

On maximal frequent and minimal infrequent sets in binary matrices^{*}

E. Boros and V. Gurvich

RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway New Jersey 08854-8003, USA (`{boros,gurvich}@rutcor.rutgers.edu`)

L. Khachiyan

Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey, 08854-8019, USA (`leonid@cs.rutgers.edu`)

K. Makino

Division of Systems Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka, 560-8531, Japan (`makino@sys.es.osaka-u.ac.jp`)

Abstract. Given an $m \times n$ binary matrix A , a subset C of the columns is called t -frequent if there are at least t rows in A in which all entries belonging to C are nonzero. Let us denote by α the number of maximal t -frequent sets of A , and let β denote the number of those minimal column subsets of A which are not t -frequent (so called t -infrequent sets). We prove that the inequality $\alpha \leq (m - t + 1)\beta$ holds for any binary matrix A in which not all column subsets are t -frequent. This inequality is sharp, and allows for an incremental quasi-polynomial algorithm for generating all minimal t -infrequent sets. We also prove that the analogous generation problem for maximal t -frequent sets is NP-hard. Finally, we discuss the complexity of generating closed frequent sets and some other related problems.

Keywords: data mining, frequent sets, infrequent sets, independent sets, hitting sets, transversals, dualization.

1. Introduction

Let us consider an $m \times n$ binary matrix $A : \mathcal{R} \times \mathcal{C} \rightarrow \{0, 1\}$, and an integral threshold value $t \in \{1, \dots, m\}$. To each subset $C \subseteq \mathcal{C}$ of the columns, let us associate the subset $R(C) \subseteq \mathcal{R}$ of all those rows

^{*} This research was supported in part by the National Science Foundation (Grant IIS-0118635), the Office of Naval Research (Grant N00014-92-J-1375), and the Scientific Grant in Aid of the Ministry of Education, Science, Sports, Culture and Technology of Japan. Visits of the second author to Rutgers University were also supported by DIMACS, the National Science Foundation's Center for Discrete Mathematics and Theoretical Computer Science.

An extended abstract of this work was published at the *19th International Symposium on Theoretical Aspects of Computer Science (STACS 2002)*. (H. Alt and A. Ferreira, eds., Antibes Juan-les-Pins, France, March 14-16, 2002), Lecture Notes in Computer Science 2285 (2002) pp. 133-141, (Springer Verlag, Berlin, Heidelberg, New York).

$r \in \mathcal{R}$, for which $A(r, c) = 1$ in every column $c \in C$. The cardinality $|R(C)|$ is called the *support* of the set C . Let us call a subset $C \subseteq \mathcal{C}$ of the columns *frequent* (or more precisely, *t-frequent*) if its support is at least the given integral threshold t , i.e. if $|R(C)| \geq t$, and let us denote by \mathcal{F}_t the family of all t -frequent subsets of the columns of the given binary matrix A . Let us further call a subset $C \subseteq \mathcal{C}$ *infrequent* (or *t-infrequent*) if its support does not exceed the given threshold t , i.e. if $|R(C)| < t$. Clearly, subsets of frequent sets are also frequent, and supersets of infrequent sets are also infrequent. Let us denote by $\mathcal{M}_t \subseteq \mathcal{F}_t$ the family of all maximal t -frequent sets (i.e. those which are t -frequent, but no superset of them is t -frequent), and by \mathcal{I}_t the family of all minimal t -infrequent sets (i.e. those which are infrequent but all proper subsets of them are t -frequent.)

The generation of frequent sets of a given binary matrix A is an important task of knowledge discovery and data mining, e.g. it is used for mining association rules [1, 2, 6, 18, 26, 27], correlations [9], sequential patterns [3], episodes [28], emerging patterns [12], and appears in many other applications. Most practical procedures to generate \mathcal{F}_t are based on the anti-monotone *Apriori* heuristic (see [2]) and build frequent sets in a bottom-up way, in time proportional to the number of frequent sets. It was also demonstrated recently in [10] that these methods are inadequate in practice when there are (many) frequent sets of large size (see also [4, 20, 24]), due to the fact that $|\mathcal{F}_t|$ can be exponentially larger than $|\mathcal{M}_t|$.

These results show that it is perhaps more important to find the *boundary* of the frequent sets, i.e. the families of maximal frequent and minimal infrequent sets $\mathcal{M}_t \cup \mathcal{I}_t$ (proposed e.g. in [31]), and use those as a condensed representation of the data set, as suggested in [26]. There are several other examples presented in [26] to show the usefulness of maximal frequent sets and minimal infrequent sets, e.g. providing error bounds for the confidence of an arbitrary Boolean rule, in terms of minimal infrequent sets. Furthermore, it can easily be seen (see e.g. [18]) that no algorithm using membership queries “ $X \in \mathcal{F}_t?$ ” (i.e. using the matrix A only to compute the support for column subsets $X \subseteq \mathcal{C}$) can generate all maximal frequent sets by asking fewer than $|\mathcal{M}_t \cup \mathcal{I}_t|$ such queries.

In this short paper we prove the following inequality.

THEOREM 1. *If $\mathcal{I}_t \neq \emptyset$ then*

$$|\mathcal{M}_t| \leq (m - t + 1)|\mathcal{I}_t|. \quad (1)$$

Note that the requirement that \mathcal{I}_t be non-empty is necessary because for $|\mathcal{I}_t| = 0$ we would have $\mathcal{M}_t = \{\mathcal{C}\}$ and hence $|\mathcal{M}_t| = 1$ in contradic-

tion with (1). The condition $\mathcal{I}_t \neq \emptyset$ thus excludes the degenerate case when the entire columns set of A is t -frequent.

Before proceeding further, let us mention some algorithmic implications of (1).

Given a hypergraph \mathcal{H} on a finite base set V , a subset $S \subseteq V$ is called a *transversal* of it if $S \cap H \neq \emptyset$ for all $H \in \mathcal{H}$. Let \mathcal{H}^d denote the family of all minimal transversals of \mathcal{H} . It was observed independently by several authors (see [5, 18, 19]) that the completeness of given subfamilies $\mathcal{X} \subseteq \mathcal{H}$ and $\mathcal{Y} \subseteq \mathcal{H}^d$ (i.e., the equalities $\mathcal{X} = \mathcal{H}$ and $\mathcal{Y} = \mathcal{H}^d$) is equivalent with $\mathcal{X}^d = \mathcal{Y}$. Here the problem of checking if $\mathcal{X}^d = \mathcal{Y}$ is called *hypergraph transversal* problem (for definitions and related results see e.g. [13]). Though the exact complexity of the hypergraph transversal problem is still open, it is known to be solvable in quasi-polynomial time [15], implying thus an incremental quasi-polynomial generation of $\mathcal{H} \cup \mathcal{H}^d$ (assuming \mathcal{H} is represented by an efficiently computable membership oracle). (For more on joint generation of hypergraphs and their transversals and the other related results, see [8].)

To apply the above ideas to frequent set generation, let us observe first that minimal t -infrequent sets are exactly the minimal transversals of the complements of maximal t -frequent sets. Introducing $\mathcal{H}^c = \{V \setminus H \mid H \in \mathcal{H}\}$ to denote the family of complementary edges of a hypergraph \mathcal{H} , we can thus see that $\mathcal{I}_t = (\mathcal{M}_t^c)^d$. Since the given matrix A provides us with an efficient membership oracle for both families \mathcal{I}_t and \mathcal{M}_t (and equivalently for \mathcal{M}_t^c), the above results imply an incremental quasi-polynomial joint generation of the families \mathcal{I}_t and \mathcal{M}_t . Specifically, it follows from [15] that for each $k \leq |\mathcal{M}_t \cup \mathcal{I}_t|$, we can generate k sets in $\mathcal{M}_t \cup \mathcal{I}_t$ in $\text{poly}(n, m) + k^{o(\log k)}$ time. The above inequality (1) clearly implies that if we can generate $\mathcal{M}_t \cup \mathcal{I}_t$ in time $T(|\mathcal{M}_t \cup \mathcal{I}_t|)$, then the entire set \mathcal{I}_t can be generated in time $T((m - t + 2)|\mathcal{I}_t|)$. We thus conclude that the family of minimal infrequent sets \mathcal{I}_t can be generated in output quasi-polynomial time, i.e. in time bounded by a quasi-polynomial in $|\mathcal{I}_t|$. This can be further improved to show that the incremental complexity of generating \mathcal{I}_t is also equivalent with that of the transversal hypergraph problem (see [7, 8] for more detail). Hence

COROLLARY 1. *For each $k \leq |\mathcal{I}_t|$, we can compute k minimal t -infrequent sets of A in $\text{poly}(n, m) + K^{o(\log K)}$ time, where $K = \max\{k, m\}$.*

Let us note next that the matrix A can also be interpreted as the adjacency matrix of a bipartite graph $G = (\mathcal{R} \cup \mathcal{C}, E)$, i.e., in which $(r, c) \in E$ if and only if $A(r, c) = 1$. Then, maximal frequent sets of

A correspond to maximal complete bipartite subgraphs $K_{R,C} \triangleleft G$, where $R \subseteq \mathcal{R}$, and $C \subseteq \mathcal{C}$. Such complete bipartite subgraphs are also considered in the context of concept lattices, and are called *formal concepts*, see e.g. [11, 16].

It is known (see e.g., [22]) that computing the number of maximal complete bipartite subgraphs of a bipartite graph is a #P-complete problem, and hence by the above equivalence, computing $|\cup_{t \geq 1} \mathcal{M}_t|$ is also #P-complete. (In [14] an $O(l^3 2^{2l}(m+n))$ algorithm was presented to generate all maximal complete bipartite subgraphs of a bipartite graph on $m+n$ vertices, or equivalently, to generate all maximal frequent sets of A , where l denotes the maximum of $|C||R(C)|/(|C| + |R(C)| - 1)$, with the maximum taken over all maximal frequent sets C .) Strengthening these (negative) results and the NP-hardness result of [25, Lemma 4.2], we can show the following:

THEOREM 2. *Given an $m \times n$ matrix A , a threshold t , and a subfamily $\mathcal{S} \subseteq \mathcal{M}_t$, it is NP-complete to decide if $\mathcal{S} \neq \mathcal{M}_t$, even if $|\mathcal{S}| = O(n^\varepsilon)$ and $|\mathcal{M}_t|$ is exponentially large in n whenever $\mathcal{S} \neq \mathcal{M}_t$, where $\varepsilon > 0$ can be arbitrarily small.*

Note that deciding whether or not $\mathcal{S} \neq \mathcal{M}_t$ for *polylogarithmically* large $|\mathcal{S}|$ can be done in quasi-polynomial time. This is because $\mathcal{S} = \mathcal{M}_t$ exactly when $(\mathcal{S}^c)^d = \mathcal{I}_t$, which is equivalent with $(\mathcal{S}^c)^d \subseteq \mathcal{I}_t$, since $\mathcal{S} \subseteq \mathcal{M}_t$. Thus, we need to generate $(\mathcal{S}^c)^d$ and then check if each of those transversals are minimal t -infrequent sets, or not. To generate $(\mathcal{S}^c)^d$, let us observe that $(\mathcal{S}^c)^d \subset \bigotimes_{S \in \mathcal{S}} (\mathcal{C} \setminus S)$, and all the sets on the right hand side can easily be generated in $O(n^{|\mathcal{S}|})$ time.

Let us next remark that the inequality (1) is best possible, as the following examples show. Let A be an $m \times (m-t+1)$ matrix, in which every entry is 1, except the diagonal entries in the first $m-t+1$ rows, which are 0. Then any $m-t$ element subset of the columns is a maximal t -frequent set, while the set \mathcal{C} of all columns is the only minimal t -infrequent set. Thus we have equality in (1) for such matrices.

It is also worth mentioning that (1) stays accurate, up to a factor of $\log m$, even if $m \gg n$ and $|\mathcal{I}_t|$ is arbitrarily large. To see this, let us consider a binary matrix A with $m = 2^k$ rows and $n = 2k$ columns ($k \geq 1$, integer), having all the rows that contain exactly one 0 and one 1 in each pair of the adjacent columns $\{1, 2\}, \{3, 4\}, \dots, \{2k-1, 2k\}$. It is not difficult to see that for $t = 1$ there are 2^k maximal 1-frequent sets (every row of the matrix is the characteristic vector of a maximal 1-frequent set), and that there are only k minimal 1-infrequent sets, namely $\{2i-1, 2i\}$ for $i = 1, \dots, k$. Thus, for such examples we have $|\mathcal{M}_t| = (m/\log m)|\mathcal{I}_t|$.

The same example shows also that $|\mathcal{M}_t|$ cannot be bounded by a polynomial function of only $|\mathcal{I}_t|$ and n , the number of columns of A .

Needless to say that in general, $|\mathcal{I}_t|$ cannot be bounded by any polynomial in $|\mathcal{M}_t|$, n and m .

Let us add finally that Corollary 1 and Theorem 2 answer questions raised in [31] regarding the incremental complexity of generating maximal frequent and minimal infrequent sets.

2. Closed Frequent Sets

Following [32], let us call a subset $C \subseteq \mathcal{C}$ of the columns *closed* if $R(C') \not\subseteq R(C)$ for all $C' \supsetneq C$, that is, if $c \in C$ exactly when $A(r, c) = 1$ for all $r \in R(C)$ (see also [29, 30]). Let us further denote by \mathcal{D}_t the family of all closed t -frequent column sets. Clearly, we have

$$\mathcal{M}_t \subseteq \mathcal{D}_t \subseteq \mathcal{F}_t$$

for all $t = 1, \dots, m$.

It is clear that the family of closed sets may not be Sperner, and hence it may not correspond directly to a monotone system, in general. However, the sets $C \cup R(C)$ for closed sets $C \subseteq \mathcal{C}$ do correspond naturally to a monotone system. Namely, it is easy to see that the sets $C \cup R(C) \subseteq \mathcal{C} \cup \mathcal{R}$ for closed subsets $C \subseteq \mathcal{C}$ together with \mathcal{C} and \mathcal{R} form the family of maximal independent sets of the bipartite graph the vertex set of which is $\mathcal{C} \cup \mathcal{R}$, and in which a pair (r, c) for $r \in \mathcal{R}$ and $c \in \mathcal{C}$ is an edge if and only if $A(r, c) = 0$. Thus, according to the results of [21], closed frequent sets can be generated with polynomial delay (in terms of the number of rows and columns of A). The generation of all closed sets of course has the disadvantage that there may be too many with $|R(C)|$ small. In what follows, we show that those closed sets for which $|R(C)|$ exceeds a given threshold can also be generated efficiently.

It can be seen by the definitions that every closed t -frequent set is also a maximal t' -frequent set for some $t' \geq t$, implying the following claim.

PROPOSITION 1. $\mathcal{D}_t = \cup_{t' \geq t} \mathcal{M}_{t'}$. □

Let us note next that for $C \in \mathcal{D}_t \setminus \mathcal{D}_{t+1}$ we either have a nonempty subset $C' \subsetneq C$ with $C' \in \mathcal{D}_{t+1}$, or $A(r, c) = 0$ for all $c \in C$ and $r \notin R(C)$. For sets of the latter type we must have $R(C) = R(\{i\})$ for

some $i \in \mathcal{C}$, and thus the number of such subsets is limited by n , and it is easy to identify those in $O(mn^2)$ time. The sets in $\mathcal{D}_t \setminus \mathcal{D}_{t+1}$ of the first type can be obtained by trying to increment all sets of \mathcal{D}_{t+1} in all possible ways. For every $C \in \mathcal{D}_{t+1}$ let us consider the hypergraph on $R(C)$ defined by the columns $C' \in \mathcal{C} \setminus C$ (as characteristic vectors), and let H_1, H_2, \dots be the maximal subsets in this hypergraph. Let C_i denotes the set of columns $C' \in \mathcal{C} \setminus C$ corresponding to H_i . Then it is enough to consider $C \cup C_1, C \cup C_2, \dots$ for each $C \in \mathcal{D}_{t+1}$, as the candidates for subsets in $\mathcal{D}_t \setminus \mathcal{D}_{t+1}$. These maximal hyperedges can be determined in $O(|\mathcal{C} \setminus C|^2 |R(C)|)$ time, and thus all sets of the first type can be generated in $O(n^2 m |\mathcal{D}_{t+1}|)$ time. Therefore, all sets of $\mathcal{D}_t \setminus \mathcal{D}_{t+1}$ can be generated in $O(n^2 m |\mathcal{D}_{t+1}|)$ time.

Finally, denoting by τ the maximum number of 1's in a column of A , we can claim that $\mathcal{D}_t = \emptyset$ for all $t > \tau$, and that \mathcal{D}_τ can easily be generated in $O(nm)$ time.

Putting all these together, we can conclude that, in contrast to maximal frequent sets, closed frequent sets can be generated efficiently in incremental polynomial time.

PROPOSITION 2. *The family \mathcal{D}_t can be generated in incremental polynomial time for any $t \in \{1, \dots, m\}$.* \square

Let us finally remark that we can have $|\mathcal{D}_t|$ be exponentially larger than $|\mathcal{M}_t|$ and at the same time $|\mathcal{F}_t|$ be exponentially larger than $|\mathcal{D}_t|$. To see such an infinite family of examples, let us choose positive integers $k, l (> k)$ and t , set $m = kt, n = kl$, and define the matrix A as follows. Let $U_i = \{(i-1)l + j \mid j = 1, \dots, l\}$ for $i = 1, \dots, k$, let $\mathcal{C} = \cup_{i=1}^k U_i$, and let $a_i \in \{0, 1\}^n$ ($1 \leq i \leq k$) be the binary vector in which $a_{ij} = 0$ if $j \in U_i$, and $a_{ij} = 1$ otherwise. Finally, let $A \in \{0, 1\}^{m \times n}$ be the matrix formed by t copies, as rows, of each vectors $a_i, i = 1, \dots, k$.

It is now easy to see that the maximal t -frequent sets in this matrix are exactly the column subsets C of the form $C = \mathcal{C} \setminus U_i$ for some $1 \leq i \leq k$. Thus, $|\mathcal{M}_t| = k$. Furthermore, the column subsets of the form $C = \mathcal{C} \setminus (\cup_{i \in S} U_i)$ for nonempty subsets $S \subseteq \{1, \dots, k\}$ are exactly the closed t -frequent sets of A , therefore we have $|\mathcal{D}_t| = 2^k - 1$. Furthermore, any subset $C \subseteq \mathcal{C}$ of the columns, disjoint from at least one of the sets U_1, \dots, U_k , is a t -frequent set, implying that $|\mathcal{F}_t| > 2^{(t-1)k} > |\mathcal{D}_t|^{t-1}$. Finally, any minimal transversal of the subsets U_i for $i = 1, \dots, k$ is a minimal t -infrequent set, thus we have $|\mathcal{I}_t| = l^k$.

In summary, for these examples we have $m = kt, n = kl$ and

$$|\mathcal{M}_t| = k \ll |\mathcal{D}_t| = 2^k - 1 \ll |\mathcal{I}_t| = l^k \ll 2^{k(l-1)} < |\mathcal{F}_t|. \quad (2)$$

3. Proofs of Theorems 1 and 2

For the proof of Theorem 1 we shall need the following combinatorial lemma [8]. For completeness, we provide the proof of the lemma below.

LEMMA 1. *Given a base set V of size $|V| = m$ and a threshold $t \in \{1, \dots, m\}$, let $\mathcal{S} = \{S_1, \dots, S_\alpha\}$ and $\mathcal{T} = \{T_1, \dots, T_\beta\}$ be two families of subsets of V such that*

- (i) $|S| \geq t$ for all $S \in \mathcal{S}$, while $|T| < t$ for all $T \in \mathcal{T}$, and
- (ii) for each of the $\alpha(\alpha-1)/2$ pairs $S', S'' \in \mathcal{S}$ there exists a $T \in \mathcal{T}$, such that $S' \cap S'' \subseteq T$.

Then $\alpha \leq (m - t + 1)\beta$, whenever $\alpha \geq 2$.

Let us remark first that if $\alpha = 1$ then the family \mathcal{T} might be empty, which would violate the inequality $\alpha \leq (m - t + 1)\beta$. Let us also mention that by (ii) $\beta \geq 1$ must hold whenever $\alpha \geq 2$. In addition, conditions (i) and (ii) together imply that \mathcal{S} is a Sperner family, i.e. $S_i \not\subseteq S_j$ whenever $i \neq j$ (since otherwise $S_i = S_i \cap S_j \subseteq T_k$ would follow by (ii) for some $T_k \in \mathcal{T}$, contradicting condition (i).) Without loss of generality we can assume that \mathcal{T} is also Sperner, for otherwise we can replace \mathcal{T} by the family of all maximal sets of \mathcal{T} .

Proof of Lemma 1. We shall prove the Lemma by induction on t . If $t = 1$ then $\mathcal{T} = \{\emptyset\}$ by condition (i). In view of (ii), this implies that the sets of \mathcal{S} are pairwise disjoint, and hence $\alpha \leq m = (m - t + 1)\beta$.

In a general step, let us define subfamilies $\mathcal{S}_v = \{S \setminus \{v\} \mid S \in \mathcal{S}, v \in S\}$ and $\mathcal{T}_v = \{T \setminus \{v\} \mid T \in \mathcal{T}, v \in T\}$ for each $v \in V$. Let us further introduce the notations $\alpha_v = |\mathcal{S}_v|$, and $\beta_v = |\mathcal{T}_v|$.

For vertices $v \in V$ for which $\alpha_v \geq 2$ (and thus $\beta_v \geq 1$) the families \mathcal{S}_v and \mathcal{T}_v satisfy all the assumptions of the Lemma with $m' = m - 1$ and $t' = t - 1$, and hence

$$\alpha_v \leq (m' - t' + 1)\beta_v = (m - t + 1)\beta_v \quad (3)$$

follows by the inductive hypothesis. Let us then consider the partition $V = V_1 \cup V_2$, where $V_1 = \{v \in V \mid \alpha_v \leq 1\}$, and $V_2 = \{v \in V \mid \alpha_v \geq 2\}$. Summing up the inequalities (3) for all $v \in V_2$, we obtain

$$\sum_{v \in V_2} \alpha_v \leq (m - t + 1) \sum_{v \in V_2} \beta_v. \quad (4)$$

On the left hand side, using condition (i) and the definition of α_v we obtain

$$\alpha t - |V_1| \leq \sum_{S \in \mathcal{S}} |S| - |V_1| \leq \sum_{S \in \mathcal{S}} (|S| - |S \cap V_1|) = \sum_{S \in \mathcal{S}} |S \cap V_2| = \sum_{v \in V_2} \alpha_v, \quad (5)$$

where the first inequality follows by $|S| = \alpha$ and $|S| \geq t$ for $S \in \mathcal{S}$, while the second one is implied by $|V_1| \geq \sum_{S \in \mathcal{S}} |S \cap V_1|$, which follows from the definition of V_1 .

On the right hand side of (4) we can write

$$\sum_{v \in V_2} \beta_v = \sum_{T \in \mathcal{T}} |T \cap V_2| \leq \sum_{T \in \mathcal{T}} |T| \leq \beta(t-1), \quad (6)$$

where the first equality follows by the definition of β_v and \mathcal{T}_v , and the last inequality follows by the conditions $|T| < t$ for $T \in \mathcal{T}$.

Putting together (4), (5) and (6) we obtain $\alpha t - |V_1| \leq (m-t+1)(t-1)\beta$, or equivalently that

$$\alpha \leq \frac{|V_1|}{t} + \frac{t-1}{t}(m-t+1)\beta. \quad (7)$$

If $|V_1| \leq m-t+1$, then

$$\frac{|V_1|}{t} + \frac{t-1}{t}(m-t+1)\beta \leq (m-t+1)\beta,$$

and hence $\alpha \leq (m-t+1)\beta$ by (7). On the other hand, if $|V_1| > m-t+1$, then for each set $S \in \mathcal{S}$ we have $|S \cap V_1| \geq |S| - |V_2| \geq t - |V_2| > 1$. Now by the definition of the set V_1 we obtain $\alpha \leq |V_1|/(t - |V_2|) = (m - |V_2|)/(t - |V_2|) \leq m - t + 1 \leq (m - t + 1)\beta$. \square

Proof of Theorem 1. Assume without loss of generality that $|\mathcal{M}_t| \geq 2$, for otherwise (1) readily follows from the assumption of the theorem that $|\mathcal{I}_t| \geq 1$. Let us recall that to any subset $C \subseteq \mathcal{C}$ of the columns we have associated the subset $R(C)$ of those rows $r \in \mathcal{R}$ for which $A(r, c) = 1$ for every column $c \in C$. Thus, by definition we have $R(C) = \bigcap_{y \in C} R(\{y\})$, implying

$$R(C' \cup C'') = R(C') \cap R(C'') \quad \text{for all } C', C'' \subseteq \mathcal{C}. \quad (8)$$

In its turn, (8) implies that the mapping $C \mapsto R(C)$ is anti-monotone, i.e. $R(C') \supseteq R(C'')$ whenever $C' \subseteq C''$. Furthermore, $|R(F)| \geq t$ for every maximal t -frequent set $F \in \mathcal{M}_t$, while $|R(U)| < t$ for every minimal t -infrequent set $U \in \mathcal{I}_t$. It is also easy to see that the restriction of the above mapping on \mathcal{M}_t is injective, i.e. $R(F') \neq R(F'')$ for any two distinct maximal t -frequent sets of columns $F', F'' \in \mathcal{M}_t$. If $F', F'' \in \mathcal{M}_t$ then their union $F' \cup F''$ is not t -frequent, and hence there exists

a minimal t -infrequent set $U \in \mathcal{I}_t$, for which $R(F') \cap R(F'') = R(F' \cup F'') \subseteq R(U)$. Thus, the families $\mathcal{S} = \{R(F) \mid F \in \mathcal{M}_t\}$ and $\mathcal{T} = \{R(U) \mid U \in \mathcal{I}_t\}$ satisfy the conditions of Lemma 1 with $V = \mathcal{R}$, which implies the inequality $|\mathcal{S}| \leq (m - t + 1)|\mathcal{T}|$. Since the mapping $C \mapsto R(C)$ is a one-to-one correspondence between \mathcal{M}_t and \mathcal{S} , we have $|\mathcal{S}| = |\mathcal{M}_t|$. Now (1) follows from the trivial inequality $|\mathcal{T}| \leq |\mathcal{I}_t|$. \square

Proof of Theorem 2. We reduce our problem from the following well-known NP-complete problem: Given a graph $G = (V, E)$ and an integer threshold t , determine if G contains an independent vertex set of size at least t . Let us first substitute every vertex $v \in V$ of G by two new vertices v' and v'' connected by an edge, i.e., consider the graph $G' = (V', E')$, where $V' = \{v', v'' \mid v \in V\}$ and $E' = \{(v', v'') \mid v \in V\} \cup \{(v', u'), (v', u''), (v'', u'), (v'', u'') \mid (u, v) \in E\}$. Clearly, $G' = (V', E')$ has an independent set of size t if and only if G has one, moreover, if G' has one, then it has at least 2^t .

Let us now associate a matrix $A = A_{G', t}$ to G' as follows. Let $\mathcal{C} = V'$ be the set of columns of the matrix A . To every edge $(v, w) \in E'$ we assign $t - 2$ identical rows in A containing 0 in the columns v and w , and 1 in all other columns. Furthermore, to every vertex $v \in V'$ we assign one row containing 0 in the column v and 1 in all other columns. Thus, A has $m = (t - 2)|E'| + |V'| = t|V| + 4(t - 2)|E|$ rows, and $n = |V'| = 2|V|$ columns.

Clearly, for every edge $e = (v, w) \in E'$ the set $C_e = \mathcal{C} \setminus \{v, w\}$ is a maximal t -frequent set of A . Let $\mathcal{S} = \{C_e \mid e \in E'\}$. We claim that $\mathcal{S} \neq \mathcal{M}_t$ for this matrix if and only if there exists an independent set I of size $|I| \geq t$ in the graph G' .

To see this claim, let us assume first that $I \subseteq V'$ is an independent set of G' such that $|I| \geq t$. Then $R(V' \setminus I)$ contains all rows corresponding to vertices $v \in I$, and hence $|R(V' \setminus I)| \geq |I| \geq t$. Since I does not contain an edge of the graph, the set $C = V' \setminus I$ is not a subset of the member of \mathcal{S} , and thus it is contained by a maximal t -frequent set of the matrix A , which does not belong to \mathcal{S} .

For the other direction, let us assume now that $C \subseteq \mathcal{C} = V'$ is a maximal t -frequent set of A , not contained by any member of \mathcal{S} . This latter implies that $I = V' \setminus C$ does not contain an edge of G' , i.e. that I is an independent set of G' . This also implies that $R(C)$ cannot contain any of the rows corresponding to an edge of G' , and hence $|R(C)| = |V' \setminus C| = |I|$. Thus, $|I| \geq t$ follows by our assumption that C is a t -frequent set, i.e. I is an independent set of size at least t .

Let us recall finally that the maximum independent set problem remains NP-complete, even if the input is restricted to cubic planar graphs (see e.g. [17]), i.e. we can assume $|E| = O(|V|)$. Therefore, we

have $|\mathcal{S}| = |E'| = |V| + 4|E| = O(|V|)$, and either we have $\mathcal{S} = \mathcal{M}_t$, or $|\mathcal{M}_t| \geq |\mathcal{S}| + 2^t$. Since we can assume without loss of generality that $t = \Theta(|V|)$, we obtain the the statement of the theorem for $|\mathcal{S}| = O(n)$, that is for $\varepsilon = 1$. For smaller values of ε it suffices to add $n^{1/\varepsilon}$ isolated vertices to G' . \square

4. Conclusions

In this paper we considered various types of frequent sets, their generation, and connections between them.

It is well-known that the family \mathcal{F}_t of t -frequent sets of a given $m \times n$ matrix A can efficiently be generated incrementally. Unfortunately, this family may be exponentially larger than the families of maximal t -frequent sets, or closed t -frequent sets, the knowledge of which would be sufficient for most practical purposes (e.g. finding "interesting" association rules). As alternative approaches, we considered in this paper the generation of maximal frequent sets, boundary sets, and closed frequent sets.

We established that the family \mathcal{M}_t of maximal t -frequent sets cannot be generated efficiently, unless $P=NP$. On the other hand, our results imply that the family \mathcal{I}_t of minimal t -frequent sets, as well as the boundary set $\mathcal{M}_t \cup \mathcal{I}_t$ can both be generated in incremental quasi-polynomial time. More precisely, the incremental complexity of generation for these families is polynomially equivalent with the complexity of the hypergraph transversal problem, which is still an open problem, and for which a quasi-polynomial algorithm was presented in [15].

An alternative approach is the generation of \mathcal{D}_t , the family of closed t -frequent sets. Since most publications about closed frequent sets do not discuss the complexity of their generation, for completeness we included an easy argument showing that \mathcal{D}_t can be generated in incremental polynomial time.

For the purpose of determining the family \mathcal{M}_t , generating the boundary $\mathcal{M}_t \cup \mathcal{I}_t$ and closed t -frequent sets \mathcal{D}_t are the two most competitive, and not fully comparable approaches. Our examples show that there are infinitely many examples when $\mathcal{M}_t \cup \mathcal{I}_t$ is exponentially larger than \mathcal{D}_t (see the examples in Section 2, as in (2)), and vice versa. For the latter case let us consider the examples $A_{G,t}$ as defined in the previous section in the proof of Theorem 2. In these examples any subset of the columns avoiding both endpoints of an edge of G is t -frequent and also closed (due to the first n rows). Maximal t -frequent sets are the complements of edges of G as well as stable sets of size t , while minimal

t -infrequent sets correspond to complements of stable sets of size $t - 1$ as well as maximal stable sets of size at most $t - 1$. Thus for graphs G with polynomially many maximal stable sets and with $t = \alpha(G)$ (e.g. threshold graphs, chordal graphs, etc.) or for graphs having small stable sets only we have

$$\max\{|\mathcal{M}_t|, |\mathcal{I}_t|\} \ll 2^{n-2} < |\mathcal{F}_t| = |\mathcal{D}_t|. \quad (9)$$

Thus, in these cases the incremental quasi-polynomial generation of the boundary is still much faster for obtaining all maximal t -frequent sets, than the incrementally polynomial generation of \mathcal{D}_t .

Let us also remark that the incremental polynomial generation of \mathcal{D}_t in examples when $|\mathcal{D}_t| \ll |\mathcal{I}_t|$ is seemingly in contradiction with the claim we cited earlier, namely that a membership-oracle based generation of \mathcal{M}_t needs at least $|\mathcal{M}_t \cup \mathcal{I}_t|$ steps, since from \mathcal{D}_t we could easily generate \mathcal{M}_t , much faster in such a case than the $|\mathcal{M}_t \cup \mathcal{I}_t|$ bound. The potential contradiction is resolved however if we note that the efficient generation of \mathcal{D}_t is possible only by utilizing the matrix A in a much more substantial way than a membership-oracle would do (see Section 2).

Let us finally remark that t -frequent sets can also be viewed as $m' \times n'$ submatrices of A with $m' \geq t$ in which every entry is nonzero. It is interesting to note that finding such a submatrix of maximum perimeter, i.e. maximizing $m' + n'$ is a polynomially solvable optimization problem, due to the König-Egerváry theorem, since it corresponds to a maximum independent set in the bipartite graph defined by the zeros in A . On the other hand, finding a completely 1 submatrix of A having the largest area, i.e. maximizing $m' \times n'$ is an NP-hard optimization problem [23].

References

1. R. Agrawal, T. Imielinski and A. Swami, Mining associations between sets of items in massive databases. In: *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data*, pp. 207-216.
2. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy eds., *Advances in Knowledge Discovery and Data Mining*, 307-328, AAAI Press, Menlo Park, California, 1996.
3. R. Agrawal and R. Srikant, Mining sequential patterns. In: *Proceedings of the 11th International Conference on Data Engineering, 1995*, pp.3-14.
4. R. J. Bayardo, Efficiently mining long patterns from databases. In: *Proceedings of the 1998 ACM-SIGMOD International Conference on Management of Data*, pp. 85-93.

5. J. C. Bioch and T. Ibaraki, Complexity of identification and dualization of positive Boolean functions. *Information and Computation* 123 (1995) 50-63.
6. M. M. Bongard, *Problema Unznawania*, Nauka Press, Moscow, 1967. English translation: *Pattern Recognition*, Hayden Book Co., Spartan Book, Rochelle Park, New Jersey, USA, 1970.
7. E. Boros, V. Gurvich, L. Khachiyan and K. Makino, Generating partial and multiple transversals of a hypergraph. In: *Proceedings of the 27th International Colloquium on Automata, Languages and Programming (ICALP)*, (U. Montanari, J.D.P. Rolim and E. Welzl, eds.) Lecture Notes in Computer Science **1853** pp. 588-599, (Springer Verlag, Berlin, Heidelberg, New York, 2000).
8. E. Boros, V. Gurvich, L. Khachiyan and K. Makino, Dual-bounded generating problems: Partial and multiple transversals of a hypergraph. *SIAM Journal on Computing*, **30** (2001) pp. 2036-2050.
9. S. Brin, R. Motwani, and C. Silverstein, Beyond market basket: Generalizing association rules to correlations. In: *Proceedings of the 1997 ACM-SIGMOD Conference on Management of Data*, pp. 265-276.
10. S. Brin, R. Motwani, J. Ullman, and S. Tsur, Dynamic itemset counting and implication rules for market basket data. In: *Proceedings of the 1997 ACM-SIGMOD Conference on Management of Data*, pp. 255-264.
11. B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, Cambridge University Press, 1990.
12. G. Dong and J. Li, Efficient mining of emerging patterns. In: *Proceeding of the 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 43-52.
13. T. Eiter and G. Gottlob, Identifying the minimal transversals of a hypergraph and related problems. *SIAM Journal on Computing*, 24 (1995) 1278-1304.
14. D. Eppstein, Arboricity and bipartite subgraph listing algorithms. *Information Processing Letters* **51** (1994), pp. 207-211.
15. M. L. Fredman and L. Khachiyan, On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithms*, 21 (1996) 618-628.
16. B. Ganter and R. Wille, *Formal Concept Analysis*, Springer, 1996.
17. M. R. Garey and D. S. Johnson, *Computers and Intractability*, Freeman, New York, 1979.
18. D. Gunopulos, R. Khardon, H. Mannila, and H. Toivonen, Data mining, hypergraph transversals and machine learning. In: *Proceedings of the 16th ACM-SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, (1997) pp. 12-15.
19. V. Gurvich and L. Khachiyan, On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions. *Discrete Applied Mathematics*, 1996-97, issue 1-3, (1999) 363-373.
20. J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation. In: *Proceedings of the 2000 ACM-SIGMOD Conference on Management of Data*, pp. 1-12.
21. D. S. Johnson, M. Yannakakis and C. H. Papadimitriou, On generating all maximal independent sets, *Information Processing Letters*, **27** (1988) 119-123.
22. S. O. Kuznetsov, Interpretation on graphs and complexity characteristics of a search for specific patterns, *Nauchn. Tekh. Inf., Ser. 2 (Automatic Document. Math. Linguist.)* **23**(1), (1989) pp. 23-37.
23. V. Levit, private communication, 2000.

24. D. Lin and Z. M. Kedem, Pincer-search: a new algorithm for discovering the maximum frequent set. In: *Proceedings of the Sixth European Conference on Extending Database Technology*, to appear.
25. K. Makino and T. Ibaraki, Inner-core and outer-core functions of partially defined Boolean functions, *Discrete Applied Mathematics*, 1996-97, issue 1-3 (1999), 307-326.
26. H. Mannila and H. Toivonen, Multiple uses of frequent sets and condensed representations. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, (1996) pp. 189-194.
27. H. Mannila and H. Toivonen, Levelwise search and borders of theories in knowledge discovery. Series of Publications C C-1997-8, University of Helsinki, Department of Computer Science (1997).
28. H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1 (1997), 259-289.
29. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Discovering frequent closed itemsets for association rules. *Proc. of the 7th ICDT Conference*, Jerusalem, Israel, January 10-12, 1999; *Lecture Notes in Computer Science*, **1540**, pp. 398-416, Springer Verlag, 1999.
30. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Closed set based discovery of small covers for association rules. *Proc. 15emes Journees Bases de Donnees Avancees, BDA*, pp. 361-381, 1999.
31. R. H. Sloan, K. Takata, G. Turan, On frequent sets of Boolean matrices. *Annals of Mathematics and Artificial Intelligence* 24 (1998) 1-4.
32. M. J. Zaki and M. Ogihara, Theoretical foundations of association rules. *3rd SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, June 1998.

