Stochastic Blockmodeling for the Analysis of Big Data

Gabriella Schoier(✉) and Giuseppe Borruso

DEAMS – Department of Economic, Business, Mathematic and Statistical Sciences "Bruno de Finetti", University of Trieste, Tigor 22, 34100 Trieste, Italy
{gabriella.schoier,giuseppe.borruso}@deams.units.it

Abstract. The aim of this paper is to consider the stochastic blockmodel to obtain clusters of units as regards patterns of similar relations; moreover we want to analyze the relations between clusters. Blockmodeling is a technique usually applied in social network analysis focussing on the relations between "actors" i.e. units. In our time people and devices constantly generate data. The network is generating location and other data that keeps services running and ready to use in every moment. This rapid development in the availability and access to data has induced the need for better analysis techniques to understand the various phenomena. Blockmodeling techniques and Clustering algorithms, can be used for this aim. In this paper application regards the Web.

Keywords: Blockmodeling · Gibbs sampling · Latent class model · Clustering algorithms · Big data

1 Introduction

Stochastic blockmodeling is a technique often used in social network analysis for studying the relationship between two factors. Its goal is to cluster objects with respect to some given observed variables using the existing relationships between actors.

The clustering problem has been considered in many contexts and by researchers in different disciplines. It is useful in several exploratory pattern-analysis, grouping, decision-making and machine-learning situations, including data mining web mining and spatial data mining.

Cluster analysis can be defined as the organization of a collection of patterns - usually represented as a vector of measurements, or a point in a multidimensional space - into clusters based on similarity [7, 11, 19, 25].

Unlike cluster analysis, which subdivides the elements of a network into groups based on the study of the properties and characteristics of the network units (actors in Social Analysis framework), blockmodel, allows the formation of homogeneous groups based on the study of the relationships existing between the actors of the network itself [1].

Different approaches have been developed in stochastic blockmodeling. In this paper we have applied the one proposed by Nowicki, Snijders and others [4, 9, 13–15, 17, 18] in order to study the navigational patterns through a website [11].

https://doi.org/10.1007/978-3-030-24302-9_33

In our analysis, the clusters, called colours, are formed by groups of units, which are the viewed web pages.

It is interesting to notice that other different applications in a big data and in a spatially big data context can be considered as the computational aspects and the visual representation are attractive tools.

Specifically, we have applied blockmodeling to establish three relationships between viewed web pages.

These relationships are: having common users, sharing the same (recoded) time in the pages and having same value as regards the impressions (that is the number of times the page has been viewed during the session divided by the number of viewed pages); they are induced directly by navigation itself, and they in turn reflect users browsing behaviour during the navigation.

Clustering the shared common web pages among users allows us to gather useful information different from, for instance, clustering of web users.

The advantage of blockmodeling is that a differentiated structure for the degree of similarity within and between clusters is allowed. To explain this an example on a web site of a portal for children is considered.

The paper aims to divide a set of Web pages into homogeneous groups on the base of three known relationships existing between the pages. The peculiarity of this application is the simultaneous study of several relationships and the fact that the input data are not of dichotomous type but can assume a wider range of values. As regards the study of the three relations, Ucinet [2, 3] and Stocknet (in particular the Block module [10, 16, 17] programs have been used.

2 The Methodology: Blockmodel and Stochastic Blockmodel

Graphs and oriented graphs have been used as mathematical models for social and physical phenomena where the relationships between the various units are known.

Two important types of graph models (oriented or not) are the blockmodel and the stochastic models. An integration of these two approaches has been proposed by Wassermann [22–24].

Blockmodels allow to elaborate the results of a positional analysis providing a simplified representation of the links and interactions present in a complex social network.

First it is necessary to briefly explain what a positional analysis means. Starting from a multirelational set, the final aim of a positional analysis is to group the "actors" in positions, so that individuals who can be considered similar, according to a strict definition, are inserted in the same block, obtaining a complete partition of all the actors belonging to the social network.

It is also necessary to give a definition of "role" and "position", keywords in a positional analysis. "Position" is a set of individuals that interact from and to other actors in the social network in the same way. "Role" is a system of associations between relations between individuals or between positions.

A blockmodel consists of a description of how the actors are assigned to the positions, *i.e*. a partition of the actors in a discrete number of subsets, the "positions",

one or more image matrixes depending on the number of relations considered which indicate the presence or absence of a link for each pair of positions considered. The image matrix (one for every relation) is a square matrix whose size is the number of positions of the social network. It refers to positions and not to single individuals.

Let us consider a set of R binary sub-matrices defined on n actors belonging to N, which describe the relationships existing between individuals; there are as many submatrices as the considered relationships.

Consider $C_1, \ldots, C_c$, with $c < n$, an exclusive and exhaustive partition of N in c positions and consider the map function Φ (•) such that Φ (i) = C_k indicates that the actor i belongs to the class C_k.

Let us consider the relationships, no longer between the single actors (units), but between the positions through a matrix C (image matrix) whose elements c_{klr} can assume value 1 (oneblock) or 0 (zeroblock) depending if the relation r between the positions k and l exists or not.

A blockmodel is therefore a matrix of size *(cxcxR)* with values 1 and 0. It is the result of an empirical procedure based on the idea that units in a network can be grouped into equivalent sets, under a given definition of equivalence.

On the base on the type of the considered equivalence, a distinction is made between deterministic and stochastic blockmodel.

The deterministic blockmodel is based on the concept of structural equivalence: two actors are defined equivalent if they perfectly possess the same relational ties [11].

This approach has the disadvantage of not using statistical tests to determine how well the blockmodel adapts to real values. To overcome this problem, a stochastic approach has been developed, precisely what we will consider [1, 21].

In the stochastic blockmodel a stochastic equivalence is considered:

Two actors i and i' are stochastically equivalent if the probability that i is in relation (to and from) with every other actor is the same also for the actor i' *i.e.* if the probability of an event concerning X does not change by substituting i with i'.

We have to notice that: structural equivalence => stochastich equivalence (but *the vice versa* is not true).

As mentioned above a blockmodel is formed by a probability distribution *p(x)* and a map function Φ.

Depending on how the map function is found the stochastic blockmodel is distinguished between:

- *a priori blockmodel* where it is assumed that the map function is previously known and that it depends on exogenous characteristics of the actors in relation to the studied relations,
- *a posteriori blockmodel* where the map function is the result of the application of the data on the relations.

In general different approaches can be applied.

- Approach on the base of the *p1* model (Wassermann and Anderson [24] and Wasserman, Anderson and Faust [23]).

The stochastic blockmodel can be defined as a probability distribution (or a family of distributions) for graphs (oriented or not) in which the vertices set is divided into

subset called blocks (or colors), which satisfy the property that the distribution of probability of the graph remains unchanged following permutations of the vertices within the block to which they belong. The probability that a bond is present between two vertices depends only on the color of the vertex, i.e. the block to which it belongs. Two vertices belonging to the same block are called stochastically equivalent.

Depending on whether the attributes of the vertices and therefore the blocks are known or not, we speak of a priori or a posterior stochastic blockmodel. The latter is much more complex than the former.

Wassermann and Anderson [24] and Wasserman, Anderson and Faust [23] have studied the a posterior blockmodel with respect to the *p1* family. This is a log-linear exponential family of probability distribution for graphs.

In a first phase the vertices are "blocked" through an ML estimation of the vertices parameters themselves, then they are grouped on the basis of multiple comparisons of the estimated parameters, i.e. the vertices that have approximately similar estimates of the two parameters considered (productivity parameter and popularity parameter) are put in the same group.

In general, the *p1* model in statistical inference has the problem that the number of parameters increases with the increase in the number of vertices; this problem is solved with the combination of p1 and blockmodel, since even if the number of vertices increases, the number of blocks remains unchanged, as instead of considering the single vertices we consider the blocks.

Another disadvantage of the posterior blockmodeling based on the p1 model derives from its too restrictive nature. In fact, in the p1 model the vertices having a high productivity parameter are relatively more likely to have outbound links, i.e. to other vertices (→), while vertices with a high popularity have a high probability of having inbound links, *i.e.* from other vertices. This excludes the important case of oriented graphs with vertices classes where the density of relations is elevated within the class and low among different classes.

– Bayesian approach (Snjders and Nowicki [17]).

This is a more generic approach than the previous one, because it is not related to the *p1* function.

Each vertex of the observed graph belongs to a block, however the structure of the blocks is not observed. Moreover the relationships are independent, conditioned only by the block structure.

In particular, two methods are considered:

– in the case of a graph with few vertices (<20) we can use both the method of maximum likelihood (ML estimation) and the Bayesian estimates implemented using Gibbs sampling,
– in the case of a higher number of vertices, even if very high, the Bayes method can be used.

3 The Application

3.1 The Data and the Preliminary Phase

The objective of this application is to use the posteriori stochastic blockmodeling according to a Bayesian approach developed by Snijders and Nowicki adapting it to the case of an analysis based on the study of several relationships observed on the same set of actors.

The environment in which this analysis has been developed is the Web Mining [8, 20]. In particular, an analysis of three relationships between users and Web pages has been considered. The objective has been to divide the various pages into groups whose elements are considered stochastically equivalent [12].

The analysis regards the log files of the web site www.girotondo.com, a portal for children. In this site there are seven different sessions: *Bacheca*, *Corso*, *Favolando*, *Giochi*, *Links*, *News*, *Percome*, it has 362 jhtml pages.

The period of observation is from the 29/11/2000 to the 18/01/2001. The original file contained 3000,000 records. Record of log files containing information about any object (with .gif, .jpeg., etc. extension) that is not its Internet address are cancelled.

The log-file information taken into account in the analysis concerns:

- IP address and page visited, that is if the user having IP address i has visited or not the page j of the site (data expressed in a dichotomous form),
- time spent by user *i* on page *j*,
- impressions, *i.e.* the ratio between the number of times a page has been viewed in a session and the total number of pages viewed within the session.

In this way we obtain a file indicating the Internet address for every visited page.

We have proceeded into a recodification of the Web pages transforming their URL into a number in order to handle them easily, in so doing 117 pages have been considered. After the pre-processing of the data we have obtained a file with 10,000 records. The data considered consist of a finite set of vertices (visited pages) on which $R = 3$ relational variables (having more than one user in common, having users which stay the same interval of time, having users with the same value for the impression) are measured; this is a *network N* (set of units and relation(s) defined over it). These variables are collected into three sets of (10000×117) matrices $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ called sociomatrices which represent three 2-mode networks (users x pages).

The first problem that had to be considered is related to the determination of the inputs required by Stocnet and in particular the applicative Block [17], *i.e.* a single network representing a matrix of adjacencies, therefore of one-mode type. For this purpose, the free Ucinet program has been used [2].

Once the three matrices have been obtained and precisely one matrix for visited pages, one for times and one for the impressions they have had to be re-coded.

At this point these three rectangular matrices which represent three 2-mode networks have been changed into three square matrices representing three 1-mode networks (pages x pages).

An 1-mode matrix is a matrix in which both the rows and the columns refer to the same set of objects (vertices), while in a 2-mode matrix the rows and columns refer to two sets of different objects, in our case respectively IP addresses and web pages.

Ucinet allows to pass from a 2-mode matrix to a 1-mode matrix, that is a matrix called *actor-by-actor* that counts the number of events that each pair of actors has in common, or an *event-by-event* matrix, as in our case, which counts the number of actors accessing both pages, or, in general, both event.

It has been decided to recode the matrices after each Ucinet application. So after transforming the 2-mode matrices into 1-mode matrices these have been made dichotomous; in particular, for the time and impression matrices, a value of 0 was assigned for the values included in the interval [0, 10] and value 1 for the elements > 10. As far as regards the third matrix a distinction between cases in which two pages have 0 or 1 users in common, to which the value 0 has been assigned, and those with more than one user in common recoded with the value 1 has been considered.

Ucinet allows to have multiple matrices within the same data-file. It is possible to insert each matrix in a different spreadsheet sheet. In our case we have got a spreadsheet with three sheets.

Finally the three matrices have been aggregated in one matrix. Aggregation means the transformation of more than one matrix into one that contains their characteristics.

Ucinet allows different aggregation methods. The one used in this case is the procedure that assigns a single value to each unique combination of values between the relations. A single network file has thus obtained to be used as input in the Stocnet application. To be able to use the matrix in Stocnet you have to delete the labels and transform the file into text format. From now on, therefore, the various pages will no longer be indicated by their name, but by numbers.

3.2 Stochastic Blockmodel Application

A stochastic blockmodeling technique has been applied using the free program Stocknet that gives a graphical interface for different modules (the one for stochastic block modeling is Block). In Block we have recodified the values of the input matrix so to have four values: 0, 1, 2, 3.

A Blockmodeling allows to describe and to interpret a dataset through a block structure so to give a simplified representation of the existing ties and relation(s).

The primary tool of this technique is the *blockmodel* which, in our case, consists of a mapping of approximately equivalent 117 units or vertices (in our analysis visited pages) into discrete subsets called blocks and a statement regarding the relations between the positions or clusters or colours (in our case the three relations).

This represents a partition of the vertices into blocks and a mapping function $\Phi(\cdot)$ which describes the subdivision of the vertices through the positions.

One of the main procedural goals of *blockmodeling* is to identify clusters of units that share structural characteristics defined in terms of the *R* relations where the optimal value of a criterion function has to be found. The criterion function can be constructed indirectly as a function of compatible (dis)similarity measure between pairs of units (see e.g. for the case of Web data [11]), or directly as a function measuring the fit of a clustering to an ideal one with perfect relations within each cluster and between

clusters. In this paper we will consider this second possibility, moreover the approach we will adopt is the stochastic.

Let us consider the matrix $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, it is called super-sociomatrix. The probability distribution for $\boldsymbol{X}$, gives the probability that various relational linkages between actors across all relations are equal to a value $\boldsymbol{x}$.

A stochastic blockmodel is based either on the probability distribution for $\boldsymbol{X}$ and on the mapping function $\Phi(\cdot)$. There are two categories of stochastic blockmodels: *a priori* and *a posteriori*.

In this paper we will consider the *a posteriori* stochastic blockmodeling structure proposed by Snijders and Nowicki, (see these article for a more detailed description).

Assume that a set of 117 vertices is given; this is divided in B positions, classes or *colours*. The colours assumed by the vertices are described by the attribute vector: $\mathbf{s} = (s_1, \ldots\ldots, s_{117})$, where s_i is the attribute of the i-th vertex. Conditional on the vector $\boldsymbol{s}$, $\boldsymbol{x}$ is modelled. The model can be regarded as a mixture model.

The stochastic blockmodel is given by the joint distribution of ($\boldsymbol{X}$, $\boldsymbol{S}$). In terms of cluster analysis the fact that the heterogeneity is modelled by stochastic membership of the classes makes it analogous to a mixture rather than a discrete classification model. In input there is the matrix of the relations, in output a partition of the vertices in classes such that all the vertices belong to a certain class have the same probability of having a certain relation with other vertices belonging to other classes.

The final conclusions consist on the probability for the vertices to belong to a certain group and in the probability distribution of the relation, these estimations have been obtained with the Gibbs sampling [5, 6]. A non informative prior has been chosen.

In order to assess the suitableness of the adopted model Snijders and Nowicki [16] present two indices I_d H_s (see [9] for the formulas).

After a preliminary analysis (based on 50,000 iterations) on the base of the results of the indices I_d H_s presented in Table 1 the analysis has been fixed on 5/6/7/8 classes; for each of them a Gibbs sampling of 10,000 iterations have been performed.

Table 1. Different partitions

B = classes	Gibbs sequences	I_d	H_s
B = 5	1	0.2287	0.0438
B = 6	1	0.2190	0.0379
B = 7	1	0.1667	0.0277
B = 8	1	0.1414	0.0166

The two indices, estimated by the posterior mean, may give different solutions, in this case Nowicki and Snijders suggest to give more importance to the conclusions derived by index H_s.

The result are reported in Table 1, on the base of these data a partition of eight colours has to be preferred.

In more detail we can see from Table 2 the pages belonging to different colours/blocks:

Table 2. Colours/blocks

COLOUR 1	COLOUR 2	COLOUR 4	COLOUR 6
PERCHE'	AIUTO\AIUTO	BAZAR\MYCOMPUTER\01	FAVOLANDO\03_00\01
PRESENTAZIONE	BACHECA\BACHECA	CORSO\FATTO\01	FAVOLANDO\04_00\01
SICURO	FAVOLANDO\07_00\02	CORSO\FATTO\03	FAVOLANDO\06_00\01
FAVOLANDO\04_00\02	FAVOLANDO\08_00\03	CORSO\NAVIGA01	FAVOLANDO\06_00\02
FAVOLANDO\06_00\03	FAVOLANDO\09_00\02	GIOCHI\GIOCHI\09_00\GIOCHI2	FAVOLANDO\07_00\01
GIOCHI\GIOCHI\08_00\GIOCHI2	FAVOLANDO\11_00\02	GIOCHI\GIOCHI\09_00\GIOCHI3	FAVOLANDO\07_00\03
GIOCHI\GIOCHI\09_00\GIOCHI4	MAPPA\01	GIOCHI\GIOCHI\09_00\SCARICA \REBUS\REBUS2	FAVOLANDO\07_00\04
LINKS\02		GIOCHI\GIOCHI\12_00\GIOCHI2	FAVOLANDO\08_00\01
LINKS\07_00\02	COLOUR 3	LINKS\ARCHIVIO	FAVOLANDO\08_00\03
LINKS\09_00\02		LINKS\LINKS	FAVOLANDO\08_00\04
LINKS\09_00\022	BAZAR\01	LINKS\03_00\01	FAVOLANDO\09_00\01
LINKS\11_00\011	GIOCHI\GIOCHI08_01\GIOCHI1	LINKS\06_00\011	FAVOLANDO\09_00\03
LINKS\11_00\02	GIOCHI\GIOCHI\09_01\GIOCHI1	LINKS\07_00\01	FAVOLANDO\09_00\04
NEWS\03_00\011	LINKS\12_00\022	LINKS\07_00\022	FAVOLANDO\10_00\01
NEWS\04_00\01		LINKS\12_00\011	FAVOLANDO\10_00\02
NEWS\04_00\022	COLOUR 5	NEWS\02	FAVOLANDO\10_00\03
NEWS\07_00\02		NEWS\03_00\01	FAVOLANDO\11_00\01
NEWS\08_00\011	FAVOLANDO\ARCHIVIO	NEWS\05_00\01	FAVOLANDO\11_00\03
NEWS\09_00\011	FAVOLANDO\05_00\01	PERCOME\02	FAVOLANDO\11_00\04
NEWS\12_00\01 *	FAVOLANDO\05_00\02	PERCOME\03_00\01	FAVOLANDO\11_00\05
PERCOME\03_00\022	FAVOLANDO\12_00\01	PERCOME\03_00\02	FAVOLANDO\11_00\06
PERCOME\05_00\02	FAVOLANDO\12_00\02	PERCOME\LIBRINEWS\LIBRO	FAVOLANDO\12_00\04 *
PERCOME\07_00\02	FAVOLANDO\12_00\05	PERCOME\LIBRINEWS\10_00\01	FAVOLANDO\12_00\08
PERCOME\08_00\022	FAVOLANDO\12_00\06		COLOUR 7
PERCOME\09_00\02	FAVOLANDO\12_00\07	COLOUR 8	
PERCOME\09_00\022	GIOCHI\GIOCHI\01_01\GIOCHI1		FAVOLANDO\05_00\03
PERCOME\10_00\02		LINKS\01	FAVOLANDO\05_00\05
PERCOME\10_00\022		NEWS\01	FAVOLANDO\10_00\04
PERCOME\11_00\02		PERCOME\01	FAVOLANDO\10_00\06
PERCOME\11_00\022		GIOCHI\GIOCHI\05_00\GIOCHI1	FAVOLANDO\12_00\09
PERCOME\LIBRINEWS\10_00\01		GIOCHI\GIOCHI\11_00\GIOCHI1	GIOCHI\GIOCHI\04_00\GIOCHI1
		GIOCHI\GIOCHI\12_00\GIOCHI1	GIOCHI\GIOCHI\06_00\GIOCHI1
			GIOCHI\GIOCHI\07_00\GIOCHI1

As one can see from Table 2 we have obtained eight colours whose pages can be summarized in the following scheme:

(1) COLOUR 1/block1: in this cluster, the most numerous, the pages are obtained through a random navigation, it contains some pages of the sections *News*, *Links* and *Percome* and a few pages of the section *Links*,

(2) COLOUR 2/block2: in this cluster there are the initial pages of two sections: *Favolando* and *Bacheca* and the help,

(3) COLOUR 3/block3: in this cluster there are the other pages of the section *Links* it is not homogeneous,

(4) COLOUR 4/block4: in this cluster there are pages of the sections *Corso*, *Giochi* and *Links*,

giuseppe.borruso@deams.units.it

(5) COLOUR 5/block5: in this cluster there are the pages of the section *Favolando* and one page of *Giochi,*
(6) COLOUR 6/block6: in this cluster there are pages of the sections *Favolando*, *Links* and *Percome*,
(7) COLOUR7/block7: in this cluster there are the most recent Web pages of the site, the last published tale and the last game introduced,
(8) COLOUR 8/block8: in this cluster there are a few pages of section *Giochi*, and only one pages of the sections *News*, *Links* and *Percome.*

The positions and the relational ties between positions for a stochastic blockmodel need to be represented in order to interpret the model; there are two common way of representation: *density tables* and *reduced graphs*.

Density tables contain the probabilities that vertices relate to and are related to by other vertices when the vertices are in the same or different positions, each row and column of these tables correspond to a position. Reduced graphs give a graphical representation of the situation.

In Table 3 the density table represents the final probabilities estimated for the case of the multiple relation 2 = (1,1) which can be summarized as: "the pages share a few users that stay a little on them and don't visit them a lot of times during the session".

These probabilities are high in the case in which both the pages belong to the same colour except for block (3), (4) and (7), the relation is sure when: one page has colour (1) and the other has colour (2), (3) or (8), one page has colour (2) and the other has colour (5), (6) or (8), one page has colour (5) and the other colour (6). Ambiguous situations regards one page belonging to colour (7) and the other to colour (8).

Table 3. Density table

	1	2	3	4	5	6	7	8
1	0.99	0.96	0.95	0.02	0.02	0.00	0.01	0.98
2	0.96	0.76	0.60	0.10	0.93	0.98	0.13	0.91
3	0.95	0.60	0.17	0.14	0.14	0.02	0.09	0.83
4	0.02	0.10	0.14	0.06	0.04	0.00	0.06	0.18
5	0.02	0.93	0.14	0.04	0.88	0.98	0.85	0.85
6	0.00	0.98	0.02	0.00	0.98	0.99	0.13	0.05
7	0.01	0.13	0.09	0.06	0.85	0.13	0.38	0.50
8	0.98	0.91	0.83	0.18	0.85	0.05	0.50	0.85

The results may be viewed by the reduced graph in Fig. 1 which consists of nodes corresponding to the positions and lines or arcs corresponding to the relations.

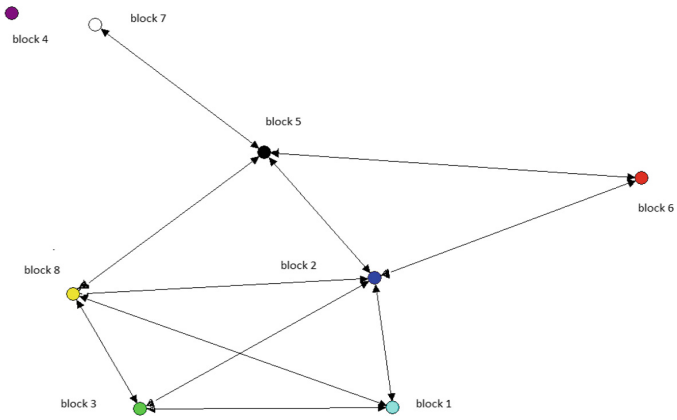

Fig. 1. Reduced graph

In Table 4 the density table represents the final probabilities estimated for the case of the multiple relation 1 = (0, 0) which can be summarized as: "the pages share no users that stay a little on them and don't visit them a lot of times during the session".

In this case we have found a certain ambiguity as regards the probability of relation between group 7 and group 8 (0, 46) and between group 3 and 5 (0, 76). As one can see from Table 4.

Table 4. Density table

	1	2	3	4	5	6	7	8
1	0.00	0.01	0.01	0.98	0.97	0.99	0.98	0.01
2	0.01	0.04	0.03	0.88	0.01	0.01	0.84	0.02
3	0.01	0.03	0.09	0.83	0.76	0.96	0.84	0.03
4	0.98	0.88	0.83	0.93	0.95	0.99	0.93	0.80
5	0.97	0.01	0.76	0.95	0.02	0.00	0.13	0.12
6	0.99	0.01	0.96	0.99	0.00	0.00	0.86	0.94
7	0.98	0.84	0.84	0.93	0.13	0.86	0.57	0.46
8	0.01	0.02	0.03	0.80	0.12	0.94	0.46	0.05

The results may be viewed by the reduced graph in Fig. 2 which consist of the nodes corresponding to the positions and lines or arcs corresponding to the relations.

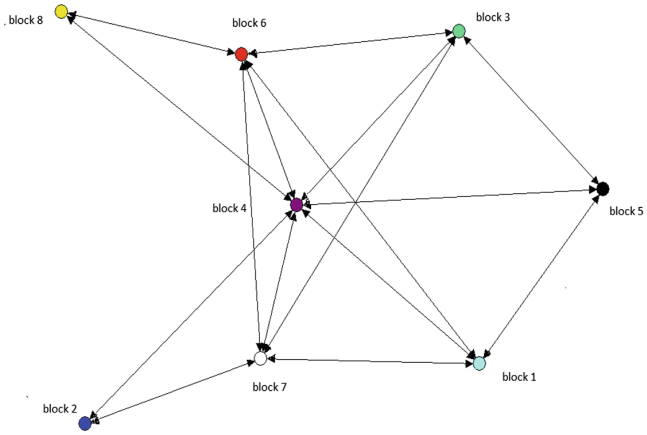

Fig. 2. Reduced graph

In Tables 5 and 6 the density table representing the final probabilities estimated for the case of the multiple relation 3 = (2, 2) which can summarise as: "the pages share enough users that stay a certain time on them and visit them enough times during the session" and for the multiple relation 4 = (6, 6) which can be summarized as: "the pages share many users that stay a lot on them and visit them a lot of times during the session".

As one can see from the density tables (Tables 5 and 6) these two relations do not present interesting relations among blocks.

Table 5. Density table

	1	2	3	4	5	6	7	8
1	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.01
2	0.03	0.16	0.34	0.01	0.01	0.01	0.01	0.02
3	0.03	0.34	0.65	0.01	0.07	0.01	0.05	0.03
4	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
5	0.00	0.04	0.02	0.02	0.02	0.00	0.01	0.02
6	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
7	0.00	0.01	0.05	0.01	0.01	0.00	0.00	0.02
8	0.01	0.02	0.03	0.02	0.02	0.01	0.01	0.05

Table 6. Density table

	1	2	3	4	5	6	7	8
1	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
2	0.00	0.04	0.03	0.01	0.01	0.01	0.01	0.02
3	0.01	0.03	0.09	0.02	0.07	0.01	0.05	0.03
4	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.01
5	0.00	0.01	0.07	0.00	0.02	0.00	0.01	0.02
6	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
7	0.00	0.01	0.05	0.00	0.01	0.00	0.03	0.02
8	0.01	0.02	0.03	0.01	0.02	0.01	0.02	0.05

The advantage of stochastic blockmodeling with respect to a classical cluster analysis is clear: a differentiated structure for the degree of similarity within and between the clusters is allowed. Different clusters present low internal similarity; moreover, the pattern between clusters similarities is interesting and varied; such results are more difficult to obtain with classical cluster analysis.

4 Conclusions

In this paper we present the Bayesian analysis based on the program Stocknet (module Block) and applied it in the case of Web Mining. We obtained an useful tool to understand how an user navigates through the site, which pages are more attractive and which are less interesting. For this reason it seems to be a potential tool especially if applied in the case of commercial sites.

The results presented in the previous section show that stochastic blockmodeling may be useful in order to improve the comprehension of different problems, for instance in this application we consider the type of behaviour of the users of a site.

It is interesting to notice that other different applications in a big data and in a spatially big data context can be considered as the computational aspects and the visual representation of this methodology are attractive tools.

References

1. Anderson, C.J., Wasserman, S., Faust, K.: Building stochastic blockmodels. Soc. Netw. **14**, 137–161 (1992)
2. Borgatti, S.P., Everet, M.G.: Freeman: UCINET for windows software for social network analysis harvard: analytic technologies. http://www.analytictech.con/ucinet_5_description.htm
3. Borgatti, S.P., Foster, P.C.: The network paradigm in organizational research: a review and typology. J. Manag. **29**, 991–1013 (2003)
4. Burk, W.J., Steglich, C.E.G., Snijders, T.A.B.: Beyond dyadic interdependence: actor-oriented models for co-evolving social networks and individual behaviors. Int. J. Behav. Dev. **31**, 397–404 (2007)
5. Gelman, A., Carlin, J.B., Stern, H., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall, London (1995)
6. Gilks, W.R., Richardson, S., Spiegelhalter, D.: Markov Chain Monte Carlo in Practice. Chapman and Hall, London (1996)
7. Jan, A.K.: Data clustering. 50 years beyond K-means. Pattern Recogn. Lett. **31**, 651–666 (2010)
8. Mobasher, B., Doi, H., Luo, T., Nakagawa, M., Sung, Y., Wiltshire, Y.: Discovery of aggregate usage profiles for web personalization. Conference on Knowledge Discovery in Databases (2000). http://www.maya.cs.depaul.edu/∼mobasher/personalization
9. Nowicki, K., Snijders, T.A.: Estimation and prediction for stochastic blockstructures. J. Am. Stat. Assoc. **96**, 1077–1087 (2001)
10. Ripley R., Snijders T.A., Boda, Z., Vörös, A., Preciad, P.: Manual for SIENA version 4.0. Department of Statistics, University of Oxford, Oxford (2017). http://www.stats.ox.ac.uk/∼snijders/siena/

11. Schoier, G., Borruso, G.: A methodology for dealing with spatial big data. Int. J. Bus. Intell. Data Min. **12**(1), 1–13 (2017)
12. Schoier, G.: Blockmodeling techniques for web mining. In: Haerdle, W., Roenz, B. (eds.) Compstat. Springer, Berlin (2002). https://doi.org/10.1007/978-3-642-57489-4_26
13. Snijders, T.A.: Stochastic actor-oriented models for network dynamics. Annu. Rev. Stat. Appl. **4**, 343–363 (2017). https://doi.org/10.1146/annurev-statistics-060116-054035
14. Snijders, T.A., Lomi, A., Torló, V.: A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. Soc. Netw. **35**, 265–276 (2013)
15. Snijders, T.A., van de Bunt, G.G., Steglich, C.E.G.: Introduction to stochastic actor-based models for network dynamics. Soc. Netw. **32**, 44–60 (2010)
16. Snijders, T.A., Nowicki, K.: Manual for blocks (2001)
17. Snijders, T.A., Boer, P., Huisman, M., Zeggelink, E.P.H.: StOCNET: an open software for the advanced statistical analysis of social networks (2001)
18. Snijders, T.A., Nowicki, K.: Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J. Classif. **14**, 75–100 (1997)
19. Steinbach, M., Ertöz, L., Kumar, V.: The challenges of clustering high dimensional data (2003). http://www-users.cs.umn.edu/~kumar/papers/high_dim_clustering_19.pdf
20. Srivastava, J., Colley, R., Deshpande, M., Ton, P.: Web usage mining: discovery and applications of usage patterns from web data (2000). http://www.maya.cs.depaul.edu/~mobasher/personalization
21. Wang, Y.J., Wong, G.Y.: Stochastic blockmodels for directed graphs. J. Am. Stat. Assoc. **82**, 8–18 (1987)
22. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)
23. Wasserman, S., Faust, K.: Blockmodels: interpretation and evaluation. Soc. Netw. **14**, 5–61 (1992)
24. Wasserman, S., Anderson, C.: Stochastic a posteriori blockmodels: construction and assessment. Soc. Netw. **9**, 1–36 (1987)
25. Xu, R., Wunsch II, D.: Survey of clustering algorithms (2005) http://ieeexplore.ieee.org/iel5/72/30822/01427769.pdf