Some Experimental Results on Randomly Generating Formal Contexts

Daniel Borchmann[1,2] and Tom Hanika[3]

[1] Chair of Automata Theory
Technische Universität Dresden, Germany
[2] Center for Advancing Electronics Dresden
Technische Universität Dresden, Germany
[3] Knowledge & Data Engineering Group
University of Kassel, Germany

daniel.borchmann@tu-dresden.de, tom.hanika@cs.uni-kassel.de

Abstract We investigate different simple approaches to generate random formal contexts. To this end, we consider for each approach the empirical correlation between the number of intents and pseudo-intents. We compare the results of these experiments with corresponding observations on real-world use-cases. This comparison yields huge differences between artificially generated and real-world data sets, indicating that using randomly generated formal contexts for applications such as benchmarking may not necessarily be meaningful. In doing so, we additionally show that the previously observed phenomenon of the "Stegosaurus" does not express a real correlation between intents and pseudo-intents, but is an artifact of the way random contexts are generated.

Keywords: Formal Concept Analysis, Pseudo-Intents, Closure Systems

1 Introduction

In the early times of Formal Concept Analysis [1], the study of lattices represented as the concept lattice of a particular formal context $\mathbb{K}$ was one of the main driving motivations. For this one has to solve the computational task of determining all formal concepts of $\mathbb{K}$, one of the first algorithmic challenges in the field of FCA. Since then, many algorithms have been developed to solve this task.

With the rise of a multitude of algorithms it became increasingly important to be able to *compare* these algorithms. One of the first comparisons was done in 2002 by Kuznetsov [2]. The data sets used in this comparison were all "randomly generated", a notion that up to today is not completely understood. Consequently, [2] regrets that there is no deeply investigated algorithm for generating random contexts.

From its original motivation, Formal Concept Analysis has since then evolved into an active research area with many connections to fields outside the scope of this original approach. Nevertheless, the study of properties of lattices in terms

of corresponding formal contexts is still one of the main lines of research. One of the earliest observations in this direction was that every concept lattice can be understood as the lattice of all closed sets of the valid implications of the underlying formal context. This observation did not only open up connections to fields like data-base theory, data-mining, and logic. It also fostered research on finding efficient algorithms for extracting small *bases* of implications of a given formal context. One of those bases, called the *canonical base*, stands out as base of minimal size for which an explicit construction is known. Recall that for a formal context $\mathbb{K} = (G, M, I)$ the canonical base $\mathcal{L}(\mathbb{K})$ is the set of implications defined by

$$\mathcal{L}(\mathbb{K}) := \{P \to P'' \mid P \text{ is pseudo intent of } \mathbb{K}\},$$

where *pseudo intents of* $\mathbb{K}$ are subsets of M such that $P \neq P''$ and for all pseudo intents $Q \subsetneq P$ it is true that $Q'' \subseteq P$. This recursive definition of pseudo intents makes theoretical investigations of the canonical base rather difficult. Indeed, Babin and Kuznetsov [3] showed that recognizing pseudo-intents is coNP-complete.

Although there are bases whose computation may be more worthwhile in practice, the canonical base is still of major interest for both research and applications. In 2011, Bazhanov and Obiedkov [4] made a performance comparison of the known algorithms to compute canonical bases. For this they used seven distinct real world contexts. More recently is a parallel approach by Borchmann and Kriegel [5]. To evaluate their algorithm they used random contexts as well as real-world contexts from the `fcarepository.com` (which disappeared recently).

It emerges that evaluating the performance of algorithms for computing the set of formal concepts as well as computing the canonical base heavily depends on the choice of the available data sets. Because obtaining real-world data sets may be a challenging endeavor, one often resolve to use artificially-generated "random contexts" instead. However, a thorough theory of randomly generated formal contexts is missing, and even experimental studies are hard to find. This is where this work tries to step in. In particular, it aims to shed some light on a phenomenon we shall call the *Stegosaurus-phenomenon*, a surprising empirically observed correlation between the number of pseudo intents and the number of formal concepts of formal contexts. We shall show that the phenomenon depends strongly on the method used for generating random contexts. Other random context generators show similar, but substantially different phenomena.

Finally, we want to compare our approaches of randomly generating formal contexts with two data sets constructed from real world data, namely from BibSonomy and form the Internet Movie Database. Not surprisingly, the correlation between the number of intents and pseudo-intents in these data sets differs considerably to those observed in the randomly generated contexts. This reminds of an obvious but too rarely stated meme from the early days of formal concept analysis: don't invent data!

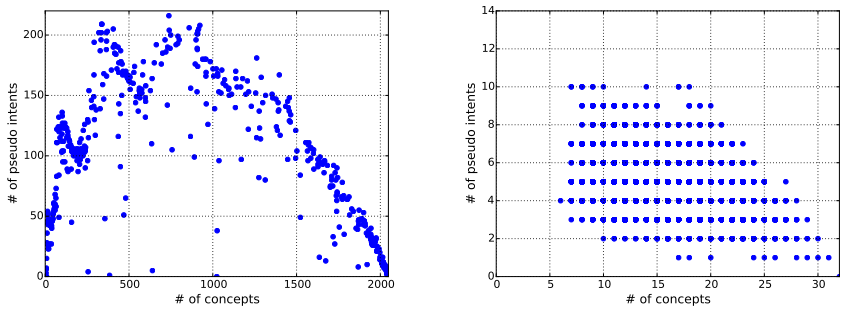

Figure 1. Experimentally observed correlation between the number of intents and pseudo-intents of randomly generated formal contexts on twelve attributes (left), plot of all formal contexts on five attributes (right).

2 Related Work

The original observation of a correlation between the number of intents and pseudo-intents first appeared in [6]. This work was originally not concerned with investigating this relationship, but with representing closure operators on sets by means of formal contexts of minimal size. However, during the experiments on the efficiency of this approach, a correlation between the number of intents and the number of pseudo-intents of randomly generated formal contexts was discovered. The original phenomenon is shown in Figure 1 and has subsequently been called the *Stegosaurus* (because, with some fantasy, the shape of Figure 1 resembles the one of this well-known dinosaur).

Further investigation was conducted in a talk at the in Formal Concept Analysis Workshop in 2011. There not only the experimental setup was discussed in more detail, but also questions were raised that are connected to the experiment. Most importantly, it was asked whether the phenomenon really exists, or whether it was just a programming error or an artifact of the experimental setup. Indeed, using a reimplementation[4], the second author was later able to independently verify the outcome of the experiment.

Another question raised in this investigation was whether the way the formal contexts were generated has an impact on the outcome of the experiment. The problem here is that although in the original experiment the formal contexts were generated in a uniformly random manner, the underlying closure systems were not. This is because closure systems can have multiple representations by means of formal contexts, and the number of those contextual representations may differ widely between different closure systems. Therefore, uniformly choosing a formal contexts does not mean to choose a closure system in a uniform way.

A first attempt to remove the shortcomings of the way random formal contexts are generated was conducted by Ganter [7]. In this work an approach was

[4] `https://github.com/tomhanika/fcatran`

investigated to correctly generate closure systems on a finite set with a uniform random distribution. However, while the proposed algorithm was conceptually simple, it turned out that it is not useful for our experiment. Indeed, it has been shown that the proposed algorithm is only practical for closure systems on sets of up to 7 elements, whereas the original experiment needs a size of at least 9 or 10 to exhibit the characteristic pattern of Figure 1. This is also the reason why an earlier computation of all reduced formal contexts on five attributes, shown in Figure 1, was not helpful to investigate the phenomenon.

3 Experiments

The purpose of this section is to present different experimental approaches to enhance our understanding of the Stegosaurus phenomenon. For this purpose, we shall first recall the original experiment that first exhibited the Stegosaurus. After this, we shall discuss an alternative approach of randomly generating formal contexts that fixes the number of attributes per object. Then we shall consider another method proposed [8]. Finally, we compare our findings against experiments on real world data.

All computations presented in this section were conducted using `conexp-clj`[5].

3.1 Original Experiment

The original experiment that first unveiled the Stegosaurus-phenomenon randomly generated formal contexts as follows. For a given number of attributes N and some $p \in [0, 1]$, first the number of objects is randomly chosen between 1 and 2^N. Then for each pair (g, m) of an object g and an attribute m, a biased coin with probability p was used to determine whether g has attribute m.

Applying this algorithm to generate 1000 formal contexts with $N = 10$ leads to the picture in Figure 2. The result does not change qualitatively by repetition. The provided generating algorithm seems biased towards creating contexts that lie on some idiosyncratic curve. This curve exhibits multiple spikes (in the given picture at least 4 can be identified) and a general skew to the left. Contexts beneath that curve are hit infrequently, above that curve even less. The behavior at the right end of the plot is expected, since when almost every subset of M is an intent, the number of pseudo intents must be low: the number of pseudo-intents of a formal context $\mathbb{K} = (G, M, I)$ is at most $2^{|M|}$ minus the number of intents of $\mathbb{K}$. On the other hand, the behavior in the rest of the picture is not as easily explained and still eludes proper understanding.

We also plotted a histogram in Figure 2 which contains a bin for every occurring number of pseudo intents. By the height of the erected rectangle above each bin we can observe the frequency of appearance of a formal context with that particular number of pseudo intents. The distribution shown in Figure 2 has an expected spike at zero: while generating a random formal context with a high

[5] `https://github.com/exot/conexp-clj`

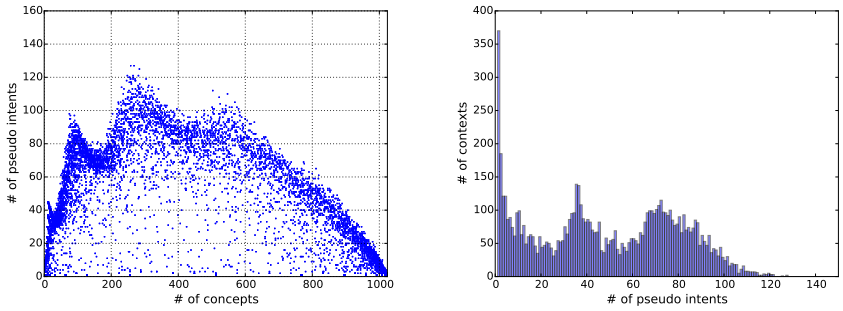

Figure 2. Experimentally observed correlation: Between the number of intents and pseudo-intents (left) and the distribution of the number of contexts having a given number of pseudo intents (right), for 1000 randomly generated formal contexts with ten attributes

probability of crosses, the chances of hitting the ten object-vectors spanning a contra-nominal-scale context is high. Apart from that, there is an cumulation of contexts for approximately 40 and 70 pseudo intents. For some reason the algorithm favors context with those pseudo intent numbers. This could also mean that the same context is generated for multiple times.

These unexpected results lead to many more questions to generate a deeper understanding of the connection between the number of formal concepts and pseudo intents.

One of these questions is what happens if a lot of contexts are generated that way. To address this question, we created five million random contexts using the introduced method. This led to the result shown in Figure 3. In contrast to Figure 2 we see a filled picture. Almost all combinations below the characteristic curve have been realized by at least one context. Only a small seam of not realized combinations is left at the bottom. At a second glance we observe that the whole characteristic curve seems shifted up by approximately ten to twenty pseudo intents. Even more interestingly, a fifth spike can be imagined at about 800 concepts. Furthermore, even in this figure there are still some random context hoovering even above the spikes. This leads to the conjecture that there are contexts with even larger canonical bases that cannot be computed feasibly by the applied method.

In Figure 3 we also plotted the according histogram like we did in Figure 2. The distribution of contexts is of course shifted up since more contexts are generated. But it still resembles the one in Figure 2. In particular, for contexts with about 50 pseudo-intents, a plateau can be observed.

Another question is how far the number of attributes we have chosen for our experiments has an influence on the shape of the Stegosaurus. Since in the first discovery of the Stegosaurus was made with a context that has eleven attributes, the question about the influence of N on the phenomenon is natural. To investigate this question, we computed, still using the same method, several

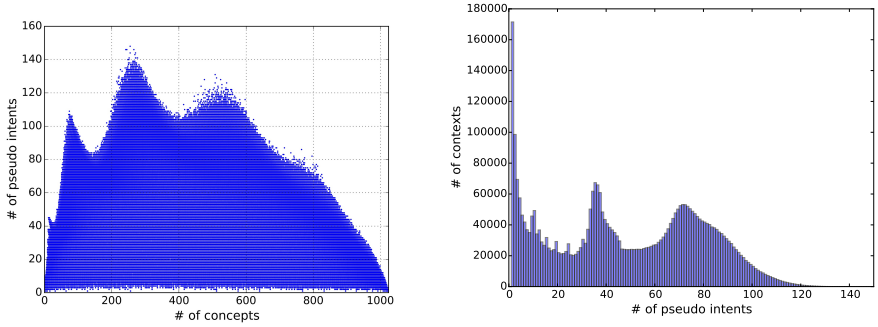

Figure 3. Experimentally observed correlation: Between the number of intents and pseudo-intents (left) and the distribution of the number of contexts having a given number of pseudo intents (right), for five million randomly generated formal contexts with ten attributes, using experiment in Section 3.1.

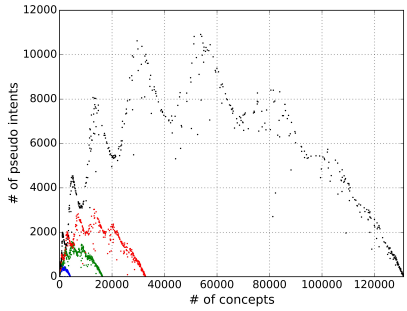

Figure 4. The influence of increasing m for the original experiment.

formal contexts with up to seventeen attributes. As can be see in Figure 4, the characteristic Stegosaurus curve is present in all of them. However, we also can see an increase in spikes.

Therefore, we conjecture that the occurrence of the Stegosaurus phenomenon seems independent from the value of N.

3.2 Increasing the number of pseudo-intents

As described in the previous section, in the original experimental setup the number of pseudo-intents of randomly generated formal contexts increases with the number of iterations. A natural question is whether we can find an upper bound on the number of pseudo-intents a formal context can have given that the number of intents is fixed. For this purpose, we investigate an alternative approach of generating formal contexts that is described in this section.

Let us say that a formal context $\mathbb{K} = (G, M, I)$ has *fixed row-density* if the number of attributes for each object $g \in G$ is the same. In other words, for

all $g, h \in G$ we have $|g'| = |h'|$. It is clear how to obtain such formal contexts: let $k, n \in \mathbb{N}$ with $0 \leq k < n$. Let $M = \{1, \ldots, n\}$ and choose $G \subseteq \binom{M}{k}$. Then the formal context (G, M, I), where $(S, i) \in I$ if and only if $i \in S$, has fixed row-density. Let us call a formal context $\mathbb{K}$ with fixed row-density *object maximal* if $\mathbb{K}$ is object clarified and no new object can be added to $\mathbb{K}$ such that the formal context is still object clarified and has fixed row-density. In other words, $\mathbb{K}$ is object maximal with fixed row-density if and only if $\mathbb{K}$ is isomorphic to $\mathbb{K}_{n,k} := (\binom{M}{k}, M, \ni)$, where $M = \{1, \ldots, n\}$.

Formal contexts with fixed row-density have been used by Kuznetsov in his performance comparison of concept lattice generating algorithms [2]. The following observation had already been hinted at (so we suppose) in [9], when it was claimed that constructing formal contexts with as much as $\binom{|M|}{\lfloor |M|/2 \rfloor}$ pseudo-intents is easy.

Proposition 1. *Let $k < n - 1$. The number of pseudo intents of $\mathbb{K}_{n,k}$ is $\binom{n}{k+1}$.*

Proof. Let $M = \{1, \ldots, n\}$. For all $P \subseteq M$ with $|P| = k + 1$ we see that $P \subsetneq P'' = M$. For all proper subsets $Q \subsetneq P$ it is clear that Q is an intent of $\mathbb{K}_{n,k}$, as it can be represented as an intersection of subsets of M of size k. Therefore, the subsets of M of cardinality $k+1$ are in fact pseudo-intents of $\mathbb{K}_{n,k}$, and there are $\binom{n}{k+1}$ many of them. Because each $k+1$-elemental subset $P \subseteq M$ satisfies $P'' = M$, we also have that there are no other pseudo-intents in $\mathbb{K}_{n,k}$.

In fact, for any attribute set M, object maximal formal contexts with fixed row-density are the contexts with the largest canonical base we discovered in our experiments so far. The results of applying this algorithm for $N = 10$ for various k can be seen in Figure 5. We observe the highest peak in the plot for $k = 4$, as Proposition 1 implies. For $k = 1$ we notice the ten possible formal contexts are plotted in between one to ten concepts, as expected, with up to 45 pseudo intents. In contrast to that, we find the ten possible contexts in the $k = 9$ case stringed along the axis for contexts with one pseudo intent, as expected for contexts resembling a contra-nominal scale.

An overlay of all those plots is shown in Figure 6, together with an overlay for $N = 11$ which, despite the thin and high spikes, both are reminiscent of Figure 2.We observe multiple sharp spikes, seven in the case of $N = 10$ and eight in the case of $N = 11$. The top of each spike is the object maximal formal context with fixed row-density for the corresponding k. For every k we observe a hump in the graph before the spike starts. The reasons for that hump as well as for the dale afterwards are unclear.

The curiosity about the Stegosaurus-phenomenon increases even more after overlaying Figure 6 with Figure 3. In contrast to the observation so far, now some spikes seem to "grow" out of dales in the original Stegosaurus plot. In particular, the question if the upper bound in the original Stegosaurus plot states some inherent correlation between the number of pseudo intents and the number of intents can be safely negated at this point.

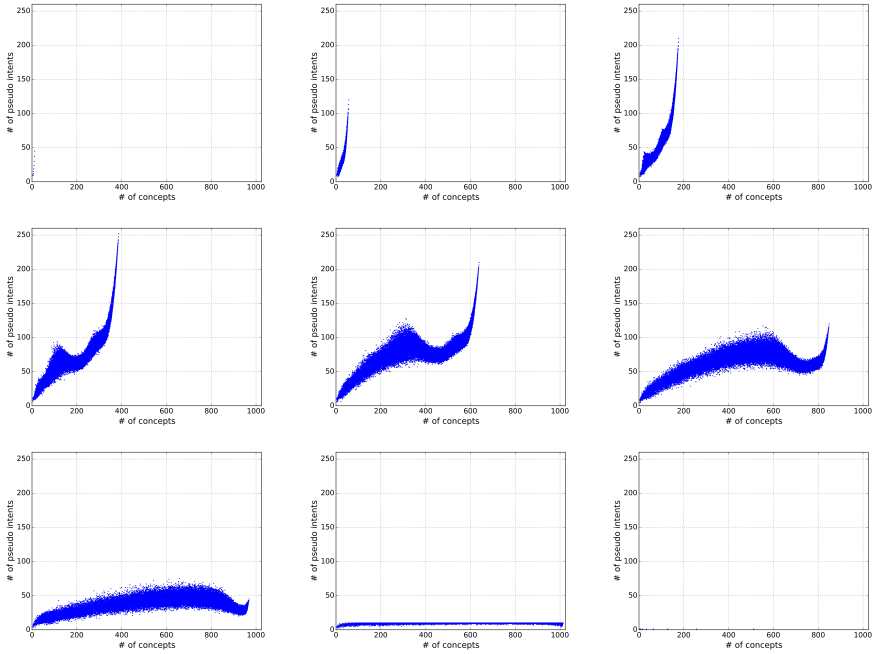

Figure 5. 100,000 random fixed row-density contexts for $|M| = 10$, plotted for $k = 1$ (upper left) up to $k = 9$ (down right).

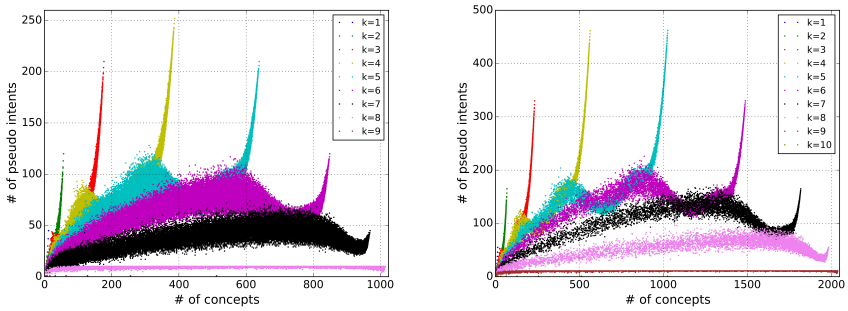

Figure 6. 100,000 random fixed row-density context for $m = 10$ (left) and $m = 11$ (right) for various k (best looked at in color).

3.3 SCGaz-Contexts

In 2013, Rimsa et al. [8] contributed a synthetic formal context generator named SCGaz[6]. The goal for this generator was to create random object irreducable formal contexts with a chosen density that have no full or empty rows and columns. The authors employ four different algorithms for each phase of the generation process, i.e., reaching minimum density, regular filling, coping with problems near the maximum density, and brute force. Since the interactions of these algorithms is rather involved and not possible to describe in short, we refer the reader to [8].

When this tool is invoked with a fixed number of attributes, a number (or an interval) of objects must be provided, as well as a density. In cases when the provided density does not fit with the other parameters, the density is interchanged with 0.5. For example, the request to generate a context with 32 objects, 5 attributes and density 0.9 is impossible, since there is only one object clarified context with those parameters, an object maximal formal context with fixed row-density, which has a density of 0.5.

This particularity in the usage of SCGaz made it tiring to generate a large number of random formal contexts, since a correct density had to be pre-calculated. We did so and generated a set of 3.5 Million contexts for a set of ten attributes, varying number of objects, and three different densities per object-attribute-number combination. The result is shown in Figure 7.

The first thing to observe is again a spike structure. However, the previously observed skew as in Figure 2 is gone, and the upper bound of the plot is significantly higher than in Figure 2. Furthermore, there seems to be a unnatural gap in the plot. This missing piece is an artifact of our parameter generation for invoking SCGaz, in particular the density bound calculations. We verified this by generating a small number of contexts using random densities which led to contexts resembling the same behavior as in Figure 7, but without the missing piece. However, in favor of the more filled plot we decided to include Figure 7 instead of the smaller sample.

Comparing the results from SCGaz with the one obtained in Section 3.2, we observe that some spikes stemming from context with fixed row-density emerge from dales in the SCGaz plot and others adapt closely to the SCGaz spikes. Nevertheless, all spikes we have seen in Section 3.2 outnumber the ones from SCGaz by the number of pseudo intents.

3.4 Real-World Contexts

The purpose of this section is to compare our observations about artificially generated formal contexts with results from experiments based on real-world data sets. The actual experiment is the same as before: we compute for a collection of formal contexts the number of intents and pseudo-intents and plot the result. However, in contrast to our previous experiments, we do not generate the formal

[6] `https://github.com/rimsa/SCGaz`

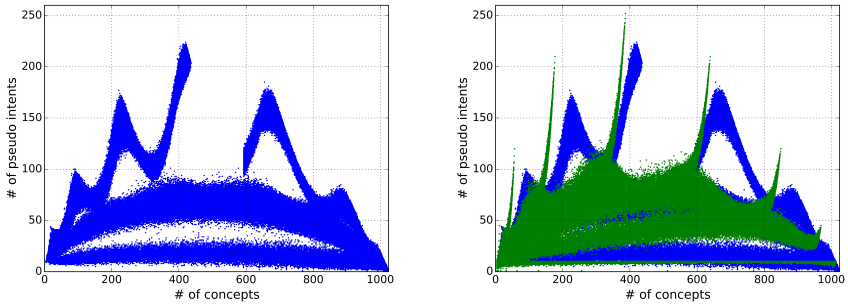

Figure 7. 3.5 Million random contexts generated by SCGaz, using ten attributes and varying density and number of objecst (left) and the same overlain by Figure 6 (right).

contexts using a designated procedure, but use some readily available data sets for this.

The first data sets stems from the BibSonomy project[7]. BibSonomy is a social publication sharing system that allows a user to tag publications with arbitrary tags. Using the publicly available anonymized data sets [10] of BibSonomy[8], we created 2835 contexts as follows. For every user u we defined a set of attributes M_u consisting of the twelve most frequently used tags of the user. The set of objects per user is the set of all the publications stored in BibSonomy. The incidence relation then is the obvious relation between publications and their tags.

The results are depicted in Figure 8. Note that even if the cardinality of the attribute set is twelve, the plot is shown only for up to 1024 intents, because no contexts with more than 1024 intents are contained in the data set.

The majority of the contexts seem to lie near a linear function of the number of concepts. Hence, it looks like the left part of the Stegosaurus phenomenon. Even a first spike can be accounted for with about 60 pseudo intents.

The distribution of contexts, however, behaves very differently. Of course, the first spike for contexts with no pseudo-intents is missing, as the contra-nominal scale is not common in real world data. Furthermore, we can find that there is no wide dale in the graph, like it is observed in Figure 2.

For our second real world data set we chose to use the Internet Movie Database[9]. We created 57582 formal contexts using the following approach. For every actor (context) we took the set of his movies (objects) and the related star-votes. Every movie can be rated from one to ten, and the ten bins of votes were considered as attributes. Every rate-bin that has at least 10% of the total amount of votes was considered as being present for an object. The resulting graphs are shown in Figure 9.

[7] `http://bibsonomy.org`
[8] `http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/`
[9] `http://www.imdb.com`

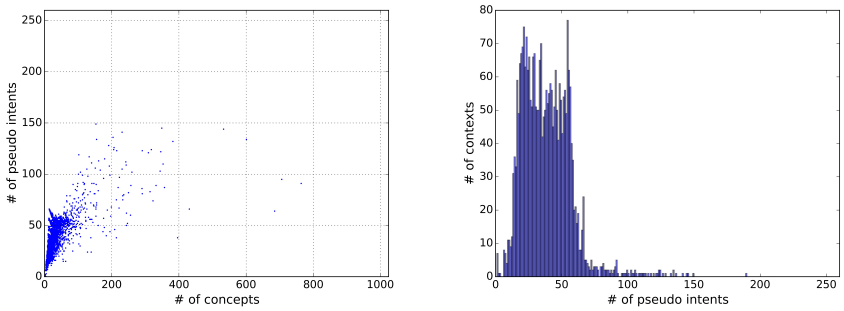

Figure 8. 2835 contexts created using the public BibSonomy data set.

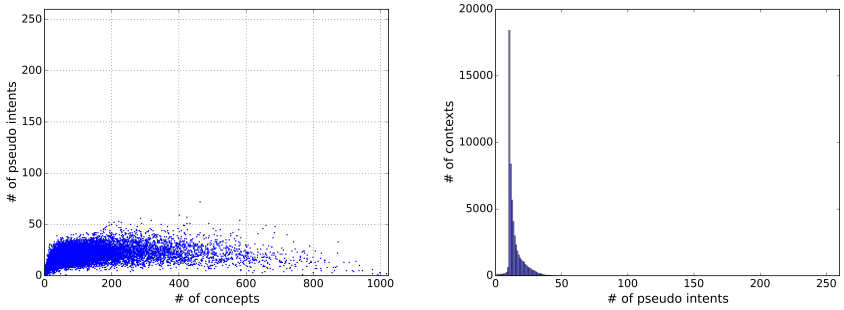

Figure 9. Formal contexts created using the Internet Movie Database.

We observe a quite different behavior to that of the classical Stegosaurus as well as to that of the BibSonomy data set. These contexts fill the area for infrequent contexts of the experiment in Section 3.1. Their canonical bases are mostly below 50 pseudo intents and the number of formal concepts goes up to 400 for a majority, and contexts around 1000 concepts are hit three times.

3.5 Discussion of the Experiments

Throughout our experiments, we observed that the Stegosaurus phenomenon seems to be more associated with the actual algorithm of constructing the formal contexts than with any unknown correlation between the number of pseudo intents and the number of formal concepts. Also, the upper bound which was suggested by the phenomenon appears vacuous for a deeper understanding of the correlation in question.

In particular, the experiments concerning formal contexts with fixed row-density nourished our understanding what actually can be the reason for the original phenomenon. Since the algorithm in Section 3.1 uses a constant probability for generating crosses, the row density in a context does not vary much.

Indeed, with N attributes and a cross probability of p, the expected number of attributes per object is pN. Therefore, in most cases, the algorithm generates an "approximation" of a context with fixed row-density. If one imagines Figure 6 without the thin spikes, the result resembles a lot the one of Figure 2.

At this point we cannot explain the result of the SCGaz context generator with respect to our experimental setup. However, in Figure 7 we see in the overlay plot that the dales are artificial since spikes are running right through them.

The final investigation using real world data sets leads to the question if all discussed random context generators miss the point of creating contexts that behave like real world data, making them unsuitable for real-world benchmarking. For the BibSonomy data set one could still argue that Figure 8 resembles the very left part of Figure 2 and Figure 7. However, in the case of the IMDB data set, strange capping of the number of pseudo intents can be observed that does not appear in any of our approaches of randomly generating formal contexts.

4 Conclusions and Outlook

At his first discovery, the Stegosaurus phenomenon raised a lot of questions. Is it a programming error, is it a systematic error, is it a hint to enhance the understanding of canonical bases? At this point, we feel confident to state that it is "just" a systematic bias in generating the contexts. Therefore, benchmarking FCA-algorithms using random contexts created by the original algorithm seems unreasonable. The SCGaz generator can be tuned to generate more diverse samples. However, this tuning needs some effort and there is still some unaccounted bias. In any way, the question what a truly "random context" is and how it can be sampled remains open.

Recalling the results of the real world data sets, one can conclude that the idea of randomly generating test data for algorithms needs some reconsideration. Like simple random generated graphs in general do not resemble a social network graph, randomly generated contexts might not reproduce real world contexts. In the case of randomly generating social graphs, the method of *preferential attachment* led to better results [11]. Hence, random context generators trying to sample formal contexts with the characteristics of some class of real world contexts would be an improvement in the realm of random contexts.

Still, new algorithmic ideas need to be tested. Therefore, a set of specialized random context generators, as proposed by Kuznetsov [2], producing contexts of a particular class would be an improvement. On the other hand, a standard set of formal contexts to test against should be compiled as well. To this end, the authors have obtained the abandoned domain `fcarepository.com` to revive the idea of a central repository of formal contexts in the next months.

References

1. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (1999)

2. Kuznetsov, S.O., Obiedkov, S.: Comparing performance of algorithms for generating concept lattices. Journal of Experimental and Theoretical Artificial Intelligence **14** (2002) 189–216
3. Babin, M.A., Kuznetsov, S.O.: Recognizing pseudo-intents is conp-complete. In: Proceedings of the 7th International Conference on Concept Lattices and Their Applications, Sevilla, Spain, October 19-21, 2010. (2010) 294–301
4. Bazhanov, K., Obiedkov, S.A.: Comparing performance of algorithms for generating the duquenne-guigues basis. In: Proceedings of The Eighth International Conference on Concept Lattices and Their Applications, Nancy, France, October 17-20, 2011. (2011) 43–57
5. Kriegel, F., Borchmann, D.: Nextclosures: Parallel computation of the canonical base. CLA 2015 (2015) 181
6. Borchmann, D.: Decomposing Finite Closure Operators by Attribute Exploration, University of Nicosia (May 2011)
7. Ganter, B.: Random extents and random closure systems. In Napoli, A., Vychodil, V., eds.: CLA. Volume 959 of CEUR Workshop Proceedings., CEUR-WS.org (2011) 309–318
8. 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: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. (Oct 2013) 3464–3470
9. Ganter, B.: Two Basic Algorithms in Concept Analysis. In: Formal Concept Analysis: 8th International Conference, ICFCA 2010, Agadir, Morocco, March 15-18, 2010. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 312–340
10. Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G.: The social bookmark and publication management system bibsonomy. The VLDB Journal **19**(6) (December 2010) 849–875
11. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature (393) (1998) 440–442