

Ideal-Type Model for Random Networks

by

Jean-Jacques Daudin and Laurent Pierre



Research Report No. 22
January 2009

STATISTICS FOR SYSTEMS BIOLOGY GROUP
Jouy-en-Josas/Paris/Evry, France
<http://genome.jouy.inra.fr/ssb/>

IDEAL-TYPE MODEL FOR RANDOM NETWORKS

BY JEAN-JACQUES DAUDIN, LAURENT PIERRE

UMR AgroParisTech/INRA518 and University Paris X

E-MAIL: jean-jacques.daudin@agroparistech.fr

Abstract A new model for heterogeneous random graphs is presented with an algorithm for obtaining the maximum likelihood estimates of the parameters. This model is related to the concept of Ideal-Type in social science, but is general and applicable to any network.

1. Introduction. Data sets giving not only information about items but also information about the relation between them are more and more studied in different domains such as social sciences and biology. The data size is proportional to the square of the number of individuals, so that it is necessary to summarize the information in a simpler form. The network representation of the data is graphically attractive, but is not readable for $n > 100$. There are two ways for producing a synthetic representation of such data: multidimensional scaling where position in a metric space is assigned to each item (Handcock et al., 2007, (2)), and clustering of the items using a mixture model (Nowicki and Snijder, 2001 (3) and Daudin, 2008 (1)). In this paper we present a new method, which has some flavor of mixture and some flavor of multidimensional scaling but is really different of both. We restrict our interest to the case of pure relational information, putting aside any information on items. The intensity of relation may be continuous or binary. We restrict our interest on the binary case. These two restrictions are made for sake of simplicity. The model we propose may be extended to the general case, but this is not done in this paper.

We define the Ideal-Type Model (IDTM) in Section 2. In Section 3, we give a maximum-likelihood estimation algorithm. Some simulations are provided in section 4 and an example is studied in Section 5.

2. Ideal-Type Model.

2.1. Model IDT. Vertices Consider a graph with n vertices, labeled in $\{1, \dots, n\}$. The model is based on Q hypothetical unobserved Ideal-Type vertices. Ideal-Type, also known as pure type or Idealtyp in the original German, is a typological term most closely associated with sociologist Max Weber. An ideal type is formed from

AMS 2000 subject classifications: Primary 62F10; secondary 62F30

Keywords and phrases: Random Graph, Mixture Model, Maximum Likelihood

characteristics and elements of the given phenomena, but it is not meant to correspond to all of the characteristics of any one particular case. It is not meant to refer to perfect things, moral ideals nor to statistical averages but rather to stress certain elements common to most cases of the given phenomena. Weber wrote: "An ideal type is formed by the one-sided accentuation of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified analytical construct". Although the term Ideal-Type has been proposed first for social science, in this paper we use it for all types of networks.

Each vertex i is the weighted mean of Q Ideal-Types, with weights given by $Z_i = (z_{i1}, \dots, z_{iQ})$, with $z_{iq} \geq 0$ and $\sum_q z_{iq} = 1$. The Q Ideal-Type vertices are put at the end of the canonical unit vectors $(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$ in \mathbb{R}^Q in an arbitrary order. The set of vertices $\{1, \dots, n\}$ is contained in the simplex $S_Q = \{x, \in [0, 1]^Q, \sum_{q=1, Q} x_q = 1\}$, so that the Ideal-Type vertices are hypothetical extreme vertices. Each Ideal-Type hypothetical vertex is supposed to be typical of the group of vertices which are near from it in S_Q , with more extremal properties than its neighbor real vertices.

Edges Each edge from a vertex i to a vertex j is associated to a binary random variable X_{ij} following a Bernoulli distribution with probability P_{ij} . The probability that the Ideal-Type q sends an edge to the Ideal-Type l is a_{ql} . The connectivity properties of each vertex i are a mixture of the connectivity properties of the Ideal-Types so that P_{ij} may be expressed using the weights z_{iq} and z_{jl} and the connectivity matrix A between the Ideal-Types:

$$(2.1) \quad P_{ij} = \sum_{q,l=1,Q} z_{iq} a_{ql} z_{jl}$$

which gives the matrix relation

$$(2.2) \quad P = ZAZ',$$

with

- P the (n, n) matrix containing the p_{ij} ,
- Z the (n, Q) matrix containing the z_{iq} and Z' the transpose of Z , $Z \in S_Q^n$,
- and $A \in [0, 1]^{Q^2}$, the (Q, Q) matrix containing the a_{ql} , the connectivity matrix between Ideal-Types.

The random variables X_{ij} are assumed to be independent. Let X be the (n, n) matrix containing the random variables X_{ij} . Finally the model is summarized by

$$(2.3) \quad X \sim \mathbb{B}(Z'AZ)$$

where \mathbb{B} denotes the Bernoulli distribution, $Z \in S_Q^n$ and $A \in [0, 1]^{Q^2}$.

The parameters of the model are A and Z . In a sense this model may be classified in the set of the semi-parametric statistical models, for each individual (vertex) has is proper set of parameters (z_{i1}, \dots, z_{iQ}) . Using statistical models, it is generally impossible to estimate as much parameters as the number of individuals. Moreover there are $Q^2 + n(Q - 1)$ parameters, so that this number tends to infinity with n . However, the number of observations contained in X is not proportional to n but to n^2 , so that the ratio of the number of parameters with the number of observations tends to 0 when $n \rightarrow \infty$. In practice, for each vertex i , there are n data, (x_{i1}, \dots, x_{in}) , available to estimate the $Q - 1$ parameters (z_{i1}, \dots, z_{iQ}) .

We can choose whether the graph is oriented or not by letting the X_{ij} loose or setting $X_{ij} = X_{ji}$ for all i, j . If the graph is not oriented, A is symmetric. Note that we assume in the following that there is no self-loop ($X_{ii} = 0$, for $i = 1, n$).

2.2. Relation between IDTM and other models.

2.2.1. Relation between IDTM and mixture model. In a mixture model for random graphs (Nowicki et al., 2001 (3) and Daudin, 2007 (1)), the variables Z are random and are equal to 0 or to 1. In the IDT model the variables Z are fixed parameters, and take their values in the simplex S_Q^n . In a mixture model, each item is assumed to pertain to only one group. The mixture model is a mixture of populations of "pure" items. In the IDT model, each item is a compound of Ideal-Types, so that the mixture is at the individual level. However there are two practical applications of the two models:

- the clustering of the items, i. e. the classification of each item in a group. The key element is $E(Z/X = x)$ in the mixture model and directly Z for IDTM. Note that $E(Z/X = x)$ in the mixture model, and Z in IDTM, take their values in the same set S_Q^n .
- The connectivity matrix A is the key element for the description and interpretation of groups in the two models (see Daudin (1)). However in the mixture model, A is the mean connectivity matrix in the sense that the probability of connection is the weighted mean of the connections between the vertices. On the opposite A represents an extreme connectivity matrix in the IDTM, so that A is more contrasted in IDTM than the one obtained with a mixture model.

Therefore, although the mixture model and IDTM are definitively different, their practical use is very similar.

2.2.2. Relation between IDTM and multidimensional scaling . The multidimensional scaling (MS) method, applied to the similarity matrix P , consists in

positioning each item in a metric space so that the similarity between items is approximatively kept. The underlying model is $P = TT'$, where T contains the coordinates of the items in a k -dimensional metric space. This model is similar to (2.2), with $k = Q - 1$. There are two major differences:

- T lies in \mathbb{R}^{kn} and $Z \in S_Q^n$.
- No Ideal-Type is modeled in the MS method and there is no connectivity matrix A , so that the clustering objective is not reached by MS. A further step of clustering the points in the metric space is necessary if one wants to obtain groups of vertices.

Note that (2.2) implies that $P \in [0, 1]^{n^2}$, but this is not true for $P = TT'$ so that the elements of P are not probabilities (i.e. contained in $[0, 1]$). The logit transformation has been used by Handcock et al. (2007) in order to handle this point, but their model is then definitively different from IDTM.

2.3. Model identifiability. As defined till now the model is not identifiable. Let P be a known matrix and assume that A and Z exist so that $P = ZAZ'$. It is generally possible to find other sets of parameters \tilde{A} and \tilde{Z} so that $P = \tilde{Z}\tilde{A}\tilde{Z}'$. Let H be a (Q, Q) matrix with the following properties:

1. H^{-1} exist
2. $H\mathbf{1}_Q = \mathbf{1}_Q$, with $\mathbf{1}_Q = (1 \dots 1)'$, with Q ones
3. $\tilde{Z} = ZH \geq 0$
4. $\tilde{A} = H^{-1}AH'^{-1} \in [0, 1]^{Q^2}$

Then we have:

- $\tilde{Z}\tilde{A}\tilde{Z}' = ZHH^{-1}AH'^{-1}H'Z' = P$
- $\tilde{Z}\mathbf{1}_Q = ZH\mathbf{1}_Q = Z\mathbf{1}_Q = \mathbf{1}_Q$ so that $\tilde{Z} \in S_Q^n$
- $\tilde{A} \in [0, 1]^{Q^2}$ by condition 4.

so that \tilde{Z} and \tilde{A} and Z and A are equivalent admissible sets of parameters.

Such matrix H do exist:

Assume that two columns l, q , of Z are strictly positive. Let

$$H = I_Q + B$$

where I_Q is the (Q, Q) identity matrix and B has all its coefficients equal to zero excepted $b_{qq} = b_{ll} = \epsilon \in]0, 1[$ and $b_{ql} = b_{lq} = -\epsilon$. We have

1. H^{-1} has the same structure as H with $b'_{qq} = b'_{ll} = -\frac{\epsilon}{1+2\epsilon}$ and $b'_{ql} = b'_{lq} = \frac{\epsilon}{1+2\epsilon}$
2. $H\mathbf{1}_Q = \mathbf{1}_Q$
3. $\tilde{Z} = ZH \geq 0$ because ϵ may be taken sufficiently small so that no term in \tilde{Z} becomes negative

4. $\tilde{A} = H^{-1}AH'^{-1} \in [0, 1]^{\mathcal{Q}^2}$ because some terms of \tilde{A} are equal to the corresponding terms of A and the other ones are weighted means of terms of A with positive weights summing up to one.

PROPOSITION 1. *The H -matrix operation increases $Tr(Z'Z)$ and decreases $V(A) = \sum (a_{iq} - \bar{a})^2$, so that it amplifies the differences between the coefficients of the same row of Z and decreases the differences between the coefficients of A .*

PROOF. Assume that the H -matrix is such that $l = 1$ and $q = 2$ (this trick is made only for simplicity of notations and does not imply any loss of generality). $Tr(Z'Z) = \sum_i \sum_q z_{iq}^2$, and $Tr(\tilde{Z}'\tilde{Z}) = \sum_i \sum_q \tilde{z}_{iq}^2$. For any fixed i we have

$$\begin{aligned}
 \sum_q \tilde{z}_{iq}^2 &= ((1 + \epsilon)z_{i1} - \epsilon z_{i2})^2 + ((1 + \epsilon)z_{i2} - \epsilon z_{i1})^2 + \sum_{q \geq 3} z_{iq}^2 \\
 &= (1 + 2\epsilon + 2\epsilon^2)z_{i1}^2 + (1 + 2\epsilon + 2\epsilon^2)z_{i2}^2 + \sum_{q \geq 3} z_{iq}^2 \\
 &= z_{i1}^2 + z_{i2}^2 + 2\epsilon(1 + \epsilon)(z_{i1} - z_{i2})^2 + \sum_{q \geq 3} z_{iq}^2 \\
 &= \sum_{q \geq 1} z_{iq}^2 + 2\epsilon(1 + \epsilon)(z_{i1} - z_{i2})^2 \\
 &\geq \sum_{q \geq 1} z_{iq}^2.
 \end{aligned}$$

The property $V(A) \geq V(\tilde{A})$ comes from the argument given in point 4. □

If no more than one column of Z is strictly positive, and at least two columns of A are strictly positive and strictly less than one, the same argument applies, using the matrix H^{-1} in place of H . Such H^{-1} -matrix operation decreases the differences between the coefficients of the same row of Z and increases the differences between the coefficients of A .

For illustration purpose, let us see the example with $n = 9$ vertices clustered in $Q = 3$ groups given in Table 1. This example works with any matrix A and two cases of matrices A are presented in Table 1. This example shows two equivalent sets of parameters, giving the same value for P . The first set is the version with the most contrasted possible values for A and medium coordinates Z for the vertices. On the opposite, the second parametrization has extremal values for Z and medium values for A . These are the two extremal possible parameterizations, with a continuous range of intermediate ones. The two extremal parameterizations have their own advantages and drawbacks: the first allows a very clear interpretation of

the connectivity matrix between clusters, A , but the possibility that most of the observed vertices are near the barycentre and no vertex near from the Ideal-Types. In the second one, the Ideal-Types are near (and in some cases such as in Table 1 exactly equal to) real vertices, but the connectivity matrix between the clusters is less clear-cut. We propose to choose Z which maximizes $Tr(ZZ')$ among the equivalent versions of model (2.4). The choice is motivated by two reasons:

- this constraint implies unicity of (Z, A) provided that $n \gg Q$ and the n vertices are different.
- the Ideal-Type should not be too far from real vertices in order to assess to them some reality. This closeness between Ideal-Type and some vertices is naturally provided by the maximization of $Tr(ZZ')$.

Finally the model is now:

$$(2.4) \quad X \sim \mathbb{B}(Z'AZ)$$

where \mathbb{B} denotes the Bernoulli distribution, $Z \in S_Q^n$, $A \in [0, 1]^{Q^2}$ and $Tr(Z'Z)$ is maximum.

3. Parameter Estimation. The log-likelihood is

$$(3.1) \quad L = \sum_{i,j} x_{ij} \log \left(\sum_{q,l=1,Q} z_{iq} A_{ql} z_{jl} \right) + (1 - x_{ij}) \log \left(1 - \sum_{q,l=1,Q} z_{iq} A_{ql} z_{jl} \right)$$

which may be written

$$(3.2) \quad L(A, Z) = Tr(X' \log(ZAZ')) + Tr((J - X)' \log((J - ZAZ')))$$

with J the (n, n) matrix composed of 1, $\log(ZAZ')$ is the matrix composed of the log of each element of ZAZ' , and the constraints on the parameters are

$$(3.3) \quad A \in [0, 1]^{n^2}$$

$$(3.4) \quad Z \in S_Q^n$$

Note that the the set of admissible solutions, $[0, 1]^{n^2} \times S$, is convex.

3.1. *Log-likelihood derivatives.* After some algebraic manipulations we obtain

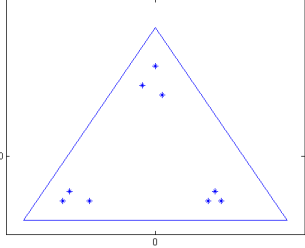
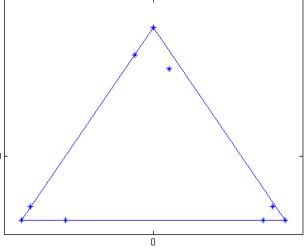
$$(3.5) \quad \frac{\partial L}{\partial Z} = RZA' + R'ZA$$

with R a (n, n) matrix with $r_{ij} = \frac{x_{ij} - p_{ij}}{p_{ij}(1 - p_{ij})}$, and

$$(3.6) \quad \frac{\partial L}{\partial A} = Z'RZ$$

TABLE 1

Example of two different sets of parameters (A, Z) giving the same connectivity matrix, P . The H matrix is the inverse of the first three lines of Z . Two cases of matrices A are presented.

	A and Z		$\tilde{A} = H^{-1}AH'^{-1}$ and $\tilde{Z} = ZH$	
A	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.66 & 0.24 & 0.17 \\ 0.24 & 0.54 & 0.17 \\ 0.17 & 0.17 & 0.66 \end{pmatrix}$	$\begin{pmatrix} 0.815 & 0.875 & 0.85 \\ 0.275 & 0.695 & 0.31 \\ 0.15 & 0.21 & 0.43 \end{pmatrix}$
V(A)	0.25	0.25	0.0475	0.0872
Z	$\begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.20 & 0.70 & 0.10 \\ 0.10 & 0.10 & 0.80 \\ 0.70 & 0.20 & 0.10 \\ 0.75 & 0.10 & 0.15 \\ 0.20 & 0.65 & 0.15 \\ 0.25 & 0.65 & 0.10 \\ 0.20 & 0.10 & 0.70 \\ 0.15 & 0.20 & 0.65 \end{pmatrix}$		$\begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0 & 0 & 1.0 \\ 0.833 & 0.167 & 0.0 \\ 0.929 & 0.0 & 0.071 \\ 0.012 & 0.917 & 0.071 \\ 0.083 & 0.917 & 0.0 \\ 0.143 & 0.0 & 0.857 \\ 0.048 & 0.167 & 0.786 \end{pmatrix}$	
	$Tr(Z'Z) = 5$		$Tr(Z'Z) = 7.68$	
Triangular representation of Z and \tilde{Z} in the simplex S_3				

The second order derivatives are more cumbersome:

$$(3.7) \quad \frac{\partial L}{\partial a_{ql} \partial a_{uv}} = - \sum_{ij} r_{ij}^2 z_{iq} z_{iu} z_{jl} z_{jv}$$

$$(3.8) \quad \frac{\partial L}{\partial a_{ql} \partial z_{iu}} = \delta_{qu} \sum_j r_{ij} z_{jl} + \delta_{lu} \sum_j r_{ji} z_{jq} - \sum_{jv} (r_{ij}^2 z_{jl} z_{jv} a_{uv} z_{iq} + r_{ji}^2 z_{jq} z_{jv} a_{vu} z_{iq})$$

$$(3.9) \quad \frac{\partial L}{\partial z_{iq} \partial z_{jl}} = r_{ij} a_{ql} + r_{ji} a_{lq} - \delta_{ij} \sum_{kuv} [r_{ik}^2 z_{ku} a_{lu} z_{kv} a_{qv} + r_{ki}^2 z_{ku} a_{ul} z_{kv} a_{vq}] - \sum_{uv} (r_{ij}^2 z_{iu} a_{ul} z_{jv} a_{qv} + r_{ji}^2 z_{iu} a_{lu} z_{jv} a_{vq})$$

with $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$.

3.2. Estimation Algorithm.

3.2.1. *Algorithm* . The constraints on the parameters are linear, but the log-likelihood is not linear. A linearization of (3.1) leads to a simple linear programming problem.

Let $A^{(k)}$ and $Z^{(k)}$ be the parameter estimates at step k , $P^{(k)} = Z^{(k)} A^{(k)} Z^{(k)'}$ and $R^{(k)}$ a (n, n) matrix with $r_{ij}^{(k)} = \frac{x_{ij} - p_{ij}^{(k)}}{p_{ij}^{(k)}(1 - p_{ij}^{(k)})}$.

The linear approximation of the log-likelihood at point $(A^{(k)}, Z^{(k)})$ is

$$\begin{aligned} L(A, Z) &\approx L(A^{(k)}, Z^{(k)}) + Tr \left[(A - A^{(k)})' \frac{\partial L}{\partial A}(A^{(k)}, Z^{(k)}) \right] + Tr \left[(Z - Z^{(k)})' \frac{\partial L}{\partial Z}(A^{(k)}, Z^{(k)}) \right] \\ &\approx L(A^{(k)}, Z^{(k)}) + Tr \left[(A - A^{(k)})' Z^{(k)'} R^{(k)} Z^{(k)} \right] + Tr \left[(Z - Z^{(k)})' (R^{(k)} Z^{(k)} A^{(k)'} + R^{(k)'} Z^{(k)} A^{(k)}) \right] \end{aligned}$$

The algorithm is the following:

- Find initializing values $(A^{(0)}, Z^{(0)})$
- At step (k) use a linear programming algorithm to maximize the function in (A, Z) :

$$(3.10) \quad f_k(A, Z) = Tr \left[A' Z^{(k)'} R^{(k)} Z^{(k)} \right] + Tr \left[Z' (R^{(k)} Z^{(k)} A^{(k)'} + R^{(k)'} Z^{(k)} A^{(k)}) \right]$$

under the constraints:

$$(3.11) \quad A \in [0, 1]^{n^2}$$

$$(3.12) \quad Z \in S_Q^n$$

- We can use two alternative stopping rules:

$$(3.13) \quad \|A^{(k)} - A^{(k-1)}\| + \|Z^{(k)} - Z^{(k-1)}\| <$$

$$(3.14) \quad |L(A^{(k)}, Z^{(k)}) - L(A^{(k-1)}, Z^{(k-1)})| < \alpha$$

In practice it is valuable to limit the difference between two successive steps of the algorithm by adding the constraints $|Z^{(k)} - Z^{(k-1)}| < \epsilon$ and $|A^{(k)} - A^{(k-1)}| < \epsilon$. This is easy to do, because coordinates of Z and A are in $[0, 1]$, so that it is reasonable (and easy using linear programming) to bound the absolute values of these differences. This trick prevents a possible oscillating behavior of the algorithm. An alternative algorithm of non-linear optimization using the second order derivatives of the log-likelihood given in 3.7, 3.8 and 3.9 would be possible, but we have found that the above algorithm is easy to implement, robust and very efficient.

3.2.2. *Initialization* . A good starting value for the algorithm is given by the following process:

1. Use Principal Coordinate on the similarity matrix X to obtain a $Q - 1$ dimensional representation space. This is obtained by taking the $Q - 1$ first eigenvectors of $W = X'X + XX'$.
2. Find an approximately minimal convex hull of the n points in \mathbb{R}^{Q-1} with Q vertices.
3. Build A using the values of X for the nearest point of each vertex, so that $A = X(1 : Q, 1 : Q)$ where the matrix X is sorted so that the first vertex of X is the nearest from one of the Q vertices of the convex hull, the second vertex of X is the nearest from another vertex of the convex hull, and so on till vertex Q .

3.2.3. *Assessment of the identification of the model*. The model is not identifiable as it stands (see 2.3). In practice we have not seen any problem coming from the lack of identifiability when using the above algorithm. However, we advise to use the stopping rule (3.14) in order to avoid to the algorithm to fluctuate between equivalent solutions (A, Z) . After convergence, we obtain a unique instance of the equivalent class of parameters (A, Z) by maximizing $Tr(Z'Z)$ under the constraint that $ZZ' = Z^{(k)}A^{(k)}Z^{(k)}$, with k the iteration number at convergence.

3.3. *Choice of the Number of Groups*. We use the AIC or BIC criteria:

$$(3.15) \quad AIC(Q) = -2L(\hat{A}_Q, \hat{Z}_Q) + 2(Q^2 + n(Q - 1))$$

$$(3.16) \quad BIC(Q) = -2L(\hat{A}_Q, \hat{Z}_Q) + Q^2 \log(n(n - 1)) + n(Q - 1) \log(2n)$$

(\hat{A}_Q, \hat{Z}_Q) are the maximum likelihood estimates of (A, Z) for Q groups. There are $n(n - 1)$ observations for the edges and $Q^2 + n(Q - 1)$ parameters.

TABLE 2
Results of three simulations. Parameter estimate (Asymptotic standard error estimate)

case	1	2	3
$\hat{a}_{11} (\hat{\sigma}(\hat{a}_{11}))$	0.612 (0.028)	0.594 (0.0013)	0.78 (0.02)
\hat{a}_{12}	0.100 (0.019)	0.094 (0.001)	0.049 (0.016)
\hat{a}_{21}	0.203 (0.023)	0.194 (0.001)	0.61 (0.02)
\hat{a}_{22}	0.000 (0.020)	0.000 (0.0003)	0.889 (0.02)
iterations	34	4	32
time	0.6s	1.6s	3.3s

4. Simulation Study. We have tested our algorithm on some simulated examples. Here we consider three cases:

1. $n = 50, Q = 2, A = \begin{pmatrix} 0.6 & 0.1 \\ 0.2 & 0 \end{pmatrix}$
2. $n = 1000, Q = 2, A = \begin{pmatrix} 0.6 & 0.1 \\ 0.2 & 0 \end{pmatrix}$
3. $n = 200, Q = 4, A = \begin{pmatrix} 0.7 & 0.1 & 0.2 & 0.3 \\ 0.6 & 0.8 & 0.0 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.8 \\ 0.5 & 0.2 & 0.4 & 0.1 \end{pmatrix}$

The results are given in the Table 2 and Figure 1.

5. Example. We analyze the well known data set about the relations between members of a karate-club from Zachary (1977) (4). The network is presented in Figure 2.

Using AIC, the best choice is $Q = 4$. The estimates for A and Z are given in Tables 3 and 4. The 4 groups are the circles's hubs (vertices 33 and 34), the circles, the squares's hubs (vertices 1 and 2) and the squares. The four Ideal-Types are characterized by the connectivity matrix \hat{A} : they are

- a circle's hub connected with himself and with the circles
- a circle connected only to its hub
- a square's hub connected with himself and with the squares
- a square connected only to its hub

Some vertices are very near to an Ideal-Type (see vertices 1, 8, 10, 12, ...34), but others have intermediate values, such as vertex 2 which is a mixture between a squares's hub and a simple square. Vertex 1 is the most important vertex for the squares. The same comment may be applied to vertices 33 and 34. Some vertices hesitate between the two sides such as vertices 9, 17 and 25. However there is no classification error coming from the unsupervised clustering made by the IDT model: the Karate-club has been divided in two known groups, the circles and the

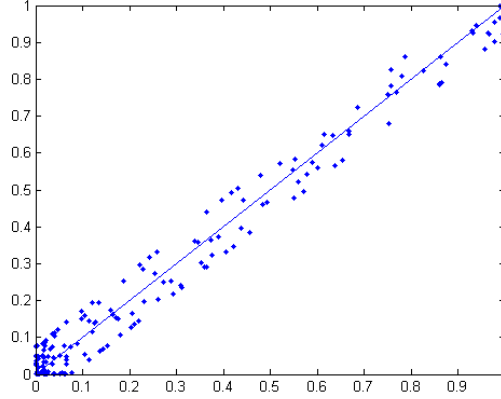


FIGURE 1. Plot of the estimated values of Z_1 (y-axis) versus the true ones (x-axis) for Ideal-Type 1, in the third simulated case

TABLE 3
Connectivity matrix estimate \hat{A} for the Karate-club network, (asymptotic estimate of the standard-deviation of \hat{A})

group	circles's hubs	circles	squares's hubs	squares
circles's hubs	1 (0.87)	1 (0.16)	0.002 (1.1)	0 (0.20)
circles	1 (0.16)	0 (0.05)	0.002 (0.24)	0 (0.13)
squares's hubs	0.002 (1.1)	0.002 (0.24)	1 (1.68)	1 (0.06)
squares	0 (0.20)	0 (0.13)	1 (0.06)	0 (0.24)

squares. This division is exactly recovered by the IDTM if we bring together groups 1 and 2 on one side, and groups 3 and 4 on the other side. The particular position of vertex 3 is interesting: it is mostly a simple square but also in part a squares's hub and a circles'hub, having a strategic position between the two sides.

The asymptotic estimates of the standard errors for the parameters have been computed using the inverse of the information matrix. Some of them are high, for example the probability of connection between two circles's hubs is not precise: its estimated value is equal to 1 but the standard error is estimated to 0.87. This high value may be explained by the fact that there is few information to estimate this parameter, because there are only two vertices which are concerned by this Ideal-Type. The same is true for the probability of connection between the squares's hubs. On the opposite the relation between the squares's hubs and the squares is

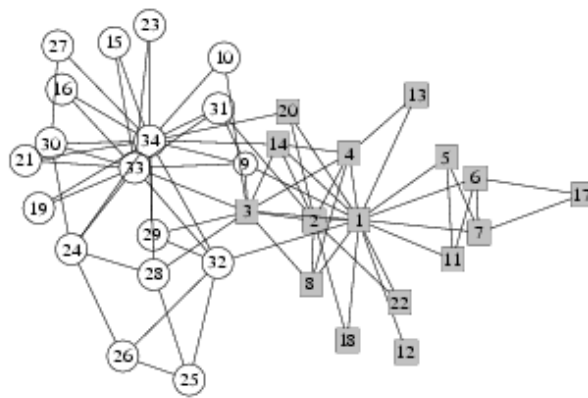


FIGURE 2. *Karate-club network from Zachary. The club has split in two subsets: squares and circles.*

TABLE 4
Z'values for the karate-club network

vertex	circles's hubs	circles	squares's hubs	squares
i	z_{i1}	z_{i2}	z_{i3}	z_{i1}
1	0	0	1	0
2	0	0	0.503	0.497
3	0.267	0	0.246	0.486
4	0	0	0.279	0.719
5	0	0	0.109	0.890
6	0	0	0.150	0.848
7	0	0	0.079	0.921
8	0	0	0	1
9	0.072	0.559	0	0.369
10	0	1	0	0
11	0	0	0.016	0.984
12	0	0	0	1
13	0	0	0	1
14	0	0.2096	0	0.7884
15	0	1	0	0
16	0	1	0	0
17	0	0.426	0.090	0.483
18	0	0	0	10
19	0	1	0	0
20	0	0.286	0	0.713
21	0	1	0	0
22	0	0	0	10
23	0	1	0	0
24	0.196	0.802	0	0
25	0.155	0.534	0	0.311
26	0	0.832	0	0.167
27	0	1	0	0
28	0	1	0	0
29	0	1	0	0
30	0.097	0.902	0	0
31	0	0.773	0	0.226
32	0.181	0.547	0	0.272
33	0.638	0.361	0	0
34	1	0	0	0

better known (standard error = 0.06) because there are about 15 vertices which bear information on this probability of connection.

6. Conclusion. The mixture models is a method of choice for modelling heterogeneous random graphs, because it contains most of the known structures of heterogeneity: hubs, hierarchical structures or community structure. One of the weakness of mixture models on random graphs is that there is no theoretically completely satisfying estimation method. The variational EM estimate used in (1) is not consistent (although it works pretty well in many examples) and it does not allow to get the variances of the estimates. The bayesian method used in (2) is computationally highly intensive and works only for small networks. The discrete nature of Z implies that one has to explore a space of dimension Q^n , a task which is clearly impossible. In the IDT model the discrete Z are replaced by continuous ones, which leads to an easier optimization problem and allows to obtain the maximum-likelihood estimates with an efficient algorithm. However some additional work is necessary to understand the behavior of the maximum-likelihood estimates of n parameters and n^2 observations when $n \rightarrow \infty$.

References.

- [1] DAUDIN, JJ., PICARD, F. and ROBIN, S. (2007). A mixture model for random graphs. *Statist. Comput.* **18**(2), 173–183
- [2] HANDCOCK, MS., RAFTERY, AE. and TANTRUM, JM. (2007). Model-based clustering for social networks. *JRSSA* **54**, 301-354
- [3] NOWICKI, K. and SNIJDERS, T. (2001). Estimation and prediction for stochastic block-structures. *J. Am. Stat. Assoc.* **96**, 1077-1087
- [4] ZACHARY, WW. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* **33**, 452–473

AGROPARISTECH, 16 RUE CLAUDE BERNARD 75231 PARIS CEDEX05, FRANCE
UNIVERSITY PARIS X, 200 AVENUE DE LA REPUBLIQUE, 92001 NANTERRE CEDEX, FRANCE