Publication IV

Markus Ojala, Gemma C. Garriga, Aristides Gionis, and Heikki Mannila. 2010. Evaluating query result significance in databases via randomizations. In: Proceedings of the 10th SIAM International Conference on Data Mining (SDM 2010). Columbus, Ohio, USA. 29 April - 1 May 2010. Society for Industrial and Applied Mathematics. Pages 906-917.

Evaluating Query Result Significance in Databases via Randomizations

Markus Ojala* Gemma C. Garriga* Aristides Gionis† Heikki Mannila*

Abstract

Many sorts of structured data are commonly stored in a multi-relational format of interrelated tables. Under this relational model, exploratory data analysis can be done by using relational queries. As an example, in the Internet Movie Database (IMDb) a query can be used to check whether the average rank of action movies is higher than the average rank of drama movies.

We consider the problem of assessing whether the results returned by such a query are statistically significant or just a random artifact of the structure in the data. Our approach is based on randomizing the tables occurring in the queries and repeating the original query on the randomized tables. It turns out that there is no unique way of randomizing in multi-relational data. We propose several randomization techniques, study their properties, and show how to find out which queries or hypotheses about our data result in statistically significant information and which tables in the database convey most of the structure in the query. We give results on real and generated data and show how the significance of some queries vary between different randomizations.

1 Introduction

The question of evaluating whether certain hypotheses made from observed data are significant or not, is one of the oldest problems in statistics. Statistical significance reduces an observed result (statistic) to a p-value that tells about the probability of observing the same result at random when a certain null hypothesis is true. If this p-value is sufficiently small, we can assume that the null hypothesis is false. The technical challenge of defining an exact p-value for a given hypothesis is typically resolved by studying the null distribution of the test analytically; for example, the well known chi-squared test is based on statistics that follow a chi-square distribution under the null hypothesis. Alternatively, when analytical solutions are not possible or hard to state exactly, the null distribution can be defined via permutation tests.

These useful statistical concepts have been used for years in experimental fields such as medicine, biology, geology or physics, to name a few. Many of these considerations have been extended as well to the data mining and database community. In a very first paper about association rules, Brin et al. [18] considered measuring the significance of rules via the chi-squared test, and from there many other papers followed—see e.g. [19] for a comprehensive survey. More recently, the approach of defining randomization tests to assess data mining results was introduced for binary data [10], and for real-valued data [15].

$$\begin{aligned}
\text{GM} &= \{(\text{Romance}, m_1), (\text{Romance}, m_2), (\text{Drama}, m_3),\\
&\quad (\text{Drama}, m_4), (\text{Drama}, m_5), (\text{Drama}, m_6),\\
&\quad (\text{Drama}, m_7), (\text{History}, m_6), (\text{History}, m_7)\}\\
\text{MD} &= \{(m_1, \text{C. Waitt}), (m_2, \text{C. Waitt}), (m_3, \text{C. Waitt}),\\
&\quad (m_4, \text{C. Waitt}), (m_5, \text{C. Waitt}), (m_6, \text{T. George}),\\
&\quad (m_7, \text{T. George})\}\\
\text{DA} &= \{(\text{C. Waitt}, 30), (\text{T. George}, 60)\}
\end{aligned}$$

Figure 1: A toy example of a multi-relational database with three binary relations: movies classified by genres, GM; movies directed by directors, MD; and ages of directors, DA.

Abstracting a bit from the question of how significant patterns are in the data, we introduce here the statistical testing framework to databases and the exploratory task of querying the relations of the database. The question of understanding what we know and what we believe about our dataset becomes tricky when the data is highly structured and interrelated. Structured data is everywhere: examples are the Internet Movie Database (IMDb), or the DBLP computer science bibliography, and indeed, most of today's information systems are actually relational databases. In IMDb, for example, basic entities are directors, movies, genres, ranks or years; in addition, we have relations such as directors direct movies, movies are classified by a genre, movies are ranked with some quality criteria, and directors are born in a certain year. Each of these relations is represented in a separate table which relates to others through their common attribute values. A simple toy example is given in Figure 1.

In multi-relational databases, users and applications access the data via queries. E.g., a query can be made to check the average age of directors of history movies, or the average age of directors of romance movies. In the toy exam-

*HIIT, Helsinki University of Technology, Finland
†Yahoo! Research, Barcelona, Spain

ple of Figure 1, the first query returns a value of 60, while the second query returns a value of 30. Usually, the answer returned by the query is assumed as a fact, thus implying some conventional wisdom—for this toy example we might be tempted to believe that the directors of romance movies are younger than the directors of history movies. But, should we really believe that this hypothesis is significant from the data? If we knew that all history movies are also classified as drama movies, would the value of 60 still have the same importance? Or, if we knew that the same director has participated in both romance and drama movies?

We study whether the results returned by queries are significant or just a random artifact due to the structure in the data. Our statistical tool is randomizations and the approach is simple: randomize certain relations occurring in the queries and repeat the original query in the random samples. This provides an empirical p-value, and, as in basic statistics, we can reject or accept our hypothesis linked to the query. The goal behind this idea is to provide an understanding of how the structure of the data affects the significance of the information we derive from our queries. If certain structures or patterns remain after simple randomizations (e.g., the fact that history movies are also drama movies in the toy example), the answers of a query that rely on such patterns should be regarded as not significant.

It turns out that there is no unique way of randomizing in multi-relational data. We study several randomization methods and show the combinatorial properties of the null distributions on multiple tables. Each randomization method tests a different property of the original data, thus giving unique information of the result of the query. Our contribution makes a first step towards understanding how the significance of a query is linked to the structure hidden in the data; randomizations are a sound statistical tool to make such a connection. We believe this is an important problem of interest to both the database and data mining communities. We present experimental results on synthetic data, and show the usability of the method for several queries in real datasets.

2 Problem Statement

Let A be a binary relation between sets I and J, $A \subseteq I \times J$. In the market basket application, for example, I could be a set of customers and J a set of products. A binary relation $A \subseteq I \times J$ identifies which customers from I buy which products from J. We denote with $(i, j) \in A$ a pair $i \in I$ and $j \in J$ belonging to A. Notice that every binary relation can be seen as a binary matrix describing the occurrences between the row set I and column set J. Examples of such representations are given in Figure 1 and Figure 2.

Let $\{A_1, \ldots, A_n\}$ be a set of n binary relations representing some structured data. This relational model is very general. It applies, for example, to a movie database system, as shown in Figure 2. The representation of the same example as a sequence of bipartite graphs is depicted in Figure 3.

	m_1	m_2	m_3	m_4	m_5	m_6	m_7
Romance	1	1	0	0	0	0	0
Drama	0	0	1	1	1	1	1
History	0	0	0	0	0	1	1

(a) Genre $\times$ Movie

	C. Waitt	T. George
m_1	1	0
m_2	1	0
m_3	1	0
m_4	1	0
m_5	1	0
m_6	0	1
m_7	0	1

(b) Movie $\times$ Director

	30	60
C. Waitt	1	0
T. George	0	1

(c) Director $\times$ Age

Figure 2: The binary table representation of the toy database in Figure 1: (a) GM; (b) MD; and (c) DA.

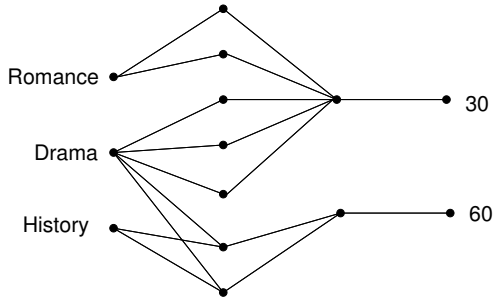


Figure 3: The bipartite graph representation of the movie database shown in Figure 2. The graph shows all the possible paths from the source nodes, Genre, to the destination nodes, Age.

The basic operator to combine relations is the *natural join*. Conceptually, a join between two relations A and B, denoted $A \bowtie B$, combines all entries from A and B that share common attribute values to return a composition of the relations. For example, given $(i, j) \in A$, $(j, k) \in B$ and $(j, k') \in B$, we have $(i, j, k) \in A \bowtie B$ and also $(i, j, k') \in A \bowtie B$. The join operator is associative over a set of relations and its result explicitly represents all existing paths between the occurring relations. For example, the natural join of the three tables in Figure 2 returns a tuple for each path there is between Genre and Age. For an *ordered* subset of binary relations from the database $S \subseteq \{A_1, \ldots, A_n\}$, we use $\bowtie S$ to denote the final join between all elements in S. The order in S is to ensure a join of consistent relations; we assume that S in $\bowtie S$ is always implicitly ordered.

A *query* q is applied to the join of a subset of the

relations in the database $S \subseteq \{A_1, \ldots, A_n\}$. The result of a query is denoted by $q(\bowtie S)$. We say that S is the set of relations occurring in the query. A query can be described with the operators of projection and selection [16], applied to a join $\bowtie S$. Projection is a unary operator $\pi_X(\bowtie S)$ that restricts tuples of $\bowtie S$ to attributes in X. Selection is a unary operator $\sigma_\varphi(\bowtie S)$ where φ is a propositional formula. The operator selects all tuples in the relation $\bowtie S$ for which φ holds.

Consider the movie database in Figure 2. A possible query is: select drama movies and project movie and age of its director. We can write this query as follows,

$$q_1 = \pi_{\text{Movie,Age}}(\sigma_{\text{Genre = Drama}}(\text{GM} \bowtie \text{MD} \bowtie \text{DA}))$$

The result of query q_1 is a set of pairs: $\{(m_3, 30), (m_4, 30), (m_5, 30), (m_6, 60), (m_7, 60)\}$. Another very similar query is: select drama movies and project age only. That is,

$$q_2 = \pi_{\text{Age}}(\sigma_{\text{Genre = Drama}}(\text{GM} \bowtie \text{MD} \bowtie \text{DA}))$$

Query q_2 returns:$\{30, 60\}$. Although queries q_1 and q_2 are very similar, the projection made by q_2 only on Age, has eliminated repeated values. The result of query q_1 tells us how many paths there are between directors of Drama and Age, while in query q_2 we only know if a path exists or not.

Our goal is to assess whether the results returned by a query provide significant information about our hypothesis on the data. For simplicity, a *statistic* f is required to map the results of a query to a single real value. We assume this function f is provided by the user together with the query (or a set of queries); they define the hypothesis on the data the user wants to test. Examples of this statistic are the average of the returned results, or the number of tuples in the answer, but indeed f can be any general function returning a real value.

For example, the average value of Age in query q_1 is 42.5 (i.e., the average age of directors of Drama weighted by the number of directed movies). Then, we may want to know whether that average age is interesting or not. Another two-tailed hypothesis is whether that average is significantly different from the average age of directors of romance movies.

Formally, our problem reads as follows.

PROBLEM 1. *Given a set of multiple binary relations* $\{A_1, \ldots, A_n\}$ *corresponding to some structured data and a query* q *on some occurring* $S \subseteq \{A_1, \ldots, A_n\}$*, is the value of* $f(q(\bowtie S))$ *for a statistic* f*, significant (in some sense to be made more specific later)?*

3 Overview of the Approach

In this section, we present an overview of the approach and describe the intuition behind it. We show how our method can be used to test the significance of queries and to uncover the structurally important relations in the data.

3.1 Significance Testing via Randomizations We approach the problem of testing the statistical significance of the results of the query via randomizations. The general idea under significance testing is to evaluate a null hypothesis against an alternative hypothesis: the alternative hypothesis relates to observations derived from the data, while the null hypothesis assumes that observations come from a random distribution without any structure. Randomizations have been widely used as a method to generate samples from null distributions. For example, in medical studies it is customary to measure the effect of a certain drug via permutation tests between the control group and the case group [11].

For short, let $R = \bowtie S$ for some $S \subseteq \{A_1, \ldots, A_n\}$. To assess the significance of $f(q(R))$, we generate randomized versions of R and run the same query over the samples. Let $\hat{\mathcal{R}} = \{\hat{R}_1, \ldots, \hat{R}_k\}$ be a set of randomizations of R. We will specify in Section 4.1 how to generate such randomized versions of R. Then the one-tailed *empirical* p*-value* of $f(q(R))$ with the hypothesis of $f(q(R))$ being small is [11],

$$\frac{|\{\hat{R} \in \hat{\mathcal{R}} : f(q(\hat{R})) \leq f(q(R))\}| + 1}{k + 1}. \tag{3.1}$$

This definition represents the fraction of randomized samples having a smaller value of the statistic f as the original data when using the same query. If the p-value is small, e.g., below a threshold value $\alpha = 0.05$, we can say that the value of $f(q(R))$ is significant in the original data. The threshold value α is a compromise between the power of the test, i.e., the ability to reject the null-hypothesis, and the possibility of Type I error, i.e., the error of rejecting a null-hypothesis when it is actually true. The one-tailed p-value with the hypothesis of f being large and the two-tailed p-value are defined similarly.

3.2 Multiple Hypotheses Testing If we test multiple hypotheses at the same time, for example if we want to test whether the average rate of movies by the several genres in the database is small, it would be imprudent to use the same threshold value α for inferring the statistical significance of the results as is used for determining the significance of one hypothesis. As the number of hypotheses increases, the probability of incorrectly rejecting one null-hypothesis increases.

There exist various methods for controlling the error made in statistical inference in multiple hypotheses testing. The classical method by Bonferroni controls the *familywise error rate* (FWER), i.e., the probability of making one or more false discoveries, by using a transformed threshold value alpha $\alpha_0 = \alpha/N$, where N is the number of the tests. This is a very conservative approach—the probability of making Type II error, i.e., the error of failing to reject a null-hypothesis when it is not true, is high. The extended Holm-Bonferroni method alleviates this problem slightly [12].

	30	60
Romance	1	0
Drama	1	1
History	0	1

(a) GM · MD · DA

	30	60
Romance	2	0
Drama	3	2
History	0	2

(b) GM ∗ MD ∗ DA

Figure 4: (a) Binary relation Genre × Age obtained via boolean product between GM · MD · DA of Figure 2; (b) Contingency table of paths between Genre and Age obtained via matrix product of GM ∗ MD ∗ DA.

The *false discovery rate* (FDR) is a less conservative procedure for correcting multiple comparisons. It measures the expected false positive rate, i.e., the proportion of incorrectly rejected null-hypotheses. The FDR is good for selecting a list of rejected null-hypothesis, especially in exploratory data analysis. For example, the method by Benjamini-Hochberg is a simple way to limit the FDR below the chosen threshold α. [1]

In this paper, we will not use any correction for multiple comparisons to keep the experimental results simple and easily interpretable. The main contribution of this paper resides in the new approach for significance testing of queries, not multiple hypothesis testing. Any multiple hypothesis testing correction can be directly used to correct the p-values obtained, as explained here.

3.3 Where to Randomize? The challenge is how to generate the set $\hat{\mathcal{R}}$, that is, the different randomized versions of $R = \bowtie S$, to compute the empirical p-value as defined above. Consider the toy example in Figure 2. Suppose we want to evaluate whether the average age of the directors of drama movies, as in query q_2 of Section 2, is young (small value in the data). A first naive approach is to consider randomizing directly the binary matrix obtained from the boolean product of all relations from Genre to Age. The boolean product tells us whether there is a path from the set of nodes of Genre to the set of nodes of Age, exactly as required by query q_2.

A traditional permutation test[1] on this new matrix shown in Figure 4(a) can produce only two possible random samples: either the original matrix, or a matrix where the age values between Romance and History are swapped. For the particular case of romance movies with the hypothesis of having small age, we would obtain a p-value close to 0.5 (i.e. 50% of the randomized samples would have the same value as the original). Thus the result is not significant. Indeed, under such randomization none of the three genres would test significantly small, nor large, nor different.

[1] A traditional permutation test would swap any values in the matrix, while keeping the row and column sums fixed. In binary data this is called swap randomization.

Algorithm 1 Query significance in multi-relational data

Input: A set of binary relations $S \subseteq \{A_1, \ldots, A_n\}$, a query $q(\bowtie S)$ and a hypothesis over the statistic $f(q(\bowtie S))$
Output: A set of p-values
1: **for each** binary relation $A \in S$ **do**
2: Obtain k random samples of A, $\hat{\mathcal{A}} = \{\hat{A}_1, \ldots, \hat{A}_k\}$
3: Let $\hat{\mathcal{R}}_A = \{\bowtie \mathrm{T} \cup \hat{A} \mid \hat{A} \in \hat{\mathcal{A}} \text{ and } T = S \backslash A\}$
4: Compute the p-value using the random samples $\hat{\mathcal{R}}_A$
5: **end for**

Alternatively, we could apply a traditional permutation test on the contingency table of paths [6], shown in Figure 4(b). This table gives the number of paths between the Genre and Age, as required by q_1. The hypothesis related to our queries under those permutation tests would never be significant.

The problem of these direct approaches is that they ignore the structure of the relations occurring in the query. In our toy example there are three binary relations participating in the query: GM, MD and DA. Indeed, these relationships convey some structure on the data: the relation MD shows that all history movies are also drama movies; the relation MD shows that all movies from Drama and Romance have been directed by the same person. How do these structures affect the significance of the results in a query?

In queries involving multiple binary relations, there is no unique way to randomize. To assess the structural effect that each particular binary table from S has over the query $q(\bowtie S)$, we should randomize only that corresponding binary relation. That is, the different randomizations of $\bowtie S$ are obtained by randomizing a single relation $A \in S$ while keeping the rest fixed.

More formally, the random samples of $\bowtie S$, when only $A \in S$ is randomized, are defined as follows:

$$\hat{\mathcal{R}}_A = \{\bowtie \mathrm{T} \cup \hat{A} \mid \hat{A} \in \hat{\mathcal{A}} \text{ and } T = S \backslash A\},$$

where $\hat{\mathcal{A}} = \{\hat{A}_1, \ldots, \hat{A}_k\}$ is the set of randomized versions of the original $A \in S$. In Section 4.1 we describe the different randomization techniques to obtain such samples. Finally, these randomized samples $\hat{\mathcal{R}}_A$ will be used to compute the corresponding p-value, as described in Equation (3.1).

Observe that for a given query involving relations in S, we can obtain one p-value for each $A \in S$ we randomize (while keeping $S \backslash A$ fixed). Each p-value is interesting as it measures the structural effect that the participant relation A has on the significance of the result of the query.

The sketch of the method is described in Algorithm 1. The basis of our proposal can be found in traditional statistics under the name of restricted randomizations (see e.g. [11], typically to test whether a treatment variable has effect on a response variable).

3.4 Example We study now the toy example in Figure 2 to illustrate the use of the framework. Consider a query defined such as q_1 from Section 2, yet on the three different Genres.

The first hypothesis that romance movies are directed by young directors obtains a p-value of 0.131 when randomizing on GM, a p-value of 0.494 on MD, and a p-value of 0.495 on MA. The hypothesis is not significant under any randomization, but we observe that randomizing on GM obtains the smallest p-value for this query. Therefore, the structure of GM has an effect on the significance of the related query.

The hypothesis that history movies are directed by old directors obtains p-values 0.269, 0.045, 0.495 when randomizing on GM, MD, DA, respectively. Thus the hypothesis is significant considering the structure in relation MD: all non-history movies are directed by the same person.

Finally, the hypothesis that drama movies are directed by young (or old) directors is not significant in any of the randomizations, always with a p-value close to 1 when randomizing on GM or MD, and p-value of 0.495 when randomizing on DA.

In summary: the age value of 30 associated to romance movies is close to being significant when randomizing on GM because, when focusing on the movies, Romance is a non-intersecting genre to the cluster of genres Drama and History—all history movies are also drama movies in GM; the age value of 60 associated to history movies is significant when randomizing on MD because, when focusing on the directors, the history movies are non-intersecting with the romance and drama movies—all romance and drama movies are directed by the same person; also, the relation DA always swaps with equal probability, because of its one-to-one structure. In the next section we will understand better the reason of these explanations.

4 Randomizations in Multi-relational Model

This section describes how to obtain random samples for a single relation A (line 2 in Algorithm 1), and presents the combinatorial properties of combining such samples with the other relations in the query (line 3 in Algorithm 1).

4.1 Types of Randomization Given a binary relation A we use three different types of randomization to obtain random samples from A. In this paper, rows and columns of A correspond to the binary table presentation of A, e.g., as seen in Figure 2. The running times and space consumptions of the methods are linear in the size of the relation A.

(1) *Swap randomization* of A, as used in [7, 10], produces random samples of A that preserve the row and column sums. The algorithm starts from the original dataset A and performs local swaps interchanging a pair of 1's with a pair of 0's preserving the row and column sums. Technically, a local swap consists of selecting entries $(i,j),(k,l) \in A$ such that $(i,l),(k,j) \notin A$, and swapping the elements so that $(i,j),(k,l) \notin A$ and $(i,l),(k,j) \in A$. From the point of view of the bipartite graph of the relation A, a local swap represents a flip between two independent edges.

A sequence of swaps is performed until the data mixes sufficiently enough in a Markov chain approach [2, 3], and therefore, a random sample of A is obtained. We use ten times the number of ones in the matrix as the number of swaps, which suffices for the convergence of the chain [10]. We denote the set of all random samples that can be reached via swap randomization of A as $\mathsf{sw}(A)$.

(2) *Row permutation* of A corresponds to permuting the order of the rows of A. We denote the set of all random samples that can be reached via row permutation of A as $\mathsf{rp}(A)$.

(3) *Column permutation* of A corresponds to permuting the order of the columns of A. We denote the set of all random samples that can be reached via column permutation of A as $\mathsf{cp}(A)$.

While the swap randomization has been used in [10] to assess the data mining results on a single binary relation, the new randomizations, corresponding to row and column permutations, do not make sense in such a context. The row or column permutation of a matrix does not change any of the frequent pattern solutions in the new randomized matrix. These permutations only make sense in a multi-relational data model, where the permuted matrices are combined with other relations, as presented in this paper, since we are eventually interested in the the final join combination of those permuted matrices with other relations. Both row and column permutation of a single relation change the global paths from the source nodes to destination nodes in the query graph, and thus, the evaluation of the query can change on the randomized data.

4.2 Properties Next we study the properties of combining the obtained random samples with the other relations in the query. For simplicity, we study the case of queries with only two occurring relations $q(A \bowtie B)$ and use boolean product as a simplification of the natural join. The boolean product corresponds to natural join with projection where the common attribute between A and B is dropped out from the final result. For notational convenience, we overload the boolean product for the sets of binary matrices, e.g.,

$\mathsf{sw}(A) \cdot \mathsf{sw}(B)$ represents the boolean product of each pair of elements $A \in \mathsf{sw}(A)$ and $B \in \mathsf{sw}(B)$. Note, particularly, that $\mathsf{sw}(A)$, $\mathsf{rp}(A)$ and $\mathsf{cp}(A)$ refer to sets of matrices.

The relationship between swap randomizations and permutations can be stated as follows.

PROPOSITION 4.1. *Let A be a binary matrix. Then:*

- $\mathsf{rp}(A) = \mathsf{sw}(I) \cdot A$*, where I is an identity matrix;*
- $\mathsf{cp}(A) = A \cdot \mathsf{sw}(I)$*, where I is an identity matrix;*
- *if A has one 1 in each column, then $\mathsf{sw}(A) = \mathsf{cp}(A)$; if A has one 1 in each row, then $\mathsf{sw}(A) = \mathsf{rp}(A)$.*

Proof. Note that $\mathsf{sw}(I)$, for identity matrix I, can produce any permutation matrix with uniform distribution [17]. Thus, the boolean product $\mathsf{sw}(I) \cdot A$ produces all permutations for the rows of A and similarly, the boolean product $A \cdot \mathsf{sw}(I)$ produces all permutations of the columns of A. If A has exactly one 1 in each row, then each local swap corresponds to the swap of the corresponding rows, thus $\mathsf{sw}(A)$ corresponds to permuting the order of rows. Intuitively, these row (or column) permutations can be seen as a random re-assignment of the row (or column) names in A.

Next, we present several properties relating swap randomization to row and column permutations. They follow from Proposition 4.1. These properties are useful to not repeat unnecessary randomizations in applying the approach.

PROPOSITION 4.2. *Let A, B binary relations. Then:*

- $\mathsf{cp}(A \cdot B) = A \cdot \mathsf{cp}(B)$
- $\mathsf{rp}(A \cdot B) = \mathsf{rp}(A) \cdot B$
- $\mathsf{cp}(A) \cdot B = A \cdot \mathsf{rp}(B) = \mathsf{cp}(A) \cdot \mathsf{rp}(B)$

Proof. Using Proposition 4.1 we get that $\mathsf{cp}(A \cdot B) = A \cdot B \cdot \mathsf{sw}(I) = A \cdot \mathsf{cp}(B)$. Similarly for $\mathsf{rp}(A \cdot B) = \mathsf{rp}(A) \cdot B$. Finally, $\mathsf{cp}(A) \cdot B = A \cdot \mathsf{sw}(I) \cdot B = A \cdot \mathsf{rp}(B) = A \cdot \mathsf{sw}(I) \cdot \mathsf{sw}(I) \cdot B = \mathsf{cp}(A) \cdot \mathsf{rp}(B)$, because $\mathsf{sw}(I) \cdot \mathsf{sw}(I) = \mathsf{sw}(I)$.

This means that column and row permutations do not make sense in more than one relation, e.g., $A \cdot \mathsf{cp}(B \cdot C \cdot D) \cdot E = A \cdot B \cdot C \cdot \mathsf{cp}(D) \cdot E$. The last property of Proposition 4.2 states that only one permutation, either column permutation on A or row permutation on B, is indeed necessary.

Next, we give two direct implications of Proposition 4.1 that reduce the number of different randomizations considerably. These will be applied extensively in the experiments.

THEOREM 4.1. *Let A, B be binary relations. Then: $A \cdot \mathsf{sw}(I) \cdot B = \mathsf{cp}(A) \cdot B = A \cdot \mathsf{rp}(B)$, where I is an identity matrix.*

Hence, we prefer to use the notation with the identity matrix I to refer to the row and column permutations. This also removes the multiple-presentation problem seen in Proposition 4.2. The operation $A \cdot \mathsf{sw}(I) \cdot B$ randomizes the boolean product, whereas the operations $\mathsf{sw}(A) \cdot B$ and $A \cdot \mathsf{sw}(B)$ randomize the original data. From this perspective, $A \cdot \mathsf{sw}(I) \cdot B$ tells about the significance of the combination operation, while $\mathsf{sw}(A) \cdot B$ tells whether the structure in A is significant.

THEOREM 4.2. *Let A and B be binary relations. Then: If A has one 1 in each column, then $\mathsf{sw}(A) \cdot B = A \cdot \mathsf{sw}(I) \cdot B$. If B has one 1 in each row, then $A \cdot \mathsf{sw}(B) = A \cdot \mathsf{sw}(I) \cdot B$.*

In real world datasets, it is quite common to have such one-to-one relations. For example, the ages of the directors in the example in Figure 2 are one-to-one. Thus swap randomization of the relation DA produces the same set of samples as randomizing the connection between the relations MD and DA.

Next, we show that the sets of randomizations $\mathsf{sw}(A) \cdot B$, $A \cdot \mathsf{sw}(I) \cdot B$ and $A \cdot \mathsf{sw}(B)$ are different in general.

THEOREM 4.3. *Let A, B be binary matrices. Then:*

- $A \cdot B \subseteq \mathsf{sw}(A) \cdot B \subseteq \mathsf{sw}(A) \cdot \mathsf{sw}(B)$*;*
- $A \cdot B \subseteq A \cdot \mathsf{sw}(B) \subseteq \mathsf{sw}(A) \cdot \mathsf{sw}(B)$*;*
- $A \cdot B \subseteq A \cdot \mathsf{sw}(I) \cdot B \subseteq \mathsf{sw}(A) \cdot \mathsf{sw}(I) \cdot \mathsf{sw}(B)$.

Especially, in general, the sets $\mathsf{sw}(A) \cdot B$, $A \cdot \mathsf{sw}(B)$ and $A \cdot \mathsf{sw}(I) \cdot B$ are not subsets of each other.

Proof. The set $\mathsf{sw}(A)$ of swap randomizations of A contains always the original matrix A, i.e., $A \subseteq \mathsf{sw}(A)$, thus the given inclusions hold. To show that $\mathsf{sw}(A) \cdot B$, $A \cdot \mathsf{sw}(B)$ and $A \cdot \mathsf{sw}(I) \cdot B$ are different and not subsets of each other, consider the following matrices A and B.

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A \cdot B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

There exist three swap randomized versions of A, which correspond to permuting the order of the rows, as well as three swap randomized versions of B. For the 2×2 identity matrix I, there are two swap randomizations. The corresponding resulting sets of the three randomizations are:

$$\mathsf{sw}(A) \cdot B = \left\{ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \right\}$$

$$A \cdot \mathsf{sw}(B) = \left\{ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} \right\}$$

$$A \cdot \mathsf{sw}(I) \cdot B = \left\{ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \right\}$$

Thus the only common randomization in the sets $\mathsf{sw}(A) \cdot B$, $A \cdot \mathsf{sw}(B)$ and $A \cdot \mathsf{sw}(I) \cdot B$ is the original $A \cdot B$. The example can also be generalized to arbitrarily large relations.

Theorem 4.3 tells us that the set of samples that can be obtained by randomizing two relations is larger than by randomizing only one relation. As discussed in Section 3.3, we prefer to randomize a single table at a time in order to control much better the structural effect the randomized relation has on the query. Additionally, we know that the sets of randomized samples $\mathsf{sw}(A) \cdot B$, $A \cdot \mathsf{sw}(B)$ and $A \cdot \mathsf{sw}(I) \cdot B$ are different from each other, thus it makes sense to do them all separately if the conditions of Theorem 4.2 do not hold.

To sum up, we have the following result:

COROLLARY 4.1. *For a query $q(A \bowtie B)$, there exist three different randomizations: (i) $\mathsf{sw}(A)$ while keeping B fixed; (ii) $\mathsf{sw}(B)$ while keeping A fixed; (iii) $\mathsf{sw}(I)$ where I is an identity relation between the columns of A and the rows of B (this corresponds to $\mathsf{cp}(A)$ and $\mathsf{rp}(B)$).*

Notice that if A or B are one-to-one relations, then randomization (iii) will be the same as (i) or (ii), respectively. Each randomization provides a set of samples from which we can compute a p-value for our query (hypothesis on the data). Every p-value is interesting as it shows how the structure of the randomized relation affects the significance.

4.3 Example Revisited The p-values reported in Section 3.4 for the toy example in Figure 2, correspond to swap randomization of the binary tables GM, or MD, or DA, respectively. Indeed, because MD has one single 1 in each row, we have that GM $\cdot$ $\mathsf{sw}(I)$ $\cdot$ MD $\cdot$ DA is equal to GM $\cdot$ sw(MD) $\cdot$ DA. Similarly, because DA is a one-to-one relation, we have GM $\cdot$ MD $\cdot$ $\mathsf{sw}(I)$ $\cdot$ DA equals GM $\cdot$ MD $\cdot$ sw(DA). Thus, for this example, only swap randomization in the three tables is necessary.

Interestingly, we can understand better now the p-values reported in Section 3.4. On the relation GM, drama movies and history movies have no independent edges to swap between them. Therefore, the pattern of History implying Drama tends to remain in random samples. As a result, the p-value of the hypothesis related to history or drama movies is not significant. On the other hand, the p-value related to romance movies becomes close to being significant because, for this genre, the null distribution diverges more from the original. The fact that there are only two romance movies raises this p-value slightly above the 0.05 threshold.

Similar explanation applies when randomizing MD. When looking at MD, local swaps can interchange at most two edges between movies of the young director C. Waitt and movies of the not-so-young director T. George. Actually, in all random samples coming from MD we observe that C. Waitt has always at least three movies from either drama or romance. As a result, neither drama nor romance can be significant—in the null distribution they are always closely linked to a young director as in the original data. Yet, history movies directed by T. George have more local swaps that would create a diverging null distribution—most of the samples in the null distribution have the history movies connected to the age of 30. The hypothesis of history movies being directed by a not-so-young person is then significant.

$E[\mathsf{sw}(\text{GM}) * \text{MD} * \text{DA}]$ $E[\text{GM} * \mathsf{sw}(\text{MD}) * \text{DA}]$ $E[\text{GM} * \text{MD} * \mathsf{sw}(\text{DA})]$

$$\begin{pmatrix} \mathbf{0.849} & \mathbf{1.151} \\ 3.269 & 1.731 \\ 0.882 & 1.118 \end{pmatrix} \quad \begin{pmatrix} 1.413 & 0.587 \\ 3.587 & 1.413 \\ \mathbf{1.455} & \mathbf{0.545} \end{pmatrix} \quad \begin{pmatrix} 0.984 & 1.016 \\ 2.492 & 2.508 \\ 1.016 & 0.984 \end{pmatrix}$$

Figure 5: Expectation of the number of paths when swap randomizing relation GM, MD or DA, respectively. The rows correspond to genres Romance, Drama and History, in this order, and the columns to Ages 30 and 60, in this order.

5 Studying Path Distributions

For a query $q(A \bowtie \ldots \bowtie B)$ where $A \subseteq I \times L$ and $B \subseteq J \times K$, let $P = A * \ldots * B$ be the matrix product of all relations participating in q. This corresponds to the contingency table of paths from origin I to destination nodes K. An example is shown in Figure 4(b) for the toy data of Figure 2. For all types of queries, the significance of the result is closely related to the path distributions between nodes I and K. For example, suppose we want to test whether the average age of history-movie directors is large. In the original data of Figure 3, there are two paths from History to the age of 60 and no path to the age of 30. It is sensible to assume that if we had random samples where paths are mainly swapped the other way round, it would turn the hypothesis into a significant finding.

Naturally, a simple way to visualize whether there exists an interesting finding in the data is to compare the path distribution of P with the expected path distribution on the given random samples. The larger the change, the more significant the result would tend to be. Examining path distributions gives us an idea of which statistic might be significant for the relations involved in the query.

The following three matrices in Figure 5 show the expectation of the paths when swap randomizing relation GM, MD or DA, respectively, for the example in Figure 2. The genre that swaps most of its paths under randomizations with GM is Romance. We had observed that the p-value of the hypothesis related to romance movies would drop to 0.131 in this randomization. This means that in only 131 samples out of 1000 we observed paths between Romance

and the age of 30. History swaps the paths from the age of 60 to the age of 30 when randomizing on MD. Indeed, we know that the hypothesis linked to such query becomes significant here with a a p-value of 0.045, that is only 45 samples out of 1000 preserve the path between History and the age of 60. Randomization on DA distributes paths fifty-fifty for each genre. The p-values obtained there were always close to 0.5.

6 Empirical Results

In this section, we present empirical results on synthetic and real datasets. Our real dataset is **MovieLens**, which is very similar to IMDb, yet containing more binary relations. In all cases, we calculate the empirical p-values over 999 randomized samples and use the threshold of $\alpha = 0.05$ to determine the query significance. We do not apply any correction for multiple hypothesis testing, but see Section 3.2 for discussion of multiple comparisons. Any of these corrections can be applied directly to the p-values obtained here.

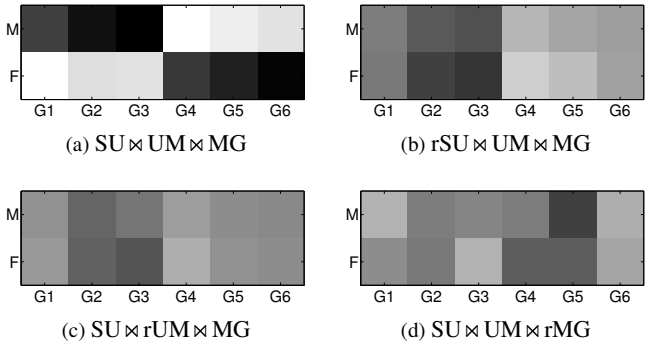
The randomization methods are fast in practice. In our experiments, producing one randomized sample took approximately the same time as evaluating the query. The time and space consumption of the methods scale linearly in the size of the relation. The sequential approach for calculating the empirical p-value by Besag and Clifford [4] can be used to determine the sufficient amount of randomized samples in large-scale applications. For example, in many cases 30 samples is usually sufficient to determine the significance of the result. In that case, performing the significance testing of the query takes approximately 30 times longer than just evaluating the plain query.

6.1 Synthetic Dataset To motivate our approach and understand better why randomizations are consistent with the inferences about our hypothesis, we generate a synthetic dataset to simulate relations of users, movies and genres. We will be interested in testing the following hypothesis.

HYP 1. *Men watch different types of movies than women.*

The relations occurring in the query are: Gender×User (SU), User×Movie (UM) and Movie×Genre (MG).

For studying the behavior of randomizations, we generate the tables SU, UM and MG to make our hypothesis clearly be significant. We let SU contain 30 men and 20 women, thus SU is a 2×50 binary table where the first 30 values in the first row and the last 20 values in the second row are 1s. We generate UM to be a 50×100 binary table where men watch any of the first 60 movies with probability of 0.40 and any of the last 40 movies with probability of 0.05. To create a strong pattern, we let the probabilities of a female watching movies be the other way round. Finally, we generate MG as a 100×6 binary table where the first three genres will be considered to be manly and the last three genres will be considered to be womanly. For each movie in the relation, we select two genres as follows: for the first 60 movies we select a genre from the manly genres with a probability of 0.9 and from the womanly genres with a probability of 0.1. For the last 40 movies the probabilities are the other way round. So, each movie has at most two genres, because if we happen to select the same genre for a movie twice, then we say that the movie has only one genre.



(a) SU ⋈ UM ⋈ MG (b) rSU ⋈ UM ⋈ MG

(c) SU ⋈ rUM ⋈ MG (d) SU ⋈ UM ⋈ rMG

Figure 6: Proportion of paths going from a gender (M=male, F=female) to a genre (G1–G6) in the different combined tables. Lighter color represents less paths, while darker more paths; to be more exact: white corresponds to the lowest value of 4.5% and black to the highest value of 30%.

Next we create the anti-tables from those above, called rSU, rUM and rMG. These anti-tables will not contain any structure at all, they are random. We let rSU be a 2×50 binary table with 30 men and 20 women where the order of the users is random. We generate rUM to be a 50×100 table with each element being 1 with a probability of $(0.40 + 0.05)/2$. Finally, we let rMG be formed similarly to MG but with the two genres for each movie assigned uniformly with replacement.

The goal of this experiment is to study how the p-values of Hyp 1 change when combining the original significant tables SU, UM and MG to one of these nonsignificant tables. Figure 6 shows the contingency table of paths from those combinations. We notice that using the original tables SU, UM and MG (Figure 6(a)) produces clearly a significant difference between the types of movies that males and females watch. By replacing one of the original tables with a random version, the pattern seems to disappear. Still, we cannot clearly see from the path distributions which of the underlying tables mainly breaks the original structure. We would like to check with our tests whether randomizing in the proper tables will tell us where the pattern is broken.

For the test, we use the following statistic.

STATISTIC 1. *L_1 distance between the distribution of genres of the movies that men and women have watched.*

This statistic is the sum of the absolute differences between the proportion of paths of men and women, as shown in Fig-

Input relations			p-values			
A	B	C	$\mathsf{sw}(I_{\mathrm{AB}})$	$\mathsf{sw}(\mathrm{B})$	$\mathsf{sw}(I_{\mathrm{BC}})$	$\mathsf{sw}(\mathrm{C})$
SU	UM	MG	0.001	0.001	0.001	0.001
rSU	UM	MG	**0.517**	0.030	0.013	0.003
SU	rUM	MG	**0.282**	**0.279**	**0.155**	0.124
SU	UM	rMG	0.001	0.001	**0.704**	**0.727**

Table 1: Significance tests for the Hyp 1 with the combined input relations A $\bowtie$ B $\bowtie$ C. The first three columns contain the relations considered as input, labeled A, B and C. Columns 4th to 7th are empirical p-values for the hypothesis when only one relation is randomized: $\mathsf{sw}(I_{\mathrm{AB}})$ randomizes the identity matrix between relations A and B, which is equivalent to $\mathsf{sw}(\mathrm{A})$, randomizing the relation A; $\mathsf{sw}(\mathrm{B})$ randomizes only on relation B; $\mathsf{sw}(I_{\mathrm{BC}})$ randomizes the identity matrix between relations B and C; $\mathsf{sw}(\mathrm{C})$ randomizes only relation C. Bold p-values correspond to randomizations which touch the anti-tables.

ure 6 for different inputs. The original value of the statistic with the tables SU, UM and MG is 1.23, implying a clear difference between males and females. When one of the tables SU, UM and MG is replaced with the corresponding anti-table, the value of the L_1 statistic is around 0.1.

In Table 1 we show the results of the several significance tests for the hypothesis Hyp 1 on the several combined tables. There is a clear connection between the structure of the relations A, B and C occurring in the query and the p-values obtained by randomizing in different relations. As expected, the empirical p-value of Hyp 1 with tables SU, UM and MG is significant with randomizations in all tables. On the other hand, when one of the clearly-structured tables SU, UM or MG is replaced by the anti-tables rSU, rUM or rMG respectively, we obtain large empirical p-values for those randomizations that touch the anti-tables (see the bold values of Table 1). This illustrates how randomizations can tell about the structural effects in the significance of a query.

6.2 MovieLens Dataset The **MovieLens** data is collected through the MovieLens web site (`movielens.umn.edu`). The downloadable data is already cleaned up, i.e., users who had less than 20 ratings or did not have complete demographic information were removed from the data set. In all, the data consists of 100,000 ratings (valued from 1 to 5) from 943 users on 1,682 movies. Each user has rated at least 20 movies and the demographic information for the users correspond to attributes of age, gender, occupation and zip code. For each movie we have title, release year and a list of genres. Furthermore, we interpret that if a user has rated a movie, it means that he or she has watched it. This corresponds to the binary table named UM. From the ratings we have calculated an average movie rating and also an average user rating—corresponding to the average value that a user has given for his/her movies. These correspond to the relations MR and UR, respectively. In Table 2, we summarize the binary relations in the **MovieLens** dataset.

Relation	Description	Rows	Cols	# of 1's/row
UM	User×Movie	943	1680	106
MG	Movie×Genre	1680	18	1.7
UO	User×Occupation	943	21	1
US	User×Gender	943	2	1
MA	Movie×Age	1680	1680	1
MR	Movie×Rating	943	943	1
UA	User×Age	943	943	1
UR	User×Rating	943	943	1

Table 2: Summary of binary relations in MovieLens dataset. The tables MA, MR, UA and UR are identity maps between movies or users and their ages or ratings. We denote a transpose by reversing the relation name.

	Mean	(Std)	p-value
SU $\bowtie$ UM $\bowtie$ MG	0.16		
sw(SU) $\bowtie$ UM $\bowtie$ MG	0.03	(0.01)	0.001
SU $\bowtie$ sw(UM) $\bowtie$ MG	0.01	(0.00)	0.001
SU $\bowtie$ UM $\bowtie$ $\mathsf{sw}(I_{\mathrm{M}})$ $\bowtie$ MG	0.03	(0.01)	0.001
SU $\bowtie$ UM $\bowtie$ sw(MG)	0.02	(0.00)	0.001

Table 3: Evaluation results of Hyp 2 on SU $\bowtie$ UM $\bowtie$ MG. Mean and std are the average and standard deviation of Statistic 2 in the original input data (first row) and several randomizations. The randomization SU $\bowtie$ $\mathsf{sw}(I_{\mathrm{U}})$ $\bowtie$ UM $\bowtie$ MG is equivalent to sw(SU) $\bowtie$ UM $\bowtie$ MG.

We handle numerical values by repetition. The tables MA, MR, UA and UR are identity maps between movies or users and their ages or ratings. Each column in these tables corresponds to the age or rating of one user or movie. Thus two different columns may correspond to the same numerical value. Handling numerical values in this way guarantees that two users or two movies having the same age or rating are not combined into a single one after a join and a projection.

Next, we go through a few queries on the dataset and analyze their significances. In the tests, we use Theorem 4.2 to identify theoretically equivalent randomizations. It applies to most of the tables: UO, US, MA, MR, UA and UR.

HYP 2. *Men watch different types of movies than women.*

To assess Hyp 2 we use the same statistic as for Hyp 1.

STATISTIC 2. *L_1 distance between the distribution of genres of the movies that men and women have watched.*

In Table 3 we give the empirical p-values for Hyp 2. Each row shows the relation being randomized for obtaining

Genre G	Orig.	sw(SU)	sw(UM)	sw(I_{M})	sw(MG)
Action	2.5	0.001	0.001	0.001	0.001
Sci-fi	1.5	0.001	0.001	0.001	0.001
Thriller	1.1	0.001	0.001	0.001	0.001
Adventure	0.8	0.001	0.001	0.001	0.001
Crime	0.6	0.002	0.001	0.001	0.002
War	0.5	0.002	0.001	0.004	0.002
Horror	0.4	0.019	0.001	0.011	0.020
Western	0.2	0.001	0.001	0.005	0.003
Film-noir	0.1	0.012	0.001	0.054	0.058
Mystery	0.0	0.392	0.395	0.424	0.469
Document.	0.0	0.404	0.391	0.468	0.489
Fantasy	-0.1	0.064	0.051	0.243	0.201
Animation	-0.2	0.032	0.001	0.027	0.018
Musical	-0.5	0.001	0.001	0.001	0.001
Children's	-1.0	0.001	0.001	0.001	0.001
Comedy	-1.3	0.001	0.001	0.001	0.001
Drama	-2.3	0.001	0.001	0.001	0.001
Romance	-2.3	0.001	0.001	0.001	0.001

Table 4: Empirical p-values for Hyp 3 on SU $\bowtie$ UM $\bowtie$ MG. The values for the associated Statistic 3 in the original relations are given in the second column. The different randomizations methods (columns 3rd to 6th) correspond to randomizing in one relation at a time from SU $\bowtie$ I_{U} $\bowtie$ UM $\bowtie$ I_{M} $\bowtie$ MG. The randomization sw(I_{U}) is equivalent to sw(SU). Genres are sorted by the value of the statistic. Significance tests say: genres over the first dashed line are more watched by men (p-values always under 0.05); genres under the second dotted line are more watched by women (p-values always under 0.05). We cannot say anything about genres in between the two dotted lines.

		sw(OU)			sw(I_{M})		
	Orig.	Mean	(Std)	p-val.	Mean	(Std)	p-val.
None	0.23	0.13	(0.05)	**0.038**	0.07	(0.01)	0.001
Librarian	0.18	0.05	(0.02)	**0.001**	0.04	(0.01)	0.001
Retired	0.18	0.10	(0.04)	**0.040**	0.05	(0.01)	0.001
Homemaker	0.17	0.14	(0.05)	0.269	0.15	(0.03)	**0.226**
Doctor	0.15	0.14	(0.05)	0.373	0.08	(0.02)	0.001
Entert.	0.14	0.09	(0.03)	0.073	0.04	(0.01)	0.001
Educator	0.13	0.04	(0.01)	**0.001**	0.03	(0.01)	0.001
Lawyer	0.13	0.11	(0.04)	0.237	0.05	(0.01)	0.001
Salesman	0.12	0.11	(0.04)	0.330	0.06	(0.01)	0.001
Healthcare	0.12	0.09	(0.03)	0.211	0.04	(0.01)	0.001
Student	0.11	0.03	(0.01)	**0.001**	0.03	(0.01)	0.001
Scientist	0.11	0.07	(0.02)	0.052	0.05	(0.01)	0.001
Artist	0.10	0.07	(0.03)	0.130	0.04	(0.01)	0.001
Technician	0.10	0.07	(0.03)	0.183	0.03	(0.01)	0.001
Programmer	0.08	0.05	(0.02)	**0.025**	0.03	(0.01)	0.001
Engineer	0.08	0.05	(0.02)	**0.034**	0.03	(0.01)	0.001
Marketing	0.08	0.07	(0.03)	0.340	0.05	(0.01)	0.006
Writer	0.08	0.06	(0.02)	0.122	0.03	(0.01)	0.001
Executive	0.07	0.07	(0.02)	0.337	0.04	(0.01)	0.001
Administr.	0.05	0.04	(0.02)	0.367	0.02	(0.01)	0.001
Other	0.04	0.04	(0.01)	0.483	0.02	(0.00)	0.002

Table 5: Evaluation results of Hyp 4 on OU $\bowtie$ UM $\bowtie$ MG. The original values of Statistic 4, with mean and std of randomized samples are given. The randomization sw(I_{U}) is equivalent to sw(OU). The results on randomizations sw(UM) and sw(MG) were similar to sw(I_{M}). Bold p-values are significant with sw(OU) and nonsignificant with sw(I_{M}).

the corresponding p-value. The query associated to the hypothesis traverses the relations Gender $\times$ User $\times$ Movie $\times$ Genre, corresponding to relations SU, UM and MG. There are five different types of randomizations of the query. However, the randomizations SU $\bowtie$ sw(I_{U}) $\bowtie$ UM $\bowtie$ MG and sw(SU) $\bowtie$ UM $\bowtie$ MG are equivalent by Theorem 4.2. The results in Table 3 show that Hyp 2 is significant wrt all different randomizations.

Indeed, the results on Hyp 2 seem to indicate that men watch movies with different genres than women. All randomizations are consistent. We will next analyze which genres separate men and women. We repeat the following hypothesis (with associated query) for each genre G.

HYP 3. *Men watch genre G more (or less) than women.*

STATISTIC 3. *The difference between the %-proportions of the movies from genre G among all the movies men and women have watched.*

Notice this statistic is similar to Statistic 2 but now we only look at the difference for the specific genre G. The empirical p-values of the significance testings of Hyp 3 are given in Table 4. Again, we find out that randomizations in different relations produce fairly similar results in general. We can observe that men watch significantly more, for example, action and sci-fi movies than women, whereas women watch significantly more romance and drama movies than men. Interestingly, we can say the popularity of mystery and documentary movies do not depend on the gender. Actually the genres which have the smallest amount of movies are the least significant ones. The genres with fewest number of movies are fantasy (with 22 movies), film-noir (24), western (27), animation (41) and documentary (50).

Next we study users by their occupation.

HYP 4. *The users with occupation O watch different types of movies than other users with different occupations.*

STATISTIC 4. *L_1 distance between the distributions of genres of the movies watched by users with occupation O and users with other occupations.*

The results of the significance testings are given in Table 5. When evaluating the associated query, we find that randomizing in different relations matters for that query. For most of

	Orig.	$\mathsf{sw}(\mathrm{AU})$	$\mathsf{sw}(\mathrm{UM})$	$\mathsf{sw}(I_{\mathrm{M}})$	$\mathsf{sw}(\mathrm{MG})$
Film-noir*	35.8	0.001	0.001	0.003	0.001
Documentary	35.0	**0.134**	0.001	0.001	0.001
Mystery	34.3	**0.197**	0.001	0.004	0.001
War	34.2	**0.308**	0.001	0.004	0.001
Drama	34.1	**0.493**	0.001	0.001	0.001
Western	33.8	**0.307**	0.001	**0.168**	**0.060**
Romance*	33.4	0.024	0.001	0.039	0.002
Musical	33.0	0.016	**0.253**	**0.469**	**0.257**
Crime	32.6	0.001	0.001	**0.181**	**0.411**
Comedy*	32.5	0.001	0.001	0.003	0.007
Thriller*	32.2	0.001	0.001	0.003	0.004
Adventure*	32.0	0.001	0.001	0.001	0.006
Fantasy	32.0	0.002	0.001	**0.130**	**0.164**
Children's*	31.8	0.001	0.001	0.002	0.001
Sci-fi*	31.8	0.001	0.001	0.001	0.003
Action*	31.7	0.001	0.001	0.001	0.001
Horror*	31.1	0.001	0.001	0.001	0.001
Animation*	30.9	0.001	0.001	0.004	0.002

Table 6: Empirical p-values for Hyp 5 on $\mathrm{AU} \bowtie \mathrm{UM} \bowtie \mathrm{MG}$. The randomization $\mathrm{AU} \bowtie \mathsf{sw}(I_{\mathrm{U}}) \bowtie \mathrm{UM} \bowtie \mathrm{MG}$ is equivalent to $\mathsf{sw}(\mathrm{AU}) \bowtie \mathrm{UM} \bowtie \mathrm{MG}$. Genres with a star are significant with all randomizations. Bold p-values are nonsignificant.

the occupations, Hyp 4 is not significant when randomizing on $\mathsf{sw}(\mathrm{OU}) \bowtie \mathrm{UM} \bowtie \mathrm{MG}$ nor $\mathrm{OU} \bowtie \mathsf{sw}(I_{\mathrm{U}}) \bowtie \mathrm{UM} \bowtie \mathrm{MG}$. For the other randomizations we have that all occupations, except homemakers, exhibit significance of the hypothesis. We observe that the largest occupation groups of librarians (51), educators (95) and students (196) have the most significant empirical p-values for the query, with all type of randomizations. See the bold p-values. We could infer that those type of users watch different genres than other users.

HYP 5. *Average age of the users who have watched movies of a given genre is significant.*

STATISTIC 5. *Weighted average age of the users who have watched movies of the given genre.*

The results of assessing Hyp 5 are given in Table 6. The empirical p-values of the queries depend largely on the type of randomization used. By randomizing the ages of the users, that is, $\mathsf{sw}(\mathrm{AU}) \bowtie \mathrm{UM} \bowtie \mathrm{MG}$, the movies whose average age of watchers has originally been around 34 years are not significant. This makes sense when it is compared to the average of all users which is 34.1 years. Notice that in the query the average is weighted by the number of movies watched by the user. Thus randomizing the table AU tests the connection between the ages and the users. Other randomization points tell us that the results on western, romance, crime and fantasy are not significant, whereas the results on other genres are significant. Thus the inner

	Mean	(Std)	p-value
$\mathrm{AU} \bowtie \mathrm{UM} \bowtie \mathrm{MA}$	0.16		
$\mathsf{sw}(\mathrm{AU}) \bowtie \mathrm{UM} \bowtie \mathrm{MA}$	-0.00	(0.03)	0.001
$\mathrm{AU} \bowtie \mathsf{sw}(\mathrm{UM}) \bowtie \mathrm{MA}$	-0.00	(0.03)	0.001
$\mathrm{AU} \bowtie \mathrm{UM} \bowtie \mathsf{sw}(\mathrm{MA})$	-0.00	(0.09)	0.033

Table 7: Evaluation results of Hyp 6 on $\mathrm{AU} \bowtie \mathrm{UM} \bowtie \mathrm{MA}$. The randomization $\mathsf{sw}(I_{\mathrm{U}})$ is equivalent to $\mathsf{sw}(\mathrm{AU})$, and the randomization $\mathsf{sw}(I_{\mathrm{M}})$ is equivalent to $\mathsf{sw}(\mathrm{MA})$.

	Mean	(Std)	p-value
$\mathrm{RU} \bowtie \mathrm{UM} \bowtie \mathrm{MR}$	0.40		
$\mathsf{sw}(\mathrm{RU}) \bowtie \mathrm{UM} \bowtie \mathrm{MR}$	-0.00	(0.03)	0.001
$\mathrm{RU} \bowtie \mathsf{sw}(\mathrm{UM}) \bowtie \mathrm{MR}$	0.07	(0.03)	0.001
$\mathrm{RU} \bowtie \mathrm{UM} \bowtie \mathsf{sw}(\mathrm{MR})$	-0.00	(0.07)	0.001

Table 8: Evaluation results of Hyp 7 on $\mathrm{RU} \bowtie \mathrm{UM} \bowtie \mathrm{MR}$. The randomization $\mathsf{sw}(I_{\mathrm{U}})$ is equivalent to $\mathsf{sw}(\mathrm{RU})$ and the randomization $\mathsf{sw}(I_{\mathrm{M}})$ is equivalent to $\mathsf{sw}(\mathrm{MR})$.

structure of the User$\times$Movie and Movie$\times$Genre relations explain the results of our query. The average ages of the users of the following genres were significant with all types of randomizations: film-noir, romance, comedy, thriller, adventure, children's, sci-fi, action, horror and animation.

HYP 6. *Old people watch old movies.*

STATISTIC 6. *Correlation between the age of the movies and the age of the users who have watched the movie.*

The correlation between the age of the movies and the age of the users who have watched the movie is pretty small, 0.16. Normally this would be directly regarded as insignificant correlation. In Table 7 the results on significance tests of the hypothesis are given. We notice that the hypothesis is significant according to all randomization points. However, randomizing the Movie$\times$Age relation gives an empirical p-value of 0.033 thus implying that part of this result is basically explained by the release years of the movies.

HYP 7. *The average rating of the movies a user has rated correlates with the mean of the ratings the user has given.*

STATISTIC 7. *The correlation between the average rating of the movies a user has rated and the mean of the ratings the user has given.*

The original value of the test statistic, that is the correlation, is 0.40. Thus it is again fairly small but still positive. The results of significance testings with different randomizations are given in Table 8. With all randomizations the original correlation is significant.

7 Related Work

To the best of our knowledge there is no work that addresses the problem presented in this paper, thus preventing us from comparing our approach with other methods. Obviously, there is a large amount of statistical literature about hypothesis testing [5, 11]. For the particular case of data mining, many papers work on the significance of association rules and other patterns [18, 19]. In the recent years, the framework of randomizations has been introduced to the data mining community to test significance of patterns: the papers [7, 10] deal with randomizations on binary data, and the work in [15] studies randomizations on real-valued data. For another type of approach to measuring p-values for patterns, see [20]. A related work that studies permutations on networks and how this affects significance of patterns is [14]. Sub-sampling methods such as bootstrapping [9] use randomization to study the properties of the underlying distribution instead of testing the data against some null-model. Finally, database theory studies mainly query processing and optimization in different complex data [8, 13].

8 Conclusions and Future Work

We have addressed the problem of assessing the significance of queries made for the exploratory analysis of relational databases. Each query, together with the associated statistic, define the hypothesis to test on our data. Our mathematical tool to decide the significance is via randomizations. It turns out that in multi-relational data there is no unique way to randomize. We propose to randomize tables occurring in the queries one at a time, and obtain a set of p-values for each randomization. This choice is theoretically justified by the combinatorial properties of the randomizations. Each p-value tells not only how significant is our hypothesis, but also what is the structural impact of the randomized table in the query. For example, if certain structures or patterns remain after the randomizations, the answers of a query that rely on such patterns should not be significant. Experiments with synthetic data showed that for well defined significant patterns, randomizations reveal which tables from our database convey most of the structure in the query. For real datasets, we tested several hypothesis to show the usability of the method. Still, we found out that in real data it is difficult to give a fully satisfactory answer about how to use all the obtained p-values to conclude the correct inference. However, most of the studied real datasets contained many one-to-one binary relations or very sparse tables, making the different randomizations theoretically or practically equivalent. Our contribution makes an important first step towards understanding how the structure hidden in the data makes some hypotheses more significant than others, but still, a lot of interesting future work needs to be done: study of the combinatorial properties and its connection to the significance of queries and patterns.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[2] J. Besag. Markov chain Monte Carlo methods for statistical inference. `http://www.ims.nus.edu.sg/Programs/mcmc/files/besag_tl.pdf`, 2004.

[3] J. Besag and P. Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76(4):633–642, 1989.

[4] J. Besag and P. Clifford. Sequential Monte Carlo p-values. *Biometrika*, 78(2):301–304, 1991.

[5] G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, 2001.

[6] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu. Sequential MC methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–120, 2005.

[7] G. W. Cobb and Y-P. Chen. An application of Markov chain Monte Carlo to community ecology. *The American Mathematical Monthly*, 110:265–288, 2003.

[8] O. de Moor, D. Sereni, P. Avgustinov, and M. Verbaere. Type inference for datalog and its application to query optimisation. In *PODS'08*, pages 291–300, 2008.

[9] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[10] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM TKDD*, 1(3), 2007.

[11] P. I. Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses; Springer series in statistics.*, volume 2nd. Springer, 2000.

[12] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[13] A. Jha, V. Rastogi, and D. Suciu. Query evaluation with softkey constraints. In *PODS'08*, pages 119–128, 2008.

[14] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[15] M. Ojala, N. Vuokko, A. Kallio, N. Haiminen, and H. Mannila. Randomization methods for assessing data analysis results on real-valued matrices. *Statistical Analysis and Data Mining*, 2(4):209–230, 2009.

[16] R. Ramakrishnan and J. Gehrke. *Database Management Systems*. McGraw-Hill Higher Ed., 2003.

[17] H. J. Ryser. Combinatorial properties of matrices of zeros and ones. *Canad. J. Math.*, 9:371–377, 1957.

[18] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *DMKD*, 2(1):39–68, 1998.

[19] P-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD '02*, pages 32–41, 2002.

[20] G. I. Webb. Discovering significant patterns. *Mach. Learn.*, 68(1):1–33, 2007.