Average Size of Implicational Bases

Giacomo Kahn[1] and Alexandre Bazin[2]

[1] LIMOS & Université Clermont Auvergne, France
[2] Le2i - Laboratoire Electronique, Informatique et Image, France
giacomo.kahn@isima.fr, contact@alexandrebazin.com

Abstract Implicational bases are objects of interest in formal concept analysis and its applications. Unfortunately, even the smallest base, the Duquenne-Guigues base, has an exponential size in the worst case. In this paper, we use results on the average number of minimal transversals in random hypergraphs to show that the base of proper premises is, on average, of quasi-polynomial size.

Keywords: Formal Concept Analysis, Implication Base, Average Case Analysis.

1 Introduction

Computing implication bases is a task that has been shown to be costly [6], due to their size and to the enumeration delay. Even the smallest base (the Duquenne-Guigues base) is, in the worst case, exponential in the size of the relation [12]. While the extremal combinatorics of implicational bases is a well studied subject, up to now, the average case has not received a lot of attention.

In this paper, we adapt the results presented in [5] to provide some average-case properties of implicational bases. We consider the base of proper premises and the Duquenne-Guigues base. We bound the average size of the base of proper premises under two statistical models and show that it is, on average, quasi-polynomial. This implies that the size of the Duquenne-Guigues base is on average at most quasi-polynomial. We then give an almost sure lower bound for the number of proper premises.

The paper is organized as follows: in section 2 we introduce the definitions and notations that we use in the remainder of the paper. Section 3 contains the main results of this work. In section 4, we discuss randomly generated contexts and the models that are used in this paper. We then conclude and discuss future works.

2 Definitions and Notations

In this section, we provide the definitions and results that will be used in this paper. Most of the FCA definitions can be found in [10]. From now on, we will omit the brackets in the notation for sets when no confusion is induced by this simplification.

2.1 Formal Concept Analysis

A *formal context* is a triple $\mathcal{C} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$ in which $\mathcal{O}$ and $\mathcal{A}$ are finite sets of objects and attributes and $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$ is a binary relation between them. A pair $(o, a) \in \mathcal{R}$ is read "object o has attribute a". Formal contexts can naturally be represented by cross tables, where a cross in the cell (o, a) means that $(o, a) \in \mathcal{R}$.

Table 1. Toy context $\mathcal{C}$.

	a_1	a_2	a_3	a_4	a_5
o_1	×	×			
o_2		×		×	×
o_3		×	×	×	
o_4			×		×
o_5				×	×

Table 1 shows a toy context with 5 objects and 5 attributes. It will serve as a running example throughout this paper.

Let O be a set of objects and A a set of attributes, we denote by O' the set of all attributes that are shared by all objects of O and A' the set of all objects that have all the attributes of A. More formally, $O' = \{a \in \mathcal{A} \mid \forall o \in O, (o, a) \in \mathcal{R}\}$ and $A' = \{o \in \mathcal{O} \mid \forall a \in A, (o, a) \in \mathcal{R}\}$.

The composition of those two operators, denoted $\cdot''$, forms a closure operator. A set $X = X''$ is said to be closed. A pair (O, A) with $O \subseteq \mathcal{O}$, $A \subseteq \mathcal{A}$, $A' = O$ and $O' = A$ is called a *(formal) concept* of the (formal) context $\mathcal{C}$. In this case, we also have that $A'' = A$ and $O'' = O$.

The set of all the concepts of a context, ordered by inclusion on either their sets of attributes or objects forms a complete lattice. Additionally, every complete lattice is isomorphic to the one formed by the concepts of a particular context.

Definition 1. *An* implication *(between attributes) is a pair of sets $X, Y \subseteq \mathcal{A}$. It is noted $X \rightarrow Y$.*

Definition 2. *An implication $X \rightarrow Y$ is said to* hold *in a context $\mathcal{C}$ if and only if $X' \subseteq Y'$.*

In an implication $X \rightarrow Y$, X is called the premise and Y the conclusion. Many implications are redundant, that is if an implication $a \rightarrow c$ holds, then $ab \rightarrow c$ holds and is redundant. The number of implications that hold can be quite large [12]. It is necessary to focus on the interesting ones.

Definition 3. *An implication set that allows for the derivation of all implications that hold in a context, and only them, through the application of Armstrong's axioms is called an implication base of the context.*

Definition 4 (Duquenne-Guigues Base). *An attribute set P is a pseudo-intent if and only if $P \neq P''$ and $Q'' \subset P$ for every pseudo-intent $Q \subset P$. The set of all the implications $P \to P''$ in which P is a pseudo-intent is called the Duquenne-Guigues Base.*

The Duquenne-Guigues Base, also called *canonical* base, or *stem* base has first been introduced in [11] and is the smallest (cardinality-wise) of all the bases. Here, we denote this base as Σ_{stem}. The complexity of enumerating the elements of this base is studied in [6].

Base of Proper Premises While the Duquenne-Guigues Base is the smallest base, the *base of proper premises*, or *Canonical Direct Base*, noted here Σ_{Proper}, is the smallest base for which the logical closure can be computed with a single pass. The Canonical Direct Base was initially known under five independent definitions, shown to be equivalent by Bertet and Montjardet in [2].

For a set X of attributes, let $X^\bullet$ be the set of attributes that are contained in X'' but not in the closure of any proper subset of X, that is

$$X^\bullet = X'' \setminus \left(X \cup \bigcup_{S \subset X} S'' \right).$$

X is called a *proper premise* for attribute a if $X^\bullet$ is not empty and $a \in X^\bullet$.

2.2 Hypergraphs and Transversals

Let V be a set of vertices. A hypergraph $\mathcal{H}$ is a subset of the powerset 2^V. Each $E \in \mathcal{H}$ is called an (hyper)edge of the hypergraph. A set $S \subseteq V$ is called a hypergraph transversal of $\mathcal{H}$ if it intersects every edge of $\mathcal{H}$, that is $S \cap E \neq \emptyset, \forall E \in \mathcal{H}$. A set $S \subseteq V$ is called a minimal hypergraph transversal of $\mathcal{H}$ if S is a transversal of $\mathcal{H}$ and S is minimal with respect to the subset inclusion among all the hypergraph transversals of $\mathcal{H}$. The set of all minimal hypergraph transversals of $\mathcal{H}$ forms a hypergraph, that we denote $Tr(\mathcal{H})$ and that is called the transversal hypergraph.

2.3 Proper Premises as Hypergraph Transversals

In this section, we introduce a definition of the base of proper premises based on hypergraph transversals.

Proposition 1 (from [10]). *$P \subseteq \mathcal{A}$ is a premise of $a \in \mathcal{A}$ if and only if $(\mathcal{A} \setminus o') \cap P \neq \emptyset$ holds for all $o \in \mathcal{O}$ such that $(o, a) \notin \mathcal{R}$. P is a proper premise for a if and only if P is minimal with respect to subset inclusion for this property.*

Proposition 23 from [10] uses $o \swarrow a$ instead of $(o, a) \notin \mathcal{R}$. It is a stronger condition that involves a maximality condition that is not necessary here.

The set of proper premises of an attribute is equivalent to the minimal transversals of a hypergraph induced from the context with the following proposition:

Proposition 2 (From [17]). *P is a premise of a if and only if P is a hypergraph transversal of $\mathcal{H}_a$ where*

$$\mathcal{H}_a = \{\mathcal{A} \setminus o' | o \in \mathcal{O}, (o, a) \notin \mathcal{R}\}$$

The set of all proper premises of a is exactly the transversal hypergraph $Tr(\mathcal{H}_a)$.

To illustrate this link, we show the computation of the proper premises for some attributes of Context 1. We compute the hypergraph $\mathcal{H}_a$ for a_1, a_2 and a_5. Let's begin with attribute a_1. We have to compute $\mathcal{H}_{a_1} = \{\mathcal{A} \setminus o' \,|o \in \mathcal{O}, (o, a_1) \notin \mathcal{R}\}$ and $Tr(\mathcal{H}_{a_1})$. In $\mathcal{C}$, there is no cross for a_1 in the rows o_2, o_3, o_4 and o_5. We have :

$$\mathcal{H}_{a_1} = \{\{a_1, a_3\}, \{a_1, a_5\}, \{a_1, a_2, a_3\}, \{a_1, a_2, a_4\}\}$$

and

$$Tr(\mathcal{H}_{a_1}) = \{\{a_1\}, \{a_2, a_3, a_5\}, \{a_3, a_4, a_5\}$$

We have the premises for a_1, which give implications $a_2a_3a_5 \rightarrow a_1$ and $a_3a_4a_5 \rightarrow a_1$. $\{a_1\}$ is also a transversal of $\mathcal{H}_{a_1}$ but can be omitted here, since $a \rightarrow a$ is always true.

In the same way, we compute the hypergraph and its transversal hypergraph for the other attributes. For example,

$$\mathcal{H}_{a_2} = \{\{a_1, a_2, a_3\}, \{a_1, a_2, a_4\}\} \text{ and } Tr(\mathcal{H}_{a_2}) = \{\{a_1\}, \{a_2\}, \{a_3, a_4\}\}$$

$$\mathcal{H}_{a_5} = \{\{a_1, a_5\}, \{a_3, a_4, a_5\}\} \text{ and } Tr(\mathcal{H}_{a_5}) = \{\{a_5\}, \{a_1, a_3\}, \{a_1, a_4\}\}$$

The set of all proper premises of a_i is exactly the transversal hypergraph $Tr(\mathcal{H}_{a_i})$, $\forall i \in \{1, \ldots, 5\}$, to which we remove the trivial transversals (a_i is always a transversal for $\mathcal{H}_{a_i}$). The base of proper premises for context $\mathcal{C}$ is the union of the proper premises for each attributes:

$$\Sigma_{Proper}(\mathcal{C}) = \bigcup_{a \in \mathcal{A}} Tr(\mathcal{H}_a) \setminus \{a\}$$

3 Average Size of an Implication Base

In [17], Distel and Borchmann provided expected numbers of proper premises and concept intents. Their approach, like the one in [5], uses the Erdős-Rényi model [8] to generate random hypergraphs. However, in [17], the probability for each vertex to appear in a hyperedge is a fixed 0.5 (by definition of the model) whereas the approach presented in [5] consider this probability as a variable of the problem and is thus more general.

3.1 Single Parameter Model

In the following, we assume all sets to be finite, and that $|\mathcal{O}|$ is polynomial in $|\mathcal{A}|$. We call p the probability that an object o has an attribute a. An object having an attribute is independent from other attributes and objects. We denote by $q = 1 - p$ the probability that $(o, a) \notin \mathcal{R}$. The probability of an attribute that is not a appearing in a hyperedge of $\mathcal{H}_a$ is also q.

The hypergraphs that we consider in the following are sub-hypergraphs constructed from $\mathcal{H}_a$ by removing a and removing all the hyperedges that contained only a. The transversal hypergraph of a hypergraph constructed in this way is exactly $Tr(\mathcal{H}_a) \setminus \{a\}$. This allows us to consider the transversal hypergraph without adding a as a premise for a. The average number of hyperedges of this hypergraph is $m = |\mathcal{O}| \times q \times (1 - p^{|\mathcal{A}|-1})$. Indeed, there is one hyperedge for each object o for which $(o, a) \notin \mathcal{R}$ and there exists an attribute a_2 such that $(o, a_2) \notin \mathcal{R}$ (otherwise the edge would be empty and, as such, removed). We note n the number of vertices of $\mathcal{H}_a \setminus \{a\}$. At most all attributes appear in $\mathcal{H}_a \setminus \{a\}$, except a, so $n \leq |\mathcal{A}| - 1$.

Proposition 3 (Reformulated from [5]). *In a random hypergraph with m edges and n vertices, with $m = \beta n^\alpha, \beta > 0$ and $\alpha > 0$ and a probability p that a vertex appears in an edge, there exists a positive constant c such that the average number of minimal transversals is*

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

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

Proposition 3 bounds the average number of minimal transversals in a hypergraph where the number of edges is polynomial in the number of vertices. In [5], the authors also prove that this quantity is quasi-polynomial.

From Prop. 3 we can deduce the following property for the number of proper premises for an attribute.

Proposition 4. *In a random context with $|\mathcal{A}|$ attributes, $|\mathcal{O}|$ objects and probability p that $(o, a) \in \mathcal{R}$, the number of proper premises for an attribute is on average:*

$$O\left((|\mathcal{A}| - 1)^{\left(d(\alpha) log_{\frac{1}{p}} \left(|\mathcal{O}| \times (q \times (1 - p^{|\mathcal{A}|-1}))\right) + c \ln \ln\left(|\mathcal{O}| \times (q \times (1 - p^{|\mathcal{A}|-1}))\right)\right)}\right)$$

and is quasi-polynomial in the number of objects.

Proposition 4 states that the number of proper premises of an attribute is on average quasi-polynomial in the number of objects in a context where the number of objects is polynomial in the number of attributes.

As attributes can share proper premises, $|\Sigma_{Proper}|$ is on average less than $|\mathcal{A}| \times O\left((|\mathcal{A}|-1)^{\Big(d(\alpha)log_{\frac{1}{p}}\big(|\mathcal{O}|\times q\times(1-p^{|\mathcal{A}|-1}))\big)+c\ln\ln\big(|\mathcal{O}|\times q\times(1-p^{|\mathcal{A}|-1})))\big)\Big)}\right)$

Since $|\Sigma_{stem}| \leq |\Sigma_{Proper}|$, Prop. 4 immediately yields the following corollary:

Corollary 1. *The average number of pseudo-intents in a context where the number of objects is polynomial in the number of attributes is less than or equal to:*

$$|\mathcal{A}| \times O\left((|\mathcal{A}|-1)^{\Big(d(\alpha)log_{\frac{1}{p}}\big(|\mathcal{O}|\times q\times(1-p^{|\mathcal{A}|-1})\big)+c\ln\ln\big(|\mathcal{O}|\times q\times(1-p^{|\mathcal{A}|-1}))\big)\Big)}\right)$$

Corollary 1 states that in a context where the number of object is polynomial in the number of attributes, the number of pseudo-intents is on average at most quasi-polynomial.

3.2 Almost Sure Lower Bound on the Number of Proper Premises

An almost sure lower bound is a bound that is true with probability close to 1. In [5], the authors give an almost sure lower bound for the number of minimal transversals.

Proposition 5 (Reformulated from [5]). *In a random hypergraph with m edges and n vertices, and a probability p that a vertex appears in an edge, the number of minimal transversals is almost surely greater than*

$$\mathcal{L}_{MT} = n^{log_{\frac{1}{q}} m + O(\ln\ln m)}$$

Proposition 5 states that in a random context with probability p that a given object has a given attribute, one can expect at least $\mathcal{L}_{MT}$ proper premises for each attribute.

Proposition 6. *In a random context with $|\mathcal{A}|$ attributes, $|\mathcal{O}|$ objects and probability q that a couple $(o, a) \notin \mathcal{R}$, the size of Σ_{Proper} is almost surely greater than*

$$|\mathcal{A}| \times (|\mathcal{A}|-1)^{\Big(log_{\frac{1}{p}}\big(|\mathcal{O}|\times q\times(1-p^{|\mathcal{A}|-1})\big)+O(\ln\ln\big(|\mathcal{O}|\times q\times(1-p^{|\mathcal{A}|-1})\big)))\Big)}$$

As Prop 6 states a lower bound on the number of proper premises, no bound on the number of pseudo-intents can be obtained this way.

3.3 Multi-parametric Model

In this section we consider a multi-parametric model that fits real life data better. In this model, each attribute j has a probability p_j of appearing in the description of a given object. All the attributes are not equiprobable.

We consider a context with m objects and n attributes. The set of attributes is partitioned into 3 subsets:

- The set U contains the attributes that appear in a lot of objects' descriptions (ubiquitous attributes). For all attributes $u \in U$ we have $q_u = 1 - p_u < \frac{x}{m}$ with x a fixed constant.
- The set R represents rare events, i.e. attributes that rarely appear. For all attributes $r \in R$, we have that $p_r = 1 - \frac{1}{\ln n}$ tends to 0.
- The set $F = \mathcal{A} \setminus (U \cup R)$ of other attributes.

Proposition 7 (Reformulated from theorem 3 [5]). *In the multi-parametric model, we have:*

- *If* $|F \cup R| = O(\ln |\mathcal{A}|)$*, then the size of the base of proper premises is on average at most polynomial.*
- *If* $|R| = O((\ln |\mathcal{A}|)^c)$*, then the size of the base of proper premises is on average at most quasi-polynomial.*
- *If* $|R| = \Theta(|\mathcal{A}|)$*, then the size of the base of proper premises is on average at most exponential on* $|R|$*.*

Proposition 7 states that when most of the attributes are common (that is, are in the set U), $|\Sigma_{Proper}|$ is on average at most polynomial. When there is a logarithmic quantity of rare attributes (attributes in R), $|\Sigma_{Proper}|$ is on average at most quasi-polynomial (in the number of objects). When most of the attributes are rare events, $|\Sigma_{Proper}|$ is on average at most exponential.

As in the single parameter model, Prop. 7 also yields the same bounds on the number of pseudo-intents.

4 Discussion on Randomly Generated Contexts

The topic of randomly generated contexts is important in FCA, most notably when used to compare performances of algorithms. Since [13], a few experimental studies have been made. In [4], the authors investigate the Stegosaurus phenomenon that arises when generating random contexts, where the number of pseudo-intents is correlated with the number of concepts [3].

As an answer to the Stegosaurus phenomenon raised by experiments on random contexts, in [9], the author discusses how to randomly and uniformly generate closure systems on 7 elements.

In [16], the authors introduce a tool to generate less biased random contexts, avoiding repetition while maintaining a given density, for any number of elements. However this tool doesn't ensure uniformity.

The partition of attributes induced by the multi-parametric model allows for a structure that is close to the structure of real life datasets [5]. However, we can't conclude theoretically on whether this model avoids the stegosaurus phenomenon discussed in [4]. This issue would be worth further theoretical and experimental investigation.

5 Conclusion

In this paper, we used results on average-case combinatorics on hypergraphs to bound the average size of the base of proper premises. Those results concerns only the proper premises, and can't be applied on the average number of pseudo-intents. However, as the Duquenne-Guigues base is, by definition, smaller than the base of proper premises, the average size of the base of proper premises can serve as an average bound for the number of pseudo-intents.

This approach does not give indications on the number of concepts. However, there exists some works on this subject [1, 7, 15].

As the average number of concepts is known [7, 15], and this paper gives some insight on the average size of some implicational bases, future works can be focused on the average number of pseudo-intents. It would also be interesting to study the average number of n-dimensional concepts or implications, with $n \geq 3$ [14, 18].

Acknowledgments

This research was partially supported by the European Union's "*Fonds Européen de Développement Régional (FEDER)*" program.

References

1. Albano, A., Chornomaz, B.: Why concept lattices are large: extremal theory for generators, concepts, and vc-dimension. Int. J. General Systems **46**(5), 440–457 (2017). https://doi.org/10.1080/03081079.2017.1354798, `https://doi.org/10.1080/03081079.2017.1354798`
2. Bertet, K., Monjardet, B.: The Multiple Facets of the Canonical Direct Unit Implicational Basis. Theor. Comput. Sci. **411**(22-24), 2155–2166 (2010). https://doi.org/10.1016/j.tcs.2009.12.021, `https://doi.org/10.1016/j.tcs.2009.12.021`
3. Borchmann, D.: Decomposing finite closure operators by attribute exploration. ICFCA 2011, Supplementary Proceedings (2011)
4. Borchmann, D., Hanika, T.: Some experimental results on randomly generating formal contexts. In: Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016. pp. 57–69 (2016), `http://ceur-ws.org/Vol-1624/paper5.pdf`
5. David, J., Lhote, L., Mary, A., Rioult, F.: An Average Study of Hypergraphs and their Minimal Transversals. Theor. Comput. Sci. **596**, 124–141 (2015). https://doi.org/10.1016/j.tcs.2015.06.052, `https://doi.org/10.1016/j.tcs.2015.06.052`

6. Distel, F., Sertkaya, B.: On the complexity of enumerating pseudo-intents. Discrete Applied Mathematics **159**(6), 450–466 (2011). https://doi.org/10.1016/j.dam.2010.12.004, https://doi.org/10.1016/j.dam.2010.12.004
7. Emilion, R., Lévy, G.: Size of random galois lattices and number of closed frequent itemsets. Discrete Applied Mathematics **157**(13), 2945–2957 (2009). https://doi.org/10.1016/j.dam.2009.02.025, https://doi.org/10.1016/j.dam.2009.02.025
8. Erdős, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**(1), 17–60 (1960)
9. Ganter, B.: Random extents and random closure systems. In: Proceedings of The Eighth International Conference on Concept Lattices and Their Applications, Nancy, France, October 17-20, 2011. pp. 309–318 (2011), http://ceur-ws.org/Vol-959/paper21.pdf
10. Ganter, B., Wille, R.: Formal Concept Analysis - Mathematical Foundations. Springer (1999)
11. Guigues, J. L., D.V.: Familles Minimales d'Implications Informatives Résultant d'un Tableau de Données Binaires. Mathématiques et Sciences Humaines **95**, 5–18 (1986), http://eudml.org/doc/94331
12. Kuznetsov, S.O.: On the Intractability of Computing the Duquenne-Guigues Base. J. UCS **10**(8), 927–933 (2004). https://doi.org/10.3217/jucs-010-08-0927, https://doi.org/10.3217/jucs-010-08-0927
13. Kuznetsov, S.O., Obiedkov, S.A.: Comparing performance of algorithms for generating concept lattices. J. Exp. Theor. Artif. Intell. **14**(2-3), 189–216 (2002). https://doi.org/10.1080/09528130210164170, https://doi.org/10.1080/09528130210164170
14. Lehmann, F., Wille, R.: A Triadic Approach to Formal Concept Analysis. In: Conceptual Structures: Applications, Implementation and Theory, Third International Conference on Conceptual Structures, ICCS '95, Santa Cruz, California, USA, August 14-18, 1995, Proceedings. pp. 32–43 (1995). https://doi.org/10.1007/3-540-60161-9-27, https://doi.org/10.1007/3-540-60161-9-27
15. Lhote, L., Rioult, F., Soulet, A.: Average number of frequent (closed) patterns in bernouilli and markovian databases. In: Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA. pp. 713–716 (2005). https://doi.org/10.1109/ICDM.2005.31, https://doi.org/10.1109/ICDM.2005.31
16. Rimsa, A., Song, M.A.J., Zárate, L.E.: Scgaz - A synthetic formal context generator with density control for test and evaluation of FCA algorithms. In: IEEE International Conference on Systems, Man, and Cybernetics, Manchester, SMC 2013, United Kingdom, October 13-16, 2013. pp. 3464–3470 (2013). https://doi.org/10.1109/SMC.2013.591, https://doi.org/10.1109/SMC.2013.591
17. Ryssel, U., Distel, F., Borchmann, D.: Fast Algorithms for Implication Bases and Attribute Exploration Using Proper Premises. Ann. Math. Artif. Intell. **70**(1-2), 25–53 (2014). https://doi.org/10.1007/s10472-013-9355-9, https://doi.org/10.1007/s10472-013-9355-9
18. Voutsadakis, G.: Polyadic Concept Analysis. Order **19**(3), 295–304 (2002). https://doi.org/10.1023/A:1021252203599, http://dx.doi.org/10.1023/A:1021252203599