# Interaction Graphs for Multivariate Binary Data

Johan Van Horebeek[1] and Jesús Emeterio Navarro-Barrientos[2]

[1] Centro de Investigación en Matemáticas, CIMAT,
Ap. Postal 402, 36000 Guanajuato, GTO, Mexico
[2] School of Mathematical and Statistical Sciences, Arizona State University,
P.O. Box 871804, Tempe, AZ
horebeek@cimat.mx, jnavarrob@asu.edu

**Abstract.** We define a class of graphs that summarize in a compact visual way the interaction structure between binary multivariate characteristics. This allows studying the conditional dependency structure between the underlying stochastic variables at a finer scale than with classical probabilistic Graphical Models. A model selection strategy is derived based on an iterative optimization procedure and the consistency problem is discussed. We include the analysis of two data-sets to illustrate the proposed approach.

**Keywords:** Graphical Models, conditional dependency, data visualization.

## 1 Introduction

Graphical elements are of increasing importance in a variety of methods to study the interaction structure of multivariate data. In this context, graphs as employed in *Graphical Models* [9], have turned out to be useful instruments. Especially in Machine Learning they have become a standard tool in areas like probabilistic expert systems, statistical learning or data mining [3].

Graphical Models were originally conceived to describe families of multivariate distributions by means of a graph where nodes represent the variables and the absence of edges between variables reflects particular conditional independencies in the underlying interaction structure.

The use of graphs contributed significantly to the success of these models: on one hand, they provide an intuitively clear interface between the data, its underlying distribution and the user; on the other hand they allow to formulate and derive many efficient inference algorithms in terms of characteristics of the graph like cliques and separability.

Despite of the above, relatively little attention has been paid to how to make Graphical Models more informative and suitable for exploratory purposes. Probably the oldest example can be found in [14] where the width of the edges is made variable in order to reflect the support of the corresponding conditional independence hypothesis. This serves as a guide in data understanding and model building. However, it reflects only an overall summary at the level of variables.

A. Kuri-Morales and G.R. Simari (Eds.): IBERAMIA 2010, LNAI 6433, pp. 433–442, 2010.

*Split Models* [7] and *Generalized Graphical Models* [12] extend Graphical Models by introducing a way to incorporate information about conditional (in)dependencies which depend on the values of the variables in the conditional part. The latter adds labels to the edges, the former builds a tree of graphs; an example is shown in Fig. 1 for the case $X_1 \perp X_2 | X_3 = 0$ and $X_1 \not\perp X_2 | X_3 = 1$, i.e., only when $X_3$ equals 0, the variables $X_1$ and $X_2$ are conditionally independent. A similar approach is taken in [7]; it is directly formulated in terms of the underlying clique structure but with no visual representation.

In other approaches like [8] additional information is included about the presence of symmetry in the underlying interaction structure.

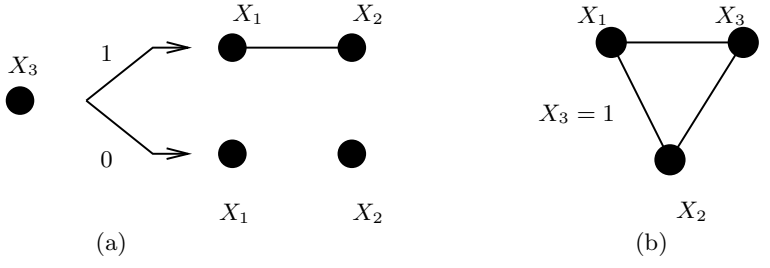


**Fig. 1.** Visual representations for the case $X_1 \perp X_2 | X_3 = 0$ and $X_1 \not\perp X_2 | X_3 = 1$.: (a) Split Model and (b) Generalized Graphical Model.

In this contribution we extend [10] and [12], following the original motivation for Graphical Models and using ideas proposed by e.g. [6]; this will lead to a class of graphs called *interaction graphs*. Our approach characterizes the interaction structures between multivariate variables at a finer scale including more information in the graph to facilitate model selection.

We consider the particular case of the multivariate binary stochastic variables because of its importance in many Computer Science applications.

In the sequel $X$ denotes a $d$-dimensional binary stochastic variable $(X_1, \cdots, X_d)$ where $X_i \in \{0, 1\}$ and we suppose that all cell probabilities $p_{x_1,\cdots,x_d} := P(X_1 = x_1, \cdots, X_d = x_d)$ are strictly positive. The vector with all cell probabilities will be denoted by $P$. The symbol '-' before an index of a vector refers to the exclusion of the corresponding entry: e.g. $X_{-i,-j} = \{X_k | k = 1, ..., d; k \neq i, j\}$. To refer simultaneously to a set $A$ of entries of a vector we will write $X_A$. Finally, if the conditional independence $X_i \perp X_j | X_{-i,-j} = x_{-i,-j}$ is valid for all values $x_{-i,-j}$, we write $X_i \perp X_j | X_{-i,-j}$.

The paper is organized as follows: Section 2 reviews briefly some aspects of dependencies between variables and of Graphical Models. Section 3 introduce Interaction Graphs and Section 4 shows two applications. Finally, Section 5 summarizes our main conclusions and discusses areas of further study.

## 2 Dependencies and Graphical Models

### 2.1 Odds-ratio

The *(log) odds-ratio* is a fundamental concept to quantify the dependency between binary variables and is the starting point of other important association measures like, e.g., Yule's $Q$ and $Y$ measure (see [2]). It is defined as:

$$\theta_{i,j}(x_{-i,-j}) = \frac{P(X_i=1,X_j=1,X_{-i,-j}=x_{-i,-j})P(X_i=0,X_j=0,X_{-i,-j}=x_{-i,-j})}{P(X_i=1,X_j=0,X_{-i,-j}=x_{-i,-j})P(X_i=0,X_j=1,X_{-i,-j}=x_{-i,-j})}.$$

In the binary case, the odds-ratio $\theta_{i,j}(x_{-i,-j})$ summarizes completely the conditional interaction between two variables:

$$X_i \perp X_j | X_{-i,-j} = x_{-i,-j} \text{ iff } \theta_{i,j}(x_{-i,-j}) = 1 \text{ iff } \log(\theta_{i,j}(x_{-i,-j})) = 0 \quad (1)$$

As will be shown in Section 3, for our purpose it is useful and illustrative to represent a discrete binary distribution as a hypercube where the vertices correspond to the cell probabilities. As shown in Fig. 2, each odds-ratio is in 1-1 correspondence with a particular face of the hypercube, defined by the cell probabilities associated with that face.

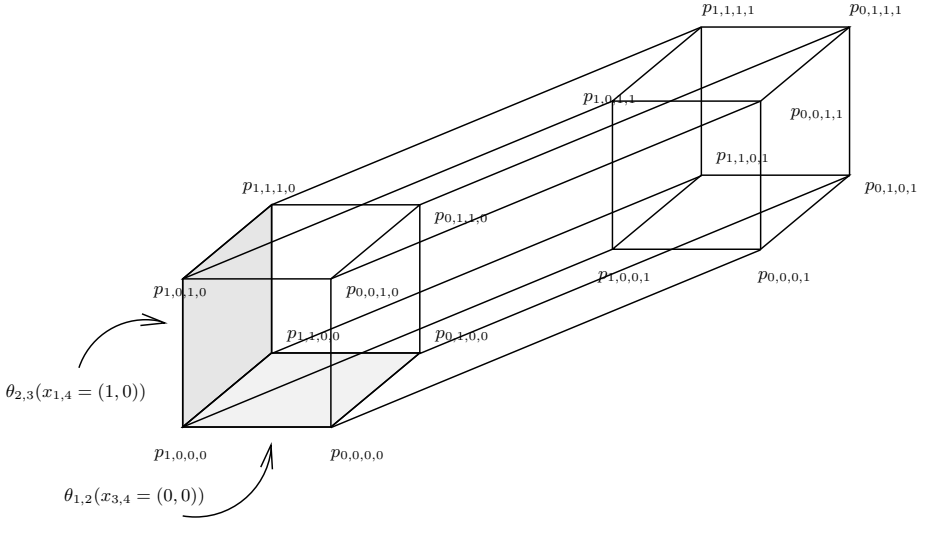


**Fig. 2.** Representation of a multivariate binary distribution ($d = 4$)

In the sequel, we will also make use of the fact that $\log(\theta_{i,j}(x_{-i,-j})) = 0$ imposes an orthogonality constraint of the vector $\log P$ to a vector $v$; e.g., if $d = 3$:

$$\log(\theta_{1,2}(x_3 = 0)) = 0 \;\Leftrightarrow\; X_1 \perp X_2 | X_3 = 0 \;\Leftrightarrow \tag{2}$$
$$(\log p_{1,1,0}, \log p_{0,0,0}, \log p_{1,0,0}, \log p_{0,1,0}, \log p_{1,1,1}, \log p_{0,0,1}, \log p_{1,0,1}, \log p_{0,1,1})$$
$$\perp (1, 1, -1, -1, 0, 0, 0, 0)$$

As the odds-ratios are related to each other, they do not constitute an independent set of parameters and we will need other parameterizations of $P$; probably the best known is the *Log-linear model* which we explain briefly in the next section.

### 2.2 Graphical Models

Graphical Models are a subclass of Log-linear models. The latter were introduced by Birch [1] as of particular way of parameterizing $P(X_1 = x_1, \cdots, X_d = x_d)$. The most generic form for the binary case is

$$\log P(X_1 = x_1, \cdots, X_d = x_d) = \sum_i \beta_i f(x_{A_i}) \;\text{ with } A_i \subset \{1, \cdots, d\} \tag{3}$$

or, equivalently written in vector form, for a fixed *design matrix* $\mathbb{A}$:

$$\log P = \mathbb{A} \cdot \beta. \tag{4}$$

Graphical Models are Log-linear models satisfying a set of independence statements $S$ of the form

$$X_i \perp X_j | X_{-i,-j}, \tag{5}$$

i.e., independencies which do not depend on the values of the variables in the conditional part.

Given $S$, a graph is constructed: each $X_i$ represents a node in the graph and the absence of an edge between node $i$ and $j$ corresponds to an independence of the form specified in Eq. (5). One tacitly assumes that all independencies of the form (5) that do not belong to $S$ are not present in the distribution (4). See Fig. 6 for an example. The graph is used as a summary of the main structure of the data.

One of the advantages of working only with independencies of the form (5) is that they traduce immediately in straightforward restrictions on which terms should enter in (3). A disadvantage is that this limitation often will be too restrictive, as shown in the following example.

To this end, consider the following data-set from a survey about the circumstances in which accidents occurred between American football players [4]. Table 1 shows the data-set for the following three variables: $X_1$ the accident occurred in a defensive play (value 0) or in an attack (value 1), $X_2$ the accident occurred while throwing the ball (value 0) or not (value 1) and $X_3$ the accident occurred in a tackle (value 0) or in a block (value 1). It can be shown that for this data-set the only Graphical Model that can be accepted corresponds to a complete connected graph (all hypothesis tests of the form (5) have

**Table 1.** Contingency table of accidents occurred to football players

| | $X_3 = 0$ | | $X_3 = 1$ | |
|---|---|---|---|---|
| | $X_2 = 0$ | $X_2 = 1$ | $X_2 = 0$ | $X_2 = 1$ |
| $X_1 = 0$ | 125 | 85 | 129 | 31 |
| $X_1 = 1$ | 216 | 62 | 61 | 16 |

a $p$-value $\leq 0.01$). Nevertheless, if we look at the estimated log odds-ratios and their 95% confidence intervals:

$\log(\theta_{1,2}(x_3 = 0)) = -0.86$ $[-1.26, -0.46]$ $\log(\theta_{1,2}(x_3 = 1))$= $\mathbf{0.09}$ $[\mathbf{-0.59}, \mathbf{0.76}]$
$\log(\theta_{1,3}(x_2 = 0)) = -1.3$ $[-1.67, -0.92]$ $\log(\theta_{1,3}(x_2 = 1))$= $\mathbf{-0.34}$ $[\mathbf{-1.03}, \mathbf{0.34}]$
$\log(\theta_{2,3}(x_1 = 0)) = -1.04$ $[-1.52, -0.56]$ $\log(\theta_{2,3}(x_1 = 1))$ = $\mathbf{-0.09}$ $[\mathbf{-0.71}, \mathbf{0.53}]$

we observe that some log odds-ratios (marked in bold) might be zero; hence some conditional independencies for the particular values of some variables might be plausible. As they cannot be captured by Graphical Models, the question raises how this might be included in those graphs. This will be explained in the next section.

## 3 Interaction Graphs

### 3.1 Visual Representation

The underlying idea of Interaction Graphs is to incorporate in a graph information about the (log) odds-ratios to reflect the dependency structure.

We start with a graph as used in a classical Graphical Model and subdivide the edge between each pair of nodes $i$ and $j$ in $l = 2^{d-2}$ segments, each one associated with a particular assignation of values to $X_{-i,-j}$. Fig. 3 shows an example of an edge segmentation in a data-set with five variables where the correspondence is made by means of a lexicographic ordering of the values of $x_{-1,-2}$.

An absence of a segment corresponds to an independence of the form as in Eq. (1). As illustrated in Fig. 4, the lexicographic ordering has the advantage that independencies of the form

$$X_1 \perp X_2 | X_{-1,-2,A} = x_{-1,-2,-A}, X_A, \tag{6}$$

correspond to easily detectable patterns as those shown in Fig. 4.

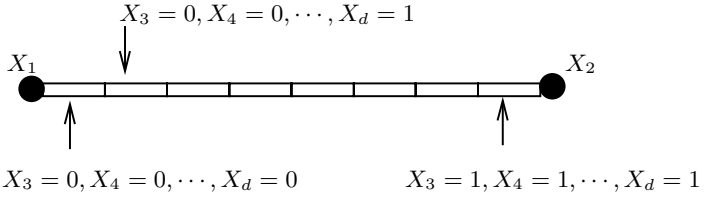


**Fig. 3.** Edge segmentation between variables for $X_1 \perp X_2 | X_{-1,-2} = x_{-1,-2}$ when $d = 5$


**Fig. 4.** Edge segmentation patterns in a data-set with five variables for the following conditional independence between $X_1$ and $X_2$, respectively: $X_1 \perp X_2 | X_5 = 0, X_{3,4}$; $X_1 \perp X_2 | X_4 = 1, X_{3,5}$ and $X_1 \perp X_2 | X_3 = 0, X_{4,5}$.

The graph will not only be used to summarize a given set of independencies but also to suggest new ones. To this end, we attach to the segments a color and make their width variable: the width is inversely proportional to the $p$-value of the test of the corresponding conditional independence using a chosen goodness-of-fit statistic. The segment is colored light gray if the corresponding $p$-value is above a user specified threshold, meaning that the segment may be taken out of the graph; otherwise, it is colored green or red depending whether the odds-ratio is significantly greater than or smaller than one, meaning a positive or negative correlation, respectively. See Fig. 5 for an example.

### 3.2 Estimation in Interaction Graphs

In order to adjust a particular model defined by a given set of independencies as in Eq. (1), we make use of the fact that each conditional independence Eq. (5) can be expressed as an orthogonality constraint of the vector $\log P$ to a vector $v_i$ as explained in Section 2.1. In this way, we obtain a set $V = \{v_1, \cdots, v_m\}$ to which $\log P$ should be orthogonal. Next, we calculate a basis for the space $V^{\perp}$. This is used to define the matrix $\mathbb{A}$ in Eq. (4).

Using a Newton-Raphson type optimization algorithm we calculate the maximum likelihood estimator for $\beta$ (taking also into account the normalization constraint on $P$). This value is used in a Power Divergence Goodness of Fit statistic [11] to calculate the $p$-value of the hypothesis that the given set of independencies hold.

### 3.3 Consistency Problem

Opposite to Graphical Models, not every graph corresponds to a valid model. This is a consequence of the odds-ratios being related to each other. In the multiplication of two odds-ratios whose corresponding faces in the hypercube (cf. Fig. 2) have an edge in common, the middle terms will cancel out, e.g.,

$$\theta_{1,2}(x_{3,4} = (0,0)).\theta_{2,3}(x_{1,4} = (1,0)) = \frac{p_{0,0,0,0}p_{1,1,0,0}}{p_{0,1,0,0}p_{1,0,0,0}} \frac{p_{1,0,0,0}p_{1,1,1,0}}{p_{1,1,0,0}p_{1,0,1,0}} = \frac{p_{0,0,0,0}p_{1,1,1,0}}{p_{0,1,0,0}p_{1,0,1,0}}.$$

As a consequence, if we repeat this for a whole sequence, eventually an odds-ratio can be expressed in terms of others. E.g., it is easy to check that:

$$\theta_{1,2}(x_3 = 0) = (\theta_{1,3}(x_2 = 1))^{-1}\, \theta_{1,2}(x_3 = 1)\, \theta_{1,3}(x_2 = 0).$$

This means that if the odds-ratios on the RHS are equal to 1, the LHS also equals one. In other words, independencies (associated to the RHS) will imply another one (associated to the LHS). Using a representation like in Fig. 2, the RHS corresponds to a loop of faces with the associated odds-ratios equal to 1, starting and ending on some face (whose odds-ratio corresponds to the LHS).

These relationships between the odds-ratios lead to a way to detect implied independencies (and hence inconsistencies) for a given set of independencies.

## 4 Examples of Using Interaction Graphs

Fig. 5 (a) shows an Interaction Graph for the saturate model of the data-set *Football Accidents*. The red segments indicate negative (conditional) correlations between the variables. Moreover, there are three gray segments suggesting that the following independencies are plausible:

$$X_1 \perp X_2 | X_3 = 1 \quad X_1 \perp X_3 | X_2 = 1 \quad X_2 \perp X_3 | X_1 = 1.$$

Fig. 5 (b) shows the final Interaction Graph.

The second example we consider is the data-set *Women and Mathematics* [5]. The data-set is based on a survey of 1190 New Jersey high school students about their interest in Mathematics; some of the students attended a conference that promotes Mathematics. The variables are: $X_1$ type of school (0=suburban, 1=urban), $X_2$ sex of the person (0=female, 1=male), $X_3$ assistance to the conference (0= yes, 1=no), $X_4$ Future Plans (0=University, 1=work), $X_5$ course

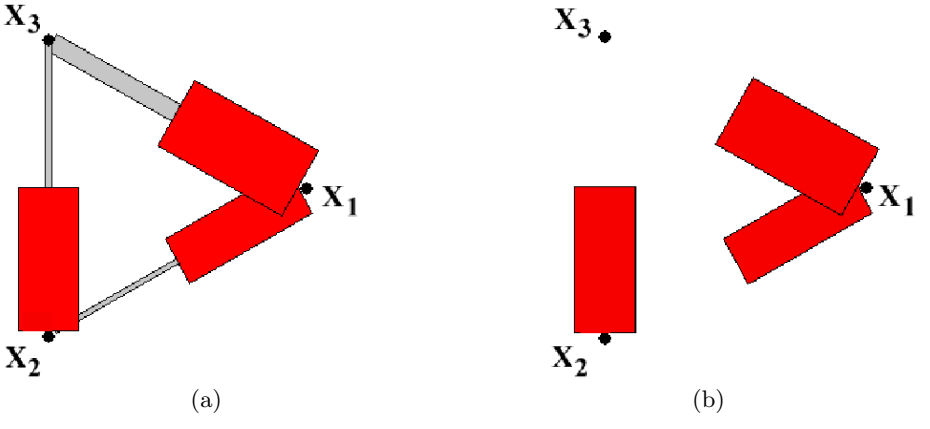


**Fig. 5.** Interaction Graph for the data-set *Football Accidents*: (a) saturated model and (b) final accepted model with a $p$-value of 0.28946

preferences (0=Mathematics, 1=Liberal Arts), $X_6$ necessity of Mathematics in the future (0=agree, 1=disagree).

Fig. 6 is a plausible classical Graphical Model (it has a $p$-value of 0.18). The graph shows e.g. that $X_3$ "assistance to the conference" is independent to all other variables. On the other hand, there seems to be a statistical significant dependence between e.g. $X_1$ "type of school" and $X_4$ "future plans", and between $X_6$ "necessity of Mathematics in the future" and $X_4$ "future plans". Fig. 7 shows an Interaction Graphs for these data; its $p$-value is 0.1228.

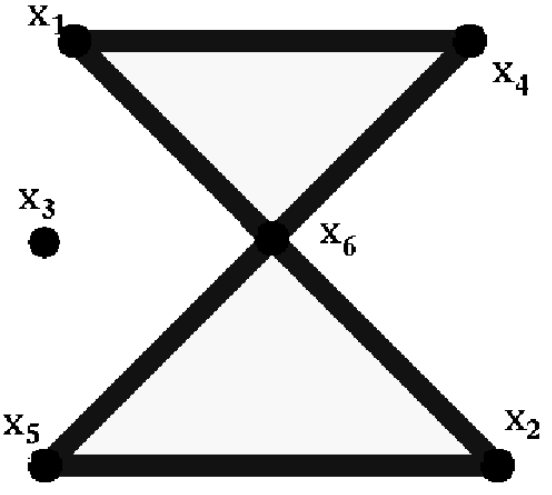


**Fig. 6.** Classical Graphical Model for data survey *Women and Mathematics*

Apart of the independencies from the previous graph, we learn in addition that the correlation between $X_1$ and $X_4$ is mainly negative. The dependency between $X_1$ and $X_4$ seems to depend on the value of a particular variable:

$$X_4 \not\perp X_6 | X_1 = 1, X_{2,3,5} \quad X_4 \perp X_6 | X_1 = 0, X_{2,3,5}.$$

Keeping in mind the patterns of Fig. 4, one discovers:

$$X_1 \perp X_6 | X_4 = 0, X_{2,3,5}$$
$$X_2 \perp X_6 | X_5 = 1, X_{1,3,4}$$
$$X_4 \perp X_6 | X_1 = 0, X_{2,3,5}$$

In practice, the software guides the user in identifying the corresponding variables. Observe that by weakening the independencies Graphical Models are working with, the set of potential distributions increases dramatically; this leads to a practical limit for the visualization.

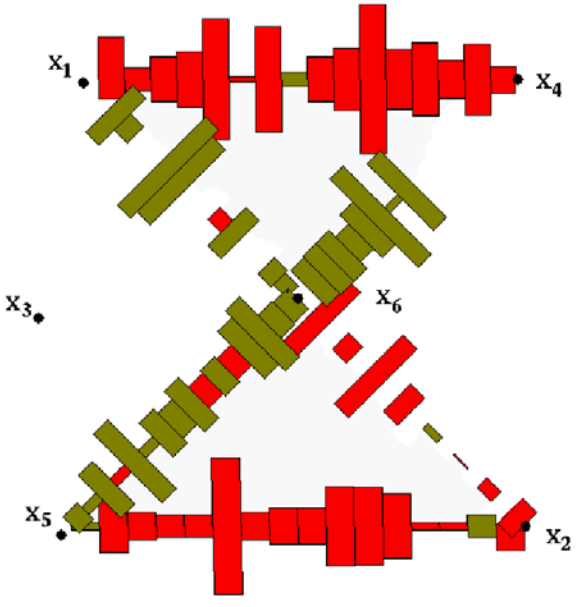


**Fig. 7.** Interaction Graph for data-set *Women and Mathematics* (its $p$-value is 0.1228)

## 5 Conclusions

In close connection with the basic philosophy of Graphical Models, we proposed in this paper a particular class of graphs that facilitate the study of conditional interaction structures between binary multivariate stochastic variables. To this end, we proposed the use of odds-ratios as the basic features of the underlying distribution. In the first place, the resulting graph pretends to offer an informative tool for exploratory purposes. A next step is to develop appropriate methods for probabilistic inference as required by e.g. probabilistic expert systems.

## References

1. Birch, M.W.: Maximum Likelihood in Three-way Contingency Tables. Journal of the Royal Statistical Society B 25, 220–233 (1963)
2. Bishop, Y., Fienberg, S., Holland, W.: Discrete Multivariate Analysis. MIT Press, Cambridge (1990)
3. Borgelt, C., Kruse, R.: Graphical Models: Methods for Data Analysis and Mining. John Wiley, Chichester (2002)
4. Buckley, W.: Concussions in Football: a Multivariate Analysis. American Journal of Sport Medicine 16, 609–617 (1988)
5. Fowlkes, E.B., Freeny, A.E., Landwehr, J.M.: Evaluating Logistic Models for Large Contingency Tables. JASA 83, 611–622 (1989)
6. Friendly, M.: Visualizing Categorical Data. SAS Institute, Cary (2000)

7. Højsgaard, S.: Split Models for Contingency Tables. Computational Statistics and Data Analysis 42, 621–645 (2003)
8. Højsgaard, S.: Inference in Graphical Gaussian Models with Edge and Vertex Symmetries with the gRc Package for R. Journal of Statistical Software 23 (2007)
9. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT Press, Cambridge (2009)
10. Navarro-Barrientos, J.E.: Modelos Gráficos Clásicos y Generalizados para el Análisis de Datos Binarios y su Aplicación en Datos de Accesos de Internet. M.Sc. Thesis CIMAT (2001)
11. Read, T., Cressie, N.: Goodness of Fit Statistics for Discrete Multivariate Data. Springer, Heidelberg (1988)
12. Teugels, J.L., Van Horebeek, J.: Generalized Graphical Models for Discrete Data. Statistics & Probability Letters 38, 41–47 (1998)
13. Wermuth, N., Lauritzen, S.: Graphical and Recursive Models for Contingency Tables. Biometrika 70, 537–552 (1983)
14. Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley Series in Probability and Mathematical Statistics (1989)