

## **Part II**

# **The Markov Cluster Process**



## Notation and definitions

This chapter introduces the terminology needed for graphs, (dis)similarity spaces, clusterings, and matrices. Single link and complete link clustering are discussed in some greater detail, because these are methods typically applied to dissimilarity data derived from attribute spaces, and are yet often formulated in graph-theoretical terms.

### 4.1 Graphs

DEFINITION 1. Let  $V$  be a finite collection of elements, enumerated  $v_1, \dots, v_t$ .

- i) A **weighted graph**  $G$  on  $V$  is a pair  $(V, w)$ , where  $w$  is a function mapping pairs of elements of  $V$  to the nonnegative reals:  $w : V \times V \rightarrow \mathbb{R}_{\geq 0}$ .
  - a)  $G$  is called **undirected** if  $w$  is symmetric, it is called **directed** otherwise.
  - b)  $G$  is said to be **irreflexive** if there are no loops in  $G$ , that is,  $w(v, v) = 0, \forall v \in V$ .
- ii) A **dissimilarity space**  $D = (V, d)$  is a pair  $(V, d)$ , where  $s$  is a symmetric function mapping  $V \times V$  to  $\mathbb{R}_{\geq 0}$ , satisfying  $s(u, v) = 0 \iff u = v$ . The function  $d$  is called a **dissimilarity measure or dissimilarity coefficient**.
- iii) A **similarity space** is a pair  $(V, s)$ , where  $s$  is a symmetric function mapping  $V \times V$  to  $\mathbb{R}_{> 0} \cup \{\infty\}$ , satisfying  $s(u, v) = \infty \iff u = v$ . The function  $s$  is called a **similarity measure or similarity coefficient**.

The elements in  $V$  are called the **nodes** of  $G$ . The **dimension** of the graph  $G$  is defined as the cardinality  $t$  of its node set  $V$ .

In this thesis, I shall use similarity coefficients in the exposition of  $k$ -path clustering in Chapter 5.

Let  $G = (V, w)$  be a weighted directed graph with  $|V| = t$ . The **associated matrix** of  $G$  lying in  $\mathbb{R}_{\geq 0}^{t \times t}$ , denoted  $\mathcal{M}_G$ , is defined by setting the entry  $(\mathcal{M}_G)_{pq}$  equal to  $w(v_p, v_q)$ . Given a matrix  $M \in \mathbb{R}_{\geq 0}^{N \times N}$ , the **associated graph** of  $M$  is written  $\mathcal{G}_M$ , which is the graph  $(V, w)$  with  $|V| = N$  and  $w(v_p, v_q) = M_{pq}$ .

An equivalent way of representing a weighted graph  $G$  is by identifying  $G$  with a triple  $(V, E, w)$ , where the *edge set*  $E$  is a subset of  $V^2$  and where  $w$  is a positive weight function defined on  $E$  only. A graph represented by such a triple  $(V, E, w)$  is in 1-1 correspondence with a graph representation  $(V, w')$  (according to Definition 1), by setting  $w'(u, v) = a > 0$  iff  $e = (u, v) \in E$  and  $w(e) = a$ , and setting  $w'(u, v) = 0$  iff  $e = (u, v) \notin E$ .

The second representation leads to the generalization of graphs called **hypergraph**. A weighted hypergraph is a triple  $(V, E, w)$  where the hyperedge set  $E$  is a subset of the powerset  $\mathcal{P}(V)$ , and where  $w$  is a weight function on  $E$  as before.

Matrices and graphs of dimension  $N$  are indexed using indices running from 1 to  $N$ . If  $u, v$  are nodes for which  $w(u, v) > 0$ , I say that there is an arc going from  $v$  to  $u$  with weight  $w(u, v)$ . Then  $v$  is called the **tail node**, and  $u$  is called the **head node**. The reason for this ordering lies in the fact that graphs will be transformed later on into stochastic matrices, and that I find it slightly more convenient to work with column stochastic matrices than with row stochastic matrices. The **degree** of a node is the number of arcs originating from it. A graph is called **voidfree** if every node has degree at least one.

A **path** of length  $p$  in  $G$  is a sequence of nodes  $v_{i_1}, \dots, v_{i_{p+1}}$  such that  $w(v_{i_{k+1}}, v_{i_k}) > 0$ ,  $k = 1, \dots, p$ . The path is called a **circuit** if  $i_1 = i_{p+1}$ , it is called a **simple path** if all indices  $i_k$  are distinct, i.e. no circuit is contained in it. A circuit is called a **loop** if it has length 1. If the weight function  $w$  is symmetric then the arcs  $(v_k, v_l)$  and  $(v_l, v_k)$  are not distinguished, and  $G$  is said to have an **edge**  $(v_l, v_k)$  with weight  $w(v_l, v_k)$ . The two nodes  $v_l, v_k$  are then said to be connected and to be **incident to the edge**. A **simple graph** is an undirected graph in which every nonzero weight equals 1. The simple graph on  $t$  nodes in which all node pairs  $u, v, u \neq v$ , are connected via an edge (yielding  $t(t-1)$  edges in all) is denoted by  $K_t$ , and is called the **complete** graph on  $t$  nodes. A weighted directed graph for which  $w(u, v) > 0, \forall u \neq v$ , is called a **weighted complete** graph. A weighted directed graph for which  $w(u, v) = 0$  for some (or many) pairs  $(u, v)$  is called a **weighted structured** graph.

Let  $G = (V, w)$  be a directed weighted graph. A **strongly connected component** of  $G$  is a maximal subgraph  $H$  such that for every ordered pair of nodes  $x, y$  in  $H$  there is a path from  $x$  to  $y$  in  $H$ . If  $G$  is undirected, then the strongly connected components are just called the **connected components**, and  $G$  is called **connected** if there is just one connected component (equalling  $G$  itself). For  $G$  directed, a **weakly connected components** is a maximal subgraph  $H$  containing at least one strongly connected component  $C$  and all nodes  $x$  in  $G$  such that there is a path in  $G$  going from  $x$  to an element of  $C$  (and thus to all elements of  $C$ ). Weakly connected components can thus overlap, but they always contain at least one strongly connected component not contained in any of the other weakly connected components.

Let  $G = (V, w)$  be a directed weighted graph  $G = (V, w)$ . In this thesis the interpretation of the weight function  $w$  is that the value  $w(u, v)$  gives the *capacity* of the arc (path of length 1) going from  $v$  to  $u$ . Let  $G$  be a simple graph, let  $M = \mathcal{M}_G$  be its associated matrix. The capacity interpretation of the weight function  $w$  is very natural in view of the fact that the  $pq$  entry of the  $k^{\text{th}}$  power  $M^k$  gives exactly the number of paths of length  $k$  between  $v_p$  and  $v_q$ . This can be verified by a straightforward computation. The given interpretation of the entries of  $M^k$  extends to the class of weighted directed graphs, by replacing the notion ‘number of paths between two nodes’ with the notion ‘capacity between two nodes’.

The graph which is formed by adding all loops to  $G$  is denoted by  $G + I$ . In general, if  $\Delta$  is a nonnegative diagonal matrix, then  $G + \Delta$  denotes the graph which results from adding to each node  $v_i$  in  $G$  a loop with weight  $\Delta_{ii}$ .

## 4.2 Partitions and clusterings

A **partition** or **clustering** of  $V$  is a collection of pairwise disjoint sets  $\{V_1, \dots, V_d\}$  such that each set  $V_i$  is a nonempty subset of  $V$  and the union  $\cup_{i=1, \dots, d} V_i$  is  $V$ . A partition  $\mathcal{P}$  is called (**top** respectively **bottom**<sup>1</sup>) **extreme** if respectively  $\mathcal{P} = \{V\}$  and  $\mathcal{P} = \{\text{singletons}(V)\} = \{\{v_1\}, \dots, \{v_t\}\}$ . A partition of the form  $\{S, S^c\}$  (where  $S^c$  is the complement of the set  $S$  in  $V$ ) is called a **bipartition** of  $V$ , it is called **balanced** if  $|S| = |S^c|$ . For a bipartition the set notation is omitted, and it is sloppily written as  $(S, S^c)$ . Given a bipartition  $(S, S^c)$ , the corresponding **characteristic difference vector**  $x$  is defined by  $x_i = 1$  for  $v_i \in S$ , and  $x_i = -1$  for  $v_i \in S^c$ .

A **hierarchical clustering** of  $V$  is a finite ordered list of partitions  $\mathcal{P}_i, i = 1, \dots, n$  of  $V$ , such that for all  $1 \leq i < j \leq n$  the partition  $\mathcal{P}_j$  can be formed from  $\mathcal{P}_i$  by conjoining elements of  $\mathcal{P}_i$ , where  $\mathcal{P}_1 = \{\text{singletons}(V)\} = \{\{v_1\}, \dots, \{v_t\}\}$  and  $\mathcal{P}_n = \{V\}$ .

An **overlapping clustering** of  $V$  is a collection of sets  $\{V_1, \dots, V_d\}$ ,  $d \in \mathbb{N}$ , such that each set  $V_i$  is a nonempty subset of  $V$ , the union  $\cup_{i=1, \dots, d} V_i$  is  $V$ , and each subset  $V_i$  is not contained in the union of the other subsets  $V_j, j \neq i$ . The latter implies that each subset  $V_i$  contains at least one element not contained in any of the other subsets, and this in turn implies the inequality  $d \leq t$ .

Let  $s$  be a similarity coefficient defined on  $V = \{v_1, \dots, v_t\}$ . Let  $s_1, \dots, s_n$  be the row of different values that  $s$  assumes on  $V \times V$ , in strictly descending order and with the value 0 added. Remember that  $s(u, u) = \infty, u \in V$ . Thus,  $\infty = s_1 > s_2 > \dots > s_n = 0$ .

The **single link clustering** of the pair  $(V, s)$  is the nested collection of partitions  $\mathcal{P}_i, i = 1, \dots, n$ , where each  $\mathcal{P}_i$  is the partition induced by the transitive closure of the relation in which two elements  $u, v$ , are related iff  $s(u, v) \geq s_i$ . According to this definition, subsequent partitions may be equal,  $\mathcal{P}_1 = \{\text{singletons}(V)\}$ , and  $\mathcal{P}_n = \{V\}$ . The fact that at each similarity level  $s_i$  the single link clustering results from taking the transitive closure implies that the clustering coincides with the connected components of the **threshold graph** of  $(V, s)$  at threshold level  $s_i$ . This is simply the graph<sup>2</sup> on  $t$  nodes where there is an edge between  $u$  and  $v$  iff  $s(u, v) \geq s_i$ .

The **complete link clustering** of the pair  $(V, s)$ , is usually procedurally defined as follows. The bottom partition  $\mathcal{P}_1$  is again taken as  $\{\text{singletons}(V)\}$ . Each clustering  $\mathcal{P}_k, k > 1$ , is subsequently defined in terms of  $\mathcal{P}_{k-1}$  by uniting the two clusters  $C_x$  and  $C_y$  of  $\mathcal{P}_{k-1}$  for which the threshold level  $s$  such that [*the subgraph on  $C_x \cup C_y$  in the threshold graph of  $(V, s)$  at level  $s$  is complete*] is maximal. Equivalently,  $C_x$  and  $C_y$  are such that

<sup>1</sup>The set of all partitions forms a lattice of which these are the top and bottom elements.

<sup>2</sup>Usually threshold graphs are presented in the setting of dissimilarity spaces, using the edge defining inequality  $s(u, v) \leq s_i$ .

the maximum of the minimal similarity in the restriction of the similarity space  $(V, s)$  to  $C_X \cup C_Y$ , is assumed for  $X = x$  and  $Y = y$ . It is not very satisfactory from a mathematical point of view that the clusterings at a given level depend on the previous clusterings. It would be more elegant to define a clustering at a given threshold level as all maximal cliques in the corresponding threshold graph. The drawback is that it will in general result in an overlapping clustering with many clusters. Moreover, different clusters may have large overlap and small symmetric difference. Many variants of this type of complete linkage have been suggested [89, 95, 121], by first forming all maximal cliques at a given threshold level, and subsequently joining clusters (which are cliques) under the transitive closure of some similarity between clusters, e.g. sharing at least  $k$  neighbours. The computational requirements of such methods are huge, and they are mostly presented as an exercise in mathematical thought.

### 4.3 Matrices

A **column stochastic** matrix is a nonnegative matrix in which the entries of each column sum to one. A matrix is called **column allowable** if its associated graph is voidfree, that is, it has no zero columns. Note that a column stochastic matrix is by definition column allowable.

Let  $M$  be a matrix in  $\mathbb{R}^{n \times n}$ , let  $\alpha$  and  $\beta$  both be sequences of distinct indices in the range  $1 \dots n$ . The **submatrix** of  $M$  corresponding with row indices from  $\alpha$  and column indices from  $\beta$  is written  $M[\alpha|\beta]$ . The determinant of a square submatrix is called a **minor**. The **principal submatrix** with both row and column indices from  $\alpha$  is written  $M[\alpha]$ . Let  $\alpha$  be a sequence of distinct indices in the range  $1 \dots n$ , denote by  $\alpha^c$  the sequence of indices in  $1 \dots n$  which are not part of  $\alpha$ . The matrix  $M$  is called **irreducible** if for all  $\alpha$  containing at least 1 and at most  $n - 1$  indices the submatrix  $M[\alpha|\alpha^c]$  has at least one nonzero coordinate. Otherwise  $M$  is called **reducible**. This can also be stated in terms of graphs. Associate a simple graph  $G = (V, w)$  with  $M$  (note that it is not assumed that  $M$  is nonnegative<sup>3</sup>) by setting  $w(v_p, v_q) = 1$  iff  $M_{pq} \neq 0$ . Then the existence of a sequence  $\alpha$  such that  $M[\alpha|\alpha^c]$  has only zero entries implies that there are no arcs in  $G$  going from the subgraph defined on the nodes corresponding with  $\alpha^c$  to the subgraph defined on the nodes corresponding with  $\alpha$ . Thus  $M$  is reducible if the node set of the associated simple graph  $G$  can be split into two (non-empty) parts such that there are no arcs going from the first part to the other, and it is irreducible otherwise.

The eigenvalues of a square matrix  $M$  of dimension  $n$  are written  $\lambda_1(M), \dots, \lambda_n(M)$ . If the eigenvalues are real, then they are written in decreasing order, thus  $\lambda_1(M) \geq \lambda_2(M) \geq \dots \geq \lambda_n(M)$ . This is a strict rule, carried through for the **Laplacian** (introduced in Chapter 8) of a graph as well. In general, the modulus of the largest eigenvalue of  $M$  is called the **spectral radius** of  $M$  and is written  $\rho(M)$ .

Let  $S$  be some subset of the reals. Denote the operator which raises a square matrix  $A$  to the  $t^{\text{th}}$  power,  $t \in S$ , by  $\text{Exp}_t$ . Thus,  $\text{Exp}_t A = A^t$ . This definition is put in such general terms because the class of diagonally *psd* matrices (to be introduced later) allows the introduction of fractional matrix powers in a well-defined way. The entry-wise product

---

<sup>3</sup>The concepts of reducibility and irreducibility are in fact usually defined in the more general setting of complex matrices, but this is not needed here.

between two matrices  $A$  and  $B$  of the same dimensions  $m \times n$  is called the **Hadamard-Schur product** and is denoted by  $A \circ B$ . It has dimensions  $m \times n$  and is defined by  $[A \circ B]_{pq} = A_{pq}B_{pq}$ . The Hadamard power (with exponent  $r$ ) of a matrix  $A$  of dimensions  $m \times n$  has the same dimensions, is written  $A^{\circ r}$ , and is defined by  $[A^{\circ r}]_{pq} = (A_{pq})^r$ .

Let  $A$  be square of dimension  $n$ , and assume some ordering on the set of  $k$ -tuples with distinct indices in  $\{1, \dots, n\}$ . The  $k^{\text{th}}$  **compound** of a square matrix  $A$  is the matrix of all minors of order  $k$  of  $A$ , and is written  $\text{Comp}_k(A)$ . It has dimension  $\binom{n}{k}$ . Its  $pq$  entry is equal to  $\det A[u_p|u_q]$ , where  $u_i$  is the  $i^{\text{th}}$   $k$ -tuple of distinct indices in the given ordering.

**Diagonal matrices** (square matrices for which all off-diagonal entries are zero) are written as  $d_\nu$ , where  $\nu$  is the vector of diagonal entries. A **circulant matrix** is a matrix  $C$  such that  $C_{kl} = C_{k+1, l+1}$  for all  $k$  and  $l$  (counting indices modulo the dimension of the matrix). This implies that the first (or any) column (or row) of a circulant defines the matrix. A circulant is written as  $C_x$ , where  $x$  is its first column vector.

Given a symmetric (hermitian) matrix  $A$ , and a real (complex) vector  $x$  of fitting dimension, the scalar  $x^*Ax$  is called a **symmetric (hermitian) form**, where  $x^*$  denotes the hermitian conjugate of  $x$ .

The **Perron root** of a nonnegative matrix (i.e. a matrix for which all elements are nonnegative) is its spectral radius. It is a fundamental result in Perron-Frobenius theory ([19], page 26) that the Perron root of a nonnegative matrix  $A$  is an eigenvalue of  $A$ . The corresponding eigenvector is called the **Perron vector** and is guaranteed to be nonnegative. If  $A$  is irreducible then the Perron vector has no zero coordinates and the Perron root is a simple eigenvalue of  $A$ . An excellent monograph on this and other subjects in the field of nonnegative matrices is [19].

There are many different subjects in matrix analysis and many textbooks and monographs on different collections of subjects. I found the following books especially useful: *Matrix Analysis* by Horn and Johnson [86], *Topics in Matrix Analysis* by the same authors [87], *Nonnegative Matrices In The Mathematical Sciences* by Berman and Plemmons [19], *Nonnegative Matrices* by Minc [130], *Non-negative matrices and Markov chains* by Seneta [149], *Special matrices and their applications in numerical mathematics* by Fiedler [57], and *Matrix Computations* by Golub and Van Loan [67].

#### 4.4 Miscellanea

Numerical experiments are described in this thesis, which means that the realm of finite precision arithmetic is entered. Numerical expressions denote floating point numbers if and only if a dot is part of the expression. Expressions in which single indices or subscripted or superscripted simple expressions are enclosed in parentheses denote the object which results from letting the index run over its natural boundaries. E.g.  $e_{(i)}$  denotes a vector or a row (the context should leave no doubt which of the two),  $T_{k(i)}$  denotes the  $k^{\text{th}}$  row of the matrix  $T$ , and  $(T^{(i)})_{kl}$  denotes the set of  $kl$  entries of the powers of  $T$ . The fact that each of the entries in a row  $e_{(i)}$  equals the same constant  $c$  is concisely written as  $e_{(i)} \underline{\underline{c}}$ .

