BOLYAI SOCIETY
MATHEMATICAL STUDIES, 17

Horizons of Combinatorics
Balatonalmádi
pp. 37–66.

Statistical Inference on Random Structures

VILLŐ CSISZÁR, LÍDIA REJTŐ and GÁBOR TUSNÁDY

Introduction

Randomness for a statistician must have some structure. In traditional combinatorics the word random means uniform distribution on a set which may be the set of all graphs with n vertices, the set of all permutations of the numbers $N = (1, 2, \ldots, n)$, the set of all partitions of N, or any other set of simple structure. In practice the statistician meets a subset of the structures and she or he is interested in the question, what was the mechanism which generated the sample. Uniform distribution and independence are shapeless and they have low complexity for catching the character of samples produced by real life situations. In [12] Persi Diaconis investigated a sample consisting of the votes in an election of the American Psychological Association. The sample was investigated by others but without achieving a reasonable goodness of fit, because the present collection of distribution of permutations is not large enough. Investigating the sample we found a hidden property leading to a new class of distributions of permutations.

Classical statistics developed around the multidimensional Gaussian distribution. Even in Euclidean space the family of useful distributions is still meager. On other sample spaces the collection of distributions is much less developed. Graphs appear in applications as structured relations. In many cases rather heavy simplifications are needed for reducing the complexity of the investigated situation to a graph. One source of our interest in graphs is the system of metabolic interactions, which may have some fractal structure: the enzymatic interactions may be leveled, they may be sensitive for situations, their control might be hierarchic. Changes of the concentration

of different enzymes in a cell follow their dynamical rule what is reflected imperfectly in the graph of enzymatic interactions.

In modern combinatorics the stochastic method is rapidly extending. We shall use the ideas of papers [7], [8] and [9] written by Christian Borgs, Jennifer Chayes, László Lovász, Vera T. Sós, Balázs Szegedy and Katalin Vesztergombi in defining new classes of random permutations.

SVD of real matrices. Let M be an arbitrary digital picture: a face, a tree, a hill or some other natural object which is not very complicated. Let us suppose that the colours are ordered according to their wave lengths and M is an m times n real matrix containing the codes of the colours in the individual pixels. Let α be a random permutation of the integers $1, \dots, m$ and β of $1, \dots, n$. Let

$$R(i,j) = M\big(\alpha(i), \beta(j)\big)$$

be the randomly reordered copy of M. How can we reconstruct M from R?

One possible method is the singular value decomposition (SVD) of R which is invariant under random permutations. The singular values of matrices M and R are identical. We refer to them as the *spectra* of the corresponding matrix. If the picture is simple, then the spectra is J-shaped: there are few large singular values and the corresponding singular vectors concentrate the majority of the relevant information in M. The coordinates of the leading singular vectors of M reflect the topology of M, while the coordinates of the singular vectors of R follow the permutations α, β. It implies that the traveling salesman problem may be easily solved in the space of leading eigenvectors independently of rows and columns.

Microarray analysis. The previous problem arises in microarray analysis where the rows are genes and the columns are the different conditions used in the experiment for controlling the expression of the genes. It is natural to postulate that the genes and conditions are embedded in Euclidean spaces and the expression level is a continuous function of the embedding. Sometime we get well defined clusters when applying SVD of microarray data: clusters in genes come from the metabolic networks of the proteins they code and the clusters of conditions come from the structure of the plan of the experiments. The phenomenon is known in the literature as the *chequerboard structure:* after appropriate reordering, gene-expression matrices become chequerboard like. Batches of genes express similarly under

batches of conditions. Interestingly, rather good reorderings are supplied by simple hierarchical clusterings of rows and columns simultaneously.

Graphs

Graph complexity. There are natural ways to assign matrices to a graph: the off-diagonal entries reflect the connectivity and the diagonal entries may be set to zero or to the degree multiplied by -1. In the second case the sum in each row is zero and a non-zero vector with equal coordinates is an eigenvector with zero eigenvalue. All eigenvalues are non-positive in the second case. We call the matrix in first case the adjacency matrix and the second one the Laplacian ([3], [4], [5], [10], [18], [23], [35]). For regular graphs the spectra of the two matrices differ only by a constant.

An arbitrary graph is a free sequence of $\binom{n}{2}$ bits. Without fathoming the inner structure of the graph we can not catch the complexity of a graph. In the simplest case the spectra is J-shaped: there is some topology on the vertices and the edges follow that. For Albert–Barabási graphs ([1], [6]) the topology comes from preference: the degrees of the vertices control the choice of the edges. According to Wigner's semicircle law ([17], [20], [25]) for random graphs the spectra of the adjacency matrix forms a semicircle, which is definitely not J-shaped. Incidentally: we do not know what is the asymptotic for the spectra for random symmetrical matrices with i.i.d. off-diagonal entries but putting the sums (multiplied by -1) in the diagonal. If the entries of a random matrix are independent Wiener processes, the eigenvalues $\lambda_i = \lambda_i(t)$ follow the system of stochastic differential equation

$$d\lambda_i = dW_i + dt \prod_{j \neq i} \frac{1}{\lambda_i - \lambda_j}, \qquad i = 1, \ldots, n$$

showing that the eigenvalues repel each other. Do eigenvalues of random graphs repel each-other? Does this depend on which eigenvalue definition we use and what model of random graphs?

Fractals. An other intriguing question is, whether there are fractals in large graphs? To catch the fractal behavior we propose the following potential defined for connected graphs. For a given vertex x let y be the vertex closest to x of degree not smaller than that of x, and let D_x be the set of vertices

different from x that are strictly closer to x than y is. This D_x is the *estate* and its size the *asset* of x. (If there is only one vertex with maximal degree then its estate is empty.) The *wealth* V_x of x is the sum of the assets of all vertices in D_x. Finally, the potential of the graph Γ is

$$\mathcal{Q}(\Gamma) = \sum V_x^\alpha V_y^\alpha d^\beta(x,y),$$

where the summation runs on all pairs (x,y) of vertices, d is the distance on the graph and $\alpha, \beta > 0$ are fixed constants. What is the graph which maximizes this potential for fixed number of vertices? For $n = 254$, $\alpha = 0.75$, $\beta = 0.25$ we constructed several graphs. Revealing the structure of optimal graphs created by exhaustive stochastic search we generated the graph presented in the Appendix. For this graph $Q(\Gamma) = 14,343$. The structure of the graph is shown in Figure 1. The empty circles represent virtual vertices, which help only in building up the structure. We tend to believe that real complexity is connected with the repelling property of the eigenvalues, while the concentration of the eigenvalues comes from the equivalence of the vertices.

Equivalent vertices. Equivalence of vertices have two features:

- equivalent vertices may prefer each other: the edge-density inside equivalent clusters is larger than outside
- vertices belonging to equivalent clusters behave similarly.

The first case is reflected by the spectra of the Laplacian and the second case is Szemerédi's regularity property ([14], [21], [27], [34], [37]): we say that the bipartite graph with vertex sets A, B is ε-regular if

$$\left| \frac{E(X,Y)}{|X|\,|Y|} - \Delta \right| \le \varepsilon,$$

holds true for all $X \in A$, $Y \in B$ such that $|X| \ge \varepsilon|A|$, $|Y| \ge \varepsilon|B|$, where

$$\Delta = \frac{E(A,B)}{|A|\,|B|}$$

is the edge-density in the whole graph.

Regularity lemma for a statistician. Roughly speaking, Szemerédi's regularity lemma states that the vertices of *every* graph may be clustered

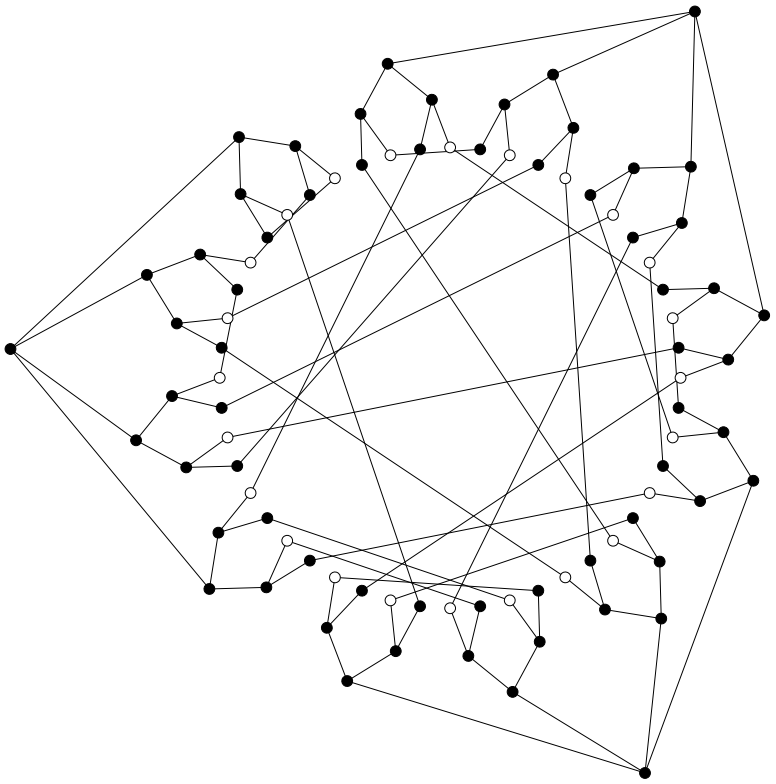

Fig. 1

in such a way that the bipartite graphs corresponding to different clusters are ε regular with a small exceptional fraction of the pairs if the number of vertices is large enough.

For a statistician the condition in the definition of ε-regularity is a statistical test resembling to Rényi's version of the Kolmogorov test. Let n be an arbitrary number, for integers i between 1 and n let $\alpha(i)$ be arbitrary integers between 1 and k, where $k < n$. Let $p_{i,j}$, $1 \leq i, j \leq k$ be an arbitrary symmetric matrix with $0 \leq p_{i,j} \leq 1$. We call the random graph *checkerboard graph* if vertices i, j, where $1 \leq i \leq j \leq n$ are connected with probability $p_{\alpha(i),\alpha(j)}$ and the edges are independent. At first instance, the regularity lemma seems to state that the collection of checkerboard graphs is *bold enough* for having the power to generate all graphs. The striking effect of the lemma is its simplicity: the random mechanisms used in a possible rigorous formalization are quite natural and, what is more, they are not capable of catching all the possible information out of a graph.

The riddle of Szemerédi's lemma is hidden in the definition of regularity. It fixes, prescribes a test on graphs for the use of testing the hypothesis that the graph comes from the class of checkerboard distributions. Being true statisticians we propose to develop other tests, possibly with relevant power for testing the hypothesis. One natural aspirant is the spectra of the adjacency matrix: for checkerboard graphs it has to be J-shaped, *and* the eigenvectors have to show clear clusters. Any deviation from these properties may lead to rejecting the hypotheses.

A universal lemma might state that *any* maximum likelihood estimate is bold enough to have the property that it is optimal for all measures in the statistical field. You can never use the picture given by a maximum likelihood estimate for testing the hypothesis concerning the completeness of the investigated measures. Inside the world the statistical field they have to be bold enough just by definition of the maximum likelihood estimate. But we can test the hypothesis by other accordingly chosen statistics which are usually orthogonal to the logic of the likelihood. Recently one of the most interesting fields for an extension of the lemma are the hypergraphs. Accordingly we have to learn the precise use of the *stochastic method:* it is better to formulate minor sets of conditions under which a useful theorem of stochastics holds true and extend it to as wide a territory as possible but we can never forget effectiveness. In case of Szemerédi's lemma it is the blow up property.

Blow up property states that in a large enough graph all the small graphs appear with a frequency proportional with their probabilities. The statement is also called the *Counting Lemma*. Taking a large enough distance from the details of the affair investigated we think that the situation resembles quantum physics: first you choose what you are interested in, then the analytic machinery answers your question as you like it. If we want to ensure ε-regular colouring for all graphs then we have to choose the number of colours enormously large. But according to our experience, checkerboard graphs are applicable to small graphs too. The "bold enough" property appears only for large graphs, but the property being universal for large graphs may be present for a special family of small graphs.

Other models. Let $f(x, y)$ be a differentiable function for $0 \le x, y \le 1$ such that $0 \le f(x, y) \le 1$. Let $x_1, x_2, \ldots, x_n$ be arbitrary numbers in $(0, 1)$. The

random graph connecting the vertices i, j independently with probability $p_{i,j} = f(x_i, x_j)$ represents the function f and numbers x_i. We can try to reconstruct the model parameters $f, x_1, \ldots, x_n$ by a maximum likelihood method. Maximizing the likelihood the following two-phase algorithm is applicable:

- for given f the gradient method applies to the x_i-s
- for given x_i-s the function $f(x, y)$ may be estimated by the edge-density for $|x_i - x| < \varepsilon$, $|x_j - y| < \varepsilon$.

We say that the function f is the *face* of the graph and the x_i-s are its *core*. For large n and $x_i = \frac{i-0.5}{n}$ the spectra of the random graph is close to the spectra of f. If $f(x, y) = \frac{x+y}{2}$ the eigengenvalues are uniformly distributed in $\left(\frac{1}{4}, \frac{3}{4}\right)$ in contrast with the chequerboard case when they are clustered around a few points. It goes without saying that the uniform distribution may be approximated by a discrete distribution concentrated on finitely many points, but we may detect the difference with appropriate statistics. What is the case with Szemerédi's statistics

$$\frac{\big(E(X,Y) - F(X,Y)\big)^2}{G(X,Y)},$$

where

- $E(X, Y)$ is the number of edges between the disjoint sets X, Y
- $F(X, Y) = \sum_{x_i \in X, x_j \in Y} p_{i,j}$ is the expected value of $E(X, Y)$
- $G(X, Y) = \max\big(1, \sum_{x_i \in X, x_j \in Y} p_{i,j}(1 - p_{i,j})$ is the truncated variance of $E(X, Y)$?

Of course it has to be applicable to detect the difference, but in the regularity lemma the constants are chosen loosely for that aim. The spectra of the adjacency matrix shows more characteristic effect of the checkerboard structure than the Laplacian, but a rigid SVD of the matrix $p_{i,j}$ is usually not flexible enough for detecting real structures because it poorly approximates matrices with entries in the interval $(0, 1)$. The logistic transform $p_{i,j} = b/\big(c + \exp(a_{i,j})\big)$ offers an easy bridge between real numbers and the $(0, 1)$ interval. More generally, we can use any monotone increasing function for this role. The nonparametric maximum likelihood estimator is a step function usually with a small number of steps and a remarkable portion of the edges has probability zero or one and thus the fitted model has moderately random character only on the borderline of the two subsets of edges where we can explicitly predict their existence.

Dynamics. The most complicated matrix $p_{i,j}$ is unable to reflect fine interactions between the edges. We can build up systematically stochastic models starting with a joint distribution of two or three edges or subsets of vertices, but in case of graphs presented by real life situations the structure of stochastic interactions is mostly multifactorial. Firstly, gathering all the available information, we can try to describe with words the characteristic features of the investigated graph. Next we translate our own words to mathematical formulas and we define some potential function measuring the perfection of individual graphs and we develop algorithms to maximize the potential following a kind of Darwinian path. The algorithms may resemble to the mechanisms creating the studied graphs. But typically the optimization procedure reveals something that is rather far from our ideals formulated originally in words. In such situation the whole procedure starts again and we should recycle it until convergence.

The potential $\mathcal{Q}(\Gamma)$ defined by assets and wealths led to the following procedure. We start with one vertex. Step by step, each vertex in the graph is divided into two daughters, and in the new graph

- we join two daughter points with probability $p = 0.06$, if their mothers were joined
- otherwise we join them with probability $q = 0.005$, and
- we join them with probability $p = 0.03$ if they have the same mother.

The fractal structure is imprinted in the algorithm. The reason for the low probabilities is that the potential prefers spare graphs. One source of the potential is

$$\Psi(\Gamma) = \sum_{(x,y)\in\mathcal{E}} f_x f_y,$$

where f_x is the degree of x. In Albert–Barabási dynamics Ψ is maximal among graphs with given degrees, which is unnatural in the majority of cases: the hubs are in most cases separated, they are far from other hubs.

Permutations

The Thurstonean. Let $F_1, F_2, \ldots, F_n$ be arbitrary continuous real distributions, for each $1 \le i \le n$ let X_i be a random variable with distribution

F_i and let the variables $X_1, X_2, \dots, X_n$ be independent. Let

$$\pi = \big(\pi(1), \pi(2), \dots, \pi(n)\big)$$

be the permutation ordering the X_i-s monotone increasingly:

$$X_{\pi(1)} < X_{\pi(2)} < \cdots < X_{\pi(n)}.$$

The model was proposed by Louis Leon Thurstone in [38] (see also in [33]) thus we call the distribution defined by the model *Thurstonean.*

It is easy to see that if the distributions F_i are exponentials with parameters λ_i then

$$P\big(\pi(a+1) = t \mid \pi(1), \pi(2), \dots, \pi(a)\big) = \frac{\lambda_t}{\sum_{i \notin L_a} \lambda_i},$$

$$a = 0, 1, \dots, n-1, \quad t \notin L_a,$$

where $L_a = \big\{\pi(1), \pi(2), \dots, \pi(a)\big\}$ with $L_0 = \emptyset$.

If Y_t, $t = 1, 2, \dots$, are i.i.d. with distribution

$$P(Y_1 = t) = p_t, \quad t = 1, 2, \dots, n,$$

and we delete all elements from the sequence that we have seen earlier, then the remaining random numbers form a permutation in N with the same distribution as the exponential Thurstonean one, whenever

$$p_t = \frac{\lambda_t}{\sum_{i=1}^{n} \lambda_i}.$$

Interestingly, in these two models the EM-algorithm [24] leads to different iterations. In the general case the Baum–Welch algorithm [29] leads to the following iteration. For the sake of simplicity let us suppose that π is the identity. In this case we have to calculate the conditional distributions

$$Q_i(t) = P(X_i < t \mid X_1 < X_2 < \cdots < X_n), \quad i = 1, 2, \dots, n.$$

In the forward phase of the algorithm we calculate recursively the conditional distributions

$$G_i(t) = P(X_i < t \mid X_1 < X_2 < \cdots < X_i)$$

by

$$g_i(t) = f_i(t) G_{i-1}(t),$$

where $f_i = F_i'$, $g_i = G_i'$. Similarly, for

$$H_i(t) = P(X_i < t \mid X_i < X_{i+1} < \cdots < X_n)$$

$$h_i(t) = f_i(t)H_{i+1}(t)$$

holds true where $h_i = H_i'$ while $G_1 = F_1$, $H_n = F_n$. Then

$$q_i(t) = f_i(t)G_{i-1}(t)H_{i+1}(t),$$

where $q_i = Q_i'$ and $G_0 = H_{n+1} = 1$.

Let us denote by $\mathcal{T}_n$ the set of all Thurstonean random permutations with n elements. A possible generalization is to drop the independence of the X_i-s. Let $\mathcal{G}_n$ be the set of n-dimensional Gaussian distributions with expectation μ and covariance Σ. $\mathcal{G}_n$ is described by

$$\binom{n+1}{2} + n - 2$$

parameters, which suggests that for $n = 2, 3$ the model is overparametrized. Indeed, for $n = 2$, Σ may be reduced to the identity matrix and $\mathcal{G}_2 = \mathcal{T}_2 = \mathcal{P}_2$ where $\mathcal{P}_n$ stands for the set of all possible distributions of permutations on N. If $n = 3$, then the distribution rendering half probability to the permutations $(1, 2, 3)$, $(3, 2, 1)$ is definitely not in $\mathcal{T}_3$ yet, it is in $\mathcal{G}_3$ for μ equals zero and with a covariance ensuring that $X_1 = 2X_2, X_3 = 0$ or $X_1 = -X_3, X_2 = 0$. For large n the set $\mathcal{T}_n$ should be larger than $\mathcal{G}_n$, for the number of degree of freedoms goes to infinity more rapidly in the first case. In the Gaussian case a possible reduction of the number of parameters is the control on the rank of Σ as it is usual in factor analysis and principal component analysis. But the covariance of π is unable to catch the rank of Σ, it is visible only in the covariance of π^{-1}. Permutations in practice mostly come from some one-to-one correspondence between two different unordered sets. The row-ordering and column-ordering of the chequerboard representing the permutations usually is lurking behind. If X has some multidimensional stochastic structure one cannot find it in π, because $\pi(i)$ gives the *coordinates* of the i-th element of the ordered sample answering the question: *who* stays on the i-th position. But the order of coordinates in X is arbitrary. In contrary, $\pi^{-1}(j)$ tells us *where* the j-th coordinate X_j is in the ordered sample which is a nearly linear function of the values of the coordinates, hence the covariances of X and π^{-1} are close to each other.

Statistics. Models and statistics on a structure are the two legs of any inference. All models have their natural statistics or sufficient statistics and for a given family we can test the goodness of fit of the whole family. For permutations the primary marginals are the positions of a subsets of the elements among the whole set: here a void mark is substituted for the elements outside the group, this means that the order of the elements of the chosen group is filled in with some void marks:

$$* * * 3 * 5\, 1 * * * * * 4 * 2 *$$

means that $\pi(4) = 3$, $\pi(6) = 5$, $\pi(7) = 1$, $\pi(12) = 4$, $\pi(14) = 2$. Dropping the stars we get the permutation of the chosen elements which is another marginal. We say that the permutation 35142 is the *shrunken* version of the original one into the set $(1, 2, 3, 4, 5)$. For one element only the filled marginals contain information, one is tempted to use these one-element positions as aspirants for the unknown distributions in the Thurstonean case. Turning to the inverse, other one-element marginals appear and the distribution, having simultaneously a given row marginal and a given column marginal is of the form

$$P(\pi) = \kappa \prod_{i=1}^{n} a_{i,\pi(i)}^{\Delta},$$

where the matrix $A = (a_{i,j})$ is an arbitrary doubly stochastic matrix, Δ is a positive number, and κ is the normalizing factor. It is well known that for any doubly stochastic A there is at least one permutation with positive probability. We call the distribution *simple rook* distribution because representing the permutations on a chequerboard the probability of the permutation is proportional to the product of the numbers in the occupied pixels. For simple rook distributions the sufficient statistics are

$$\nu(i, j) = \#\big\{\, \pi(i) = j \big\}, \quad 1 \le i, j \le n,$$

which are simultaneously the matrices of unnormed row and column marginals. The statistics $\nu(i, j)$ are useful for distributions

$$P(\pi) = \kappa \exp\big(- d^2(\pi, \pi_0)/T \big),$$

where d is some distance function, π_0 is the centrum of the distribution, T is a positive constant, and κ is the norming factor. The family was introduced by Mallows in [22] (see also [15], [28] and [36]).

Row cuttings. Let us say that an element of $\mathcal{P}_n$ has the property of *row cutting at* a if

$$P(\pi \mid L_a) = f\big(\pi(1), \pi(2), \dots, \pi(a)\big)\ g\big(\pi(a+1), \dots, \pi(n)\big)$$

holds true with some a variate function f and $(n-a)$ variate function g, where $2 \le a \le n-2$. We denote by $\mathcal{R}_a$ the set of all distributions with the property row cutting at a. Row cutting at a means that the permutations $\big(\pi(1), \pi(2), \dots, \pi(a)\big)$, $\big(\pi(a+1), \pi(a+2), \dots, \pi(n)\big)$ are conditionally independent on the statistics L_a. We say that a random permutation is *row-free* if it has the row cutting property for all a. It is easy to see that row-free random permutations have the form

$$P(\pi) = \prod_{a=0}^{n-1} c\big(\pi(a+1), L_a\big),$$

where the conditional probabilities $c(u, V)$ are concentrated on $u \in N \setminus V$. The degree of freedom of the set $\mathcal{R}$ of row-free permutations is

$$r_n = \sum_{a=1}^{n} (a-1)\binom{n}{a} = \Big(\frac{n}{2} - 1\Big)\, 2^n + 1.$$

The set $\mathcal{C}_b$ is the set of all distributions with the property column cutting at b, and the set $\mathcal{C}$ of column-free permutations is similarly defined with substituting π^{-1} for π. A possible representation of n element permutations is putting rooks on the n by n chequerboard: here the properties of row- and column-freeness are symmetrical. The sample presented by Persi Diaconis happens to be in a certain sense inside of the intersection of the sets $\mathcal{R}$ and $\mathcal{C}$. Our main theorem states that the degree of freedom of the intersection is

$$\nu_n = \sum_{a=1}^{n-1} a^2.$$

We call the elements in the intersection *free* distributions. Exponential Thurstonean distributions are row-free and the simple rook distribution is free. A possible set of sufficient statistics for free distributions is the following:

$$\nu(a,b) = \sum_{i=1}^{a} \mathcal{I}\big(\pi(i) \le b\big), \quad 1 \le a, b \le n-1,$$

$$\mu(a,b) = \sum_{i=1}^{a} \mathcal{I}\big(\pi(i) \le b, \pi(a+1) = b+1\big), \quad 1 \le a, b \le n-2.$$

If all permutations have positive probabilities then a free distribution has the form

$$P(\pi) = \prod_{a=1}^{n-2} g(a, \pi(a+1), \mu\big(a, \pi(a+1)\big) \prod_{a=1}^{n-1} \prod_{b=1}^{n-1} f\big(a, b, \nu(a,b)\big)$$

In the intersection of the sets $\mathcal{R}_a$, $\mathcal{C}_b$ the distributions have the form

$$P(\pi) = \alpha(\pi_a^b)\beta\big(\pi_a^{\bar{b}}\big)\,\gamma\big(\pi_{\bar{a}}^b\big)\,\delta\big(\pi_{\bar{a}}^{\bar{b}}\big)\,,$$

where $\alpha, \beta, \gamma, \delta$ are positive functions, and

π_a^b denotes the shrunken version of $\big(\pi(1), \pi(2), \pi(a)\big)$ to the set $(1, 2, \ldots, b)$

$\pi_a^{\bar{b}}$ denotes the shrunken version of $\big(\pi(1), \pi(2), \pi(a)\big)$ to the set $(b+1, b+2, \ldots, n)$

$\pi_{\bar{a}}^b$ denotes the shrunken version of $\big(\pi(a+1), \pi(a+2), \pi(n)\big)$ to the set $(1, 2, \ldots, b)$

$\pi_{\bar{a}}^{\bar{b}}$ denotes the shrunken version of $\big(\pi(a+1), \pi(a+2), \pi(n)\big)$ to the set

$$(b+1, b+2, \ldots, n).$$

The product of the four functions in $P(\pi)$ means that random permutations in the intersection of $\mathcal{R}_a$ and $\mathcal{C}_b$ have the property that the events in the four quarters of the chequerboard are conditionally independent whenever the subsets of rows and columns occupied inside them is given and the occupied rows and columns in the left upper quarter are conditionally independent from the ones in the right lower quarter under the condition that the number of rows and columns is given. (Observe that the number of rows should be equal to the number of columns.)

Structural zeros in row-free distributions may appear independently: any conditional probability $C(u, V)$ may be zero as long as there is at least one permutation with positive probability. For free distributions structure zeros may be generated by the parameters f, g, but the intersection of row-free and column-free distributions with structural zeros is larger than this set. For $n = 4$ the uniform distribution concentrated on permutations such that only the permutations 1234, 2341, 2413, 2431, 3124, 3142, 3241, 4321 is free. However, for any set of the structural parameters $f(a, b, c)$, $g(a, b, c)$ such that the probabilities of the above eight permutations are positive, all permutations have positive probabilities.

Estimation of the parameters of row-free random permutations is straightforward: the estimators of the conditional probabilities $c(u, V)$ are the corresponding conditional relative frequencies. For free random permutations there are two iterative procedures:

- we can use alternating divergence projections on the sets $\mathcal{R}, \mathcal{C}$ or
- we can apply iterative fitting procedures on the statistics $\nu(a, b), \mu(a, b)$.

An exact implementation of these algorithms consume $n!$ steps what renders them to small n-s. Metropolis algorithm and Bayes machine apply both for generating i.i.d. free samples and estimating model parameters. As an estimator of the expectations in the likelihood equations we may use the averages of i.i.d. samples generated by the iteratively changing parameters.

Let $M \subset N$ be such that $2 \leq |M| \leq n-2$. Let us denote by L_M the set $\{\pi(i),\, i \in M\}$. We say that the random permutation has the M-cutting property if the ordered numbers $(\pi(i),\, i \in M)$ and $(\pi(i),\, i \notin M)$ are conditionally independent on L_M. Random permutations having M-cutting property for all M are the simple rook distributions. The number of model parameters may be reduced by controlling the rank of the matrix A. The rank has to be at least 2 because if it is equal to 1 then all elements of A are equal.

Partitions

Once upon a time there was a party with 14 participants labeled by integers from 1 to 14. As it is usual in parties they formed groups which were sensed and recorded by devices offered by our modern technology. The data can be found on the home page of G. Tusnády as SIRP DATA (http://www.renyi.hu/~tusnady/). The first part is given in Table 1.

Each record of the data represents one grouping (partition) formed in the course of the party. The first number means the time in hours when the actual grouping occurred and the next 14 integers denote the partition. Each set of a partition is labelled by its smallest number what we call leading member.

For example

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.900159	1	1	1	4	5	5	5	5	5	5	11	5	11	5

Table 1. First 24 records of SIRP.DATA

TIME	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.019238	1	1	1	4	5	6	5	6	5	5	5	6	13	5
0.064438	1	1	1	4	5	6	5	6	5	5	5	6	4	5
0.107385	1	1	1	4	5	6	5	6	5	5	5	6	13	5
0.119281	1	1	1	4	5	6	7	6	5	7	7	6	13	7
0.127421	1	1	1	4	5	5	7	5	5	7	7	5	13	7
0.159595	1	1	1	4	5	6	7	6	5	7	7	5	13	7
0.244247	1	1	1	4	5	6	1	6	5	1	1	5	13	1
0.246863	1	1	1	4	5	4	1	4	5	1	1	5	13	1
0.393910	1	2	2	4	5	4	1	4	5	1	1	5	13	1
0.466802	1	1	1	4	5	4	1	4	5	1	1	5	13	1
0.506604	1	1	1	4	5	4	7	4	5	7	7	5	13	7
0.518243	1	1	1	4	5	6	7	6	5	7	7	5	13	7
0.519593	1	1	1	4	5	4	7	4	5	7	7	5	13	7
0.576503	1	1	1	4	5	4	5	4	5	5	5	5	13	5
0.707155	1	1	1	4	5	4	5	4	5	5	11	5	13	5
0.716638	1	1	1	4	5	6	5	6	5	5	11	5	13	5
0.727348	1	1	1	4	5	6	5	6	5	5	11	5	11	5
0.733247	1	1	1	4	5	4	5	4	5	5	11	5	11	5
0.834109	1	1	1	4	5	6	5	6	5	5	11	5	11	5
0.900159	1	1	1	4	5	5	5	5	5	5	11	5	11	5
0.918424	1	1	1	4	5	6	5	6	5	5	11	5	11	5
0.998953	1	1	1	4	5	6	5	6	5	5	11	5	13	5
1.155627	1	1	1	4	5	1	5	1	5	5	11	5	13	5
1.252516	1	1	1	4	5	6	5	6	5	5	11	5	13	5

means that after 0.900159 hours from the beginning of the party the following groups were sensed by our detectors: $1+2+3$, 4, $5+6+7+8+9+10+12+14$, $11+13$. Poor 4, seemingly a lonely person walked alone, the noisy central body $5+6+7+8+9+10+12+14$ was situated around the dinner table, while $1+2+3$ had a very important discussion in a secret corner and $11+13$ were playing table tennis. There is a natural way to order a graph to partitions: the vertices are the participants and they are connected whenever they belong to the same group. However, partitions are special graphs, because they contain only disjunct complete subgraphs called sometime a clique.

Visualization of partitions. In multidimensional data analysis a general idea is to compress objects whenever they have something common. The trouble is that without any constraints the population shrinks to a single point. We use multidimensional covariance standardization as a constraint: the data are centered by subtracting their average, dividing them by the standard deviation and using covariances to keep the scales finite and non-zero. The effect resembles opening an umbrella: the wires spread out what the canopy pulls together.

In a good party there are appropriate places for people willing to do something together. But to use different positions for each subset is prohibitive: there is a combinatorial explosion. We restrict our algorithm to pairs: all pairs of the participants have same special meeting point and the groups are located at the average of the positions of their pairs. The formal description of the algorithm is the following.

Let S be the number of different partitions occurring in the party and let x_k, y_k be the coordinates of the point representing the k-th partition ($k = 1, 2, .., S$). The initial values of the coordinates are random standard normal numbers. The iteration consists of the following steps:

Step 1. Opening the umbrella (Schmidt orthogonalization):

$$\tilde{x}_k = \frac{x_k - \overline{x}}{w(x)},$$

where

$$\overline{x} = \frac{1}{S}\sum_{k=1}^{S} x_k,$$

and

$$w(x) = \sqrt{\sum_{k=1}^{S} (x_k - \overline{x})^2};$$

$$\tilde{y}_k = \frac{y_k^* - c(xy) * \tilde{x}_k}{w(y)},$$

where

$$y_k^* = y_k - \overline{y},$$

$$\overline{y} = \frac{1}{S}\sum_{k=1}^{S} y_k,$$

$$c(xy) = \sum_{k=1}^{S} y_k^* * \tilde{x}_k,$$

$$w(y) = \sqrt{\sum_{k=1}^{S} \left(y_k^* - c(xy) * \tilde{x}_k\right)^2}.$$

Step 2. Positioning pairs of persons (averaging the partitions where the given pair happens to be in the same group):

$$u_{i,j} = \frac{\sum_{k:\ p(i,k)=p(j,k)} \tilde{x}_k * t_k}{\sum_{k:\ p(i,k)=p(j,k)} t_k},$$

$$v_{i,j} = \frac{\sum_{k:\ p(i,k)=p(j,k)} \tilde{y}_k * t_k}{\sum_{k:\ p(i,k)=p(j,k)} t_k},$$

where $p(i,k)$ denotes the leading person of the group containing the i-th person in the k-th partition and t_k is the duration of the k-th partition.

Step 3. Dynamics (relocating the partitions with the gradient of the pairs they unite):

$$x_k^{\text{new}} = \tilde{x}_k - \gamma * xx_k,$$

where γ denotes a small positive constant (it controls the speed of the algorithm) and

$$xx_k = \sum_{i,j:\ p(i,k)=p(j,k)} u_{i,j},$$

$$y_k^{\text{new}} = \tilde{y}_k - \gamma * yy_k,$$

where

$$yy_k = \sum_{i,j:\ p(i,k)=p(j,k)} v_{i,j}.$$

The pair potential model. Our data are generated by the distribution

$$P_A(\pi) = \frac{1}{\Gamma(A)} \exp\big(Q(\pi, A)\big),$$

where the potential $Q(\pi, A)$ is defined by

$$Q(\pi, A) = \sum_{1 \le i < j \le c:\ \pi(i) = \pi(j)} a_{i,j},$$

and

$$\Gamma(A) = \sum \exp\big(Q(\pi, A)\big)$$

is the scaling factor where the summation runs over all partitions π. The matrix $A = a_{i,j}$ is symmetric and given by Table 2. The maximum of the potential $Q(\pi, A)$ is 39.64 and it is attained for the partition

$$\pi = \{1, 1, 1, 4, 5, 6, 5, 6, 5, 5, 5, 6, 13, 5\}.$$

The distribution can be sampled by the Metropolis algorithm [26], which is based on a graph where the vertices are the partitions. We say that two partitions are connected by an edge whenever one is formed by the other with uniting two of its groups. The price of the edge is the product of the numbers of persons in the united groups. The distance of two arbitrary vertices is the price of the cheapest path between them. This is

$$d(\pi_1, \pi_2) = \sum_{i=1}^{n} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \big(\nu(i,j)\nu(i,k) + \nu(j,i)\nu(k,i)\big),$$

where

$$\nu(i,j) = \sum_{k=1}^{n} \mathcal{I}\big(\pi_1(k) = i,\ \pi_2(k) = j\big),$$

where n is the number of persons.

Concerning the partition function $\Gamma(A)$ one can prove that

$$E_A \exp\big(Q(\pi, B)\big) = \frac{\Gamma(A+B)}{\Gamma(A)},$$

and

$$\Gamma(A) \le \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \big(1 + \exp(a_{i,j})\big),$$

Table 2. Model parameters of SIRP.DATA

0.00	−3.08	−0.40	0.23	0.75	−1.23	1.13	−0.02	0.51	−2.62	0.78	2.93	0.05	−2.66
−3.08	0.00	−4.63	1.18	1.43	1.19	4.73	−0.59	2.61	−5.01	2.08	1.94	3.91	1.94
−0.40	−4.63	0.00	4.31	1.65	2.94	−1.17	0.87	−1.37	−0.72	1.66	−0.92	0.19	0.01
0.23	1.18	4.31	0.00	2.15	−1.10	−2.77	1.63	4.13	6.79	6.29	1.51	0.41	2.92
0.75	1.43	1.65	2.15	0.00	−0.06	−2.31	−0.35	−1.52	−2.30	2.48	−2.23	2.25	−0.72
−1.23	1.19	2.94	−1.10	−0.06	0.00	2.75	−4.47	1.29	−4.41	0.38	−4.07	3.75	2.54
1.13	4.73	−1.17	−2.77	−2.31	2.75	0.00	−2.86	−2.50	−6.20	−3.35	0.30	3.98	−0.68
−0.02	−0.59	0.87	1.63	−0.35	−4.47	−2.86	0.00	−1.16	3.12	−0.27	3.61	5.84	2.25
0.51	2.61	−1.37	4.13	−1.52	1.29	−2.50	−1.16	0.00	3.29	5.46	−1.05	1.64	−7.29
−2.62	−5.01	−0.72	6.79	−2.30	−4.41	−6.20	3.12	3.29	0.00	−5.14	−2.20	−1.12	−1.77
0.78	2.08	1.66	6.29	2.48	0.38	−3.35	−0.27	5.46	−5.14	0.00	5.45	−0.32	−4.20
2.93	1.94	−0.92	1.51	−2.23	−4.07	0.30	3.61	−1.05	−2.20	5.45	0.00	3.07	1.74
0.05	3.91	0.19	0.41	2.25	3.75	3.98	5.84	1.64	−1.12	−0.32	3.07	0.00	2.82
−2.66	1.94	0.01	2.92	−0.72	2.54	−0.68	2.25	−7.29	−1.77	−4.20	1.74	2.82	0.00

but we do not have an explicit form for $\Gamma(A)$. We generated the matrix A as random Gaussian number with zero expectation and standard deviation 2.5 thus the model has the flavor of spin glass processes: there is an abundance of local maxima of the potential and the process spends the majority of the time in the potential valleys with short time jumps between them. This might be the case with real world parties where the different partitions are evaluated by the well-being of the persons inside the actual groups. Our model is the simplest possible one because it is based on pair-relations only. Generalization to higher order interactions is straightforward.

Estimation of model parameters. The pair-potentials $a_{i,j}$ can be estimated by the maximum likelihood equation (see in [2], [16])

$$\frac{1}{S}\sum_{k:\ p(i,k)=p(j,k)} 1 = P_A\big(\pi(i)=\pi(j)\big) \qquad 1 \le i < j \le n,$$

or by simple weighted linear regression. The probabilities $b_{i,j} = P_A\big(\pi(i) = \pi(j)\big)$ are given in Table 3.

Corresponding relative frequencies $\beta_{i,j}$ are given in Table 4. The pair potential model is loglinear: the logarithms of probabilities $P_A(\pi)$ are linear functions of the model parameters $a_{i,j}$. There is no direct relation between $a_{i,j}$ and $b_{i,j}$, and the $\beta_{i,j}$ relative frequencies are closer to the theoretical probabilities $b_{i,j}$ than the estimators of $a_{i,j}$ to $a_{i,j}$. The estimation of the model parameters is a typical ill-conditioned problem and to compare different data sets, the $b_{i,j}$ parameters may be more useful.

The specific feature of our data is that successive partitions can be either the union or the splitting of the previous one. For the first part of our data

Table 3. Probabilities of equivalence

.000	.750	.710	.018	.075	.087	.238	.108	.111	.413	.319	.024	.022	.294
.750	.000	.937	.006	.006	.051	.080	.085	.035	.256	.161	.034	.010	.123
.710	.937	.000	.001	.025	.029	.101	.066	.068	.249	.155	.055	.016	.149
.018	.006	.001	.000	.005	.142	.011	.087	.001	.000	.000	.076	.301	.000
.075	.006	.025	.005	.000	.279	.541	.287	.672	.415	.342	.494	.002	.530
.087	.051	.029	.142	.279	.000	.161	.663	.139	.208	.148	.548	.001	.014
.238	.080	.101	.011	.541	.161	.000	.279	.452	.792	.776	.129	.004	.682
.108	.085	.066	.087	.287	.663	.279	.000	.257	.167	.210	.244	.000	.101
.111	.035	.068	.001	.672	.139	.452	.257	.000	.278	.268	.363	.006	.678
.413	.256	.249	.000	.415	.208	.792	.167	.278	.000	.816	.123	.010	.592
.319	.161	.155	.000	.342	.148	.776	.210	.268	.816	.000	.021	.058	.588
.024	.034	.055	.076	.494	.548	.129	.244	.363	.123	.021	.000	.005	.156
.022	.010	.016	.301	.002	.001	.004	.000	.006	.010	.058	.005	.000	.002
.294	.123	.149	.000	.530	.014	.682	.101	.678	.592	.588	.156	.002	.000

Table 4. Relative frequencies of equivalence

.000	.773	.737	.012	.077	.084	.243	.100	.111	.423	.326	.023	.016	.297
.773	.000	.948	.004	.006	.054	.085	.079	.032	.266	.167	.030	.007	.130
.736	.948	.000	.001	.027	.030	.105	.058	.066	.259	.161	.055	.011	.153
.012	.004	.001	.000	.002	.118	.011	.070	.000	.000	.000	.064	.282	.000
.077	.006	.027	.002	.000	.292	.563	.314	.712	.430	.366	.510	.001	.554
.084	.054	.030	.118	.292	.000	.166	.688	.154	.208	.151	.566	.000	.015
.243	.085	.105	.011	.563	.166	.000	.287	.483	.794	.783	.138	.001	.706
.100	.079	.058	.070	.314	.688	.287	.000	.274	.173	.213	.274	.000	.108
.111	.032	.066	.000	.712	.154	.483	.274	.000	.307	.300	.384	.003	.692
.423	.266	.259	.000	.430	.208	.794	.173	.307	.000	.823	.126	.007	.610
.326	.169	.161	.000	.366	.151	.783	.213	.300	.823	.000	.029	.052	.607
.023	.030	.055	.064	.510	.566	.138	.274	.384	.126	.029	.000	.002	.165
.016	.007	.011	.282	.001	.000	.001	.000	.003	.007	.052	.002	.000	.001
.297	.130	.153	.000	.554	.015	.706	.108	.692	.610	.607	.165	.001	.000

the operations are given in Table 5. Having the information at hand that the data were generated by the Metropolis [26] algorithm, one may develop more efficient estimators. The abundance of inverted pairs of union and splitting among the operators is remarkable here.

Table 5. Operators of data in Table 1

NAME	FIRST GROUP	SIGN	SECOND GROUP
UNION	4	+	13
SPLITTING	4	‖	13
SPLITTING	5, 9	‖	7, 10, 11, 14
UNION	5, 9	+	6, 8, 12
SPLITTING	5, 9, 12	‖	6, 8
UNION	1, 2, 3	+	7, 10, 11, 14
UNION	4	+	6, 8
SPLITTING	1, 7, 10, 11, 14	‖	2, 3
UNION	1, 7, 10, 11, 14	+	2, 3
SPLITTING	1, 2, 3	‖	7, 10, 11, 14
SPLITTING	4	‖	6, 8
UNION	4	+	6, 8
UNION	5, 9, 12	+	7, 10, 11, 14
SPLITTING	5, 7, 9, 10, 12, 14	‖	11
SPLITTING	4	‖	6, 8
UNION	11	+	13
UNION	4	+	6, 8
SPLITTING	4	‖	6, 8
UNION	5, 7, 9, 10, 12, 14	+	6, 8
SPLITTING	5, 7, 9, 10, 12, 14	‖	6, 8
SPLITTING	11	‖	13
UNION	1, 2, 3	+	6, 8
SPLITTING	1, 2, 3	‖	6, 8

Independent participants. One may ask at this point whether any simpler stochastic model would be able to generate the same $\beta_{i,j}$ frequencies. In the above model the participants are intrinsically correlated because the whole matrix A is involved forming the probabilities of groups. The following model emerges from the idea of independence. Let us offer the possibility to the participants of the party to choose *independently* from finitely many options, like:

- to have a delicate food,
- to play hide and seek,
- to watch TV,
- to discuss Shakespeare,

– to make a small excursion.

With the program in hand, people having their preferences make their choices independently in the programs and the groups are formed in a natural way by the programs. Let us denote by $w(i,r)$ the probability that the i-th participant chooses the r-th possibility then

$$P\big(\pi(i)=\pi(j)\big)=\sum_{r=1}^{R} w(i,r)w(j,r),$$

where R denotes the number of possibilities.

Nonnegative matrix factorization was investigated in [13]. We have the constraint

$$\sum_{r=1}^{R} w(i,r)=1, \qquad 1\le i\le n,$$

which leads to a poor fit of our data. Interestingly, dropping the constraint, the frequencies $\beta_{i,j}$ has a good factorization with $w(i,r)$ given in the Table 6. If the number of participants goes to infinity, the size of groups is the most important feature of the distribution. It may remain bounded or slowly increasing as it is the case in politics when the groups are the political parties having the tendency to become of small number mostly because preference choice. The second possibility is the square root law: the size of groups and their number both are around the square root of n. Third possibility is represented in chemistry: the size of groups remains small and the number of groups increases with c for example for proteins. In the independent model the situation is easily controlled by R but in the case of pair-potential model we do not know the answer. Our guess is the third possibility on the argument that $Q(\pi,A)$ may achieve the size n^2 for π with small groups. A natural way to control the size of groups is to add a constant to the pair-potentials, i.e. to apply the pair-potentials $\tilde{a}_{i,j}=a_{i,j}+\Delta$. Negative Δ shrinks the groups and positive Δ increases them. When all $\tilde{a}_{i,j}$ become positive, all participants are in the same group with large probability. As a matter of fact, independence is not far from the pair-potential model: it is equivalent to the random graph model conditioned on the restriction to graphs representing only partitions. For large c we substituted all of the $a_{i,j}$-s with zero and used the parameter Δ only. According to our computer experiments $\Delta=0.03$ seems to be the critical value for $n=10\,000$.

Table 6. Factorization of equivalence probabilities

0.23545	0.77188	–	0.06091	0.01306	0.05126	0.01362
–	0.99511	0.00546	0.03724	0.00529	–	–
0.00937	0.94614	0.03934	0.00191	0.00865	0.02693	–
–	–	0.01632	0.10669	0.21538	–	–
0.25981	–	0.62682	0.03831	–	0.37697	0.45547
0.04035	0.01241	0.37992	1.09130	–	–	–
0.72835	0.08377	0.12272	0.08269	–	0.14629	0.51251
–	0.05948	0.16962	0.56842	–	0.00001	0.41019
0.15838	0.03633	0.34507	0.02085	0.00001	0.63947	0.46404
0.86278	0.26262	0.15723	0.09302	0.00299	–	0.21790
0.79914	0.16356	–	0.10403	0.03747	0.00339	0.34943
–	0.01224	0.65654	0.28753	0.00381	0.22826	–
–	–	–	–	1.30074	–	0.00391
0.60732	0.13486	–	–	–	0.72583	0.26880

Checkerboard model. Parties and hypergraphs are appeared as early as 1941 in the literature [11] where 18 ladies attending on 14 parties are investigated. It is a special case of partitions when only two groups are considered, whether each person is present or absent in a party. In the next table we show the application of checkerboard model to the Table 7. We reordered slightly the ladies and parties and grouped them into 8 and 6 clusters respectively. The probabilities that a lady belonging to the ith cluster takes part in the party belonging to the jth cluster are given in Table 8.

There are 22 pairs of (i, j)-s with probability zero, for example $i = 3$, $j = 6$, accordingly ladies 5, 6, 7 did not attended in parties 10, 12, 13, 14. There are 12 pairs of (i, j)-s with probability one, for example $i = 1$, $j = 2$. accordingly ladies 1, 3 attended in parties 3, 5, 6. For the remaining 14 pairs the range of probabilities are between 0.11 and 0.83. There is only one pair ($i = 7$, $j = 4$) with probability 0.5 which means the maximal uncertainty. In microarray analysis this is the so-called chequerboard structure. We characterize the uncertainty of the data with the reciprocal of the delogarithmized averaged log-likelihood which is 1.205627. In case this quantity equals 2, the uncertainty is maximal, for all (i, j) pairs the probabilities are equal to 0.5. We can test the power of the model by mixing randomly the bits in the data. In this case the uncertainty is between 1.36 and 1.42.

Table 7. Davis – Gardner – Gardner data

		1	2	4	3	5	6	7	8	9	11	10	12	13	14
		1	1	1	2	2	2	3	3	4	5	6	6	6	6
1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0
3	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0
2	2	1	1	0	1	1	1	1	1	0	0	0	0	0	0
4	2	1	0	1	1	1	1	1	1	0	0	0	0	0	0
5	3	0	0	1	1	1	0	1	0	0	0	0	0	0	0
6	3	0	0	0	1	1	1	0	1	0	0	0	0	0	0
7	3	0	0	0	0	1	1	1	1	0	0	0	0	0	0
8	4	0	0	0	0	0	1	0	1	1	0	0	0	0	0
9	4	0	0	0	0	1	0	1	1	1	0	0	0	0	0
10	5	0	0	0	0	0	0	1	1	1	0	0	1	0	0
11	5	0	0	0	0	0	0	0	1	1	0	1	1	0	0
16	5	0	0	0	0	0	0	0	1	1	0	1	1	0	0
12	6	0	0	0	0	0	0	0	1	1	0	1	1	1	1
13	6	0	0	0	0	0	0	1	1	1	0	1	1	1	1
14	7	0	0	0	0	0	1	1	0	1	1	1	1	1	1
15	7	0	0	0	0	0	0	1	1	0	1	1	1	1	1
17	8	0	0	0	0	0	0	0	0	1	1	0	0	0	0
18	8	0	0	0	0	0	0	0	0	1	1	0	0	0	0

Table 8. Structural probabilities of checkerboard model

	1	2	3	4	5	6
1	0.83	1	0.75	1	0	0
2	0.67	1	1	0	0	0
3	0.11	0.78	0.67	0	0	0
4	0	0.33	0.75	1	0	0
5	0	0	0.67	1	0	0.42
6	0	0	0.75	1	0	1
7	0	0.17	0.75	0.50	1	1
8	0	0	0	1	1	0

Cluster numbers 8 and 6 seem to be large, considering the numbers of ladies and parties but with smaller cluster numbers we were unable to present satisfactory clustering. In statistical investigations, in cluster analysis partitions appear mostly in the following two different aspects:

– we may form groups from the investigated objects on the basis that any connection is possible only inside the groups
– we may form the groups of similar objects

The second possibility is used in checkerboard model. Its extension to the pair-potential model is a numbering $f(k)$, $k = 1, \ldots, n$, of the participants such that

– $1 \leq f(k) \leq g$; $k = 1, \ldots, n$
– for all $1 \leq j \leq g$ there is a $1 \leq i \leq c$ such that $f(i) = j$
– there is a $g * g$ matrix D with entries $d_{u,v}$ such that $a_{i,j} = d_{f(i),f(j)}$ for all $1 \leq i < j \leq n$

this is called blown-up of the matrix D into matrix A. We investigated partition-clustering in [30] and interactive networks in [31]. A widely investigated process on partitions is Kingman's coalescent process [19]. In the Table 9. we give the groups where the first participants spent the most time.

The number of partitions. There is a recursion for P_n which denotes the number of partitions of n elements:

$$P_{n+1} = \sum_{j=0}^{n} \binom{n}{j} P_j, \qquad n = 0, 1, \ldots,$$

where $P_0 = 1$. (Especially $P_{14} = 190,899,322$.) There is an explicit form as well for P_n,

$$P_n = \sum_{j=1}^{n} j^n \frac{\delta(n-j)}{j!},$$

where

$$\delta(k) = \sum_{s=0}^{k} \frac{(-1)^s}{s!}.$$

Acknowledgements. We thank to R. Albert, D. Bancroft, L. Barabási, I. Bárány, M. Bolla, T. Breuer, I. Csiszár, K. Friedl, P. Hussami, M. Ispány, J. Komlós, A. Krámli, L. Lovász, K. Marton, I. Miklós, M. Simonovits, G. Simonyi, V. T. Sós, E. Szemerédi and T. Vicsek the enlightening conversations.

Table 9. Most frequent partitions in SIRP.DATA

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
7271.18	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1370.10	1	1	1	0	0	0	0	0	0	1	0	0	0	0
1134.26	1	1	1	0	0	0	1	0	0	1	1	0	0	1
1071.91	1	0	0	0	0	0	1	0	0	1	1	0	0	1
724.33	1	1	1	0	0	0	0	0	0	1	1	0	0	1
713.87	1	0	0	0	1	0	1	0	1	1	1	0	0	1
666.55	1	1	1	0	0	0	0	0	0	1	1	0	1	0
654.82	1	1	1	0	0	0	0	0	0	1	1	0	0	0
536.01	1	1	1	0	0	0	0	1	0	0	0	0	0	0
445.13	1	0	0	0	1	0	1	0	0	1	1	0	0	1
407.04	1	0	0	0	0	0	0	0	0	0	1	0	0	1
329.61	1	1	1	0	0	1	0	0	0	1	0	1	0	0
313.40	1	1	1	0	0	0	0	0	1	0	0	0	0	1
261.53	1	0	0	0	0	0	0	0	1	0	0	0	0	1
241.93	1	1	0	0	0	1	0	1	0	1	0	0	0	0
229.54	1	0	0	0	0	0	0	0	0	0	0	0	0	0
204.26	1	0	0	0	0	1	1	1	0	1	1	0	0	0
168.22	1	0	0	0	0	1	0	1	0	0	0	0	0	0
159.01	1	1	1	0	0	0	1	0	0	1	1	0	0	0
154.32	1	0	0	0	0	0	1	0	1	1	1	0	0	1
118.99	1	1	1	0	0	1	0	0	0	1	0	0	0	0
118.45	1	1	0	0	0	0	0	0	0	0	0	0	0	0

References

[1] A. L. Barabási and R. Albert, Emergence of scaling in random networks, *Science,* **286** (1999), 509–512.

[2] O. Barndorff-Nielsen, *Information and Exponential families in statistical theory,* Chichester (Wiley, 1978).

[3] M. Bolla, Distribution of the eigenvalues of random block-matrices, *Linear Algebra and its Applications,* **377** (2004), 219–240.

[4] M. Bolla, Recognizing linear structure in noisy matrices, *Linear Algebra and its Applications,* **402** (2005), 228–240.

[5] M. Bolla and G. Tusnády, Spectra and optimal partitions of weighted graphs, *Discrete Mathematics,* **128** (1994), 1–20.

[6] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády, The degree sequence of a scale-free random graph, *Random Structures Algorithms,* **18** (2001), 279–290.

[7] C. Borgs, J. Chayes, L. Lovász, V. T. Sós, B. Szegedy and K. Vesztergombi, Counting graph homomorphisms, in: *Topics in Discrete Mathematics* (eds. M. Klazar, J. Kratochvil, M. Loebl, J. Matousek, R. Thomas, P. Valtr), 315–371 (Springer, 2006).

[8] C. Borgs, J. Chayes, L. Lovász, V. T. Sós and K. Vesztergombi, *Graph limits and parameter testing,* STOC (2006).

[9] C. Borgs, J. Chayes, L. Lovász, V. T. Sós and K. Vesztergombi, *Convergent sequences of dense graphs I: subgraph frequencies, metric properties and testing,* arXiv:math.CO/0702004v1 31Jan2007.

[10] F. Chung, *Spectral graph theory,* revised (2006).

[11] A. Davis, B. B. Gardner and M. R. Gardner, *Deep south: a social anthropological study of caste and class,* University of Chicago Press (1941).

[12] P. Diaconis, A generalization of spectral analysis with applications to ranked data, The 1987 Wald memorial lectures, *The Annals of Statistics,* **17** (1989), 949–979.

[13] L. Finesso and P. Spreij, Nonnegative matrix factorization and I-divergence alternating minimization, *Linear Algebra and its Applications,* **416** (2006), 270–287.

[14] A. Frieze and R. Kannan, *Quick approximation to matrices and applications,* manuscript (2006).

[15] J. Gupta and P. Damien, Conjugacy class prior distributions on metric-based ranking models, *Journal of the Royal Statistical Society Series B,* **64** (2002), 433–445.

[16] X. Guyon, *Random fields on a network, modelling, statistics and applications,* Springer (1995).

[17] F. Hiai and D. Petz, The semicircle law, free random variables and entropy, *AMS Mathematical Surveys and Monograph,* **77** (2000).

[18] O. Khorunzhiy, W. Kirch and P. Müller, Lifschitz tails for spectra of Erdős–Rényi random graphs, *The Annals of Applied Probability,* **16** (2006), 295–309.

[19] J. F. C. Kingman, Origins of the coalescent: 1974–1982, *Genetics,* **156** (2000), 1461–1463.

[20] W. König, Orthogonal polynomial ensembles in probability theory, *Probability Survey,* **2** (2005), 385–447.

[21] L. Lovász and B. Szegedy, *Szemerédi's regularity lemma for the analyst,* manuscript (2006).

[22] C. L. Mallows, Non null ranking models I. *Biometrika,* **44** (1957), 114–130.

[23] B. D. McKay, The expected eigenvalue distribution of a large regular graph, *Linear Algebra and its Applications,* **40** (1981), 203–216.

[24] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions,* John Wiley &Sons (New York, 1997).

[25] M. L. Mehta, *Random matrices,* 3rd edn, New York Cademic Press (2004).

[26] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, Equations of state calculations by fast computing machines, *Journal of Chemical Physics,* **21** (1953), 1087–1092.

[27] T. Nepusz, L. Négyessy, G. Tusnády and F. Bazsó, *Predicting key areas and uncharted connections in the cerebral cortex using Szemerédi's regularity lemma,* manuscript (2006).

[28] R. L. Plackett, The analysis of permutations, *Appl. Statist.,* **24** (1975), 193–202.

[29] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE,* **77** (1989), 257–286.

[30] L. Rejtő and G. Tusnády, Clustering methods in microarrays, *Periodica Mathematica Hungarica,* **50** (2005), 199–221.

[31] L. Rejtő and G. Tusnády, Reconstruction of Kauffman networks applying trees, *Linear Algebra and Application,* **417** (2006), 220–244.

[32] C. G. Small, Multidimensional medians arising from geodesics on graphs, *Annals of Statistics,* **25** (1997), 478–494.

[33] H. Stern, Models for distributions on permutations, *Journal of the American Statistical Association,* **85** (1990), 558–564.

[34] E. Szemerédi, Regular partitions of graphs, *Colloquies Internationale C.N.R.S.,* **260**, Proalèkes Combinatorics et Théories des Graphes, Oracy (1976), pp. 399–401.

[35] A. D. Szlam, M. Maggioni, R. R. Coifman and J. C. Bremer, Jr.: Diffusion-driven multiscale analysis on manifolds and graphs: top-down and bottom-up constructions, *SPIE Wavelets XI,* **5914** (2005).

[36] G. M. Tallis and B. R. Dansie, An alternative approach to the analysis of permutations, *Appl. Statist.,* **32** (1983), 110–114.

[37] T. Tao, *Szemerédi's regularity lemma revisited,* arXiv:math.CO/0504472v2 16Nov2005.

[38] L. L. Thurstone, A law for comparative judgement, *Psychological Review,* **34** (1927), 278–286.

Villő Csiszár, Lídia Rejtő & Gábor Tusnády

Rényi Institute
Budapest, Hungary

e-mail: villo@renyi.hu, rejto@renyi.hu, tusnady@renyi.hu

Appendix

sign: vertex label d: degree, a: asset, w: wealth, n_i: i-th neighbor

sign	d	a	w	n_1	n_2	n_3	n_4	n_5	n_6	sign	d	a	w	n_1	n_2	n_3	n_4	n_5
A	6	109	321	A1	A2	A3	A4	A5	A6	u	5	16	16	f5	g5	h5	i5	j5
B	6	134	403	B1	B2	B3	B4	B5	B6	v	5	17	12	p5	q5	r5	s5	t5
C	6	152	431	C1	C2	C3	C4	C5	C6	w	5	16	12	f6	g6	h6	i6	j6
D	6	149	421	D1	D2	D3	D4	D5	D6	x	5	17	16	p6	q6	r6	s6	t6
E	6	167	459	E1	E2	E3	E4	E5	E6	a6	4	0	0	05	A2	A6		
F	6	159	451	F1	F2	F3	F4	F5	F6	B1	4	0	0	b2	l1	c4		
G	6	158	440	G1	G2	G3	G4	G5	G6	B2	4	0	0	b2	l2	d4		
H	6	143	428	H1	H2	H3	H4	H5	H6	b2	4	0	0	B3				
I	6	151	397	I1	I2	I3	I4	I5	I6	c1	4	0	0	25	35	C1		
J	6	155	467	J1	J2	J3	J4	J5	J6	c2	4	0	0	08	11	C2		
a	5	16	20	a1	b1	c1	d1	e1		D1	4	0	0	d3	d1	n1		
b	5	15	20	k1	l1	m1	n1	o1		d2	4	0	0	11	D2	02		
c	5	17	8	a2	b2	c2	d2	e2		n2	4	0	0	11	26	D2		
d	5	16	16	k2	l2	m2	n2	o2		D3	4	0	0	d3	a1	n3		
e	5	17	12	a3	b3	c3	d3	e3		d3	4	0	0	26				
f	5	15	24	k3	l3	m3	n3	o3		e3	4	0	0	03	14	E3		
g	5	15	20	a4	b4	c4	d4	e4		f1	4	0	0	25	28	F1		
h	5	15	24	k4	l4	m4	n4	o4		F2	4	0	0	f3	f2	s3		
i	5	15	24	a5	b5	c5	d5	e5		f3	4	0	0	08	F3			
j	5	12	16	k5	l5	m5	n5	o5		q2	4	0	0	05	07	11		
k	5	16	16	a6	b6	c6	d6	e6		g4	4	0	0	04	28	G4		
l	5	15	24	k6	l6	m6	n6	o6		g6	4	0	0	07	14	G6		
m	5	17	12	f1	g1	h1	i1	j1		q6	4	0	0	08	G6	02		
n	5	15	12	p1	q1	r1	s1	t1		r2	4	0	0	H2	H3	40		
o	5	14	16	f2	g2	h2	i2	j2		H4	4	0	0	r4	g2	h4		
p	5	18	12	p2	q2	r2	s2	t2		r4	4	0	0	04	28			
q	5	16	16	f3	g3	h3	i3	j3		r5	4	0	0	03	13	H5		
r	5	13	16	p3	q3	r3	s3	t3		r6	4	0	0	13	35	H6		
s	5	16	20	f4	g4	h4	i4	j4		s2	4	0	0	08	13	I2		
t	5	16	16	p4	q4	r4	s4	t4		i5	4	0	0	07	25	I5		

Appendix (continued)

sign	d	a	w	n_1	n_2	n_3	n_4	sign	d	a	w	n_1	n_2	n_3	n_4	sign	d	a	w	n_1	n_2
s5	4	0	0	35	I1	I5		32	4	4	0	a5	l3	d1	f4	E5	3	0	0	e5	
j1	4	0	0	03	07	J1		33	4	4	0	q5	h3	s3	t6	o5	3	0	0	A4	
t1	4	0	0	26	J1	40		34	4	4	0	k2	b1	o4	g5	E6	3	0	0	e6	o6
J5	4	0	0	i1	j5	t5		35	4	0	0	h3				F1	3	0	0	p1	
j6	4	0	0	05	14	J6		36	4	4	0	m2	p1	g1	r1	F3	3	0	0	p3	
01	4	4	0	m6	p2	t4	j5	37	4	4	0	l4	m3	o6	h6	p4	3	0	0	F4	
03	4	0	0	a2				38	4	4	0	d4	n6	o2	m5	f5	3	0	0	F5	
04	4	0	0	a4	g5			39	4	4	0	f4	q1	g2	i2	p5	3	0	0	H2	
05	4	0	0	p6				A1	3	0	0	k1	l5			f6	3	0	0	F6	
06	4	4	0	o3	h4	h5	i1	A2	3	0	0	k2				g1	3	0	0	G1	
09	4	4	0	a4	l1	c6	f5	a3	3	0	0	A3				g3	3	0	0	40	
10	4	4	0	l4	b5	e1	i6	k3	3	0	0	E5				q3	3	0	0	G5	
12	4	4	0	e2	p3	g3	t2	k5	3	0	0	A5				G4	3	0	0	q4	
13	4	0	0	j3				A6	3	0	0	k6				G5	3	0	0	q5	
14	4	0	0	q1				b1	3	0	0	D4				H1	3	0	0	r1	j4
15	4	4	0	a3	k4	n1	d6	B3	3	0	0	b3				h1	3	0	0	J2	
16	4	4	0	o1	p6	j2	j3	B4	3	0	0	b4	o2			h2	3	0	0	G2	
17	4	4	0	n3	e1	e6	q4	b5	3	0	0	B5				H3	3	0	0	t6	
18	4	4	0	b3	n4	s1	t5	B6	3	0	0	l6	e4			r3	3	0	0	J3	
19	4	4	0	k1	b4	h1	n5	C1	3	0	0	m1				h5	3	0	0	I6	
20	4	4	0	a2	l5	c5	o5	C2	3	0	0	m2				h6	3	0	0	I4	
21	4	4	0	m1	p5	i3	s6	C3	3	0	0	c3	m3			I1	3	0	0	s1	
22	4	4	0	k5	k6	l2	c4	m4	3	0	0	C4				I2	3	0	0	i2	
23	4	4	0	a1	a5	q3	t3	c5	3	0	0	D6				i3	3	0	0	02	
24	4	4	0	m6	p2	s4	s6	c6	3	0	0	E1				I4	3	0	0	s4	
25	4	0	0	b6				n4	3	0	0	i4				I6	3	0	0	i6	
26	4	0	0	d6				d5	3	0	0	D5				j2	3	0	0	G3	
27	4	4	0	m4	e4	p4	i4	D6	3	0	0	n6				t2	3	0	0	I3	
28	4	0	0	k4				E1	3	0	0	o1				J3	3	0	0	t3	
29	4	4	0	b6	e5	f6	j4	e2	3	0	0	E2				t4	3	0	0	J4	
30	4	4	0	k3	l6	c3	f2	E3	3	0	0	o3									
31	4	4	0	l3	d5	h2	r3	o4	3	0	0	E4									