About the Lossless Reduction of the Minimal Generator Family of a Context

T. Hamrouni[1,2,3], P. Valtchev[2,3], S. Ben Yahia[1], and E. Mephu Nguifo[4]

[1] URPAH, Faculté des Sciences de Tunis, Tunis, Tunisie
{tarek.hamrouni, sadok.benyahia}@fst.rnu.tn
[2] LATECE, Université du Québec à Montréal, Montréal, Québec, Canada
petko.valtchev@uqam.ca
[3] DIRO, Université de Montréal, Montréal, Québec, Canada
{hamrount, valtchev}@iro.umontreal.ca
[4] CRIL-CNRS, IUT de Lens, Lens, France
mephu@cril.univ-artois.fr

Abstract. Minimal generators (MGs), *aka* minimal keys, play an important role in many theoretical and practical problem settings involving closure systems that originate in graph theory, relational database design, data mining, etc. As minima of the equivalence classes associated to closures, MGs underlie many compressed representations: For instance, they form premises in *canonical* implication/association rules – with closures as conclusions – that losslessly represent the entire rule family of a closure system. However, MGs often show an intra-class combinatorial redundancy that makes an exhaustive storage and use impractical. In this respect, the *succinct system of minimal generators* (SSMG) recently introduced by Dong *et al.* is a first step towards a lossless reduction of this redundancy. However, as shown elsewhere, some of the claims about SSMG, *e.g.*, its invariant size and lossless nature, do not hold. As a remedy, we propose here a new succinct family which restores the losslessness by adding few further elements to the SSMG core, while theoretically grounding the whole. Computing means for the new family are presented together with the empirical evidences about its relative size *w.r.t.* the entire MG family and similar structures from the literature.

1 Introduction and Motivations

Minimal generators (MGs) [1], *aka* minimal keys, play an important role in many theoretical and practical problem settings involving closure systems that originate in graph theory, database design and data mining, to cite but a few. Standing at the "antipodes" of the closures within their respective equivalence classes in the Boolean lattice [2] – MGs are the smallest elements of a class while the closures are the largest – they help delimit the classes and hence ease their detection/traversal.

From a computational viewpoint, the MG set is often the intermediate step in the construction of structures that are either larger or lay higher in the Boolean lattice: the *frequent* itemset family [3], *frequent* closed itemset (CI) family [4,5], an iceberg lattice [6], etc. Underlying their computational importance is a structural property of the

S.O. Kuznetsov and S. Schmidt (Eds.): ICFCA 2007, LNAI 4390, pp. 130–150, 2007.

MG set, *i.e.*, its *order ideal* shape [5]. Indeed, the MGs are the first elements of their respective equivalence classes to be reached by a breadth-first climb in the Boolean lattice. This fact is essentially exploited both by level-wise [4,7] and depth-first [8] itemset mining algorithms in achieving better performances.

Beside their impact on computation and efficiency, the use of MGs brings gains on the semantic level, especially in decision support environments (*e.g.*, medical diagnosis). As they are usually strictly smaller than the closures (unless themselves closed), MGs offer minimal combinations of tests/exams/answers necessary to identify a class of situations and hence reduce the economic cost of the decision process. For example, they were shown to be highly instrumental in applications involving rule induction and classification [8].

On the structural side, various concise representations of the frequent itemset family have been defined in terms of MGs [9,10,11]. More interestingly, MGs underly a variety of compact subsets of the implication/association rule families of a context [1,12,13], which are hence called by some *generic bases*. Traditionally, a generic basis is considered as an irreducible nucleus of the underlying rule family, although some redundancy clearly persists. In fact, given two MGs g_1 and g_2 of the same equivalence class, there is a one-to-one correspondence between rules of the basis involving g_1 and those involving g_2.

In this respect, a study of intra-class redundancies in MGs was initiated by Dong *et al.*, who recently proposed a way to derive MGs from other ones in the same equivalence class [14]. The overall reduction principle may be grossly summarized as follows: an arbitrary total order is defined on the itemset family and the unique minimal members of the respective equivalence classes are kept. This results in a split of the global MG family into *succinct* and *redundant* parts. Thus, the succinct system of minimal generators (SSMG) was introduced as a concise representation from which the entire MG family can be retrieved without any information loss.

However, contrary to the authors' claim, the SSMG as defined in [14] proved to be loss-prone, *i.e.*, in some cases *a priori* redundant MGs are impossible to derive. Furthermore, the different SSMGs of a context (emerging through different orders) do not necessarily share the same size, again contradicting what was stated in [14]. As an attempted improvement on both issues, a new construct was hence proposed by Hamrouni *et al.* in [15]. Unfortunately, the new family lost the order ideal structure what greatly complicates its extraction.

In this paper, we propose a third system that overcomes the worst limitations of the previous ones. We present its definition and show that it preserves the precious order ideal property together with further structural properties that underly a lossless reduction mechanism. The presentation is organized as follows: the next section defines the basic constructs to be used throughout the remainder of the text. Section 3 is a detailed study of the SSMG as defined by Dong *et al.*, whereas Section 4 sketches that of Hamrouni *et al*. Section 5 expands on our own definition as well as its structural properties. An algorithm extracting the family is sketched in Section 6, while the empirical evidences about the utility of the approach is provided in Section 7.