MiSoSouP: Mining Interesting Subgroups with Sampling and Pseudodimension*

MATTEO RIONDATO†, Amherst College, USA
FABIO VANDIN, Università di Padova, Italy

We present MiSoSouP, a suite of algorithms for extracting high-quality approximations of the most interesting subgroups, according to different popular interestingness measures, from a random sample of a transactional dataset. We describe a new formulation of these measures as functions of averages, that makes it possible to approximate them using sampling. We then discuss how pseudodimension, a key concept from statistical learning theory, relates to the sample size needed to obtain an high-quality approximation of the most interesting subgroups. We prove an upper bound on the pseudodimension of the problem at hand, which depends on characteristic quantities of the dataset and of the language of patterns of interest. This upper bound then leads to small sample sizes. Our evaluation on real datasets shows that MiSoSouP outperforms state-of-the-art algorithms offering the same guarantees, and it vastly speeds up the discovery of subgroups w.r.t. analyzing the whole dataset.

CCS Concepts: • **Mathematics of computing → Probabilistic algorithms**; • **Information systems → Data mining**; • **Theory of computation → Sketching and sampling**; **Sample complexity and generalization bounds**.

Additional Key Words and Phrases: Pattern Mining, Statistical Learning Theory

ACM Reference Format:
Matteo Riondato and Fabio Vandin. 2020. MiSoSouP: Mining Interesting Subgroups with Sampling and Pseudodimension. *ACM Trans. Knowl. Discov. Data.* 14, 5, Article 56 (June 2020), 30 pages. https://doi.org/10.1145/3385653

"Miso makes a soup loaded with flavour that saves you the hassle of making stock." – Y. Ottolenghi [21]

1 INTRODUCTION

A fundamental task within data mining is *subgroup discovery* [11, 13, 37], which requires to identify *interesting subsets* (the subgroups) of a dataset, for which the distribution of a specific feature (the *target*) within the subgroup largely differs from the distribution of that feature in the entire dataset. The notion of *interestingness* is captured by a formally-defined measure of *quality* that combines the frequency of the subgroup in the dataset and the difference between the mean of the target within the subgroup and the mean of the target in the entire dataset. Subgroup discovery is a broadly applicable task and is relevant in many domains: in market basket analysis, it uncovers groups of customers with a particular interest in buying a product; in social networks, it identifies members attracted to a given topic; in biomedicine, it discovers groups of patients associated with a clinical phenotype, such as response to therapy.

*A preliminary version of this work appeared in the proceedings of ACM KDD'18 as [29].
†Part of the work done while affiliated to Two Sigma Labs.

Authors' addresses: Matteo Riondato, Department of Computer Science, Amherst College, 25 East Drive, Amherst, MA, 01002, USA, mriondato@amherst.edu; Fabio Vandin, Department of Information Engineering, Università di Padova, Via G. Gradenigo 6/B, Padova, IT-35131, Italy, fabio.vandin@unipd.it.

Many exact algorithms for subgroup discovery have been proposed [13, 37] (see also the comprehensive reviews by Herrera et al. [8] and Atzmueller [2]). They naturally require to process the entire dataset, but the sheer amount of data may render such full computation infeasible. A general approach to deal with very large datasets is to only analyze a *small random sample* of the data. Random sampling has been successful in many areas of knowledge discovery, such as frequent itemsets mining [25, 26] and graph analysis [27]. The main challenge in using sampling for subgroup discovery is understanding how close the qualities of the subgroups observed in the sample are to their exact values, which are unknown as they can only be obtained by processing the entire dataset. Solving this challenge requires the derivation of a sample size S such that, with high probability (over the possible samples), on a sample of size S, all the subgroup qualities measured on the sample are within ε from their value in the dataset, where ε is an user-specified parameter, controlling the maximum allowed error and to be fixed with domain knowledge.

The derivation of such sample size for subgroup discovery is more complex than in other scenarios such as frequent itemsets mining [25, 26], since estimating the quality of a subgroup requires to approximate both the frequency of the subgroup in the dataset and the mean of the target *within the subgroup*. The latter is an especially challenging inferential task since it amounts to estimating a *conditional* expectation. Additionally, there are many measures of interestingness for subgroups, and each needs a "personalized" approach. This increased complexity is reflected in the lack of rigorous sampling algorithms for subgroup discovery, with even popular approaches [30] not providing rigorous quality guarantees on their output, as we discuss in App. A.

1.1 Contributions

The main focus of this work is the extraction of a high-quality approximation of the top-k most interesting subgroups from a random sample of the dataset. Our contributions are the following.

- We precisely define the concept of *ε-approximation* of the set of top-k subgroups according to many popular interestingness measures, extending and strengthening an existing definition by Scheffer and Wrobel [30, Def. 2]. The user-defined parameter ε controls the quality of the approximation (see Def. 3.3).
- We give a *new formulation of the 1-quality* (see Sect. 4.1), one of the key measures of subgroup interestingness, and as a consequence also of other fundamental measures. This novel formulation is crucial to enable the estimation of the interestingness of subgroups from a sample. It is also at the basis of our approach to compute high-quality approximation for the collection of interesting subgroups according to other measures, such as 2- and 1/2-qualities.
- We present MiSoSouP, a suite of algorithms that use *random sampling* to extract, with probability at least $1 - \delta$ (for some user-specified $\delta \in (0, 1)$, which controls the confidence in the results) over their runs, ε-approximations of the set of top-k interesting subgroups from a small random sample of the dataset (see Sect. 4). We present specialized algorithms for different interestigness measures, showcasing the generality and the power of our approach. Ours are the first algorithms able to obtain such approximations, while previous work [30] does not actually provide rigorous guarantees (see App. A). The only parameters of MiSoSouP are ε, k, and the failure probability δ, which are all easily interpretable, therefore making MiSoSouP algorithms very practical.
- We use *pseudodimension* [23], a key concept from statistical learning theory [35] (see Sect. 3.2), to derive the sample sizes employed in MiSoSouP. We show an upper bound to the pseudodimension of the task of subgroup discovery (see Sect. 4.1.4). This bound is independent from the size of the dataset and only depends on properties of the set of possible subgroups (known as the *language*) and on the number of columns of the dataset. The computation of

the upper bound is essentially cost-free. We also show an almost matching lower bound (see Lemma 4.11). To the best of our knowledge, ours is the first application of pseudodimension to the field of subgroup discovery, and in general to pattern mining.
- We perform an extensive experimental evaluation (see Sect. 5) showing that MiSoSouP identifies rigorous approximations to the most interesting subgroups using a small fraction of the dataset, and it provides a significant speed-up w.r.t. other sampling approaches with the same guarantees.

2 RELATED WORK

Many measures for evaluating the quality (i.e., interestingness) of subgroups have been proposed in the literature, and many subgroup discovery algorithms are available. We discuss some of the measures in Sect. 3, and refer the reader to the surveys by Herrera et al. [8] and Atzmueller [2] for details about the algorithms. In this work we treat these algorithms as black-boxes: we run them on a small random sample of the dataset and we are interested in how well the so-obtained collection of interesting subgroups approximates the one we would obtain by mining the whole dataset.

Scheffer and Wrobel [30] first studied the use of sampling for subgroups: they present GSS, a progressive sampling algorithm to compute an approximation of the most interesting subgroups. Unfortunately, the analysis of GSS has some issues. The first concern is that the quantities of interest (e.g., the number of subgroups at iteration i) are *random* variables, while the analysis assumes that they are *fixed* values, i.e., it is essentially conditioning on the outcome. Another major issue is that the analysis uses a Chernoff bound [19, Ch. 4] for the probability of deviation for the in-sample *unusualness* of a subgroup from its expectation (i.e., the unusualness of that subgroup in the whole dataset), but applying a Chernoff bound is improper, since the unusualness is a *conditional* probability, hence it cannot be obtained as the average of a binary function over *all* transactions in the sample. These and other issues are discussed more in depth together with partial possible solutions in App. A. Even when (partially) corrected, the analysis of GSS relies on the availability of probabilistic confidence intervals on the estimated quality of each subgroup under consideration, and then on a union bound [19, Lemma 1.2]over all possible subgroups, in order to obtain simultaneous guarantees on the confidence intervals of all subgroups. The union bound is, by design, loose in many practical situations, effectively assuming that the considered events are independent. As a results, the stopping condition used by GSS cannot be satisfied at small sample sizes. MiSoSouP instead relies on pseudodimension [23], which allows us to use very small sample sizes.

Some works focused on the issue of the statistical significance of subgroups. Duivesteijn et al. [6] designed a permutation-based approach to estimate the distribution of false discoveries, which is used to assess the ability of various quality measures to distinguish between statistically significant patterns and false discoveries. van Leeuwen and Ukkonen [34] showed that several real datasets contain large numbers of high-quality subgroups, many more than are expected from randomly drawn subgroups. Terada et al. [32] introduced LAMP, a method to identify a minimum generality threshold to find subgroups while bounding the family-wise error rate (FWER), where the significance of a subgroup is given by its association with a binary target variable as assessed by Fisher's exact test. Minato et al. [18] subsequently improved LAMP by employing a more efficient mining strategy. We do not investigate the issue of statistical significance of subgroups, but one of the quality measures we study (i.e., the 1/2-quality measure, see Sect. 3) is a proxy for the z-score, a well-defined measure of statistical significance.

Our approach is orthogonal to heuristic approaches that sample *subgroups* to speed-up the discovery of interesting subgroups [4, 20]. In contrast, MiSoSouP samples *transactions* while providing rigorous guarantees on the relation between the qualities of the subgroups obtained from

the sample and their exact qualities, i.e., those one would measure on the entire dataset. MiSoSouP can use any exact or heuristic algorithm to mine the sample, while maintaining the aforementioned guarantees for the resulting subgroups. The use of sampling is also orthogonal to techniques that aim at reducing the redundancy in the output collection of subgroups [33]. Indeed these approaches could be applied to the collection of subgroups obtained by MiSoSouP.

Pseudodimension [23] is a key concept from statistical learning theory [35]. Like many other measures of sample complexity, such as Rademacher averages, it has long been considered only of theoretical interest, but recent applications [7, 10, 24, 26, 27] of these quantities have shown that they can be extremely useful in practice, especially on very large datasets. Pseudodimension is closely related to the concept of Vapnik-Chervonenkis dimension that has been used in the context of frequent itemsets mining by Riondato and Upfal [25]. Despite the relative similarity between subgroup discovery and frequent itemset mining, using pseudodimension for the former presents significant challenges, such as lack of anti-monotonicity in the quality measures, that do not allow to use the same approach by Riondato and Upfal [25]. To the best of our knowledge, ours is the first application of concepts from statistical learning theory to the task of subgroup discovery.

This version of the work differs from the preliminary one that appeared in the proceedings of ACM KDD'18 [29]. The most important addition is the presentation of algorithms for all three quality measures that we study, rather than just for one. This change shows the flexibility of our approach. The second major change is the inclusion of additional experiments for these measures. A third major change is that we show an almost matching (up to the *additive* constant 1) lower bound for our upper bound to the pseudodimension of the task at hand. We also added running examples to all our definitions to make them more concrete.

3 PRELIMINARIES

In this section we introduce the core definitions and theorems that we use throughout the article. The main notation that we use is reported also in Table 1.

3.1 Subgroup discovery

We now define the fundamental concepts of subgroup mining [12] and the quality measures used to rank the subgroups.

Let $\mathcal{D}$ be a *dataset*, i.e., a bag of $(z+1)$-dimensional tuples, known as *transactions*, over the attributes $\{A_1, \ldots, A_z, T\}$. The attributes A_i, $1 \leq i \leq z$ are known as *description* attributes, while T is the *target* attribute. Transactions take value in $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_z \times \mathcal{Y}_T$, where each $\mathcal{Y}_i$ is the (categorical or numerical) domain of attribute A_i, while $\mathcal{Y}_T$ is Boolean (i.e., $\mathcal{Y}_T = \{0, 1\}$). Table 2 shows an example of a dataset with four transactions over four attributes.

A *subgroup* is a *conjunction* of disjunctions of conditions on the description attributes. An example of subgroup is

$$(A_1 = \text{“blue”} \vee A_1 = \text{“red”}) \wedge (A_2 \geq 4) \ . \tag{1}$$

A transaction $t \in \mathcal{D}$ *supports* a subgroup Y if the values of t's attributes satisfy Y. The *cover* $\mathsf{C}_{\mathcal{D}}(Y)$ *of* Y *on* $\mathcal{D}$ is the bag of transactions in $\mathcal{D}$ that support Y. For example, the first two transactions in the example dataset in Table 2 support the subgroup defined in (1), while the other two transactions do not support it, so the cover of this subgroup is the set containing only the first two transactions.

Symbol	Description
$\mathcal{D}$	dataset
A_i	i-th description attribute in the dataset
T	target attribute
Y	generic subgroup
$\mathcal{L}$	language of subgroups of interest
$\mathsf{C}_{\mathcal{D}}(Y)$	cover of Y on $\mathcal{D}$
$g_{\mathcal{D}}(Y)$	generality of Y on $\mathcal{D}$
$\mu_{\mathcal{D}}(Y), \mu(\mathcal{D})$	target mean of the cover of Y on $\mathcal{D}$, and target mean of $\mathcal{D}$, respectively
$\mathsf{u}_{\mathcal{D}}(Y)$	unusualness of Y on $\mathcal{D}$
$\mathsf{q}^{(p)}_{\mathcal{D}}(Y)$	p-quality of Y on $\mathcal{D}$
$\mathsf{r}^{(p)}_{\mathcal{D}}(k)$	p-quality on $\mathcal{D}$ of the top-k-th subgroup
$\mathcal{L}_{\mathcal{D}}$	subset of $\mathcal{L}$ of the subgroups actually appearing in $\mathcal{D}$
$\mathcal{F}$	family of functions from a domain $\mathcal{H}$ to $[a, b] \subset \mathbb{R}$
$\mathsf{PD}(\mathcal{F})$	pseudodimension of $\mathcal{F}$
$\mathsf{m}_Z(f)$	mean of the function f on a set Z
$\mathcal{S}$	uniform, independent sample of transactions from $\mathcal{D}$
$\tilde{\mathsf{q}}^{(p)}_{\mathcal{S}}(Y)$	approximate p-quality of Y on $\mathcal{S}$
$\tilde{\mathsf{r}}^{(p)}_{\mathcal{S}}(k)$	approximate p-quality on $\mathcal{S}$ of the top-k-th subgroup (w.r.t. this measure)
$\mathcal{P}$	family of functions associated to $\mathcal{L}$ on $\mathcal{D}$

Table 1. Main notation used in this work

A_1	A_2	A_3	T
blue	4	circle	1
red	7	square	0
blue	3	square	1
green	2	square	1

Table 2. Example dataset

The *generality* $\mathsf{g}_{\mathcal{D}}(Y)$ *of a subgroup* Y *on* $\mathcal{D}$ is the ratio between the size of the cover of Y on $\mathcal{D}$ and the size of $\mathcal{D}$:

$$\mathsf{g}_{\mathcal{D}}(Y) = \frac{|\mathsf{C}_{\mathcal{D}}(Y)|}{|\mathcal{D}|} \ .^{1}$$

For example, the subgroup from (1) has generality $1/2$ on the dataset in Table 2.

Given a bag $\mathcal{B}$ of transactions, let

$$\mu(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} t.T$$

be the *target mean* of $\mathcal{B}$, where $t.T$ denotes the value in the target attribute of the tuple t. If $\mathcal{B} = \emptyset$, $\mu(\mathcal{B}) = 0$. The *target mean of a subgroup* Y *on* $\mathcal{D}$ is

$$\mu_{\mathcal{D}}(Y) = \mu(\mathsf{C}_{\mathcal{D}}(Y)) \ .$$

[1]We use $|B|$ to denote the size of a bag B, i.e., the number of elements in B, counting repeated elements multiple times.

The *unusualness*[2] $\mathsf{u}_{\mathcal{D}}(Y)$ *of* Y *on* $\mathcal{D}$ is the difference between the target mean of Y and the target mean of $\mathcal{D}$:

$$\mathsf{u}_{\mathcal{D}}(Y) = \mu_{\mathcal{D}}(Y) - \mu(\mathcal{D}) \ .$$

The unusualness of the subgroup from (1) on the dataset in Table 2 is

$$\frac{1}{2} - \frac{3}{4} = -\frac{1}{4} \ .$$

The generality and the unusualness are used to define quality measures for the subgroups (see Sect. 3.1.1).

A *description language* $\mathcal{L}$ is a set of subgroups that are of potential interest, and is *fixed in advance* by the user before analyzing the dataset. It could be a superset or a subset of the subgroups that actually appear in the dataset, and it expresses the constraint that only subgroups in the description language should be considered in the mining process. For example, given some integer m, one may consider the description language of all and only the subgroups composed of up to m conjunctions of equality conditions on the attributes. Another example is the language of subgroups composed of up to two conjunctions of disjunctions of no more than two conditions. The subgroup in (1) belongs to this language.

3.1.1 Quality measures. A *quality measure* for the subgroups in $\mathcal{L}$ on a dataset $\mathcal{D}$ is a function $\phi_{\mathcal{D}} : \mathcal{L} \to \mathbb{R}$ which assigns a numerical score to each subgroup $Y \in \mathcal{L}$ based on its generality and unusualness. In this work we consider the most popular subgroup quality measures [13], which differ from each other for the relative weight given to generality and unusualness.

Definition 3.1 ([11, 22, 37]). Let $p \in \{1/2, 1, 2\}$. The *p-quality of a subgroup Y on a dataset $\mathcal{D}$* is

$$\mathsf{q}^{(p)}_{\mathcal{D}}(Y) = (\mathsf{g}_{\mathcal{D}}(Y))^p \, \mathsf{u}_{\mathcal{D}}(Y) \ .$$

The subgroup in (1) has 1-quality equal to $-1/8$ on the dataset from Table 2.

The 1-quality is also known as *Weighted Relative Accuracy* (WRAcc).[3] The $1/2$-quality is proportional to the z-score[4] for the statistic $|\mathsf{C}_{\mathcal{D}}(Y)|\mu_{\mathcal{D}}(Y)$, which can be used to test whether a subgroup shows statistical association with the target variable. Thus, the $1/2$-quality can be used as a proxy for the statistical significance of subgroup Y [11, 30, 34]. The domain $\mathcal{Y}_T$ of the target attribute is Boolean, thus $\mathsf{q}^{(p)}_{\mathcal{D}}(Y) \in [-1, 1]$ for any subgroup Y. There exist variants of the p-qualities that consider the *absolute value* of the unusualness [30]. MiSoSouP can be easily adapted to work with such measures.

3.1.2 Subgroup discovery task. Fix $p \in \{1/2, 1, 2\}$. Let $\mathcal{L}_{\mathcal{D}}$ be the subset of $\mathcal{L}$ containing only the subgroups of $\mathcal{L}$ that actually appear in $\mathcal{D}$, i.e., those with generality strictly greater than zero. We do not assume to know $\mathcal{L}_{\mathcal{D}}$: it is only needed for the following definition. Assume to sort the subgroups in $\mathcal{L}_{\mathcal{D}}$ in decreasing order according to their p-quality in $\mathcal{D}$, ties broken arbitrarily. Let $k > 0$ be an integer and let $\mathsf{r}^{(p)}_{\mathcal{D}}(k)$ be the p-quality of the k-th subgroup in the sorted order.

[2] Scheffer and Wrobel [30] use the term *statistical unusualness*. We choose to drop the adjective to avoid confusion with *statistical significance*.

[3] van Leeuwen and Ukkonen [34] denote the $1/2$-quality as "WRAcc", but all other references we found (e.g., [8, 15]) use this name to denote the 1-quality.

[4] The z-score for a test statistic X is $(X - \mathbb{E}[X])/\sigma_X$, where $\mathbb{E}[X]$ is the expectation and σ_X is the standard deviation of X under the null hypothesis. For subgroups, under the null hypothesis of no association of a subgroup Y with the target, the z-score of $|\mathsf{C}_{\mathcal{D}}(Y)|\mathsf{m}_{\mathcal{D}}(Y)$ is:

$$\frac{|\mathsf{C}_{\mathcal{D}}(Y)|\mu_{\mathcal{D}}(Y) - |\mathsf{C}_{\mathcal{D}}(Y)|\mu(\mathcal{D})}{\sqrt{\mathsf{C}_{\mathcal{D}}(Y)\mu(\mathcal{D})(1-\mu(\mathcal{D}))}} = \frac{(\mathsf{C}_{\mathcal{D}}(Y))^{\frac{1}{2}}\mathsf{u}_{\mathcal{D}}(Y)}{\sqrt{\mu(\mathcal{D})(1-\mu(\mathcal{D}))}} = \mathsf{q}^{(1/2)}_{\mathcal{D}}(Y)\sqrt{\frac{|\mathcal{D}|}{\mu(\mathcal{D})(1-\mu(\mathcal{D}))}} \ .$$

Definition 3.2. The *subgroup discovery* task consists in extracting the set $\mathsf{TOP}_p(k, \mathcal{D})$ of the *top-k subgroups in $\mathcal{L}_\mathcal{D}$ w.r.t. the p-quality in $\mathcal{D}$* , i.e., the set of subgroups with p-quality at least $\mathsf{r}^{(p)}_\mathcal{D}(k)$:

$$\mathsf{TOP}_p(k, \mathcal{D}) = \left\{ Y \in \mathcal{L}_\mathcal{D} \ : \ \mathsf{q}^{(p)}_\mathcal{D}(Y) \geq \mathsf{r}^{(p)}_\mathcal{D}(k) \right\} \ .$$

$\mathsf{TOP}_p(k, \mathcal{D})$ may contain more than k elements when many subgroups have p-quality equal to $\mathsf{r}^{(p)}_\mathcal{D}(k)$.[5]

A variant of the task allows the user to specify a constraint on the minimum generality of returned subgroups. MiSoSouP can handle this case with minor modifications.

3.1.3 Approximations. We want to obtain an *ε-approximation to the set* $\mathsf{TOP}_p(k, \mathcal{D})$ from a small *random sample* of the dataset, where $\varepsilon \in (0, 1)$ is an user-defined parameter that controls the maximum acceptable error. Formally this concept is defined as follows.

Definition 3.3. Let $\varepsilon \in (0, 1)$. An ε-approximation to $\mathsf{TOP}_p(k, \mathcal{D})$ is a set $\mathcal{B}$ of pairs (Y, q_Y) where Y is a subgroup and q_Y is a value in $[-1, 1]$, and $\mathcal{B}$ is such that:

(1) for any $Y \in \mathsf{TOP}_p(k, \mathcal{D})$, there is a pair $(Y, q_Y) \in \mathcal{B}$; and
(2) there is no pair $(Y, q_Y) \in \mathcal{B}$ such that $\mathsf{q}^{(p)}_\mathcal{D}(Y) < \mathsf{r}^{(p)}_\mathcal{D}(k) - \varepsilon$; and
(3) for each pair $(Y, q_Y) \in \mathcal{B}$, $|\mathsf{q}^{(p)}_\mathcal{D}(Y) - q_Y| \leq \varepsilon/4$.

MiSoSouP computes (with high probability) an ε-approximation from a random sample of the dataset. For the $1/2$-quality, we define slightly different conditions for an ε-approximation (presented in Sect. 4.3).

An ε-approximation can act as a set of candidates for $\mathsf{TOP}_p(k, \mathcal{D})$, as it contains a pair (Y, q_Y) for each subgroup Y in this set. Scheffer and Wrobel [30, Definition 2] present a slightly different definition of approximation. Such an approximation is *not* a set of candidates for $\mathsf{TOP}_p(k, \mathcal{D})$, and in particular its intersection with this set may be empty. On the other hand, if we sort the pairs in an ε-approximation by decreasing order of their second component, ties broken arbitrarily, the set of the subgroups in the first k pairs according to this order is an approximation in the sense defined by Scheffer and Wrobel [30]. The choice of ε, similarly to the choice of k in Def. 3.2, must be informed, at least in part, by domain knowledge. For example, when $p = 1$, one may want to set ε so that $\varepsilon \leq \mathsf{r}^{(1)}_\mathcal{D}(k) + \mu(\mathcal{D})$ in order to avoid a trivial approximation containing all subgroups, and the condition can be verified after obtaining the approximation. In addition, the quantity $1 - \mu(\mathcal{D})$ can act, in some sense, as an upper bound to the possible choice of ε, as no subgroup can have 1-quality greater than this quantity.

3.2 Pseudodimension

We now introduce the main concepts and results on *VC-dimension* [36] and *pseudodimension* [23], specializing some of them to our settings.[6]

3.2.1 VC-dimension. Let $\mathcal{W}$ be a finite domain and let $\mathcal{R} \subseteq 2^\mathcal{W}$ be a collection of subsets of $\mathcal{W}$, where $2^\mathcal{W}$ is the set of all subsets of $\mathcal{W}$. We call $\mathcal{R}$ a *rangeset on* $\mathcal{W}$, and call its members *ranges*. The set $A \subseteq \mathcal{W}$ is *shattered* by $\mathcal{R}$ if $\{R \cap A \ : \ R \in \mathcal{R}\} = 2^A$. The *VC-dimension* $\mathsf{VC}(\mathcal{W}, \mathcal{R})$ of $(\mathcal{W}, \mathcal{R})$ is the size of the largest subset of $\mathcal{W}$ that can be shattered by $\mathcal{R}$.

[5] This definition of the task is therefore slightly different from the one given in [30, Definition 1], where the size of $\mathsf{TOP}_p(k, \mathcal{D})$ is limited to exactly k elements.

[6] For an in-depth discussion of these topics see, e.g., the books by Shalev-Shwartz and Ben-David [31] and by Anthony and Bartlett [1].

The following example of VC-dimension is by Riondato and Upfal [27, Sect. 3.3]. Let $D = \mathbb{R}$ and let $\mathcal{R}$ be the collection of closed intervals of $\mathbb{R}$, i.e.,

$$\mathcal{R} = \{[a, b], a < b \in \mathbb{R}\} \ .$$

A set $A = \{c, d\}$ of two distinct reals $c, d \in \mathbb{R}$ can be shattered as follows. W.l.o.g., let $c < d$, and define $g = c + (d - c)/2$, and let h_1 and h_2 be such that $h_1 < h_2 < c$. Consider the ranges $[h_1, h_2]$, $[c, g]$, $[g, d]$, $[c, d]$. Each intersection of each of one of these ranges with A is a different subset of $\{c, d\}$, and for each B of the four subsets of A there is one range R_B of the four above such that $A \cap R_A = B$. Thus $P_{\mathcal{R}}(A) = 2^A$, i.e., the set A is shattered by $\mathcal{R}$.

Consider now a set $C = \{c, d, f\}$ of three different points $c < d < f \in \mathbb{R}$. There is no range $R \in \mathcal{R}$ such that $R \cap C = \{c, f\}$. Indeed all intervals that contain c and f must also contain d. Thus, C cannot be shattered, because it must be $P_{\mathcal{R}}(C) \neq 2^C$. This fact holds for all sets C of three points, so the VC-dimension of $\mathcal{R}$ is $\mathsf{VC}(\mathcal{R}) = 2$.

3.2.2 Pseudodimension. Pseudodimension [23] is an extension of VC-dimension [36] to *real-valued* functions, defined as follows.

Let $\mathcal{F}$ be a family of functions from a domain $\mathcal{H}$ onto $[a, b] \subset \mathbb{R}$. In this work $\mathcal{H}$ will be the dataset $\mathcal{D}$, and $\mathcal{F}$ will contain one function f_Y for each subgroup $Y \in \mathcal{L}$ (see Sect. 4.1.1). Consider, for each $f \in \mathcal{F}$, the subset R_f of $\mathcal{H} \times [a, b]$ defined as

$$R_f = \{(x, t) \ : \ t \leq f(x)\} \ .$$

Let

$$\mathcal{F}^+ = \{R_f, f \in \mathcal{F}\},$$

be a rangeset on $\mathcal{H} \times [a, b]$. The *pseudodimension* $\mathsf{PD}(\mathcal{F})$ *of* $\mathcal{F}$ is the VC-dimension of $(\mathcal{H} \times [a, b], \mathcal{F}^+)$ [1, Sect. 11.2]:

$$\mathsf{PD}(\mathcal{F}) = \mathsf{VC}(\mathcal{H} \times [a, b], \mathcal{F}^+) \ .$$

The following example of pseudodimension is by Riondato and Upfal [27, Sect. 3.3]. Consider the family $\mathcal{F}$ of functions from $(0, 1]$ to $[0, 1]$ defined as

$$\mathcal{F} = \{f_k(x) = kx, \text{ for } 0 < k \leq 1\} \ .$$

The pseudodimension of $\mathcal{F}$ is $\mathsf{PD}(\mathcal{F}) = 1$. For each $f_k \in \mathcal{F}$, i.e., for each $0 < k \leq 1$, the set $R_{f_k} = R_k$ is

$$R_k = \{(x, y), 0 \leq x \leq 1 \text{ and } y \leq kx\} \ .$$

It is a useful exercise to check how to shatter a set containing a single point (x, y), $0 \leq x, y \leq 1$.

To show that $\mathsf{PD}(\mathcal{F}) = 1$ we need to show that no set A of two pairs (x_1, y_1) and (x_2, y_2) can be shattered by $\mathcal{F}^+$. First of all, notice that it must be $y_1 \leq x_1$ and $y_2 \leq x_2$ because there is no range R_k that contains (x, y) if $y > x$. Assume now w.l.o.g. that $x_1 \leq x_2$. If $y_1 > y_2$, then there is no $k \in [0, 1]$ such that $kx_1 \geq y_1$ and $kx_2 < y_2$, thus there is no range R_k such that $A \cap R_k = \{(x_1, y_1)\}$. If instead $y_1 \leq y_2$, then let $z = y_2/x_2$. We have to consider two sub-cases:

(1) if $y_1 > zx_1$, then there is no $k \in [0, 1]$ such that $kx_1 \geq y_1$ and $kx_2 < y_2$, thus there is no range R_k such that $A \cap R_k = \{(x_1, y_1)\}$. To see this, assume that such a k exists. Then it would hold that $k > z$ because $kx_1 \geq y_1 > zx_1$, thus $kx_2 > zx_2 = y_2$, which is a contradiction.
(2) if $y_1 \leq zx_2$, then there is no $k \in [0, 1]$ such that $kx_1 < y_1$ and $kx_2 \geq y_2$, thus there is no range R_k such that $A \cap R_k = \{(x_2, y_2)\}$. To see this, assume that such a k exists. Then it would hold that $k < z$ because $kx_1 < y_1 \leq zx_1$, thus $kx_2 < zx_2 = y_2$, which is a contradiction.

Hence, the set A cannot be shattered, implying $\mathsf{PD}(\mathcal{F}) = 1$.

3.2.3 Uniform convergence. Let $\mathcal{S} = \{x_1, \ldots, x_\ell\}$ be a *bag* of elements of $\mathcal{H}$, sampled independently and uniformly at random, with replacement. For each $f \in \mathcal{F}$, define

$$\mathsf{m}_{\mathcal{H}}(f) = \frac{1}{|\mathcal{H}|} \sum_{x \in \mathcal{H}} f(x) \quad \text{and} \quad \mathsf{m}_{\mathcal{S}}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} f(x_i) \ .$$

We call $\mathsf{m}_{\mathcal{S}}(f)$ the *empirical average of f on $\mathcal{S}$*. It holds $\mathbb{E}[\mathsf{m}_{\mathcal{S}}(f)] = \mathsf{m}_{\mathcal{H}}(f)$. The following result connects an upper bound to the pseudodimension of $\mathcal{F}$ to the number of samples needed to simultaneously approximate all the expectations of all the functions in $\mathcal{F}$ using their sample averages.

Theorem 3.4 ([16]). *Let $\mathsf{PD}(\mathcal{F}) \leq d$. Fix $\xi, \eta \in (0,1)$. When $\mathcal{S}$ is a collection of*

$$|\mathcal{S}| = \frac{(b-a)^2}{\xi^2} \left(d + \log \frac{1}{\eta} \right) \tag{2}$$

elements sampled independently and uniformly at random with replacement from $\mathcal{H}$, then, with probability at least $1-\eta$ over the choice of $\mathcal{S}$, it holds

$$|\mathsf{m}_{\mathcal{H}}(f) - \mathsf{m}_{\mathcal{S}}(f)| < \xi, \text{ for every } f \in \mathcal{F} \ .$$

The following two lemmas by Riondato and Upfal [27, Lemmas 3.7 and 3.8] are useful when proving upper bounds to the pseudodimension of a family of functions.

Lemma 3.5. *If $B \subseteq \mathcal{H} \times [a,b]$ is shattered by $\mathcal{F}^+$, it may contain* at most one *element $(d,x) \in \mathcal{H} \times [a,b]$ for each $d \in \mathcal{H}$.*

Lemma 3.6. *If $B \subseteq \mathcal{H} \times [a,b]$ is shattered by $\mathcal{F}^+$, it cannot contain* any *element in the form (d,a), for any $d \in \mathcal{H}$.*

4 ALGORITHMS

We now present MiSoSouP, our suite of algorithms to compute ε-approximations of $\mathsf{TOP}_p(k, \mathcal{D})$.

4.1 MiSoSouP for 1-quality

We start by introducing a family $\mathcal{P}$ of functions which we use to give a novel expression for the 1-quality of a subgroup. We then present a sufficient condition for extracting an ε-approximation from a sample, and derive bounds to the sample size sufficient to ensure that the condition holds with high probability. Finally ,we describe the algorithm.

4.1.1 A novel formulation of the 1-quality. The family $\mathcal{P}$ contains one function ρ_Y from $\mathcal{D}$ to $\{-\mu(\mathcal{D}), 0, 1-\mu(\mathcal{D})\}$ for each subgroup $Y \in \mathcal{L}$, defined, for $t \in \mathcal{D}$, as:

$$\rho_Y(t) = \begin{cases} 1-\mu(\mathcal{D}) & \text{if } t \in \mathsf{C}_{\mathcal{D}}(Y) \text{ and } t.T = 1 \\ -\mu(\mathcal{D}) & \text{if } t \in \mathsf{C}_{\mathcal{D}}(Y) \text{ and } t.T = 0 \\ 0 & \text{otherwise} \end{cases} \ . \tag{3}$$

We assume to know the exact value of $\mu(\mathcal{D})$, which is a standard and reasonable assumption (made also by Scheffer and Wrobel [30]), since $\mu(\mathcal{D})$ can be computed with a very quick scan of the target attribute on $\mathcal{D}$, or kept up-to-date while collecting the data.

The 1-quality of a subgroup Y can be expressed as the average over the transactions in the dataset of the function ρ_Y:

$$\begin{aligned}
\mathsf{m}_{\mathcal{D}}(\rho_Y) &= \frac{1}{|\mathcal{D}|}\sum_{t\in\mathcal{D}}\rho_Y(t)\\
&= \frac{1}{|\mathcal{D}|}\left((1-\mu(\mathcal{D}))\mu_{\mathcal{D}}(Y)|\mathsf{C}_{\mathcal{D}}(Y)| - \mu(\mathcal{D})|\mathsf{C}_{\mathcal{D}}(Y)|(1-\mu_{\mathcal{D}}(Y))\right)\\
&= \frac{|\mathsf{C}_{\mathcal{D}}(Y)|}{|\mathcal{D}|}\left(\mu_{\mathcal{D}}(Y)-\mu(\mathcal{D})\right) = \mathsf{g}_{\mathcal{D}}(Y)\mathsf{u}_{\mathcal{D}}(Y) = \mathsf{q}^{(1)}_{\mathcal{D}}(Y)\ . \qquad (4)
\end{aligned}$$

This equivalence is a novel insight of *crucial importance* to enable the efficient estimation of the 1-quality from a sample of the dataset.

Let now $\mathcal{S} = \{t_1, \ldots, t_\ell\}$ be a collection of transactions sampled uniformly and independently at random with replacement from $\mathcal{D}$. It holds, following the same steps as in (4), that

$$\mathsf{m}_{\mathcal{S}}(\rho_Y) = \frac{1}{\ell}\sum_{i=1}^{\ell}\rho_Y(t_i) = \mathsf{g}_{\mathcal{S}}(Y)\left(\mu_{\mathcal{S}}(Y)-\mu(\mathcal{D})\right)\ .$$

This quantity is different from $\mathsf{q}^{(1)}_{\mathcal{S}}(Y)$, as it uses $\mu(\mathcal{D})$ rather than $\mu(\mathcal{S})$. As mentioned earlier, it is reasonable to assume knowledge of $\mu(\mathcal{D})$. We define the *approximate 1-quality of Y on $\mathcal{S}$* as

$$\tilde{\mathsf{q}}^{(1)}_{\mathcal{S}}(Y) = \mathsf{m}_{\mathcal{S}}(\rho_Y)\ .$$

4.1.2 Sufficient condition for an ε-approximation. We now show a condition on the sample $\mathcal{S}$ that is sufficient to allow the computation of an ε-approximation of $\mathsf{TOP}_1(k,\mathcal{D})$ from $\mathcal{S}$. Assume to sort the subgroups in $\mathcal{L}$ in decreasing order by their approximate 1-quality on $\mathcal{S}$, ties broken arbitrarily. Let $\tilde{\mathsf{r}}^{(1)}_{\mathcal{S}}(k)$ be the *approximate* 1-quality on $\mathcal{S}$ of the k-th subgroup in this order.

Theorem 4.1. *If $\mathcal{S}$ is such that*

$$|\tilde{\mathsf{q}}^{(1)}_{\mathcal{S}}(Y) - \mathsf{q}^{(1)}_{\mathcal{D}}(Y)| \le \frac{\varepsilon}{4} \text{ for every } Y\in\mathcal{L}, \qquad (5)$$

then the set

$$\mathcal{B} = \left\{\left(Y, \tilde{\mathsf{q}}^{(1)}_{\mathcal{S}}(Y)\right) \ :\ \tilde{\mathsf{q}}^{(1)}_{\mathcal{S}}(Y) \ge \tilde{\mathsf{r}}^{(1)}_{\mathcal{S}}(k) - \frac{\varepsilon}{2}\right\} \qquad (6)$$

is an ε-approximation to $\mathsf{TOP}_1(k,\mathcal{D})$.

Proof. Equation (5) holds in particular for subgroups appearing in the pairs in $\mathcal{B}$. Thus, $\mathcal{B}$ satisfies Property 3 from Def. 3.3. It holds

$$\tilde{\mathsf{r}}^{(1)}_{\mathcal{S}}(k) \ge \mathsf{r}^{(1)}_{\mathcal{D}}(k) - \frac{\varepsilon}{4} \qquad (7)$$

because all the subgroups in $\mathsf{TOP}_1(k,\mathcal{D})$, which are at least k, have, from (5), approximate 1-quality in $\mathcal{S}$ at least $\mathsf{r}^{(1)}_{\mathcal{D}}(k) - \varepsilon/4$. Another consequence of (5) is that

$$\tilde{\mathsf{r}}^{(1)}_{\mathcal{S}}(k) \le \mathsf{r}^{(1)}_{\mathcal{D}}(k) + \frac{\varepsilon}{4} \qquad (8)$$

because only subgroups with exact 1-quality in $\mathcal{D}$ strictly greater than $\mathsf{r}^{(1)}_{\mathcal{D}}(k)$ can have an approximate 1-quality in $\mathcal{S}$ strictly greater than $\mathsf{r}^{(1)}_{\mathcal{D}}(k) + \varepsilon/4$, and there are only at most $k-1$ such subgroups. It then holds from (8) and (5) that

$$\tilde{\mathsf{q}}^{(1)}_{\mathcal{S}}(Z) \ge \tilde{\mathsf{r}}^{(1)}_{\mathcal{S}}(k) - \frac{\varepsilon}{2} \text{ for all } Z\in\mathsf{TOP}_1(k,\mathcal{D})\ .$$

Thus $\mathcal{B}$ satisfies Property 1 of Def. 3.3.

Let now Y be any subgroup with $\mathrm{q}^{(1)}_{\mathcal{D}}(Y) < \mathrm{r}^{(1)}_{\mathcal{D}}(k) - \varepsilon$. It follows from (5) that

$$\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(Y) \leq \mathrm{r}^{(1)}_{\mathcal{D}}(k) - 3\varepsilon/4,$$

and using (7) we get

$$\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(Y) < \tilde{\mathrm{r}}^{(1)}_{\mathcal{S}}(k) - \varepsilon/2,$$

hence $\left(Y, \tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(Y)\right) \notin \mathcal{B}$, as required by Property 2 of Def. 3.3. □

4.1.3 Loose bounds to the sufficient sample size. Intuition correctly suggests that if the sample $\mathcal{S}$ is large enough, then with high probability over the choice of $\mathcal{S}$, $\mathcal{S}$ satisfies the condition in (5), thus allowing the computation of an ε-approximation of $\mathsf{TOP}_1(k, \mathcal{D})$ from $\mathcal{S}$. To warm up, and as a baseline, we first present a loose bound on how large $\mathcal{S}$ should be for the above to happen.

Theorem 4.2. *Let $\delta \in (0,1)$, $\varepsilon \in (0,1)$, and $k \geq 1$. Let $\mathcal{S}$ be a collection of*

$$|\mathcal{S}| \geq \frac{16}{\varepsilon^2}\left(\ln|\mathcal{L}_{\mathcal{D}}| + \ln\frac{2}{\delta}\right) \tag{9}$$

transactions sampled uniformly at random with replacement from $\mathcal{D}$. With probability at least $1-\delta$ (over the choice of $\mathcal{S}$), the set

$$\mathcal{B} = \left\{\left(Y, \tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(Y)\right) \;:\; \tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(Y) + \frac{\varepsilon}{2} \geq \tilde{\mathrm{r}}^{(1)}_{\mathcal{S}}(k)\right\}$$

is an ε-approximation to $\mathsf{TOP}_1(k, \mathcal{D})$.

To prove this result, we first recall the two-tailed Hoeffding's inequality [9].

Theorem 4.3. *Let f be a function from a domain $\mathcal{Y}$ to $[a,b] \subseteq \mathbb{R}$. Let $\mathcal{S} = (x_1, \ldots, x_\ell)$ be a collection of independent samples from $\mathcal{Y}$, and let $\xi \in (0,1)$. Then*

$$\Pr\left(|\mathrm{m}_{\mathcal{S}}(f) - \mathbb{E}\left[\mathrm{m}_{\mathcal{S}}(f)\right]| \geq \xi\right) \leq 2\exp\left(-\frac{n\xi^2}{(b-a)^2}\right) \;.$$

Then Thm. 4.2 is a straightforward application of Thm. 4.3, using $\varepsilon/4$ as ξ, and the fact that $b - a = 1$ for the functions in $\mathcal{P}$, and the union bound [19, Lemma 1.2].

The quantity in (9) is a loose upper bound to the sample size sufficient to probabilistically obtain an ε-approximation, due to the use of the union bound. It is also somewhat intuitive that the sample size should not depend on just the size of $\mathcal{L}_{\mathcal{D}}$, but on a quantity that better describes the relationship between the language and the dataset, as will be the case for the sample size used by MiSoSouP. Another drawback is that the sample size in (9) can only be computed when the *size* of $\mathcal{L}_{\mathcal{D}}$ is known, which is almost never the case. A loose upper bound to $|\mathcal{L}_{\mathcal{D}}|$ can be computed with a full scan of the dataset, which is potentially expensive (see details in Sect. 5). The sample size used by MiSoSouP, presented next, does not suffer from these downsides.

4.1.4 Bounds to the pseudodimension and to the sample size. In this section we present a novel upper bound to the number of samples needed to satisfy the condition in (5), and therefore compute an high-quality approximation of $\mathsf{TOP}_1(k, \mathcal{D})$. It relies on the following bound to the *pseudodimension* [23] (see Sect. 3.2) of the family $\mathcal{P}$ introduced in Sect. 4.1.1.

Theorem 4.4. *Let d be the maximum number of subgroups from $\mathcal{L}$ that may appear in a transaction of $\mathcal{D}$. Then, the pseudodimension $\mathsf{PD}(\mathcal{P})$ of $\mathcal{P}$ satisfies:*

$$\mathsf{PD}(\mathcal{P}) \leq \lfloor \log_2 d \rfloor + 1 \;.$$

We need some intermediate results before proving this theorem. Define, for every subgroup $Y \in \mathcal{L}$, the range

$$R_Y = \{(t, x) \ : \ t \in \mathcal{D} \text{ and } x \leq \rho_Y(t)\},$$

and let $\mathcal{R} = \{R_Y, Y \in \mathcal{L}\}$ be a rangeset on $\mathcal{D} \times [-\mu(\mathcal{D}), 1 - \mu(\mathcal{D})]$.

Lemma 3.6 tells us that only subsets of $\mathcal{D} \times (-\mu(\mathcal{D}), 1 - \mu(\mathcal{D})]$ may be shattered by $\mathcal{R}$. The following lemmas further restrict the collection of sets that may be shattered.

For any $x \in (-\mu(\mathcal{D}), 1 - \mu(\mathcal{D})]$ let

$$\mathrm{c}(x) = \begin{cases} 1 - \mu(\mathcal{D}) & \text{if } 0 < x \leq 1 - \mu(\mathcal{D}) \\ 0 & \text{if } -\mu(\mathcal{D}) < x \leq 0 \end{cases} .$$

Lemma 4.5. *A set $B \subseteq \mathcal{D} \times (-\mu(\mathcal{D}), 1 - \mu(\mathcal{D})]$ is shattered by $\mathcal{R}$ if and only if the set*

$$B' = \{(t, \mathrm{c}(x)) \ : \ (t, x) \in B\}$$

is also shattered by $\mathcal{R}$. It holds $|B| = |B'|$.

Proof. It follows from the definition of R_Y, $Y \in \mathcal{L}$, that (t, x) belongs to all and only the R_Y's that $(t, \mathrm{c}(x))$ belongs to. Hence if B is shattered then the same ranges that shatter it also shatter B', and vice versa.

The equality $|B| = |B'|$ follows from 1) the fact that clearly it is impossible that $|B'| > |B|$; and 2) Lemma 3.5 as it ensures that if B is shattered then it cannot contain more than a single element (t, y) for a fixed $t \in \mathcal{D}$ and some $y \in (-\mu(\mathcal{D}), 1 - \mu(\mathcal{D})]$, hence it is impossible that two or more elements of B are mapped by $\mathrm{c}(\cdot)$ to the same element of B'. □

Lemma 4.6. *Let $t \in \mathcal{D}$ be any transaction such that $t.T = 0$. No $B \subseteq \mathcal{D} \times \{0, 1 - \mu(\mathcal{D})\}$ such that $(t, 1 - \mu(\mathcal{D})) \in B$ can be shattered by $\mathcal{R}$.*

Proof. There is no subgroup $Y \in \mathcal{L}$ such that $(t, 1 - \mu(\mathcal{D})) \in R_Y$, thus, for any B containing $(t, 1 - \mu(\mathcal{D}))$, it is impossible to find an $Y \in \mathcal{L}$ such that $R_Y \cap B = \{(t, 1 - \mu(\mathcal{D}))\}$, hence B cannot be shattered. □

Lemma 4.7. *Let $t \in \mathcal{D}$ be any transaction such that $t.T = 1$. No $B \subseteq \mathcal{D} \times \{0, 1 - \mu(\mathcal{D})\}$ such that $(t, 0) \in B$ can be shattered by $\mathcal{R}$.*

Proof. The element $(t, 0)$ belongs to R_Y for any $Y \in \mathcal{L}$, so for any B containing $(t, 0)$, it is impossible to find an $Y \in \mathcal{L}$ such that $R_Y \cap B = \emptyset$, hence B cannot be shattered. □

It follows from Lemmas 3.6 and 4.5 to 4.7 that, to prove Thm. 4.4, we can focus our attention only on trying to shatter subsets of $\mathcal{D} \times [-\mu(\mathcal{D}), 1 - \mu(Ds)]$ containing elements that are either in the form $(t, 1 - \mu(\mathcal{D}))$ with $t.T = 1$, or in the form $(t, 0)$ with $t.T = 0$. The two following lemmas show upper bounds to the sizes of such subsets that can be shattered by $\mathcal{R}$. Theorem 4.4 is then an immediate consequence.

Lemma 4.8. *Let $B \subseteq \mathcal{D} \times \{0, 1 - \mu(\mathcal{D})\}$ be a set that is shattered by $\mathcal{R}$ and such that B contains an element $(t, 1 - \mu(\mathcal{D}))$, for some $t \in \mathcal{D}$. Then it must be*

$$|B| \leq \lfloor \log_2 d \rfloor + 1,$$

for d as in Thm. 4.4.

Proof. The proof is in part inspired by the one for [25, Theorem 4.5]. Consider one of the elements in the form $(t, 1 - \mu(\mathcal{D}))$ belonging to B. By hypothesis there is at least one such element. Let us denote it as $a = (t, 1 - \mu(\mathcal{D}))$.

Denote the $2^{|B|-1}$ non-empty subsets of B containing a as C_i, $1 \leq i \leq 2^{|B|-1}$, labelling them in an arbitrary order. Since B is shattered, for each of the C_i's there must be a subgroup Y_i such that $R_{Y_i} \cap B = C_i$. Since $C_i \neq C_j$ for each $i \neq j$, $1 \leq i, j \leq 2^{|B|-1}$, it must hold $R_{Y_i} \neq R_{Y_j}$. The element a belongs to each R_{Y_i}, $1 \leq i \leq 2^{|B|-1}$. From Lemma 4.6 it follows that, since B is shattered, then it must be $t.T = 1$. Thus the element a belongs to all and only the ranges R_Z for $Z \in \mathcal{L}$ such that $t \in \mathrm{C}_{\mathcal{D}}(Z)$. There are at most d such Z's, hence it must be $2^{|B|-1} \leq d$. □

Lemma 4.9. *Let $B \subseteq \mathcal{D} \times \{0, 1 - \mu(\mathcal{D})\}$ be a set that is shattered by $\mathcal{R}$ and such that B contains an element $(t, 0)$, for some $t \in \mathcal{D}$. Then it must be*

$$|B| \leq \lfloor \log_2 d \rfloor + 1,$$

for d as in Thm. 4.4.

Proof. Consider one of the elements in the form $(t, 0)$ that belong to B. By hypothesis there is at least one such element. Let us denote it as $a = (t, 0)$. The proof is similar to the one for Lemma 4.8, but with one profound difference, i.e., we essentially consider the subsets of B that *do not* contain a.

Denote the $2^{|B|-1}$ subsets of B *not* containing a as C_i, $1 \leq i \leq 2^{|B|-1}$, labelling them in an arbitrary order. There must be an i such that $C_i = \emptyset$. Since B is shattered, for each of the C_i's there must be a subgroup Y_i such that $R_{Y_i} \cap B = C_i$. Since $C_i \neq C_j$ for each $i \neq j$, $1 \leq i, j \leq 2^{|B|-1}$, it must hold $R_{Y_i} \neq R_{Y_j}$. The element a does not belong to any R_{Y_i}, $1 \leq i \leq 2^{|B|-1}$. From Lemma 4.5 it follows that, since B is shattered, then it must be $t.T = 0$. Thus the element a does not belong only to the ranges R_Z for $Z \in \mathcal{L}$ such that $t \in \mathrm{C}_{\mathcal{D}}(Z)$. There are at most d such Z's, hence it must be $2^{|B|-1} \leq d$. □

It is an interesting open question how to compute d efficiently for a generic $\mathcal{L}$. It is common to choose $\mathcal{L}$ to be the set of subgroups involving conjunctions of simple *equality* conditions on up to c attributes, for some $c \geq 1$. The following corollary is a reformulation of Thm. 4.4 using the maximum number of subgroups from $\mathcal{L}$ that may appear in a transaction of $\mathcal{D}$ for such cases.

Corollary 4.10. *Let z be the number of* description *attributes in $\mathcal{D}$ (i.e.,* not *counting the target attribute). Let $\mathcal{L}$ be the set of subgroups of conjunctions of equality conditions on up to c attributes, for some $1 \leq c \leq z$. Then*

$$\mathrm{PD}(\mathcal{P}) \leq \left\lfloor \log_2 \sum_{i=1}^{c} \binom{z}{i} \right\rfloor + 1 \ . \tag{10}$$

These upper bounds to the pseudodimension are *almost* thight, in the sense that there are datasets that almost attain the bounds, as shown in the following Lemma, which present an almost matching lower bound (up to the additive constant 1) to the pseudodimension.

Lemma 4.11. *Let z be a positive integer and let $\mathcal{L}$ be the language of subgroups of conjunctions of simple equality conditions on up to z attributes. There exists a dataset $\mathcal{D}$ with z description attributes such that*

$$z - 1 = \left\lfloor \log_2 \sum_{i=1}^{z} \binom{z}{i} \right\rfloor \leq \mathrm{PD}(\mathcal{P}) \leq \left\lfloor \log_2 \sum_{i=1}^{z} \binom{z}{i} \right\rfloor + 1 = z \ .$$

Proof. The equalities at the extreme of the thesis come from the definition of z, and the rightmost inequality come from corollary 4.10, so we only have to show that we can build a dataset for which

$$z - 1 \leq \mathrm{PD}(\mathcal{P}) \ .$$

For $z = 1$, in any dataset with at least two different transactions clearly is possible to shatter a set $\{t\}$ of one transaction t with the language of subgroups composed by a single equality condition: we only need the two subgroups "$A_1 = t.A_1$" and "$A_1 = t'.A_1$", where t' is any transaction with

value in A_1 different than the value of t in A_1. So for $z = 1$ the lower and the upper bound to the pseudodimension actually match.

For $z > 1$, consider any dataset $\mathcal{D}$ with z description attributes, such that $\mathcal{D}$ contains the set $\mathcal{W} = \{t_1, \ldots, t_{z-1}\}$ of $z - 1$ transactions where

$$t_i = (\underbrace{1, \ldots, 1}_{A_1,\ldots,A_{i-1}}, 0, \underbrace{1, \ldots, 1}_{A_{i+1},\ldots,A_z}, 0) \text{ for each } i = 1, \ldots, z-1 \ .$$

All the transactions in $\mathcal{W}$ have $A_z = 1$, and $T = 0$. We now show that $\mathcal{W}$ is shattered by $\mathcal{R}$. For each non-empty $C \subsetneq \mathcal{W}$, let $I_C = \{i \ : \ t_i \notin C\}$ be the set of indices of the transactions of $\mathcal{W}$ that are *not* in C, and define the subgroup

$$Y_C = \bigwedge_{i \in I_C} (A_i = 1)$$

composed of conjunctions on $|I_C| < z$ attributes. Clearly $Y_c \in \mathcal{L}$. For any non-empty $C \subsetneq \mathcal{W}$, it holds

$$R_{Y_C} \cap \mathcal{W} = C \tag{11}$$

because

- each transaction $t \in C$ has $t.A_i = 1$ for every $i \in I_C$, so t supports Y_C hence $t \in R_{Y_C}$. Therefore $C \subseteq R_{Y_C}$, and since $C \subseteq \mathcal{W}$, it must be
$$C \subseteq R_{Y_C} \cap \mathcal{W}; \text{ and} \tag{12}$$
- for each transaction $t \in \mathcal{W} \setminus C$ there exists an $i_t \in I_c$ such that $t.A_{i_t} = 0$, thus t does not support Y_C, hence $t \notin R_{Y_C}$, which means that
$$R_y \cap (\mathcal{W} \setminus C) = \emptyset \ . \tag{13}$$

By combining (12) and (13) we obtain (11), which holds for each *non-empty* $C \subsetneq \mathcal{W}$. Define now the subgroups $Y_{\mathcal{W}} = (A_z = 1)$ and $Y_\emptyset = (A_z = 0)$, both of which belong to $\mathcal{L}$. It holds $\mathcal{W} = R_{Y_{\mathcal{W}}} \cap \mathcal{W}$ and $\emptyset = R_{Y_\emptyset} \cap \mathcal{W}$ by construction of the transactions in $\mathcal{W}$. Thus $\mathcal{W}$, which contains $z - 1$ transactions, is shattered by $\mathcal{R}$, and $\mathrm{PD}(\mathcal{P}) \geq z - 1$. □

Understanding whether the looseness is in the upper bound or in the lower bound, and improving either to obtain matching bounds to the pseudodimension, are very interesting open questions.

Bounds to the sample size. By combining Thm. 4.4 with Thm. 3.4 we obtain the following result on the number of samples needed to guarantee, with high probability the sufficient condition to obtain an ε-approximation.

Theorem 4.12. *Let $\delta \in (0, 1)$, $\varepsilon \in (0, 1)$, and $k \geq 1$. Let d as in Thm. 4.4. Let*

$$S = \frac{16}{\varepsilon^2}\left(\lfloor \log_2 d \rfloor + 1 + \ln \frac{1}{\delta}\right). \tag{14}$$

The probability that a collection $\mathcal{S}$ of S transactions sampled independently and uniformly at random with replacement from $\mathcal{D}$ satisfies (5) *is at least $1 - \delta$.*

The improvement of (14) over (9) is evident: $\lfloor \log_2 d \rfloor + 1$ is usually much much smaller, potentially orders of magnitude so, than $\ln |\mathcal{L}_\mathcal{D}|$. The quantity d depends on both the dataset and the language: it is intuitively more "natural" that the sample size should depend on the relationship between the two, rather than just on the language.

4.1.5 The algorithm. We now have all the ingredients to describe and analyze MiSoSouP-1, our algorithm for extracting, with probability at least $1-\delta$ (over the runs of the algorithm), an ε-approximation to $\mathsf{TOP}_1(k,\mathcal{D})$. The pseudocode is shown in Alg. 1 The input of the algorithm is the tuple $(\mathcal{D},\mathcal{L},k,\varepsilon,\delta)$.

Algorithm 1: MiSoSouP-1

Input: dataset $\mathcal{D}$, language $\mathcal{L}$, integer $k \geq 1$, reals $\varepsilon,\delta \in (0,1)$
Output: a set $\mathcal{B}$ of pairs (Y,c), where Y is a subgroup in $\mathcal{L}$, and $c \in [-1,1]$ such that $\mathcal{B}$ is, with probability at least $1-\delta$, an ε-approximation of $\mathsf{TOP}_1(\mathcal{D},k)$.

1 $d \leftarrow \texttt{maxSubgroupsInTransaction}(\mathcal{D},\mathcal{L})$
2 $S \leftarrow 16\left(\lfloor \log_2 d \rfloor + 1 + \ln 1/\delta\right)/\varepsilon^2$
3 $\mathcal{S} \leftarrow \texttt{uniformRandomSample}(\mathcal{D},S)$
4 $\mathrm{r}^{(1)}_{\mathcal{S}}(k) \leftarrow \texttt{getTopKQuality}(\mathcal{S},k)$
5 $\mathcal{B} \leftarrow \texttt{mineSubgroupsWithThreshold}(\mathcal{S},\mathrm{r}^{(1)}_{\mathcal{S}}(k)-\varepsilon/2)$
6 **return** $\mathcal{B}$

After having computed the (upper bound to the) maximum number of subgroups of $\mathcal{L}$ in a transaction of $\mathcal{D}$ and from it the sample size S as in (14), MiSoSouP-1 creates the sample $\mathcal{S}$ by drawing S transactions independently and uniformly at random with replacement from $\mathcal{D}$. An exact algorithm for subgroup discovery is used to extract from $\mathcal{S}$ the set $\mathcal{B}$ defined in (6). Any exact algorithm can be used for the discovery step, but it needs to be slightly modified to use $\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(Y)$ as measure for the interestingness of a subgroup Y, instead of $\mathrm{q}^{(1)}_{\mathcal{S}}(Y)$. This modification is straightforward. The set $\mathcal{B}$ is then returned in output. By combining Thm. 4.1 with Thm. 4.12 we obtain the following result on the quality guarantees of MiSoSouP-1.

Theorem 4.13. *With probability at least* $1-\delta$ *(over its runs),* MiSoSouP-1 *outputs an* ε*-approximation to* $\mathsf{TOP}_1(k,\mathcal{D})$*.*

4.2 MiSoSouP for 2-quality

In this section we present MiSoSouP-2, the variant of MiSoSouP for computing an ε-approximation to $\mathsf{TOP}_2(k,\mathcal{D})$. To obtain upper bounds on the number of samples needed by MiSoSouP-2, we will combine the result obtained in the previous section with variants of results by Riondato and Vandin [28] on the number of samples needed to compute a high-quality approximation of the frequent itemsets from a random sample of a transactional dataset.

We start by defining another family $\mathcal{G}$ of functions, in addition to $\mathcal{P}$ defined in Sect. 4.1.1. The domain of the functions in $\mathcal{G}$ is $\mathcal{D}$. For each subgroup $A \in \mathcal{L}$ there is one function $g_A \in \mathcal{G}$, defined as follows, for $t \in \mathcal{D}$:

$$g_A(t) = \mathbb{1}_{\mathrm{C}_{\mathcal{D}}(A)}(t) = \begin{cases} 1 & \text{if } t \in \mathrm{C}_{\mathcal{D}}(A) \\ 0 & \text{otherwise} \end{cases} . \tag{15}$$

It holds

$$\mathrm{m}_{\mathcal{D}}(g_A) = \frac{1}{|\mathcal{D}|}\sum_{t\in\mathcal{D}} g_A(t) = \frac{|\mathrm{C}_{\mathcal{D}}(A)|}{|\mathcal{D}|} = \mathrm{g}_{\mathcal{D}}(A) .$$

Let now $\mathcal{S} = \{t_1,\ldots,t_\ell\}$ be a collection of transactions sampled uniformly and independently at random with replacement from $\mathcal{D}$. It holds

$$\mathrm{m}_{\mathcal{S}}(g_A) = \frac{1}{\ell}\sum_{i=1}^{\ell} g_A(t_i) = \mathrm{g}_{\mathcal{S}}(A)$$

For any subgroup $A \in \mathcal{L}$, we define the *approximate 2-quality of A on $\mathcal{S}$* as

$$\tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) = \mathrm{m}_{\mathcal{S}}(g_A)\mathrm{m}_{\mathcal{S}}(\rho_A) = \mathrm{g}_{\mathcal{S}}(A)\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(A) \ .$$

As was the case for $p = 1$, it holds $\tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) \neq \mathrm{q}^{(2)}_{\mathcal{S}}(A)$.

The following lemma shows how a bound on the deviations of the values taken by the functions in $\mathcal{G}$ on $\mathcal{S}$ from their values on $\mathcal{D}$, together with a bound on the deviations of the approximate 1-qualities on $\mathcal{S}$ of all subgroups, give a bound to the approximate 2-qualities on $\mathcal{S}$ of all subgroups.

Lemma 4.14. *Let $\varepsilon \in (0, 1)$. If*

(1) $\sup_{A\in\mathcal{L}} \left|\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(A) - \mathrm{q}^{(1)}_{\mathcal{D}}(A)\right| \leq \sqrt{1 + \varepsilon/4} - 1$*; and*

(2) $\sup_{A\in\mathcal{L}} |\mathrm{g}_{\mathcal{S}}(A) - \mathrm{g}_{\mathcal{D}}(A)| \leq \sqrt{1 + \varepsilon/4} - 1$,

then

$$\sup_{A\in\mathcal{L}} \left|\tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) - \mathrm{q}^{(2)}_{\mathcal{D}}(A)\right| \leq \frac{\varepsilon}{4} \ .$$

Proof. We show the proof for a more general case. For any η_{q} and η_{g}, assume it holds

(1) $\sup_{A\in\mathcal{L}} \left|\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(A) - \mathrm{q}^{(1)}_{\mathcal{D}}(A)\right| \leq \eta_{\mathrm{q}}$; and

(2) $\sup_{A\in\mathcal{L}} |\mathrm{g}_{\mathcal{S}}(A) - \mathrm{g}_{\mathcal{D}}(A)| \leq \eta_{\mathrm{g}}$.

From the above, it holds, for any $A \in \mathcal{L}$,

$$\begin{aligned}
\tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) = \mathrm{g}_{\mathcal{S}}(A)\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(A) &\leq (\mathrm{g}_{\mathcal{D}}(A) + \eta_{\mathrm{g}}) \left(\mathrm{q}^{(1)}_{\mathcal{D}}(A) + \eta_{\mathrm{q}}\right) \\
&\leq \mathrm{q}^{(2)}_{\mathcal{D}}(A) + \mathrm{q}^{(1)}_{\mathcal{D}}(A)\eta_{\mathrm{g}} + \mathrm{g}_{\mathcal{D}}(A)\eta_{\mathrm{q}} + \eta_{\mathrm{g}}\eta_{\mathrm{q}} \\
&\leq \mathrm{q}^{(2)}_{\mathcal{D}}(A) + \eta_{\mathrm{g}} + \eta_{\mathrm{q}} + \eta_{\mathrm{g}}\eta_{\mathrm{q}} \ . \qquad (16)
\end{aligned}$$

Additionally, again from the hypothesis, for any $A \in \mathcal{L}$ it holds

$$\begin{aligned}
\tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) = \mathrm{g}_{\mathcal{S}}(A)\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(A) &\geq (\mathrm{g}_{\mathcal{D}}(A) - \eta_{\mathrm{g}}) \left(\mathrm{q}^{(1)}_{\mathcal{D}}(A) - \eta_{\mathrm{q}}\right) \\
&\geq \mathrm{q}^{(2)}_{\mathcal{D}}(A) - \mathrm{q}^{(1)}_{\mathcal{D}}(A)\eta_{\mathrm{g}} - \mathrm{g}_{\mathcal{D}}(A)\eta_{\mathrm{q}} + \eta_{\mathrm{g}}\eta_{\mathrm{q}} \\
&\geq \mathrm{q}^{(2)}_{\mathcal{D}}(A) - \eta_{\mathrm{g}} - \eta_{\mathrm{q}} - \eta_{\mathrm{g}}\eta_{\mathrm{q}} \ . \qquad (17)
\end{aligned}$$

Thus, by combining, we obtain

$$\sup_{A\in\mathcal{L}} \left|\tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) - \mathrm{q}^{(2)}_{\mathcal{D}}(A)\right| \leq \eta_{\mathrm{g}} + \eta_{\mathrm{q}} + \eta_{\mathrm{g}}\eta_{\mathrm{q}} \ .$$

The thesis follows by setting $\eta_{\mathrm{q}} = \eta_{\mathrm{g}} = \sqrt{1 + \varepsilon/4} - 1$. □

This lemma sheds light on how to obtain ε-approximations of $\mathrm{TOP}_2(k, \mathcal{D})$ from a sample $\mathcal{S}$: the sample must offer *simultaneous guarantees on two family of functions*, $\mathcal{P}$ and $\mathcal{G}$. We discussed the case for family $\mathcal{P}$ when presenting MiSoSouP-1, so we now focus on $\mathcal{G}$. An application of the union bound will ensure the simultaneous guarantees.

4.2.1 Loose bounds to the sufficient sample size. Similarly to what we did in Sect. 4.1.3, we show in Thm. 4.15 a loose upper bound to the sample size needed to obtain an ε-approximation to $\mathrm{TOP}_2(k, \mathcal{D})$ from a sample $\mathcal{S}$. This bound is obtained through an additional application of Hoeffding's inequality to bound the deviation of the estimation of the generality from its expectation, and an application of the union bound.

Theorem 4.15. *Let $\delta \in (0,1)$, $\varepsilon \in (0,1)$, and $k \geq 1$. Let*

$$\xi = \sqrt{1+\varepsilon/4} - 1 \ .$$

With probability at least $1-\delta$, if $\mathcal{S}$ is a collection of

$$|\mathcal{S}| \geq \frac{1}{\xi^2}\left(\ln|\mathcal{L}| + \ln\frac{4}{\delta}\right) \tag{18}$$

transactions sampled uniformly at random with replacement from $\mathcal{D}$, then

$$\tilde{B} = \left\{\left(A, \tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A)\right) \ : \ \tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) + \frac{\varepsilon}{2} \geq \tilde{\mathrm{r}}^{(2)}_{\mathcal{S}}(k)\right\}$$

is an ε-approximation top $\mathsf{TOP}_2(k,\mathcal{D})$.

Proof (Sketch). With probability at least $1-\delta$ it holds simultaneously that

$$\sup_{A\in\mathcal{L}} |\tilde{\mathrm{q}}^{(1)}_{\mathcal{S}}(A) - \mathrm{q}^{(1)}_{\mathcal{D}}(A)| \leq \xi$$

and

$$\sup_{A\in\mathcal{L}} |\mathrm{g}_{\mathcal{S}}(A) - \mathrm{g}_{\mathcal{D}}(A)| \leq \xi \ .$$

An application of Lemma 4.14 with $\eta_{\mathrm{g}} = \eta_{\mathrm{q}} = \xi$ concludes the proof. □

4.2.2 The algorithm. Consider a rangeset $\mathcal{R}_{\mathrm{g}}$ containing a range

$$R_A = \{t \in \mathcal{D} \ : \ A \in t\}$$

for each $A \in \mathcal{L}$ appearing in at least one transaction of $\mathcal{D}$.

As shown by Riondato and Vandin [28, Sect. 4], the VC-dimension of this rangeset is also upper bounded by $\lfloor \log_2 d \rfloor + 1$, for d as in Thm. 4.4. An equivalent of Thm. 3.4 also holds for families of binary functions with bounded VC-dimension, with the same sample size as in (2), and therefore relates the sample size of $\mathcal{S}$ with the maximum deviation from the second condition of Lemma 4.14.

We are now ready to describe MiSoSouP-2. The input is the same tuple $(\varepsilon, \delta, k, \mathcal{D})$ as in MiSoSouP-1. After creating a sample $\mathcal{S}$ of size

$$S = \frac{1}{\left(\sqrt{1+\varepsilon/4}-1\right)^2}\left(\lfloor \log_2 d \rfloor + 1 + \ln\frac{2}{\delta}\right), \tag{19}$$

MiSoSouP-2 runs on $\mathcal{S}$ a modified variant of an exact algorithm for subgroup discovery that uses $\tilde{\mathrm{q}}^{(2)}$ as the interestingness measure. Let $\tilde{\mathrm{r}}^{(2)}_{\mathcal{S}}(k)$ be the top-k highest *approximate* 2-quality on $\mathcal{S}$, ties broken arbitrarily. The output set $\mathcal{B}$ is defined as

$$\mathcal{B} = \left\{\left(A, \tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A)\right) \ : \ \tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) + \frac{\varepsilon}{2} \geq \tilde{\mathrm{r}}^{(2)}_{\mathcal{S}}(k)\right\} \ .$$

The following theorem states the guarantees of MiSoSouP-2.

Theorem 4.16. *With probability at least $1-\delta$ (over its runs),* MiSoSouP-2 *outputs an ε-approximation to* $\mathsf{TOP}_2(k,\mathcal{D})$.

sketch. An application of the union bound and of Theorem 3.4 and its corresponding version for VC-dimension gives that the probability that $\mathcal{S}$ satisfies the hypothesis of Lemma 4.6 is at least $1-\delta$. When that is the case, then the thesis of that lemma holds, i.e.,

$$\sup_{A\in\mathcal{L}} \left|\tilde{\mathrm{q}}^{(2)}_{\mathcal{S}}(A) - \mathrm{q}^{(2)}_{\mathcal{D}}(A)\right| \leq \frac{\varepsilon}{4} \ .$$

From here, the proof continues essentially as the one for Theorem 4.12. □

4.3 MiSoSouP for 1/2-quality

We now present MiSoSouP-1/2, the variant of MiSoSouP for computing an approximation of the top-k subgroups in $\mathcal{D}$ w.r.t. 1/2-quality and *with generality* $\mathrm{g}_{\mathcal{D}}(Y) \geq \sigma$, where σ is a user-defined threshold. Let $\mathcal{L}_\sigma$ be the set of subgroups Y in $\mathcal{L}$ with $\mathrm{g}_{\mathcal{D}}(Y) \geq \sigma$. Assume to rank the subgroups in $\mathcal{L}_\sigma$ in decreasing order according to their 1/2-quality in $\mathcal{D}$, ties broken arbitrarily. Let $k > 0$ be an integer and let $\mathrm{r}_{\mathcal{D}}^{(1/2)}(k)$ be the p-quality of the k-th subgroup in the ranking. We then define

$$\mathsf{TOP}_{1/2}(k, \sigma, \mathcal{D}) = \left\{ Y \in \mathcal{L}_\sigma \ : \ \mathrm{q}_{\mathcal{D}}^{(1/2)}(Y) \geq \mathrm{r}_{\mathcal{D}}^{(1/2)}(k) \right\} \ .$$

The additional constraint on the generality of the subgroups is needed for technical purposes of the analysis. Other algorithms based on sampling for computing approximations of the original set $\mathsf{TOP}_{1/2}(k, \mathcal{D})$ [30] do not require this additional constraint but they also do not actually offer the promised quality guarantees (see App. A).

The guarantees offered by MiSoSouP-1/2 take into account the additional constraint as follows.

Definition 4.17. Let $\varepsilon \in (0, 3/4\sigma)$. A ε-approximation to the set $\mathsf{TOP}_{1/2}(k, \sigma, \mathcal{D})$ is a set $\mathcal{B}$ of pairs (Y, q_Y) where Y is a subgroup and:

(1) for any $Y \in \mathsf{TOP}_{1/2}(k, \sigma, \mathcal{D})$, there is a $(Y, q_Y) \in \mathcal{B}$;
(2) there is no pair $(Y, q_Y) \in \mathcal{B}$ such that $Y \in \mathcal{L}_\sigma$ and $\mathrm{q}_{\mathcal{D}}^{(1/2)}(Y) < \frac{1}{4}\mathrm{r}_{\mathcal{D}}^{(1/2)}(k) - \varepsilon$;
(3) there is no pair $(Y, q_Y) \in \mathcal{B}$ such that $\mathrm{g}_{\mathcal{D}}(Y) < \sigma - \varepsilon/2$;
(4) for each pair $(Y, q_Y) \in \mathcal{B}$ with $Y \in \mathcal{L}_\sigma$, $\frac{1}{\sqrt{2}} q_Y - \varepsilon/2 \leq \mathrm{q}_{\mathcal{D}}^{(1/2)}(Y) \leq 2q_Y + \varepsilon/2$.

As for $p = 1, 2$, an ε-approximation can be used as a set of candidates for $\mathsf{TOP}_{1/2}(k, \sigma, \mathcal{D})$, as it contains a pair (Y, q_Y) for each subgroup Y in this set.

The following lemma shows that, similarly to the 2-quality, a bound to the approximate 1/2-qualities on $\mathcal{S}$, for subgroups Y in $\mathcal{L}_\sigma$, is obtained by combining a bound on the deviations of the values taken by the functions in $\mathcal{G}$ (defined in Sect. 4.2) on $\mathcal{S}$ from their values on $\mathcal{D}$, together with a bound on the deviations of the approximate 1-qualities on $\mathcal{S}$ of all subgroups in $\mathcal{L}_\sigma$ and a minor requirement on the relationship between σ and ε.

Lemma 4.18. *If*

(1) $\sup_{Y \in \mathcal{L}_\sigma} |\mathrm{g}_{\mathcal{S}}(Y) - \mathrm{g}_{\mathcal{D}}(Y)| \leq \frac{\varepsilon}{4}$; *and*
(2) $\sup_{Y \in \mathcal{L}_\sigma} \left| \tilde{\mathrm{q}}_{\mathcal{S}}^{(1)}(Y) - \mathrm{q}_{\mathcal{D}}^{(1)}(Y) \right| \leq \frac{\varepsilon\sqrt{\sigma}}{4}$; *and*
(3) $\varepsilon \leq \frac{3}{4}\sigma$

then

(1) $\sup_{Y \in \mathcal{L}_\sigma} \left| \tilde{\mathrm{q}}_{\mathcal{S}}^{(1/2)}(Y) - 2\mathrm{q}_{\mathcal{D}}^{(1/2)}(Y) \right| \leq \frac{\varepsilon}{2}$; *and*
(2) $\sup_{Y \in \mathcal{L}_\sigma} \left| \mathrm{q}_{\mathcal{D}}^{(1/2)}(Y)/\sqrt{2} - \tilde{\mathrm{q}}_{\mathcal{S}}^{(1/2)}(Y) \right| \leq \frac{\varepsilon}{2}$.

Proof. We prove the general case for positive reals η_q and $\eta_g < 3/4$. Assume $\sup_{Y \in \mathcal{L}_\sigma} |g_{\mathcal{S}}(Y) - g_{\mathcal{D}}(Y)| \leq \eta_g$, $\sup_{Y \in \mathcal{L}_\sigma} \left|\tilde{q}^{(1)}_{\mathcal{S}}(Y) - q^{(1)}_{\mathcal{D}}(Y)\right| \leq \eta_q$. From the hypothesis it holds, for any $Y \in \mathcal{L}_\sigma$,

$$\begin{aligned}
\tilde{q}^{(1/2)}_{\mathcal{S}}(Y) = \tilde{q}^{(1)}_{\mathcal{S}}(Y)/\sqrt{g_{\mathcal{S}}(Y)} &\leq \frac{q^{(1)}_{\mathcal{D}}(Y) + \eta_q}{\sqrt{g_{\mathcal{D}}(Y) - \eta_g}} \\
&\leq \frac{q^{(1)}_{\mathcal{D}}(Y)}{\sqrt{g_{\mathcal{D}}(Y) - \eta_g}} + \frac{\eta_q}{\sqrt{g_{\mathcal{D}}(Y) - \eta_g}} \\
&\leq 2q^{(1/2)}_{\mathcal{D}}(Y) + 2\frac{\eta_q}{\sqrt{g_{\mathcal{D}}(Y)}} \\
&\leq 2q^{(1/2)}_{\mathcal{D}}(Y) + 2\frac{\eta_q}{\sqrt{\sigma}},
\end{aligned}$$

where the third inequality follows from

$$g_{\mathcal{D}}(Y) - \eta_g \geq g_{\mathcal{D}}(Y) - \frac{3}{4}g_{\mathcal{D}}(Y) = \frac{g_{\mathcal{D}}(Y)}{4} \quad .$$

In addition, again from the hypothesis, for any $Y \in \mathcal{L}_\sigma$ it holds

$$\begin{aligned}
\tilde{q}^{(1/2)}_{\mathcal{S}}(Y) = \tilde{q}^{(1)}_{\mathcal{S}}(Y)/\sqrt{g_{\mathcal{S}}(Y)} &\geq \frac{q^{(1)}_{\mathcal{D}}(Y) - \eta_q}{\sqrt{g_{\mathcal{D}}(Y) + \eta_g}} \\
&\geq \frac{q^{(1)}_{\mathcal{D}}(Y)}{\sqrt{g_{\mathcal{D}}(Y) + \eta_g}} - \frac{\eta_q}{\sqrt{g_{\mathcal{D}}(Y) + \eta_g}} \\
&\geq \frac{q^{(1)}_{\mathcal{D}}(Y)}{\sqrt{2}\sqrt{g_{\mathcal{D}}(Y)}} - \frac{\eta_q}{\sqrt{g_{\mathcal{D}}(Y)}} \\
&\geq \frac{1}{\sqrt{2}}q^{(1/2)}_{\mathcal{D}}(Y) - \frac{\eta_q}{\sqrt{\sigma}},
\end{aligned}$$

where the third inequality follows from $\eta_g \leq \frac{3}{4}g_{\mathcal{D}}(Y)$.

The thesis follows by setting $\eta_g = \frac{\varepsilon}{4}$ and $\eta_q = \frac{\varepsilon\sqrt{\sigma}}{4}$. □

4.3.1 The algorithm. We now describe MiSoSouP-1/2. The input is a tuple $(\varepsilon, \delta, k, \mathcal{D}, \sigma)$, where ε, δ, k, and $\mathcal{D}$ are the same as in MiSoSouP-1 and MiSoSouP-2, while σ is a minimum generality threshold for the subgroups of interest. In addition, MiSoSouP-1/2 requires that $\varepsilon \leq \frac{3}{4}\sigma$.

After creating a sample of $\mathcal{S}$ of size

$$S = \frac{16}{\varepsilon^2\sigma}\left(\lfloor \log_2 d \rfloor + 1 + \ln\frac{2}{\delta}\right),$$

MiSoSouP-1/2 runs on $\mathcal{S}$ a variant of an exact algorithm for subgroup discovery that uses $\tilde{q}^{(1/2)}$ as interesting measure. Let $\tilde{r}^{(1/2)}_{\mathcal{S}}(k)$ be the top-k highest *approximate* 1/2-quality on $\mathcal{S}$ for groups Y with $g_{\mathcal{S}}(Y) \geq \sigma - \frac{\varepsilon}{4}$, ties broken arbitrarily. The output set $\mathcal{B}$ is defined as

$$\begin{aligned}
\mathcal{B} = \Big\{\Big(Y, \tilde{q}^{(1/2)}_{\mathcal{S}}(Y)\Big) \;:\; & g_{\mathcal{S}}(Y) \geq \sigma - \varepsilon/4, \\
& \tilde{q}^{(1/2)}_{\mathcal{S}}(Y) + (\varepsilon/2) \geq \tilde{r}^{(1/2)}_{\mathcal{S}}(k)\Big\} \quad .
\end{aligned}$$

The following establishes the theoretical guarantees of MiSoSouP-1/2.

Table 3. Characteristics of the datasets

Dataset	Size	Attributes	Max. Length
Car	6912×10^4	6	4
Mushroom	32496×10^4	22	4
Tic-Tac-Toe	3832×10^4	9	5

Theorem 4.19. *With probability at least* $1 - \delta$ *(over its runs),* MiSoSouP-1/2 *outputs an* ε*-approximation to* $\mathsf{TOP}_{1/2}(k, \sigma, \mathcal{D})$.

Sketch. The proof follows the same lines as the proof for Thm. 4.16, but leveraging on Lemma 4.18 instead of Lemma 4.6. □

5 EXPERIMENTAL EVALUATION

We now discuss our experimental evaluation to assess the performances of MiSoSouP.

5.1 Goals

Our experiments have two goals: *1.* evaluate the speed-up of MiSoSouP w.r.t. sampling-based approximation algorithms offering the same quality guarantees; and *2.* evaluate the quality of the approximations returned by MiSoSouP, in terms of the accuracy of the estimates of the quality of the returned subgroups, and of the number of returned subgroups.

5.2 Baselines

We compare the performances of MiSoSouP against a class UB of baseline algorithms.[7] We use UB-1 and UB-2 to denote the variant of UB for 1- and 2-quality respectively. Like MiSoSouP, UB computes, with probability at least $1 - \delta$, an ε-approximation to $\mathsf{TOP}_p(k, \mathcal{D})$ by analyzing a sample of the dataset. The *only difference* between MiSoSouP-1 (resp. MiSoSouP-2) and UB-1 (resp. UB-2) is that UB-1 (resp. UB-2) uses, as sample size, the r.h.s. of (9) (resp. of (18)). In other words, the pseudocode for UB-1 would be very similar to the one presented in Alg. 1 for MiSoSouP-1, with difference only in the computation of the sample size S (line 1 and line 2), in the sense that on line 1, UB-1 would have to compute the number of subgroups in $\mathcal{L}$ that actually appear in $\mathcal{D}$ (or an upper bound to such number) and in line 2, UB-1 would compute the sample size S using (9). The rest of the code would not change w.r.t. MiSoSouP-1, and neither would the input or output (including the guarantees). In our experimental evaluation we consider description languages of conjunctions of *equality* conditions on up to *maxlen* attributes, for some value *maxlen* (see Table 3). In this case, an *upper bound* to the number of subgroups in $\mathcal{L}$ that appear in $\mathcal{D}$ can be computed by considering the size of the (effective) domains of the columns in the dataset, and taking the sum, over all r-subsets C of columns, for r from 1 to *maxlen*, of the products of the sizes of the column domains in C. Computing the sizes of the column domains requires a linear scan of the dataset. Despite the fact that this step can be relatively expensive and its cost grows with the size of the dataset, we do not include the time for such computation in the reported runtime of UB, therefore favoring UB in our comparisons. MiSoSouP relies on (10) to compute the upper bound d to the pseudodimension used in (14) to obtain its sample size. The cost of evaluating the r.h.s. of (10) is essentially nil, as all values are known by MiSoSouP, since $\mathcal{L}$ and thus c are fixed in advance, and the number of columns of $\mathcal{D}$ is an immediately available quantity.

[7] The UB algorithms were not presented before in the literature. We introduce them only for comparison with MiSoSouP, which, as we will see, offers several practical advantages.

We do not compare MiSoSouP with algorithms that mine the whole dataset and output the exact collection $\text{TOP}_p(k, \mathcal{D})$ because MiSoSouP (and also UB) have sample sizes that are independent on the size of the dataset, while an exact algorithm would take time proportional to this quantity. As a result, on modern-sized datasets, an exact algorithm is always much slower than a sampling-based algorithm. We also do not to compare against GSS [30] because the algorithm does not actually offer the claimed guarantees (see App. A), and an implementation is not available.

5.3 Datasets and languages

We use datasets from the UCI repository [17]. Since these datasets are quite small for today's standards, we replicate them 20,000 times (i.e., each transaction is copied 20,000 times) and then shuffle the order of the transactions in the replicated copy. This way, we obtain significantly larger datasets while *preserving the distribution* of the p-qualities of the subgroups appearing in the original datasets. This approach does not change the search space of any algorithm and does not give any advantage to MiSoSouP over UB. Table 3 shows the descriptive statistics of the datasets we used. We consider the description language $\mathcal{L}$ of subgroups of up to "Max. Length" conjunctions of *equality* conditions.

5.4 Implementation and environment

We implemented MiSoSouP and UB in C++17. Our code is available from http://matteo.rionda.to/software/misosoup.tbz2. The implementation uses a simple exhaustive search algorithm for extracting the subgroups from the sample. Any algorithm can be used for this step, we just found it more practical to write our own implementation than to modify an existing implementation of a more efficient algorithm. We run our experiments on a cluster of GNU/Linux machines, except for the timing experiments, which were performed on a machine with an AMD Phenom$^{\text{TM}}$ II X4 955 processor and 16GB of RAM, running FreeBSD 12.

5.5 Parameters

We report results for $k \in \{10, 50, 100, 200, 500, 1000, 2000\}$, $\varepsilon \in \{0.05, 0.02, 0.01, 0.0075\}$,[8] and for $\delta = 0.1$. We tested different values for δ, but given that both MiSoSouP and UB have (the same) logarithmic dependence on δ, varying δ has limited quantitative effect and no qualitative effect. We run MiSoSouP and UB five times for each combination of parameters: the results were extremely stable and we report them for a randomly chosen run among the five.

5.6 Results for $p = 1$

We first show the results on runtime and sample sizes (Sect. 5.6.1), then discuss the accuracy of the estimates of the 1-qualities obtained by MiSoSouP-1 (Sect. 5.6.2), and finally analyze the number of false positives it reports (Sect. 5.6.3).

5.6.1 Sample size reduction and speed-up. We compare the number of samples used by MiSoSouP-1 and by UB-1 as ε varies. In both cases, the sample size is independent from k: k enters into play only when computing the final output, so it can be chosen after the "sampling phases" of the algorithms have run. The results are presented in the 3$^{\text{rd}}$ and 4$^{\text{th}}$ column from the left of Table 4. W.r.t. the whole dataset (whose size is reported in Table 3), MiSoSouP-1 looks at a small fraction of the transactions, and *this quantity does not grow as the dataset grows*, which is one of the main

[8] With the exception of $\varepsilon = 0.0075$ for the Tic-Tac-Toe dataset for $p = 2$, because running MiSoSouP-2 with these parameters would take too long.

Table 4. Sample size and runtime evaluation for MiSoSouP-1

Dataset	ε	$\lvert\mathcal{S}\rvert$	Reduction w.r.t. UB-1	Runtime (s)	Reduction w.r.t. UB-1
Car	0.05	53137	-25.07%	1.50	-1.5%
	0.02	332104		2.13	-10.55%
	0.01	1328414		4.42	-17.68%
	0.0075	2361625		6.67	-20.16%
Mushroom	0.05	104337	-11.98%	88.66	-8.64%
	0.02	652104		467.97	-13.86%
	0.01	2608414		1816.05	-11.45%
	0.0075	4637180		3274.01	-10.70%
Tic-Tac-Toe	0.05	72337	-17.35%	2.34	-12.96%
	0.02	452104		9.72	-16.66%
	0.01	1808414		35.36	-17.47%
	0.0075	3214958		59.31	-19.72%

advantages of sampling-based approaches.[9] MiSoSouP-1 achieves a very large reduction in the sample size w.r.t. UB-1 (only a single number is reported for each dataset because the two sample sizes have the same dependency on ε and δ, and do not depend on k). The reduction is extremely significant because, especially when ε is small, UB-1 would require to analyze a sample *larger* than the original dataset, defeating the whole purpose of sampling, while MiSoSouP-1 would still shine.[10] Hence, MiSoSouP-1 can be used with success in situations where UB-1 would be useless. There are other scenarios where UB-1 would not work but MiSoSouP-1 would: if given just a sample and no information on the *size* of the language, UB-1 would not be able to compute the sample size, while MiSoSouP-1 would have no issues. Thus, MiSoSouP-1 requires fewer transactions than UB-1, while being more flexible.

The runtime of MiSoSouP-1 and the reduction over UB-1 are reported in the 5$^{\text{th}}$ and 6$^{\text{th}}$ columns of Table 4. We remark once again that the runtime of UB-1 did not include the time to compute an upper bound to the size of language, which on large datasets is significant. Thus the improvement of MiSoSouP-1 over UB-1 is actually even larger than reported. At small sample sizes (i.e., large values of ε), both algorithms have fixed costs that dominate over the part of the running time that depends on the size of the sample, thus the reduction in MiSoSouP-1's runtime w.r.t. UB-1's is not proportional to the reduction in the sample size. The sample-size-dependent costs dominate when ε is small (larger sample sizes) and in these cases the speed-up becomes essentially equal to the reduction in the sample size.

5.6.2 Accuracy. We evaluate the accuracy of the output of MiSoSouP-1 by measuring, for each subgroup A in the output, the *absolute error* on the sample $\mathcal{S}$: $\mathrm{err}_{\mathcal{S}}^{(p)}(A) = \left|\tilde{\mathrm{q}}_{\mathcal{S}}^{(p)}(A) - \mathrm{q}_{\mathcal{D}}^{(p)}(A)\right|$. The results are reported in Table 5, where we show the minimum, 1$^{\text{st}}$ quartile, median (i.e., 2$^{\text{nd}}$ quartile), 3$^{\text{rd}}$ quartile, and maximum of this error. The quality guarantees of MiSoSouP-1 ensure that, with probability at least $1-\delta$, the absolute error is bounded by $\varepsilon/4$ for all subgroups. A first important

[9]This property of sampling-based approaches is also the reason why we did not perform evaluate the scalability of MiSoSouP as the dataset size grows.

[10]For extremely small values of ε and only moderately large datasets, MiSoSouP-1 would also require a sample size larger than the datasets. This weakness is implicit in all sampling-based approaches, but for MiSoSouP-1, it appears at much smaller values of ε than for UB-1.

Table 5. Accuracy (absolute error) evaluation — $p = 1$

		Absolute error ($\times 10^4$)														
		Car					Mushroom					Tic-Tac-Toe				
k	$\frac{\varepsilon}{4} \times 10^4$	Min	1st Q	Med	3rd Q	Max	Min	1st Q	Med	3rd Q	Max	Min	1st Q	Med	3rd Q	Max
10	125	< 0.01	0.32	0.76	1.87	25.50	18.72	22.96	29.99	35.95	39.19	0.03	3.95	8.59	16.19	48.88
	50	0.01	1.14	2.48	4.50	8.20	7.29	7.70	13.51	15.96	16.37	< 0.01	2.17	4.66	9.38	28.58
	25	0.66	2.69	3.44	5.10	10.78	0.16	0.55	0.74	1.06	2.32	0.43	1.37	2.46	4.43	5.91
	18.75	0.01	0.77	1.88	2.61	4.45	4.46	5.30	5.41	5.64	6.09	0.09	0.48	1.35	1.75	5.14
50	125	< 0.01	0.32	0.76	1.88	25.50	5.82	14.92	20.23	25.12	39.19	< 0.01	0.82	1.93	3.90	48.88
	50	< 0.01	0.13	0.30	0.74	8.20	2.16	8.73	13.49	15.98	18.83	< 0.01	1.73	3.68	6.70	28.58
	25	< 0.01	0.95	1.81	2.87	10.78	0.16	0.93	2.32	2.91	4.71	0.01	1.03	1.73	3.02	6.01
	18.75	< 0.01	0.56	1.13	2.07	4.45	3.01	3.95	5.00	6.09	6.45	0.04	0.56	1.18	1.73	5.14
100	125	< 0.01	0.32	0.76	1.88	25.50	3.79	12.90	20.23	26.09	44.35	< 0.01	0.85	1.99	3.95	48.88
	50	< 0.01	0.14	0.32	0.72	8.20	1.36	4.86	9.14	13.89	18.83	< 0.01	1.65	3.31	5.91	28.58
	25	< 0.01	0.46	0.99	1.89	10.78	0.05	1.37	2.53	3.76	4.71	0.01	0.80	1.36	2.44	6.01
	18.75	< 0.01	0.43	0.84	1.46	4.98	2.44	3.85	4.54	5.30	6.79	0.04	0.66	1.18	1.95	5.14
200	125	< 0.01	0.32	0.76	1.88	25.50	2.07	8.60	17.17	23.29	45.75	< 0.01	0.86	2.04	4.07	48.88
	50	< 0.01	0.14	0.32	0.72	8.20	1.36	4.14	6.76	11.36	19.98	< 0.01	1.11	2.29	3.92	28.58
	25	< 0.01	0.06	0.14	0.42	10.78	0.05	1.47	2.87	3.76	5.30	0.01	0.73	1.35	2.29	6.01
	18.75	< 0.01	0.26	0.60	1.09	4.98	2.44	3.76	4.46	5.20	8.51	< 0.01	0.49	1.01	1.66	5.14
500	125	< 0.01	0.32	0.76	1.88	25.50	2.07	8.14	12.56	20.77	45.75	< 0.01	0.87	2.04	4.11	48.88
	50	< 0.01	0.14	0.32	0.73	8.20	0.40	5.31	6.56	8.41	22.51	< 0.01	0.32	0.74	1.52	28.58
	25	< 0.01	0.07	0.15	0.36	10.78	0.05	1.81	3.27	4.31	5.53	< 0.01	0.41	0.87	1.63	6.01
	18.75	< 0.01	0.05	0.11	0.26	4.98	0.86	3.56	4.10	4.88	8.85	< 0.01	0.35	0.79	1.36	5.14
1000	125	< 0.01	0.32	0.76	1.88	25.50	0.17	8.22	13.35	21.09	45.75	< 0.01	0.87	2.04	4.11	48.88
	50	< 0.01	0.14	0.32	0.73	8.20	0.40	5.63	6.56	8.45	22.86	< 0.01	0.34	0.77	1.53	28.58
	25	< 0.01	0.07	0.15	0.36	10.78	0.05	2.53	4.27	4.63	7.45	< 0.01	0.32	0.68	1.21	7.86
	18.75	< 0.01	0.05	0.11	0.26	4.98	0.05	3.41	3.77	4.43	8.85	< 0.01	0.29	0.64	1.11	5.14
2000	125	< 0.01	0.32	0.76	1.88	25.50	0.01	6.15	11.67	17.09	45.75	< 0.01	0.87	2.04	4.11	48.88
	50	< 0.01	0.14	0.32	0.73	8.20	0.02	4.14	6.56	9.10	22.86	< 0.01	0.34	0.78	1.56	28.58
	25	< 0.01	0.07	0.15	0.36	10.78	0.02	1.73	3.28	4.38	7.45	< 0.01	0.17	0.39	0.76	7.86
	18.75	< 0.01	0.05	0.11	0.26	4.98	< 0.01	2.59	3.56	4.30	8.85	< 0.01	0.21	0.44	0.80	5.14

result is that the above was true in *all* the thousands of runs of MiSoSouP-1 we performed, i.e., not just with probability $1 - \delta$. Hence MiSoSouP-1 has, in practice, even higher confidence than it guarantees theoretically. We will further comment later on this aspect.

We can see that not only the maximum absolute error was approximately between two to seven times smaller than the maximum allowed ($\varepsilon/4$), but the majority of the distribution of the error (over the subgroups) is highly concentrated around values that are often orders of magnitude smaller, with the median being at times even more than 100 times smaller than $\varepsilon/4$. Additionally we see how, as ε decreases, the distribution of the error becomes more concentrated, with the maximum values decreasing faster than the third quartiles and the medians.

A possible explanation for the fact that the estimation of the 1-qualities is much better than what is guaranteed by the theory is that the analysis uses an *upper bound* to the pseudodimension, which itself is a *worst-case* measure of complexity. This looseness is somewhat inevitable, but it suggests that there is room for improvement in the analysis. We plan to investigate the use of Rademacher averages [14] to obtain tighter sample-dependent bounds to the deviations of the sample qualities from their exact values.

5.6.3 Output properties. The set of subgroups returned by MiSoSouP-1 is a superset of $\mathrm{TOP}_p(k, \mathcal{D})$. This was always the case in all the runs, so the *recall* of MiSoSouP-1 is, in practice, 100%. MiSoSouP-1 therefore effectively exceeds the theoretical guarantees it offers. As for the precision, we must remark that a sampling-based algorithm can obviously not guarantee 100% precision, especially if it gives 100% recall like MiSoSouP-1 does.

Nevertheless, MiSoSouP-1 guarantees that *False Positives (FP)*, i.e., subgroups not in $\mathrm{TOP}_p(k, \mathcal{D})$ that may be included in the output, can only be among those subgroups with 1-quality in $\mathcal{D}$ at

least $r^{(1)}_{\mathcal{D}}(k) - \varepsilon$, i.e., at most ε less than the 1-quality of the top-k-th subgroup in $\mathcal{D}$. The number of these "acceptable" FP depends on the distribution of the 1-qualities in the dataset, and cannot be controlled by the algorithm. Thus, the precision may be very low if there are many (potentially $\gg k$) subgroups that would be acceptable FP, and these FP are the price to pay for the speed-up in analyzing the dataset. It is arguable that in these cases the exact choice of k becomes somewhat arbitrary, because there are many subgroups with p-qualities very close to each other. In any case, the output of MiSoSouP-1 is a superset of $\mathsf{TOP}_1(k, \mathcal{D})$ and can be refined to obtain this set with a fast linear scan of the dataset.

We report in Table 6, the number of FP in the output and to what percentage of the acceptable FP that number corresponds to. As expected, for a fixed value of k, the number of FP included in the output decreases as ε becomes smaller, but notice that the percentage may not decrease because the set of acceptable FP changes with ε. The absolute number of FP tends to grow with k, because the number of acceptable FP also tends to grow with k, which is a consequence of the power-law distribution of the qualities of the subgroups.

For the Car dataset, the distribution of the FP is denser than in the other datasets, therefore the percentages are often high. It is in such cases that the choice of k reveals its arbitrary nature w.r.t. the qualities of the subgroups considered of interest, and this nature is made visible but not caused by the properties of MiSoSouP-1.

In the end, the amount of FP is either a small number (either in absolute terms or relatively to k) or a relatively small fraction of the total number of acceptable FP. This fact can be explained by the "excessive" accuracy of MiSoSouP-1 in estimating the 1-quality of the subgroups, as discussed in Sect. 5.6.2. As mentioned, MiSoSouP-1 gives no guarantees that only a small subset of the acceptable FP would be included in the output, so the fact that in most cases less than half of them are actually present is a witness to the good performances of the algorithm.

5.7 Results for $p = 2$

We now discuss the results for $p = 2$. Most of the results are qualitatively similar to those for $p = 1$, so we only give additional details where the results differ.

5.7.1 Sample size reduction and speed-up. The results for the sample size and the runtime are reported in Table 7. We compare the performances of MiSoSouP-2 with those of UB-2, which uses (18) to compute the sample size. We can see that the improvement of MiSoSouP-2 over UB-2 in terms of the used sample size is similar to the case for $p = 1$. In absolute, the sample sizes are actually quite larger than those used by MiSoSouP-1 and UB-1, due to the different dependency on ε. When comparing the runtimes, it is now even clearer than in the case for $p = 1$ how the runtime improvement converges fast to the improvement in the sample size.

5.7.2 Accuracy. In Table 8 we present the statistics on the absolute error in the estimation of the 2-quality of the subgroups in the output of MiSoSouP-2. A comparison of the Max. and the $\varepsilon/4$ columns reveals that MiSoSouP-2 is between 4 and 13 times more accurate than guaranteed, even more than MiSoSouP-1. The whole distribution of the error is actually more concentrated towards zero than it was the case for $p = 1$. This fact can be explained by the additional looseness in the derivation of the sample size used by MiSoSouP-2.

5.7.3 Output properties. We report the results on the False Positives included in the output of MiSoSouP-2 in Table 9. Qualitatively, they are the same as for the $p = 1$ case, so we do not comment them further.

Table 6. Output evaluation — $p = 1$

		Car			Mushroom			Tic-Tac-Toe		
ε	k	$\lvert\text{TOP}_1(k,\mathcal{D})\rvert$	FP	% of all Acceptable FP	$\lvert\text{TOP}_1(k,\mathcal{D})\rvert$	FP	% of all Acceptable FP	$\lvert\text{TOP}_1(k,\mathcal{D})\rvert$	FP	% of all Acceptable FP
0.05	10	10	3218	99.41	10	29	19.72	11	386	1.01
	50	54	3188	99.84	50	120	22.64	53	28100	73.39
	100	105	3139	99.90	100	232	25.86	107	36738	96.06
	200	211	3034	99.93	200	399	32.83	207	37887	99.31
	500	549	2696	99.93	546	764	34.80	509	37778	99.82
	1000	1340	1905	99.90	1013	850	17.63	1003	37302	99.86
	2000	2357	888	99.78	2004	4030	15.26	2017	36312	99.93
0.02	10	10	74	2.35	10	14	56.00	11	23	11.50
	50	54	2599	81.58	50	48	57.83	53	230	5.96
	100	105	3051	97.26	100	57	40.42	107	373	1.38
	200	211	2970	97.99	200	141	51.64	207	2058	5.57
	500	549	2679	99.48	546	361	59.66	509	29526	78.50
	1000	1340	1900	99.74	1013	284	42.83	1003	34868	93.69
	2000	2357	883	99.44	2004	949	31.96	2017	35090	96.75
0.01	10	10	15	20.00	10	2	14.28	11	6	27.27
	50	54	157	6.04	50	17	36.95	53	54	26.21
	100	105	551	18.06	100	26	45.61	107	101	27.01
	200	211	2392	80.54	200	46	34.07	207	245	13.02
	500	549	2604	97.20	546	67	18.55	509	2301	7.76
	1000	1340	1841	96.89	1013	129	41.88	1003	6486	18.61
	2000	2357	824	93.32	2004	455	48.50	2017	21745	61.92
0.0075	10	10	11	40.74	10	2	100.00	11	4	66.67
	50	54	77	13.25	50	8	34.78	53	53	58.89
	100	105	301	11.81	100	26	78.78	107	36	19.15
	200	211	609	20.69	200	35	35.00	207	186	29.43
	500	549	2376	90.27	546	47	16.60	509	1160	14.10
	1000	1340	1816	96.19	1013	92	46.23	1003	3062	11.70
	2000	2357	799	91.73	2004	246	36.71	2017	8947	27.01

Table 7. Sample size and runtime evaluation for MiSoSouP-2

Dataset	ε	$\lvert\mathcal{S}\rvert$	Reduction w.r.t. UB-2	Runtime (s)	Reduction w.r.t. UB-2
Car	0.05	213873	-29.49%	2.19	-8.07%
	0.02	1331733		5.86	-20.20%
	0.01	5320295		17.46	-28.43%
	0.0075	9455351		30.30	-25.10%
Mushroom	0.05	419951	-15.15%	573.03	-13.29%
	0.02	2614931		3446.71	-15.11%
	0.01	10446693		13562.23	-14.85%
	0.0075	18566105		23857.04	-15.03%
Tic-Tac-Toe	0.05	291152	-21.34%	11.52	-18.04%
	0.02	1812932		65.82	-20.50%
	0.01	7242694		244.89	-20.02%

6 CONCLUSIONS

We introduced MiSoSouP, the first family of algorithms based on random sampling that compute probabilistically-guaranteed high-quality approximations of the collection of the top-k most interesting subgroups in a dataset, according to different popular interestingness measures. To achieve this result, we show a novel formulation of 1-quality as an average of a carefully tailored function, which we then extend to 2-quality and 1/2-quality, in order to cover important practical uses and showcase the flexibility of our approach. Our analysis relies on pseudodimension, a fundamental

Table 8. Accuracy (absolute error) evaluation — $p = 2$

		Absolute error ($\times 10^4$)														
		Car					Mushroom					Tic-Tac-Toe				
k	$\frac{\varepsilon}{4} \times 10^4$	Min	1st Q	Med	3rd Q	Max	Min	1st Q	Med	3rd Q	Max	Min	1st Q	Med	3rd Q	Max
10	125	< 0.01	0.20	0.45	1.07	38.30	2.20	10.17	12.86	14.45	20.21	< 0.01	< 0.01	0.01	0.02	10.77
	50	0.02	1.27	2.85	5.47	9.30	0.05	0.96	1.48	2.41	4.27	< 0.01	< 0.01	< 0.01	0.01	3.73
	25	0.18	1.15	1.99	3.10	5.80	0.01	0.21	0.36	0.43	0.58	0.01	0.13	0.24	0.39	1.87
	18.75	0.24	0.69	1.78	2.94	5.46	0.22	0.44	0.68	0.83	1.17					
50	125	< 0.01	0.20	0.45	1.10	38.30	1.08	8.19	12.80	13.69	20.21	< 0.01	< 0.01	0.01	0.02	10.77
	50	< 0.01	0.07	0.16	0.38	9.30	0.05	1.47	2.41	3.08	4.47	< 0.01	< 0.01	< 0.01	0.01	3.73
	25	< 0.01	0.56	1.33	2.14	6.97	0.01	0.09	0.36	0.49	1.6	< 0.01	< 0.01	< 0.01	< 0.01	1.875
	18.75	0.01	0.59	1.34	2.06	5.46	0.22	0.45	0.71	1.00	1.66					
100	125	< 0.01	0.20	0.45	1.10	38.30	1.08	5.60	10.68	13.41	20.21	< 0.01	< 0.01	0.01	0.02	10.77
	50	< 0.01	0.07	0.17	0.40	9.30	0.05	1.41	1.97	2.71	4.72	< 0.01	< 0.01	< 0.01	0.01	3.73
	25	< 0.01	0.30	0.68	1.29	6.97	0.01	0.30	0.51	0.80	1.65	< 0.01	< 0.01	< 0.01	< 0.01	1.87
	18.75	0.01	0.35	0.75	1.48	5.46	0.22	0.74	1.07	1.29	1.74					
200	125	< 0.01	0.20	0.45	1.10	38.30	1.08	4.75	5.89	12.88	20.21	< 0.01	< 0.01	0.01	0.02	10.77
	50	< 0.01	0.07	0.17	0.41	9.30	0.05	1.48	1.97	2.65	4.72	< 0.01	< 0.01	< 0.01	0.01	3.73
	25	< 0.01	0.04	0.08	0.18	6.97	0.01	0.23	0.49	0.68	1.65	< 0.01	< 0.01	< 0.01	< 0.01	1.87
	18.75	< 0.01	0.20	0.45	0.86	5.46	0.22	0.76	1.07	1.29	1.74					
500	125	< 0.01	0.20	0.45	1.10	38.30	< 0.01	1.43	2.57	4.81	20.21	< 0.01	< 0.01	0.01	0.02	10.77
	50	< 0.01	0.07	0.17	0.41	9.30	0.01	0.17	0.32	1.68	4.72	< 0.01	< 0.01	< 0.01	0.01	3.73
	25	< 0.01	0.04	0.09	0.20	6.97	0.01	0.24	0.51	0.88	2.02	< 0.01	< 0.01	< 0.01	< 0.01	1.87
	18.75	< 0.01	0.03	0.07	0.15	5.46	0.05	0.79	1.08	1.25	1.74					
1000	125	< 0.01	0.20	0.45	1.10	38.30	< 0.01	0.04	0.09	0.26	20.21	< 0.01	< 0.01	0.01	0.02	10.77
	50	< 0.01	0.07	0.17	0.41	9.30	0.01	0.18	0.42	1.60	4.72	< 0.01	< 0.01	< 0.01	0.01	3.73
	25	< 0.01	0.04	0.09	0.20	6.97	0.01	0.42	0.60	0.68	2.02	< 0.01	< 0.01	< 0.01	< 0.01	1.87
	18.75	< 0.01	0.03	0.07	0.16	5.46	0.03	1.03	1.09	1.15	1.74					
2000	125	< 0.01	0.20	0.45	1.10	38.30	< 0.01	0.04	0.09	0.26	31.26	< 0.01	< 0.01	0.01	0.02	10.77
	50	< 0.01	0.07	0.17	0.41	9.30	< 0.01	0.09	0.28	0.50	4.72	< 0.01	< 0.01	< 0.01	0.01	3.73
	25	< 0.01	0.04	0.09	0.20	6.97	< 0.01	0.19	0.44	0.64	2.02	< 0.01	< 0.01	< 0.01	< 0.01	1.87
	18.75	< 0.01	0.03	0.07	0.16	5.46	< 0.01	0.31	0.55	1.09	1.74					

Table 9. Output evaluation — $p = 2$

		Car			Mushroom			Tic-Tac-Toe		
ε	k	$\lvert\text{TOP}_1(k, \mathcal{D})\rvert$	FP	% of all Acceptable FP	$\lvert\text{TOP}_1(k, \mathcal{D})\rvert$	FP	% of all Acceptable FP	$\lvert\text{TOP}_1(k, \mathcal{D})\rvert$	FP	% of all Acceptable FP
0.05	10	10	6195	98.95	10	140	22.01	13	38343	100.00
	50	54	6212	99.92	51	386	21.53	51	38305	100.00
	100	105	6161	99.92	103	490	5.59	107	38249	100.00
	200	211	6058	99.97	201	672	0.67	213	38143	100.00
	500	500	5769	99.97	501	4191	4.21	515	37841	100.00
	1000	1019	5250	99.96	1010	412867	417.08	1001	37355	100.00
	2000	2808	3461	99.94	2000	422868	431.50	2045	36311	100.00
0.02	10	10	60	0.97	10	24	32.88	13	38317	99.93
	50	54	5623	90.55	51	54	18.82	51	38301	99.99
	100	105	6075	98.60	103	235	65.10	107	38245	99.99
	200	211	5995	99.01	201	209	62.02	213	38139	99.99
	500	500	5752	99.76	501	837	44.22	515	37839	99.99
	1000	1019	5245	99.90	1010	856	11.41	1001	37355	100.00
	2000	2808	3456	99.86	2000	9503	9.70	2045	36311	100.00
0.01	10	10	17	22.67	10	19	79.17	13	40	0.10
	50	54	164	2.92	51	18	38.30	51	38273	99.93
	100	105	789	12.99	103	108	41.70	107	38233	99.97
	200	211	5420	90.42	201	99	47.60	213	38127	99.97
	500	500	5677	98.70	501	145	17.32	515	37827	99.97
	1000	1019	5161	98.40	1010	486	55.16	1001	37343	99.97
	2000	2808	3397	98.29	2000	1482	15.70	2045	36299	99.97
0.0075	10	10	11	40.74	10	19	100.00			
	50	54	74	11.42	51	12	37.50			
	100	105	306	5.49	103	102	63.75			
	200	211	1030	17.26	201	33	18.13			
	500	500	5448	95.50	501	97	25.94			
	1000	1019	5156	98.53	1010	442	71.64			
	2000	2808	3372	97.91	2000	952	23.29			

concept from statistical learning theory. This connection is novel for subgroup discovery. We show upper and almost-matching lower bounds to the pseudodimension of the task, and show that it depends on characteristics of the dataset and of the language of interest.

Our experimental evaluation shows that MiSoSouP requires much smaller sample sizes than state-of-the-art solutions to obtain approximations with the same guarantees, therefore providing the first viable tool to efficiently identify the most interesting subgroups for ever-more-massive datasets.

Our algorithms hinge on defining quality measures as averages of specific functions. This approach can likely be used in concert with Rademacher averages to design progressive-sampling methods for subgroups discovery, as done for other mining tasks [26]. Investigating this approach is an interesting direction for future research.

ACKNOWLEDGMENTS

This work is supported, in part by the National Science Foundation grant IIS-1247581 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1247581) and by the University of Padova grants SID2017 and STARS: Algorithms for Inferential Data Mining.

REFERENCES

[1] Martin Anthony and Peter L. Bartlett. 1999. *Neural Network Learning – Theoretical Foundations*. Cambridge University Press.

[2] Martin Atzmueller. 2015. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5, 1 (2015), 35–49.

[3] Michele Borassi and Emanuele Natale. 2016. KADABRA is an ADaptive Algorithm for Betweenness via Random Approximation. In *24th Annual European Symposium on Algorithms (ESA '16)*. 20:1–20:18.

[4] Guillaume Bosc, Jean-François Boulicaut, Chedy Raïssi, and Mehdi Kaytoue. 2018. Anytime discovery of a diverse set of patterns with Monte Carlo tree search. *Data Mining and Knowledge Discovery* 32, 3 (2018), 604–650.

[5] Victor De la Peña and Evarist Giné. 1999. *Decoupling: from dependence to independence*. Springer.

[6] Wouter Duivesteijn, Ad Feelders, and Arno Knobbe. 2012. Different slopes for different folks: mining for exceptional regression models with Cook's distance. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*. ACM, 868–876.

[7] Tapio Elomaa and Matti Kääriäinen. 2002. Progressive Rademacher Sampling. In *AAAI/IAAI*, Rina Dechter and Richard S. Sutton (Eds.). AAAI Press / The MIT Press, 140–145.

[8] Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. 2011. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems* 29, 3 (2011), 495–525.

[9] Wassily Hoeffding. 1963. Probability Inequalities for Sums of Bounded Random Variables. *J. American Statistical Assoc.* 58, 301 (1963), 13–30.

[10] Matti Kääriäinen, Tuomo Malinen, and Tapio Elomaa. 2004. Selective Rademacher Penalization and Reduced Error Pruning of Decision Trees. *Journal of Machine Learning Research* 5 (Dec. 2004), 1107–1126.

[11] Willi Klösgen. 1992. Problems for knowledge discovery in databases and their treatment in the Statistics Interpreter Explora. *International Journal of Intelligent Systems* 7 (1992), 649–673.

[12] Willi Klösgen. 1995. Assistant for knowledge discovery in data. In *Assisting Computer: A New Generation of Support Systems*, P. Hoschka (Ed.).

[13] Willi Klösgen. 1996. Explora: A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, 249–271.

[14] Vladimir Koltchinskii. 2001. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory* 47, 5 (July 2001), 1902–1914.

[15] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. 2009. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research* 10, Feb (2009), 377–403.

[16] Yi Li, Philip M. Long, and Aravind Srinivasan. 2001. Improved Bounds on the Sample Complexity of Learning. *J. Comput. System Sci.* 62, 3 (2001), 516–527.

[17] M. Lichman. 2013. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

[18] 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 *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 422–436.

[19] Michael Mitzenmacher and Eli Upfal. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.
[20] Sandy Moens and Mario Boley. 2014. Instant exceptional model mining using weighted controlled pattern sampling. In *International Symposium on Intelligent Data Analysis*. Springer, 203–214.
[21] Yotam Ottolenghi. 2012. Yotam Ottolenghi's recipes for char-grilled sprouting broccoli with sweet tahini, plus gingery fish balls in miso soup. https://www.theguardian.com/lifeandstyle/2012/feb/03/grilled-broccoli-fishball-soup-recipes.
[22] Gregory Piatetsky-Shapiro. 1991. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases* (1991), 229–248.
[23] David Pollard. 1984. *Convergence of stochastic processes*. Springer-Verlag.
[24] Theodoros Rekatsinas, Manas Joglekar, Hector Garcia-Molina, Aditya Parameswaran, and Christopher Ré. 2017. SLiMFast: Guaranteed Results for Data Fusion and Source Reliability. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17)*. ACM, New York, NY, USA, 1399–1414.
[25] Matteo Riondato and Eli Upfal. 2014. Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees. *ACM Trans. Knowl. Disc. from Data* 8, 4 (2014), 20. https://doi.org/10.1145/2629586
[26] Matteo Riondato and Eli Upfal. 2015. Mining Frequent Itemsets through Progressive Sampling with Rademacher Averages. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, 1005–1014.
[27] Matteo Riondato and Eli Upfal. 2018. ABRA: Approximating Betweenness Centrality in Static and Dynamic Graphs with Rademacher Averages. *ACM Trans. Knowl. Disc. from Data* 12, 5 (2018), 1–38.
[28] Matteo Riondato and Fabio Vandin. 2014. Finding the True Frequent Itemsets. In *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, Mohammed Javeed Zaki, Zoran Obradovic, Pang-Ning Tan, Arindam Banerjee, Chandrika Kamath, and Srinivasan Parthasarathy (Eds.). SIAM, 497–505. https://doi.org/10.1137/1.9781611973440.57
[29] Matteo Riondato and Fabio Vandin. 2018. MiSoSouP: Mining Interesting Subgroups with Sampling and Pseudodimension. In *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Disc. and Data Mining (KDD '18)*. ACM, 2130–2139.
[30] Tobias Scheffer and Stefan Wrobel. 2002. Finding the most interesting patterns in a database quickly by using sequential sampling. *J. Mach. Learn. Res.* 3 (Dec. 2002), 833–862.
[31] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
[32] Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. 2013. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences* 110, 32 (2013), 12996–13001.
[33] Matthijs van Leeuwen and Arno Knobbe. 2011. Non-redundant subgroup discovery in large and complex data. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD '11)*. 459–474.
[34] Matthijs van Leeuwen and Antti Ukkonen. 2016. Expect the Unexpected – On the Significance of Subgroups. In *Proceedings of Discovery Science (DS '16)*.
[35] Vladimir N. Vapnik. 1998. *Statistical learning theory*. Wiley.
[36] Vladimir N. Vapnik and Alexey J. Chervonenkis. 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications* 16, 2 (1971), 264–280. https://doi.org/10.1137/1116025
[37] Stefan Wrobel. 1997. An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '97)*. 78–87.
[38] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. 2016. Adaptive Concentration Inequalities for Sequential Decision Problems. In *Advances In Neural Information Processing Systems (NIPS '16)*. 1343–1351.

A ON THE CORRECTNESS OF THE GSS ALGORITHM

In this section we present our concerns on the analysis of the GSS algorithm by Scheffer and Wrobel [30]. We borrow the same notation used in [30].

A.1 Random stopping time and hypothesis sets

The proofs of [30, Lemmas 13 and 14] do not appear to be correct.

Let's start with the proof for [30, Lemma 14] first, as it is easier to explain and it will clarify the situation for [30, Lemma 13]. The issue is that the value i_{max} is a random variable, and so is the set $H_{i_{max}}$, and thus its size. All these quantities are functions of the sequence of random variables $(t_i)_{i\geq 1}$, where t_i is the transaction sampled by the algorithm at time i.[11] In [30, Lemma 14], all these

[11] To be specific, i_{max} is a *stopping time* [19, Section 12.2] for the sequence of random variables $(t_i)_{i\geq 1}$.

quantities are assumed to be fixed, i.e., the results are obtained *conditioned* on the realized values of i_{max} and $H_{i_{max}}$. One may wish to believe that it is therefore sufficient to apply the law of total probability [19, Thm. 1.6] over all possible values of i_{max} and $H_{i_{max}}$, and given that each conditional probability is bounded by $\delta/2$, then so is the unconditional probability. As observed by Borassi and Natale [3, Section 3], this approach is not justified a priori. The solution presented by Borassi and Natale [3, Section 2], although for a different problem, can only partially be adapted to GSS, as it does not solve the problem that the set H_i of hypotheses under consideration at iteration i is a random variable for every $i > 1$. One may be tempted to try to fix the correctness of the algorithm, at least as far as this issue is concerned, by removing the possibility for the algorithm to stop when $E(i, \delta/(2|H_i|))$ is less than or equal to $\varepsilon/2$ (third condition on line 3 of the pseudocode in [30, Table 1]), and replace it with a condition that is satisfied when the number of sampled transactions is equal to M (defined on line 2 of [30, Table 1]), but these changes would not be sufficient to guarantee the correctness of the algorithm, due to additional issues that we discuss in App. A.2.

For more details on the topic of randomly stopped sequences of random variables, we refer the reader to the book by De La Peña and Giné [5, Ch. 2] and for recent developments to the paper by Zhao et al. [38].

The issue with [30, Lemma 13] is similar to the one described above: the event in the probability on the line above [30, Equation 131] is defined conditioned on the sets H_i, but this conditioning must be justified and handled appropriately. In order to *try* to fix this issue, one could consider the worst case when all the H_i equal H. Then the assumption A1 would be modified to have $|H|$ instead of $|H_i|$ and the same modification would be made in the algorithm on lines $3.e.i$ and $3.e.ii$ in [30, Table 1], and to [30, Equations 85 and 104]. Even these changes are not sufficient due to the additional issues that we discuss in the following section.

A.2 Use of the Chernoff bounds

Scheffer and Wrobel [30, Section 4.1] discuss how to use Chernoff bounds [19, Chapter 4] to compute a probabilistic tail bound for the deviation of a relative frequency (i.e., a sample average of a 0–1 binary function) from its expectation. The function $E(m, \delta)$ in [30, Equation 4] gives a value such that the probability that a sample average computed from a sample of m transactions deviates from its expectation by more than $E(m, \delta)$ is at most δ, as shown in [30, Equations 5–7].

The correctness of the above statement, i.e., of the properties of $E(m, \delta)$ depends crucially on the fact that the sample average is the average of m values. Such is definitively the case for the functions presented in [30, Section 4.1], but not for the functions presented in [30, Sections 4.2, 4.3, 4.4]. This fact is recognized by Scheffer and Wrobel [30], who develop sequences of insightful inequalities to show how to upper bound the probability that a sample estimate of the p-quality, $p \in \{1/2, 1, 2\}$, deviates from its expectations by more than some quantity with a sum of the probabilities that the sample estimates of the generality and the usefulness deviate from their expectation by more than some other quantities. The derivations of these upper bound are presented in [30, Equation 16–22] for the 1-quality, [30, Equation 36–43] for the 2-quality,[12] and [30, Equation 59–67] for the 1/2-quality.

The issue in these derivations is that Scheffer and Wrobel[30] consider the sample estimate for the unusualness (which they denote as $\hat{p}$) as the *average of a 0-1 function over the m elements in the sample* (the function is 0 if the target attribute is 0, 1 otherwise), but this is not correct. For a subgroup $A \in \mathcal{L}$, one could see $\hat{p}$ of A as the average of such 0-1 function *only over the cover of A on the sample*, i.e., over *at most m* transactions, but potentially (and, in practice, often) many fewer than m. The consequences are quite impactful. For example, the rightmost probability in [30,

[12]There is a typo in the second line of [30, Equation 41]: the ">" sign should be "<".

Equation 42] is (implicitly) upper bounded in [30, Equation 43] using an exponential quantity obtained from the Chernoff bound, but this quantity is valid only for averages of 0-1 functions over m elements, which $\hat{p}$, as discussed, is not. Essentially the same issue is observed in the passage between [30, Equation 63] and [30, Equation 64], and in the passage between [30, Equation 21] and [30, Equation 22]. In the analysis of MiSoSouP, we solve these issue by defining the function ρ_A, $A \in \mathcal{L}$, as a non-binary function over all transactions, not just those in the cover of A.

Because of the issues mentioned in the previous paragraph, the "instantiation" of the GSS algorithm for the p-qualities are not correct. One can instead use the function $\rho_A(t)$ defined in (3): its average over *all* transactions in the sample is an estimate for the 1-quality (see Sect. 4.1), and this can be combined with the sample average of the generality g_A to obtain an estimate for the 2-quality, as we do in Sect. 4.2. Fixing the result for the 1/2-quality seems more complicate, as it was also deriving MiSoSouP-1/2.

Additionally, instead of using the Chernoff bound to compute tail bounds for the sample estimation of unusualness, one must use Hoeffding's inequality, as we discuss in Sect. 4.1.3.