Measures of clustering in binary choice models

Paulina Hetman
Andrzej Janutka
Piotr Magnuszewski
Andrzej Radosz

Institute of Physics
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
50–370 Wrocław, Poland

Wroclaw, June 2006

Contents

Chapter 1

Introduction

1.1 Network. Basic notions and properties

Let us consider a population of N individuals (nodes) labeled as $1, 2, \ldots, N$ with a certain configuration of links between them that form a network. The network structure (links) is represented by a $N \times N$ matrix $\mathbf{D} = (d_{ij})$ where d_{ij} stands for the "distance" from the node i to the node j. In general d has not to have the sense of the distance as it lacks symmetry in the directed graph case. One of the possible distance definitions is as follows

$$d_{ij} = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{if } i \text{ is connected to } j \\ \infty, & \text{if } i \text{ is not connected to } j \end{cases} \tag{1.1}$$

with the additional assumption that i is connected to j $(i \mapsto j)$ iff $j \mapsto i$ yielding the symmetric matrix $\mathbf{D}$, i.e. $\mathbf{D} = \mathbf{D}^T$.

Because of the uncomfortable infinity symbol in $\mathbf{D}$ we introduce the matrix of connectivity (nearness [10]) $\Delta = (\delta_{ij})$ such that

$$\delta_{ij} = \begin{cases} 0, & d_{ij} \neq 1, \\ 1, & d_{ij} = 1. \end{cases}$$

Notice that each node is considered unconnected with itself.

For some of the further presented measures of clustering the generalization of $\mathbf{D}$ is possible in the sense that $\mathbf{D}$ contains the weighted distances between nodes and not only the binary information on the presence or lack of connection.

- **Neighborhood**
  Each node i in the network has its neighborhood denoted $\Gamma(i)$ which is a set of nodes that are directly connected with i. In general case of (the directed graph) we define the in-neighborhood and the out-neighborhood as, respectively, the set of nodes that are connected to i and the set of nodes that i is connected to. More precisely

$$\Gamma_{in}(i) = \{j \in \{1, 2, \ldots, N\} : d_{ij} = 1\}$$

$$\Gamma_{out}(i) = \{j \in \{1, 2, \ldots, N\} : d_{ji} = 1\}$$

or alternatively

$$\Gamma_{in}(i) = \{j \in \{1, 2, \ldots, N\} : \delta_{ij} = 1\}$$

$$\Gamma_{out}(i) = \{j \in \{1, 2, \ldots, N\} : \delta_{ji} = 1\}$$

If the connections in the network form an undirected graph $\Gamma_{in} = \Gamma_{out} = \Gamma$.

- **Degree. In-degree. Out-degree**
  If the matrix $\mathbf{D}$ is symmetric (undirected case) the degree of ith node k_i is the number of elements in $\Gamma(i)$, $k_i = \overline{\overline{\Gamma(i)}}$. In the general case the notions of in-degree k^{in} and out-degree k^{out} are distinguished [10]. In the matrix notation $\mathbf{k^{out}} = \mathbf{\Delta 1}$, $\mathbf{k^{in}} = \mathbf{\Delta}^T\mathbf{1}$, where $\mathbf{1}$ is a $N \times 1$ column vector of ones. In the symmetric case $\mathbf{k^{in}} = \mathbf{k^{out}} = \mathbf{k}$. The total number of connections (pairs) S if all the $i \mapsto j$ and $j \mapsto i$ links are counted separately is given as $S = \mathbf{1^T \Delta 1} = \mathbf{k^T 1}$.

- **Regular lattice**
  In some cases we will refer to the notion of regular lattice with periodic boundary conditions. For convenience, we consider only regular network with the degree n being an even number. An exemplary connectivity matrix Δ for the 4-neighbors regular lattice with 8 nodes is of the form

$$\mathbf{\Delta} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

  which gives an idea of the construction of $\mathbf{\Delta}$.

  Notice that in the n-regular matrix the total number of pairs $S = nN$.

- **Binary choice**
  Each node (i) at a fixed time is characterized by its decision (attitude), being the result of its preliminary binary choice (see 1.2). The decision is denoted by variable σ_i taking the values -1 or 1. Some authors prefer the 0 and 1 distinction. We use the following notation:

$$\sigma = [\sigma_1, \ldots, \sigma_N]^T$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_N \end{bmatrix}$$

The majority of clustering measures bases on the notion of the number of pairs of neighbors that share the same attitude (pairs of common neighbors). Either the total number of such pairs (with or without the +1 and −1 pairs distinction) is of interest or the distribution over all nodes is considered. The latter case correspond to the distribution of the number of neighbors that share the attitude (see 2.2.1). We will use the following terms:

- $\mathbf{Q}^{(11)}$ the column vector of the number of neighbors (for each node) that **share** the attitude 1. More precisely $Q_i^{(11)}$ is zero if $\sigma_i = -1$ and if $\sigma_i = 1$ $Q_i^{(11)}$ equals the number of nodes from $\Gamma(i)$ that take the value 1 if $\sigma_i = 1$, i.e.

$$Q_i^{(11)} = \frac{\sigma_i + 1}{2} \sum_{j=1}^{N} \delta_{ij} \frac{1 + \sigma_i \sigma_j}{2}. \tag{1.2}$$

- Analogously, we define $\mathbf{Q}^{(-1-1)}$ for the number of neighbors that share the attitude −1,

$$Q_i^{(-1-1)} = \frac{-\sigma_i + 1}{2} \sum_{j=1}^{N} \delta_{ij} \frac{1 + \sigma_i \sigma_j}{2}. \tag{1.3}$$

- The sum $\mathbf{Q} = \mathbf{Q}_{11} + \mathbf{Q}_{-1-1}$, gives the number of pairs in which both neighbors share the same attitude.

- The number of pairs of neighbors with opposite attitude (here, we count together the $-1, 1$ and $1, -1$ cases) is:

$$Q_i^{-11} = k_i - Q_i. \tag{1.4}$$

In the matrix notation we have

$$\mathbf{Q} = \frac{1}{2}(\mathbf{\Sigma}\mathbf{\Delta}\sigma + \mathbf{k}) \tag{1.5}$$

$$\mathbf{Q}^{(11)} = \frac{1}{2}(\mathbf{\Sigma} + \mathbf{I})\mathbf{Q} \tag{1.6}$$

$$\mathbf{Q}^{(-1-1)} = \frac{1}{2}(-\mathbf{\Sigma} + \mathbf{I})\mathbf{Q} \tag{1.7}$$

$$\mathbf{Q}^{(-11)} = \mathbf{k} - \mathbf{Q} \tag{1.8}$$

The total number of pairs of each type is given by:

$$q = \mathbf{Q}^T\mathbf{1} \tag{1.9}$$

$$q^{(11)} = (\mathbf{Q}^{(11)})^T\mathbf{1} = \frac{1}{4}(\sigma + \mathbf{1})^T\mathbf{\Delta}(\sigma + \mathbf{1}) \tag{1.10}$$

$$q^{(-1-1)} = (\mathbf{Q}^{(-1-1)})^T\mathbf{1} = \frac{1}{4}(\mathbf{1} - \sigma)^T\mathbf{\Delta}(\mathbf{1} - \sigma) \tag{1.11}$$

$$q^{(-11)} = S - q. \tag{1.12}$$

1.2 Binary choice

In standard formulations of binary choice models each individual is making a choice σ_i between two alternatives usually coded as -1 and 1 or 0 and 1 (obviously any other codification is possible). The decision is a result of the comparison of two choices: the more profitable/effortless is better. Of course the choice is done from the individual's point of view and does not have to be objectively better. In general the choice is being made in the environment created by the neighbors of individual, the external influence and the former attitude of the individual. Moreover, it may be biased by some noise (random-originated fluctuations of decisions). Formally, σ_i has to fulfill

$$U_i(\sigma_i) \leq U_i(-\sigma_i)$$

where U_i may be denominated as a utility function or a pay-off function. Generally U_i depends on σ_i, σ_j, for $j \in \Gamma(i)$ and h_i - the external randomly biased field. For the purpose of this report (computer simulations in use) we limit ourselves to the utility function of the form,

$$U_i(\sigma_i') = \sigma_i' \left(h + s(b\sigma_i + \sum_{j \in \Gamma(i)} \sigma_j) \right) \tag{1.13}$$

where $s, b > 0$ are the strength parameters (sb determines the self-supporting strength). For n-regular network, we put $s = \frac{1}{n}$. The external random has the random logistic distribution with the mean value h_0, i.e.

$$\Pr(h < z) = \frac{1}{1 + \exp(-2/T(z - h_0))}$$

where the parameter $T > 0$ accounts for the the variance of distribution and is identified with the so called 'social temperature'. For details see [2, 4, 6, 7, 9].

1.3 Clustering

Two main approaches to the clustering analysis appear from the literature studies. The first one bases on the intuitive definition of the cluster and points out the cluster-size distribution in the given network structure. It deals with the clusters physically present (in the sense that they are to be shown one by one) in the system. The clustering is described by means of clustering density and the notion of the average cluster size.

The other class of clustering measures bases on the idea of counting the **pairs** of neighbors that share the same attitude. The clustering reads the degree to which spatial neighborhood and the decisions are correlated. In the random configuration of a given proportion of the decisions +1 and 1 the total number of pairs of nodes that are of the same sign should be significantly smaller than that in the configuration that displays clustering.

Chapter 2

Measures of clustering

2.1 Cluster size density. Average cluster size

In our definition we want to assume "one" to be the minimum size of a cluster that yields each node being a part of certain cluster. Secondly, we want the clusters to be disjunctive such that each node belongs to one and only one cluster.

Let i be the node in the network. **The cluster $C(i)$ generated by i may is defined by the following recursive formula**

1. $i \in C(i)$

2. if $j \in C(i)$ **and for any** $k \in \Gamma(j)$ $\sigma_k = \sigma_i$ **then** $k \in C(i)$

This way, in the first turn, all the neighbors of i that share i's attitude are included into the cluster $C(i)$ and in next turns these neighbors of newly added members of $C(i)$ that share the attitude σ_i are subsequently jointed.

Each node is attributed to a cluster of a certain size N being the number of its elements. Firstly, we focus on the cluster-size distribution. Let $L(s, N)$ be the number of clusters of size s in the network of size N nodes. Notice that $sL(s, N)$ is the number of nodes that belong to the clusters of size s and $\sum_s sL(s) = N$. This quantity depends on the network size N and therefore the normalized one — cluster number density given as

$$l(s, N) = L(s, N)/N, \tag{2.1}$$

is introduced [3]. In this notation $sn(s, l)$ is the probability that a given node belongs to the s-cluster. We may distinguish the $+1$ and -1 cluster-size distribution (the clusters with attitude $+1$) or examine them jointly. It seems interesting to examine the data in respect to the power-law characterizing the cluster-size density.

2.1.1 Average cluster size

The average cluster size [1, 3] is defined by the following formula

$$\chi = \frac{\sum\limits_s s \cdot sL(s, N)}{\sum\limits_s L(s, N)s} = \frac{\sum\limits_s s^2 L(s, N)}{N}$$

In the denominator we have the total number of nodes, in the numerator the sum taken over all nodes of the size of clusters they belong to. Again, we may consider $+1$ and -1 separately or/and jointly. See fig. 2.1

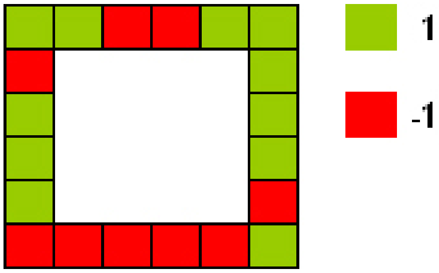


Figure 2.1: Counting from the left bottom corner in a clockwise motion the nodes (20) belong to the clusters of sizes: 2, 2, 2, 2, 5, 5, 5, 5, 5, 1, 1, 5, 5, 5, 5, 5, 3, 3, 3, 1. $\chi = \frac{1}{20}(2{\cdot}2{+}2{\cdot}2{+}5{\cdot}5{+}1{\cdot}1{+}1{\cdot}1{+}5{\cdot}5{+}3{\cdot}3{+}1{\cdot}1) = 3.5$; $\chi_1 = \frac{1}{11}(2 \cdot 2 + 5 \cdot 5 + 1 \cdot 1 + 3 \cdot 3) = 3.55$ $\chi_{-1} = \frac{1}{9}(2 \cdot 2 + 1 \cdot 1 + 5 \cdot 5 + 1 \cdot 1) = 3.44$

2.1.2 Examples

The figures 2.2 and 2.4 shows the cluster size density in the stationary state in a model, respectively, with and without external field and fluctuations. On the figures 2.3 and 2.5 one can see the evolution in time (converging to stationary state) of the average cluster size.

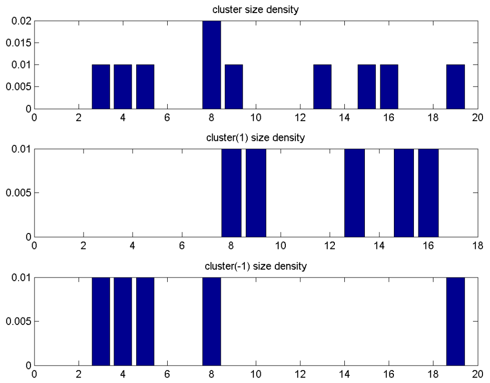


Figure 2.2:

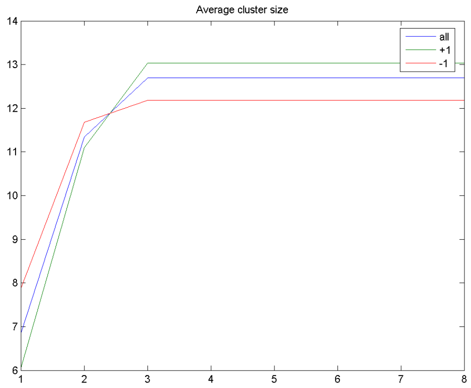


Figure 2.3:

Table 2.1: The binary choice parameters for figures 2.2, 2.3: $h = 10$, $T = 100$, network $=$ 4-regular, $b = 0$.

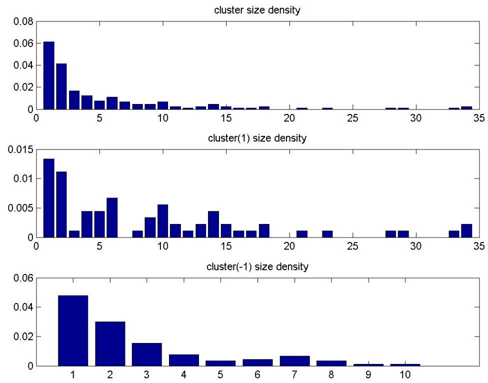


Figure 2.4:

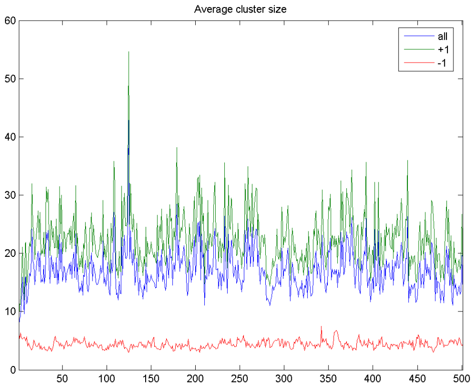


Figure 2.5:

Table 2.2: The binary choice parameters for figures 2.2, 2.3: $h = 10$, $T = 100$, network = 4-regular, $b = 0$.

2.2 Measures based on the number of common pairs

2.2.1 Cluster probability

Cluster probability (CP) is a measure adopted from the analysis of spatial data and, originally, is a factor of the association of attitudes on the two dimensional array with a von Neumann neighborhood (the cells "to the north, east, south and west"). This approach introduces a definition of a cluster, i.e.

> **a node belongs to a cluster if its four nearest neighbors all share its attitude.**

The cluster probability is defined as a **fraction of points that are a part of cluster** or in other words it is a **probability that an arbitrary point is a part of a cluster** [11].

It seems natural to generalize CP by replacing in the cluster definition 'four nearest neighbors' by 'all nearest neighbors' no matter the matrix $\mathbf{\Delta}$, formally

$$CP = \frac{\sum\limits_{i=1}^{N} f_i\left(Q_i\right)}{N} \tag{2.2}$$

where

$$f_i(x) = \begin{cases} 1 & x = k_i, \\ 0, & x \neq k_i. \end{cases} \tag{2.3}$$

or in the matrix notation

$$CP = \frac{1}{N}\left(\mathbf{f}(\mathbf{Q})\right)^T \mathbf{1}, \tag{2.4}$$

where $\mathbf{f}(\mathbf{x}) = \mathbf{f}([x_1, \ldots, x_N]^T) = [f_1(x_1), \ldots, f_N(x_N)]^T$.
Technically, we count these nodes that share their attitude with all their neighbors and divide

this number by the total number of nodes.

The next step would be to assign certain non-zero coefficients to the nodes that share their attitude with some majority of their neighbors, i.e. we set f_i such that $f_i(x) = f(x, k_i) : N \times N \to [0, 1]$, is for a fixed i a nondecreasing function of x.

Notice that for a n-regular lattice, CP may be expressed as

$$CP = \sum_{i=1}^{n} \Pr(M = i)a_i \tag{2.5}$$

where $a_i = f(i, n) = f(i/n)$ is a nondecreasing sequence of positive numbers and M is the (random) number of common neighbors for an arbitrary node. In particular if we set $f(1) = 1$ and $f(x) = 0$ for $x < 1$ we get the 'classical' definition of CP. Therefore CP is a quantity determined by the M distribution.

As a simple example let us consider binary choice model with no external field nor fluctuations, $h0 = 0, T = 0$. One can show that

$$\sigma_i' = \text{sgn}\left(b\sigma_i + \sum_{j\in\Gamma(i)} \sigma_j\right) \tag{2.6}$$

For n-regular network (even n) $\sum_{j\in\Gamma(i)} \sigma_j$ takes the values $-2n, -2n+2, \ldots, 0, \ldots, 2n-2, 2n$. Therefore, the decision rule depends on the sign of $\pm b + 2k$, where $k = -n, \ldots, n$ that does not change for b in intervals $[0, 2)$, $[2, 4)$,... $[2n, \infty)$. See tables 2.3, 2.4, 2.5.

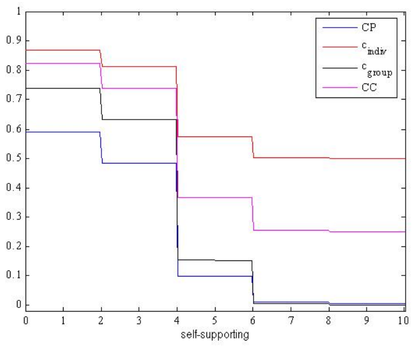


Figure 2.6: (Clustering measures in stationary state as a function of self-supporting b parameter

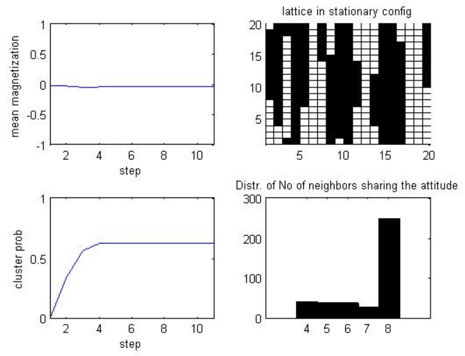


Figure 2.7: $b = 0$: Mean magnetization in time, the visualization in stationary state, cluster probability in time and the histogram of common neighbors distribution

Table 2.3: The results for 8-regular network, $N = 400$ nodes, $h_0 = 0$, $T = 0$.

2.2.2 Individual- and group-level index of clustering.

An instance of the generalized cluster probability is the clustering individual-level index (C_{indiv}) introduced by Nowak and Latane [7]. C_{indiv} is the proportion of neighbors sharing the

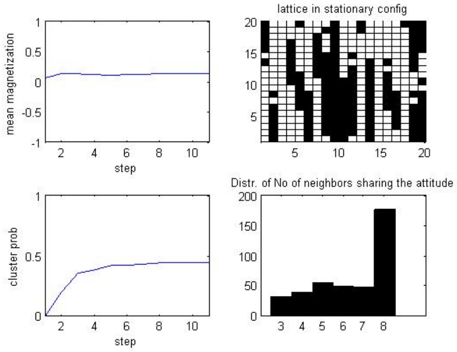


Figure 2.8: $b = 2$: Mean magnetization in time, the visualization in stationary state, cluster probability in time and the histogram of common neighbors distribution

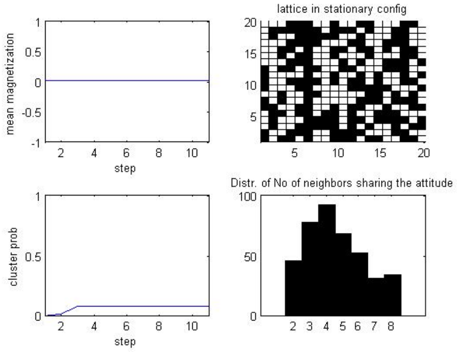


Figure 2.9: $b = 4$: Mean magnetization in time, the visualization in stationary state, cluster probability in time and the histogram of common neighbors distribution

Table 2.4: The results for 8-regular network, $N = 400$ nodes, $h_0 = 0$, $T = 0$.

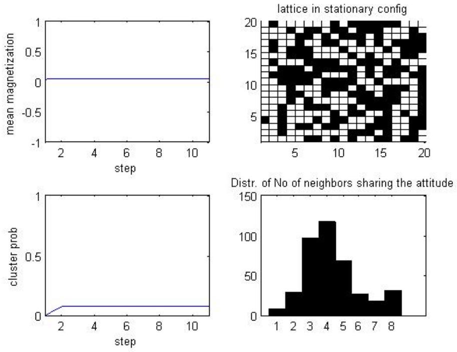


Figure 2.10: $b = 6$: Mean magnetization in time, the visualization in stationary state, cluster probability in time and the histogram of common neighbors distribution

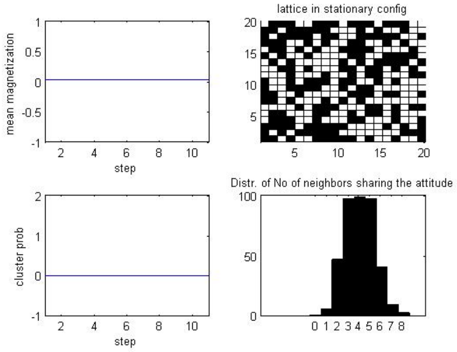


Figure 2.11: $b = 8$: Mean magnetization in time, the visualization in stationary state, cluster probability in time and the histogram of common neighbors distribution

Table 2.5: The results for 8-regular network, $N = 400$ nodes, $h_0 = 0$, $T = 0$.

same attitude summed over all nodes, i.e.

$$C_{indiv} = \frac{q}{S} \tag{2.7}$$

leading to (2.4) for $f_i(x) = \frac{N}{S}x$.

If the network is the n-regular lattice $f_i(x) = \frac{x}{n}$ and (2.7) may be expressed as

$$C_{indiv} = \frac{1}{N}\sum_{i=1}^{N}\frac{Q_i}{n}, \quad Q_i \in \{1, 2, \ldots, n\} \tag{2.8}$$

showing that in this case each node contributes to the cluster probability proportionally to the number of neighbors that share his attitude.

Notice that for an arbitrary network and proportion of the decisions ±1 even in the completely random configuration the index C_{indiv} need not to be close to 0. Moreover C_{indiv} need not to be close 1 if the system is in the maximum possible order. However, this is what one would expect of a 'good' measure of clustering. To normalize the index so that the above two conditions hold, Nowak and Latane define the 'group-level index of clustering'

$$C_{group} = \frac{C_{indiv} - C_{chance}}{C_{max} - C_{chance}}, \tag{2.9}$$

where C_{chance} is the proportion of pairs of common neighbors expected if attitudes are distributed randomly (but the proportion of $+1$ and -1 are maintained) and C_{max} is the maximum possible proportion of common neighbors. C_{chance} and C_{max} in some cases may be calculated analytically (regular networks). In general C_{chance} may be easily computed numerically as an average proportion taken over a number of permutations of the attitudes. In the n-regular network case:

$$C^R_{chance} = (P * (P-1) + Q * (Q-1))/(N * (N-1))$$

where $P = 1/2(\sigma + \mathbf{1})\mathbf{1}$, $Q = 1/2(-\sigma + \mathbf{1})\mathbf{1}$ are the numbers of nodes with 1 and -1, respectively. The C^R_{max} corresponds to the following configuration:

$$\sigma = [\underbrace{1, \ldots, 1}_{P}, \underbrace{-1, \ldots, -1}_{Q}]^T.$$

How to obtain the value of C_{max} seems somehow more complicated and is still an open question.

2.2.3 Spatial autocorrelation

The term 'spatial autocorrelation' refers to the correlation of locational and attribute similarities among spatial objects and has its roots in the area data analysis. It resembles the Pearson's R^2 correlation coefficient. Under the null hypothesis of no spatial autocorrelation, I has an expected value near zero for large n, $E(I) = -\frac{1}{n-1}$. I varies (though not strictly) between the values '-1' and '$+1$.' Positive spatial autocorrelation indicates that like values tend to cluster in space whereas the negative SA suggests that neighbors are dissimilar. The random patterns exhibit zero spatial autocorrelation, see fig. 2.12.

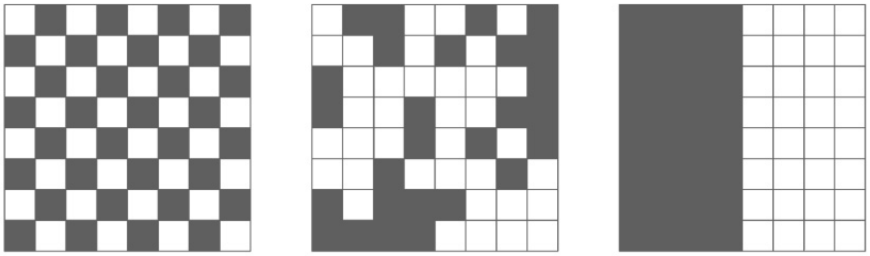


Figure 2.12: From the left: negative, zero and positive spatial autocorrelation pattern.

There is a number of indices of spatial autocorrelation, one of the oldest was proposed by Moran (1950) and is still widely in use. Although it is mainly applied to the continuous type of spatial data using it in a binary case is also possibly and does not pose the difficulties. The formula proposed by Moran (Moran's I) is as follows

$$I = \frac{N \sum\limits_i \sum\limits_j W_{ij} \left(X_i - \overline{X}\right) \left(X_j - \overline{X}\right)}{(\sum\limits_i \sum\limits_j W_{ij}) \sum\limits_i \left(X_i - \overline{X}\right)^2} \tag{2.10}$$

where X_i, $i = 1, \ldots, N$ is the variable value at i location, $\overline{X}$ is the average value taken over all locations and W_{ij} is a weight applied to the comparison between locations i and j. The weight matrix depicts the relation between an element and its surrounding elements. Weights can be based, for example, on contiguity relations or distance. In a weight matrix based on contiguity, a value unequal to zero in the matrix represents pairs of elements with a certain contiguity relation and a zero represents pairs without contiguity relation. in the context of binary choice in network reads

$$I = \frac{N \sum\limits_i \sum\limits_j \delta_{ij}(\sigma_i - m)(\sigma_j - m)}{S \sum\limits_i (\sigma_i - m)^2}. \tag{2.11}$$

Formula (2.11) may be rewritten as a function of a weighted sum of $q^{(11)}$ and $q^{(-1-1)}$.

$$I = \frac{(1 - M/N)q^{(11)} + (M/N)q^{(-1-1)}}{SM/N(1 - M/N)} - 1 \tag{2.12}$$

where M is the number of $+1$ nodes, i.e. $M = 1/2\,(\sigma + \mathbf{1})^T\,\mathbf{1}$. In the terms of the mean magnetization $m = \frac{1}{N}\sum_{i=1}^{N} \sigma_i = 1/N \sigma\,\mathbf{1}$ formula (2.12) reads

$$I = \frac{(1 - m)q^{(11)} + (1 + m)q^{(-1-1)}}{2S(1 - m^2)} - 1. \tag{2.13}$$

In general, we compare the observed value of spatial autocorrelation with the value that we would expect under the randomness of the locations of values $+1$ and -1. The first way to assess the significance of the Moran's I statistics bases on the convergence to Gaussian and is restricted to the data assumed to follow approximately a normal distribution for which the expectation and variance can be calculated. In this case

$$\frac{I - \mathrm{E}(I)}{S_{\mathrm{E}(I)}} \overset{d}{\sim} N(0, 1),$$

where

$$\mathrm{E}(I) = -\frac{1}{N - 1}$$

and

$$S_{\mathrm{E}(I)} = \sqrt{\frac{N^2 \sum\limits_{ij} W_{ij}^2 + 3(\sum\limits_{ij} W_{ij})^2 - N \sum\limits_i (\sum\limits_j W_{ij})^2}{(N^2 - 1)(\sum\limits_{ij} W_{ij})^2}}$$

or for the contiguity matrix

$$S_{\mathrm{E}(I)} = \sqrt{\frac{N^2S + 3S^2 - N\sum\limits_i (k_i)^2}{(N^2-1)S^2}}.$$

The p-value corresponding to the null hypothesis of no spatial autocorrelation is of the form

$$p = 2\left(1 - F\left(\frac{I + (N-1)^{-1}}{S_{\mathrm{E}(I)}}\right)\right),$$

where F is the standard Gaussian cumulative distribution function.

Significant spatial autocorrelation is indicated by p-value p lower than the significance level (usually 0.05). Obviously, the lower the significance level, the more the data must diverge from the null hypothesis to be significant.

The other way bases on the idea of sampling. There are at least two ways of defining spatial randomness: free sampling and nonfree sampling (randomization/permutation). Consequently, there are two types of null hypotheses. Under the null hypothesis in free sampling the nodes are assigned $+1$ or -1 independently with the probabilities M/N and $1 - M/N$ respectively. In non-free sampling (exact test, permutation test) the nodes are assigned $+1$ or -1 with the constraint that the total number of $+1$ is maintained (the data is permuted). In both cases of sampling the value of the test statistic (here, I) is compared to reference distribution being the distribution of the test statistic assuming the null hypothesis is true. The p-value is the proportion of the distribution that is at least as extreme than the observed statistic. If the p-value is smaller than the required significance level then the null hypothesis is rejected and an alternative hypothesis is rendered more plausible.

2.2.4 Join-count analysis

Spatial patterns for binary data (e.g. presence/ absence) from adjacent sampling units (e.g. parcels) or regions (e.g. counties) can be assessed also using joincount statistics [5, 10]. For the binary case, the null hypothesis states that neighboring regions are more likely to be of the same category, say '-1' ('0', white) or '$+1$' (black), and therefore not described by a pattern of randomness. The observed join-count statistics count the number of join encounters in adjacent regions having the same category (already introduced $q^{(11)}$ and $q^{(-1-1)}$); another corresponding join-count statistic counts the number of adjacent regions not having the same category ($q^{(-11)}$). Hence the $q^{(11)}$ and $q^{(-1-1)}$ statistics assess the presence of positive spatial autocorrelation, while $q^{(-11)}$ assesses the presence of negative spatial autocorrelation. As it was already stressed, the Moran's I spatial autocorrelation statistic (2.12) may be expressed as a weighted sum of $q^{(11)}$ and $q^{(-1-1)}$. As a consequence, the Moran's I receives a high value also when $+1$ are clustered but -1 are not (or on the contrary).Therefore, one finds some advantage to use the measures $q^{(11)}$ and $q^{(-1-1)}$ which are separately related to groups of both attitudes.

It is redundant to use all the statistics together as $q^{(11)} + q^{(-1-1)} + q^{(-11)} = const = S$. In general, either only $q^{(-11)}$ is used or a pair – both $q^{(11)}$ and $q^{(-1-1)}$, if the different patterns of clustering for '$+1$' and '-1' are of interest. As for the Moran's I, we compare the observed

q (the one of interest) with its probability distribution under complete randomness (obtained from either free or non-free sampling), and accept or reject the null hypothesis. The significance probability - p-value based on B random permutations is given as the proportion of events such that the join-count statistic of randomly permuted vector σ surpasses the corresponding q, namely

$$p = \frac{\overline{\overline{\{q^{\pi} \geq q\}}}}{B}. \tag{2.14}$$

If free sampling is considered, we proceed analogously – instead of permute σ, the spatial autocorrelation is calculated for vectors randomly drawn within the rules formerly mentioned.

2.3 Clustering coefficient

The term 'clustering' is also used in the context of the transitivity of network structure [8]. Although this meaning of clustering is not the subject of the consideration of this paper, the definition of clustering coefficient (the classical transitivity measure) seems to be applicable after some modifications. In the context of network topology, transitivity (or clustering) means a heightened number of triangles: sets of three nodes where each node is connected with two others. The clustering coefficient is defined as the ratio

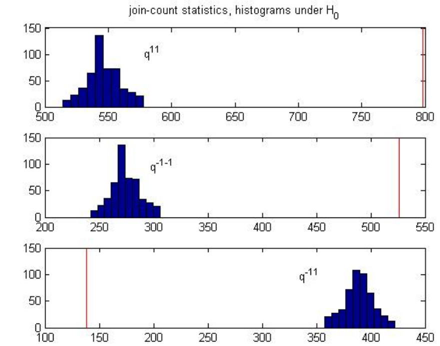
$$CC = \frac{3 \times \text{total number of triangles in the network}}{\text{total number of triples in the network}},$$

where triple is a single node with two links to an unordered pair of other nodes. We propose the following form of clustering coefficient:

$$CC = \frac{\text{number of triangles with common attitude}}{\text{total number of triangles}}. \tag{2.15}$$

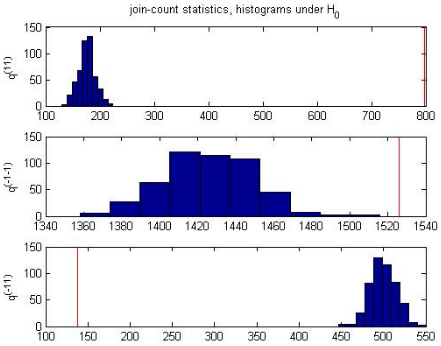
In this approach we do not count the number of pairs of common neighbors but the triangles of common neighbors. Notice, that for the 2-neighbors regular lattice (chain) the triangles do not occur and thus the clustering coefficient does not apply. Moreover it seems that it would not work properly for more complex network that are partly composed of chains. Let us consider the following example: the network that is composed of two parts: $\mathtt{N}_1$ and $\mathtt{N}_2$ such that $\mathtt{N}_1 \cap \mathtt{N}_2 = \{k\}$ (the subnetworks $\mathtt{N}_1$ and $\mathtt{N}_2$ are joint by a single node k). Let $\mathtt{N}_1$ be characterized by the the clustering coefficient CC_1, cluster probability CP_1 and number of nodes N_1. As regards $\mathtt{N}_2$, we assume that it is a chain of the length (number of nodes) M and that all the nodes in $\mathtt{N}_2$, have the same value as σ_k, say $+1$. Intuitively $\mathtt{N}_1 \cup \mathtt{N}_2$ is more clustered than $\mathtt{N}_1$. The cluster probability increases, $CP_{1 \cup 2} = \frac{CP_1 N + m}{N+m} \geq CP_1$ so does the average cluster size and the individual level index. At the same time the cluster coefficient does not change $CC_{1 \cup 2} = CC_1$ as the number of triangles is constant.

We present (see Tab. 2.6,2.7,2.8) computer calculations: the clustering statistics (cluster probability, clustering individual-level index, clustering coefficient, p-values for join-count statistics $q^{(11)}$, $q^{(-1-1)}$ and $q^{(-11)}$, spatial autocorrelation (I) and p-value for I) for three particular networks



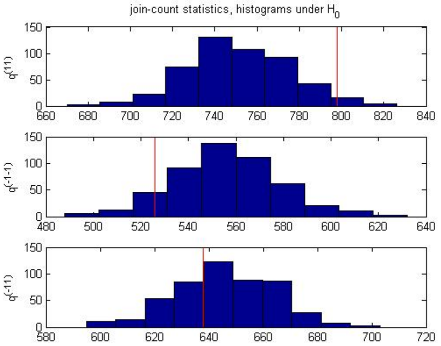
```
       cp: 0.5525
c_indiv: 0.8275
       cc: 0.7700
      p11: 0
    p_1_1: 0
     p_11: 0
        I: 0.6447
   pvalue: 0
```

Table 2.6: $\mathtt{N}_1$ — the 4-neighbors regular network composed of $N_1 = 400$ nodes with the values $\sigma = [\sigma_1, \ldots, \sigma_{400}]^T$ such that the clustering measures are as shown on the right. On the left, distribution of join-count statistics under H_0 and the observed values (red lines).



```
       cp: 0.8011
c_indiv: 0.8938
       cc: 0.7700
      p11: 0
    p_1_1: 0
     p_11: 0
        I: 0.9736
   pvalue: 0
```

Table 2.7: The clustering statistics for $\mathtt{N}_1 \cup \mathtt{N}_2$ where $\mathtt{N}_2$ is a chain of $N_2 = 500$ nodes which are all common with the node that joins $\mathtt{N}_1$ and $\mathtt{N}_2$, i.e. $\sigma_{401} = \ldots = \sigma_{900} = \sigma_{400} = -1$, On the left, distribution of join-count statistics under H_0 and the observed values (red lines).



```
       cp: 0.2456
c_indiv: 0.5092
       cc: 0.7700
      p11: 0.0320
    p_1_1: 0.9160
     p_11: 0.3280
        I: 0.0084
   pvalue: 0.6283
```

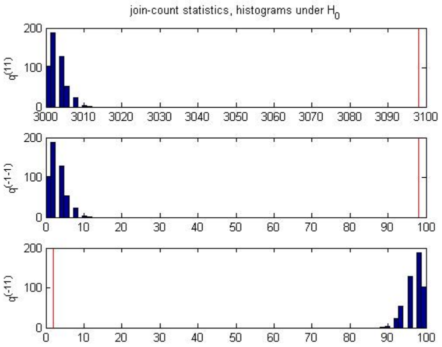
Table 2.8: For $\mathtt{N}_1 \cup \mathtt{N}_2$ such that $[\sigma_{401}, \sigma_{402}, \ldots, \sigma_{899}, \sigma_{900}]^T = [1, -1, \ldots, 1, -1]^T$. On the left, distribution of join-count statistics under H_0 and the observed values (red lines).

Chapter 3

Results and comments

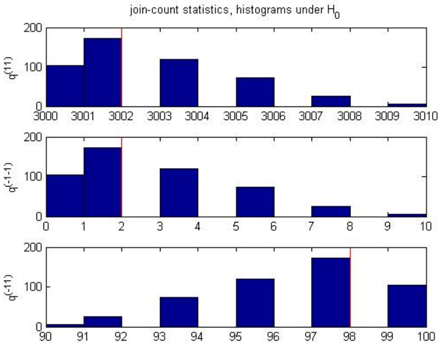
As it has been already mentioned a 'good' measure of clustering should indicate whether the spatial pattern differs from the pattern under random configuration. Notice, that if the mean magnetization m is close to its extreme values ($+1$ or -1) neither cluster probability nor individual level index satisfy this condition. The extreme values of mean magnetization corresponds to high disproportion majority/minority, the minority is very small and cluster probability takes high values no matter the spatial pattern that minority exhibits.

Let us consider the examples of network with mean magnetization close to 1, in particular we consider respectively 2-regular and 4-regular network. In all cases $N = 1600$, the number of minority nodes $N_{-} = 50$ yielding $m = \frac{1600-50}{1600} = 0.9688$. For each network, we calculate different measures of clustering for two configuration: all minority nodes are gathered in a single cluster and the random configuration. In all cases the values of cluster probability, individual-level index and clustering coefficient are high. The randomness is detected by c_group that takes values close to 0 for random configuration. The p-values corresponding to Moran's I autocorrelation and join-count statistics also detect the randomness. On the figures (left) one can see how 'far from randomness' the values of join-count q are.



```
N=100, m=50, 2-regular
          cp: 0.9975
     c_indiv: 0.9988
       c_group: 1
   p-value (q): 0
             I: 0.9794
   p-value (I): 0
```

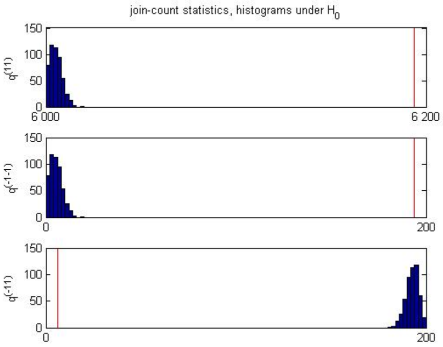
Table 3.1: All minority nodes in one cluster in 2-regular network. On the left, distribution of join-count statistics under H_0 and the observed values (red lines). On the right, clustering measures.



```
          cp: 0.9081
     c_indiv: 0.9387
     c_group: -0.0088
 p-value (q): 0.7920
           I: -0.0116
 p-value (I): 0.5347
```

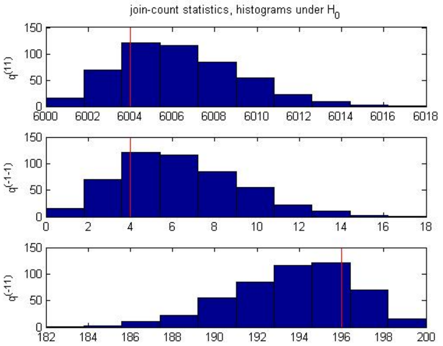
Table 3.2: Random configuration of minority nodes in 2-regular network. On the left, distribution of join-count statistics under H_0 and the observed values (red lines). On the right, clustering measures.

As another example let us consider the time evolution of clustering measures for a binary choice with high values of external field, ($h_0 = 100$) temperature ($T = 100$) and self-supporting ($b = 10$) in 4-regular network. Again, CP and c_{indiv} take high values though the spatial pattern is random. The peaks of c_{group} (fig. 3.2) and the undeterminable autocorrelation (fig. 3.3) correspond to the mean magnetization equal exactly 1 (all nodes have the same sign).



```
N=100, m=50, 4-regular
         cp: 0.9950
    c_indiv: 0.9981
    c_group: 1
         cc: 0.9975
p-value (q): 0
          I: 0.9690
  p-value (I): 0
```

Table 3.3: All minority nodes in one cluster in 4-regular network. On the left, distribution of join-count statistics under H_0 and the observed values (red lines). On the right, clustering measures.



```
         cp: 0.8500
    c_indiv: 0.9387
    c_group: -0.0115
         cc: 0.9081
p-value (q): 0.8280
          I: -0.0116
          Z: -0.8768
p-value (I): 0.3806
```

Table 3.4: Random configuration of minority nodes in 4-regular network. On the left, distribution of join-count statistics under H_0 and the observed values (red lines). On the right, clustering measures.

3.1 Conclusions

We have presented several measures of clustering which we divided in two main groups: cluster size distribution and common neighbors distribution measures. Additionally, we introduced clustering coefficient analogous to the measure of network transitivity. The 'good' measure of clustering should indicate whether and/or how far from randomness the observed spatial pattern is. This condition is satisfied by the measures that use the statistics methods of hypotheses testing (join-count, spatial autocorrelation) and by the group-level index of clustering. However, the latter is not (so far) easily determinated in non-regular network case.

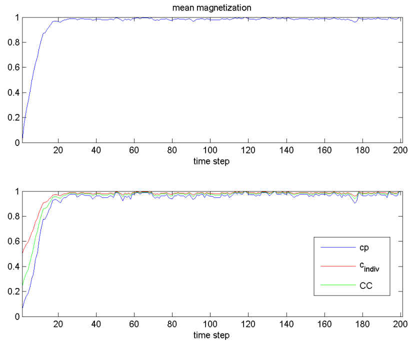


Figure 3.1: Mean magnetization (top) in time. Cluster probability, individual-level index an clustering coefficient in time (bottom).

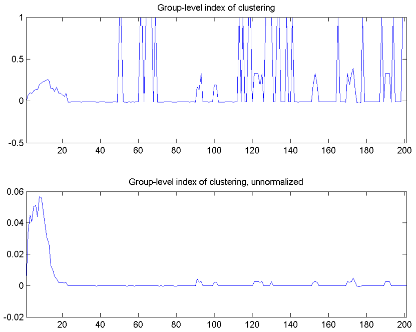


Figure 3.2: Group-level of clustering.

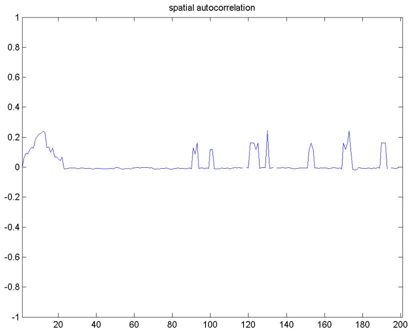


Figure 3.3: Moran's I spatial autocorrelation in time.

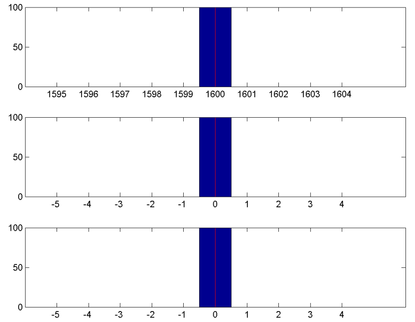


Figure 3.4: P-values for join-count statistics in time.

Table 3.5: The binary choice parameters: $h_0 = 100$, $T = 170$, $b = 10$, 4-regular network.

Bibliography

[1] Ph. Blanchard *et al.*, *Cluster percolation in O(n) spin models* J. Phys. A: Math. Gen. 33 (2000) 8603-8613.

[2] W.A. Brock, and S.N. Durlauf, *Discrete Choice With Social Interactions*, Rev. Econ. Stud., 68 (2001) 235-260.

[3] K.Christensen, N.Moloney, *Complexity and Criticality*, Imperial college Press, London, 2005.

[4] How can statistical mechanics contribute to social science? S.N. Durlauf, *How can statistical mechanics contribute to social science?*, PNAS 96 (1999) 10582–10584.

[5] M.J. Fortin, M.R.T. Dale, J. ver Hoef, *Spatial analysis in ecology, in Encyclopedia of Environmetrics* Wileys, Chichester, (2002) (and the references therein).

[6] J.A. Holyst, K. Kacperski, F. Schweitzerb, *Phase transitions in social impact models of opinion formation*, Physica A 285 (2000) 199-210.

[7] B. Latane, A. Nowak, *Measuring emergent social phenomena: dynamism, polarization, clustering as order parameters of social systems*, Behav. Sci. 39 (1994).

[8] M.E.J. Newman, *The Structure and Function of Complex Network,* SIAM Rev. 45 (2003) 167-256.

[9] Physica A 287 (2000) 613-630. A. Nowak, M.Kus, J. Urbaniak, T. Zarycki, *Simulating the coordination of individual economic decisions*, Physica A 287 (2000) 613-630

[10] J. Nyblom, S. Borgatti, J. Roslakka, M.A. Salo, *Statistical analysis of network data. An application to diffusion of innovation*, Social Networks 25 (2003), 175-195.

[11] J.C. Sprott , J. Bolliger , D.J. Mladenoff, *Self-organized criticality in forest-landscape evolution*, Phys. Lett. A 297 (2002) 267-27