as the IS measure, the SCWS value of a rare large combination is still very large. In addition, the same as Two-way Support, the SCWS value of the

Table VI. The Conditional Probability Table for Variables A and B

	B	$\bar{B}$
A	f_{11}/f_{1+}	f_{10}/f_{1+}
$\bar{A}$	f_{01}/f_{0+}	f_{00}/f_{0+}

negatively correlated itemset gets closer to the constant number C when the Support of this itemset gets closer to 0, which is not qualified for the negative correlation search.

2.2.2. Pair-Only Correlation Measures. Given the typical 2×2 contingency table for a pair of binary variables in Table I, the commonly used pair-only-type correlation measures are calculated as follows.

ϕ Correlation Coefficient. The ϕ Correlation Coefficient [Reynold 1977] is derived from Pearson's Correlation Coefficient for binary variables. The formula of the ϕ Correlation Coefficient is as follows: $(f_{00}f_{11} - f_{01}f_{10})/\sqrt{f_{0+}f_{1+}f_{+0}f_{+1}}$. It measures the linear relationship between two binary variables.

Relative Risk. Relative Risk [Sistrom and Garvan 2004] is the ratio of the probability of the event occurring in the exposed group versus a nonexposed group. It is often used to compare the risk of developing a side effect in people receiving a drug versus people not receiving the treatment. Given Table I, the Relative Risk for the event B within the two situations defined by A and $\bar{A}$ is $\frac{f_{11}/f_{1+}}{f_{01}/f_{0+}}$.

Odds Ratio. The Odds Ratio [Mosteller 1968] is a measure of effect size, describing the strength of nonindependence between two binary variables and comparing them symmetrically. It plays an important role in logistic regression. The Odds Ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. In Table I, the odds for B within the two subpopulations defined by A and $\bar{A}$ are defined in terms of the conditional probabilities in Table VI. Thus, the Odds Ratio is $(\frac{f_{11}/f_{1+}}{f_{10}/f_{1+}})/(\frac{f_{01}/f_{0+}}{f_{00}/f_{0+}}) = \frac{f_{11}*f_{00}}{f_{10}*f_{01}}$. The final expression is easy to remember as the product of the concordant cells ($A = B$) divided by the product of the discord cells ($A \neq B$). Since Relative Risk is a more intuitive measure of effectiveness, the distinction is important especially in cases of medium to high probabilities. If action A carries a risk of 99.9% and action B a risk of 99.0% then the Relative Risk is just over 1, while the odds associated with action A are almost 10 times higher than the odds with B. In medical research, the Odds Ratio is favored for case-control studies and retrospective studies. Relative Risk is used in randomized controlled trials and cohort studies.

Conviction. Conviction [Brin et al. 1997b] is calculated as $f_{1+} * f_{+0}/f_{10}$. Logically, $A \rightarrow B$ can be rewritten as $\neg(A \wedge \neg B)$. Similar to Lift, $f_{11}/(f_{1+} * f_{+1})$, which seeks the deviation from independence between A and B, Conviction examines how far $A \wedge \neg B$ deviates from independence. In other words, Conviction looks for the correlation underlying the rule $A \rightarrow B$ instead of the pair A and B.

Added Value. Similar to Conviction, Added Value [Zhong et al. 1999] is a measure for rules instead of pairs. Given the rule $A \rightarrow B$, Added Value measures the difference between $Support(B)$ in the whole population and that in the population A. Specifically, $AddedValue(A \rightarrow B) = P(B|A) - P(B)$. If we transform the formula, we get $AddedValue(A \rightarrow B) = (P(A \wedge B) - P(A) \cdot P(B))/P(A) = Leverage(A, B)/P(A)$. It is the Leverage tuned by $P(A)$. When $P(A)$ is small, $Leverage(A, B)$ will also be small. The Added Value tries to divide the Leverage by $P(A)$ to prompt the correlated pattern in the small population.

Table VII. Formulas of Correlation Measures

Correlation Measure	Formula
Support	tp
Any-confidence	$\frac{tp}{min(P(I_1),P(I_2),...,P(I_m))}$
All-confidence	$\frac{tp}{max(P(I_1),P(I_2),...,P(I_m))}$
Bond	$\frac{tp}{P(I_1 \cup I_2 \cup \cdots \cup I_m)}$
Simplified χ^2-statistic	$n \cdot \frac{(tp-ep)^2}{ep}$
Probability Ratio	$ln\ \frac{tp}{ep}$
Leverage	$tp - ep$
Likelihood Ratio	$n \cdot [tp \cdot ln\ \frac{tp}{ep} + (1-tp) \cdot ln\ \frac{1-tp}{1-ep}]$
BCPNN	$ln\ \frac{tp+cc}{ep+cc}$
SCWCC	$n \cdot \frac{(tp-ep)^2}{ep+cc}$
IS	$\frac{tp}{\sqrt{ep}}$
Two-way Support	$tp \cdot ln\ \frac{tp}{ep}$
SCWS	$n \cdot tp \cdot \frac{(tp-ep)^2}{ep}$
ϕ-coefficient	$\frac{f_{00} \cdot f_{11} - f_{10} \cdot f_{01}}{\sqrt{f_{1+} \cdot f_{0+} \cdot f_{+1} \cdot f_{+0}}}$
Relative Risk	$\frac{f_{11}/f_{1+}}{f_{01}/f_{0+}}$
Odds Ratio	$\frac{f_{11} \cdot f_{00}}{f_{10} \cdot f_{01}}$
Conviction	$\frac{f_{1+} \cdot f_{+0}}{n \cdot f_{10}}$
Added Value	$\frac{n \cdot f_{11} - f_{1+} \cdot f_{+1}}{n \cdot f_{1+}}$

2.2.3. Summary of Correlation Measures. We have categorized both general and pair-only correlation measures. The general type can be further divided into three subcategories: suboptimal measures, basic measures, and adjusted measures. Support, Any-confidence, All-confidence, and Bond are the suboptimal measures. All of them have no direct comparison with the expected probability and violate more than one mandatory correlation property. The basic correlation measures, derived from simple statistical theories, include Simplified χ^2, Probability Ratio, Leverage, and Likelihood Ratio. They satisfy all three mandatory properties, but they might violate some desirable correlation properties. The measures adjusted by continuity correction are BCPNN and Simplified χ^2 with Continuity Correction. They use the shrinkage technique to reduce the volatility of a measure by trading a bias to no correlation for decreased variance. In this way, we modify the basic correlation measures to satisfy all the desirable properties. The measures adjusted by Support include IS, Two-way Support, and SCWS. They try to adjust the basic measures by multiplying Support to suppress the increase from correlation measures when Support is close to 0. Table VII shows the original formulas of measures, and Table VIII is a summary of the original version measures with regard to all 10 properties.

2.3. The Upper and Lower Bounds of Measures

Among 18 correlation measures we study, the three mandatory properties proposed by Piatetsky-Shapiro only screen out five measures. The two desired properties proposed by us together with the three mandatory properties can successfully narrow down the

Table VIII. Properties of Correlation Measures

Correlation Measure	P1	P2	P3	P4	P5	P6	P7	AP1	AP2	AP3
Support		X		X		X				
Any-confidence		X		X		X			X	
All-confidence		X		X		X			X	
Bond				X		X			X	
Simplified χ^2-statistic	X	X	X		X	X	X			
Probability Ratio	X	X	X			X				
Leverage	X	X	X	X	X	X		X		
Likelihood Ratio	X	X	X	X	X	X	X			
BCPNN	X	X	X	X	X	X				
SCWCC	X	X	X	X	X	X	X			
IS		X	X			X			X	
Two-way Support	X	X	X	X	X					
SCWS	X	X	X		X		X			
ϕ-coefficient	X	X	X	X		X		X		
Relative Risk	X	X	X			X				
Odds Ratio	X	X	X			X		X		X
Conviction	X	X	X			X				
Added Value	X	X	X	X		X				

candidate list to five measures: Leverage, Likelihood Ratio, BCPNN, Two-way Support, and SCWCC. Since the candidate list still has five measures, the natural question is, "Do they retrieve the same results? If not, what are the differences?" In order to check the difference, we study the upper and lower bounds of different measures when tp is fixed and discuss the tradeoff between Support and itemset size in this section.

Support. Since $Support(S) = tp$, both the upper bound and the lower bound of $Support(S)$ is tp.

Any-confidence. Since $tp \leq P(I_i) \leq 1$ for each item I_i in S, the minimal value of $min(P(\{I_i|I_i \in S\}))$ is tp. Suppose the maximal value of $min(P(\{I_i|I_i \in S\}))$ is x; then we need to find the maximal x that satisfies $P(I_1) \geq x$, $P(I_2) \geq x, \ldots,$ and $P(I_m) \geq x$. In order to maintain the value of tp, x is the maximal value that all the $P(I_i)$ can reach simultaneously. According to Theorem 1, we get $x = (m-1+tp)/m$. Therefore, the upper bound of Any-confidence(S) is $tp/tp = 1$ and the lower bound is $tp/((m-1+tp)/m) = m \cdot tp/(m-1+tp)$.

All-confidence. Since $tp \leq P(I_i) \leq 1$ for each item I_i in S and $P(I_1 \cap I_2 \cap \ldots \cap I_m) = tp$ holds when each $P(I_i) = tp$, the minimal value of $max(P(\{I_i|I_i \in S\}))$ is tp when each $P(I_i) = tp$. When a certain $P(I_i) = 1$, it is still possible for $P(I_1 \cap I_2 \cap \ldots \cap I_m) = tp$. Therefore, the maximal value of $max(P(\{I_i|I_i \in S\}))$ is 1. Then, the upper bound of All-confidence(S) is $tp/tp = 1$ and the lower bound is $tp/1 = tp$.

Bond. Since $tp \leq P(I_i) \leq 1$ for each item I_i in S and $P(I_1 \cap I_2 \cap \ldots \cap I_m) = tp$ holds when each $P(I_i) = tp$, the minimal value of $P(I_1 \cup I_2 \cup \cdots \cup I_m)$ is tp when each $P(I_i) = tp$. When a certain $P(I_i) = 1$, it is still possible for $P(I_1 \cap I_2 \cap \ldots \cap I_m) = tp$. Therefore, the maximal value of $P(I_1 \cup I_2 \cup \cdots \cup I_m)$ is 1. Then, the upper bound of Bond(S) is $tp/tp = 1$ and the lower bound is $tp/1 = tp$.

Correlation measures satisfying Property 3. In the following, we study the upper and lower bounds of the correlation measure satisfying Property 3.

THEOREM 1. *Given an itemset $S = \{I_1, I_2, \ldots, I_m\}$ with the actual probability tp, its expected probability ep is no less than tp^m and no more than $((m-1+tp)/m)^m$.*

PROOF. (1) According to definition, $ep = \prod_{i=1}^{m} P(I_i = 1)$. For each item I_i in S, $tp \leq P(I_i = 1) \leq 1$. When the actual probability of each item I_i reaches the lower bound tp and all the items occur together, the expected probability ep reaches its lower bound tp^m.

(2) Given the itemset $\{I_1, I_2, \ldots, I_m\}$, we have

$$\sum_{I_1=0}^{I_1=1} \sum_{I_2=0}^{I_2=1} \cdots \sum_{I_m=0}^{I_m=1} P(I_1, I_2, \ldots, I_m) = 1, \tag{1}$$

and the Support $P(I_1 = 1)$ for each item I_1 is

$$\sum_{I_2=0}^{I_2=1} \sum_{I_3=0}^{I_3=1} \cdots \sum_{I_m=0}^{I_m=1} P(I_1 = 1, I_2, \ldots, I_m).$$

Given the cell $\{I_1 = 1, I_2, \ldots, I_p = 0, \ldots, I_q = 0, \ldots, I_m\}$ with more than two items having value 0, if its probability is greater than 0, we can decrease its probability to 0 and increase the probability of the cell $\{I_1 = 1, I_2, \ldots, I_p = 1, \ldots, I_q = 0, \ldots, I_m\}$ (or $\{I_1 = 1, I_2, \ldots, I_p = 0, \ldots, I_q = 1, \ldots, I_m\}$). By doing that, we keep $P(I_1 = 1)$ the same but increase $P(I_p = 1)$ (or $P(I_q = 1)$).

Since $ep = \prod_{i=1}^{m} P(I_i = 1)$, ep can be increased by adjusting the probability of the cell with more than two absent items to 0. Therefore, in order to get the maximal ep, we can simplify Equation (1) to

$$P(I_1 = 1, I_2 = 1, \ldots, I_m = 1) + \sum_{i=1}^{m} P(I_1 = 1, I_2 = 1, \ldots, I_i = 0, \ldots, I_m = 1) = 1.$$

Since we know $P(I_1 = 1, I_2 = 1, \ldots, I_m = 1) = tp$, then

$$\sum_{i=1}^{m} P(I_1 = 1, I_2 = 1, \ldots, I_{i-1} = 1, I_i = 0, I_{i+1} = 1 \ldots, I_m = 1) = 1 - tp.$$

Therefore, we have

$$P(I_i = 1) = 1 - P(I_1 = 1, I_2 = 1, \ldots, I_{i-1} = 1, I_i = 0, I_{i+1} = 1 \ldots, I_m = 1)$$

and

$$\begin{aligned} \sum_{i=1}^{m} P(I_i = 1) &= m - \sum_{i=1}^{m} P(I_1 = 1, \ldots, I_{i-1} = 1, I_i = 0, I_{i+1} = 1 \ldots, I_m = 1) \\ &= m - 1 + tp. \end{aligned}$$

In order to get the maximal $ep = \prod_{i=1}^{m} P(I_i = 1)$ when $\sum_{i=1}^{m} P(I_i = 1) = m - 1 + tp$, we have $P(I_1 = 1) = P(I_2 = 1) = \cdots = P(I_m = 1) = (m-1+tp)/m$. Therefore, the upper bound of ep is $((m - 1 + tp)/m)^m$. □

THEOREM 2. *The lower bound of ep, tp^m, is no more than tp and its upper bound $(\frac{m-1+tp}{m})^m$ is no less than tp.*

PROOF. (a) Since $0 \leq tp \leq 1$ and m is a positive integer larger than 1, $tp^m \leq tp$.

(b) Let $f(tp) = (\frac{m-1+tp}{m})^m - tp$ be a function of tp; then we have $f'(tp) = (\frac{m-1+tp}{m})^{(m-1)} - 1$. Since $0 \leq tp \leq 1$, we have $\frac{m-1}{m} \leq \frac{m-1+tp}{m} \leq \frac{m}{m}$. Therefore, $(\frac{m-1+tp}{m})^{(m-1)} \leq 1$ and $f'(tp) \leq 0$. Thus, $f(tp) \geq f(1) = 0$. We get $(\frac{m-1+tp}{m})^m - tp \geq 0$, that is, $(\frac{m-1+tp}{m})^m \geq tp$. □

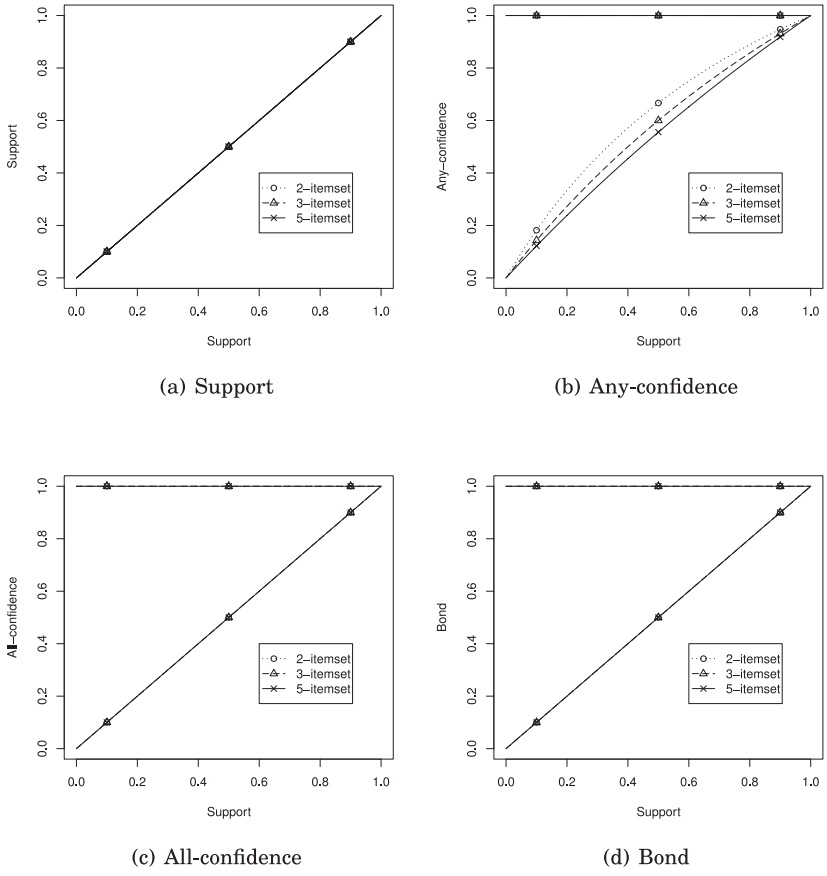
Table IX. Bounds of Correlation Measures

Correlation Measure	Upper Bound	Lower Bound
Support	tp	tp
Any-confidence	1	$\frac{m \cdot tp}{m-1+tp}$
All-confidence	1	tp
Bond	1	tp
Simplified χ^2-statistic	$\frac{(tp-tp^m)^2}{tp^m}$	$-(tp-(\frac{m-1+tp}{m})^m)^2 \cdot (\frac{m}{m-1+tp})^m$
Probability Ratio	$\frac{tp}{tp^m}$	$tp \cdot (\frac{m}{m-1+tp})^m$
Leverage	$tp - tp^m$	$tp - (\frac{m-1+tp}{m})^m$
Likelihood Ratio	$tp \cdot ln\ \frac{tp}{tp^m} + (1-tp) \cdot ln\ \frac{1-tp}{1-tp^m}$	$-tp \cdot ln\ \frac{tp \cdot m^m}{(m-1+tp)^m} - (1-tp) \cdot ln\ \frac{(1-tp) \cdot m^m}{m^m-(m-1+tp)^m}$
BCPNN	$\frac{tp+cc}{tp^m+cc}$	$\frac{(tp+cc) \cdot m}{m-1+tp+cc \cdot m}$
SCWCC	$\frac{(tp-tp^m)^2}{tp^m+cc}$	$-\frac{(tp \cdot m^m-(m-1+tp)^m)^2}{m^m \cdot (m-1+tp)^m+cc \cdot m^{2m}}$
IS	$tp^{(1-m/2)}$	$tp \cdot (\frac{m}{m-1+tp})^{m/2}$
Two-way Support	$(1-m) \cdot tp \cdot ln(tp)$	$tp \cdot ln(tp) - m \cdot tp \cdot ln\ \frac{m-1+tp}{m}$
SCWS	$\frac{tp \cdot (tp-tp^m)^2}{tp^m}$	$-\frac{tp \cdot (tp-((m-1+tp)/m)^m)^2}{((m-1+tp)/m)^m}$
ϕ-coefficient	1	$-\frac{1-tp}{1+tp}$
Relative Risk	∞	tp
Odds Ratio	∞	0
Conviction	∞	tp
Added Value	$1-tp$	$-\frac{(1-tp)^2}{2(1+tp)}$

THEOREM 3. *Given any correlation measure that satisfies Property 3 and an itemset* $S = \{I_1, I_2, \ldots, I_m\}$ *with fixed tp, the correlation measure reaches its upper bound when* $ep = tp^m$ *and reaches its lower bound when* $ep = ((m-1+tp)/m)^m$.

PROOF. Given any correlation measure that satisfies Property 3, its correlation value should monotonically decrease with the increase of ep when tp is fixed. In other words, this measure reaches its upper bound when ep reaches the lower bound tp^m, given the itemset size m and the actual probability tp. Similarly, any correlation measure that satisfies Property 3 reaches its lower bound when ep reaches the upper bound. □

Pair-only correlation measures. All the pair-only correlation measures in this article satisfy Property 3. Given tp and $m = 2$, $tp^2 \leq ep \leq ((1+tp)/2)^2$. They reach their upper bound when $ep = tp^2$. When $ep = tp^2$, we have $f_{10} = 0$, $f_{01} = 0$, and $f_{00} = n - f_{11}$. According to the formulas shown in Table VII, it is easy to get $\phi_{ub} = 1$, $RelativeRisk_{ub} = \infty$, $OddsRatio_{ub} = \infty$, $Conviction_{ub} = \infty$, and $AddedValue_{ub} = 1 - tp$. In order to get their lower bounds, there are three promising situations: (1) $f_{11} = tp$, $f_{10} = 1 - tp - \epsilon$, $f_{01} = \epsilon$, and $f_{00} = 0$, where ϵ is a very small positive number; (2) $f_{11} = tp$, $f_{10} = \epsilon$, $f_{01} = 1 - tp - \epsilon$, and $f_{00} = 0$, where ϵ is a very small positive number; (3) $f_{11} = tp$, $f_{10} = (1-tp)/2$, $f_{01} = (1-tp)/2$, and $f_{00} = 0$. ϕ reaches its lower bound $-\frac{1-tp}{1+tp}$ in Case 3. Relative Risk reaches its lower bound tp in Case 1. Odds Ratio reaches its lower bound 0 when $f_{00} = 0$. Conviction reaches its lower bound tp in Case 2. Added Value reaches its lower bound $-\frac{(1-tp)^2}{2 \cdot (1+tp)}$ in Case 3.



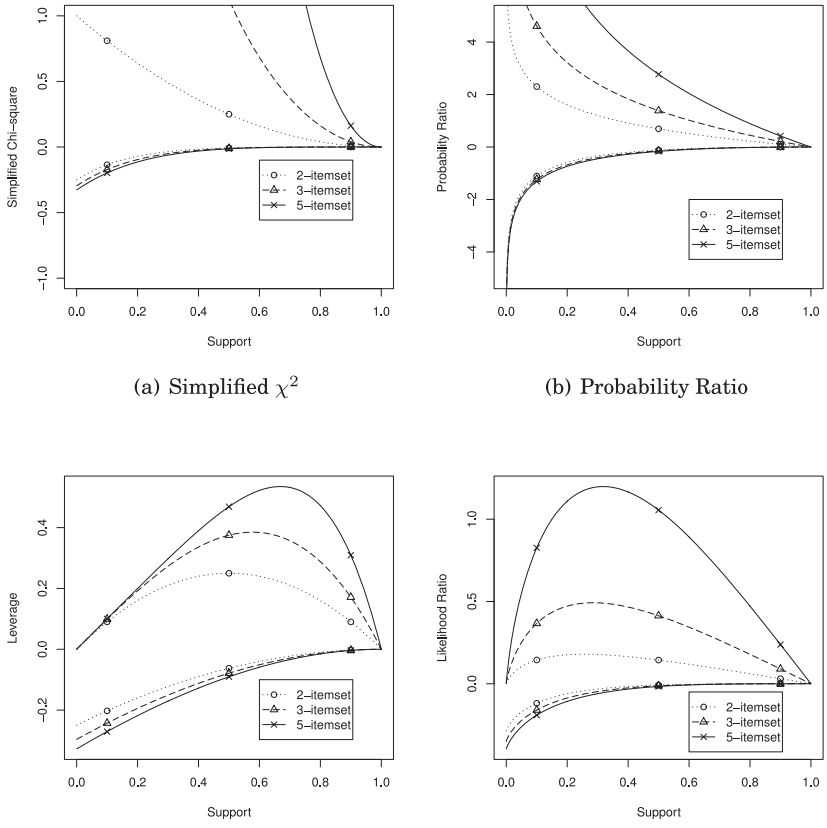
(a) Support (b) Any-confidence

(c) All-confidence (d) Bond

Fig. 1. Upper and lower bounds of subcategory 1.

2.3.1. Upper and Lower Bound Summary. Table IX shows the upper and lower bounds of all the measures given tp. Figures 1, 2, 3, 4, and 5 show the upper and lower bounds of the various measures with respect to different Support and itemset sizes. It is easy to see that different measures favor itemsets within different Support ranges.

Suboptimal measures. Since both the upper bound and lower bound of Support are Support itself, Support strictly favors a high Support itemset. In Figure 1, Any-confidence has the fixed upper bound 1, but the lower bound of Any-confidence increases with the increase of tp. Given an itemset with the fixed Support tp and the fixed itemset size m, we assume its Any-confidence follows a certain distribution between its upper bound 1 and the lower bound $m \cdot tp/(m - 1 + tp)$. The expected Any-confidence increases with the increase of tp when m is fixed, and the expected Any-confidence decreases with the increase of m when tp is fixed. Any-confidence favors high Support and small-size itemsets. Similar to Any-confidence, All-confidence and Bond favor high Support itemsets. Though the lower bounds of All-confidence and Bond have nothing to do with the itemset size m, Support favors small-size itemsets by its nature. Therefore, All-confidence and Bond favor small-size itemsets indirectly. In addition, the lower bounds of All-confidence and Bond are lower than that of Any-confidence. Therefore, All-confidence and Bond favor higher Support itemsets than Any-confidence.



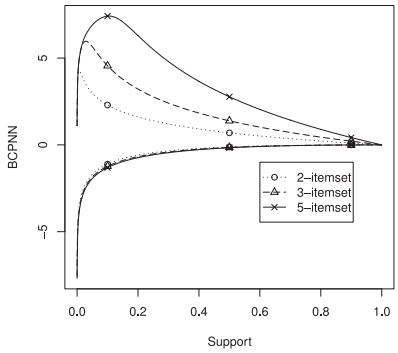
(a) Simplified χ^2 (b) Probability Ratio

(c) Leverage (d) Likelihood Ratio

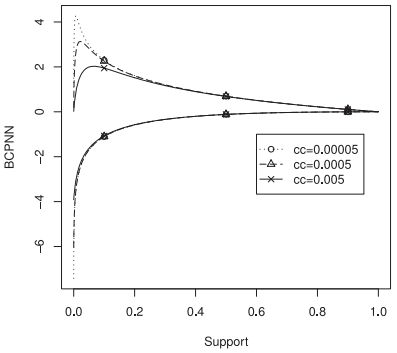
Fig. 2. Upper and lower bounds of subcategory 2.

Other general measures. The upper bounds of the Simplified χ^2-statistic and Probability Ratio increase to infinity when Support is close to 0, which means they favor coincidences rather than significant patterns. The only special situation is that the upper bound of the Simplified χ^2-statistic is equal to 1 instead of ∞ when the itemset size is 2. That explains why χ^2 works well for pairs but poorly for larger itemsets. As we increase the itemset size, the upper bounds of the Simplified χ^2-statistic and Probability Ratio get higher as Support approaches 0. Compared to the Simplified χ^2-statistic, Probability Ratio is more biased to low Support itemsets with large size.

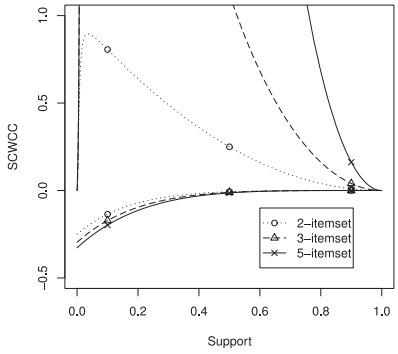
Leverage, Likelihood Ratio, BCPNN, Two-way Support, and SCWCC reach their highest upper bound when tp is between 0 and 1. For Leverage, the maximal value is reached when tp is between 0.5 and 0.8. In other words, the itemset with tp between 0.5 and 0.8 has a better chance to get a higher value. As we can see, different measures favor different tp regions. According to Figures 2 and 3, BCPNN favors lowest Support itemsets, followed by SCWCC, Likelihood Ratio, Two-way Support, and Leverage. The favored Support range of BCPNN and SCWCC can be adjusted by different continuity correction numbers according to Figure 3. Normally, we recommend $cc = 0.5/n$ for clean datasets. If the dataset is noisy, we might use a larger value to favor a relatively high Support region to suppress false-positive correlations from the noisy data in the low Support region, but the large value will also suppress true positive correlations



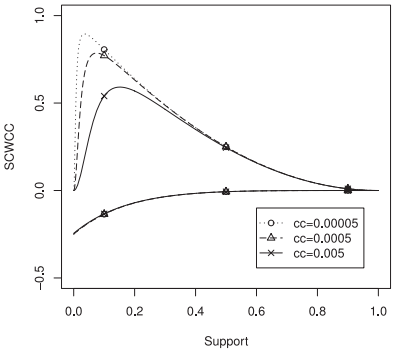
(a) BCPNN for different itemset size when $cc = 0.00005$



(b) BCPNN for different cc when itemset size is 2



(c) Simplified χ^2 with Continuity Correction for different itemset size when $cc = 0.00005$



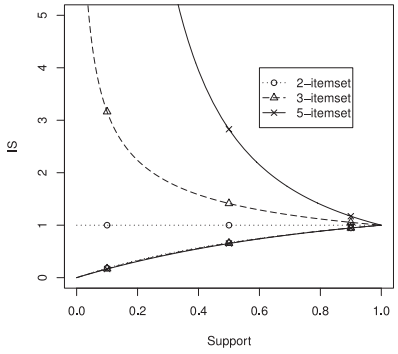
(d) Simplified χ^2 with Continuity Correction for different cc when itemset size is 2

Fig. 3. Upper and lower bounds of subcategory 3.

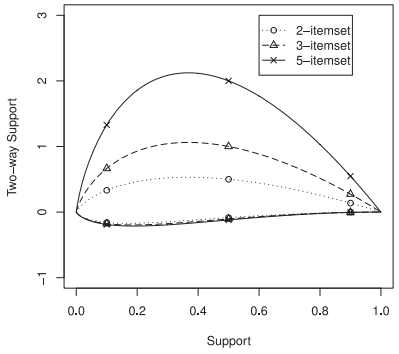
in the low Support region at the same time. Therefore, improper *cc* adjustment will degrade the effectiveness of BCPNN and SCWCC.

Tan et al. [2000] purposed IS by hoping the additional Support can suppress the increase of Probability Ratio when Support is close to 0. It works for 2-itemsets but fails for large-size itemsets. Better than IS, Two-way Support successfully decreases the upper bound when Support is close to 0. However, the lower bound trend is also reversed when the Support is close to 0, which is not what we want. We expect the highest negatively correlated itemsets to come from the low Support region. The Simplified χ^2 with Support has both the disadvantage of IS and the disadvantage of Two-way Support. It successfully suppresses the upper bound of 2-itemsets and 3-itemsets, but not for itemsets with size greater than 3. In addition, the trend of lower bound is reversed when the Support is close to 0.

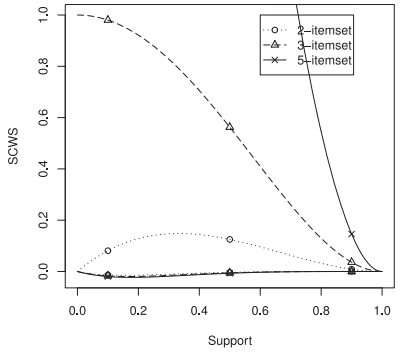
Pair-only measures. The upper bound of ϕ-coefficient, Relative Risk, Odds Ratio, and Conviction is a fixed number; they don't favor any region. The highest value can come from anywhere. The upper bound of Added Value decreases with the increase of Support, and it favors the low Support region.



(a) IS

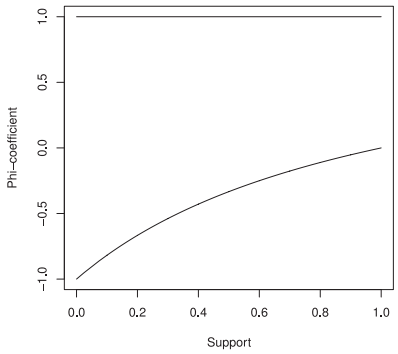


(b) Two-way Support

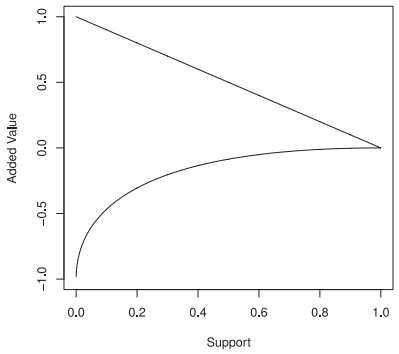


(c) Simplified χ^2 with Support

Fig. 4. Upper and lower bounds of subcategory 4.



(a) ϕ-coefficient



(b) Added Value

Fig. 5. Upper and lower bounds of pair-only measures.

3. EXPERIMENTS

There is no ground truth for us to compare with for real-life datasets as correlation search is an unsupervised learning procedure. Therefore, we will make use of the characteristics of different datasets to evaluate the performance of different measures. Although correlated pairs are a special case of correlated itemsets, we will check the performance on correlated pair search and correlated itemset search separately because some correlation functions can only measure pairs.

3.1. Experiments on Correlated Pair Search

3.1.1. OMOP Dataset. The Observational Medical Outcomes Partnership (OMOP[2]) is a public–private partnership to help monitor drug safety. For observational analysis, we want to find a correlation between drugs and conditions from a population. To facilitate the methodological research, it typically requires some "gold standard" to measure performance. However, the ground truth may not be absolutely determined because most observational data sources are poorly characterized, and clinical observations may be insufficiently recorded or poorly validated. Because of these issues, OMOP developed a simulation procedure to supplement the methods evaluation.

The simulated dataset has predefined associations between drugs and conditions. For each condition, each synthetic patient has a prevalent probability of having it. When the patient takes the related drugs for a certain condition, the probability of having it will increase. The dataset contains 10 million persons, more than 90 million drug exposures from 5,000 unique drugs, and more than 300 million condition occurrences from 4,500 unique conditions over a span of 10 years. In order to simulate reality, most drugs and conditions only happen a few times. Therefore, only 1/3 of the predefined associations are observed in the simulated dataset. In addition, among those predefined associations being observed, most of them only occur a small number of times.

A key step in the application of our correlation measure is the mapping of the data into drug–condition two-by-two tables, and then we can calculate the correlation between drugs and conditions. Among different ways of constructing two-by-two tables for longitudinal data (such as claims databases or electronic health records), we only use the "Modified SRS" method [OMOP 2010] to construct the two-by-two contingency table, which is the benchmark proposed by OMOP.

We check the bias and performance for each measure. First, we calculated the average Support of the top-K pairs retrieved by each measure. If the value is large, the measure favors frequent correlation patterns. Second, the mean average precision (MAP), a commonly used metric in the field of information retrieval, is used to evaluate each method. It measures how well a system ranks items and emphasizes ranking true positive items higher. Let y_{dc} be equal to 1 if the dth drug causes the cth condition, and 0 otherwise. Let $M = \sum_{d,c} y_{dc}$ denote the number of causal combinations and $N = D \times C$ the total number of combinations. Let z_{dc} denote the correlation value for the dth drug and the cth condition. For a given set of correlation values $\overrightarrow{z} = (z_{11}, \ldots, z_{DC})$, we define "precision-at-J" denoted $P^J(\overrightarrow{z})$ as the fraction of causal combinations among the J largest predicted values in $\overrightarrow{z}$. Specifically, let $z_1 > \cdots > z_N$ denote the ordered value of $\overrightarrow{z}$. Then, $P^J(\overrightarrow{z}) = \frac{1}{J}\sum_{i=1}^{J} y_i$, where y_i is the true status of combination corresponding to z_i. The MAP is calculated as $\frac{1}{M}\sum_{J=1}^{N}(P^J(\overrightarrow{z}) \cdot y_J)$. The MAP is very similar to the area under the precision-recall curve, which penalizes both types of misclassification: identifying a correlation when no relationship exists (false positive) and failing to identify true correlations (false negative). Table X shows the average Support of the top 1,000 pairs and the MAP for each measure.

[2]http://omop.fnih.org/.

Table X. Evaluation Result for OMOP Data

Type	Measures	Average Support of the Top 1000 Pairs	Mean Average Precision
Suboptimal measures	Support	55,849.62	0.0344
	Any-confidence	9,377.08	0.0925
	All-confidence	32,425.03	0.0668
	Bond	33,330.55	0.0694
Basic measures	Simplified χ^2-statistic	31,838.57	0.2258
	Probability Ratio	71.19	0.1102
	Leverage	44,955.40	0.1472
	Likelihood Ratio	42,298.72	0.2505
Adjusted measures	BCPNN	2,984.61	0.2440
	SCWCC	35,370.45	0.2415
	IS	32,531.93	0.0961
	Two-way Support	43,695.10	0.1876
	SCWS	41,345.57	0.1983
Pair-only measures	ϕ-coefficient	31,855.09	0.2256
	Relative Risk	89.62	0.1070
	Odds Ratio	6,928.90	0.0482
	Conviction	4,344.26	0.1020
	Added Value	4,344.00	0.1016

Since only 1/3 of the predefined associations are observed in the dataset, no method can achieve MAP beyond 0.33 unless it can infer unobserved drug–condition correlations. In Section 2.1, correlation properties are categorized into four groups: mandatory, desired, optional, and pair-only. Here, we study the effectiveness of these properties in terms of selecting good correlation measures. First, Support, Any-confidence, All-confidence, Bond, and IS violate some mandatory properties, and all their MAPs are below 0.1. If the correlation measures satisfy all the mandatory properties, all their MAPs are above 0.1 except Odds Ratio, which is frequently used for case-control studies and retrospective studies. Second, among all the measures satisfying all the three mandatory properties, if they satisfy two desired properties proposed in this article, their MAPs are generally better. In order to simulate reality, most drugs and conditions only happen a few times; therefore, the Support of most predefined associations is small. However, Leverage favors the high Support region, and that is why Leverage doesn't work well in this dataset. According to the average Support of the top 1,000 pairs, Leverage favors the highest Support pairs, followed by Two-way Support, Likelihood Ratio, SCWCC, and BCPNN, which is consistent with Figures 2, 3, and 4. Here, we are measuring the performance of pair search. If we only consider the upper bound for pairs, SCWS and Simplified χ^2-statistic also satisfy two desired properties and their MAPs are good. This also explains why statisticians usually use χ^2-statistic for pairs but doubt the performance of χ^2-statistic for large-size itemsets. When searching for correlated pairs, satisfying Property 5 is unnecessary since it regulates the pattern for itemsets. Therefore, the performance of ϕ-coefficient is good as long as it satisfies Property 4. The upper bound of Probability Ratio increases to infinity when $P(S)$ is close to 0. It favors coincidences rather than significant correlations in the data, which is verified by its small average Support of the top 1,000 pairs. Third, Simplified χ^2-statistic, Likelihood Ratio, and SCWCC satisfy all the optional properties. Here, we are not interested in negative correlations and search only for positive correlations in one centered dataset. Therefore, the optional properties won't help us to identify good correlation measures in this experiment. Fourth, satisfying the first and third additional properties cannot help us to identity good correlation measures. In addition,

Table XI. Evaluation Result for Facebook Data

Measures	Mean Average Precision	Mean Personal MAP
Support	0.4415	0.7084
Any-confidence	0.3657	0.6106
All-confidence	0.4865	0.5920
Bond	0.5062	0.6302
Simplified χ^2-statistic	0.5029	0.6563
Probability Ratio	0.1800	0.4312
Leverage	0.4579	0.7278
Likelihood Ratio	0.5287	0.7363
BCPNN	0.5168	0.7342
SCWCC	0.4970	0.7327
IS	0.5067	0.6582
Two-way Support	0.5177	0.7360
SCWS	0.5540	0.7104
ϕ-coefficient	0.5033	0.6564
Relative Risk	0.1275	0.1977
Odds Ratio	0.1609	0.3278
Conviction	0.2420	0.5874
Added Value	0.3224	0.7278

only bad correlation measures satisfy the second additional properties. Therefore, we doubt that the three additional properties qualify as desired properties.

3.1.2. Facebook Dataset. We crawled Facebook for the University of Iowa community. The resulting dataset contains 41,073 people and their friend information within this community. If we consider each friend list as a transaction and people as items in the transaction, we can calculate the correlation between any two people in this community and get a ranking list of how strongly people are correlated with each other. However, we don't have the ground truth of whether two given people are correlated or not in this real-life dataset. Therefore, we can only naively assume two given people are correlated with each other if they are friends of each other. Since the ground truth is not perfect, the experimental result is only complementary to the OMOP result. By using the friend relationships within this community, we can calculate the MAP to evaluate the friend ranking list as we do in the OMOP dataset. Another interesting question related to this dataset is how other people are correlated to a particular person. The ranking list of this type is useful for friend recommendations in Facebook. Similarly, we can calculate the MAP for each ranking list related to a particular person and then average all the personal MAP values. All the evaluation values are shown in Table XI.

Surprisingly, suboptimal measures work pretty well in this application. The result supports why the Facebook friend recommender system recommends the person with the most common friends. The reason that suboptimal measures work well in this application is as follows. If two people know 90% of the people in this community and have a lot of common friends, the chance for them to know each other is high. Knowing each other doesn't mean a high correlation with each other. We use friendship as an indicator for correlation because there is no better indicator for this dataset. Although friendship biases to Support, measures like Simplified χ^2-statistic, Leverage, Likelihood Ratio, BCPNN, SCWCC, Two-way Support, and SCWS, which satisfy all the mandatory and relaxed desired properties, can still do slightly better than Support. Another interesting measure that works well for mean personal MAP in this application is Added Value. The formula, $P(B|A) - P(B)$, indicates that B can achieve a high score if

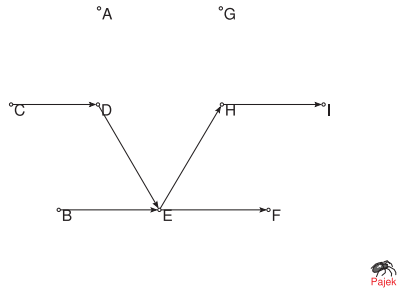


Fig. 6. An example of simulated datasets with nine items.

B knows most of the people A knows. In social activity, B usually has a tight connection with A first in order to know most of the people A knows.

3.2. Experiments on Correlated Itemset Search

3.2.1. Simulated Dataset. Since the exact ground truth for real-life datasets is impossible to get, we will use simulated datasets to test the performance of different measures for correlated itemset search. Our simulation procedure is as follows:

(1) A set of occurrence probabilities p_i for each item following a power law distribution is generated.
(2) A set of cause–effect pairs is randomly selected one by one from all the possible pairs, and each cause–effect pair is associated with a randomly assigned probability. If the current pair generates a chain cycle with previous pairs, we will discard the current pair instead of selecting it as a cause–effect pair. For example, if $A \to B$ and $B \to C$ have been selected previously and we currently select $C \to A$, the pair $C \to A$ will be discarded. We don't allow a circle because it might magnify correlation.
(3) In the beginning, each item is generated independently according to its occurrence probability p_i for each transaction.
(4) For each cause–effect pair $I_{cause} \to I_{effect}$, we search for the transactions in which I_{cause} occurs but I_{effect} does not. For each of the related transactions, the status of I_{effect} might be changed to occurrence with the probability associated with this cause–effect pair in Step 2.

Figure 6 shows an example of a simulated dataset with nine items. It has six cause–effect pairs: $B \to E$, $C \to D$, $D \to E$, $E \to F$, $E \to H$, and $H \to I$. For the two pairs $I_{cause1} \to I_{effect1}$ and $I_{cause2} \to I_{effect2}$, we can connect them to form a chain $I_{cause1} \to I_{effect1} \to I_{effect2}$ if $I_{effect1}$ and I_{cause2} are the same item. If a chain $I_1 \to I_2 \to \cdots \to I_m$ cannot be expanded by adding any item in the beginning or the end according to the cause–effect pairs, we call it a maximal chain. In this example, there are four maximal chains: $B \to E \to F$, $B \to E \to H \to I$, $C \to D \to E \to F$, and $C \to D \to E \to H \to I$. For any itemset S that is a subset of any maximal chain, S is a correlated itemset according to the way we simulate. For example, $\{C, D, H\}$ is a subset of $C \to D \to E \to H \to I$. It is a correlated itemset in this example.

We use this simulation procedure to simulate datasets with 100 items and 100,000 transactions. In the simulated dataset, 50 pairs are randomly selected as cause–effect pairs. The simulated dataset has $2^{100} - 1 - 100$ itemsets that contain no less than two items, and we can calculate their correlation and rank them. Since the dataset is simulated by us, we know all the correlated itemsets in it. Suppose the number of correlated itemsets is k; we can check what percentage of the top-k itemsets in the

Table XII. Evaluation Result for Simulated Data

Measures	Average Support of the Top-k Itemsets	Average Size of the Top-k Itemsets	Precision at k
Support	3,860.66	2.33	0.2131
Simplified χ^2-statistic	1.00	14.56	0.0000
Probability Ratio	1.00	14.56	0.0000
IS	1.00	14.56	0.0000
SCWS	1.00	14.56	0.0000
Leverage	2,663.55	2.84	0.3572
Two-way Support	2,276.69	3.25	0.3311
Likelihood Ratio	1,773.80	3.42	0.3074
SCWCC	725.39	4.02	0.1914
BCPNN	369.25	4.56	0.0810

ranking list for each correlation measure correspond to actual correlated itemsets. This percentage is called precision at k. If one correlation measure can perfectly rank all the correlated itemsets in the top-k list, it achieves the perfect score of 1. In addition, we check the average occurrence and size of the top-k itemsets in the ranking list for each correlation measure in order to check their bias on the Support region and the itemset size. Table XII shows the experimental result. First, Simplified χ^2-statistic, Probability Ratio, IS, and SCWS, which violate the desirable Property 4, all get the precision at k of 0 and the top-k itemsets that all are large size and occur once only. According to their upper bound graph, their upper bound increases when tp is close to 0. Their upper bound also increases with the size of the itemset. Both experimental results and theoretical analysis indicate that Simplified χ^2-statistic, Probability Ratio, IS, and SCWS favor the large-size itemsets that rarely occur. Second, the precision at k of the correlation measures that satisfy all the three mandatory and the two desirable properties are generally good. However, the performance of each measure in this dataset is totally different from that in the OMOP dataset. For example, Leverage achieves the best score among Leverage, Two-way Support, Likelihood Ratio, SCWCC, and BCPNN in this dataset, but the worst score among the five measures in the OMOP dataset. It is because many correlated itemsets frequently occur in this dataset, while most correlated pairs in OMOP occur few times. Therefore, we need to understand the dataset in order to select the best measure for it. According to the average Support of the top 1,000 pairs, Leverage favors the highest Support itemsets, followed by Two-way Support, Likelihood Ratio, SCWCC, and BCPNN, which is consistent with our upper bound analysis.

3.2.2. Netflix Dataset. Here, we will make use of the characteristics of the Netflix dataset[3] to evaluate the effectiveness of different correlated itemset search methods. Since the Netflix dataset contains 17,770 movies and has around 480,000 transactions, it is impossible to find the top-k correlated itemsets due to the computational cost. Therefore, we create a subset of Netflix that only contains the first 100 movies (according to the code sequence in the Netflix dataset) and use the brute-force search method to find the top-k correlated itemsets in this subset.

We show the top-3 correlated itemsets of typical measures and the Support for each itemset in Table XIII. First, all three patterns retrieved by Support contain the movie *Something's Gotta Give*. It is the most popular movie in the subset. Considering that the probability of liking *Dragonheart* conditioned on liking *Something's Gotta Give* is lower than the probability of liking *Dragonheart*, Support is a bad measure for correlation

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

Table XIII. Top-3 Correlated Itemsets for Netflix Data

Measures	Top-3 correlated itemsets	Support
Support	*Lilo and Stitch* (2002), *Something's Gotta Give* (2003)	9,574
	Something's Gotta Give (2003), *Silkwood* (1983)	5,248
	Something's Gotta Give (2003), *Dragonheart* (1996)	3,736
Simplified χ^2-statistic, Probability Ratio, IS, and SCWS	A set with 18 movies	1
	A set with 17 movies	1
	A set with 17 movies	1
Leverage	*Lilo and Stitch* (2002), *Dragonheart* (1996)	3,108
	Dragonheart (1996), *Congo* (1995)	1,091
	Spitfire Grill (1996), *Silkwood* (1983)	1,207
Two-way Support	*Lilo and Stitch* (2002), *Dragonheart* (1996)	3,108
	Dragonheart (1996), *Congo* (1995)	1,091
	Spitfire Grill (1996), *Silkwood* (1983)	1,207
Likelihood Ratio	*Dragonheart* (1996), *Congo* (1995)	1,091
	Lilo and Stitch (2002), *Dragonheart* (1996), *Congo* (1995)	501
	The Rise and Fall of ECW (2004), *WWE: Royal Rumble* (2005)	153
SCWCC	*My Favorite Brunette* (1947), *The Lemon Drop Kid* (1951)	103
	The Rise and Fall of ECW (2004), *WWE: Royal Rumble* (2005)	153
	Screamers (1996), *Dragonheart* (1996), *Congo* (1995)	120
BCPNN	*My Favorite Brunette* (1947), *The Lemon Drop Kid* (1951)	103
	The Rise and Fall of ECW (2004), *WWE: Royal Rumble* (2005), *ECW: Cyberslam '99* (2002)	41
	WWE: Armageddon (2003), *WWE: Royal Rumble* (2005)	47

search. Second, Simplified χ^2-statistic, Probability Ratio, IS, and SCWS violate the desired Property 4, and they retrieve the same three long patterns that only happen once. The upper bounds of these four measures become steeper when the itemset size increases from 2 to 3 to 5, and their upper bounds increase to infinity when Support is close to 0. In other words, they favor rare itemsets with large size. The upper-bound graphs are consistent with the experimental result. If the measure violates the desired properties proposed by us, the performance might be good for pair search, but they are all bad for itemset search. Third, for the measures satisfying all the mandatory and desired properties, Leverage favors frequent correlation patterns followed by Two-way Support, Likelihood Ratio, SCWCC, and BCPNN according to the Support of retrieved itemsets.

In the Netflix data, we can assume movies in the same series are strongly correlated. Since this subset only contains a few movies in the same series, most retrieved patterns are not movies in the same series, which is hard to justify. It would be better if we could find the top-k patterns in the whole Netflix dataset. Therefore, we make use of the maximal fully correlated itemset framework [Duan and Street 2009] to find the top-5 patterns in the whole Netflix dataset for four typical measures in Table XIV. Only one of the five patterns retrieved by Support contains movies in the same series. All the five patterns retrieved by Leverage, Likelihood Ratio, and BCPNN are movies in the same series. Leverage and Likelihood Ratio find popular movie series, while BCPNN retrieves unpopular movie series, which is consistent with our correlation analysis.

4. CONCLUSION

In this article, we did a comprehensive study on effective correlation search for binary data. First, we studied 18 different correlation measures and proposed two desirable

Table XIV. Top 5 Patterns for Netflix Data

Measure	Maximal Fully Correlated Itemsets
Support	*The Lord of the Rings: The Fellowship of the Ring* (2001), *The Lord of the Rings: The Two Towers* (2002), *The Lord of the Rings: The Return of the King* (2003)
	Forrest Gump (1994), *The Green Mile* (1999)
	The Lord of the Rings: The Two Towers (2002), *Pirates of the Caribbean: The Curse of the Black Pearl* (2003)
	The Lord of the Rings: The Fellowship of the Ring (2001), *Pirates of the Caribbean: The Curse of the Black Pearl* (2003)
	Forrest Gump (1994), *The Shawshank Redemption: Special Edition* (1994)
Leverage	*The Lord of the Rings: The Fellowship of the Ring* (2001), *The Lord of the Rings: The Two Towers* (2002), *The Lord of the Rings: The Return of the King* (2003)
	Raiders of the Lost Ark (1981), *Indiana Jones and the Last Crusade* (1989)
	Star Wars: Episode V: The Empire Strikes Back (1980), *Star Wars: Episode VI: Return of the Jedi* (1983)
	The Lord of the Rings: The Fellowship of the Ring: Extended Edition (2001), *The Lord of the Rings: The Two Towers: Extended Edition* (2002)
	Star Wars: Episode IV: A New Hope (1977), *Star Wars: Episode V: The Empire Strikes Back* (1980)
Likelihood Ratio	*The Lord of the Rings: The Fellowship of the Ring: Extended Edition* (2001), *The Lord of the Rings: The Two Towers: Extended Edition* (2002), *The Lord of the Rings: The Return of the King: Extended Edition* (2003)
	Star Wars: Episode IV: A New Hope (1977), *Star Wars: Episode V: The Empire Strikes Back* (1980), *Star Wars: Episode VI: Return of the Jedi* (1983)
	The Lord of the Rings: The Fellowship of the Ring (2001), *The Lord of the Rings: The Two Towers* (2002), *The Lord of the Rings: The Return of the King* (2003)
	Harry Potter and the Sorcerer's Stone (2001), *Harry Potter and the Chamber of Secrets* (2002)
	Kill Bill: Vol. 1 (2003), *Kill Bill: Vol. 2* (2004)
BCPNN	*Roughnecks: The Starship Troopers Chronicles: The Homefront Campaign* (2000), *Roughnecks: The Starship Troopers Chronicles: The Klendathu Campaign* (2000)
	Now and Then, Here and There: Vol. 1: Discord and Doom (1999), *Now and Then, Here and There: Vol. 2: Flight and Fall* (2002), *Now and Then, Here and There: Vol. 3: Conflict and Chaos* (1999)
	Dragon Ball: King Piccolo Saga: Part 1 (1986), *Dragon Ball: King Piccolo Saga: Part 2* (1986) *Dragon Ball: Piccolo Jr. Saga: Part 1* (1995), *Dragon Ball: Piccolo Jr. Saga: Part 2* (1995)
	Dragon Ball: Red Ribbon Army Saga (2002), *Dragon Ball: Commander Red Saga* (2002)
	Dragon Ball: Piccolo Jr. Saga: Part 1 (1995), *Dragon Ball: Piccolo Jr. Saga: Part 2* (1995) *Dragon Ball: Red Ribbon Army Saga* (2002)

properties to help us select good correlation measures. The experiments with both simulated and real-life datasets show that the two desirable properties are very successful for selecting good correlation measures to tell correlated patterns from uncorrelated patterns. Second, we studied different techniques to adjust original correlation measures and used them to propose two new correlation measures: the Simplified χ^2 with Continuity Correction and the Simplified χ^2 with Support. Third, we studied the upper and lower bounds of different measures to find their different favorable Support region. Although the region favor itself is not a good or bad property, the user can choose the measure that favors his or her preference. Last, we made use of the characteristics of different datasets to validate our conclusions. In Section 1, we mentioned that different correlation measures are favored in different domains by studying the literature related to correlation, and it can be explained by our correlation analysis now. First of all, all of them satisfy our two desirable properties. In text mining, too frequent or rare words don't have too much discriminative power for classification. Therefore, they favor Likelihood Ratio to find the pattern that is not too rare or too frequent [Dunning 1993]. In medical domains, the associations between frequent diseases and frequent symptoms have already been observed in the clinical trials. The big issue is how to find the correlation between rare diseases and rare symptoms. That is why they use BCPNN [Bate et al. 1998]. In social networks, people are more interested in the patterns affecting a larger population. Therefore, they favor Leverage [Clauset et al. 2004; Duan et al. 2014]. In all, we recommend Leverage for searching obvious patterns in the dataset that we know nothing about; Likelihood Ratio, Simplified χ^2 with Continuity Correction, and Two-way Support for searching typical patterns in the dataset that we know something of; and BCPNN for searching rare patterns in the dataset that we know well.

ACKNOWLEDGMENTS

The authors would like to thank Jieqiu Chen and Qiwei Sheng of the University of Iowa for helping with the math proof of several correlation properties.

REFERENCES

R. Agrawal, T. Imieliński, and A. Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'93)*. ACM, New York, NY, 207–216. DOI:http://dx.doi.org/10.1145/170035.170072

A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas. 1998. A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology* 54, 4 (1998), 315–321.

S. Brin, R. Motwani, and C. Silverstein. 1997a. Beyond market baskets: Generalizing association rules to correlations. In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'97)*. ACM, New York, NY, 265–276. DOI:http://dx.doi.org/10.1145/253260.253327

S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. 1997b. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'97)*. ACM, New York, NY, 255–264. DOI:http://dx.doi.org/10.1145/253260.253325

A. Clauset, M. E. J. Newman, and C. Moore. 2004. Finding community structure in very large networks. *Physical Review E* 70, 6 (Dec. 2004), 066111+. DOI:http://dx.doi.org/10.1103/PhysRevE.70.066111

L. Duan and W. Nick Street. 2009. Finding maximal fully-correlated itemsets in large databases. In *Proceedings of the International Conference on Data Mining (ICDM'09)*. 770–775.

L. Duan, W. Nick Street, and Y. Liu. 2013. Speeding up correlation search for binary data. *Pattern Recognition Letters* 34, 13 (2013), 1499–1507. DOI:http://dx.doi.org/10.1016/j.patrec.2013.05.027

L. Duan, W. Nick Street, Y. Liu, and H. Lu. 2014. Community detection in graphs through correlation. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (ACM SIGKDD'14)*.

W. Dumouchel. 1999. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *American Statistician* 53, 3 (1999), 177–202.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 1 (1993), 61–74.

H. Everett. 1957. 'Relative State' formulation of quantum mechanics. *Reviews of Modern Physics* 29 (1957), 454–462.

L. Geng and H. J. Hamilton. 2006. Interestingness measures for data mining: A survey. *Computing Surveys* 38, 3 (2006), 9. DOI:http://dx.doi.org/10.1145/1132960.1132963

C. Jermaine. 2005. Finding the most interesting correlations in a database: How hard can it be? *Information Systems* 30, 1 (2005), 21–46. DOI:http://dx.doi.org/10.1016/j.is.2003.08.004

R. H. Johnson and D. W. Wichern. 2001. *Applied Multivariate Statistical Analysis*. Prentice Hall.

M. Liu, E. R. M. Hinz, M. E. Matheny, J. C. Denny, J. S. Schildcrout, R. A. Miller, and H. Xu. 2013. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *Journal of the American Medical Informatics Association* 20, 3 (2013), 420–426. DOI:http://dx.doi.org/10.1136/amiajnl-2012-001119

F. Mosteller. 1968. Association and estimation in contingency tables. *Journal of the American Statistical Association* 63, 321 (1968), 1–28.

P.-N. Tan and V. Kumar. 2000. Interestingness measures for association patterns: A perspective. In *Proceedings of the KDD 2000 Workshop on Postprocessing in Machine Learning and Data Mining*.

G. N. Norén, A. Bate, J. Hopstadius, K. Star, and I. R. Edwards. 2008. Temporal pattern discovery for trends and transient effects: Its application to patient records. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, NY, 963–971. DOI:http://dx.doi.org/10.1145/1401890.1402005

E. R. Omiecinski. 2003. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering* 15, 1 (2003), 57–69. DOI:http://dx.doi.org/10.1109/TKDE.2003.1161582

OMOP. 2010. Methods section for the disproportionality paper. (September 2010). `http://omop.fnih.org/MethodsLibrary`.

G. Piatetsky-Shapiro. 1991. *Discovery, Analysis, and Presentation of Strong Rules*. AAAI/MIT Press, 229–248.

H. T. Reynold. 1977. *The Analysis of Cross-Classifications*. Free Press.

C. L. Sistrom and C. W. Garvan. 2004. Proportions, odds, and risk. *Radiology* 230, 1 (2004), 12–19.

P.-N. Tan, V. Kumar, and J. Srivastava. 2004. Selecting the right objective measure for association analysis. *Information Systems* 29, 4 (2004), 293–313. DOI:http://dx.doi.org/10.1016/S0306-4379(03)00072-3

P.-N. Tan, M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining*. Addison Wesley.

C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton. 2013. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery* (2013), 1–42. DOI:http://dx.doi.org/10.1007/s10618-013-0326-x

H. Xiong, M. Brodie, and S. Ma. 2006a. TOP-COP: Mining top-K strongly correlated pairs in large databases. In *Proceedings of the International Conference on Data Mining (ICDM'06)*. Washington, DC, 1162–1166. DOI:http://dx.doi.org/10.1109/ICDM.2006.161

H. Xiong, S. Shekhar, P.-N. Tan, and V. Kumar. 2006b. TAPER: A two-step approach for all-strong-pairs correlation query in large databases. *IEEE Transactions on Knowledge and Data Engineering* 18, 4 (2006), 493–508. DOI:http://dx.doi.org/10.1109/TKDE.2006.68

J. Zhang and J. Feigenbaum. 2006. Finding highly correlated pairs efficiently with powerful pruning. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management (CIKM'06)*. ACM, New York, NY, 152–161. DOI:http://dx.doi.org/10.1145/1183614.1183640

N. Zhong, C. Liu, and S. Ohsuga. 2001. Dynamically organizing KDD processes. *International Journal of Pattern Recognition and Artificial Intelligence* 15, 3 (2001), 451–473.

N. Zhong, Y. Y. Yao, and S. Ohsuga. 1999. Peculiarity oriented multi-database mining. In *Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'99)*. Springer-Verlag, London, UK, 136–146.

Received August 2013; revised January 2014; accepted June 2014