Discovering Significant Patterns under Sequential False Discovery Control

Sebastian Dalleiger
CISPA Helmholtz Center for Information Security
Saarbruecken, Germany
sebastian.dalleiger@cispa.de

Jilles Vreeken
CISPA Helmholtz Center for Information Security
Saarbruecken, Germany
jv@cispa.de

ABSTRACT

We are interested in discovering those patterns from data with an empirical frequency that is significantly differently than expected. To avoid spurious results, yet achieve high statistical power, we propose to *sequentially* control for false discoveries *during* the search. To avoid redundancy, we propose to update our expectations whenever we discover a significant pattern. To efficiently consider the exponentially sized search space, we employ an easy-to-compute upper bound on significance, and propose an effective search strategy for sets of significant patterns. Through an extensive set of experiments on synthetic data, we show that our method, Spass, recovers the ground truth reliably, does so efficiently, and without redundancy. On real-world data we show it works well on both single and multiple classes, on low and high dimensional data, and through case studies that it discovers meaningful results.

CCS CONCEPTS

• **Mathematics of computing → Information theory**; **Probabilistic inference problems**; **Exploratory data analysis**; **Statistical paradigms**; • **Information systems → Data mining**.

KEYWORDS

pattern mining, binomial test, multiple hypothesis testing, sequential hypothesis testing, false discovery rate, family-wise error rate, maximum entropy distribution

ACM Reference Format:
Sebastian Dalleiger and Jilles Vreeken. 2022. Discovering Significant Patterns under Sequential False Discovery Control. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539398

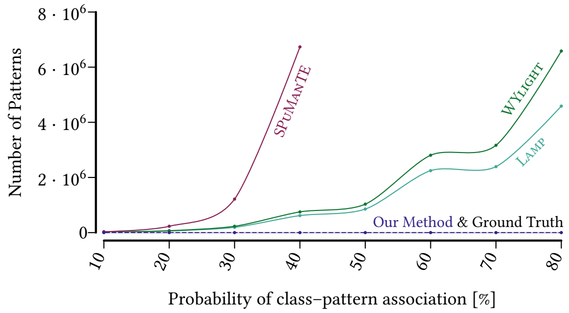


Figure 1: Our method recovers the ground truth where competitors struggle. We show the number of *statistically significant* patterns discovered at an FWER of 0.05 on two-class data where we vary the probability of associating 50 ground truth patterns with its classes.

1 INTRODUCTION

A cornerstone of many scientific problems is discovering statistically significant associations between features from data. In the biomedical domain, for example, researchers are interested in identifying combinations of genetic markers that are associated with specific phenotypes [22, 23, 45], studying combinations of mutations caused by cancer [40], or analyzing correlated markers that together indicate a high survival chance of a patient [32]. *Statistically significant pattern mining* is a branch of data mining in which we are after those patterns that are *statistically significant* with regard to some null hypothesis. Thus, it is particularly well-suited to meet the needs of many scientific domains.

A key issue plaguing significant pattern mining is the *multiple hypothesis testing problem*: If we test a single pattern for significance, the probability of falsely rejecting the null hypothesis is bounded by its p-value. This probability quickly converges to 1 as we test more hypotheses, however, and since the pattern search space is exponential in the number of binary features, we drown in spurious results unless we use some form of false discovery control. One option is to limit risk of making at least one false discovery, also known as controlling the Family-Wise Error Rate (FWER), and another is to limit the expected number of false discoveries, which is known as controlling the False Discovery Rate (FDR). Most work in the field focuses on finding a good balance between *statistical power* and *computational efficiency*.

Although recent work achieves impressive results, it falls short when it comes to reporting succinctly and without redundancy. To illustrate this problem, we run three recent statistically significant pattern miners, Lamp [37], WYlight [24], and SPuManTE [29], on synthetic two-class data using only 50 ground truth patterns, randomly associated to each class. The higher the class association probability, the more patterns participate in generating the class. At 100%, for example, all 50 patterns are used to generate each classes

KDD '22, August 14–18, 2022, Washington, DC, USA.

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539398

(see Sec. 5 for more details). In Fig. 1, we show the number of patterns discovered by Lamp, WYlight, SPuManTE, and our method at a significance level of 0.05. There, Lamp, WYlight, SPuManTE identify orders of magnitude more patterns as significant as we originally used to generate the data. Although not incorrectly, since subsets or combinations of ground-truth patterns might also be significant. However, these redundant results drown the analyst in patterns.

To achieve concise and informative, rather than redundant results, we propose to test patterns for significance against our expectation, based on the patterns we have discovered so far. To prevent spurious results, yet achieve high statistical power, we sequentially control for either family-wise error rate or false discovery rate. That is, we iteratively adjust the significance level α during the search, factoring in what part of the space we explored and what hypotheses we have rejected. Our method, Spass (significant Pattern Associations), automatically associates patterns to those classes for which they are significant, thereby immediately exposing the similarities and differences between the classes. This allows us to handle data with one or more classes, while existing methods can only handle data with one or two classes.

Through an extensive set of experiments, we show that Spass performs well in practice. While its competitors drown the analyst in large numbers of highly redundant patterns, we demonstrate that Spass reliably recovers the ground truth in synthetic data and discovers succinct and non-redundant patterns in real-world data. In two detailed case studies on real-world data, we illustrate that the patterns identified by Spass are also meaningful. Furthermore, Spass is very fast, taking only seconds up to a few minutes in our experiments where competitors take hours, days, or even weeks.

Our main contributions are as follows. In particular, we

(1) suggest to iteratively test for significance against a probabilistic model of the data based on our most current knowledge of the data,
(2) propose a novel sequential FWER control and introduce the first pattern mining procedure under sequential online FDR control,
(3) introduce the Spass algorithm to efficiently discover non-redundant sets of statistically significant patterns using an easy-to-compute Chernoff bound, and
(4) provide an extensive empirical evaluation on synthetic and real-world data.

The remainder of the paper is structured as follows. In Section 2, we state our problem formally. Next, we introduce and analyze our method Spass in Section 3. Related work is discussed in Section 4, and we empirically evaluate Spass in Section 5. After discussing our method in Section 6, we conclude our paper in Section 7. We provide additional details for reproducibility in the supplementary, and make all code and data publicly available.[1]

2 SIGNIFICANT PATTERN SETS

In this section we define our problem, starting with notation, after which we informally describe the problem. We then move to the statistical test for one hypothesis, its efficient inference, and afterwards introduce a sequential hypothesis testing procedure.

2.1 Notation

We consider data $\mathcal{X}$ of n samples over m binary features, or items, $\mathcal{I}$, where every sample $y \in \mathcal{X}$ is independently drawn from the powerset $2^{\mathcal{I}}$ of all possible samples, or itemsets. We assume that the $\mathcal{X}$ is partitioned into $k \geq 1$ classes $\{X_1, \ldots, X_k\}$. Each class X_i in itself is a multiset of $|X_i|$ independently drawn itemsets.

As patterns, we consider itemsets $x \subseteq \mathcal{I}$, i.e. the powerset $2^{\mathcal{I}}$ is the set of all possible patterns. A data point $y \in \mathcal{X}$ *supports* a pattern x iff $x \subseteq y$. The empirical frequency q of an itemset x in a class X_i is

$$q_{X_i}(x) = |\{y \in X_i \mid x \subseteq y\}| \, / |X_i| \ .$$

A pattern set $S \subseteq 2^{\mathcal{I}}$ is simply a set of patterns. We will maintain one pattern set S_i for each class X_i. A pattern $x \in S_i$ is said to be associated with class X_i. Combined with empirical frequencies, a pattern set S_i is the sufficient statistics of our probability distribution p over $2^{\mathcal{I}}$.

With our notation in place, we are now ready to informally state the problem.

2.2 The Problem, Informally

Overall, our goal is to discover those patterns whose empirical frequencies in the data differ significantly from what we would expect, based on what we already know about the data. We strive to do this for datasets with one or multiple classes over the same set of binary features, such that we find not only patterns that are distributed significantly differently in general but also patterns that are distributed significantly differently in one or multiple classes.

We explicitly seek to prevent redundant results, and hence require that every reported pattern is significant in light of all previously discovered patterns. This formulation has the benefit that it allows us to *sequentially* control for false discoveries by adjusting the significance threshold during the search, based on what part of the search space we have considered so far.

Existing statistical pattern mining approaches are reporting every significant pattern, often including the subsets or combinations of true pattern, which introduces redundancy. The key idea of our approach is that we discover non-redundant results by testing each pattern for significance against a model of the data based on prior discoveries, and do so using an appropriately adjusted significance level.

A bit more formally, our goal is to discover one pattern set S_i for each class X_i, such that the empirical frequency $q_{S_i}(x)$ of each pattern $x \in S_i$ diverges significantly from our expected frequency $p_{S_i \setminus \{x\}}(x)$ based on patterns already accepted prior to x, while controlling for false discoveries.

With this intuition in mind, we next describe our probabilistic model and the statistical test for *one* hypothesis. Afterwards, we will show how to sequentially control for false discoveries when testing multiple hypotheses using either family-wise error rate (FWER) or false discover rate (FDR).

2.3 Testing for Significance

To infer an expected frequency $p_{S_i}(x)$, we need a probability distribution p. We choose p by the principle of maximum entropy [20],

[1] eda.mmci.uni-saarland.de/spass

which states that the distribution f which uses no additional information beyond what it was explicitly provided, is the distribution that satisfies the given constraints, but otherwise maximizes Shannon entropy. Accordingly, we define the expectation $p_{S_i}(x)$ as the expected probability

$$\mathbb{E}_f[x] = \sum_{\substack{y \in 2^{\mathcal{I}} \\ x \subseteq y}} f(y \mid S_i) \ ,$$

under the distribution f that maximizes the Shannon entropy

$$\arg\max_f - \sum f_x \log f_x \ ,$$

subject to linear equality constraints $\mathbb{E}_f[x] = q_{X_i}(x)$ for all elements in S_i. We know from Csiszár [9] that f has the form

$$f(x \mid S) = \theta_0 \prod_{y_j \in S} \theta_j^{\mathbf{1}\{y_j \subseteq x\}} \ ,$$

where $\mathbf{1}$ is the indicator function. Although convex, finding the maximizer requires exponentially many inferences, is known to be a PP-hard problem [36], and thus cannot easily be approximated. To still allow for efficient inference, we factorize p into independent and fast-to-infer factors [11, 12], starting as the product of marginal frequencies. Since these marginals are the building blocks of our factorization, we explicitly model the frequencies of each singleton $y \in \mathcal{I}$ per class.

Formally, we state the *null hypothesis* that the distribution of an $x \subseteq \mathcal{I}$ in class X_i follows the expectation p given $S_i \subseteq 2^{\mathcal{I}}$ as

$$H : \quad q_{X_i}(x) = p_{S_i}(x) \ ,$$

and the *alternative hypothesis* that it is distributed differently as

$$H^a : \quad q_{X_i}(x) \neq p_{S_i}(x) \ .$$

A pattern x either occurs in a sample in X_i or it does not, and under the null hypothesis it is hence Bernoulli distributed with success probability $p_{S_i}(x)$. By convention, we assume that each sample in X_i is independently drawn, by which the number of samples in which x occurs becomes binomially distributed. Under the null hypothesis, the p-value $\mathbb{P}[H]$ of observing a pattern x with a more extreme frequency $q_{X_i}(x)$ than our expectation $p_{S_i}(x)$ is

$$\mathbb{P}\left[n_x \geq \hat{n}_x \mid H\right] \ ,$$

where $n_x = |X_i| \cdot q_{X_i}(x)$ is the *observed* number of data points that supports x, and $\hat{n}_x = |X_i| \cdot p_{S_i}(x)$ is the *expected* number of such data points. To infer these p-values exactly, we can use the binomial cumulative distribution function

$$F(s, n; p) = \sum_{k=s}^{n} \binom{n}{k} p^k (1-p)^{n-k} \ , \tag{1}$$

for the number of successes s, number of trials n, and success probability p. More precisely, if $q_{X_i}(x) \geq p_{S_i}(x)$ we can infer $\mathbb{P}[H]$ by computing $F(n_x, |X_i|; p_{S_i}(x))$, or else, we do so with $F(0, n_x; p_{S_i}(x))$. Although mathematically convenient, as we may have to infer the CDF exponentially often, computing F exactly during our search is impractical. We therefore propose to approximate F using the easy-to-compute Chernoff bound [8],

$$F(n \cdot a, n; b) \leq \exp\left[-nD(a \| b)\right] \ ,$$

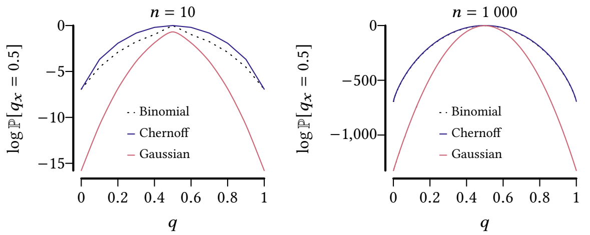


Figure 2: Chernoff closely approximates the binomial CDF. For the fixed probability $p(x)$ of 0.5, we show p-values for Chernoff, Gaussian, and the exact binomial CDF over 10 (left) and 1 000 (right) samples.

where $D(a\|b)$ is the Kullback-Leibler divergence

$$a \log a/b + (1-b) \log(1-a)/(1-b)$$

of the two Bernoulli distributions a and b. To illustrate how well the Chernoff bound approximates the binomial CDF in comparison to the popular standard normal approximation, we show p-values for 10 and 1 000 samples for a success probability of 0.5 in Fig. 2. There, we see that the Chernoff bound tightly approximates exactly computed p-values, even for few samples.

2.4 Controlling for False Discoveries

If we test a single hypothesis, the probability of falsely rejecting the null hypothesis H is bounded by its p-value $\mathbb{P}[H]$. The more hypotheses we test, the probability of falsely rejecting at least one null hypothesis converges to 1, that is, unless we carefully control for testing multiple hypotheses. We consider two approaches to false discovery control, namely, one targeting the family-wise error rate and one targeting the false discovery rate. Both have in common that, rather than testing each hypothesis at the same significance level α, they test hypotheses at *adjusted* significance levels $\alpha_t < \alpha$. In a nutshell, in both cases, we consider a sequence of hypotheses

$$H_1, H_2, \dots \ ,$$

for which we compute the corresponding sequence of p-values

$$\mathbb{P}[H_1], \mathbb{P}[H_2], \dots \ .$$

We decide to reject the t^{th} hypothesis H_t if its p-value $\mathbb{P}[H_t]$ is less than the *adjusted* test level α_t, i.e.,

$$\mathbb{P}[H_t] < \alpha_t \ ,$$

and denote the set of all rejections as $\mathcal{R} = \{H_t \in \mathcal{H} \mid \mathbb{P}[H_t] \leq \alpha_t\}$, where $\mathcal{H}$ is the set of all hypotheses. Regardless of how we control α_t, we ideally want to maximize the statistical power, also known as the *true discovery rate* (TDR)

$$\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}^a|}{\max\{1, |\mathcal{H}^a|\}}\right] ,$$

where $\mathcal{H}^a = \{H \in \mathcal{H} \mid H^a = 1\}$ is the unknown set of truly alternative hypotheses. In the following, we discuss how to determine a test level sequence α_t that achieves a high TDR while controlling for false discoveries, starting with the conservative family-wise error rate and then moving on to the less conservative false discovery rate.

Controlling FWER. We start with the adjustment of the significance levels α_t to guarantee a FWER of at most α. The *Family-Wise Error Rate*, or FWER for short, is the probability

$$\mathbb{P}[|\mathcal{R} \cap \mathcal{H}_0|] > 0$$

of making at least one false discovery, where $\mathcal{H}_0 = \{H \in \mathcal{H} \mid H = 1\}$ is the unknown set of all hypotheses that are truly null. We can keep the FWER below α by testing against an adjusted significance threshold $\alpha_t = \alpha/N$, where we simply divide α by the number N of hypotheses in $\mathcal{H}$. This is known as *Bonferroni correction* [6]. While Bonferroni correction works well when testing relatively few hypotheses, in our case, $N = k \cdot 2^m$ is exponential in m, and hence, for any non-trivial value of m, the adjusted values α_t will be so low that the probability of a true discovery is (almost) zero.

In statistically significant pattern mining, one common approach to increase the TDR is by outright excluding hypothesis if their *minimally attainable p-value* is above the significance threshold [35]. This is *Tarone's exclusion principle* [27]. Since, the minimally attainable p-value in our case shrinks exponentially with number of samples in a class (cf. the infimum of Eq. (1)), it becomes so small that we cannot excluding many hypotheses in advance. We can, however, exploit the fact that we typically do not test all patterns but rather a much smaller set C of candidate patterns. Hence, it suffices to adjust α by the size of C, rather than N, since $|C| \ll N$. Unfortunately, we do not know C in advance. But fortunately, we do know how we *generate* C. We, therefore, make our significance level adjustment *search space aware* [4, 42, 43]. That is, we *sequentially* adjust the significance levels α_t while we iteratively generate C.

To do so, we need to impose structure on the search space. That is, we organize the hypotheses (i.e., patterns) as a lattice, such that layer l contains all patterns of length l [2]. If we now search for significant patterns layer by layer, we only have to correct for up to and including the current layer l, which is at most the sum of all binomials up to l. While easy to compute for small l, for larger values, this sum is computationally costly, and hence, we resort to the upper bound

$$\sum_{k=1}^{l} \binom{m}{k} < \sum_{k=1}^{l} \frac{m^k}{k!} \leq \mathrm{e}^l (m/l)^l \ . \tag{2}$$

To obtain the adjustment factor we need for the t^{th} hypothesis, we multiply the right-hand side of Eq. (2) with the number of classes k. We summarize the above in the following lemma.

Lemma 1. For any sequence of p-values, we control for the FWER by adjusting the test levels, for the t^{th} hypothesis using

$$\alpha_t = \min\nolimits_{s<t} \left\{\alpha_s,\ \alpha[k \cdot \mathrm{e}^l (m/l)^l]^{-1}\right\} , \tag{3}$$

where l is the highest layer in the search lattice explored so far, $m = |\mathcal{I}|$ is the number of features which coincides with the highest lattice level, and k is the number of classes under consideration.

Proof. At each level $l \in \{1, \ldots, m\}$, we adjust for all possible hypotheses up to that layer, which grows to at most $l = m$. By summing Eq. (2) up to m, we achieve equality with Bonferroni correction. □

Although many domains require FWER, there are problems do not need such strict error control. In such cases, we therefore control for the less conservative FDR, which we describe in the following.

Controlling FDR. The *False Discovery Rate* [5], or FDR for short, is an alternative approach to false discovery control. To permit a higher statistical power than FWER, the FDR is controls for the *expected number* of false discoveries

$$\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{1, |\mathcal{R}|\}}\right] ,$$

rather than the probability of at least one false discovery.

To ensure an FDR of at most α, we can determine α_t using so-called generalized α-investing (GAI) rules [3]. These rules "invest" a fraction of our available "α-budget" in each significance test we perform, thus decreasing the available α-budget, and reward each discovery by increasing the available α-budget. In short, we start with a budget of $0 < w_0 < \alpha$, decrease the budget when testing the t^{th} hypothesis with a penalty ϕ_t, and increase the budget with a reward ψ_t when we reject it. Thus, we can continue testing as long as we make discoveries. Formally, our budget develops as

$$w_{t+1} \leftarrow w_t - \phi_t + \mathbf{1}[\mathbb{P}[H_t] < \alpha_t] \cdot \psi_t \ . \tag{4}$$

Since our p-values are statistically dependent, and we seek high statistical power, we employ a variant of the LORD-update rules proposed by Javanmard and Montanari [19]. We start with an initial budget of $w_0 = \alpha/2$. For every discovery, we receive a constant reward $\psi_t = \alpha - w_0$. To prevent that we use all available budget on a single hypothesis, we set the penalty ϕ_t and the level α_t to a fraction

$$\alpha_t \leftarrow \gamma_t \cdot w_\tau$$

of the budget w_τ available at the most recent discovery time

$$\tau = \operatorname*{arg\,max}_{s<t} \mathbf{1}\{\mathbb{P}[H_s] < \alpha_s\} = 1 ,$$

using a non-increasing sequence $(\gamma_t)_{t\geq1}$ as the fraction of our budget w_t that invest into the test in iteration t. To choose such a sequence $(\gamma_t)_{t\geq1}$ for *arbitrarily dependent* p-values, we resort to Thm. 3.7 from Javanmard and Montanari [19]. In essence, this theorem states that any non-increasing sequence $(\gamma_t)_{t\geq1}$ guarantees a bounded FDR if

$$\sum_{t}^{\infty} \gamma_t (1 + \log t) \leq \alpha/w_0 ,$$

holds. In particular, this is true for the sequence

$$\gamma_t = \frac{6}{t^2\pi^2} \frac{\alpha/w_0}{(1 + \log t)}$$

which we summarize in the following lemma.

Lemma 2. For $\gamma_t = \frac{6}{t^2\pi^2} \frac{\alpha/w_0}{(1+\log t)}$, the generalized α-investing rules described above control FDR for arbitrarily dependent p-values.

Proof. By substituting γ_t in Thm. 3.7 from [19], we observe that the factor $1 + \log t$ cancels out. Since $\sum_t^\infty \frac{6}{t^2\pi^2}$ converges to 1, the series converges exactly to α/w_0. □

Algorithm 1: Spass

```
   Input: classes $X_1, \ldots, X_k$, test level $\alpha \in [0, 1]$
   Output: patterns sets $S_1, \ldots, S_k$
1  $S_i \leftarrow \emptyset \; \forall i \in \{1, \ldots, k\}$
2  $C \leftarrow \{x \subseteq \mathcal{I} \mid |x| = 2\}$
3  $t \leftarrow 1$
4  while $C \neq \emptyset$
5      $\hat{x}, C \leftarrow$ Search$(C)$
6      foreach class $X_i$ do
7          $t \leftarrow t + 1$
8          adjust test level $\alpha_t \leftarrow$ [Eq. (3) or Eq. (4)]
9          hypothesize $H_t : p_{S_i}(\hat{x}) = q_{X_i}(\hat{x})$
10         if $\mathbb{P}[H_t] < \alpha_t$
11             $S_i \leftarrow S_i \cup \{\hat{x}\}$
12             estimate coefficients for $p_{S_i}$
13 return $S_1, \ldots, S_k$
```

3 THE SPASS ALGORITHM

Now, we introduce our algorithm Spass for efficiently discovering significant, non-redundant patterns under false discovery control. We give the pseudocode of Spass in Algorithm 1.

We start with an empty pattern set S_i for each class X_i (l. 1), an initial search space C that contains all itemsets of length two (l. 2), and set the significance tests counter to 1 (l. 3). Then, we expand the search-space C using the Search algorithm, detailed below, selecting that candidate

$$\hat{x} \leftarrow \underset{x \in C}{\arg\min} \; \mathbb{E}[p_S(x) = q_X(x)] \tag{5}$$

which has the lowest expected chance (l. 5)

$$\mathbb{E}[p_S(x) = q_X(x)] = \sum_{i=1}^{m} \mathbb{P}[p_{S_i}(x) = q_{X_i}(x)]$$

of resulting in false discoveries, across all distributions. We test the significance of $\hat{x}$ (l. 9) for every class X_i (l. 6) against a significance level that is appropriately adjusted according to either FWER or FDR (l. 7–8). If the candidate is significant, we reject the null hypothesis, add $\hat{x}$ to S_i, and re-infer the distribution $p_{S_i}(\cdot)$ (l. 10–11). We iterate until convergence, and finally return the k pattern sets (l. 12).

To identify the next pattern to test, we use Algorithm 2. Given the current search space C, we first find the most promising candidate $\hat{x} \in C$ using Eq. (5) (l. 1). We then expand C with all combinations of $\hat{x}$ and singletons $y \in \mathcal{I}$ (l. 2). Note that this corresponds to exploring (part of) layer $l + 1$ of the lattice, where $l = |\hat{x}|$ is the layer which $\hat{x}$ resides. If there exists an x in the now-expanded search space C that is better than $\hat{x}$, we recurs (l. 4), and otherwise, we return the best candidate $\hat{x}$ and the search space C (l. 6) we explored so far.

The computational complexity of Spass depends on the size of C, which grows binomially with each layer of expansion, and can reach up to 2^m elements in total. By assuming that the complexity of inferring the expectations p is bounded by a constant, the worst-case complexity is hence in $O(2^m)$. Since in practice we never explore the entire lattice, so Algorithm 1 has a complexity of $O(\mathrm{e}^l (m/l)^l)$ after reaching the l^{th} layer of the lattice. As we do not expand layers fully either, Spass is still more efficient in practice.

Algorithm 2: Search

```
   Input: search space $C \subseteq 2^{\mathcal{I}}$
   Output: candidate $\hat{x}$ and expanded search space $C$
1  $\hat{x} \leftarrow \arg\min_{x \in C} \mathbb{E}[p_S(x) = q_X(x)]$
2  $C \leftarrow C \cup \{\hat{x} \cup \{y\} \mid y \in \mathcal{I}\}$
3  if $\min_{x \in C} \mathbb{E}[p_S(x) = q_X(x)] < \mathbb{E}[p_S(\hat{x}) = q_X(\hat{x})]$
4      return Search$(C)$
5  else
6      return $\hat{x}, C \setminus \hat{x}$
```

4 RELATED WORK

Pattern mining is arguably one of the most important and well-studied areas of data mining. Traditional approaches, such as *frequent itemset mining* [2], aim for completeness, and return all patterns that satisfy some user-defined constraints. By scoring patterns individually, the results of traditional methods are typically very large, highly redundant, and mostly spurious [1].

Pattern set mining aims to search only a small and non-redundant set of patterns that together generalize the data well. Typical quality measures include probabilistic [14] or information-theoretic scores [41], and algorithms have been used for characterizing data with one [13] or multiple classes [7, 11]. Although these methods discover succinct, non-redundant sets of patterns that have been proven useful, the results come without statistical guarantees, which bars their application in critical domains, such as genetics [22, 23, 40, 45] or survival analysis [32], or network analysis [34].

Significant pattern mining provides statistical guarantees by using statistical tests to prune out spurious results. There exist many significance tests, and hence almost as many dedicated statistically significant pattern mining methods, e.g., for Fisher's exact [17, 37], Mann-Whitney-Wilcoxon [37], conditional permutation [45], Westfall-Young permutation [24, 30], Cochran-Mantel-Haenszel [28], Barnard's unconditional [29], Poisson [21], swap randomization [16], or the Likelihood ratio test [33]. Each of these methods corrects for multiple hypothesis testing mostly targeting FWER and using Bonferroni [6] correction. Some methods use Tarone's exclusion principle [35] to increase the statistical power. Another approach to cope with the low statistical power exhibited under Bonferroni correction is to make the adjustment "search-aware" [4, 42, 43] and directly adjust it, without necessarily knowing the total number of hypotheses to adjust for in advance. A search-aware significance level adjustment is also used for the search of non-redundant top-k statistically tested-to-be informative patterns [44].

Although these methods rigorously control for false discoveries, they still test against a static null hypothesis and as a result report every significant pattern, and consequently, they tend to discover many and highly redundant results—often orders of magnitude more than there are samples in the data.

Our goal with Spass is to discover concise, non-redundant sets of statistically significant patterns. Here, we combine the best of pattern set mining and statistical significant pattern mining. Our approach is unique in that it marries sophisticated probabilistic modelling to rigorous statistical testing, while accounting for the

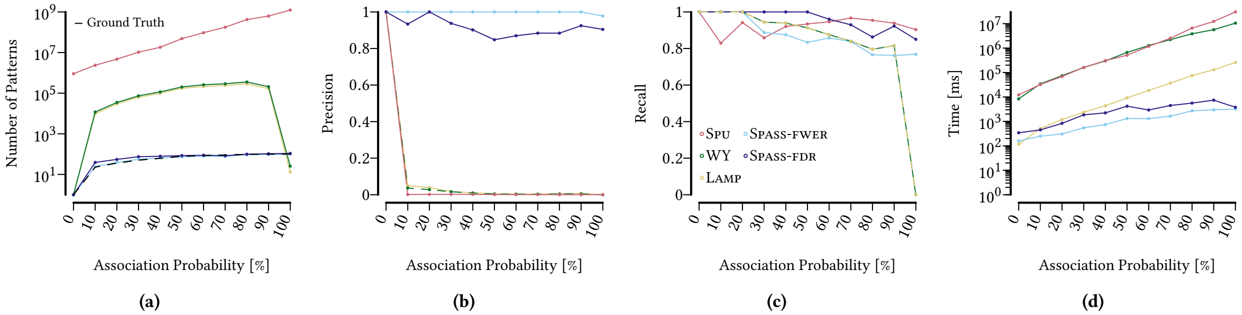


Figure 3: Our method efficiently recovers the ground truth with high precision and recall. Given is the (a) number of significant discoveries, (b) precision, (c) recall, and (d) runtime, for Lamp, WYlight, SPuManTE, and our method, Spass, on synthetic data over 500 unique items, with two classes of 5 000 samples each, in which we plant up to 100 ground truth patterns overall. We vary the association probability p_a by which we independently associate patterns per class; for $p_a = 0$ no patterns are planted at all, while for $p_a = 1$ every pattern is present in both classes.

multiple hypothesis testing problem using a sequential and search-aware significance level adjustment that can target either FWER or FDR.

5 EXPERIMENTS

We implement Spass in C++, and run experiments on an Intel Xeon E5-2643 CPU, reporting wall clock time. To differentiate between FWER and FDR control, we write Spass-fwer and Spass-fdr, respectively. We provide the source code, datasets, synthetic dataset generator, and additional information needed for reproducibility.[2] We compare Spass to three methods for significant pattern mining, Lamp [37], WYlight [24] and SPuManTE [29]; two methods for non-redundant pattern set mining, mtv [26] and desc [11]; and to one statistically non-redundant pattern miner opus [44]. All our competitors represent the state-of-the-art in their respective fields. We report results at a significance level α of 0.05.

5.1 Synthetic Data

To validate that Spass recovers true patterns, we start by evaluating the algorithm on two-class data with known ground truth. To this end, we generate synthetic data as follows. First, we sample 100 random patterns of up to 5 items from an alphabet of $|\mathcal{I}| = 500$ items and insert them into ground truth pattern set S_i^* with an "association" probability varying in between 0% (no patterns planted at all) and 100% (all patterns are shared among all S_i^*). Then, we randomly draw 5 000 samples for each class X_i, in such a way that the ground truth patterns $x \in S_i^*$ all have a randomly chosen frequency of in between 15% and 30%. Afterwards, we add additive noise of 5% and destructive noise of 1%. To account for random fluctuations, we average over 10 samples per 10% increment in probability.

We run significant pattern miners on each dataset and report the average number of statistically significant discoveries in Fig. 3a. At 0% association probability (i.e., no patterns, pure noise) SPuManTE is the only method that wrongfully discovers patterns. Across the board, we see that Lamp, WYlight, and SPuManTE all report orders of magnitude more patterns as significant as the number of patterns used to generate the data. As subsets or combinations of significant patterns are often significant as well, this is not incorrectly per se. Spass, in contrast, almost matches the ground truth in number. At 100% association probability, there are no contrasting patterns and only shared patterns. Lamp and WYlight only identifies that there are almost no contrasting patterns, whereas Spass correctly identifies that we have shared all patterns with all classes.

To assess the quality of the discovered patterns, we measure precision and recall with respect to the ground truth as follows. We match each discovered pattern x with the best-matching ground truth pattern y in terms of set similarity $|x \cap y| / |x \cup y|$. Reporting precision in Fig. 3b and recall in Fig. 3c, we see that all methods obtain good recall, but due to their huge result sets, Lamp, WYlight, and SPuManTE have very low precision. The sequential redundancy control of Spass, however, prevents the exponential growth in the cardinality of its output, and consequently is both precise and often orders of magnitude faster than the competition (see Fig. 3d).

High-Dimensional Synthetic Data. Having ensured that Spass results in non-redundant discoveries under either FDR or FWER

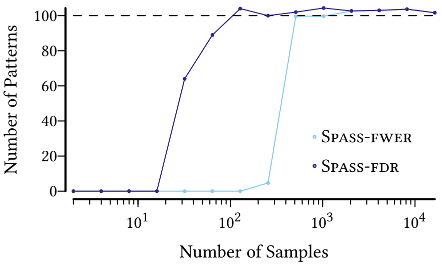
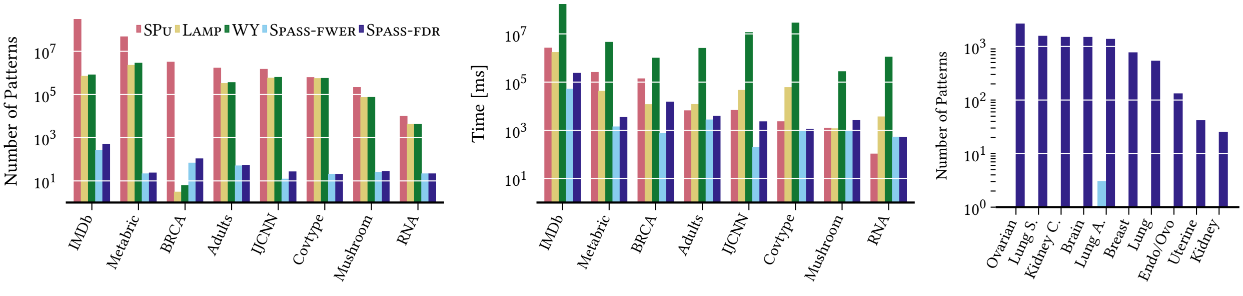


Figure 4: FWER is more conservative than FDR. We show the number of significant patterns discovered by Spass controlling for resp. FWER (blue) and FDR (green) at $\alpha = 0.05$, on synthetic data over 20 000 items with 100 ground truth patterns (dashed line) while varying the number of samples.

[2] eda.mmci.uni-saarland.de/spass



(a) (b) (c)

Figure 5: Unlike competitors, Spass efficiently discovers concise pattern sets, and only Spass-fdr can handle high-dimensional genomics data. We show the number of significant discoveries (a) and runtime needed (b) by SPuManTE, WYlight, Lamp and Spass for eight real-world, two-class datasets. In panel (c), we show the number of discoveries on high-dimensional, one-class and multi-class cancer-genomics data from Spass-fwer and Spass-fdr only, since no competitor discovered patterns.

control, we investigate how much of a difference FWER and FDR can make on high-dimensional synthetic data.

From Eq. (3), it follows that for a very large number of features m or a particularly high search depth l, FWER control becomes very strict. This means that a high dimensionality or very large patterns are particularly challenging. We generate data as above, but now over an alphabet $\mathcal{I}$ of 20 000 items in which we plant 100 random patterns. We run Spass with FWER resp. FDR on data with varying numbers of samples, and report the number of significant discoveries in Fig. 4. We see that both variants converge to the correct number of patterns, but Spass-fdr does so much more quickly, requiring between one and two orders of magnitude less data. As expected, FDR seems to be better-suited for high-dimensional data.

5.2 Real World Data

Now that we have validated that Spass works well on synthetic data, we explore how it performs in a wide range of real-world scenarios. Without access to the ground truth, we cannot compute precision and recall. We can, however, assess the number of discoveries and runtime of Spass relative to its competitors Lamp, WYlight, and SPuManTE, which we report in Fig. 5. In Fig. 5a, we see that the competitors deem orders of magnitude more patterns as significant as Spass. Furthermore, we find that our method discovers fewer patterns when controlling for the more conservative FWER instead of the FDR. From Fig. 5b, we observe that this tendency is reflected in the runtime of Spass-fwer, which is typically lower than that of Spass-fdr. Regardless of the control method, Spass is also almost always faster than its competitors.

Having ascertained that Spass efficiently discovers concise pattern sets from real-world datasets, we turn to case studies to answer three specific questions: (i) Does Spass work on *high-dimensional* real-world data? (ii) Does Spass discover *meaningful* patterns in real-world data? (iii) Can Spass compete with the state-of-the-art in statistical pattern mining on *one-class* real-world data?

5.2.1 High-Dimensional Real-World Data. To verify whether Spass works on high-dimensional real-world datasets, we consider ten gene-expression datasets concerning *Ovarian*, *Lung*, *Kidney*, *Brain*, and *Breast* cancer.[3] Together, these data span a wide range of different sizes, numbers of classes, and numbers of samples, with the shared trait that they are high-dimensional. We run Lamp, WYlight, SPuManTE, and Spass on each dataset, but Lamp, WYlight, and SPuManTE did not finish within 24 hours of runtime, thus reporting no significant patterns. In Fig. 5c, we report the number of discoveries by Spass. Here, as in our experiments on high-dimensional synthetic data, we see that FWER is much more stringent than FDR. For the *Lung A.* dataset, Spass-fwer only discovers 3 significant patterns—its highest achievement—while when we control for FDR, it makes substantially more discoveries and discovers 1 353 patterns. In the *Brain Cancer* dataset, for example, Spass-fdr discovers 1 471 patterns. According to a high-level analysis, the top pattern in the *Brain Cancer* dataset consists of genes involved in neural activities

$$\{\,\textit{A2BP1, CAMK2A, GABRA1, GABRB2, NRGN, PACSIN1,}\\ \textit{SLC12A5, SNAP25, SULT4A1, SYN2, TMEM130, VSNL1}\,\}\,.$$

In contrast, the top pattern from the *Breast Cancer* dataset

$$\{\,\textit{AOC3, AQP7, BTNL9, CIDEC, ERG, GYG2, HSPB6, ITGA7,}\\ \textit{KCNIP2, LPL, PLIN1, PPP1R1A, SLC19A3, TUSC5}\,\}\,,$$

represents high co-expression of 14 membrane-related genes. We conclude that Spass manages to discover interpretable patterns also on high-dimensional real-world data.

5.2.2 Sentiment Analysis. Next, we take a closer look at the quality of the patterns that Spass discovers. To this end, we consider the IMDb movie review dataset [25], which consists of positive and negative movie reviews as text. We run Spass on this data and report associations of natural language patterns to positive or negative sentiments. After eliminating stop words, lemmatizing the corpus, and removing infrequent words, the dataset consists of 50 000 rows with 8 124 features, in which Spass discovers 215 significant patterns under FWER control, which we rank according to their significance. The top-3 patterns,

$$\{\textit{great fantastic}\},\ \{\textit{music sound}\},\ \{\textit{film plot twist}\}\,,$$

are uniquely positive, whereas the most contrasting top-4 patterns between the two classes are

[3] https://www.cancer.gov/tcga

{seen worst}, {piece crap}, {world reality}, {painful watch} .

Regardless of the sentiment, reviewers are concerned with *special effects*, which is the highest-ranked shared pattern.

5.2.3 Clinical Survival Analysis. To further analyze the interpretability of our results, we consider the problem of clinical survival analysis. In particular, we analyze the *METABRIC* breast cancer dataset [10]. It consists of 2 000 samples (patients) over 124 binarized features, with binary labels marking the survival status of each patient. By using Spass, we discover 65 patterns at an FDR of at most 0.05. On average, these patterns consist of 4.5 ± 2.3 items, with the longest pattern having length 12. Among the patterns that are easiest to understand, we identify

$$\{\ \textit{Relapse: Recurred, Patient Died of Disease}\ \}\ \text{and}$$
$$\{\ \textit{Relapse: Not Recurred, Patient Died of Other Causes}\ \}\ ,$$

as only significant for the deceased class, and

$$\{\ \textit{Survival of}\ 49\ \textit{Months, Relapse Free for}\ 31\ \textit{Months}\ \},$$

as significant for the class of survivors.

The Nottingham Prognostic Index (NPI) is an estimate of the survival chance after breast surgery, with low numbers indicating a high chance of survival. Spass discovers that a low NPI, combined with small and early-stage tumors,

$$\{\ \textit{NPI:}\ [1.0, 3.04),\ \textit{Tumor Size:}\ [1, 15),\ \textit{Tumor Stage:}\ [0, 1)\ \}\ ,$$

is associated with survivors, while high values of NPI, together with a cancerous lymph system,

$$\{\ \textit{NPI:}\ [5.06, 7.2),\ \textit{Lymph Nodes Positive:}\ [3, 45)\ \}\ ,$$

are associated with deceased patients. Significant for both classes is the association of radiotherapy and surgery type,

$$\{\ \textit{Surgery: Conserving, Radiotherapy: Yes}\ \}\ ,$$
$$\{\ \textit{Surgery: Mastectomy, Radiotherapy: No}\ \}\ ,$$

which corresponds to clinical practice.

We further discover that cancer cells which do not respond well to hormone therapy, *ER:- (by IHC)*, are typically treated with *Chemotherapy: Yes*. Spass returns two variants of this pattern: one with *Overall Survival* $\leq$ 49 *months* and one with *Inferred Menopausal State: Pre*, both significant for the class of deceased patients. It also discovers a pattern significant for deceased patients that characterizes the situation in which cancer cells that are hard to differentiate from regular cells (i.e., they have a high histological grade) do not respond to hormone therapy,

$$\{\ \textit{ER:- (by IHC), ER:-, PR:-, Neoplasm Hist. Grade:}\ 2\ \}\ .$$

In these cases, hormone therapy tends to fail, surgery is very hard to perform, and hence, patients have low survival rates.

Overall, our experiments on real-world data from different domains therefore demonstrate that Spass not only efficiently discovers non-redundant sets of significant patterns, outperforming even specialized state-of-the-art methods, it also identifies meaningful patterns in practice.

5.2.4 One-Class Real-World Data. Finally, we want to evaluate how well Spass works on one-class data. Methods like Lamp, WYlight, or SPuManTE all require two classes, and are not applicable in this setting. We therefore compare to opus [44], which discovers self-sufficient patterns from data using Fisher's exact test while bounding FWER. Self-sufficient patterns are those with a frequency that is statistically significant compared to the frequencies of all subsets. opus requires the user to set a maximum number k of how many patterns it may report. As we are primarily interested in how well opus filters *redundant* patterns, we set $k = 10\,000$, which is high enough for it to discover any truly significant and non-redundant pattern.

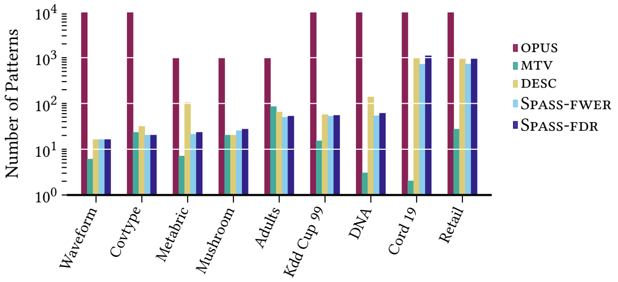


Figure 6: Self-sufficiency is insufficient to discover non-redundant patterns. We show the number of statistically tested-to-be non-redundant discoveries of mtv, desc, opus, and Spass on one-class datasets

By considering one-class data, we are in the application domain of pattern set mining, which strives to discover a non-redundant set of patterns to identify informative feature co-occurrences. We compare to two state-of-the-art methods, mtv [26] and desc [11], that also rely on maximum entropy modeling.

In Fig. 6, we show the number of patterns discovered by mtv, desc, opus, and Spass on 9 one-class datasets. There, we see that opus almost always reports all k patterns as significant, whereas the dedicated pattern *set* miners mtv and desc, as well as Spass, all report similarly concise results. Closer inspection confirms that despite a rigorous FWER control, opus still returns subsets of patterns as significant discoveries. That means, testing for self-sufficiency alone is insufficient to discover a set of non-redundant and significant patterns. We attribute this observation to the fundamental limitation of the self-sufficiency property, which tests each pattern *independently of prior discoveries* and conclude that the usage of past discoveries helps to alleviate redundant results.

6 DISCUSSION

In our experiments, we demonstrated that Spass discovers concise pattern sets and scales well to high-dimensional data. It works well in practice, it still leaves see room for future work.

Our method is essentially a framework which permits to (i) plug in a data-appropriate probabilistic model that dependents on past discoveries; (ii) choose a statistical test; and (iii) select one of the myriad FWER or FDR control techniques [18, 19, 31, 38, 39, 43]. As such, it is easily adaptable: One can simply exchange building blocks to accommodate different types of data, such as graphs, sequences, or continuous data, or to incorporate background knowledge beyond pattern frequencies. Replacing the binomial test with the standard normal approximation, for example, yields the Z-test.

Albeit our sequential FWER control is much less conservative than Bonferroni correction, we see room for increasing the statistical power even further. Recent work, for example, introduces a novel online FWER control [39], which might yield a statistically powerful sequential FWER control. However, since this work still controls for the strict FWER, it will not replace the online FDR control, which could also be improved further. For example, we might overburden our "α-budget" by paying for each test, including tests of hypothesis that have very high p-values, and thus will almost surely never result in discoveries. Therefore, we might as well outright discard (not reject) these hopeless hypotheses [38].

Furthermore, we want to maintain *one* FDR for *all* classes. It is nevertheless straightforward to adapt Spass to maintain *independent* FDR budgets per class. Since in our experiments, we have not noticed a practical performance difference from maintaining independent budgets, we present the slightly simpler version.

7 CONCLUSION

We considered the problem of discovering statistically significant patterns under false discovery control. To avoid redundancy, we proposed to statistically test whether observed frequencies match with expectation, given past discoveries. To achieve high statistical power, we proposed to sequentially control for either FWER or FDR. To efficiently discover significant patterns, we introduced the Spass algorithm that uses an easy-to-compute Chernoff bound to permits efficient significance testing. Through extensive experiments, we demonstrated that our method returns concise result sets, recovers the ground truth from synthetic data, works well on data with many dimensions and any number of classes, and identifies interesting and meaningful patterns in practice, and consistently outperforming the state-of-the-art.

REFERENCES

[1] Charu C. Aggarwal and Jiawei Han. 2004. *Frequent Pattern Mining.* Springer.

[2] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules. In *VLDB*, Vol. 1215. Morgan Kaufmann, 487–499.

[3] Ehud Aharoni and Saharon Rosset. 2013. Generalizedα-investing: definitions, optimality results and application to public databases. *J. R. Stat. Soc. B* 76, 4 (2013), 771–794.

[4] Stephen D. Bay and Michael J. Pazzani. 2001. Detecting Group Differences: Mining Contrast Sets. *Data Min. Knowl. Discov.* 5, 3 (2001), 213–246.

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Royal Stat. S B (Methodological)* 57, 1 (1995), 289–300.

[6] C. E. Bonferroni. 1936. Teoria Statistica Delle Classi e Calcolo Delle Probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), 3–62.

[7] Kailash Budhathoki and Jilles Vreeken. 2015. The Difference and the Norm - Characterising Similarities and Differences Between Databases. In *ECML PKDD (LNCS, Vol. 9285)*. Springer, 206–223.

[8] Herman Chernoff. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. *Ann Math Stat* 23, 4 (1952), 493–507.

[9] Imre Csiszár. 1975. I-Divergence Geometry of Probability Distributions and Minimization Problems. *Ann. Probab.* 3, 1 (1975), 146–158.

[10] Christina Curtis and et.al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 7403 (2012), 346–352.

[11] Sebastian Dalleiger and Jilles Vreeken. 2020. Explainable Data Decompositions. In *AAAI*. 3709–3716.

[12] Sebastian Dalleiger and Jilles Vreeken. 2020. The Relaxed Maximum Entropy Distribution and its Application to Pattern Discovery. In *ICDM*. IEEE, 978–983.

[13] Jonas Fischer and Jilles Vreeken. 2020. Discovering Succinct Pattern Sets Expressing Co-Occurrence and Mutual Exclusivity. In *KDD '20*. ACM, 813–823.

[14] Jaroslav Fowkes and Charles Sutton. 2016. A Bayesian Network Model for Interesting Itemsets. In *ECML PKDD*. Springer, 410–425.

[15] Cristian A Gallo, Rocio L Cecchini, Jessica A Carballido, Sandra Micheletto, and Ignacio Ponzoni. 2016. Discretization of gene expression data revised. *Briefings in bioinformatics* 17, 5 (2016), 758–770.

[16] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. 2007. Assessing Data Mining Results via Swap Randomization. *Trans. Knowl. Discov. Data* 1, 3 (2007), 14.

[17] Wilhelmiina Hämäläinen. 2012. Kingfisher: An Efficient Algorithm for Searching for Both Positive and Negative Dependency Rules with Statistical Significance Measures. *Knowl Inf Syst* 32, 2 (2012), 383–414.

[18] Adel Javanmard and Andrea Montanari. 2015. On Online Control of False Discovery Rate. *CoRR* (2015). arXiv:1502.06197

[19] Adel Javanmard and Andrea Montanari. 2018. Online rules for control of false discovery rate and false discovery exceedance. *Ann. Statist.* 46, 2 (2018), 526 – 554.

[20] E.T. Jaynes. 1982. On the Rationale of Maximum-Entropy Methods. *IEEE* 70, 9 (1982), 939–952.

[21] Adam Kirsch, Michael Mitzenmacher, Andrea Pietracaprina, Geppino Pucci, Eli Upfal, and Fabio Vandin. 2012. An Efficient Rigorous Approach for Identifying Statistically Significant Frequent Itemsets. *J. ACM* 59, 3, Article 12 (jun 2012).

[22] Felipe Llinares-López, Laetitia Papaxanthos, Dean Bodenham, Damian Roqueiro, and Karsten Borgwardt. 2017. Genome-Wide Genetic Heterogeneity Discovery with Categorical Covariates. *Bioinformatics* 33, 12 (2017), 1820–1828.

[23] Felipe Llinares-López, Laetitia Papaxanthos, Damian Roqueiro, Dean Bodenham, and Karsten Borgwardt. 2019. CASMAP: Detection of Statistically Significant Combinations of SNPs in Association Mapping. *Bioinformatics* 35, 15 (2019), 2680–2682.

[24] Felipe Llinares-López, Mahito Sugiyama, Laetitia Papaxanthos, and Karsten Borgwardt. 2015. Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing. In *KDD*. ACM, 725–734.

[25] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *AMACL*. Association for Computational Linguistics, 142–150.

[26] Michael Mampaey, Jilles Vreeken, and Nikolaj Tatti. 2012. Summarizing Data Succinctly with the Most Informative Itemsets. *TKDD* 6, 4 (2012), 16.

[27] Shin-ichi Minato, Takeaki Uno, Koji Tsuda, Aika Terada, and Jun Sese. 2014. A Fast Method of Statistical Assessment for Combinatorial Hypotheses Based on Frequent Itemset Enumeration. In *Mach Learn. Know Disc. Data*. Springer, 422–436.

[28] Laetitia Papaxanthos, Felipe Llinares-López, Dean Bodenham, and Karsten Borgwardt. 2016. Finding Significant Combinations of Features in the Presence of Categorical Covariates. In *NeurIPS*. 2279–2287.

[29] Leonardo Pellegrina, Matteo Riondato, and Fabio Vandin. 2019. SPuManTE: Significant Pattern Mining with Unconditional Testing. In *KDD*. ACM, 1528–1538.

[30] Leonardo Pellegrina and Fabio Vandin. 2018. Efficient Mining of the Most Significant Patterns with Permutation Testing. In *KDD*. ACM, 2070–2079.

[31] Aaditya Ramdas, Tijana Zrnic, Martin J. Wainwright, and Michael I. Jordan. 2018. SAFFRON: an Adaptive Algorithm for Online Control of the False Discovery Rate. In *ICML*, Vol. 80. PMLR, 4283–4291.

[32] Raissa T. Relator, Aika Terada, and Jun Sese. 2018. Identifying Statistically Significant Combinatorial Markers for Survival Analysis. *BMC Med. Genomics* 11, 2 (2018), 31.

[33] Mahito Sugiyama and Karsten M. Borgwardt. 2019. Finding Statistically Significant Interactions between Continuous Features. In *IJCAI*. 3490–3498.

[34] M. Sugiyama, F. López, N. Kasenburg, and K. Borgwardt. 2015. Significant Subgraph Mining with Multiple Testing Correction. In *SDM (Proceedings)*. Society for Industrial and Applied Mathematics, 37–45.

[35] R. E. Tarone. 1990. A Modified Bonferroni Method for Discrete Data. *Biometrics* 46, 2 (1990), 515.

[36] Nikolaj Tatti. 2006. Computational Complexity of Queries Based on Itemsets. *Inform. Process. Lett.* 98, 5 (2006), 183–187.

[37] Aika Terada, Koji Tsuda, and Jun Sese. 2013. Fast Westfall-Young permutation procedure for combinatorial regulation discovery. In *BIBM*. IEEE, 153–158.

[38] Jinjin Tian and Aaditya Ramdas. 2019. ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. In *NeurIPS*. 9383–9391.

[39] Jinjin Tian and Aaditya Ramdas. 2021. Online control of the familywise error rate. *Stat. Meth. Med. R.* 30, 4 (2021), 976–993.

[40] Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. 2011. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *J. Comput. Biol.* 18, 3 (2011), 507–522.

[41] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. 2011. Krimp: mining itemsets that compress. *Data Min. Knowl. Discov.* 23, 1 (2011), 169–214.

[42] Geoffrey I. Webb. 2008. Layered Critical Values: A Powerful Direct-Adjustment Approach to Discovering Significant Patterns. *Mach. Learn.* 71, 2-3 (2008), 307–323.

[43] Geoffrey I. Webb and François Petitjean. 2016. A Multiple Test Correction for Streams and Cascades of Statistical Hypothesis Tests. In *KDD*. ACM, 1255–1264.

[44] Geoffrey I. Webb and Jilles Vreeken. 2013. Efficient Discovery of the Most Interesting Associations. *ACM Trans. Knowl. Discov. Data* 8, 3 (2013), 15:1–15:31.

[45] Qingrun Zhang, Quan Long, and Jurg Ott. 2014. AprioriGWAS, a New Pattern Mining Strategy for Detecting Genetic Variants Associated with Disease through Interaction Effects. *PLoS* 10, 6 (2014), 14.

A APPENDIX

A.1 Reproducibility

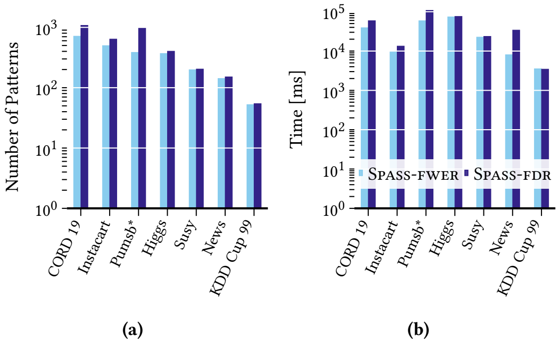
In our experiments, we compare to the statistically significant pattern miners Lamp [37], WYlight [24] and SPuManTE [29]. For SPuManTE and WYlight we use the original implementations provided by the respective authors. To limit the impact of implementation quality, we use the C++ implementation of Lamp by Llinares-López et al. [24]. In general, we use the hyperparameters suggested by the respective authors. For SPuManTE, we choose a maximum sample size of 10 000, and for WYlight we set the number of permutations to 10 000. We use the implementations of desc [11], mtv [26], opus [44], SPuManTE [29], and WYlight [24] by the respective authors, and the C++ implementation of Lamp by Llinares-López et al. [24]. We make our code, datasets, and synthetic data generator available.[4]

Dataset	$\lvert X \rvert$	dim X	Avg. Row	Density	k
Higgs	11 000 000	247	28.00 ± 0.00	0.1134	2
SUSY	5 000 000	178	18.00 ± 0.00	0.1011	2
Instacart	2 620 570	1 235	3.14 ± 2.18	0.0025	1
KDD Cup 99	1 000 000	135	16.00 ± 0.00	0.1185	1
Covtype	581 012	64	11.95 ± 0.23	0.1866	2
RNA	271 617	16	8.00 ± 0.00	0.5000	2
News	127 600	11 489	13.63 ± 4.05	0.0012	4
IJCNN	91 701	34	13.00 ± 0.00	0.3824	2
IMDb	49 969	8 125	63.95 ± 42.56	0.0079	2
Pumsb*	49 046	2 088	50.48 ± 1.98	0.0242	1
CORD-19	32 907	2 648	47.63 ± 23.87	0.0180	1
Adults	32 561	123	13.87 ± 0.48	0.1128	2
Mushroom	8 124	117	22.00 ± 0.00	0.1880	2
Breast Cancer	7 325	397	11.67 ± 13.06	0.0294	2
Metabric	1 981	124	32.32 ± 1.03	0.2606	2
Breast	1 218	20 530	3 036.89 ± 359.03	0.1863	1
Lung	1 129	20 530	3 378.75 ± 318.66	0.2043	2
Kidney	1 020	20 530	3 325.43 ± 242.96	0.2097	3
Kidney Clear	606	20 530	3 496.35 ± 371.08	0.2291	1
Lung Adeno.	576	20 530	3 053.31 ± 347.88	0.1932	1
Lung Squamous	553	20 530	3 146.87 ± 333.37	0.1972	1
Brain	530	20 530	3 099.68 ± 371.75	0.2146	1
Endo & Ovo	509	20 530	3 681.89 ± 290.47	0.2303	2
Ovarian	308	20 530	3 063.36 ± 307.32	0.2025	1
Uterine	57	20 530	3 224.40 ± 274.13	0.2253	1

Table 1: We show the number of data points, number of features, the average number of 1s per row, the overall density, and the number of classes *k* of the datasets used in our experiments.

A.2 Datasets

All datasets that we have used in our experiments are publicly available. We have taken the genomics data in Fig. 5 (c) from *The Cancer Genome Atlas Program*,[5] and binarized this dataset using a specialized method for gene-expression data [15]. We have taken *Mushroom*, and *Pumsbstar* from the Itemset Mining Dataset Repository.[6] The *AG News* dataset consists of news articles from 4 categories[7] and the *CORD 19* dataset consists of abstracts from the CORD 19 open research dataset.[8] We lemmatized the *AG News*, *CORD 19*, and *IMDb*[9], *ArXiv*, datasets and removed stop words and words with a frequency of below 0.1%. All the remaining datasets are from the UCI Machine Learning Repository[10] or from the LIBSVM repository.[11] To reduce the number of features of the *Instacart* dataset, we have combined products from the same category, e.g., we merged Spumante with Cremant to achieve the Champagne meta category.[12] We binarized each real valued feature by binning it into 5 bins of equal width, and we mapped each categorical and ordinal attribute to multiple binary features, which is often referred to as "one-to-*k*" (also called "one-hot") encoding. In Table 1, we provide basic statistics for the processed data.



(a)

(b)

Figure 7: Our method efficiently discovers concise pattern sets from large datasets. We show the number of significant discoveries (a) and runtime needed (b) of Spass-fwer and Spass-fdr.

A.3 Large Real-World Data

As a next step, we want to examine the scalability of Spass for datasets with a very large number of data points. For this, we run SPuManTE, Lamp, WYlight, and Spass on 7 large real-world datasets. However, since SPuManTE, Lamp, and WYlight could not handle these datasets, we turned to Spass. According to Figure 7, Spass-fwer deems fewer patterns as significant as Spass-fdr. This is reflected in the shorter runtime under FWER control. While competitors have not finished in 2 days, Spass only took 2 minutes.

[4] eda.mmci.uni-saarland.de/spass

[5] cancer.gov/tcga

[6] fimi.ua.ac.be/data

[7] di.unipi.it/~gulli/AG_corpus_of_news_articles

[8] allenai.org/data/cord-19

[9] ai.stanford.edu/~amaas/data/sentiment

[10] archive.ics.uci.edu/ml

[11] csie.ntu.edu.tw/~cjlin/libsvmtools/datasets

[12] instacart.com/datasets/grocery-shopping-2017