

# Finding interesting itemsets using a probabilistic model for binary databases

Tijl De Bie

University of Bristol, Department of Engineering Mathematics  
Queen's Building, University Walk, Bristol, BS8 1TR, UK  
tijl.debie@gmail.com

## ABSTRACT

A good formalization of interestingness of a pattern should satisfy two criteria: it should conform well to intuition, and it should be computationally tractable to use. The focus has long been on the latter, with the development of frequent pattern mining methods. However, it is now recognized that more appropriate measures than frequency are required.

In this paper we report results in this direction for itemset mining in binary databases. In particular, we introduce a probabilistic model that can be fitted efficiently to any binary database, and that has a compact and explicit representation. We then show how this model enables the formalization of an intuitive and tractable interestingness measure for itemsets, relying on concepts from information theory.

Our probabilistic model is closely related to the uniform distribution over all databases that can be obtained by means of swap randomization [8]. However, in contrast to the swap randomization model, our model is explicit, which is key to its use for defining practical interestingness measures.

## Categories and Subject Descriptors

H.2.8 [Database management]: Database applications—*Data mining*; I.5.1 [Pattern recognition]: Models—*Statistical*

## General Terms

Algorithms

## Keywords

Interestingness measures, frequent pattern mining, probabilistic modelling, binary databases, maximum entropy.

## 1. INTRODUCTION

Frequent Pattern Mining (FPM), and Frequent Itemset Mining (FIM) in particular, has quickly become a prototypical problem in data mining. The reasons are arguably the

elegance and efficiency of the algorithms to search for frequent patterns. Unfortunately, the frequency of a pattern is only loosely related to interestingness. The output of frequent pattern mining methods is usually an immense bag of patterns that are not necessarily interesting, and often highly redundant with each other. This has hampered the uptake of FPM and FIM in data mining practice.

Recent research has shifted focus to the search for more useful formalizations of interestingness that match practical needs more closely, while still being amenable to efficient algorithms. As FIM is arguably the simplest special case of frequent pattern mining, it is not surprising that most of the recent work has focussed on itemset patterns, see e.g. [4, 7, 13, 5, 6, 8, 15, 9], and we will follow this in this paper.

### *Statistically inspired interestingness measures.*

A sensible approach is to search for ‘large’ itemsets specified in a more comprehensive way than by means of frequency alone, but with a close eye on algorithmic scalability. For example, *tiles* (also known as *hyperrectangles* [16]) are itemset-tidset pairs covering ones in the database at their intersection, and their surface (product of the itemset’s size and frequency) has been suggested as a more effective interestingness measure than the itemset’s frequency alone [7].

Mining large tiles can be done relatively efficiently, which makes it an attractive alternative to frequent itemset mining. Still, the surface of a tile, while better than the frequency of an itemset, is often not sufficiently close to the interestingness (see also Empirical results in Sec. 4).

Other approaches are rooted more directly in statistics and in particular in hypothesis testing. Such methods attempt to compute the statistical significance of each itemset, or of a global measure that assesses the interestingness of the entirety of all itemsets found (e.g. the number of frequent itemsets) [4, 5, 6, 8, 15].

Such approaches are confronted with the challenge of specifying and computing a probabilistic null hypothesis for the database, to which the patterns are contrasted. Too naive a null hypothesis will result in the discovery of fairly trivial patterns (which are known to be present or easily spotted), rather than the interesting ones.

The null hypothesis from [8] is a particularly attractive one, being the uniform distribution over all databases with the same item frequencies and transaction sizes as in the given database. They argued that item frequencies and transaction sizes can often be thought of as prior knowledge, such that this null hypothesis is an adequate representation of the prior state of mind of the data mining practitioner.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

Unfortunately, no explicit characterization of this distribution is known. Nevertheless, random databases can be sampled from it to a good approximation and with reasonable efficiency using a Markov chain of *swap* operations. This makes it possible to statistically assess global measures of the whole of the discovered set of patterns by means of resampling methods. Still, the lack of an explicit characterization makes it useless for the assessment of discovered patterns individually, let alone for directly discovering patterns that are significant with respect to this null hypothesis.

### Contributions of this paper.

Here, we overcome some of the limitations of the above-mentioned approaches by complementing them with new ideas from statistical modelling and convex optimization theory. In particular, we show how it is possible to fit a *probabilistic model* to a database that respects the item frequencies and transaction sizes, much like in [8] using swap randomization, but here in a direct and explicit manner.

We envision that this new probabilistic model will prove useful in a variety of ways. Of particular interest in this paper, we show how it allows us to define *new interestingness measures* for tiles (or equivalently, itemsets). The proposed interestingness measure is the ratio of the information content of a tile with respect to the probability distribution fitted to the database, and the description length of the itemset-tidset pair describing the tile (see Sec. 3.1).

Moreover, like the swap randomization model, our model can be used to *assess the statistical significance of data mining results*. However, our model has a significant advantage over swap randomization: it does not rely on computation intensive Markov chain sampling techniques, of which the convergence to the desired distribution is hard to assess.

We conclude this paper with some empirical results that illustrate how the model can be efficiently fitted to a binary database; how the newly defined interestingness measure compares to two important existing measures; and how the model can be used for assessing the significance of data mining results.

## 2. THE MAXIMUM ENTROPY MODEL

In this Section, we introduce the principles of our probabilistic database model. We endeavour to create the model such that it most adequately reflects the prior knowledge of the data miner about the data and the patterns it contains.

### 2.1 The rationale of the model

It seems discussable whether one such model may have a sufficiently general applicability. Nevertheless, e.g. [8, 5, 6] have provided a convincing argumentation that the individual item frequencies, which we will refer to as the *empirical item marginals*, may be thought of as known by the practitioner. Similarly, in many cases it is warranted to regard the transaction sizes as prior knowledge. These transaction sizes, normalized by the total number of items, are also known as the *empirical transaction marginals*. Indeed, items and transactions are atomic entities that can usually be readily understood—this as opposed to correlations between items or transactions that can be much harder to understand, and certainly harder to predict.

For example, items like the plastic carrier bag or the daily newspaper typically overwhelm market basket analyses. Similarly, given that a basket is large, it is not unex-

pected that it simultaneously contains two specified items. As another example, if a text miner were told that the words ‘is’ and ‘the’ occur frequently together in the texts in his corpus, he would not receive much additional information—it is not surprising given his prior information about the language. Similarly, if his corpus consists of texts that widely vary in length, overlaps between long texts should be considered less interesting than between short texts.

In the exposition of this work we adopt the same assumption on this type of prior knowledge, following previous authors in believing that this assumption is practically relevant indeed. Nevertheless, we hope that our focus on this particular type of prior knowledge will not hide the generality of our methodology, and its potential to be applied for different types of prior knowledge.

As the dependency structure is what one typically would like to discover, our model should not encode any dependencies between items or transactions that can not be accounted for by their individual frequencies. To this end, we choose a data model in which all database entries are independent, while according to the prior knowledge about the item and transaction marginals. Still, this additional assumption does not unambiguously determine the model. Not to introduce any additional bias, we therefore choose the model with largest entropy satisfying these constraints.

### 2.2 The mathematical formalization

In this paper we will represent the database  $\mathcal{D}$  by the binary matrix  $\mathbf{D} \in \{0, 1\}^{n \times m}$ , with rows indexed by transactions  $\mathbf{t} \in \mathcal{T}$  and columns by items  $\mathbf{i} \in \mathcal{I}$ . The element on row  $\mathbf{t}$  and column  $\mathbf{i}$  is  $d_{\mathbf{t},\mathbf{i}} = 1$  iff item  $\mathbf{i}$  was part of transaction  $\mathbf{t}$ . We will now mathematically formalize the above description of the database model.

#### The independence assumption.

As we do not want to encode any spurious prior information, we will assume that all entries  $d_{\mathbf{t},\mathbf{i}}$  are independent of each other. This means that the probabilistic database model  $P$  we are looking for is the product distribution of a set of Bernoulli distributions, one for each entry  $d_{\mathbf{t},\mathbf{i}}$  in  $\mathbf{D}$ . Each such Bernoulli distribution is parameterised completely by the probability  $p_{\mathbf{t},\mathbf{i}}$  that  $d_{\mathbf{t},\mathbf{i}} = 1$ , so that we need to identify  $n \times m$  numbers  $p_{\mathbf{t},\mathbf{i}}$ .

#### Constraining the item and transaction marginals.

The empirical marginal for any item  $\mathbf{i}$  can be computed from  $\mathbf{D}$  as:

$$p_{\mathbf{i}} = \frac{1}{n} \sum_{\mathbf{t}} d_{\mathbf{t},\mathbf{i}}. \quad (1)$$

Similarly, the empirical marginal for any transaction  $\mathbf{t}$  is equal to:

$$p_{\mathbf{t}} = \frac{1}{m} \sum_{\mathbf{i}} d_{\mathbf{t},\mathbf{i}}. \quad (2)$$

In our model, we will constrain the marginal probabilities for items and transactions to be equal to these empirical estimates derived from the database, or formally:

$$\begin{aligned} \frac{1}{n} \sum_{\mathbf{t}} p_{\mathbf{t},\mathbf{i}} &= p_{\mathbf{i}}, \\ \frac{1}{m} \sum_{\mathbf{i}} p_{\mathbf{t},\mathbf{i}} &= p_{\mathbf{t}}. \end{aligned}$$

### The MaxEnt objective.

The number of constraints is only  $n+m$ , while the number of free variables in the database model  $P$  is equal to  $n \times m$ : all entry probabilities  $p_{\tau,i}$ . Hence,  $P$  is typically not fully identified by the constraints on the marginals. In such cases, in order to avoid biasing the distribution, one searches for the distribution of maximal entropy among those that satisfy the constraints [10]. We will refer to this distribution as the MaxEnt distribution. Despite the wide acceptance of the choice for MaxEnt in the absence of sufficient information, we appreciate that this choice may seem arbitrary at the moment. However, we will provide an additional justification in Sec. 2.3.

### The complete optimization problem.

Stated mathematically, the MaxEnt database model  $P$  is thus found by solving the following optimization problem:

$$\max_{p_{\tau,i}} - \sum_{\tau,i} [p_{\tau,i} \log p_{\tau,i} + (1 - p_{\tau,i}) \log(1 - p_{\tau,i})] \quad (3)$$

$$\text{s.t.} \quad \sum_{\tau} p_{\tau,i} = np_i, \quad (4)$$

$$\sum_i p_{\tau,i} = mp_{\tau}. \quad (5)$$

### The final shape of the model.

In Appendix A, we show that the optimal values of  $p_{\tau,i}$  for this optimization problem can be stated compactly as:

$$p_{\tau,i} = \frac{\exp(\lambda_i + \mu_{\tau})}{1 + \exp(\lambda_i + \mu_{\tau})},$$

where  $\lambda_i$  and  $\mu_{\tau}$  are dual variables corresponding to the primal constraints, which can be computed efficiently using standard convex optimization techniques (e.g. pseudo-Newton, see Table 4 for an algorithm). Additionally, there is a lot of redundancy among these variables: many  $\lambda_i$  are equal to each other, and similarly for  $\mu_{\tau}$ . More specifically, the number of different  $\lambda_i$  and  $\mu_{\tau}$  is upper bounded by  $\min\{m, n, \sqrt{2N}\}$ , where  $N$  is the number of nonzero entries in  $\mathbf{D}$ . Hence, the probabilistic model can be characterized using a number of variables sublinear in the size of  $\mathbf{D}$ .

## 2.3 Connections to swap randomization

Previous research [8] has looked into the use of Monte Carlo randomizations to generate databases with the same empirical transaction and item marginals. Comparing patterns found in a given database with thus randomized versions of that database allows one to assess the significance of the patterns, by estimating empirical p-values for example. These randomizations are based on swaps, which are elementary operations on a database that transform it into another database leaving the marginals unchanged.

To explain what is meant by a *swap* operation on a binary database  $\mathbf{D}$ , consider a  $2 \times 2$  submatrix for transactions  $\tau_1$  and  $\tau_2$  and items  $i_1$  and  $i_2$ , where  $d_{\tau_1,i_1} = d_{\tau_2,i_2} = 1$  and  $d_{\tau_1,i_2} = d_{\tau_2,i_1} = 0$ . A swap is defined as the operation that changes these values into  $d_{\tau_1,i_1} = d_{\tau_2,i_2} = 0$  and  $d_{\tau_1,i_2} = d_{\tau_2,i_1} = 1$ .

By carrying out swaps for randomly selected  $2 \times 2$  submatrices, a Markov chain of random databases is generated, each of which has the same empirical transaction and item marginals. In the limit, if the swaps are randomly sampled according to appropriate distributions, this Markov chain has been shown to converge to the uniform distribution over

all databases with these marginals, so that approximate uniform sampling from these databases is made possible. However, the mixing time of the Markov chain is hard to assess, such that it is unknown how many random swaps need to be done in order to arrive at the uniform distribution.

Evidently, our MaxEnt model is different from the uniform distribution over all databases with given marginals. In the MaxEnt model, the empirical marginals of sampled databases are only equal to the given empirical marginals in expectation, rather than exactly. In fact, this may be a desirable feature, as restricting the empirical marginals to be exactly equal may be an overly strong prior, and indeed the authors of [8] suggest that further work could look into approaches that allow relaxing this constraint. A second and related difference is that, while the database entries are all independent in our model, this is not the case in the uniform distribution over databases with fixed marginals.

Still, there is a very strong connection between both probabilistic models, as exemplified by the following Theorem.

**THEOREM 2.1.** *Databases with the same specific set of empirical transaction and item marginals are equally probable under the MaxEnt model. Furthermore, given constraints of the form Eq. (4,5), the MaxEnt model is the only probabilistic model with independent  $p_{\tau,i}$  for which this holds.*

In particular, this means that all databases with empirical marginals equal to those of the given database are equally probable under our MaxEnt model, just as in the uniform limit distribution obtained by swap randomizations. The proof is given in Appendix B.

## 3. FINDING INTERESTING ITEMSETS

A crucial difference between the swap randomization approach and our database modelling approach is that the latter allows one to create the model itself, rather than just to sample from it. We foresee that this has the potential of becoming the foundation on which many measures of interestingness can be based. This could be so for itemsets, but also for association rules, and possibly other patterns in transactional databases. However, in this paper we will confine our attention to the definition of itemsets in Sec. 3.1.

In Sec. 3.2, as a second (and more direct) application of our model, we will discuss its use for the statistical assessment of data mining results.

### 3.1 Information ratio: a new interestingness measure

#### The information content of a tile.

A first attempt to design a measure of interestingness that springs to mind is the probability of a tile under our probabilistic model. We deliberately use the terminology ‘tile’, an itemset together with its supporting transaction set, as it conveys the fact that items and transactions are treated on the same footing for the MaxEnt model.

For a tile  $\tau = (T, I)$  with tidset  $T$  and itemset  $I$ , this probability  $P(\tau)$  can be calculated as the product of all  $p_{\tau,i}$  for  $\tau \in T$  and  $i \in I$ . More conveniently, we can define the interestingness measure as minus the log-probability. Note that this log-probability is also equal to the Shannon information content of a tile. It is given by:

$$\text{InformationContent}(\tau) = - \sum_{\tau \in T, i \in I} \log(p_{\tau,i}).$$

Written in this way, one can see that the information content is a refinement of the surface of a tile: it would be equivalent to it if all  $p_{t,i}$  were equal to each other, which is the case when all transaction marginals are equal to each other as well as all item marginals.

While certainly relevant, the information content may not be appropriate directly as a measure of informativeness. For example, tiles with a singleton itemset (or a singleton tidset) could be considered very interesting by this measure, even though they do not reveal any associations between items (or transactions).

Rather than considering its information content, we should contrast a tile's information content with the length of the description (i.e., information) required to encode the tile. This is exactly what our proposed measure of interestingness does: the Information Ratio.

### *Information ratio: a new interestingness measure.*

A tile will contain 'net' information when it helps understanding the database better, *without itself being too complex to describe*. Hence, we need to take into account two aspects: the length by which the description of the database can be reduced after the tile is known to be present, and the description length of the tile itself.

The first of these aspects is exactly  $\text{InformationContent}(\tau)$ : the database entries covered by the tile need not be described anymore in order to describe the database, so that this sum of negative log-probabilities is what is gained in terms of the description length of the database.

The description length of a tile is harder to quantify, as we need to decide on an encoding scheme for tiles to do this. One sensible way of doing it would be to assume a probabilistic model for tiles in which transactions occur independently in  $T$  and items occur independently in  $I$ , with probabilities equal to their marginal probabilities  $p_t$  and  $p_i$ . The description length for a tile using a code based on this model is then given by

$$\begin{aligned} & \text{DescriptionLength}(\tau) \\ = & - \sum_{t \in T} \log(p_t) - \sum_{i \in I} \log(p_i) \\ & - \sum_{t \notin T} \log(1 - p_t) - \sum_{i \notin I} \log(1 - p_i). \end{aligned}$$

Putting this together, we suggest the *information ratio* as an interestingness measure, defined as:

$$\text{InformationRatio}(\tau) = \frac{\text{InformationContent}(\tau)}{\text{DescriptionLength}(\tau)}.$$

Other choices for the tile encoding will obviously yield different results. However, typically the tile description will remain linear in size to the number of items and transactions in  $I$  and  $T$ , whereas the description length of a tile entry by entry is quadratic. This means that as soon as the number of items and the number of transactions are above a certain level, the tile will be deemed more interesting—just a large tidset with a singleton itemset (or vice versa) will not suffice to receive a high interestingness score.

Remember that  $\text{InformationContent}(\tau)$  can be seen as a refinement of the surface  $|T| \cdot |I|$  of a tile. Similarly,  $\text{InformationRatio}(\tau)$  can be seen as a refinement of  $\frac{|T| \cdot |I|}{a|T| + b|I| + 1}$  with  $a, b$  positive constants. They would be equivalent if all transaction and item marginals were equal to each other.

### *The interestingness of a set of tiles.*

One is rarely only interested in the single most interesting pattern from a database. Therefore, it has been suggested to output maximally interesting *sets* of patterns. This is non-trivial: since patterns may convey overlapping information, the interestingness of a set of tiles cannot simply be reduced to the sum of the interestingness of each tile it contains.

The information content of a tile can, however, easily be generalized to the information content of a set of tiles, namely as the negative log probability of all database entries covered by the set of tiles. For the same reasons mentioned above, we should trade off this information content with the total description length of all tiles in the set, e.g. by considering the ratio of the information content and the description length of the set of tiles (as we did for a single tile).

An essentially equivalent way to trade off these quantities is to search for the set of tiles with the largest information content, with an upper bound on the description length of the set of tiles encoded as itemset-tidset pairs. This problem is an instantiation of the *budgeted maximum coverage problem* [11], a hard combinatorial optimization problem. Still, it is well known that a near-optimal solution can be found using a greedy algorithm. Applied to the context of this paper, this near-optimal greedy algorithm selects tiles in decreasing order of what could be called an *added interestingness*, an adaptation of  $\text{InformationRatio}(\tau)$ :

$$\text{InformationRatio}^+(\tau) = \frac{\text{InformationContent}^+(\tau)}{\text{DescriptionLength}(\tau)}. \quad (6)$$

Here, the numerator is defined as the additional information conveyed by the specification of the tile  $\tau$ :

$$\text{InformationContent}^+(\tau) = \sum_{\substack{t \in T, i \in I, \\ (t, i) \text{ uncovered}}} \log(p_{t,i}),$$

where the sum is only over the transaction-item pairs in the tile that have not yet been covered by already selected tiles.

The algorithm can be applied to any set of possibly interesting tiles, e.g. the set of all tiles corresponding to all sufficiently frequent closed itemsets, as found using a standard FIM algorithm. The greedy algorithm then efficiently sorts them in decreasing order of  $\text{InformationRatio}^+$ .

It can be proven for the greedy approximation that any first  $k$  tiles in this list have a total information content that is at least  $1 - \frac{1}{e}$  times the maximum that is achievable given the total description length of these tiles. Hence, each set of top- $k$  tiles would provide a good compression of the database, and therefore, we argue, a good intelligible summary.

Similar greedy approximations to set covering type problems have been used earlier, for example for the interestingness measure defined as the surface of a tile [7, 12, 16], and for interestingness measures defined using p-values [5, 6].

## 3.2 Assessing data mining results

Another direct application of our model is for the statistical assessment of data mining results by means of p-values. The probabilistic model can be used to sample a large number of databases, and we can run the data mining algorithm on these randomized versions. Then, the value of a global measure of choice (such as the number of frequent closed itemsets, the size of the largest frequent itemset, etc) on the database can be compared with the value of the same

measure as obtained on the random databases. This can be used to compute an empirical p-value, which is smaller as the pattern is more significant in the given database.

The authors of [8] generated random databases by means of a chain of swap randomizations. Our explicit database model allows to improve on this in two respects:

- Little is known about the convergence of the swap randomization process. It is therefore unclear how many random swaps need to be done in order to arrive at a sufficiently randomized database, as if it were randomly sampled from the uniform distribution over all databases with the same marginals. In contrast, the MaxEnt distribution is characterized explicitly, and can be directly sampled from.
- A MaxEnt-based sampling method is extremely simple to implement using a double for loop over the items and transactions. The overall complexity of such implementation is  $O(mn)$ , and does not depend on the mixing time of the chain of random swaps.

The main issue is probably the first one, although the authors of [8] argue that it is less relevant in practice: it seems to suffice to do roughly as many swaps as the number of non-zero entries in the database, or a small multiple thereof. As a swap (for their Self\_loop algorithm) can be done in constant time, this means that their method could be faster than  $O(mn)$  for very sparse databases, assuming sparsity does not negatively affect the number of swaps required for mixing. Still, even on sparse datasets, the computational performance of our method appears comparable. For example, for the Retail dataset [3] which has a density of 0.06%, [8] reports a computation time of 1 minute and 1 second on a 3GHz Pentium. Our method requires 1 minute and 35 seconds on a 2GHz Pentium Centrino. Furthermore, we wish to point out that our simple sampling algorithm is subject to significant improvements by relying on results in [14], but we postpone a more in depth discussion to further work.

## 4. EMPIRICAL RESULTS

We report empirical results on four datasets. Two textual datasets, and two other datasets commonly used for evaluation purposes:

**KDD** All KDD paper abstracts since 2001 (from all sessions) downloaded from the ACM website. Each abstract is represented by a transaction and words are items, after stop word removal and stemming.

**Mushroom** A publicly available item-transaction dataset [1].

**Pubmed** All Pubmed abstracts retrieved by querying with the search query “data mining” (first used in [6]), after stop word removal and stemming.

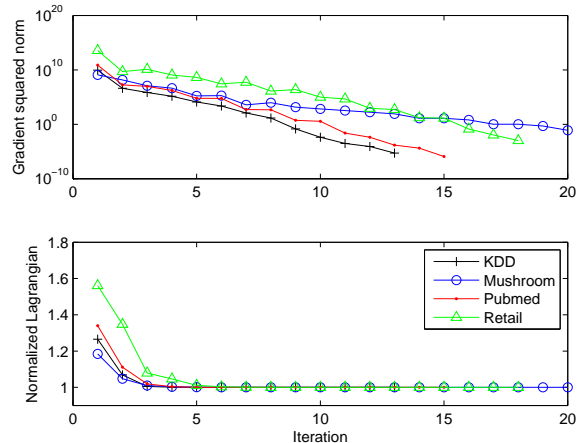
**Retail** A dataset about transactions in a Belgian supermarket store, where transactions and items have been anonymized [3].

Some statistics are gathered in Table 1. The Table also mentions support thresholds used for some of the experiments reported below, and the numbers of closed itemsets satisfying these support thresholds.

All experiments were done on a 2GHz Pentium Centrino with 1GB Memory, and implemented in C++.

**Table 1: Some statistics for the datasets investigated.**

	# items	# tids	support	# closed
KDD	6154	843	5	2,787,847
Mushroom	120	8,124	812	6,298
Pubmed	12,661	1,683	10	1,249,913
Retail	16,470	88,162	8	191,088



**Figure 1: Top: the squared norm of the dual gradient on a logarithmic scale as a function the iteration number, plotted for the four datasets investigated. This plot shows the exponential decrease of the gradient. In the second plot, the convergence of the Lagrange dual Eq. (11) is shown for the same datasets.**

### 4.1 Computing the probabilistic model

Fig. 1 shows the squared norm of the dual gradient (see Eqs. (12,13) in Appendix A for details) as it converges to zero during the first 20 iterations. Noting the logarithmic vertical axis, the convergence appears clearly exponential. The lower plot in Fig. 1 shows the convergence of the dual objective (see Eq. (11) in Appendix A) to its minimum, clearly a very fast convergence in just a few iterations.

We wish to note that the dual gradient contains as its elements the difference between the empirical marginal probabilities  $p_t$  and  $p_i$  on the one hand, and their counterparts under the estimated probabilistic model (see Eqs. (12,13)). Hence, the dual gradient norm quantifies the total violation of the constraints. This means that the norm of the gradient, normalized by the number of items and transactions, provides for a suitable stopping criterion. In all our experiments we stopped the iterations as soon as it reached  $10^{-12}$ . The number of iterations and the computation time to fit the model to the data are summarized in Table 2.

**Table 2: Number of iterations and computation time in seconds required to fit the probabilistic model.**

	# iterations	time (s)
KDD	13	0.54
Mushroom	37	0.02
Pubmed	15	1.24
Retail	18	2.80

## 4.2 Information ratio as interestingness

We implemented the greedy algorithm described in Sec. 3.1, which sorts a set of tiles in order of decreasing added interestingness, as formalized by  $\text{InformationRatio}^+$  in Eq. (6). We then applied it to the list of closed tiles as found by CHARM [17]. The support thresholds we used in running CHARM, and resulting number of tiles, are listed in Table 1.

In Fig. 2,  $\text{InformationRatio}^+(\tau)$  is plotted for the 30 top ranking tiles, as a function of their rank in the list. The dashed lines show the same result for a randomly generated dataset, sampled from the fitted probabilistic model. Note that in the random databases, the information ratio is typically smaller than 1, even for the most highly ranked tiles.

The textual datasets were included to allow for a subjective assessment of the interestingness measures. We believe that such a subjective assessment is crucial in an evaluation. Even though it may be hard to make any strong statements, it does provide one with an insight into the practical use and behaviour of the interestingness measure.

The top-10 itemsets returned for the textual datasets are shown in Table 3. For comparison, we also list the outputs given by the method based on the surface of tiles as proposed in [7], and by a method called KRIMP proposed in [13]. We would argue that  $\text{InformationRatio}^+$  achieves the most sensible results in terms of non-redundancy and interestingness of the highly ranked itemsets, many of them coinciding with major topics and concepts in data mining (KDD) and data mining applied to biological problems (Pubmed). In contrast, the tile-based method seems to favour tiles with few but individually frequent items. The KRIMP algorithm suffers from another problem when used for this purpose: the redundancy among the top-ranked itemsets is high (e.g., for the KDD abstracts almost all top-ranked itemsets contain ‘data’ or ‘paper’), and their interestingness seems worse than for the itemsets that are ranked highly by our method.

## 4.3 Assessing data mining results

Lastly, we wish to demonstrate the use of our probabilistic model for generating random databases with expected item and transaction marginals equal to the empirical ones. The availability of such random databases can then be used for the statistical assessment of data mining results.

Fig. 3 plots the number of closed itemsets retrieved on each of these four databases considered in this paper. Additionally, it shows box plots for the results obtained on databases randomly sampled from our probabilistic model fitted to the respective databases. If desired, these results allow one to extract one global measure from these results, as in [8], and to compute an empirical p-value by comparing that measure obtained on the actual data with the result on the randomized versions. However, the plots given here do not force one to make such a choice, and still they give a good idea of the patterns contained in the databases.

As a side-note, we remark that the number of closed itemsets in the Retail database is smaller than what is typically obtained in a randomly generated database. However, this seems to be attributable by the fact that large tiles, only present in the actual database, are shattered in the randomized versions, thus leading to a larger number of smaller ones. This effect is even stronger for the Mushroom database. Hence, the total number of closed itemsets above a fixed threshold would not be a good quantification of the amount of structure in the database.

## 5. CONCLUSIONS

We have introduced a probabilistic model for transactional databases, and developed a highly scalable algorithm that fits it to a given database. Such a model can be used to sample random databases much like earlier approaches, in order to assess the significance of a data mining result.

More importantly, the explicit availability of a probabilistic model for the entire database, respecting sensible prior knowledge (the item and transaction marginals), opens up new possibilities to develop measures of interestingness for patterns on such databases. Specifically, we suggested a new interestingness measure for itemsets (or tiles) in binary databases, which we referred to as the information ratio, and evaluated it on real-life data.

We hope that this work may open two new research directions. The first of these is the search for relevant interestingness measures that are based on our database model, for itemsets but also for other patterns on transactional databases, including association rules and ‘noisy’ tiles (tiles that cover also zeroes in the database).

As a second research direction, we hope that the results in this paper may lead to the specification of similar probabilistic models for more complexly structured data than transactional databases, such as sequential data, strings, graphs, and more. It is quite likely that the core ideas presented in this paper will transfer relatively easily. Indeed, a closer examination of the optimization problem in this paper reveals that the simple shape of the MaxEnt distribution, and the computational tractability, are essentially due to the linearity of the constraints on the probabilities specifying the distribution. We believe that linear constraints on probabilities (such as the constraints on the item and transaction marginals in this paper) are sufficiently rich to quantify a broad variety of prior information.

All data used in this paper not previously made available, and all code used in the experiments, will be made publicly available on the author’s website.

## 6. ACKNOWLEDGEMENTS

The author wishes to thank Bart Goethals for very interesting discussions relating to this work, and Matthijs Van Leeuwen for kindly making the code of KRIMP available.

## 7. REFERENCES

- [1] C. Blake and C. Merz. Uci repository of machine learning databases. In <http://www.ics.uci.edu/mllearn/MLRepository.html>. 1998.
- [2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [3] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: a case study. In *Proceedings of the fifth ACM SIGKDD international conference*, 1999.
- [4] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, 1997.
- [5] A. Gallo, T. De Bie, and N. Cristianini. Mini: Mining informative non-redundant itemsets. In *Proceedings of 2007 KDD*, 2007.

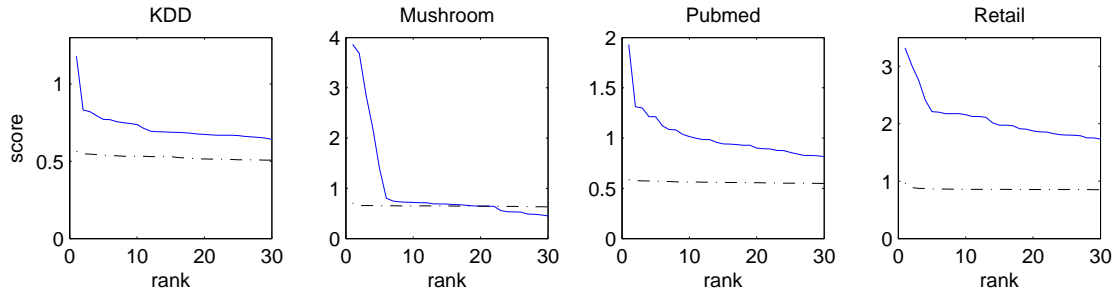


Figure 2: These plots show the scores of the ranked closed itemsets for the four datasets investigated (full line). In dotted line, the same is plotted for a randomized version of that dataset.

Table 3: The top-10 ranked itemsets in the textual datasets based on the Information Ratio interestingness measure (top), based on the surface of tiles (middle), and as outputted by KRIMP (bottom). Words were stemmed, and all itemsets are sorted alphabetically.

KDD	Pubmed
machin support svm vector art state data labeled semi supervised unlabeled algorithm effici frequent mine pattern algorithm data paper propos real synthetic graph larg network social associ mine rule express gene classifi featur machin support text vector algorithm faster magnitud order	algorithm data database drug gamma item mgps mine multi poisson report... ...safeti shrinker system chain data express gene polymerase reaction revers advers data detect drug mine reactions report signal spontan chromatography data lc liquid mass mine ms data est express gene mine sequenc tag acid amino data mine protein sequenc avail availability data mine motivation result [url] nucleotide polymorphisms singl snps studi data express gene pcr rt artifici data mine network neural
data paper algorithm propos data mine base method result show problem data set approach model present	data mine data method result analysi data express gene base studi data inform develop system approach
algorithm base data improv perform propos provid set analysi base data tool user visual visualization algorithm data mine paper propos set applic data paper problem propos set data method problem real set synthetic algorithm applic effici paper set studi algorithm base frequent larg paper set data demonstr mine paper pattern structur algorithm deriv effici method problem propos algorithm construct data paper problem relat	algorithm data drug event gamma item mine multi poisson report safeti shrinker... ...system advers algorithm data drug gamma item mine multi poisson report shrinker system advers data detect drug method mine report signal spontan conclus data determin import method mine result studi advers data event method mine report safeti signal age analysi conclus data mine result studi analysi data decis method mine result studi tree analysi combin conclus data includ method mine result analysi conclus data examin method mine rate result base clinical data decis inform mine support system

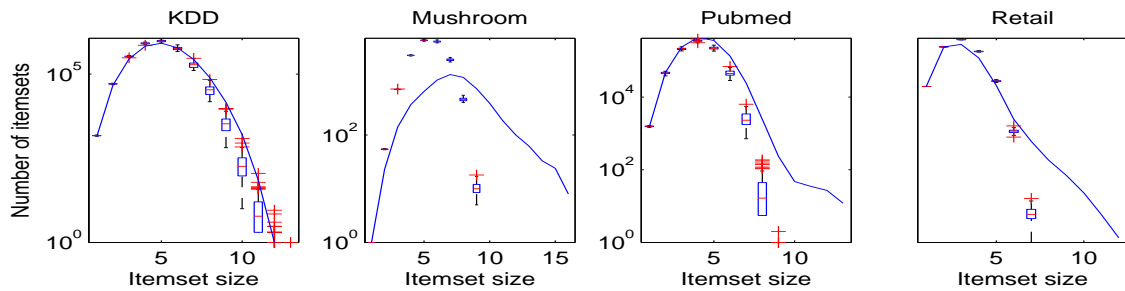


Figure 3: For the four datasets under investigation, these plots show the number of closed itemsets on a logarithmic scale, as a function of their size. Additionally, box plots are shown for the number of closed itemsets as a function of size found on 100 randomized datasets, based on our probabilistic model.

- [6] A. Gallo, A. Mammone, T. De Bie, M. Turchi, and N. Cristianini. From frequent itemsets to informative patterns. Submitted, 2009.
- [7] F. Geerts, B. Goethals, and T. Mielikäinen. Tiling databases. In *Discovery Science*, 2004.
- [8] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *TKDD*, 1(3), 2007.
- [9] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining in Knowledge Discovery*, 15(1):55–86, 2007.
- [10] E. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70, 1982.
- [11] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70, 1999.
- [12] T. Mielikäinen. *Summarization Techniques for Pattern Collections in Data Mining*. PhD thesis, Department of Computer Science, University of Helsinki, 2005.
- [13] A. Siebes, J. Vreeken, and M. van Leeuwen. Item sets that compress. In *SIAM Conference on Data Mining*, 2006.
- [14] Y. Tillé. *Sampling algorithms*. Springer, 2006.
- [15] G. Webb. Discovering significant patterns. *Machine Learning*, 68(1), 2007.
- [16] Y. Xiang, R. Jin, D. Fuhry, and F. Dragan. Succinct summarization of transactional databases: an overlapped hyperrectangle scheme. In *Proceedings of the 14th ACM SIGKDD international conference*, 2008.
- [17] M. Zaki and C. Hsiao. CHARM: An efficient algorithm for closed itemsets mining. In *Proc. of the 2nd SIAM ICDM*, 2002.

## APPENDIX

### A. EFFICIENTLY FITTING THE MODEL

#### *The shape of the MaxEnt solution.*

Optimization problem (3-5) cannot be solved analytically. However, the constraints are linear and the objective is strictly concave, so that a unique maximum must exist.

Duality theory [2] allows us to achieve better insight in the shape of the solution. Let us introduce  $m$  dual variables  $\lambda_i$  for constraints (4) and  $n$  dual variables  $\mu_t$  for constraints (5). Then problem (3) can be optimized by solving:

$$\max_{p_{t,i}} \min_{\lambda_i, \mu_t} L(p_{t,i}, \lambda_i, \mu_t), \quad (7)$$

where  $L(p_{t,i}, \lambda_i, \mu_t)$  is the Lagrangian:

$$\begin{aligned} L(p_{t,i}, \lambda_i, \mu_t) &= - \sum_{t,i} \left[ p_{t,i} \log p_{t,i} + (1 - p_{t,i}) \log(1 - p_{t,i}) \right. \\ &\quad \left. - \lambda_i p_{t,i} - \mu_t p_{t,i} \right] - n \sum_i \lambda_i p_i - m \sum_t \mu_t p_t. \end{aligned} \quad (8)$$

Thanks to convexity, strong duality holds, meaning that we can change the order of the maximization and the minimization operators in Eq. (7) without altering the result. After doing this, we can carry out the inner maximization explic-

itly, by equating the gradient of the Lagrangian with respect to the primal variables  $p_{t,i}$  to 0:

$$\begin{aligned} \frac{\partial L}{\partial p_{t,i}} = 0 &\Leftrightarrow \log p_{t,i} - \log(1 - p_{t,i}) - \lambda_i - \mu_t = 0, \\ &\Leftrightarrow \log \frac{p_{t,i}}{1 - p_{t,i}} = \lambda_i + \mu_t, \end{aligned} \quad (9)$$

$$\Leftrightarrow p_{t,i} = \frac{\exp(\lambda_i + \mu_t)}{1 + \exp(\lambda_i + \mu_t)} \quad (10)$$

Note that  $1 \geq p_{t,i} \geq 0$ , as required for probabilities. Hence we did not have to specify these constraints explicitly.

If we substitute this in the Lagrangian (Eq. (8)) we obtain the Lagrange dual of optimization problem (3-5):

$$\begin{aligned} \min_{\lambda_i, \mu_t} \sum_{t,i} \log(1 + \exp(\lambda_i + \mu_t)) \\ - n \sum_i \lambda_i p_i - m \sum_t \mu_t p_t. \end{aligned} \quad (11)$$

This is an unconstrained convex optimisation problem that can be solved efficiently by means of standard convex optimization techniques (see e.g. [2]).

#### *Efficient solution of the optimisation problem.*

In order to use widely established first or even second order methods, we need to be able to compute the gradient and second order information efficiently.

The gradient of this objective function is readily computed by:

$$\frac{\partial L}{\partial \lambda_i} = \sum_t \frac{\exp(\lambda_i + \mu_t)}{1 + \exp(\lambda_i + \mu_t)} - n p_i \quad (12)$$

$$\frac{\partial L}{\partial \mu_t} = \sum_i \frac{\exp(\lambda_i + \mu_t)}{1 + \exp(\lambda_i + \mu_t)} - m p_t \quad (13)$$

The Hessian is computed by:

$$\begin{aligned} \frac{\partial^2 L}{\partial \lambda_i \lambda_{i'}} &= 0, \\ \frac{\partial^2 L}{\partial \lambda_i^2} &= \sum_t \frac{\exp(\lambda_i + \mu_t)}{(1 + \exp(\lambda_i + \mu_t))^2}, \\ \frac{\partial^2 L}{\partial \mu_t \mu_{t'}} &= 0, \\ \frac{\partial^2 L}{\partial \mu_t^2} &= \sum_i \frac{\exp(\lambda_i + \mu_t)}{(1 + \exp(\lambda_i + \mu_t))^2}, \\ \frac{\partial^2 L}{\partial \lambda_i \mu_t} &= \frac{\exp(\lambda_i + \mu_t)}{(1 + \exp(\lambda_i + \mu_t))^2}. \end{aligned}$$

The direct computation of both the gradient and the Hessian would be highly demanding for large  $n + m$ . However, typically there is a lot of redundancy in the gradient and the Hessian. Observe that if  $p_i = p_{i'}$ , the constraints remain satisfied and value of the primal objective remains unchanged by interchanging  $p_{t,i}$  and  $p_{t,i'}$  for all  $t$ . Since the optimum is unique, this means that  $p_{t,i} = p_{t,i'}$ . Therefore, we can reduce the number of free parameters and constraints at the outset taking this into account. A similar property holds when  $p_t = p_{t'}$ .

The result is that the total number of distinct primal parameters is only equal to the number of *distinct* values of the  $p_i$  times the number of *distinct* values of the  $p_t$ , and

**Table 4: Algorithm for computing the model.**

In:	The item marginals $p_i$ and $p_t$ .
1:	Find all distinct values of $p_i$ and $p_t$ , and denote these by $p_{\bar{i}}$ and $p_{\bar{t}}$ . Compute the sets of transactions $\mathcal{T}_{\bar{t}} \subseteq \mathcal{T}$ such that $p_t = p_{\bar{t}}$ for all $t \in \mathcal{T}_{\bar{t}}$ , and similarly the sets $\mathcal{I}_{\bar{i}} \in \mathcal{I}$ . Compute the multiplicities $m_{\bar{i}} =  \mathcal{I}_{\bar{i}} $ and $n_{\bar{t}} =  \mathcal{T}_{\bar{t}} $ .
2:	Pick initial values for $\lambda_{\bar{t}}$ and $\mu_{\bar{t}}$ (e.g. = 0).
3:	Compute the gradient: $\frac{\partial L}{\partial \lambda_{\bar{i}}} = \sum_{\bar{t}} n_{\bar{t}} \frac{\exp(\lambda_{\bar{i}} + \mu_{\bar{t}})}{1 + \exp(\lambda_{\bar{i}} + \mu_{\bar{t}})} - n p_{\bar{i}}$ $\frac{\partial L}{\partial \mu_{\bar{t}}} = \sum_{\bar{i}} m_{\bar{i}} \frac{\exp(\lambda_{\bar{i}} + \mu_{\bar{t}})}{1 + \exp(\lambda_{\bar{i}} + \mu_{\bar{t}})} - m p_{\bar{t}}$ Compute the diagonal of the Hessian: $\frac{\partial^2 L}{\partial^2 \lambda_{\bar{i}}} = \sum_{\bar{t}} n_{\bar{t}} \frac{\exp(\lambda_{\bar{i}} + \mu_{\bar{t}})}{(1 + \exp(\lambda_{\bar{i}} + \mu_{\bar{t}}))^2}$ $\frac{\partial^2 L}{\partial^2 \mu_{\bar{t}}} = \sum_{\bar{i}} m_{\bar{i}} \frac{\exp(\lambda_{\bar{i}} + \mu_{\bar{t}})}{(1 + \exp(\lambda_{\bar{i}} + \mu_{\bar{t}}))^2}$
5:	Update: $\lambda_{\bar{i}} \leftarrow \lambda_{\bar{i}} - f \cdot \frac{\partial L}{\partial \lambda_{\bar{i}}} / \frac{\partial^2 L}{\partial^2 \lambda_{\bar{i}}}$ $\mu_{\bar{t}} \leftarrow \mu_{\bar{t}} - f \cdot \frac{\partial L}{\partial \mu_{\bar{t}}} / \frac{\partial^2 L}{\partial^2 \mu_{\bar{t}}}$ where $f$ is a step length found by a line search so as to decrease the Lagrangian.
6:	If the gradient's norm is small enough, go to 7. Otherwise, go to 3.
7:	For all $t \in \mathcal{T}_{\bar{t}}$ , equate $\mu_t = \mu_{\bar{t}}$ . For all $i \in \mathcal{I}_{\bar{i}}$ , equate $\lambda_i = \lambda_{\bar{i}}$ .
Out:	All $\lambda_{\bar{i}}$ and $\mu_{\bar{t}}$ , directly useable to compute $p_{t,i} = \frac{\exp(\lambda_{\bar{i}} + \mu_{\bar{t}})}{1 + \exp(\lambda_{\bar{i}} + \mu_{\bar{t}})}$ where $p_{\bar{t}} = p_t$ and $p_{\bar{i}} = p_i$ .

the number of distinct dual variables is equal to the sum of those two numbers. Typically, these numbers are much smaller than  $n$  and  $m$ , and it is not hard to show that they are both smaller than  $\min\{n, m, \sqrt{2N}\}$  where  $N$  is the total number of nonzero entries in  $\mathbf{D}$ . Hence, their number is small for sparse databases in particular, when itemset mining often proves most useful. This simplification makes the optimization problem easily amenable to real-life problems.

In principle, any standard first or second order technique for solving convex optimization problems could now be used. We found that the pseudo-Newton method did very well in particular, requiring only basic vector operations in each iteration, and converging to near machine accuracy in a few tens of iterations for all problems we considered (see Table 2).

The complete algorithm is given in Table 4, taking into account the technicalities that exploit the redundancy in the gradient and the Hessian. The most expensive step is the computation of the gradient and the diagonal of the Hessian and the line search, each taking a time proportional to the product of the number of distinct empirical item marginals and the number of distinct empirical transaction marginals.

The output of the algorithm is the optimal value for all dual variables  $\mu_t$  and  $\lambda_i$ . This is a surprisingly compact representation for the database model, linear in the size of the database dimensions. It allows computing the individual entry probabilities in constant time using Eq. (10).

## B. A LEMMA AND A PROOF OF THEOREM 2.1

LEMMA B.1. *Under the MaxEnt model, it holds that for pair of transactions  $t_1$  and  $t_2$  and pair of items  $i_1$  and  $i_2$ , the constellation*

$$d_{t_1, i_1} = d_{t_2, i_2} = 1 \quad \text{and} \quad d_{t_1, i_2} = d_{t_2, i_1} = 0$$

is equally probable as

$$d_{t_1, i_1} = d_{t_2, i_2} = 0 \quad \text{and} \quad d_{t_1, i_2} = d_{t_2, i_1} = 1.$$

In other words,

$$\begin{aligned} & p_{t_1, i_1} p_{t_2, i_2} (1 - p_{t_1, i_2}) (1 - p_{t_2, i_1}) \\ &= p_{t_1, i_2} p_{t_2, i_1} (1 - p_{t_1, i_1}) (1 - p_{t_2, i_2}). \end{aligned} \quad (14)$$

PROOF. Let us rewrite condition (14) as:

$$\begin{aligned} & \log \frac{p_{t_1, i_1}}{1 - p_{t_1, i_1}} + \log \frac{p_{t_2, i_2}}{1 - p_{t_2, i_2}} \\ &= \log \frac{p_{t_1, i_2}}{1 - p_{t_1, i_2}} + \log \frac{p_{t_2, i_1}}{1 - p_{t_2, i_1}}. \end{aligned} \quad (15)$$

Clearly, a parametric solution to this set of equations is given by Eq. (9). Hence, the maximum entropy solution is a solution for which property Eq. (15) needs to be satisfied.  $\square$

Said in other words, this means that a swap operation applied to a database does not alter its probability under the MaxEnt model. This Lemma allows us to prove Theorem 2.1.

### Proof of Theorem 2.1.

PROOF. Let us first prove the first part of the Theorem: under the MaxEnt model, all databases with the same empirical transaction and item marginals are equally probable. To this end, note that all these databases can be obtained by using a chain of swap operations starting from one such database that has the required marginals. Since Lemma B.1 states that swap operations do not alter the probability of the database, all databases with the same marginals must have the same probability.

The second part of the Theorem can be proven as follows. Observe that the number of linearly independent equations of the form of Eq. (15) is equal to  $(n-1)(m-1) = nm - (n+m) + 1$ . For example, choose the values of  $t_1$ ,  $i_1$ ,  $t_2$ , and  $i_2$  in Eq. (15) such that  $t_1 = i_1 = 1$  and  $t_2 \in \{2, 3, \dots, n\}$  and  $i_2 \in \{2, 3, \dots, m\}$ . These equations being linear and homogeneous in the  $nm$  log-odds ratios, so the solution space is an  $n + m - 1$  dimensional linear manifold. Now, also Eq. (9) defines an  $n + m - 1$  dimensional linear manifold, as parameterized by values of  $\lambda_i$  and  $\mu_t$ . (Indeed, while the number of degrees of freedom seems to be  $n + m$ , in fact it is only  $n + m - 1$ , as summing a fixed value to all  $\lambda_i$  and subtracting the same value from all  $\mu_{ttid}$  will not alter the values of the log-odds ratios). From the first part of this Theorem, it follows that any solution satisfying Eq. (9) also satisfies Eq. (15). Additionally, we have just shown that the two solution spaces are equally large. Thus, the sets of conditions Eq. (9) and Eq. (15) are equivalent.  $\square$