

# On the Complexity of Mining Itemsets from the Crowd Using Taxonomies

Antoine Amarilli<sup>1,2</sup>, Yael Amsterdamer<sup>1</sup>, and Tova Milo<sup>1</sup>

<sup>1</sup>Tel Aviv University, Tel Aviv, Israel

<sup>2</sup>École normale supérieure, Paris, France

## ABSTRACT

We study the problem of frequent itemset mining in domains where data is not recorded in a conventional database but only exists in human knowledge. We provide examples of such scenarios, and present a crowdsourcing model for them. The model uses the crowd as an oracle to find out whether an itemset is frequent or not, and relies on a known taxonomy of the item domain to guide the search for frequent itemsets. In the spirit of data mining with oracles, we analyze the complexity of this problem in terms of (i) crowd complexity, that measures the number of crowd questions required to identify the frequent itemsets; and (ii) computational complexity, that measures the computational effort required to choose the questions. We provide lower and upper complexity bounds in terms of the size and structure of the input taxonomy, as well as the size of a concise description of the output itemsets. We also provide constructive algorithms that achieve the upper bounds, and consider more efficient variants for practical situations.

## General Terms

Algorithms, Theory

## Categories and Subject Descriptors

H.2.8 [Database Applications]: Data mining

## 1. INTRODUCTION

The identification of *frequent itemsets*, namely sets of items that frequently occur together, is a basic ingredient in data mining algorithms and is used to discover interesting patterns in large data sets [1]. A common assumption in such algorithms is that the transactions to be mined (the sets of co-occurring items) have been recorded and are stored in a database. In contrast, there is data which is not recorded in a systematic manner, but only exists in human knowledge. Mining this type of data is the goal of this paper.

As a simple example, consider a social scientist analyzing the life habits of people, in terms of activities (watching TV, jogging, reading, etc.) and their contexts (time, location, weather, etc.). Typically, for large communities, there is no comprehensive database that records all transactions where an individual performs some combination of activities in a certain context. Yet, some trace of the data remains in the memories of the individuals involved. As another example, consider a health researcher who wants to identify new drugs by analyzing the practices of folk medicine (also known as traditional medicine, i.e., medicinal practice that is neither documented in writing nor tested out under a scientific protocol): the researcher may want to determine, for instance, which treatments are often applied together for a given combination of symptoms. For this purpose too, the main source of knowledge are the folk healers and patients themselves.

In a previous work [3, 4] we have proposed to address this challenge using *crowdsourcing* to mine the relevant information from the crowd. Crowdsourcing platforms (such as, e.g., [4, 14, 28, 30, 34]) are an effective tool for harnessing a crowd of Web users to perform various tasks. In [3, 4] we incorporated crowdsourcing into a *crowd mining framework* for identifying frequent data patterns in human knowledge, and demonstrated its efficiency experimentally. The goal of the present paper is to develop the theoretical foundations for crowd mining, and, in particular, to formally study the complexity of identifying frequent itemsets using the crowd.

Before presenting our results, let us explain three important principles that guide our solution.

First, in our settings, *no comprehensive database can be built*. Not only would it be prohibitively expensive to ask all the relevant people to provide all the required information, but it is also impossible for people to recall all the details of their individual transactions such as activity occurrences, illnesses, treatments, etc. [3, 7]. Hence, one cannot simply collect the transactions into a database that could be mined directly. Instead, studies show that people do remember some summary information about their transactions [7], and thus, as demonstrated in [3, 4], itemset frequencies can be learned by asking the crowd directly about them.

Second, as we want to mine the crowd by posing questions about itemset frequencies, we must define a suitable *cost model* to evaluate mining algorithms. In data mining there are two main approaches for measuring algorithm cost. The first one (see, e.g., [1]) measures running time, including the cost of accessing the database (database scans), which is not suitable for a crowd setting as there is no database that can

be accessed in this manner. The second approach (see, e.g., [26]) assumes the existence of an oracle that can be queried for insights about data patterns (frequency of itemsets, in our case); the cost is then measured by the number of oracle calls. Our setting is closer to this second approach: the crowd serves as an oracle, and we count the number of questions posed to the crowd, namely, *crowd complexity*. In addition, we study *computational complexity*, namely, the time required to compute the itemsets about which we want to ask the crowd. There is a clear tradeoff between the costs: investing more computational effort to select questions carefully may reduce the crowd complexity, and vice versa. See Section 7 for a further comparison of our work with existing approaches in data mining.

Finally, for the human-knowledge domains that we consider, one can make mining algorithms more efficient by leveraging semantic knowledge captured by *taxonomies*. A taxonomy in our context is a partial “is-a” relationship on the items relevant to the domain, e.g., tennis is a sport, sport is an activity, etc. Many such taxonomies are available, both domain-specific (e.g., for diseases [36]) and general-purpose (e.g., Wordnet [31]). The use of taxonomies in mining is twofold. First, the semantic dependencies between *items* induce a frequency dependency between *itemsets*: e.g., because tennis is a sport, the itemset {*sunglasses, sport*} implicitly appears in all transactions where {*sunglasses, tennis*} appears. Hence, if the latter itemset is frequent then so is the former. Second, with taxonomical knowledge we can avoid asking questions about semantically equivalent itemsets, such as {*sport, tennis*} and {*tennis*}. Taxonomies are known to be a useful tool in data mining [37] and we study their use under our complexity measures.

**Results.** For our theoretical results, we harness tools from three areas of computer science: *data mining*, *order theory* and *Boolean function learning* [6, 8, 16, 18, 25, 26, 37]. Order theory is relevant to our discussion, because a taxonomy is in fact a partial order over data items; and Boolean function learning is relevant since the set of frequent itemsets to identify can be represented as a Boolean function indicating whether itemsets are frequent, a connection that was also pointed out in previous works in data mining [26]. Our contribution in this paper is combining and extending these tools to characterize the complexity of crowd mining.

A summary of our main results is presented in Table 1, where we give upper and lower bounds for our two complexity measures. In the first column, we give such bounds as a function of the structure of the input taxonomy  $\Psi$ . These bounds are not affected by properties of the output, such as the actual number of frequent itemsets to be identified. In contrast, in the second column, we give complexity bounds as a function of the number of maximal frequent itemsets (MFIs) and minimal infrequent itemsets (MIIs). Intuitively, the MFIs and MIIs (to be defined formally later) are alternative concise descriptions of the frequent itemsets, and thus capture the output of the mining process.

The first row of Table 1 presents crowd complexity results. We show that, given a taxonomy  $\Psi$ , the problem of identifying the frequent itemsets has a tight bound logarithmic in  $|S(\Psi)|$  – the number of possible Boolean frequency functions, which depends on  $\Psi$ . As reflected in the inequalities at the bottom of Table 1 (and explained in Section 3),  $\log |S(\Psi)|$  is at most exponential in  $|\Psi|$ . When the output is considered, our lower

complexity bound is the sum of the numbers of MFIs and MIIs, and the upper bound adds the taxonomy size as a multiplicative factor. We provide a constructive algorithm (Algorithm 1) that achieves this bound.

In the second row of Table 1, we study computational complexity. We focus on “crowd-efficient” algorithms, which achieve the crowd complexity upper bound mentioned above. The crowd complexity lower bound is trivially a lower bound of computational complexity; however, w.r.t. the output, we can obtain a stronger hardness result by showing that the problem is *EQ-hard* in the taxonomy size and in the numbers of MFIs and MIIs. EQ is a well-known problem in Boolean function learning which is not known to be solvable in PTIME [6, 17]. As for upper bounds, from Algorithm 1, we obtain an upper computational bound polynomial in  $|I(\Psi)|$  – the number of (relevant) itemsets of  $\Psi$ . This size is at most exponential in  $|\Psi|$  (see Section 2). Algorithm 1 is not crowd-efficient w.r.t. the input alone, but for the upper computational complexity bound we relax this requirement in order to achieve a more feasible bound. Omitted from Table 1 are results for the case in which the size of itemsets is bounded by a constant  $k$ , which we also study for the problem axes mentioned above.

Finally, given the relatively high lower complexity bounds, we examine two additional approaches. The *chain partitioning* approach, following a standard technique in data mining and Boolean function learning, suggests an alternative algorithm for crowd mining. We show that while this algorithm is not crowd-efficient in general, it outperforms Algorithm 1 given certain conditions on the frequent itemsets. The *greedy* approach attempts to maximize, at each question to the crowd, the number of *itemsets* that are classified as frequent or infrequent. We show that choosing a question that maximizes this number is  $\text{FP}^{\#P}$ -hard in  $|\Psi|$ .

**Paper organization.** We start in Section 2 by formally defining the setting and the problem. Crowd and computational complexity are studied in Sections 3 and 4 respectively. We consider chain partitioning in Section 5, and a greedy approach in Section 6. Related work is discussed in Section 7 and we conclude in Section 8.

## 2. PRELIMINARIES

We now present the formal model and problem settings for taxonomy-based crowd mining. Table 2 summarizes all the introduced notation. We start by recalling some basic itemset mining definitions from [1] and explain how they apply to our settings.

Let  $\mathcal{I} = \{i_1, i_2, i_3, \dots\}$  be a finite set of distinct item names. Define an *itemset* (or *transaction*)  $A$  as a subset of  $\mathcal{I}$ . Define a *database*  $D$  as a bag (multiset) of transactions.  $|D|$  denotes the number of transactions in  $D$ . The frequency or *support* of an itemset  $A \subseteq \mathcal{I}$  in  $D$  is  $\text{supp}_D(A) := |\{T \in D \mid A \subseteq T\}| / |D|$ .  $A$  is considered *frequent* if its support exceeds a predefined threshold  $\Theta$ : we assume that  $0 < \Theta < 1$  as mining is trivial when  $\Theta \in \{0, 1\}$ . Given a database  $D$ , we define the predicate  $\text{freq}(\cdot)$  which takes an itemset as input and returns true iff this itemset is frequent in  $D$  (the dependency on  $D$  is omitted from the notation).

For example, in the domain of leisure activities,  $\mathcal{I}$  may include different activities, relevant equipment, locations, etc. Each transaction  $T$  may represent all the items involved in a particular leisure event (a vacation day, a night out, etc.). If,

		With respect to the input		With respect to the input and output	
Crowd	Lower	$\Omega(\log  \mathcal{S}(\Psi) )$	(Prop. 3.6)	$\Omega( mfi  +  mii )$	(Prop. 3.10)
	Upper	$O(\log  \mathcal{S}(\Psi) )$	(Prop. 3.9)	$O( \Psi  \cdot ( mfi  +  mii ))$	(Thm. 3.12)
Comp.	Lower	$\Omega(\log  \mathcal{S}(\Psi) )$	(Cor. of Prop. 3.6)	EQ-hard	(Prop. 4.2)
	Upper	$O( \mathcal{I}(\Psi)  \cdot ( \Psi ^2 +  \mathcal{I}(\Psi) ))$	(Cor. 4.4)	$O( \mathcal{I}(\Psi)  \cdot ( \Psi ^2 +  mfi  +  mii ))$	(Prop. 4.3)

Table 1: Summary of the main complexity results, where we have  $|\mathcal{I}(\Psi)| \leq 2^{\alpha(|\Psi|)}$  and  $|\mathcal{S}(\Psi)| \leq 2^{\alpha(|\mathcal{I}(\Psi)|)}$

e.g., the set {tennis, racket, sunglasses} is frequent, it means that a racket and sunglasses are commonly used for tennis. Or, if {indoor\_cycling, TV} is frequent, it may imply that indoor cyclists often watch TV while cycling.

In our crowd-based setting, the database of interest  $D$  is not materialized and only models the knowledge of people, so we can only access  $D$  by asking them questions. As shown in [3], we can ask people for summaries of their personal knowledge, which we can then interpret as data patterns – itemset frequencies in our case. We thus abstractly model a crowd query as follows:

**DEFINITION 2.1 (CROWD QUERY).** A crowd query takes as input an itemset  $A \subseteq \mathcal{I}$  and returns  $\text{freq}(A)$ .

When using crowdsourcing to answer this type of crowd queries, and when posing questions to the crowd in general, one must deal with imprecise or partial answers. This general problem was studied in previous crowdsourcing works [3, 4, 33]. We can employ one of their methods as a black-box and assume that each crowd query is posed to a sufficient (constant) number of users, so as to gain sufficient confidence in the obtained Boolean answer. Thus, the cost of a crowd mining algorithm can be defined as *the number of crowd queries* rather than the number of posed questions: the crowd acts as an *oracle* for itemset frequency. The cost metric does not depend on the size of the hypothetical database  $D$ , nor does it depend on the number of scans that would be necessary to determine the frequency of the queried itemsets if  $D$  were materialized.

**Itemset Dependency and Taxonomies.** The support of different itemsets can be dependent. For example, if  $A \subseteq B$  then  $B \subseteq T$  implies  $A \subseteq T$  so  $\text{supp}_D(A) \geq \text{supp}_D(B)$  for every  $D$ . This fundamental property is used by classic mining algorithms such as Apriori [1].

Moreover, as noted in [37], there may be dependencies between itemsets resulting from *semantic relations* between items. For instance, in our example from the Introduction, the itemset {sunglasses, sport} is semantically implied by any transaction containing {sunglasses, tennis}, since tennis is a sport.

Such semantic dependencies can be naturally captured by a *taxonomy* [37]. Formally, we define a taxonomy as a partially ordered set (or *poset*)  $\Psi = (\mathcal{I}, \leq)$  where  $\leq$  is a partial order over the element domain  $\mathcal{I}$ .  $i \leq i'$  indicates that item  $i'$  is more specific than  $i$  (any  $i'$  is also an  $i$ ). Observe that the antisymmetry of  $\leq$  implies that no two different items in  $\mathcal{I}$

are equivalent by  $\leq$ .<sup>1</sup> We use  $i < i'$  when  $i \leq i'$  and  $i \neq i'$ , and denote by  $\prec$  the *covering relation* of  $\leq$ :  $i \prec i'$  iff  $i < i'$  and there exists no  $i''$  s.t.  $i < i'' < i'$ .

We represent posets as DAGs, whose vertices are the poset elements, and where a directed edge  $(i, i')$  indicates that  $i \prec i'$ . This is in line with standard representations of posets such as, e.g., *Hasse diagrams*. We denote by  $|\Psi| = O(|\mathcal{I}|^2)$  the size of the taxonomy including the number of elements and pairs in  $\prec$ .

**EXAMPLE 2.2.** Consider the taxonomy  $\Psi_1$  shown in Figure 1a. We can label its elements with items, e.g.: 1. *cycling*, 2. *sport*, 3. *bicycle\_touring*, 4. *indoor\_cycling*. The interpretation of the taxonomy would then be: both bicycle touring and indoor cycling are types of cycling, and indoor cycling is also a sport.

Let us briefly define some general useful terms in the context of posets. When  $i \leq i'$  we call  $i$  an *ancestor* and  $i'$  a *descendant*. Similarly, when  $i < i'$  we call  $i$  a *parent* and  $i'$  its *child*. A *chain* is a sequence of elements  $i_1 < i_2 < \dots < i_n$ . An *antichain* is a set of elements  $A = \{i_1, \dots, i_n\}$  that are incomparable with respect to  $\leq$ , i.e., there exist no  $i_j \neq i_k \in A$  s.t.  $i_j \leq i_k$ . The *width* of a poset  $P$ , denoted by  $w[P]$ , is the size of its largest antichain. An *order ideal* (or lower set)  $A$  of a poset  $P$  is a subset of its elements s.t. if  $i \in A$  then all the ancestors of  $i$  are in  $A$ .

**EXAMPLE 2.3.** The antichains of the example taxonomy  $\Psi_1$  include the empty antichain  $\{\}$ ; singleton itemsets such as  $\{3\}$ ; and antichains of size 2 such as  $\{2, 3\}$  (since 2 and 3 are incomparable). There are no larger antichains, and thus  $w[\Psi_1] = 2$ .

We denote by  $\mathcal{AC}[\Psi]$  the domain of antichains of elements from  $\Psi$ . Antichains are concise in the sense that they contain no items implied by other items. In this way, e.g., {tennis} concisely represents {tennis, sport}, {tennis, sport, activity}, etc. Thus, unless stated otherwise, whenever we mention itemsets we assume that the items form an antichain. This is also useful for practical purposes: it would be strange, e.g., to ask users whether they simultaneously play tennis and do sport.

Based on  $\leq$ , the semantic relationship between items, we can define the corresponding relationship between *itemsets*. For itemsets  $A, B$  we define  $A \leq B$  iff every item in  $A$  is implied by some item in  $B$ . Formally:

<sup>1</sup>Such equivalence would stand for semantic synonyms such as *cycling* and *biking*. We thus assume that every group of synonyms is represented by a single item.

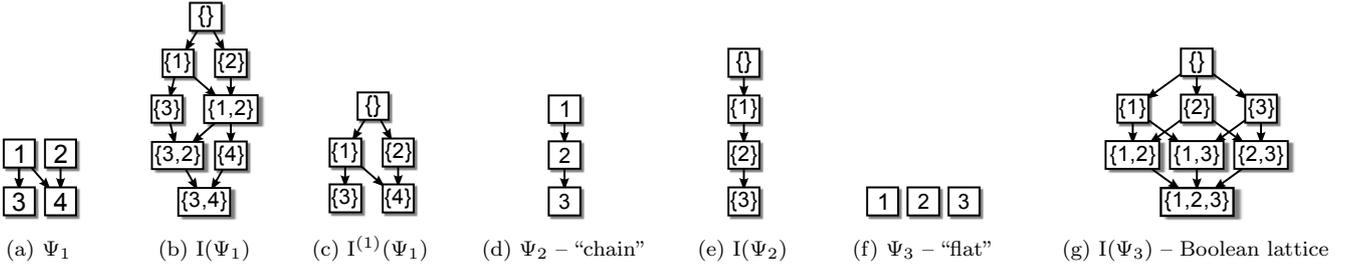


Figure 1: Example taxonomies

DEFINITION 2.4 (ITEMSET TAXONOMY). Given a taxonomy  $\Psi = (\mathcal{I}, \leq)$  we define its itemset taxonomy as the poset  $I(\Psi) = (\mathcal{AC}[\Psi], \leq)$ .

By an abuse of notation, we extend  $\leq$  to itemsets, where for every two itemsets  $A, B \in \mathcal{AC}[\Psi]$ ,  $A \leq B$  iff  $\forall i \in A, \exists i' \in B$   $i \leq i'$ . Similarly, we extend  $<$  and  $\prec$  to itemsets:  $A < B$  when  $A \leq B$  and  $A \neq B$ , and  $A \prec B$  iff  $A < B$  and there exists no  $C$  s.t.  $A < C < B$ .

Figure 1b illustrates  $I(\Psi_1)$ , the itemset taxonomy of  $\Psi_1$  from Figure 1a. Observe that, for singleton itemsets,  $\leq$  corresponds to the order on items.

Finally, we redefine support to take  $I(\Psi)$  into account.

DEFINITION 2.5 (ITEMSET SUPPORT). Let  $A \subseteq \mathcal{I}$  be an itemset. We define the support of  $A$  w.r.t. a database  $D$  and a taxonomy  $\Psi$  to be 0 if  $D$  is empty, and otherwise as

$$\text{supp}_{D, \Psi}(A) := |\{T \in D \mid A \leq T\}| / |D|$$

**Properties of the itemset taxonomy.** By construction, the itemset taxonomy  $I(\Psi)$  is not an arbitrary poset. For instance, it will always have a single “root” element, namely the empty itemset, which must precede all other elements by  $\leq$ .

More generally, the domain of all possible itemset taxonomies can be precisely characterized as the domain of all distributive lattices (see full version [2] for details).

Let us now illustrate the structure of  $I(\Psi)$  in two extreme but useful examples.

EXAMPLE 2.6. Figure 1d illustrates a total order or “chain” taxonomy, whose itemset taxonomy is a chain with one more element (Figure 1e). Figure 1f displays a “flat” taxonomy, where all the elements are incomparable. Its itemset taxonomy (Figure 1g) contains all the possible itemsets: it is the Boolean lattice structure explored by classic data mining algorithms such as Apriori [1]. Hence, if a flat  $\Psi$  has  $n$  elements,  $I(\Psi)$  has exactly  $2^n$  elements and  $\leq$  corresponds exactly to set inclusion.

**Maximal frequent itemsets.** By the definition of support,  $\text{freq}$  is a (decreasing) monotone predicate over itemsets, i.e., if  $A \leq B$  then  $\text{freq}(B)$  implies  $\text{freq}(A)$ . Consequently,  $\text{freq}$  can be uniquely and concisely characterized by a set of maximal frequent itemsets (MFIs), namely all the frequent itemsets with no frequent descendants. Equivalently, it can be characterized by a set of minimal infrequent itemsets (MIIs), which are all the infrequent itemsets with no infrequent ancestors.

MFIs and MIIs were introduced for knowledge discovery in [26] (where they are called respectively the *positive border* and *negative border*), and existing data mining algorithms such as [5] try to identify them as a concise representation of the frequent itemsets. We denote the MFIs and MIIs of a predicate  $\text{freq}$  by  $mfi$  and  $mii$  respectively, where  $\text{freq}$  is clear from the context. More generally, we call *maximal elements* the analogue of MFIs for decreasing monotone predicates over an arbitrary poset.

EXAMPLE 2.7. Consider  $I(\Psi_1)$  in Figure 1b. Assume, e.g., that we know  $\text{freq}(\{2\}) = \text{freq}(\{3\}) = \text{freq}(\{4\}) = \text{true}$ . By the monotonicity of  $\text{freq}$ , their ancestors (e.g.,  $\{1, 2\}$ ) are also frequent. Assume that  $\text{freq}$  returns false for any other itemset. Then  $\text{freq}$  can be uniquely characterized by its MFIs  $\{3\}$  and  $\{4\}$ , or by its MII  $\{3, 2\}$ : the values of  $\text{freq}$  for the other itemsets follow.

**Restricting the itemset size.** In typical crowd scenarios, there are often restrictions on the number of elements that may be presented in a crowd query [27], so it is impractical to ask users about very large itemsets. We therefore define a variant of the problem in which the itemset size is bounded from above by a constant.

DEFINITION 2.8 ( $k$ -ITEMSET TAXONOMY). We define the  $k$ -itemset taxonomy  $I^{(k)}(\Psi) := (\mathcal{AC}^{(k)}[\Psi], \leq)$  where we define  $\mathcal{AC}^{(k)}[\Psi] := \{A \in \mathcal{AC}[\Psi] \mid |A| \leq k\}$  and  $k$  is constant.

We refer to the elements of  $\mathcal{AC}^{(k)}[\Psi]$  as  $k$ -itemsets. Observe that, unlike for  $I(\Psi)$ , the number of itemsets in  $I^{(k)}(\Psi)$  is always polynomial in  $|\mathcal{I}|$ , i.e.,  $O(|\mathcal{I}|^k)$ . In addition,  $I^{(k)}(\Psi)$  need not be a distributive lattice: for instance, by setting  $k = 1$ ,  $I^{(k)}(\Psi)$  is almost identical to  $\Psi$ , i.e., it is an arbitrary poset except for the added  $\{\}$  element. Compare for example  $I^{(1)}(\Psi_1)$  (Figure 1c) with  $\Psi_1$  (Figure 1a).

**Problem statement.** Given a known taxonomy  $\Psi$  and an unknown database  $D$ , defining the (also unknown) frequency predicate  $\text{freq}$  over the itemsets in  $I(\Psi)$ , we denote by **MinFreq** the problem of identifying, using only crowd queries, all the frequent itemsets in  $D$  (or, equivalently, of identifying  $\text{freq}$  exactly). We consider *interactive* algorithms that iteratively compute, based on the knowledge collected so far, which crowd query to pose next, until **MinFreq** is solved.

As mentioned in the Introduction, we study the complexity bounds of such algorithms for two metrics. We first consider the number of crowd queries that need to be asked, namely the *crowd complexity*. Then, we study the feasibility

$\text{supp}_D(A)$	Support of itemset $A$ in database $D$
$\Theta$	Support threshold, $0 < \Theta < 1$
$\text{freq}(A)$	True iff $\text{supp}_D(A)$ exceeds $\Theta$
$\mathcal{I}$	Set of all items
$\Psi$	Taxonomy – a partial order $(\mathcal{I}, \leq)$
$i \leq i'$	Item $i'$ is more specific than $i$
$\prec$	Covering relation of $\leq$
$ \Psi $	Size of the taxonomy (as the DAG of $\prec$ )
$w[\Psi]$	Width of $\Psi$
$\mathcal{AC}[\Psi]$	Antichains of $\Psi$
$I(\Psi)$	Itemset taxonomy $(\mathcal{AC}[\Psi], \leq)$
$A \leq B$	$\forall i \in A, \exists i' \in B, i \leq i'$
$\mathcal{AC}^{(k)}[\Psi]$	Itemsets of size $\leq k$
$I^{(k)}(\Psi)$	$k$ -itemset taxonomy $(\mathcal{AC}^{(k)}[\Psi], \leq)$
$mfi$	Maximal frequent itemsets
$mii$	Minimal infrequent itemsets
$S(\Psi)$	Solution taxonomy $I(I(\Psi))$
$S^{(k)}(\Psi)$	Solution taxonomy $I(I^{(k)}(\Psi))$
<b>MineFreq</b>	Problem of identifying freq exactly

Table 2: Summary of notations

of “crowd-efficient” algorithms, by considering the *computational complexity* of algorithms that achieve the upper complexity bound. This last restriction is relaxed in the sequel.

### 3. CROWD COMPLEXITY

We now analyze the crowd complexity of **MineFreq**, first w.r.t. the input taxonomy. Then, we consider the complexity w.r.t. the output, which allows for a finer analysis depending on properties of  $\text{freq}$ . As a general remark for our analysis, note that we can always avoid querying the same itemset twice, e.g., by caching query answers. Thus, every upper bound  $O(X)$  presented in this section is actually  $O(\min\{X, |I(\Psi)|\})$ .

#### 3.1 With Respect to the Input

To illustrate the problem boundaries, consider the following specific cases of  $\Psi$  for which we know the optimal solution strategy. For a chain taxonomy (as in Figure 1d), identifying  $\text{freq}$  amounts to a binary search for the single MFI. This can be done in  $O(\log |\mathcal{I}|)$  steps. For a flat taxonomy (as in Figure 1f), for which the elements of  $I(\Psi)$  are the power set of  $\mathcal{I}$ , identifying  $\text{freq}$  is equivalent to learning a monotone Boolean function over  $n$  variables, where  $n = |\mathcal{I}|$ . For this problem, a tight bound of  $\Theta(2^n/\sqrt{n})$  is known [22, 23].

We study the solution for a *general taxonomy structure*. Let us start by defining the following:

**DEFINITION 3.1. (SOLUTION TAXONOMY).** *Given a taxonomy  $\Psi$ , we define its solution taxonomy  $S(\Psi) = I(I(\Psi))$ . The domain of elements of  $S(\Psi)$  is  $\mathcal{AC}[I(\Psi)]$ , i.e., antichains of itemsets.*

We call this construction the solution taxonomy, since its elements correspond, precisely, to the frequency predicates over  $I(\Psi)$ , i.e., all possible solutions to **MineFreq** for a given  $\Psi$ . More precisely, each element of  $S(\Psi)$  is an antichain of itemsets that is exactly the set of MFIs  $mfi$  of some

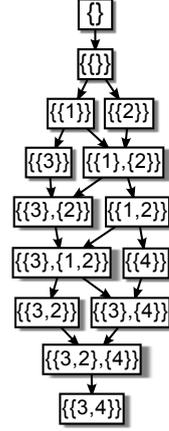


Figure 2: Example solution taxonomy  $S(\Psi_1)$

$\text{freq}$  predicate. We prove this below but first illustrate the structure via an example.

**EXAMPLE 3.2.** *Figure 2 illustrates the solution taxonomy of the running example,  $\Psi_1$ . Consider, e.g.,  $\{\{1\}, \{2\}\}$ . This element of  $S(\Psi_1)$  corresponds to the  $\text{freq}$  predicate assigning true (only) to  $\{1\}$ ,  $\{2\}$  and their ancestor  $\{\}$  in  $I(\Psi_1)$ . Similarly,  $\{\}$  in  $S(\Psi_1)$  corresponds to a predicate assigning false to every itemset, and  $\{\{3, 4\}\}$  to the one assigning true to every itemset.*

Next, we show the correspondence between  $S(\Psi)$  elements and frequency predicates. We proceed by first showing a bijective correspondence between elements of  $S(\Psi)$  and between (decreasing) monotone predicates over  $I(\Psi)$ . Then, we show that each such predicate can indeed serve as a frequency predicate for some database.

**PROPOSITION 3.3.** *There exists a bijective mapping from  $\mathcal{AC}[I(\Psi)]$  to the monotone predicates over  $I(\Psi)$ .*

The proof maps every predicate in a general poset to its (unique) set of maximal elements, which necessarily forms an antichain. Thus, every predicate over  $I(\Psi)$  can be mapped to an antichain of itemsets, which is an element of  $S(\Psi)$  (see formal proof in [2]).

**PROPOSITION 3.4.** *For every threshold  $0 < \Theta < 1$ , every monotone predicate  $F$  over  $I(\Psi)$  is the frequency predicate  $\text{freq}$  of some database  $D$ .*

**PROOF.** Given  $F$ , let  $\mathcal{M}$  be its set of maximal elements. If  $\mathcal{M}$  is empty, i.e., no itemset should be frequent,  $F$  is realized by the empty database. Otherwise, construct  $D$  to consist of the following  $d$  transactions:  $n$  “full” transactions with all the items of  $\mathcal{I}$ , one transaction per  $A \in \mathcal{M}$  containing exactly  $A$ , and  $d - n - |\mathcal{M}|$  empty transactions. We choose  $d$  and  $n$  s.t. every (non-empty) itemset  $B$  is frequent in  $D$  iff it is supported by  $> n$  transactions, or, equivalently, iff it is supported by at least one of the  $|\mathcal{M}|$  non-trivial transactions. To do that, pick an integer  $d$  that is large enough such that there exists an integer  $n$  such that  $0 \leq n/d < \Theta < (n+1)/d < (n+|\mathcal{M}|)/d \leq 1$ . When this holds, the frequent itemsets of  $D$  are exactly the ancestors of itemsets in  $\mathcal{M}$ , so  $\text{freq} = F$  as desired.  $\square$

We have now shown that the **MineFreq** problem (identifying freq) is equivalent to finding the element of  $S(\Psi)$  corresponding to freq, namely *mfi*. Before we study the complexity of this last task, let us first describe abstractly how the solutions space is narrowed down during the execution of *any* algorithm that solves **MineFreq**. In the beginning of the execution, all the elements of  $S(\Psi)$  are *possible*. The algorithm uses some decision method to pick an itemset  $A \in \mathcal{AC}[\Psi]$  and queries it. If the answer is **true** ( $A$  is frequent), this means that *mfi* contains  $A$  or one of its descendants in  $I(\Psi)$ , so we can eliminate all MFI sets of  $S(\Psi)$  that do not have this property, which we can show are exactly the non-descendants of  $\{A\}$  in  $S(\Psi)$ . Conversely, if the answer is **false** ( $A$  is infrequent), we can eliminate all the descendants of  $\{A\}$  in  $S(\Psi)$  (including  $\{A\}$ ). The last remaining element in  $S(\Psi)$  at the end corresponds to the correct freq predicate, because it is the only one consistent with the observations.

**EXAMPLE 3.5.** *Consider again  $S(\Psi_1)$  in Example 3.2. By discovering that, e.g.,  $\text{freq}(\{1\})$ , we know e.g.: that  $\{2\}$  cannot be the only MFI so we can eliminate the solution element  $\{\{2\}\}$ , that its ancestor  $\{\}$  is not an MFI so we can eliminate  $\{\{\}\}$ , and so on. In total, all non-descendants of  $\{\{1\}\}$  in  $S(\Psi_1)$  can be eliminated.*

**Lower Bound.** We now give a lower crowd complexity bound for solving **MineFreq** in terms of the input, which is proved to be tight in the sequel. The proof relies on the fact that solving **MineFreq** amounts to searching for an element in  $S(\Psi)$ ; it can be given as a simple information-theoretic argument (see [2]), but we present it in connection with [25] as we will reuse this link to obtain our upper bound.

**PROPOSITION 3.6.** *The worst-case crowd complexity of identifying freq is  $\Omega(\log(|S(\Psi)|))$ .*

**PROOF.** This is implied by the analogous claim of [25] about order ideals in general posets. By characterizing an order ideal by its maximal elements (whose descendants are not in the ideal) we obtain an antichain, which defines a bijective correspondence between order ideals and antichains. Thus, we can map antichains to order ideals, and use the observation in [25] directly to obtain the same lower bound.  $\square$

In the worst case,  $\log |S(\Psi)|$  can be linear in  $|I(\Psi)|$ , which itself may be exponential in  $|\Psi|$  (e.g., for a flat taxonomy). When this is the case, a trivial algorithm achieves the complexity bound by querying every element in  $I(\Psi)$ . However, for some taxonomy structures (e.g., chain taxonomies), the size of  $S(\Psi)$  is much smaller. We now use  $w[I(\Psi)]$  to deduce a more explicit lower bound for Prop. 3.6.

**PROPOSITION 3.7.**  $w[I(\Psi)] \geq \binom{w[\Psi]}{\lfloor w[\Psi]/2 \rfloor}$ .

**PROOF.** By definition, there exists at least one itemset in  $I(\Psi)$  of size  $w[\Psi]$ . This itemset has  $\binom{w[\Psi]}{\lfloor w[\Psi]/2 \rfloor}$  subsets of size  $\lfloor w[\Psi]/2 \rfloor$ . These itemsets are also in  $I(\Psi)$ , since they only contain incomparable items. Moreover, they are pairwise incomparable in  $I(\Psi)$ , and thus form an antichain whose size yields the lower bound.  $\square$

By replacing  $\Psi$  with  $I(\Psi)$  we get a lower bound for  $w[S(\Psi)]$ . We can thus prove the following bound which, though weaker, is more explicit than Prop. 3.6 as it is expressed in terms of the original ontology width rather than  $|S(\Psi)|$ .

**COROLLARY 3.8.** *The worst-case crowd complexity of identifying freq is  $\Omega(2^{w[\Psi]}/\sqrt{w[\Psi]})$ .*

**PROOF.** We have  $|S(\Psi)| > w[S(\Psi)] \geq \binom{w[I(\Psi)]}{\lfloor w[I(\Psi)]/2 \rfloor}$ . We obtain  $\log |S(\Psi)| \geq \Omega(\log(2^{w[I(\Psi)]}/\sqrt{w[I(\Psi)]})) = \Omega(w[I(\Psi)])$  using Stirling's approximation; and finally, using the lower bound of  $w[I(\Psi)]$  and applying the approximation again, we express the bound in terms of  $w[\Psi]$ .  $\square$

**Upper Bound.** We now state a tight upper bound (i.e., that matches the lower bound up to a multiplicative constant). The proof relies on Theorem 1.1 of [25], which shows that, in any poset, there exists an element such that the proportion of order ideals (or, in our case, elements of  $S(\Psi)$ ) that contain the element is within a constant range of  $1/2$ . Hence, a greedy strategy that queries such elements will eliminate a constant fraction of the possible solutions at each step and completes in a time that is logarithmic in the size of the search space. See [2] for proof details.

**PROPOSITION 3.9.** *The worst-case crowd complexity of identifying freq is  $O(\log |S(\Psi)|)$ .*

## 3.2 With Respect to the Input and Output

So far our results only relied on the structure and size of the input taxonomy  $\Psi$ . However, as noted in Section 2, the characteristics of the output freq predicate may have a crucial effect on the problem complexity, because, in practical scenarios, the number of MFIs and MIIs is usually small. For instance, when dealing with leisure habits, the number of activities that are commonly performed together in the population is typically very small w.r.t. all the combinations that the taxonomy allows. Hence, we next study the effect of the output on the crowd complexity boundaries of **MineFreq**.

**Lower Bound.** Since each of the sets of MFIs and MIIs uniquely represents the freq predicate, one could hope that it would be sufficient to identify only one of them to solve **MineFreq**. However, it turns out that one must query at least all the MIIs to verify that the MFIs are maximal, and vice versa. This result is well-known for Boolean lattices [16]; in our setting it follows from the more general Thm. 2 of [26] (which concerns *any* partial order rather than just distributive lattices).

**PROPOSITION 3.10.** *The worst-case crowd complexity of identifying freq is  $\Omega(|mfi| + |mii|)$ .*

Hence, though we can describe the output by its set of MFIs (or MIIs), we need to query both the MFIs and MIIs. This implies that the crowd complexity may be exponential even in the minimal output size, since the difference between  $|mfi|$  and  $|mii|$  may be large though only one suffices to describe the output. This is derived from a known result in Boolean function learning [8].

**COROLLARY 3.11.** (SEE [8]). *The worst-case crowd complexity of identifying freq is  $\Omega(2^{\min\{|mfi|, |mii|\}})$*

We note that the current lower bound is not tight: for instance, over a chain taxonomy,  $|mfi| + |mii| \leq 2$  for any freq predicate, but we already noted in Section 3.1 that the worst-case crowd complexity in this case is  $\Omega(\log |Z|)$ .

```

Data:  $\Psi$ : a taxonomy
Result:  $M = mfi$  and  $N = mii$ , for the correct freq
predicate over  $I(\Psi)$ 
 $M, N \leftarrow \emptyset$ ;
while there is an unclassified element  $A \in I(\Psi)$  do
  if  $\text{freq}(A)$  then
    /*  $A$  is an ancestor of an MFI, search for
       it by traversing  $A$ 's frequent
       descendants. */
    for  $i \in \mathcal{I}$  do
       $B \leftarrow \text{get-AC}(A \cup \text{anc}(i))$ 
      if  $A < B$  and  $\text{freq}(B)$  then  $A \leftarrow B$ ;
    /*  $A$ 's descendants are infrequent */
     $\text{mark-freq}(A)$ ; add  $A$  to  $M$ ;
  else
    /*  $A$  is a descendant of an MII, search
       for it by traversing  $A$ 's infrequent
       ancestors. */
    for  $i \in \mathcal{I}$  do
       $B \leftarrow \text{get-AC}(A \setminus \text{desc}(i))$ 
      if  $B < A$  and  $\neg \text{freq}(B)$  then  $A \leftarrow B$ ;
    /*  $A$ 's ancestors are frequent */
     $\text{mark-infreq}(A)$ ; add  $A$  to  $N$ ;
return  $M, N$ ;

```

Algorithm 1: Identify  $mfi$  and  $mii$

**Upper bound.** We next show an upper bound that is within a factor  $|\mathcal{I}|$  of the lower bound of Prop. 3.10. It generalizes known MFI and MII identification algorithms for the case where there is no underlying taxonomy, such as the monotone Boolean function learning algorithm of [16] and the Dualize and Advance algorithm of [18, 19]; see Section 7 for an in-depth comparison. Intuitively, our algorithm traverses the elements of  $I(\Psi)$  in an efficient way to identify an MFI or an MII, and repeats this process as long as there are unclassified elements in  $I(\Psi)$ , i.e., elements that are not known to be frequent or infrequent. This method can be used to identify each MFI or MII in time  $O(|\mathcal{I}|)$ , which yields the desired bound.

**THEOREM 3.12.** *Algorithm 1 identifies freq in crowd complexity  $O(|\mathcal{I}| \cdot (|mfi| + |mii|))$ .*

**PROOF.** We explain the course of Algorithm 1, prove that it is correct (i.e., identifies freq correctly), and analyze its crowd complexity.

Algorithm 1 uses a few sub-routines:  $\text{mark-freq}(A)$  (resp.,  $\text{mark-infreq}(A)$ ) classifies the itemset  $A$  and its ancestors (resp., descendants) as frequent (resp., infrequent).  $\text{get-AC}(A)$  removes from  $A$  all the items that are implied by other items (i.e., all  $i \in A$  such that  $i < i'$  for some  $i' \in A$ ) so that  $\text{get-AC}(A)$  returns an antichain representing  $A$ .  $\text{anc}(i)$  and  $\text{desc}(i)$  return, respectively, the ancestors and descendants of  $i$  in  $\Psi$  (including  $i$ ).

We argue that each iteration of the main **while** loop of Algorithm 1 identifies exactly one new MFI or MII. First, an unclassified node  $A \in I(\Psi)$  is chosen. If  $A$  is frequent (first **if** statement), it is either an MFI or an ancestor of an MFI. Since it used to be unclassified, at this point each of its descendants is unclassified or infrequent: in particular,  $A$  is not an ancestor of an already discovered MFI. We thus start traversing descendants of  $A$  by adding items from  $\mathcal{I}$  to

$A$  and using  $\text{get-AC}$  to turn the result into an antichain.<sup>2</sup> Either the current  $A$  is an MFI so all of its children are infrequent, the inner **for** loop ends, and we identify  $A$  as an MFI. Otherwise, as  $A$  is frequent but not maximal, there exists some frequent  $B \in I(\Psi)$  s.t.  $B = \text{get-AC}(A \cup \text{anc}(i'))$  for some item  $i'$ . If  $i'$  had already been considered by the **for** loop but was dismissed, it would mean that we dismissed an ancestor of  $B$  as infrequent, contradicting the assumption that  $B$  is frequent. Thus,  $i'$  cannot have been considered by the **for** loop yet, so we will replace  $A$  by  $B$  before the **for** loop terminates. Hence, at the end of the **for** loop, we identify a new MFI. In the same manner, the code within the **else** part identifies an MII by traversing infrequent ancestors until reaching an infrequent element that has only frequent parents.

**Correctness.** The above implies that the algorithm terminates, that each identified MFI and MII is correct, and that all elements are correctly marked as frequent and infrequent. To prove completeness, consider an MFI  $A$ . By the end of the algorithm,  $A$  is known to be frequent; since it has no frequent descendants,  $\text{mark-freq}(A)$  was necessarily called, which implies that  $A$  was added to  $M$ . The proof for MIIs is similar.

**Complexity.** Since Algorithm 1 identifies an MFI or MII in each **while** iteration, there can be at most  $|mfi| + |mii|$  iterations. The inner loop performs  $O(|\mathcal{I}|)$  queries, and thus the total complexity is as stated above.  $\square$

Following an idea of [18], we observe that the bound can be improved to  $O(|mii| + |\mathcal{I}| \cdot |mfi|)$  if we always choose the unclassified element  $A$  to be minimal, because this ensures that no queries need to be performed whenever we are in the **else** branch. Moreover, if we run two instances of Algorithm 1 in parallel, one choosing maximal unclassified elements for  $A$  and the other one choosing minimal unclassified elements for  $A$ , we improve the bound to

$$O(|mfi| + |mii| + |\mathcal{I}| \cdot \min\{|mfi|, |mii|\})$$

### 3.3 Restricted Itemset Size

We next consider the  $k$ -itemset taxonomy,  $I^{(k)}(\Psi)$ . Beyond the practical motivations for using  $I^{(k)}(\Psi)$  (see Section 2), restricting the number of MFIs and MIIs may naturally improve the complexity bounds.

As explained in Section 2,  $I^{(k)}(\Psi)$  is not necessarily a distributive lattice; and the size of  $I^{(k)}(\Psi)$  is always polynomial while that of  $I(\Psi)$  may be exponential (w.r.t.  $|\mathcal{I}|$ ). However, for every  $I(\Psi)$  such that  $k \geq w[I(\Psi)]$ , it holds that  $I^{(k)}(\Psi) = I(\Psi)$ .

Note that in Section 3.1 we did not make any assumptions on the itemset taxonomy structure, so our results apply to any poset and in particular to  $I^{(k)}(\Psi)$ . We obtain the following, where  $S^{(k)}(\Psi) := I^{(k)}(\Psi)$ .

**COROLLARY 3.13.** *The worst-case crowd complexity of identifying freq over  $I^{(k)}(\Psi)$  is  $\Omega(\log |S^{(k)}(\Psi)|)$ ; and there exists an algorithm to identify freq over  $I^{(k)}(\Psi)$  in crowd complexity  $O(\log |S^{(k)}(\Psi)|) \leq O(|\mathcal{I}|^k)$ .*

For the complexity w.r.t. the output over restricted itemsets, the lower bound of Thm. 3.10 holds as well, with the

<sup>2</sup>We add  $\text{anc}(i)$  to  $A$  to simplify the analysis in the next section; just adding  $i$  would also work here.

same proof. However, we cannot use Algorithm 1 directly to obtain an upper bound, as, in a  $k$ -itemset taxonomy, adding (or removing) a single item to a  $k$ -itemset may not yield a  $k$ -itemset: improving the trivial upper bound remains an open problem.

## 4. COMPUTATIONAL COMPLEXITY

We next study the feasibility of “crowd-efficient” algorithms, by considering the computational complexity of algorithms that achieve the upper crowd complexity bound. We follow the same axes as in the previous section. In all problem variants, we have the crowd complexity lower bound as a simple (and possibly not tight) lower bound. For some variants, we show that, even when the crowd complexity is feasible, the underlying computational complexity may still be infeasible.

### 4.1 With Respect to the Input

As a simple lower bound, we know that the computational complexity of `MineFreq` is higher than the crowd complexity, and is thus  $\Omega(\log(|S(\Psi)|))$ .

The problem of finding tighter bounds for computational complexity w.r.t. the input remains open. Many works [11, 13] provide efficient algorithms for computing a good split element in particular types of posets, but no efficient algorithm is known for the more general case of distributive lattices (or for arbitrary posets). We now give evidence suggesting that no such algorithm exists.

At any point of a `MineFreq`-solving algorithm, we define the *best-split* element as the element of  $I(\Psi)$  which is guaranteed to eliminate the largest number of solutions of  $S(\Psi)$  when queried. Following the proof of Prop. 3.9, if we could efficiently compute the best-split element, we would obtain a computationally efficient greedy algorithm that is also crowd-efficient. We now show that this is impossible for bounded-size itemsets (and the corresponding restricted itemset taxonomies). This, of course, does not prove that there exists no computationally efficient non-greedy algorithm, but it suggests that such an algorithm probably does not exist, because of the close relationship between finding best-split elements and counting the antichains of  $I(\Psi)$ . This relates to a result of [13], which proves that identifying a good-split element (which guarantees eliminating a constant fraction of the solutions) is computationally equivalent to a relative approximation of the number of order ideals (though this is not known to be  $\#P$ -complete).

**THEOREM 4.1.** *The problem of identifying, given  $\Psi$  and  $k$ , the best-split element in  $I^{(k)}(\Psi)$  is  $\text{FP}^{\#P}$ -complete<sup>3</sup> in  $|\Psi|$ .*

**PROOF.** (Sketch). To prove membership, we show a reduction from our problem to counting antichains in a general poset, which is known to be in  $\#P$  [35]. Using an oracle for antichain counting, we can count the number of eliminated antichains in  $S^{(k)}(\Psi)$  for every element of  $I(\Psi)$ , and thus find the best-split element.

To prove hardness, we show a reduction from the antichain counting problem (which is  $\text{FP}^{\#P}$ -hard) to our problem. Let us call *ancestor* (resp. *descendant*) solutions of  $A$  the

<sup>3</sup> $\#P$  is the class of *counting* problems that return the number of solutions of NP problems.  $\text{FP}^{\#P}$  is the class of *function* problems that can be computed in polynomial time using a  $\#P$  oracle.

solutions (elements of  $S^{(k)}(\Psi)$ ) that are eliminated if an itemset  $A$  is discovered to be frequent (resp. infrequent). For any poset  $P$  and natural number  $n$ , we show that we can construct a  $k$ -itemset taxonomy  $I^{(k)}(\Psi)$  with an itemset  $A_0$  such that, for some increasing affine function  $F$ ,  $A_0$  has  $F(|\mathcal{AC}[P]|)$  descendant solutions and  $F(n)$  ancestor solutions. As the best-split element  $A^*$  in  $I^{(k)}(\Psi)$  has a roughly equal number of ancestor and descendant solutions, comparing the position of  $A_0$  and  $A^*$  allows us to compare  $|\mathcal{AC}[P]|$  and  $n$ : if  $A^*$  is an ancestor of  $A_0$ , it has more descendant solutions than  $A_0$ , and hence  $|\mathcal{AC}[P]| < n$ . Similarly, if  $A^*$  is a descendant of  $A_0$ ,  $|\mathcal{AC}[P]| > n$ . We can then perform a binary search on values of  $n$  between 0 and  $2^{|P|}$  using this method and find the exact value of  $|\mathcal{AC}[P]|$ .  $\square$

As for upper bounds, our complexity results w.r.t. the input, namely Cor. 4.4, will follow from the results w.r.t. the input and output presented in the next section.

### 4.2 With Respect to the Input and Output

**Lower Bound.** As shown by Algorithm 1, finding an MFI or MII requires a number of queries linear in  $|\mathcal{I}|$ . However, note that the algorithm assumes that at any point we are able to determine if the set of unclassified elements of the itemset taxonomy is empty. We next show that this is a non-trivial problem. We recall the definition of problem EQ [6]. Let  $B_n = \{0, 1\}^n$  be the set of Boolean vectors of length  $n$ . Define the order  $\leq$  on  $B_n$  by  $x \leq y$  iff  $x_i \leq y_i$  for all  $i$ . For  $X \subseteq B_n$ , write  $T(X) = \{y \in B_n \mid \exists z \in X, z \leq y\}$  and  $F(X) = \{y \in B_n \mid \exists z \in X, y \leq z\}$ . Problem EQ is the following: given  $X, Y \subseteq B_n$  such that  $T(X) \cap F(Y) = \emptyset$ , decide whether  $T(X) \cup F(Y) = B_n$ .

**PROPOSITION 4.2.** *If `MineFreq` can be solved in computational time  $O(\text{poly}(|mii|, |mfi|, w[\Psi]))$  then there exists a PTIME solution for problem EQ from [6].*

It is unknown whether EQ is solvable in polynomial time (see [12, 17] for a survey); the connection between frequent itemset mining and EQ (and its other equivalent formulations, such as monotone dualization or hypergraph transversals) was already noted in [26]. Note that the proof above uses the fact that the itemset size is not restricted. For  $k$ -itemset taxonomies, finding a tighter lower bound than the trivial  $|mfi| + |mii|$  remains an open problem.

**Upper Bound.** We consider again Algorithm 1, whose crowd complexity we analyzed in Section 3.2. By completing some implementation details, we can now analyze its computational complexity as well, and obtain an upper bound. For simplicity, this bound is presented in the Introduction with  $|\Psi|$  which is  $\geq |\mathcal{I}|$ .

**PROPOSITION 4.3.** *There exists an algorithm to solve `MineFreq` in computational time*

$$O(|I(\Psi)| \cdot (|\mathcal{I}|^2 + |mfi| + |mii|))$$

**PROOF.** Algorithm 1 uses a computation of an unclassified element of  $I(\Psi)$ . Since by Prop. 4.2 this is probably non-polynomial, we can use the brute-force method of materializing the itemset taxonomy  $I(\Psi)$ . We use a hash table to find any element in the  $I(\Psi)$  structure in time linear in the element size. The implementation of mark-freq and

mark-infreq locates  $A$  in  $I(\Psi)$  using the hash table, traverses its ancestors or descendants respectively, and marks them as (in)frequent.

To compare itemsets efficiently, we represent each itemset  $A$  by an ordered list of the items in its order ideal, i.e.,  $\downarrow A = \{i \in \mathcal{I} \mid \exists i' \in A, i' \leq i\}$ . In this case,  $A \leq B$  iff  $\downarrow A \subseteq \downarrow B$ , which can be verified in time  $O(|\downarrow A| + |\downarrow B|) \leq O(|\mathcal{I}|)$ . Using this representation, we do not need the sub-routine get-AC. We generate once, for every  $i \in \mathcal{I}$ , two ordered lists:  $\text{desc}(i)$  and  $\text{anc}(i)$ , holding its descendants and ancestors respectively. These lists can be computed in time  $O(|\mathcal{I}|^2)$  by building the transitive closure of  $I(\Psi)$ , and can be used to compute  $\downarrow A \cup \text{anc}(i)$  and  $\downarrow A \setminus \text{desc}(i)$  in time  $O(|\mathcal{I}|)$ .

Let us analyze the overall complexity of the suggested implementation. We construct  $I(\Psi)$  (where each element has both its antichain and order ideal representations) in  $O(|\mathcal{AC}[\Psi]| \cdot |\mathcal{I}|^2)$  (see [2]), and construct  $\text{anc}(i)$  and  $\text{desc}(i)$  in time  $O(|\mathcal{I}|^2)$ . Now, we run  $|mfi| + |mii|$  times the body of the outer **while** loop, which 1. finds an unclassified element by a brute-force search taking time  $|\mathcal{AC}[\Psi]|$ , 2. runs  $O(|\mathcal{I}|)$  times the body of one of the **for** loops that computes  $\downarrow A \cup \text{anc}(i)$  or  $\downarrow A \setminus \text{desc}(i)$  and verifies  $\leq$  in time  $O(|\mathcal{I}|)$ , and 3. calls mark-freq or mark-infreq which takes time  $O(|\mathcal{I}| + |I(\Psi)|)$  to locate the itemset in  $I(\Psi)$  and traverse its ancestors or descendants. Summing these numbers and simplifying the expression yields the claimed complexity bound.  $\square$

Since we know that  $|mfi| + |mii| \leq |\mathcal{AC}[\Psi]|$ , we can plug  $|\mathcal{AC}[\Psi]|$  in the complexity formula and obtain an upper bound that does not depend on the numbers of MFIs and MIIs. In this manner we achieve a bound polynomial in  $|I(\Psi)|$  and improve the upper bound described in Section 4.1. However, note that this is in fact a relaxation of our requirement for crowd-efficient algorithms, since Algorithm 1 is not crowd-efficient w.r.t. the upper bound of Prop. 3.9, in terms of the input. This result is also simplified in the Introduction, replacing  $|\mathcal{I}|$  by  $|\Psi|$  which is  $\geq |\mathcal{I}|$ , and  $|\mathcal{AC}[\Psi]|$  by  $|I(\Psi)|$  which is  $\geq |\mathcal{AC}[\Psi]|$ .

**COROLLARY 4.4.** *There exists an algorithm to solve MineFreq in computational complexity*

$$O(|I(\Psi)| \cdot (|\mathcal{I}|^2 + |\mathcal{AC}[\Psi]|))$$

## 5. CHAIN PARTITIONING

Recall that in the beginning of Section 3.1 we mentioned the special case of chain taxonomies, for which a binary search achieves a tight complexity bound, both crowd and computational, of  $\Theta(\log |\mathcal{I}|)$ . We generalize this insight to solve MineFreq for taxonomies partitioned in disjoint chain taxonomies. *Chain partitioning* is a standard technique in Boolean function learning [22, 24], that splits the Boolean lattice elements into disjoint chains, and then performs a binary search for the maximal frequent element on each chain. The following easy proposition holds (we justify how the partition  $P$  is obtained at the end of the section):

**PROPOSITION 5.1.** *Given a partition  $P$  of  $I(\Psi)$  into  $w[I(\Psi)]$  chains, freq can be identified in both crowd and computational complexity  $O(w[I(\Psi)] \cdot \log |\mathcal{I}|)$ .*

The  $\log |\mathcal{I}|$  factor comes from the binary search in the chains. To understand intuitively why the length of the chains is at most  $|\mathcal{I}|$ , notice that the worst case is achieved by

the full Boolean lattice, and that, in this case, for every chain of the form  $A_0 \leq \dots \leq A_n$ , it holds that  $|A_i| + 1 \leq |A_{i+1}|$ , so that at most  $|\mathcal{I}|$  items can be added to  $A_0$  in total (see Figure 1g).

Let us compare the result of Prop. 5.1 with previous results. In terms of crowd complexity, if  $|S(\Psi)|$  is close to its lower bound,  $2^{w[I(\Psi)]}$ , then the partition binary search performs more queries by a multiplicative factor of  $\log |\mathcal{I}|$  than the upper bound of Prop. 3.9. On the other hand, since we know that the bound of Prop. 3.9 is tight, we get an upper bound for  $|S(\Psi)|$  that depends on  $w[I(\Psi)]$  (in addition to the trivial upper bound  $2^{w[I(\Psi)]}$ ).

**COROLLARY 5.2.**  $|S(\Psi)| \leq 2^{w[I(\Psi)] \log |\mathcal{I}|}$ .

When  $|mfi| + |mii| = \Omega(w[I(\Psi)])$ , the crowd complexity of the partition binary search is asymptotically smaller than that of Algorithm 1,  $O(|\mathcal{I}| \cdot (|mfi| + |mii|))$ . The intuitive explanation to this is the following: Algorithm 1, in the worst case, can traverse a full chain for every MFI and MII, taking linear time whereas the partition binary search takes logarithmic time. However, when  $|mfi| + |mii|$  is small w.r.t.  $w[I(\Psi)]$ , Algorithm 1 will consider significantly less chains and is thus more efficient.

It remains to explain how to obtain the partition  $P$ . By Dilworth's theorem, it is possible to partition the poset  $I(\Psi)$  into exactly  $w[I(\Psi)]$  chains [10]. Computing the partition can be done in  $O(\text{poly}(|I(\Psi)|))$ , by a reduction to maximum matching (or maximal join) in a bipartite graph [15]. See [2] for a discussion on the complexity of taxonomy chain partitioning.

## 6. GREEDY ALGORITHMS

In the previous sections, we have attempted to fully identify freq. The solutions that we presented try to do so by maximizing the number of eliminated solutions, or identifying MFIs or MIIs. However, we may not be able to pose enough questions to identify freq exactly. In a dynamic crowd setting we could assume, e.g., that the cost of obtaining answers from the crowd (both in terms of money and latency) is not controlled, and that the identification of freq may be interrupted at any time. In such cases, our algorithms would perform badly:

**EXAMPLE 6.1.** *Assume that the unclassified part of the itemset taxonomy  $I(\Psi)$  contains a chain  $C$  of even length  $2n$ , for some  $n > 1$ , and one incomparable itemset  $A$ . There are exactly  $2 + 4n$  antichains in this poset (one empty,  $1 + 2n$  of size 1 and  $2n$  of size 2), which is also the number of possible solutions. Asking about  $A$  eliminates exactly half of the possible solutions for freq and finds an MII or MFI. However, if we have to interrupt the computation after only one query, we have only obtained information about  $A$ . It would have been better to query a middle element of  $C$ : though this eliminates less solutions and does not identify an MII of MFI, it classifies  $\geq n$  itemsets.*

Motivated by this example, in this section, we assume that the computation can be halted at any time, and look at the intuitive strategy that tries to maximize the number of classified itemsets at halting time using the following greedy approach: compute, for every itemset, what is the worst-case (minimal) number of itemsets that could be classified if we

query it; then query the *greedy best-split itemset*, namely the itemset which *maximizes* this number. To perform the greedy best-split computation, we need to count the number of ancestors and descendants of each element; this may be done in time linear in  $|\mathbf{I}(\Psi)|$  per itemset. In terms of  $|\Psi|$ , we can show that this computation is hard.

**PROPOSITION 6.2.** *Finding the greedy best-split itemset is  $\text{FP}^{\#P}$ -hard w.r.t.  $|\Psi|$ . There exists an algorithm which finds it in time  $O(|\mathcal{AC}[\Psi]| \cdot (|\mathcal{I}|^2 + |\mathbf{I}(\Psi)|))$ .*

To prove this, we first observe that the structure of  $S^{(1)}(\Psi)$  is almost identical to that of  $\mathbf{I}(\Psi)$ . Then, for the structure used in the proof of Thm. 4.1, we show a reduction from finding the best-split element in  $S^{(1)}(\Psi)$  to finding the greedy best-split element in  $\mathbf{I}^{(1)}(\Psi)$ . The second part follows from the brute-force method described above, in combination with the complexity of materializing  $\mathbf{I}(\Psi)$ . See the full version [2] for details.

## 7. RELATED WORK

Throughout the paper, we have combined and extended results from *order theory*, *Boolean function learning* and *data mining* [6, 8, 16, 18, 19, 25], to obtain our characterization of the complexity of the crowd mining problem. We now discuss further related work.

Several recent works consider the use of crowdsourcing platforms as a powerful means of data procurement (e.g., [14, 28, 34]). As the crowd is an expensive resource, many works focus on minimizing the number of questions posed to the crowd to perform a certain task: for instance, computing common query operators such as filter, join and max [9, 21, 29, 33, 38], performing entity resolution [39], etc. The present work considers the mining of data patterns from the crowd, and thus is closely related to this line of work.

The most relevant work, by some of the present authors, is [3], which proposes a general crowd mining framework. That work focused on a technique to estimate the confidence in a mined data pattern and how much it increases if more answers are gathered: we could use this technique to implement the crowd query black-box mechanism in our context. However, [3] did not address the issue of the dependencies between rules, or study the implied complexity boundaries, which is the objective of the present paper. Another particularly relevant work is [32], which considers a crowd-assisted search problem in a graph. While it is possible to reformulate some of our problems as graph searches in the itemset and solution taxonomies, there are two important differences between our setting and theirs. First, our itemset and solution taxonomies may be exponential in the size of the original taxonomy but have a specific structure, which allows, in some cases, to perform the search without materializing them. Second, we allow algorithms for `MineFreq` to choose crowd queries interactively based on the answers to previous queries, whereas [32] studies “offline” algorithms where all questions are selected in advance. Consequently, our algorithms and complexity results are inherently different.

Frequent itemset discovery is a fundamental building block in data mining algorithms (see, e.g., [1]). The idea of using taxonomies in data mining was suggested in [37], which we use as a basis for our definitions.

Another line of works in data mining considers the discovery of interesting data patterns by performing oracle calls [26].

This work is closely connected to ours by (i) the use of oracles, which may be seen as an abstraction of the crowd (compared to our setting), and (ii) the separation between the complexity analysis of the number of oracle calls (which corresponds to crowd complexity in our case) and of the computational process. However, because our motivation is to query the crowd, we focus on the specific problem of mining under a taxonomy over the itemsets (and related variants such as limiting the itemset size) which is not studied in itself in this line of work. On the one hand, [26] studies a generalization of our setting, namely the problem of finding all interesting sentences given a specialization relation on sentences. They introduce the notion of border (corresponding to MFIs and MIIs) as a way to bound the number of oracle calls. However, in this general setting, they are not able to prove complexity bounds on the performance of applicable algorithms (e.g., Algorithm ALLMSS from [20]) to match the bounds that we obtain for the more specific setting of mining frequent itemsets under a taxonomy. On the other hand, the aforementioned papers also study the restricted case of Boolean lattices and provide some complexity bounds in this case (e.g., for the Dualize and Advance algorithm [18, 19]); however, those algorithms exploit the connection with hypergraph transversals which is very specific to the Boolean lattice. Hence, these algorithms cannot be used to mine frequent itemsets under a taxonomy, which is very natural when working with the crowd, and their complexity bounds are not applicable to our problem.

Finally, among the many works that discuss the connection of data mining and hypergraph traversals, we note the recent work [17] which is relevant to our EQ-hardness result (Prop. 4.2) as it sheds more light on the (still open) complexity of EQ.

## 8. CONCLUSION

In this paper, we have considered the identification of frequent itemsets in human knowledge domains by posing questions to the crowd, under a taxonomy which captures the semantic dependencies between items. We have studied the complexity boundaries of solutions to this problem, in terms of two cost metrics: the number of crowd queries required for identifying the frequent itemsets, and the computational complexity of choosing these queries. We identified two main factors that affect both complexities: the structure of the taxonomy; and properties of the frequency predicate.

Our results leave some intriguing theoretical questions open: in particular, we would like to find tighter complexity bounds where possible, and to further study the nature of the tradeoff between crowd and computational complexities. In addition, due to the high complexity of taxonomy-based crowd mining, practical implementations could further resort to approximations and randomized algorithms in order to identify (in expectation) a large portion of the frequent itemsets, while reducing the complexity. The greedy approach mentioned in Section 6 forms a first step in this direction of further research. A different approach involves filtering the itemsets according to a user request, which could reduce the solution search space: for instance, the user may wish to mine itemsets composed of small fragments of the taxonomy, or respecting certain constraints. We intend to investigate this approach in future work.

*Acknowledgments.* We are grateful to the anonymous referees and Toon Calders for their useful comments that have helped us improve our paper, and in particular for pointing us to related work. We also thank Pierre Senellart for his insightful ideas and comments.

This work has been partially funded by the European Research Council under the FP7, ERC grant MoDaS, agreement 291071, and by the Israel Ministry of Science.

## 9. REFERENCES

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, 1994.
- [2] A. Amarilli, Y. Amerdamer, and T. Milo. On the complexity of mining itemsets from the crowd using taxonomies. *CoRR*, 2013. Full version. <http://arxiv.org/abs/1312.3248>.
- [3] Y. Amsterdamer, Y. Grossman, T. Milo, and P. Senellart. Crowd mining. In *SIGMOD*, 2013.
- [4] Y. Amsterdamer, Y. Grossman, T. Milo, and P. Senellart. CrowdMiner: Mining association rules from the crowd. In *VLDB*, 2013.
- [5] R. J. Bayardo, Jr. Efficiently mining long patterns from databases. In *SIGMOD*, 1998.
- [6] J. Bioch and T. Ibaraki. Complexity of identification and dualization of positive Boolean functions. *Inf. Comput.*, 123(1), 1995.
- [7] N. Bradburn, L. Rips, and S. Shevell. Answering autobiographical questions: the impact of memory and inference on surveys. *Science*, 236(4798), 1987.
- [8] N. H. Bshouty. Exact learning Boolean functions via the monotone theory. *Inf. Comput.*, 123(1), 1995.
- [9] S. B. Davidson, S. Khanna, T. Milo, and S. Roy. Using the crowd for top-k and group-by queries. In *ICDT*, 2013.
- [10] R. P. Dilworth. A decomposition theorem for partially ordered sets. *Ann. Math.*, 51(1), 1950.
- [11] D. P. Dubhashi, K. Mehlhorn, D. Ranjan, and C. Thiel. Searching, sorting and randomised algorithms for central elements and ideal counting in posets. In *FSTTCS*, 1993.
- [12] T. Eiter, K. Makino, and G. Gottlob. Computational aspects of monotone dualization: A brief survey. *Discrete Applied Mathematics*, 156(11), 2008.
- [13] U. Faigle, L. Lovász, R. Schrader, and G. Turán. Searching in trees, series-parallel and interval orders. *SIAM J. Comput.*, 15(4), 1986.
- [14] M. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. CrowdDB: answering queries with crowdsourcing. In *SIGMOD*, 2011.
- [15] D. R. Fulkerson. Note on Dilworth's decomposition theorem for partially ordered sets. In *Proc. Amer. Math. Soc.*, volume 7, 1956.
- [16] D. N. Gainanov. On one criterion of the optimality of an algorithm for evaluating monotonic Boolean functions. *USSR Comp. Math. and Math. Phys.*, 24(4), 1984.
- [17] G. Gottlob. Deciding monotone duality and identifying frequent itemsets in quadratic logspace. In *PODS*, 2013.
- [18] D. Gunopulos, R. Khardon, H. Mannila, S. Saluja, H. Toivonen, and R. S. Sharma. Discovering all most specific sentences. *TODS*, 28(2), 2003.
- [19] D. Gunopulos, H. Mannila, R. Khardon, and H. Toivonen. Data mining, hypergraph transversals, and machine learning. In *PODS*, 1997.
- [20] D. Gunopulos, H. Mannila, and S. Saluja. Discovering all most specific sentences by randomized algorithms. In *ICDT*, 1997.
- [21] S. Guo, A. G. Parameswaran, and H. Garcia-Molina. So who won? Dynamic max discovery with the crowd. In *SIGMOD*, 2012.
- [22] G. Hansel. Sur le nombre des fonctions booléennes monotones de  $n$  variables. *Comptes Rendus de l'Académie des Sciences*, 262(A), 1966.
- [23] V. K. Korobkov. On monotone functions of the algebra of logic. *Problemy Kibernetiki*, 13, 1965.
- [24] B. Kovalerchuk, E. Triantaphyllou, A. S. Deshpande, and E. Vityaev. Interactive learning of monotone Boolean functions. *Inf. Sci.*, 94(1), 1996.
- [25] N. Linial and M. Saks. Every poset has a central element. *J. Combinatorial Theory*, 40(2), 1985.
- [26] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data mining and knowledge discovery*, 1(3), 1997.
- [27] A. Marcus, D. Karger, S. Madden, R. Miller, and S. Oh. Counting with the crowd. In *VLDB*, 2012.
- [28] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller. Crowdsourced databases: query processing with people. In *CIDR*, 2011.
- [29] A. Marcus, E. Wu, D. R. Karger, S. Madden, and R. C. Miller. Human-powered sorts and joins. *PVLDB*, 5(1), 2011.
- [30] Amazon Mechanical Turk. <https://www.mturk.com/>.
- [31] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 1995. <http://wordnet.princeton.edu/>.
- [32] A. Parameswaran, A. Sarma, H. Garcia-Molina, N. Polyzotis, and J. Widom. Human-assisted graph search: it's okay to ask questions. *PVLDB*, 4(5), 2011.
- [33] A. G. Parameswaran, H. Garcia-Molina, H. Park, N. Polyzotis, A. Ramesh, and J. Widom. CrowdScreen: algorithms for filtering data with humans. In *SIGMOD*, 2012.
- [34] A. G. Parameswaran, H. Park, H. Garcia-Molina, N. Polyzotis, and J. Widom. Deco: declarative crowdsourcing. In *CIKM*, 2012.
- [35] J. Provan and M. Ball. The complexity of counting cuts and of computing the probability that a graph is connected. *SIAM J. Comp.*, 12(4), 1983.
- [36] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1), 2012.
- [37] R. Srikant and R. Agrawal. Mining generalized association rules. In *VLDB*, 1995.
- [38] P. Venetis, H. Garcia-Molina, K. Huang, and N. Polyzotis. Max algorithms in crowdsourcing environments. In *WWW*, 2012.
- [39] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. CrowdER: crowdsourcing entity resolution. *PVLDB*, 5(11), 2012.