Theoretical Computer Science 596 (2015) 124–141

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

An average study of hypergraphs and their minimal transversals

Julien David [a,*], Loïck Lhote [b,*], Arnaud Mary [c,*], François Rioult [b,*]

[a] LIPN, Université Paris 13, and CNRS, UMR 7030, 99, av. J.-B. Clément, 93430 Villetaneuse, France
[b] GREYC, CNRS, ENSICAEN and Université de Caen, Caen, France
[c] Université Lyon 1, CNRS, UMR5558 LBBE, France

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form 19 May 2015
Accepted 23 June 2015
Available online 8 July 2015
Communicated by P.G. Spirakis

Keywords:
Hypergraph
Minimal transversal
Enumeration
Probabilistic analysis
Average complexity
Generic complexity

ABSTRACT

In this paper, we study some average properties of hypergraphs and the average complexity of algorithms applied to hypergraphs under different probabilistic models. Our approach is both theoretical and experimental since our goal is to obtain a random model that is able to capture the real-data complexity. Starting from a model that generalizes the Erdös–Renyi model [10,11], we obtain asymptotic estimations on the average number of transversals, irredundants and minimal transversals in a random hypergraph. We use those results to obtain an upper bound on the average complexity of algorithms to generate the minimal transversals of a hypergraph. Then we make our random model more complex in order to bring it closer to real-data and identify cases where the average number of minimal transversals is at most polynomial, quasi-polynomial or exponential.

1. Introduction

A hypergraph is a pair $\mathcal{H} = (V, \mathcal{E})$ where $V = \{1, 2, \dots, n\}$ is the set of vertices and $\mathcal{E} = (E_1, \dots, E_m)$ is the collection of hyperedges with $E_i \subseteq V$ for all i.

A *transversal* is a set of vertices that intersects all the hyperedges. A set of vertices X is said to be *irredundant* if for all vertices $i \in X$, there exists a hyperedge H such that $H \cap X = \{i\}$. X is called a *minimal transversal* when it is a transversal and none of its subset is transversal. This is equivalent to being both irredundant and a transversal.

Given a hypergraph $\mathcal{H}$, the set of all its minimal transversals forms a hypergraph called the *transversal hypergraph*.

The Transversal Hypergraph Generation problem (for short, THG-problem) consists in computing the transversal hypergraph of a given hypergraph. In the same way, the associated decision problem (in short, THD-problem) consists in deciding if a first hypergraph $\mathcal{H}_1$ is the transversal hypergraph of a second one $\mathcal{H}_2$. This problem is known to be equivalent to the famous dualization of monotone boolean functions problem (see [8]). The Transversal Hypergraph Generation problem appears in very different domains: Artificial Intelligence and Logic [6,7], Biology [2], Datamining and Machine Learning [14], mobile communications [23], *etc.* We refer to [15] for a more complete list of applications.

Since a hypergraph may have an exponential number of minimal transversals, the THG-problem does not belong to the class of polynomial problems. However, a long standing question is to decide whether there exists an algorithm to solve

[*] Corresponding authors.
E-mail address: Julien.David@lipn.univ-paris13.fr (J. David).

http://dx.doi.org/10.1016/j.tcs.2015.06.052
0304-3975/© 2015 Elsevier B.V. All rights reserved.

THG whose running time is a polynomial on the size of the hypergraph and on the number of minimal transversal. Such an algorithm is called an *output-polynomial* time algorithm.

The complexity of the THG-problem is closely related to its associated decision problem THD. Precisely, if an output-polynomial algorithm solves THG, then THD can be solved in polynomial time. In addition, THD is clearly in the class of co-NP problems but there is no evidence of its co-NP-completeness. If THD is co-NP-complete, then no output polynomial algorithm is likely to exist for the generation problem THG (unless P = co-NP) [6].

The best known algorithm to generate the transversal hypergraph is quasi-polynomial and is due to Fredman and Khachiyan in [13]. Its running time is of the form $N^{o(\log N)}$ with N the size of the input plus the output. Nevertheless, this algorithm is not efficient for practical applications. Other algorithmic solutions were proposed and a list of them can be found in [9]. In this article, we focus on the MTMiner algorithm defined by Hébert, Bretto and Crémilleux [16]. MTMiner is closely related to the mining of the frequent patterns in data mining and is clearly output-exponential in the worst-case. We will study both average complexity and generic-case [17] output-sensitive complexity of the algorithm.

In the previous quoted results, the complexity of the THG-problem and associated algorithms were mostly studied with the worst-case point of view. Indeed, very specific entries were exhibited in order to obtain worst-case lower or upper bounds on the behavior of the algorithms. But these entries do not generally occur in practice, and the existing worst-case analyses are then not sufficient to understand the practical complexity of THG. In this article, we adopt a probabilistic point of view. Though analytic combinatorics is often used to conduct an average case study, the symbolic method [12] (a generic method used to describe the recursive decomposition of combinatorial objects) does not seem to be relevant in our case, as it cannot be used to describe the patterns we are interested in.

The study of random hypergraphs under various distributions is quite common and one of the most popular is the uniform distribution on k-uniform hypergraphs [1,5,19] (in which all hyperedges have the same cardinal k). In [22], the authors study the creation and the growth of components with a continuous random hypergraph process. Recently, De Panafieu [3] studies random non-uniform hypergraphs and their structures near the birth of the complex components. In [24], the authors prove that under the uniform distribution over all the simple hypergraphs with n vertices, the THG problem is output-polynomial with probability close to 1. In fact, under this distribution, the size of the transversal hypergraph is with high probability exponential in n and even the naive algorithm that goes through the whole search space is almost surely output-polynomial. To the best of our knowledge, this is the only study on the average complexity of the THG problem.

In this paper, we consider two random models in which the number n of vertices and the number m of hyperedges are given and suppose that m is a polynomial in n. The results we obtain are original and do not intersect with [24].

Plan of the paper: Section 2 is devoted to the probabilistic models we consider. In Section 2.1, we introduce a *single-parameter* model that generalizes the Erdös–Renyi model [10,11]. In Section 2.2, we make our random model more complex so that the probability that each vertex appears in a hyperedge is given by a function. Section 3 summarizes the asymptotic results we obtain with the single-parameter model on the average number of transversals, irredundants and minimal transversals. Section 4 is devoted to the results with the second probabilistic model. Section 5 is devoted to algorithms analysis. We study the average complexity of the MTMiner algorithm and the generic-case complexity of the THG-problem. The average complexity of MTMiner is closely related to the average number of irredundants: we obtain upper bounds on the average complexity for both models. Section 6 is devoted to experimental results. Using hypergraphs obtained from real datasets, we discuss the consistency of our random models. Section 7 contains all the proofs of the main results. Conclusion is devoted to perspectives and indications on a random model that might be interesting for a future work.

2. Probabilistic models for hypergraphs

In this paper, we study the average properties of hypergraphs under two probabilistic models. For both models we suppose that:

- The number of hyperedges m is at most polynomial in the number of vertices n, precisely $\ln m = \Theta(\ln n)$. Some of our results do not require this assumption and can therefore be extended to cases where m is exponential in n ($\ln m = \Theta(n)$). In this case, most questions we study in this paper become trivial.
- A hypergraph with n vertices and m hyperedges can be seen as a binary matrix $M(\mathcal{H}) = (m_{i,j}(\mathcal{H}))_{i=1..m, j=1..n}$. Each row in the matrix encodes a hyperedge. The value $m_{i,j}$ at line i and column j is equal to 1 if the vertex j belongs to the hyperedge encoded in row i, $m_{i,j} = 0$ otherwise.
- The variables $(m_{i,j}(\mathcal{H}))_{i=1..m, j=1..n}$ form an independent family of random variables. In other words, the event that a given vertex v appears in a given hyperedge is independent from the event that a vertex u appears in the same hyperedge or from the event that v appears in another hyperedge.

2.1. The single-parameter model

We first introduce our simplest model which is close to the Erdös–Renyi model [10,11] for graphs.

Definition 1 (**HG**(**n**, **m**, **p**) *random model).* The $HG(n, m, p)$ model supposes that the family of random variables $(m_{i,j})_{i=1..m, j=1..n}$ forms an independent and identically distributed family of random variables following the same Bernoulli law of parameter p $(0 < p < 1)$ with $1 - e^{-\frac{1}{\ln n}} < p < e^{-\frac{1}{\ln n}}$.

The bounds on p are rather technical but they will be useful in the multiparametric model. The single-parameter model entails that all the vertices have the same behavior. This is a non-realistic hypothesis but this simple model allows a precise probabilistic analysis of the minimal transversal problem. In addition, the results we obtain can be used for analyses in the more realistic multiparametric model that we now introduce.

2.2. Multiparametric model

We no longer consider that the vertices occur in a hyperedge with the same probability. As we will show in Section 6, this is much more consistent with real-case datasets.

Definition 2 *(HG(n, m, g) random model).* A hypergraph $\mathcal{H}$ with m hyperedges and n vertices is seen as a binary matrix $M(\mathcal{H}) = (m_{i,j}(\mathcal{H}))_{i=1..m,j=1..n}$. The $HG(n, m, g)$ model supposes that the family of random variables $(m_{i,j})_{i=1..m,j=1..n}$ forms an independent family of random variables. In addition, for all i, j, the random variable $m_{i,j}$ follows a Bernoulli law of parameter $p_j = g(j)$ (and $q_j = 1 - g(j)$) with $g : \mathbb{N} \to [0, 1]$.

The asymptotic analyses show that the role of a vertex v is different depending on whether its probability p_v is large or small. Precisely, we partition the set V into 3 subsets:

- The set U of ubiquitous vertices. Let x be a fixed constant. For all vertices $u \in U$, we have that $q_u < \frac{x}{m}$ with $q_u = 1 - p_u$.
- The set R of rare events. For all vertices $r \in R$, we have that $p_r < 1 - e^{-\frac{1}{\ln n}}$. Note that this implies that $p_r < \frac{1}{\ln n}$. The latter bound slowly tends to zero and is relevant on experimental data (see Section 6), the former simplifies calculus.
- The set O of other events, that is $O = V \setminus \{U \cup R\}$.

We mainly use this decomposition to study the average number of minimal transversals in the $HG(n, m, g)$ model. In particular, we prove that the number of rare events is an important parameter to control the number of minimal transversals.

3. Main results with the single parameter model

A minimal transversal is both a transversal and irredundant. Then for each probabilistic model, we precisely describe the average behavior of the number of transversals, of the number of irredundants and of the number of minimal transversals. Based on these results, we study in Section 5 the average complexity of the MTMiner algorithm [16] that computes the transversal hypergraph. We conclude with an average point of view on the Transversal Hypergraph Generation problem. This section is devoted to the single parameter model whereas Section 4 deals with the multiparametric model. The proofs relative to this section are rather technical with mainly asymptotic calculations. In addition, they are not fundamental for the understanding of the results. This is why they are given in the last section (Section 7).

3.1. On the average number of transversals

In the following, T_j is the number of transversals of size j on a given hypergraph. Recall that, a given subset $X \subseteq V$ of vertices is a transversal if for all hyperedges H of the hypergraph, there exists at least one vertex $v \in X$ such that $v \in H$. We note $q = 1 - p$, the probability that a vertex does not appear in a hyperedge and that m is the number of hyperedges. The probability for a subset X of size j to be a transversal is therefore:

$$\mathbb{P}(X \text{ is a transversal}) = (1 - q^j)^m.$$

The following results, obtained by simple calculations on this probability, describe the asymptotic of T_j in function of j. The $\frac{1}{q}$ in the formula are the basis of the logs.

Proposition 1. *In the* **HG(n, m, p)** *random model, consider j of the form*

$$j = \log_{\frac{1}{q}} m + \log_{\frac{1}{q}} x, \qquad x \in \left]\frac{1}{m}, +\infty\right[.$$

1. *If $x \leq 1/\log_{\frac{1}{q}}(m \ln n)$, the average number of transversals $\mathbb{E}[T_j]$ tends to 0.*
2. *If x tends to 0 (with x lower bounded by $1/m$ in order for j to remain positive), the number of transversals is negligible compared to the number of subsets of the same size, i.e. $\mathbb{E}[T_j] = o\left(\binom{n}{j}\right)$.*
3. *If $x = \Theta(1)$, only a constant proportion of all sets of j vertices is a transversal, i.e., $\mathbb{E}[T_j] \sim \binom{n}{j} \exp(-1/x)$.*
4. *If x tends to $+\infty$, almost every sets of j vertices is a transversal, i.e., $\mathbb{E}[T_j] \sim \binom{n}{j}\left(1 - \frac{1}{x}\right)$.*

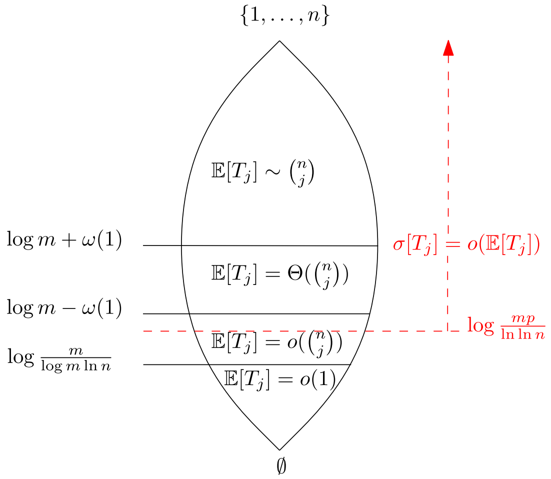
$\{1,\ldots,n\}$
$\mathbb{E}[T_j] \sim \binom{n}{j}$
$\log m + \omega(1)$
$\mathbb{E}[T_j] = \Theta(\binom{n}{j})$
$\sigma[T_j] = o(\mathbb{E}[T_j])$
$\log m - \omega(1)$
$\log \frac{mp}{\ln \ln n}$
$\log \frac{m}{\log m \ln n}$
$\mathbb{E}[T_j] = o(\binom{n}{j})$
$\mathbb{E}[T_j] = o(1)$
$\emptyset$

Fig. 1. We represented the boolean lattice on the set of vertices. As we will show in the following section, according to the size of the subsets, we can estimate the proportion of transversals. Above a given size, we also have a result on the standard-deviation.

Intuitively, the set of minimal transversals will mostly be included in cases 2 and 3. Indeed, in case 4, the probability that a given transversal does not contain a subset of case 3 which is also transversal, will intuitively be low, since the filter of transversals in case 3 will contain a huge number of transversals of case 4.

When j is sufficiently large, the next proposition shows that the standard deviation of T_j is negligible compared to its mean. Combined with Bienaymé–Chebyshev's inequality, this proves that the number of transversals of size j is almost surely equivalent to the mean number. This result will be fundamental to obtain an almost sure lower bound on the number of minimal transversals.

Proposition 2. *In the* $\mathbf{HG}(\mathbf{n}, \mathbf{m}, \mathbf{p})$ *random model, the standard deviation of the number of transversals of size* j *with* $j > \log_{\frac{1}{q}} \frac{mp}{\ln \ln n}$ *satisfies*

$$\sigma[T_j] = \mathcal{O}\left(\mathbb{E}[T_j]\frac{\ln n}{\sqrt{n}}\right).$$

The results of Propositions 1 and 2 are summarized in Fig. 1. The proofs of the propositions are given in Section 7.2.

3.2. Upper bound on the average number of irredundants

A subset X is irredundant if for all vertex $i \in X$, there exists a hyperedge H such that $H \cap X = \{i\}$. For fixed X of cardinality j, fixed i and H, the probability that $H \cap X = \{i\}$ is pq^{j-1}. For fixed X and H, the probability that $\nexists i \in X$ such that $H \cap X = \{i\}$ is equal to $1 - jpq^{j-1}$. If X is an irredundant then there exists a tuple $(k_1, \ldots, k_j)$ of positive value (i.e. at least 1) where k_i is the number of hyperedges H such that $H \cap X = \{i\}$ and the probability that X is an irredundant set is equal to:

$$\sum_{\substack{\forall i \le j,\ k_i \ge 1 \\ k_1 + \ldots + k_j = \ell \\ j \le \ell \le m}} \binom{m}{k_1, \ldots, k_j} (pq^{j-1})^{\ell} \cdot (1 - jpq^{j-1})^{m-\ell}$$

The next proposition states that the average number of irredundants is quasi-polynomial.

Proposition 3. *The average number of irredundants in a random model* $\mathbf{HG}(\mathbf{n}, \mathbf{m}, \mathbf{p})$ *is, for all* $\epsilon > 0$*, of order*

$$\mathcal{O}\left(\left((1+\epsilon)\frac{2pe^2}{q^2}\frac{mn}{\log_{\frac{1}{q}} nm}\right)^{\frac{1}{2}j_0}\right),$$

where

$$j_0 = \left\lceil \frac{1}{2}\log_{\frac{1}{q}} mn - \frac{1}{2}\log_{\frac{1}{q}}\log_{\frac{1}{q}} mn + \frac{1}{2}\log_{\frac{1}{q}} 2p \right\rceil$$

and $\lceil x \rceil$ *denotes the smallest integer greater or equal to* x *(ceiling).*

We did not succeed in performing a precise analysis of the average number of irredundants in function of their size. However, this result is sufficient to give an upper bound on the complexity of the MTMiner algorithm. The proof is rather technical and is given in Section 7.3.

3.3. Upper and lower bounds on the average number of minimal transversals

In the sequel, MT (resp. MT_j) is the random variable equal to the number of minimal transversals (resp. of size j). It is known that in the worst case, the number of minimal transversals may be exponential with respect to the size of the input hypergraph. However, for the "naive" uniform distribution on hypergraphs, the number of minimal transversal is almost surely at most linear in the size of the hypergraph. As far as we know, the $\mathbf{HG(n, m, p)}$ model leads to the first non-trivial bound on the average number of minimal transversals, as announced by the next theorem.

Theorem 1. *Consider the random model* $\mathbf{HG(n, m, p)}$ *with* $m = \beta n^{\alpha}$, $\beta > 0$ *and* $\alpha > 0$. *There exist a positive constant* $c := c(\alpha, \beta, p)$ *such that the average number of minimal transversals is*

$$\mathcal{O}\left(n^{d(\alpha)\log_{\frac{1}{q}} m + c\ln\ln m}\right),$$

with $d(\alpha) = 1$ *if* $\alpha \leq 1$ *and* $d(\alpha) = \frac{(\alpha+1)^2}{4\alpha}$ *otherwise.*

The proof can be found in Section 7.4. Once more, we did not succeed in studying the average number of minimal transversals in function of their size. Theorem 1 gives an upper bound on the average size of the output of the Transversal Hypergraph Generation problem. Using the Markov Inequality, it entails an almost sure upper bound on the size of the output. In order to prove that, in average, the Transversal Hypergraph Generation problem is output polynomial, we also need to control the size of the output and find a generic lower bound (i.e. a lower bound which is true with probability close to 1). A generic lower bound is often obtained by studying the moments of higher order or the variance. We did not succeed in studying the higher moments of the number of minimal transversals. However, we relate the number of minimal transversals to the number of transversals of a certain size and use the concentration property given in Proposition 2. We obtain the following proposition whose proof can be found in Section 7.5.

Proposition 4. *Consider* ϵ *with* $0 < \epsilon < 1$. *In the random model* $\mathbf{HG(n, m, p)}$, *the number MT of minimal transversals satisfies*

$$\mathbb{P}(MT < \epsilon\mathbb{E}[T_l]) = o(1)$$

where T_l *is the set of transversals of size* $l = \log_{\frac{1}{q}} \frac{mp}{\ln n} + 1$.

Inserting l in the expression of $\mathbb{E}[T_l] = \binom{n}{l}(1 - q^l)^m$ entails the following corollary.

Corollary 1. *In the random model* $\mathbf{HG(n, m, p)}$, *the number of minimal transversals is almost surely greater than*

$$= \Omega\left(\frac{1}{(\log_{\frac{1}{q}} m)^{3/2}\ln n}\left(\frac{e \cdot n}{\log_{\frac{1}{q}} \frac{mp}{\ln n} + 1}\right)^{\log_{\frac{1}{q}} \frac{mp}{\ln n}}\right)$$

4. Main results with the multiparametric model

The $HG(n, m, g)$ random model no longer considers that the vertices occur in a hyperedge with the same probability. Precisely, a vertex j belongs to a hyperedge with probability $p_j = g(j)$ $(q_j = 1 - g(j))$ with $g : \mathbb{N} \to [0, 1]$. Under this model, the role of the vertices may be different. Recall that we partition the set of vertices V into 3 subsets:

- The set U of ubiquitous vertices where for all $u \in U$, $q_u < \frac{x}{m}$ for some fixed constant x,
- the set R of rare events where for all $r \in R$, $p_r < 1 - e^{-\frac{1}{\ln n}}$,
- and the set O of other events, that is $O = V \setminus \{U \cup R\}$.

4.1. A lower bound on the average number of transversals

In the $HG(n, m, g)$ model, let μ the average value of the images of the function g. In other words:

$$\mu = \sum_{i=1}^{n} \frac{q_i}{n}$$

Proposition 5. *In the $HG(n,m,g)$ model, the average number of transversals is lower bounded by* $2^n - m(1+\mu)^n$.

Proof. The average number of transversals is given by the following formula:

$$\mathbb{E}[T] = \sum_{j=0}^{n} \sum_{\substack{X \subset V \\ |X|=j}} (1 - \prod_{i \in X} q_i)^m .$$

Using the Bernoulli inequality, we have:

$$\mathbb{E}[T] \geq \sum_{j=0}^{n} \sum_{\substack{X \subset V \\ |X|=j}} (1 - m \prod_{i \in X} q_i) \geq 2^n - m \sum_{j=0}^{n} \sum_{\substack{X \subset V \\ |X|=j}} \prod_{i \in X} q_i = 2^n - m \prod_{i=1}^{n} (1+q_i)$$

Then, according to the geometric inequality we have

$$\prod_{i=1}^{n} (1+q_i) \leq \left(\sum_{i=1}^{n} \frac{1+q_i}{n} \right)^n$$

and therefore

$$\mathbb{E}[T] \geq 2^n - m(1+\mu)^n$$

which concludes the proof. □

4.2. Upper bounds on the average number of irredundants

This sections gives various upper bounds on the average number of irredundants. Since a minimal transversal is irredundants, the next proposition also gives upper bounds on the number of minimal transversals.

Proposition 6. *In the $HG(n,m,g)$ model, we have the following bounds:*

1. *The average number of irredundants (and then, of minimal transversals) is bounded by* $\mathcal{O}(mn(1+\mu)^n)$.
2. *The average number of irredundants containing only vertices in $O \cup U$ is*

$$\mathcal{O}((nm \ln \sqrt{nm})^{\frac{1}{4} \ln n (\ln nm - 2 \ln \ln n - \ln \ln \sqrt{mn})})$$

3. *If $|R| = \mathcal{O}((\ln n)^c)$ where c is a constant, then the number of irredundants is quasi-polynomial.*
4. *The probability to have a polynomial number of irredundants containing ubiquitous vertices tends to* 1.

The proof can be found in Section 7.6. Result 1 in Proposition 6 is not precise but it has a certain advantage: starting from a complex model with a large number of parameters, we now have an estimation that relies on only 3 parameters. It can also be easily interpreted: if μ tends to 0 then almost all vertices appears in almost all hyperedges, hyperedges are all similar. In this case, there is few minimal transversals. If μ tends to 1 then almost each vertex appears in few hyperedges, hyperedges are all really different, that is to say the pairwise intersection of hyperedges is always small and there is an exponential number of minimal transversals.

The results 2 and 3 make the role of the rare events more precise in the average number of irredundants. Precisely if there is no rare event or if it is polylogarithmic in n, then number of irredundants is quasi-polynomial in n. Result 4 entails that ubiquitous vertices only have a polynomial role in the asymptotic number of irredundants.

4.3. On the average number of minimal transversals

Recall that M denotes the number of minimal transversals. Since the upper bounds obtained in Proposition 6 also hold for minimal transversals, we obtain the following theorem.

Theorem 2. *In the $HG(n,m,g)$ model, we have the following:*

- *If $|O \cup R| = \mathcal{O}(\ln n)$, then $\mathbb{E}[M]$ is at most polynomial.*
- *If $|R| = \mathcal{O}((\ln n)^c)$ where c is a constant, then $\mathbb{E}[M]$ is at most is quasi-polynomial.*
- *If $|R| = \Theta(n)$, then $\mathbb{E}[M]$ is at most exponential on $|R|$.*

The first point comes from Proposition 6 and the fact that the number of minimal transversals of a set of size $c \ln n$ is n^c.

Algorithm 1: The MTMiner-algorithm.

```
Data: a hypergraph $\mathcal{H}$ with $m$ hyperedges
Result: the minimal transversals of $\mathcal{H}$
$MT := \{\{v\} |\ v \in V,\ \{v\}$ is a transversal$\}$;
$N_1 := \{\{v\} |\ v \in V \setminus MT$, $v$ belongs to at least one hyperedge$\}$;
$j = 1$;
while $N_j \neq \emptyset$ do
    forall the prefix $V$ with $V \cup \{v_1\}, V \cup \{v_2\} \in N_j$ do
        $W = V \cup \{v_1\} \cup \{v_2\}$;
        if $W$ is irredundant then
            if $W$ is a transversal then
                add $W$ to $MT$;
            else
                add $W$ to $N_{j+1}$;
        $j = j + 1$;
return $MT$;
```

5. Algorithm analysis

In this section we study the average complexity of the MT-Miner algorithm. The worst-case input complexity of this algorithm is a polynomial in n times the number of irredundants. The analysis we perform and the results we obtain are also valid on any algorithm whose search space is bounded by the set of irredundants: Apriori, Dong–Li algorithm [4], Kavvadias–Stavropoulos [18], Uno–Murakami [20].

5.1. Average complexity of the MT-Miner algorithm

The MTMiner algorithm was described by Hébert, Bretto and Crémilleux in [16]. The algorithm computes all the minimal transversals of a given hypergraph using a levelwise strategy. Precisely at the jth level, the algorithm computes the *irredundants* formed with j vertices. Among the irredundants, some are minimal transversals and are stored in a data structure. The others are not minimal transversals but they might be part of one and they are used to build irredundants of size $j + 1$.

Each irredundant of size j can be extended in at most $n - j$ sets of size $j + 1$ and the minimality of each candidate set can be tested in polynomial time (w.r.t. the input size). MTMiner uses a prefix tree to optimize this generation step but even with the naive method (generate all the possible extensions), the complexity of MTMiner is $O(\text{Poly}(m, n)N)$ where N is the number of irredundants. (See Algorithm 1.)

A non-trivial upper bound on the average complexity of MTMiner follows from Proposition 3.

Proposition 7. *In the $HG(n, m, p)$ model, there exist some positive constant c such that the average complexity of* MTMiner *is*

$$\mathcal{O}\left((mn \log_{\frac{1}{q}} \sqrt{nm} \frac{p}{q^2})^{\frac{1}{4}(\log_{\frac{1}{q}} nm - \log_q p - \log_{\frac{1}{q}} \log_{\frac{1}{q}} \sqrt{nm}) + c} \right).$$

An equivalent result can be obtained in the $HG(n, m, g)$ model using Proposition 6 item 3.

Proposition 8. *In the $HG(n, m, g)$ model, if $|R| = \mathcal{O}((\ln n)^c)$ where c is a constant, then the average input-complexity of* MTMiner *is at most quasi-polynomial.*

Since MTMiner generates at least all the minimal transversals, its complexity is lower bounded by M. The generic lower bounds given by Proposition 5 and Corollary 1 entail a generic lower bound on the complexity of MTMiner.

5.2. Generic-case complexity of the THG-problem

The notion of generic-case complexity is defined in [17]. The idea is to study the worst-case complexity of an algorithm on a generic-subset of inputs. In a given random model, a subset E of inputs is said to be generic if the probability that a random input is in E tends to 1. The study of generic complexity is particularly interesting when no polynomial method is known to solve a problem in the general case (*NP*-complete problems, undecidable problems [21]), whereas there seem to be efficient methods in practice. The theoretical complexity of the THG-problem was discussed in the introduction. In particular, it is not known whether the problem is output-polynomial in the worst-case. The following theorem states that with probability that tends to 1 in the single-parameter model, the algorithm

MTMiner is output-polynomial. In other words, the set of inputs for which the algorithm is output-polynomial is a generic set.

Theorem 3. *Consider the random model* $\mathbf{HG}(\mathbf{n},\mathbf{m},\mathbf{p})$ *with* $m=\beta n^{\alpha}$, $\beta>0$ *and* $\alpha>0$. *Under this model, the generic complexity of the* THG*-problem is output-polynomial. Precisely, there exist an algorithm (*MTMiner*) such that for all* $\epsilon>0$, *the algorithm computes the minimal transversals of an input hypergraph in time* $M^{\epsilon+\frac{(\alpha+1)^2}{4\alpha}}$ *with probability asymptotically* 1 *and where* M *is the number of minimal transversals.*

Proof. To simplify the notations, we write $\gamma=\epsilon+\frac{(\alpha+1)^2}{4\alpha}$ and $a=\frac{1}{2}\mathbb{E}[T_j]$ with j as in Proposition 4. We have

$$\begin{aligned}
&\mathbb{P}(D>M^{\gamma})\\
&=\mathbb{P}([D>M^{\gamma}]\cap[M<a])+\mathbb{P}([D>M^{\gamma}]\cap[M\geq a])\\
&\leq\mathbb{P}(M<a)+\mathbb{P}(D\geq a^{\gamma})\\
&=O\left(\frac{\ln^2 mn}{n}\right)+\frac{\mathbb{E}[D]}{a^{\gamma}}
\end{aligned}$$

Alternative expressions for a^{γ} and $\mathbb{E}[D]$ are

$$\mathbb{E}[D]=n^{\frac{(\alpha+1)^2}{4\alpha}\log_{\frac{1}{q}} m+O(\ln\ln m)},$$

$$a^{\gamma}=n^{(\epsilon+\frac{(\alpha+1)^2}{4\alpha})\log_{\frac{1}{q}} m+O(\ln\ln m)}.$$

In particular, $\mathbb{E}[D]/a^{\gamma}$ is $O(n^{-\frac{\epsilon}{2}\log_{\frac{1}{q}} m})$ when $\gamma>\epsilon+\frac{(\alpha+1)^2}{4\alpha}$ and the $\mathcal{O}$-term tends to 0. This completes the proof. □

6. Comparison with real data

The following benchmark were made using datasets from the *Frequent Itemset Mining Dataset Repository*,[1] that are often used by the datamining community. The experimental results that we exhibit might therefore be well known by specialists. The objective of this section is to exhibit the links between real datasets and our probabilistic models.

We only present a selection of experimental results, that is already able to capture the diversity of hypergraphs in various contexts. It is important to note that for each example presented in this paper, one can find several datasets satisfying the same properties.

From Fig. 2 to Fig. 5, the histograms on the left side represent the number of hyperedges in which each vertex appears. Vertices are sorted according to the decreasing order of the number of hyperedges in which they appear. Histograms on the right side represents the size of hyperedges in each hypergraph.

- *mushroom.dat*: contains a few ubiquitous vertices and a few rare vertices. All the hyperedges have the same size. Those kinds of datasets validate the choice of most study to focus on k-uniform hypergraphs.
- *accident.dat*: contains a few ubiquitous vertices and a lot of rare vertices. The size of hyperedges seems to follow a Gaussian distribution.
- *pumsbstar.dat*: does not contain ubiquitous vertices and most vertices are rare.
- *T10I4D100K.dat*: all vertices are rare. Again, the size of hyperedges seems to follow a Gaussian distribution.

Our probability models seem to be more appropriate on the second and the fourth example, as the size of the hyperedges seems to follow a Gaussian distribution.

In our experiments, the generation of the minimal transversals were efficient on databases that have similar distribution to *mushroom.dat*. This result comforts us in the belief that the presence of rare events is the main parameter (by opposition to being just an important parameter amongst others) to decide whether the number of the minimal transversals is going to explode. Fig. 6 shows the distribution of the minimal transversals in *mushroom.dat*. As we can see, the number of minimal transversals $T_{min,j}$ of a given size j is maximal when $j=16$ and $T_{min,16}\sim 6\times 10^6$. Using the model $H(n,m,g)$ and Theorem 2, we could foretell that in cases *accidents.dat*, *pumsbstar.dat*, *T10I4D100K.dat* the number of minimal transversals

[1] http://fimi.ua.ac.be/data/.

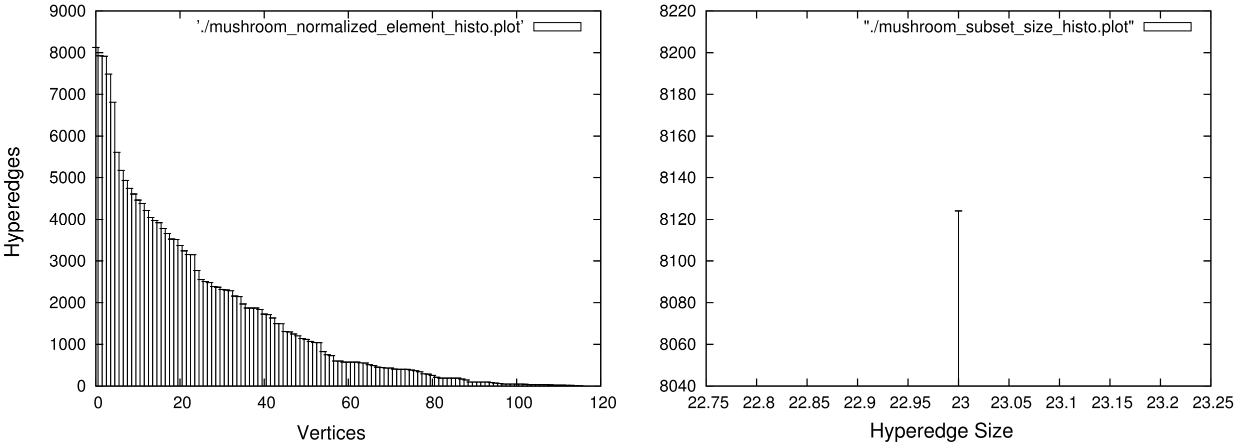

Fig. 2. *mushroom.dat* (119 vertices, 8124 hyperedges).

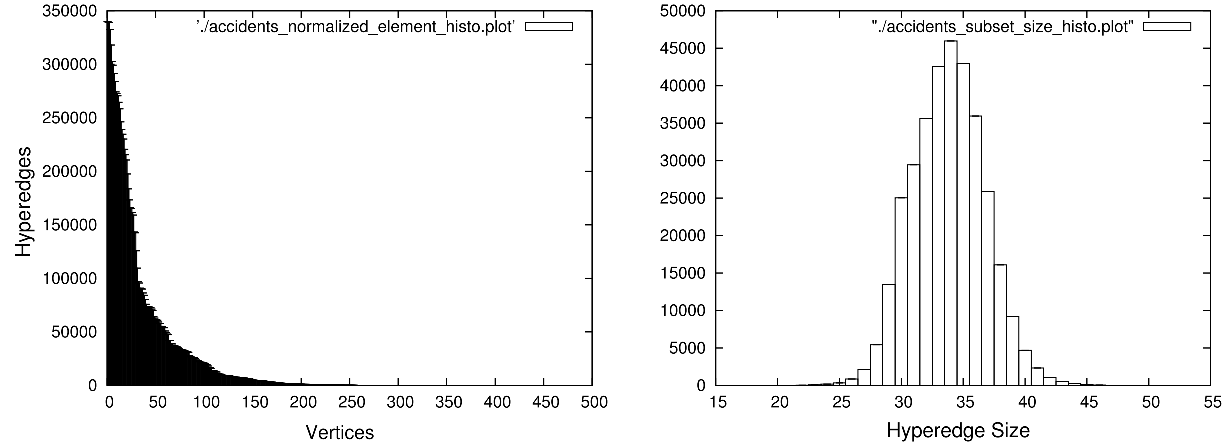

Fig. 3. *accidents.dat* (468 vertices, 340 183 hyperedges).

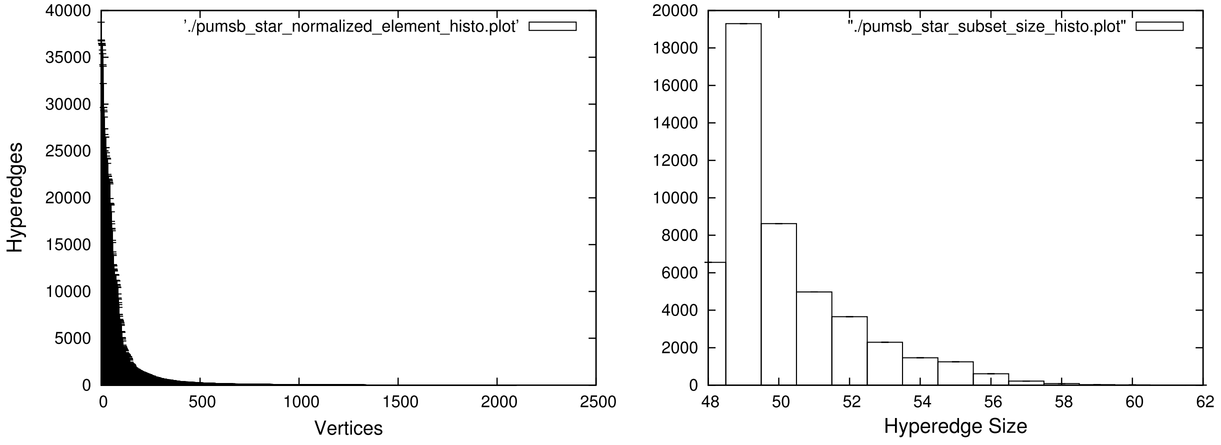

Fig. 4. *pumsbstar.dat* (7116 vertices, 49 046 hyperedges).

explodes, which seems indeed to be the case. Even after a long time execution (more than a week) on a regular computer, only a small proportion of the search space had been visited, whereas the number of minimal transversals was tremendously huge. We also tried to write the minimal transversals in a file: within a day, the program stopped because our 500 Go hard drive was full.

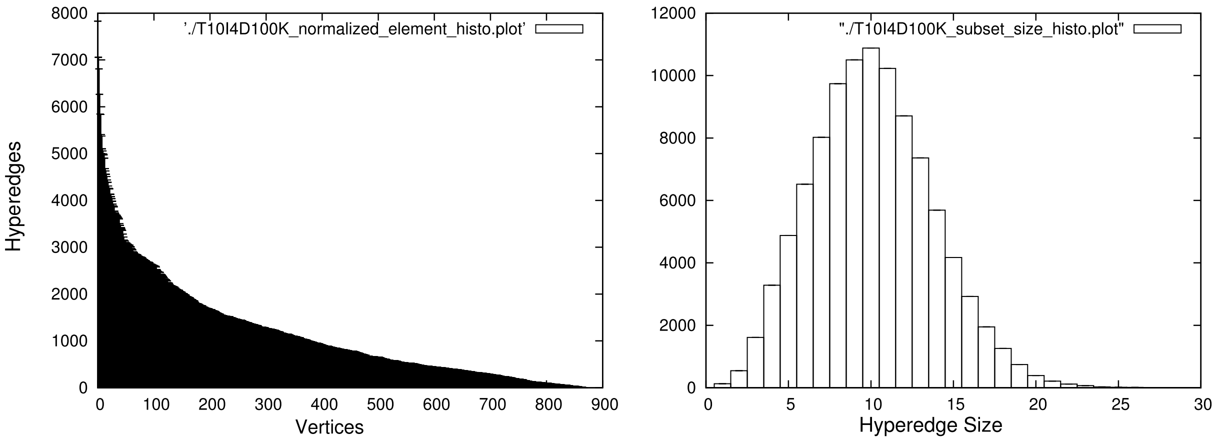

Fig. 5. *T10I4D100K.dat* (999 vertices, 100 000 hyperedges).

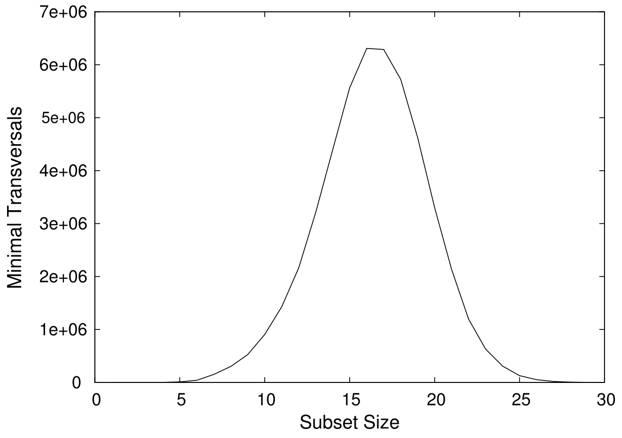

Fig. 6. Minimal transversals of *mushroom.dat*.

7. Proofs

7.1. Proof of Proposition 1

Proposition 1 describes the evolution of the number of transversals with respect to their size. Recall that T_j denotes the number of transversals of size j, m is the number of hyperedges and $q = 1 - p$ is the probability that a vertex does not appear in a hyperedge. The probability for a subset X of size j to be a transversal is

$$\mathbb{P}(X \text{ is a transversal}) = (1 - q^j)^m$$

so that the average number of transversals of size j satisfies

$$\mathbb{E}[T_j] = \binom{n}{j}(1 - q^j)^m = \binom{n}{j}\left(1 - \frac{1}{mx}\right)^m$$

with $j = \log_{\frac{1}{q}} m + \log_{\frac{1}{q}} x$ as in the proposition.

Case 1. We have $j \le j_{min}$ with

$$j_{min} = \log_{\frac{1}{q}} m - \log_{\frac{1}{q}} \log_{\frac{1}{q}} (m \ln n)\,.$$

The mean $\mathbb{E}[T_j]$ satisfies the bound

$$\mathbb{E}[T_j] \le n^j e^{-mq^j} \le e^{j_{min} \ln n - mq^{j_{min}}}\,.$$

Now, the expression in the exponent simplifies into

$$j_{min} \ln n - mq^{j_{min}} = -\ln n \times (\log_{\frac{1}{q}} \log_{\frac{1}{q}} m + \log_{\frac{1}{q}} \ln n)$$

which tends to $-\infty$ and completes the proof of Case 1.

Case 2. If x tends to 0, we have

$$\left(1 - \frac{1}{mx}\right)^m \le \exp\left(-\frac{m}{mx}\right) = \exp\left(-\frac{1}{x}\right)$$

which tends to 0. Then, $\mathbb{E}[T_j] = o\left(\binom{n}{j}\right)$.

Case 3. If $x = \Theta(1)$, we have the equivalence

$$\left(1 - \frac{1}{mx}\right)^m = \exp\left(-\frac{m}{mx} + O\left(\frac{1}{mx}\right)\right) \sim \exp\left(-\frac{1}{x}\right)$$

and $\mathbb{E}[T_j] \sim \binom{n}{j} \exp\left(-\frac{1}{x}\right)$.

Case 4. If x tends to $+\infty$, the previous equivalence remains true and $e^{-1/x} \sim 1 - (1/x)$. To conclude if x tends to 0, we have $\mathbb{E}[T_j] \sim \binom{n}{j}\left(1 - \frac{1}{x}\right)$.

7.2. *Proof of Proposition 2*

We want to prove that in the **HG**(**n**, **m**, **p**) random model, the standard deviation of the number of transversals of size j satisfies

$$\sigma[T_j] = \mathcal{O}\left(\mathbb{E}[T_j] \frac{\ln n}{\sqrt{n}}\right)$$

as soon as with $j > \log_{\frac{1}{q}} \frac{mp}{\ln \ln n}$.

For any set of vertices X, χ_X is an indicator random variable which equals to 1 if X is a transversal and 0 otherwise. Consider $j \ge 1$. The variance of T_j is by definition

$$\mathbb{V}(T_j) = \sum_{\substack{X, Y \subset V \\ |X| = |Y| = j}} \mathbb{P}[\chi_X \chi_Y = 1] - \mathbb{P}[\chi_X = 1]\mathbb{P}[\chi_Y = 1].$$

If X and Y are disjoint, the random variables χ_X and χ_Y are independent and the associated term in the previous sum is zero. We are then led to study non-disjoint subsets of vertices. If X and Y are non-disjoint, we define I, J and K as

$$K = X \cap Y, \quad I = X \backslash Y, \quad J = Y \backslash X.$$

Note that if X and Y have the same cardinality, then the same is true for I and J. Therefore we note $|K| = k$ and $|I| = |J| = j - k$. The support of K, noted $\mathcal{E}' \subseteq \mathcal{E}$, is the set of hyperedges that intersect K. If X and Y are transversals, then each of I and J intersect all the hyperedges of $\mathcal{E} \backslash \mathcal{E}'$. For a fixed set $\mathcal{E}'$, the probability to be K's support is $q^{k \cdot (m - |\mathcal{E}'|)}(1 - q^k)^{|\mathcal{E}'|}$. The probability that I (resp. J) intersects all the hyperedges of $\mathcal{E} \backslash \mathcal{E}'$ is $(1 - q^{j-k})^{m - |\mathcal{E}'|}$. Summing over all the possible cardinalities for $\mathcal{E}'$, we obtain

$$\begin{aligned} \mathbb{P}[\chi_X = \chi_Y = 1] &= \sum_{\ell=0}^{m} \binom{m}{\ell} (1 - q^k)^\ell q^{k(m-\ell)} (1 - q^{j-k})^{2(m-\ell)} \\ &= (1 - 2q^j + q^{2j-k})^m \end{aligned}$$

For fixed j and k, there are $\binom{n}{k, j-k, j-k}$ possible choices for the sets I, J and K. The probability that X (or Y) is a transversal is $(1 - q^j)^m$, and summing over all the possible k, we obtain

$$\mathbb{V}(T_j) = \sum_{k=1}^{j} \binom{n}{k, j-k, j-k} \left[(1 - 2q^j + q^{2j-k})^m - (1 - q^j)^{2m}\right].$$

Various cases are now possible.

Case (i). $j > \log_{\frac{1}{q}} mn$. Then

$$(1-2q^j+q^{2j-k})^m - (1-q^j)^{2m} = \mathcal{O}\left(\frac{1}{n}\right)$$

where the constant term in $\mathcal{O}$ only depends on n and not on j. In addition,

$$\sum_{k=0}^{j} \binom{n}{k,\, j-k,\, j-k} = \binom{n}{j}^2$$

so that, using Proposition 1 with $j > \log_{\frac{1}{q}} mn$, the variance satisfies

$$\mathbb{V}(T_j) = \mathcal{O}\left(\frac{1}{n}\binom{n}{j}^2\right) = \mathcal{O}\left(\frac{1}{n}E[T_j]^2\right)$$

and the random variable T_j is concentrated.[2]

Case (ii). If $\log_{\frac{1}{q}} \frac{mp}{\ln\ln n} \le j \le \log_{\frac{1}{q}} mn$. The variance satisfies the upper bounds

$$\begin{aligned}\mathbb{V}(T_j) &\le \sum_{k=1}^{j} \binom{n}{k,\, j-k,\, j-k} (1-2q^j+q^{2j-k})^m \\ &\le \sum_{k=1}^{j} \binom{n}{k,\, j-k,\, j-k} \exp(-2mq^j + mq^{2j-k}).\end{aligned}$$

Let α_k denotes the k-th term of this sum. The ratio of two consecutive α_k is upper bounded by

$$\begin{aligned}\frac{\alpha_{k+1}}{\alpha_k} &\le \frac{j^2}{2(n-2j+2)} e^{mpq^j} \le \frac{\log^2 mn}{2(n-2\log mn+2)} e^{mpq^{\log_{\frac{1}{q}}\left(\frac{mp}{\ln\ln n}\right)}} \\ &\le \frac{e^{\ln\ln n}\log^2 mn}{2(n-2\log^2 mn+2)} = \mathcal{O}\left(\frac{\ln^2 n}{n}\right)\end{aligned}$$

The variance satisfies $\mathbb{V}(T_j) \sim \alpha_1$ since $j\alpha_2 = o(\alpha_1)$, for $\log_{\frac{1}{q}} \frac{mp}{\ln\ln n} \le j \le \log_{\frac{1}{q}} mn$. Now, we have

$$\mathbb{V}(T_j) \sim e^{-2mq^j} \frac{n^{2j-1}}{(j-1)!^2} \sim \frac{j^2}{n} \mathrm{E}[T_j]^2.$$

The result follows for case (ii) and the proof is complete.

7.3. Proof of Proposition 3

We study a function that bounds the number of irredundants by considering the following necessary condition: let $X = \{x_1, \ldots, x_j\}$, a *selection* is a set $\{E_1, \ldots, E_j\}$ of hyperedges such that $E_i \cap X = \{x_i\}$ for all $i \le k$ (therefore $j \le min(m,n)$). X is an irredundant set if and only if there exists a selection in the hypergraph.

For each of the $\binom{n}{j}$ subsets of size j, if one can find j hyperedges amongst m such that each hyperedge contains one vertex and not the others (the order on hyperedges is not specified), then the condition is satisfied. This occurs with probability $\left(pq^{j-1}\right)^j$. The average number of irredundants of size j is therefore bounded by the function

$$f(j) = \binom{n}{j} \frac{m!}{(m-j)!} \left(pq^{j-1}\right)^j.$$

In order to exhibit the value j for which $f(j)$ is maximal, we study the ratio

$$\frac{f(j+1)}{f(j)} = \frac{(m-j)(n-j)}{j+1} pq^{2j}.$$

It is decreasing with j and then, the maximum of f is given by the smallest integer j such that the ratio is smaller than 1. Now, consider j_0 of the form

[2] That is, the standard deviation is negligible compared to the mean.

$$j_0 = \frac{1}{2}\log_{\frac{1}{q}} mn + \frac{1}{2}\log_{\frac{1}{q}} \frac{2px}{\log_{\frac{1}{q}} mn}$$

with $x \in [1, \frac{1}{q}[$ the smallest real such that j_0 is an integer. Simple asymptotic computations give

$$\begin{aligned}\frac{f(j_0+1)}{f(j_0)} &= \frac{(m-j_0)(n-j_0)}{j_0+1} pq^{2j_0} \\ &\sim \frac{2mn}{\log_{\frac{1}{q}} mn} p \frac{\log_{\frac{1}{q}} mn}{2pxmn} \\ &= \frac{1}{x}.\end{aligned}$$

In the same way, we obtain

$$\frac{f(j_0)}{f(j_0-1)} \sim \frac{1}{q}x > 1.$$

This implies that $f(j)$ is maximal when $j = j_0$. In addition, for $j = j_0 + j'$ with $|j'| < \alpha_n$ and $\alpha_n = o(j_0)$, the ratio $r(j) = f(j+1)/f(j)$ decreases geometrically by a factor q^2 as j increases ($r(j+1)/r(j) \sim q^2$) so that the number of irredundants is bounded by

$$\sum_{j=1}^{\min(m,n)} f(j) \sim \sum_{j=-\alpha_n}^{\alpha_n} f(j_0 + j) = \mathcal{O}\left(\frac{1}{1-q^2} f(j_0)\right)$$

with $\alpha_n = o(j_0)$. Since $p > 1 - e^{-1/\ln n}$, the number of irredundants is also of order of order $\mathcal{O}(f(j_0)\ln n)$. The Stirling formula combined with the insertion of j_0 into the expression of $f(j)$ give the next equivalence for $f(j_0)$,

$$f(j_0) \sim \frac{n^{j_0} m^{j_0}}{j_0!}\left(\frac{p}{q} q^{j_0}\right)^{j_0} = \frac{1}{\sqrt{2\pi j_0}}\left(\frac{e}{j_0 q}\sqrt{\frac{mnp\log_{\frac{1}{q}} mn}{2x}}\right)^{j_0}$$

Writing $j_0 = (1-\epsilon)\frac{1}{2}\log_{\frac{1}{q}} mn$ with $\epsilon = \frac{\log_{\frac{1}{q}}\log_{\frac{1}{q}} mn - \log_{\frac{1}{q}} 2px}{\log_{\frac{1}{q}} mn}$, which tends to 0, is sufficient to obtain the announced result.

7.4. Proof of Theorem 1

In the sequel, MT (resp. MT_j) is the random variable equal to the number of minimal transversals (resp. of size j). We consider once again a function that bounds the number of minimal transversals. A set X is a minimal transversal if and only if:

1. there exist a selection $\mathcal{E}'$ for X,
2. for all $F \in \mathcal{E}\backslash\mathcal{E}'$, $|F \cap X| \geq 1$.

The probability that Condition 1 holds is $(pq^{j-1})^j$ whereas the probability that Condition 2 holds is $(1-q^j)^{m-j}$. There are $\binom{n}{j}$ sets of vertices of size j and each of them have $\frac{m!}{(m-j)!}$ possible selections. Then, the average number of minimal transversals of size j is bounded by $h(j)$ with

$$h(j) = \binom{n}{j}\frac{m!}{(m-j)!}(pq^{j-1})^j(1-q^j)^{m-j}.$$

We now determine the j that maximizes $h(j)$. The ratio $\frac{h(j+1)}{h(j)}$ satisfies

$$\frac{h(j+1)}{h(j)} = \frac{(m-j)(n-j)}{j+1}\frac{p}{1-q^j}q^{2j}\left(1+\frac{pq^j}{1-q^j}\right)^{m-j-1}$$

As in the previous section, this ratio is also decreasing. The maximum of $h(j)$ is given when the ratio equals to 1. Two cases are now possible according to the value of α.
Case $\alpha \leq 1$. Suppose $j = j(x)$ is of the form

$$j(x) = \frac{1}{2}\log_{\frac{1}{q}} \frac{2pmnx}{\log_{\frac{1}{q}} mn}$$

and $x = \Theta(1)$. Then, the next relations hold (we omit the variable x):

$$q^{2j} = \frac{\log_{\frac{1}{q}} mn}{2pmnx} \quad \text{and} \quad j \sim \frac{1}{2}\log_{\frac{1}{q}} mn$$

(here, the hypotheses on p and then q are important). Inserting these relations in the ratio $h(j+1)/h(j)$ and using equivalences, we obtain

$$\frac{h(j+1)}{h(j)} \sim \frac{2mn}{\log_{\frac{1}{q}} mn} p \frac{\log_{\frac{1}{q}} mn}{2pxmn} \exp\left(mp\sqrt{\frac{\log_{\frac{1}{q}} mn}{2pxmn}}\right).$$

Since $\alpha < 1$, the term in the exponential tends to 0 so that $\frac{h(j+1)}{h(j)} \sim \frac{1}{x}$. Then, the maximum of h is given for $j = \lfloor j(1) \rfloor$ and the average number of minimal transversals satisfies

$$\mathbb{E}[M] \le \sum_{j=1}^{n} h(j) = O(nh(\lfloor j(1) \rfloor)).$$

But an equivalence of $h(j)$ around $j = \lfloor j(1) \rfloor$ is

$$h(j) \sim \frac{n^j m^j}{j!}\left(\frac{p}{q}q^j\right)^j = \exp\left(j \ln mnq^j + O\left(j \ln j + j \ln \frac{p}{q}\right)\right).$$

But,

$$j = \frac{1+\alpha}{2\alpha}\log_{\frac{1}{q}} m + O\left(\log_{\frac{1}{q}} \ln n\right) \qquad \text{and}$$

$$\ln mnq^j = \frac{1+\alpha}{2}\ln n + O\left(\ln \ln n\right)$$

and the product gives

$$h(j) = \exp\left(\frac{(1+\alpha)^2}{4\alpha}(\ln n)\log_{\frac{1}{q}} m + O\left((\ln\ln n)\log_{\frac{1}{q}} n\right)\right).$$

This completes the proof when $\alpha < 1$.

Case $\alpha > 1$. Suppose $j = j(x)$ is of the form

$$j(x) = \log_{\frac{1}{q}} \frac{pm}{x \ln \frac{m}{n}}.$$

Then,

$$q^j = \frac{x \ln \frac{m}{n}}{mp} \quad \text{and} \quad j \sim \log_{\frac{1}{q}} m$$

(once more, the hypotheses on p and then q are important). When we insert these relations into the ratio $h(j+1)/h(j)$, we obtain

$$\begin{aligned}\frac{h(j+1)}{h(j)} &\sim \frac{mn}{\log_{\frac{1}{q}} m} p \frac{x^2\left(\ln\frac{m}{n}\right)^2}{m^2p^2}\exp\left(mp\frac{x\ln\frac{m}{n}}{mp}\right)\\ &\sim \frac{x^2\left(\ln\frac{m}{n}\right)^2}{p\log_{\frac{1}{q}} m}\left(\frac{m}{n}\right)^{x-1}.\end{aligned}$$

Then, the maximum of h is given for x close to 1 (when $j = \lfloor j(1) \rfloor$) so that the average number of minimal transversals satisfies

$$\mathbb{E}[M] \le \sum_{j=1}^{n} h(j) = O(nh(\lfloor j(1) \rfloor)).$$

Around $j = \lfloor j(1) \rfloor$, $h(j)$ verifies

$$h(j) \sim \frac{n^j m^j}{j!}\left(\frac{p}{q}q^j\right)^j \exp(-mq^j) = \exp\left(j \ln mnq^j + O\left(j\ln j + j\ln\frac{p}{q} + \frac{1}{p}\ln n\right)\right).$$

Now, the next relations hold:

$$j = \log_{\frac{1}{q}} m + O\left(\log_{\frac{1}{q}} \ln n\right) \qquad \text{and}$$

$$\ln mnq^j = \ln n + O\left(\ln\ln n\right)$$

and the product gives

$$h(j) = \exp\left((\ln n)\log_{\frac{1}{q}} m + O\left((\ln\ln n)\log_{\frac{1}{q}} n\right)\right).$$

This completes the proof when $\alpha > 1$.

Case $\alpha = 1$. Suppose $j = j(x)$ is of the form

$$j(x) = \log_{\frac{1}{q}} \frac{pm}{x \ln \log_{\frac{1}{q}} m}.$$

Then,

$$q^j = \frac{x \ln \log_{\frac{1}{q}} m}{mp} \quad \text{and} \quad j \sim \log_{\frac{1}{q}} m$$

(again, the hypotheses on p and then q are important) and the ratio $h(j+1)/h(j)$ satisfies

$$\begin{aligned}\frac{h(j+1)}{h(j)} &\sim \frac{mn}{\log_{\frac{1}{q}} m} p \frac{x^2\left(\ln\log_{\frac{1}{q}} m\right)^2}{m^2p^2} \exp\left(mp\frac{x\ln\log_{\frac{1}{q}} m}{mp}\right) \\ &\sim \frac{x^2\left(\ln\log_{\frac{1}{q}} m\right)^2}{\beta p}\left(\log_{\frac{1}{q}} m\right)^{x-1}.\end{aligned}$$

Then, the maximum of h is given for x close to 1 (when $j = \lfloor j(1) \rfloor$) so that the average number of minimal transversals satisfies

$$\mathbb{E}[M] \le \sum_{j=1}^{n} h(j) = O\left(nh(\lfloor j(1) \rfloor)\right).$$

Around $j = \lfloor j(1) \rfloor$, $h(j)$ verifies

$$h(j) \sim \frac{n^j m^j}{j!}\left(\frac{p}{q}q^j\right)^j \exp(-mq^j) = \exp\left(j \ln mnq^j + O\left(j\ln j + j\ln\frac{p}{q} + \frac{1}{p}\ln\log_{\frac{1}{q}} n\right)\right).$$

Now, the next relations hold:

$$j = \log_{\frac{1}{q}} m + O\left(\log_{\frac{1}{q}} \ln n\right) \qquad \text{and}$$

$$\ln mnq^j = \ln n + O\left(\ln\ln n\right)$$

and the product gives

$$h(j) = \exp\left((\ln n)\log_{\frac{1}{q}} m + O\left((\ln\ln n)\log_{\frac{1}{q}} n\right)\right).$$

This completes the proof when $\alpha = 1$.

7.5. Proof of Proposition 4 and Corollary 1

Proof of Proposition 4. Recall that MT (resp. MT_j) is the random variable equal to the number of minimal transversals (resp. of size j). MT is lower bounded by M_j, the number of minimal transversals of size j. Among the T_j transversals of cardinality j, MT_j are irredundant and $T_j - MT_j$ are supersets of transversals of cardinality $j-1$. By definition, there are T_{j-1} transversals with cardinal $j-1$ and each of these transversals can be completed in at most $n-j+1$ transversals of cardinal j. We deduce the inequalities

$$MT \geq MT_j \qquad \text{and} \qquad T_j - (n-j+1)T_{j-1} \leq MT_j \leq T_j. \tag{1}$$

The lower bound entails that for all $0 < \epsilon < 1$,

$$\begin{aligned}\mathbb{P}(MT_j \leq \epsilon\mathbb{E}[T_j]) &\leq \mathbb{P}(T_j - (n-j+1)T_{j-1} \leq \epsilon\mathbb{E}[T_j]) \\ &= \mathbb{P}(T_j \leq \epsilon\mathbb{E}[T_j] + (n-j+1)T_{j-1}) \\ &\leq \mathbb{P}\left(T_j < \epsilon(\mathbb{E}[T_j] + (n-j+1)\mathbb{E}[T_{j-1}])\right) + \mathbb{P}(T_{j-1} < \epsilon\mathbb{E}[T_{j-1}])\end{aligned}$$

Consider $l = \log_{\frac{1}{q}} \frac{mp}{\ln n} + 1$. We have

$$(n-l+1)\frac{\mathbb{E}[T_{l-1}]}{\mathbb{E}[T_l]} = l\left(1 - \frac{pq^{l-1}}{1-q^l}\right)^m.$$

Since $\left(1 - \frac{pq^{l-1}}{1-q^l}\right)^m$ is upper bounded by $e^{-mpq^{l-1}}$ $(1-q^l$ tends to $1)$ then

$$\mathbb{E}[T_l] + (n-l+1)\mathbb{E}[T_{l-1}] = \mathbb{E}[T_l]\left(1 + O\left(\frac{\ln m}{n}\right)\right)$$

the Chernoff inequality implies that

$$\mathbb{P}(MT_l < \epsilon\mathbb{E}[T_l]) \leq e^{-\frac{\mathbb{E}[T_{l-1}](1-\epsilon)^2}{2}} + e^{-\frac{\mathbb{E}[T_l](1-(\epsilon+\epsilon\mathcal{O}(\frac{\ln m}{n})))^2}{2}}$$

Since MT is lower bounded by MT_l, this completes the proof of Proposition 4.

Proof of Corollary 1. The lower bound comes from $\mathbb{E}[T_l]$ with $l = \log_{\frac{1}{q}} \frac{mp}{\ln n} + 1$.

$$\mathbb{E}[T_l] = \binom{n}{l}(1-q^l)^m \geq \frac{n^{l-1}}{l!}e^{-q^l m}$$

According to the Stirling formula we have

$$\sim \frac{e}{\sqrt{2\pi}l^{3/2}}\left(\frac{e\cdot n}{l}\right)^{l-1} e^{-q^l m}$$

Then if we replace l by $\log_{\frac{1}{q}} \frac{mp}{\ln n} + 1$

$$\sim \frac{e}{\sqrt{2\pi}(\log_{\frac{1}{q}} \frac{mp}{\ln n} + 1)^{3/2}} \cdot \left(\frac{e\cdot n}{\log_{\frac{1}{q}} \frac{mp}{\ln n} + 1}\right)^{\log_{\frac{1}{q}} \frac{mp}{\ln n}} \cdot \frac{1}{\ln n e^{\frac{q}{p}}}$$

$$= \Omega\left(\frac{1}{(\log_{\frac{1}{q}} m)^{3/2}\ln n}\left(\frac{e\cdot n}{\log_{\frac{1}{q}} \frac{mp}{\ln n} + 1}\right)^{\log_{\frac{1}{q}} \frac{mp}{\ln n}}\right)$$

7.6. *Proof of* Proposition 6

Recall that Proposition 6 gives various bounds on the average number of irredundants.

Proof of 1. From Proposition 5 we obtain an upper bound on the average number of subset that are not transversals, that is to say $\mathcal{O}(m(1+\mu)^n)$. In the worst case, all those sets are irredundant. Notice furthermore that any minimal transversal might be obtained from some of those sets by adding a vertex. Therefore, we multiply by n in order to obtain an upper bound.

Proof of 2. In order to obtain an upper bound, we use the result in Proposition 3. In the $HG(n, m, p)$ model, the upper bound on the average number of irredundants decreases as p increases. Recall that for all vertices $a \in O \cup U$, we have $p_a \geq 1 - e^{-\frac{1}{ln}}$. The average number of irredundants in the $HG(n, m, 1 - e^{-\frac{1}{ln}})$ model is an upper bound of the average number of irredundants containing only vertices in O. If $q = e^{-\frac{1}{lnn}}$, for all y we have $\log_{1/q} y = \ln y \times \ln n$. Using this simplification, we obtain the announced upper bound.

Proof of 3. Let $Irr_{j,l}$ be the number of irredundants of size j containing exactly l rare vertices. We have:

$$Irr_{j,l} \leq Irr_{j-l,0} \times Irr_{l,l}$$

where $Irr_{j-l,0}$ is exactly the number of irredundants containing only vertices in $O \cup U$ and $Irr_{l,l}$ is the number of irredundants containing only vertices in R. Note that point 2 gives an upper bound on $Irr_{j-l,0}$. We obtain a bound on $Irr_{l,l}$ by adapting the function $f(j)$ from Section 3.2. Since for all vertices $r \in R$, we have $p < 1 - e^{-\frac{1}{\ln n}} < \frac{1}{\ln n}$, we have:

$$Irr_{j,l} \leq \binom{|R|}{l} \frac{m!}{(m-l)!} \left(\frac{1}{\ln n}\right)^l Irr_{j-l,0}$$

Now we obtain an upper bound on the average number of irredundant in the $HG(n, m, g)$ model.

$$Irr_j \leq \sum_{l=0}^{min\{j,|R|\}} \binom{|R|}{l} \left(\frac{m}{\ln n}\right)^l Irr_{j-l,0}$$

From point 2., we know that $Irr_{j-l,0}$ is at most quasi-polynomial. Since $l = \mathcal{O}(|R|) = \mathcal{O}((\ln n)^c)$, it is clear that $\binom{|R|}{l}\left(\frac{m}{\ln n}\right)^l$ is also quasi-polynomial.

Proof of 4. Let $Irr(\mathcal{H})$ denote the set of irredundants of a hypergraph $\mathcal{H}$. For a fixed vertex $e \in V$, let $Irr(\mathcal{H}, e)$ the set of irredundants containing e. Let $N(e)$ be the set of hyperedges not containing the vertex e. We have:

$$Irr(H, e) \subseteq \{X \cup \{e\} \mid X \in Irr(N(e))\}.$$

Hence we have:

$$|Irr(H, e)| \leq |Irr(N(e))|,$$

that is to say the number of irredundants containing a given vertex e is bounded by the number of irredundants of the hypergraph reduced to the hyperedges that do not contain e. The number of hyperedges that do not contain a ubiquitous vertex is bounded by x and using Poisson paradigm we know that in a random hypergraph a ubiquitous vertex is in at least $m - x\sqrt{x}$ hyperedges with probability tending to 1. The number of irredundants of a hypergraph with at most $x\sqrt{x}$ hyperedges is polynomial since x is a constant.

8. Conclusion

The models we have studied already give a partial information on the average number of minimal transversals in real context and on the average complexity of the algorithms. Indeed, our models predict the order of growth of the size of minimal transversals. Hence, we are able to tell whether the computation can be made in reasonable time and space. Though, the upper bounds we have obtained on the number of minimal transversals still seems too large compared to real-data examples.

References

[1] Dimitris Achlioptas, Cristopher Moore, On the 2-colorability of random hypergraphs, in: Proc. 6th RANDOM, Springer-Verlag, 2002, pp. 78–90.

[2] Peter Damaschke, Parameterized enumeration, transversals, and imperfect phylogeny reconstruction, Theoret. Comput. Sci. 351 (3) (2006) 337–350.

[3] Élie de Panafieu, Phase transition of random non-uniform hypergraphs, in: 24th International Workshop on Combinatorial Algorithms, IWOCA 2013, J. Discrete Algorithms 31 (2015) 26–39.

[4] Guozhu Dong, Jinyan Li, Mining border descriptions of emerging patterns from dataset pairs, Knowl. Inf. Syst. 8 (2005) 178–202.

[5] Andrzej Dudek, Alan Frieze, Loose Hamilton cycles in random k-uniform hypergraphs, Electron. J. Combin. (2010) 45.

[6] Thomas Eiter, Georg Gottlob, Identifying the minimal transversals of a hypergraph and related problems, SIAM J. Comput. 24 (6) (1995) 1278–1304.

[7] Thomas Eiter, Georg Gottlob, Hypergraph transversal computation and related problems in logic and AI, in: Sergio Flesca, Sergio Greco, Nicola Leone, Giovambattista Ianni (Eds.), JELIA, in: Lecture Notes in Computer Science, vol. 2424, Springer, 2002, pp. 549–564.

[8] Thomas Eiter, Georg Gottlob, Kazuhisa Makino, New results on monotone dualization and generating hypergraph transversals, SIAM J. Comput. 32 (2) (2003) 514–537.

[9] Thomas Eiter, Kazuhisa Makino, Georg Gottlob, Computational aspects of monotone dualization: a brief survey, Discrete Appl. Math. 156 (11) (2008) 2035–2049.

[10] P. Erdös, A. Rényi, On random graphs. I, Publ. Math. Debrecen 6 (1959) 290–297.

[11] P. Erdös, A. Rényi, On the evolution of random graphs, in: Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 1960, pp. 17–61.

[12] Philippe Flajolet, Robert Sedgewick, Analytic Combinatorics, 1st edition, Cambridge University Press, New York, NY, USA, 2009.

[13] Michael L. Fredman, Leonid Khachiyan, On the complexity of dualization of monotone disjunctive normal forms, J. Algorithms 21 (3) (1996) 618–628.

[14] Dimitrios Gunopulos, Roni Khardon, Heikki Mannila, Hannu Toivonen, Data mining, hypergraph transversals, and machine learning, in: PODS, ACM Press, 1997, pp. 209–216.

[15] Matthias Hagen, Algorithmic and computational complexity issues of MONET, Dissertation, Institut für Informatik, Friedrich-Schiller-Universität Jena, December 2008.

[16] Céline Hébert, Alain Bretto, Bruno Crémilleux, A data mining formalization to improve hypergraph minimal transversal computation, Fund. Inform. 80 (December 2007) 415–433.

[17] Ilya Kapovich, Alexei Myasnikov, Paul Schupp, Vladimir Shpilrain, Generic-case complexity, decision problems in group theory and random walks, J. Algebra 264 (2003) 665–694.

[18] Dimitris J. Kavvadias, Elias C. Stavropoulos, An efficient algorithm for the transversal hypergraph generation, J. Graph Algorithms Appl. 9 (2) (2005) 239–264.

[19] M. Lelarge, A new approach to the orientation of random hypergraphs, in: Proceedings of the Twenty-Third Annual ACM–SIAM Symposium on Discrete Algorithms, SODA '12, SIAM, 2012, pp. 251–264.

[20] Keisuke Murakami, Takeaki Uno, Efficient algorithms for dualizing large-scale hypergraphs, Discrete Appl. Math. 70 (2014) 83–94.

[21] Alexei G. Myasnikov, Alexander N. Rybalov, Generic complexity of undecidable problems, J. Symbolic Logic 73 (2) (2008) 656–673.

[22] Vlady Ravelomanana, Alphonse Laza Rijamamy, Creation and growth of components in a random hypergraph process, in: Danny Z. Chen, D.T. Lee (Eds.), Computing and Combinatorics, in: Lecture Notes in Computer Science, vol. 4112, Springer, Berlin, Heidelberg, 2006, pp. 350–359.

[23] S. Sarkar, K.N. Sivarajan, Hypergraph models for cellular mobile communication systems, IEEE Trans. Veh. Technol. 47 (2) (1998) 460–471.

[24] Ilya Shmulevich, Aleksey D. Korshunov, Jaakko Astola, Almost all monotone boolean functions are polynomially learnable using membership queries, Inform. Process. Lett. 79 (5) (September 2001) 211–213.