Available online at www.sciencedirect.com

ScienceDirect

Discrete Applied Mathematics 156 (2008) 1330–1341

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

Optimal clustering of multipartite graphs

Irène Charon, Olivier Hudry

GET-École Nationale Supérieure des Télécommunications and CNRS LTCI-UMR 5141, 46, rue Barrault, 75634 Paris cedex 13, France

Received 29 April 2004; received in revised form 10 August 2005; accepted 11 April 2007
Available online 2 June 2007

Abstract

Given a graph $G = (X, U)$, the problem dealt with in this paper consists in partitioning X into a disjoint union of cliques by adding or removing a minimum number $z(G)$ of edges (Zahn's problem). While the computation of $z(G)$ is NP-hard in general, we show that its computation can be done in polynomial time when G is bipartite, by relating it to a maximum matching problem. When G is a complete multipartite graph, we give an explicit formula specifying $z(G)$ with respect to some structural features of G. In both cases, we give also the structure of all the optimal clusterings of G.

MSC: 05C; 05C69; 05C99

Keywords: Graph theory; Clique-partitioning; Zahn's problem; Zahn index; Approximation of symmetric relations by equivalence relations; Clustering; Complexity; Matching

1. Introduction

In this paper, we deal with a problem arising in data analysis and set by Zahn in 1964 ([17]; see also [2,3] for references on Zahn's problem and related topics): given a symmetric relation S defined on a finite set X, find an equivalence relation also defined on X and which best approximates S by minimizing the number of disagreements with respect to S. In the social sciences, this problem may arise for instance in the following context. We have to decide whether some entities (the elements of X; they can represent individuals [11], animals [9,15], objects such as cars [6] or micro-computers [4], member states of the U.N.O. [9,14], companies [12], and so on; see also [16] for details and references) are similar or not. One way to decide which objects are similar and which are not consists in comparing them by pairs. Because of different reasons (diversity of the criteria, existence of thresholds, inconsistency of the deciders, and so on), the result got from these pair-wise comparisons is often a symmetric relation which is not transitive: entities x and y may be considered as similar, as well as y and z, while x and z are not considered as similar. Then, in order to cluster these entities, we look for an equivalence relation minimizing the number of disagreements with respect to results provided by the pair-wise comparison experiment.

This number of disagreements is given by the symmetric difference distance $d(S, E)$ between S and an equivalence relation E; this distance is defined by

$$d(S, E) = |S \Delta E|,$$

E-mail address: hudry@enst.fr (O. Hudry).

0166-218X/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.dam.2007.05.033

where Δ stands for the symmetric difference. Because of the symmetry of S and E, $d(S, E)$ is always even; then we shall deal with another distance $\delta(S, E) = \frac{1}{2}d(S, E)$. Notice that the number of equivalence classes is not fixed *a priori* and thus depends on the relation S.

For a given symmetric relation S, an equivalence relation minimizing δ from S is called a *Zahn equivalence relation* of S. The distance $\delta(S, E)$ between S and a Zahn equivalence relation E of S is called the *Zahn index* of S. The computation of the Zahn index of a symmetric relation is NP-hard in general [7].

It is usual to associate a simple undirected graph G with a symmetric relation S defined on X: the vertex set of G is X and there is an edge between two distinct vertices x and y if and only if x and y are in relation with respect to S. Similarly, an equivalence relation E defined on X can also be considered as a graph H with X as its vertex set and such that H is a disjoint union of cliques. (Notice that the graphs G and H are assumed to be loopless; it means that we do not take the reflexivity of S or of E into account; in fact, this simplification does not change anything to the other features of Zahn equivalence relations; thus this assumption will be done until the end of this paper.) The distance $\delta(S, E)$ between S and E becomes a distance between G and H, equal to the number of edges that we have to add to G or to remove from G in order to transform G into H. Thus, we get a graph theoretic formulation of Zahn problem (already used in [17]). For a given undirected graph G, the minimum number of edges that we have to add or to remove to transform G into a disjoint union of cliques is called the *Zahn index* $z(G)$ of G; a graph H associated with an equivalence relation and at distance $z(G)$ from G is called a *Zahn equivalence graph* of G.

In this paper, we compute the Zahn index and the set of all Zahn equivalence graphs of any bipartite graph and of any multipartite complete graph. In [13], Tomescu considered complete bipartite graphs to compute the maximum value of $z(G)$ over the set of graphs with a given number of vertices. With this respect, our study carries on Tomescu's work by considering graphs which are natural extensions of complete bipartite graphs, namely any bipartite graphs (not necessarily complete) and multipartite complete graphs (notice that dealing with any multipartite graph does not seem an easy task). Moreover, we pay attention to these kinds of graphs also because they form a family of polynomial instances of this problem, which is rather rare: to our knowledge, the only graphs known to constitute non-trivial families of polynomial instances are the ones described by Zahn in [17] and ours.

Section 2 gives the notations and the terminology used in this paper. Preliminary lemmas can be found in Section 3. Section 4 is devoted to the computation of the Zahn index and the Zahn equivalence graphs of any bipartite graph, while the same is done in Section 5 for any complete tripartite graph and more generally for complete multipartite graphs in Section 6.

2. Notations and terminology

Throughout the paper, all the graphs that we deal with are undirected. Moreover, $G = (X, U)$ will denote a graph with X as its vertex set and U as its edge set. The complementary graph of G is noted $\overline{G}$.

A graph $H = (X, U)$ which is a disjoint union of cliques is called an *equivalence graph* defined on X; H is the graph theoretic representation of an equivalence relation E defined on X; the (connected) components of H (that is, the cliques of H) are associated with the equivalence classes of E.

The distance $\delta(G, H)$ between two graphs G and H defined on the same vertex set is the one introduced in Section 1 applied to the edge sets of G and H, i.e. the cardinality of the symmetric difference between the edge sets of G and H. It is also the number of edges that must be added to G or removed from G in order to get H.

A bipartite graph is noted $G = (X = A \cup B, U)$, where $X = A \cup B$ is the set of vertices of G, such that $A \cap B$ is the empty set and every edge of G has one end in A and one in B. Similarly, a tripartite graph is noted $G = (X = A \cup B \cup C, U)$ where $\{A, B, C\}$ is a tripartition of X such that the three subgraphs induced by A, B or C are independent sets of G. More generally, a k-partite graph is noted $G = (X = A_1 \cup \cdots \cup A_k, U)$, where $\{A_1, \ldots, A_k\}$ is a k-partition of X, such that each subgraph induced by A_i $(1 \leqslant i \leqslant k)$ is an independent set of G.

A k-partite graph $G = (X = A_1 \cup \cdots \cup A_k, U)$ is *complete* if, for every i and j with $1 \leqslant i < j \leqslant k$, U contains all the possible edges with one extremity in A_i and one in A_j. Thus a complete k-partite graph is the complementary graph of the equivalence graph with k cliques and conversely.

A complete bipartite graph $G = (X = A \cup B, U)$ is *balanced* if $|A| = |B|$, and *nearly balanced* if $|A| = |B| \pm 1$.

K_n is the complete graph on n vertices; in particular, K_1 is a vertex and K_2 is the graph with two vertices linked by an edge; $K_{a,b}$ (resp. $K_{a_1,a_2,\ldots,a_k}$ for $k \geqslant 2$) is the complete bipartite (resp. k-partite) graph with a vertices in one vertex subset and b in the other one (resp. $a_1, a_2, \ldots, a_k$ vertices in the k vertex subsets).

3. Preliminary lemmas

In order to show the main theorems (Sections 4–6), we first prove four technical lemmas.

Lemma 1. *Let $G = (X = A \cup B, U)$ be a bipartite graph. Let H be an equivalence graph defined on X such that there exists one component P with at least two vertices in A and at least one vertex in B. Let A' (resp. B') be a subset of $P \cap A$ (resp. $P \cap B$) with $\emptyset \neq A' \neq P \cap A$ and $|A'| = |B'|$; let A''* (resp. B'') *denote the set of vertices $(P \cap A) - A'$ (resp. $(P \cap B) - B'$). Let H' be the equivalence graph obtained from H by splitting P into two components*: *one induced by $A' \cup B'$, the other one induced by $A'' \cup B''$ (see Fig. 1 for an illustration*). *Then*:

$$\delta(G, H') \leqslant \delta(G, H)$$

and the equality is obtained in the above relation if and only if, in G, any vertex of A' is adjacent to any vertex of B'' and any vertex of A'' is adjacent to any vertex of B'.

Proof. Fig. 1 illustrates the graphs H and H', where a line between two sets means that all the edges with one end in each set exist. Notice that the existence of A' and B' is provided by the fact that P contains at least two vertices of A and at least one vertex of B. Notice also that A'' is not empty while B'' can be, and B'' can be bigger or smaller than A''.

We set $a' = |A'|$, $a'' = |A''|$, $b' = |B'|$ and $b'' = |B''|$. Then we get

$$\begin{aligned}\delta(G, H') - \delta(G, H) = &(\text{number of edges between } A' \cup B' \text{ and } A'' \cup B'' \text{ in } G)\\ &- (\text{number of edges between } A' \cup B' \text{ and } A'' \cup B'' \text{ in } \overline{G}).\end{aligned}$$

As G is bipartite with $X = A \cup B$, it becomes

$$\delta(G, H') - \delta(G, H) \leqslant (a'b'' + b'a'') - (a'a'' + b'b'').$$

We chose $a' = b'$; so:

$$\delta(G, H') - \delta(G, H) \leqslant (a'b'' + a'a'') - (a'a'' + a'b'') = 0,$$

hence the first part of Lemma 1. Moreover, we have the equality $\delta(G, H') = \delta(G, H)$ if and only if the previous inequalities are all equalities, that is, if and only if the number of edges between $A' \cup B'$ and $A'' \cup B''$ in G is equal to $a'b'' + b'a''$ while the number of edges between $A' \cup B'$ and $A'' \cup B''$ in $\overline{G}$ is equal to $a'a'' + b'b''$. It is exactly the same as saying that, in G, any vertex of A' is adjacent to any vertex of B'' and any vertex of B' is adjacent to any vertex of A''. This completes the proof. □

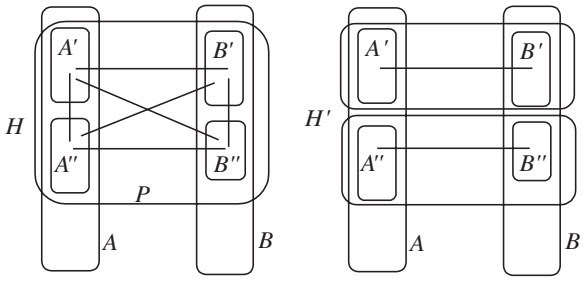

Fig. 1.

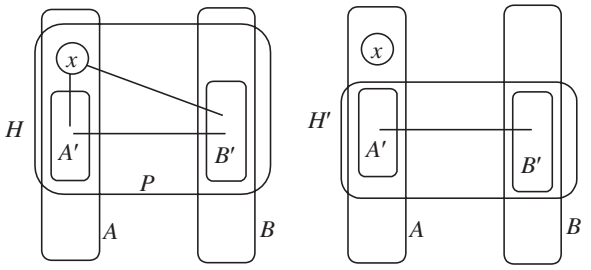

Fig. 2.

Lemma 2. *Let $G = (X = A \cup B, U)$ be a bipartite graph, and let H be any Zahn equivalence graph of G. For each component P of H, the subgraph of G generated by P is a balanced or nearly balanced complete bipartite graph.*

Proof. Let P be any component of H; suppose, without loss of generality, that we have $|A \cap P| \geqslant |B \cap P|$. Let x be a vertex belonging to $P \cap A$. Set $A' = (A \cap P) - \{x\}$, $B' = B \cap P$, $a' = |A'|$ and $b' = |B'|$: then $b' - a' \leqslant 1$. Consider the equivalence graph H' obtained from H by removing x from P and by isolating x into a clique reduced to this vertex, as shown by Fig. 2.

Then, the variation of δ is given by

$$\begin{aligned}\delta(G, H') - \delta(G, H) &= (\text{number of edges between } x \text{ and } A' \cup B' \text{ in } G)\\ &\quad - (\text{number of edges between } x \text{and } A' \cup B' \text{ in } \overline{G})\\ &= (\text{the number of edges between } x \text{ and } B' \text{ in } G)\\ &\quad - a' - (b' - \text{number of edges between } x \text{ and } B' \text{ in } G)\\ &= 2(\text{the number of edges between } x \text{ and } B' \text{ in } G) - a' - b'.\end{aligned}$$

Suppose that the number of edges between x and B' in G is strictly less than b'; then, we would have

$$\delta(G, H') - \delta(G, H) \leqslant 2b' - 2 - a' - b' = b' - a' - 2 \leqslant -1;$$

as H is a Zahn equivalence graph of G, $\delta(G, H)$ is minimum over the set of equivalence graphs defined on X, thus a contradiction with the previous inequality: so x is adjacent to every vertex of B' in G. As this is true for any x belonging to $P \cap A$, all the edges between $P \cap A$ and $P \cap B$ exist in G, and so the subgraph of G generated by P is a complete bipartite graph.

Moreover, the previous result involves also the equality:

$$\delta(G, H') - \delta(G, H) = b' - a'.$$

Once again because of the optimality of H, we must have $b' - a' \geqslant 0$. With the previous inequality $b' - a' \leqslant 1$, this involves that $b' - a'$ is equal to 0 or to 1 or, in other words, $|B \cap P| = |B'| = |A'| = |A \cap P| - 1$ or $|B \cap P| = |B'| = |A'| + 1 = |A \cap P|$: the subgraph of G generated by P is balanced or nearly balanced. This completes the proof. □

To state Lemma 3, we recall the definition of betweenness (see for instance [1,10]):

Definition. Let A, B, and C be three sets; C is said to be between A and B if $A \cap B \subseteq C \subseteq A \cup B$.

Similarly, let $G = (X, U)$, $G' = (X, U')$ and $G'' = (X, U'')$ be three graphs defined on the same vertex set X. Let U_1 (resp. U_2) be the set of edges which are in G'' (resp. G) but not in G (resp. G''): G'' is obtained from G by adding the edges of U_1 and removing the ones of U_2. If G' is obtained from G by adding a part U_1' of U_1 and removing a part U_2' of U_2, G' is said to be *between G and G''*.

Lemma 3. *Let $G = (X, U)$ be a graph, H a Zahn equivalence graph of G and $G' = (X, U')$ a graph between G and H; then, H is also at minimum distance from G' and any equivalence graph which is at minimum distance from G' is at minimum distance from G.*

Proof. Let U_1 (resp. U_2) be the set of edges which are in H (resp. G) but not in G (resp. H). Then G' is obtained from G by adding a part U'_1 of U_1 and removing a part U'_2 of U_2. Thus we have

$$\begin{aligned}\delta(G, H) &= |U_1| + |U_2| \\ &= |U'_1| + |U'_2| + |U_1 - U'_1| + |U_2 - U'_2| \\ &= \delta(G, G') + \delta(G', H).\end{aligned}$$

Let H' be any Zahn equivalence graph of $G' : \delta(G', H') \leqslant \delta(G', H)$. Then, as δ is a distance and by the optimality of a Zahn equivalence graph:

$$\delta(G, H) \leqslant \delta(G, H') \leqslant \delta(G, G') + \delta(G', H') \leqslant \delta(G, G') + \delta(G', H) = \delta(G, H).$$

So, all these inequalities are equalities, and thus:

- $\delta(G', H) = \delta(G', H')$: H is at minimum distance from G',
- $\delta(G, H') = \delta(G, H)$: H' is at minimum distance from G. □

Lemma 4. *Let $G = (X, U)$ and $G' = (X, U')$ be two graphs defined on the same vertex set X. If, for every Zahn equivalence graph H of G, G' is between G and H, then the set of Zahn equivalence graphs of G is the same as the set of Zahn equivalence graphs of G'.*

Proof. It immediately follows from the previous lemma. □

Notice in particular that (with the same notation as for Lemma 4), if G has only one Zahn equivalence graph H, then for any graph G' between G and H, H is the unique Zahn equivalence graph of G'.

4. Zahn index of bipartite graphs

In this section, we compute Zahn index for any bipartite graph G by relating it to the maximum cardinality of a matching of G. Some results are known about some complete bipartite graphs. More precisely, Tomescu [13] shows that $z(G)$ is maximum over the set of graphs G with a given number n of vertices if and only if G is $K_{(n+1)/2,(n-1)/2}$ if n is odd or $K_{n/2,n/2}$ or $K_{n/2+1,n/2-1}$ if n is even, and gives the value of $z(G)$ in these cases. Theorem 1 gives the value of $z(G)$ for any bipartite graph G and describes the structure of the Zahn equivalence graphs of G in this case.

Theorem 1. *Let $G = (X = A \cup B, U)$ be a bipartite graph, m its number of edges and v the maximum cardinality of a matching of G. Then*:

(1) $z(G) = m - v$.

(2) *An equivalence graph H is at distance $z(G)$ from G if and only if it contains a maximum matching of G and, for each component P of H, the subgraph of G generated by P is a balanced or nearly balanced complete bipartite subgraph of G.*

Proof. Let M be a matching of cardinality v. The partial graph $(A \cup B, M)$ of G is an equivalence graph at distance $m - v$ of G: thus $z(G) \leqslant m - v$.

Let us consider now an equivalence graph H defined on X at distance $z(G)$ from G. By Lemma 2, for each component P of H, the subgraph of G generated by P is a balanced or nearly balanced complete bipartite graph. According to Lemma 1, it is possible, without modifying the distance to G, to transform H into an equivalence graph H' which is a disjoint union of cliques K_1 and K_2, the edges of the cliques K_2 belonging to G: this transformation is performed by splitting the components of H iteratively, and more precisely by isolating the edges of M as long as it is possible. Then, the set of edges of H' is a matching of G: its cardinality is at most v; so: $z(G) = \delta(G, H) = \delta(G, H') \geqslant m - v$. Finally, we have: $z(G) = m - v$.

Moreover, $\delta(G, H') = m - v$ implies that the number of edges of H' is v and these edges, which are the edges of M, belong also to H: hence H contains a maximum matching of G. Conversely, consider an equivalence graph H defined on X, containing a maximum matching M of G and such that, for each component P of H, the subgraph of G

generated by P is a balanced or nearly balanced complete bipartite graph. The same transformation as above gives an equivalence graph H' of which the components are reduced to cliques K_1 or K_2, with v edges (the ones of M) and verifying $\delta(G, H') = \delta(G, H)$. As H' contains only the v edges of M, we get $\delta(G, H') = m - v$, which implies that H is at minimum distance from G. □

For complete bipartite graphs, we thus get the following corollary:

Corollary 1. *The Zahn index of $K_{a,b}$, with $a \geqslant b$, is equal to $b(a-1)$.*

Proof. It is direct from Theorem 1 and from the fact that the maximum cardinality of a matching of $K_{a,b}$ with $a \geqslant b$ is equal to b. □

Corollary 2. *If a graph G with n vertices and m edges contains a complete balanced or nearly balanced bipartite graph as a partial graph, then $z(G) = n(n-1)/2 - m$ and K_n is a Zahn equivalence graph of G.*

Proof. By Theorem 1, K_n is a Zahn equivalence graph of $K_{n/2,n/2}$ if n is even and of $K_{(n+1)/2,(n-1)/2}$ if n is odd. As G contains $K_{n/2,n/2}$ if n is even or $K_{(n+1)/2,(n-1)/2}$ if n is odd, G is between K_n and $K_{n/2,n/2}$ if n is even or $K_{(n+1)/2,(n-1)/2}$ if n is odd. Thus, by Lemma 3, K_n is a Zahn equivalence graph of G and $z(G)$ is given by the number of edges that must be added to G to get K_n, that is: $z(G) = n(n-1)/2 - m$. □

Remarks. 1. Though Zahn problem is NP-hard in general, Theorem 1 shows that bipartite graphs constitute a family of polynomial instances, since it is well-known that the computation of a maximum matching in a bipartite graph is a polynomial problem (see for instance [8]).

2. As shown in [5], it is possible to recognize in polynomial time whether a graph contains a complete balanced or nearly balanced bipartite graph as a partial graph.

3. Theorem 1 shows also that the number of Zahn equivalence graphs of a given graph (and thus the number of Zahn equivalence relations of a symmetric relation too) can be very high: for instance, the $n!$ perfect matchings of $K_{n,n}$ are Zahn equivalence graphs of $K_{n,n}$. Moreover, their structures can be also very different: for instance, K_{2n} is a Zahn equivalence graph of $K_{n,n}$ (the $n(n-1)$ missing edges are added) with $n(2n-1)$ edges and only one component while, as said above, a perfect matching with n edges and n components (the $n^2 - n$ edges which do not belong to the perfect matching are deleted) is another Zahn equivalence graph of $K_{n,n}$.

5. Zahn index of complete tripartite graphs

We consider now complete tripartite graphs $K_{a,b,c}$ for any integers a, b and c. Theorem 2 gives the description of all the Zahn equivalence graphs of $K_{a,b,c}$.

Theorem 2. *Let $G = K_{a,b,c} = (X = A \cup B \cup C, U)$ be a complete tripartite graph with $|A| = a$, $|B| = b$, $|C| = c$, $a \geqslant b$ and $a \geqslant c$. Then*:

- *if $a \leqslant b + c$, the graph $H = K_{a+b+c}$ is the unique graph at minimum distance from G;*
- *if $a \geqslant b + c + 1$, an equivalence graph H is at minimum distance from G if and only if one component of H contains $b + c$ or $b + c + 1$ vertices of A, all the vertices of B and all the vertices of C, while each other component of H is reduced to only one vertex of A.*

Proof. Consider a graph H at minimum distance from G and a maximum (with respect to cardinality) component P_1 of H; let p_1 denotes the cardinality of P_1. Let A_1, B_1 and C_1 be the sets of vertices of P_1 which belong, respectively, to A, B and C, and let a_1, b_1 and c_1 denote the respective cardinalities of these sets. We prove the statement of Theorem 2 in three steps; Step 2 is divided into six substeps.

Step 1: Let us show the inequality: $a_1 \leqslant b + c + 1$. Suppose the contrary: $a_1 > b + c + 1$. Consider a new graph H' obtained from H by removing one vertex from A_1 and by creating a new component reduced to this vertex, as illustrated by Fig. 3.

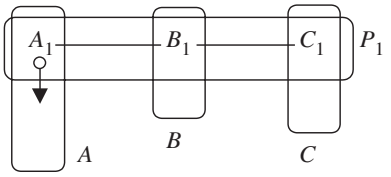

Fig. 3.

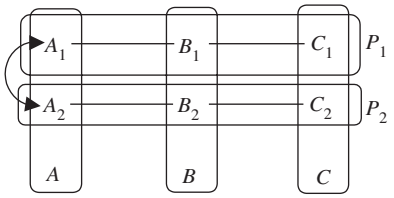

Fig. 4.

Then we get

$$\delta(G, H') - \delta(G, H) = b_1 + c_1 - (a_1 - 1) \leqslant b + c - a_1 + 1 < 0,$$

which is impossible, since H is at minimum distance from G. Hence the inequality:

$$a_1 \leqslant b + c + 1.$$

Step 2: We suppose in this step that H has at least two components. We want to show that this case is impossible if $a \leqslant b + c$ and, otherwise, that P_1 contains at least $b + c$ vertices of A and all the vertices of $B \cup C$, while each other component of H is reduced to only one vertex of A. We do this in six substeps; what we want to show in these substeps is specified at the beginning of each substep below.

For this, let P_2 be another component of H, and set $p_2 = |P_2|$. Let A_2, B_2 and C_2 be the sets of vertices of P_2 which belong, respectively, to A, B and C; let a_2, b_2 and c_2 denote the cardinalities of these sets.

Substep 1: $a_1 \geqslant a_2$, $a_1 \neq 0$, $b_1 \geqslant b_2$, $b_1 \neq 0$, $c_1 \geqslant c_2$, $c_1 \neq 0$.

Consider the equivalence graph H' obtained from H by shifting A_1 and A_2 in P_1 and P_2, as shown by Fig. 4.

Then we get

$$\begin{aligned}\delta(G, H') - \delta(G, H) &= a_1(b_1 + c_1) + a_2(b_2 + c_2) - a_1(b_2 + c_2) - a_2(b_1 + c_1)\\ &= (a_1 - a_2)[(b_1 + c_1) - (b_2 + c_2)].\end{aligned}$$

As H is at minimum distance from G : $\delta(G, H') - \delta(G, H) \geqslant 0$. So the inequality $a_1 < a_2$ would involve $b_1 + c_1 \leqslant b_2 + c_2$, and therefore $p_1 < p_2$, a contradiction with the maximality of p_1. So we have $a_1 \geqslant a_2$. Since this relation is true for any component P_2 of H and because A is not empty, we deduce that a_1 cannot be equal to 0. By similar shiftings involving B_1 and B_2 or C_1 and C_2, we can show the other relations of Substep 1: $b_1 \geqslant b_2$, $b_1 \neq 0$, $c_1 \geqslant c_2$ and $c_1 \neq 0$.

Substep 2: *At least one of the integers a_2, b_2 or c_2 is equal to zero.*

Suppose that a_2, b_2 and c_2 are not equal to zero. Consider a new graph H' obtained from H by moving simultaneously one vertex from A_2 to A_1, one vertex from B_2 to B_1 and one vertex from C_2 to C_1 (see Fig. 5).

Then we get (the first parentheses are associated with the move of the vertex of A_2, the second with the one of B_2, the third with the one of C_2)

$$\begin{aligned}\delta(G, H') - \delta(G, H) &= (a_1 - a_2 + 1 + b_2 - 1 + c_2 - 1 - b_1 - c_1)\\ &\quad + (b_1 - b_2 + 1 + a_2 - 1 + c_2 - 1 - a_1 - c_1)\\ &\quad + (c_1 - c_2 + 1 + a_2 - 1 + b_2 - 1 - a_1 - b_1)\\ &= (a_2 + b_2 + c_2) - (a_1 + b_1 + c_1) - 3 < 0,\end{aligned}$$

a contradiction with the minimality of H.

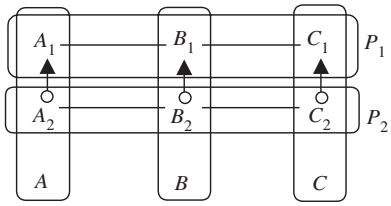

Fig. 5.

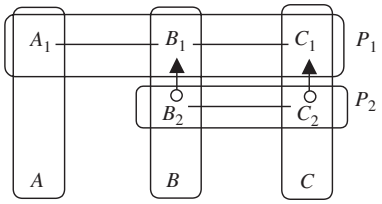

Fig. 6.

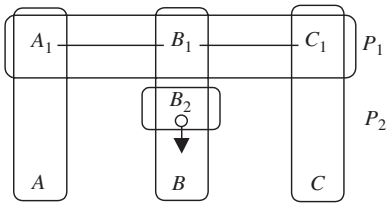

Fig. 7.

Substep 3: *Only one of the integers a_2, b_2 and c_2 is not equal to zero.*

Suppose that $a_2 = 0$, $b_2 \neq 0$ and $c_2 \neq 0$. Consider a new graph H' obtained from H by moving simultaneously one vertex from B_2 to B_1 and one vertex from C_2 to C_1, as shown by Fig. 6.

Then the variation of δ is given by (the first parentheses are associated with the move of the vertex of B_2, the second with the one of C_2)

$$\begin{aligned}\delta(G, H') - \delta(G, H) &= (b_1 - b_2 + 1 + c_2 - 1 - c_1 - a_1) + (c_1 - c_2 + 1 + b_2 - 1 - b_1 - a_1)\\ &= -2a_1 < 0.\end{aligned}$$

As H is at minimum distance from G, the previous result is impossible.

By the same way, one may prove that the relations $b_2 = 0, a_2 \neq 0, c_2 \neq 0$ are impossible simultaneously and, similarly, the relations $c_2 = 0, a_2 \neq 0, b_2 \neq 0$ are impossible simultaneously.

Substep 4: *Except P_1, any clique of H is reduced to one vertex.*

Suppose that we have $b_2 \neq 0, a_2 = c_2 = 0$. If $b_2 \geqslant 2$, consider a new graph H' obtained from H by removing one vertex from B_2 to constitute a new clique reduced to this vertex (see Fig. 7).

Then:

$$\delta(G, H') - \delta(G, H) = -b_2 + 1 < 0,$$

once again a contradiction with the choice of H at minimum distance from G. So: $b_2 = 1$.

The same result holds if we consider A or C instead of B: any clique of H different from P_1 is reduced to one vertex.

Substep 5: *The unique vertex of P_2 belongs to A.*

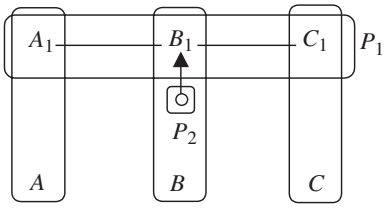

Fig. 8.

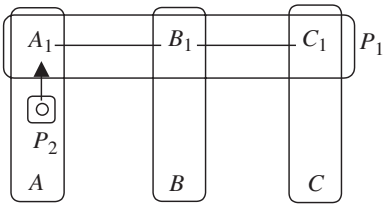

Fig. 9.

From the previous result, we know that P_2 contains only one vertex. Suppose that we have $a_2 = 0$ (the vertex of P_2 belongs to B or to C). Consider a new equivalence graph H' obtained from H by moving the vertex of P_2 from P_2 to P_1 (Fig. 8 illustrates the case for which the vertex of P_2 belongs to B).

Then, if the vertex of P_2 belongs to B (resp. to C), we get

$$\delta(G, H') - \delta(G, H) = b_1 - a_1 - c_1 \text{ (resp.} \delta(G, H') - \delta(G, H) = c_1 - a_1 - b_1).$$

As H is at minimum distance from G, the previous equality involves $b_1 \geqslant a_1 + c_1$ (resp. $c_1 \geqslant a_1 + b_1$). Notice that these two inequalities $b_1 \geqslant a_1 + c_1$ and $c_1 \geqslant a_1 + b_1$ are not compatible simultaneously with $a_1 > 0$ (Substep 1). Thus, all the components of H except P_1 are issued from the same set B or C. Then, if this set is B (resp. C), we get $a_1 = a$, $c_1 = c$, $b > b_1 \geqslant a + c$ (resp. $a_1 = a$, $b_1 = b$, $c > c_1 \geqslant a + b$), what is incompatible with $a \geqslant b$ and $c > 0$ (resp. $a \geqslant c$ and $b > 0$). Hence a contradiction: the vertex of P_2 does not belong to B nor to C, and thus it belongs to A.

Substep 6: $a > a_1 \geqslant b + c$.

From Substep 5, it appears that, in the condition of Step 2 (H has at least two components), any component P_2 of H with $P_2 \neq P_1$ is reduced to one vertex of A. Then, if we move the vertex of such a component P_2 to A_1 (see Fig. 9), we get, as above, an equivalence graph H' verifying $\delta(G, H') - \delta(G, H) = a_1 - b_1 - c_1$. Once again, the optimality of H involves the inequality $a_1 - b_1 - c_1 \geqslant 0$. Hence the relations $a > a_1 \geqslant b + c$, since $A_1 \neq A$.

Step 3. We may now conclude.

According to Step 2, if $a \leqslant b + c$, H has only one clique and therefore $H = K_{a+b+c}$ is the only optimal solution.

Suppose now that $a > b + c$. If H has only one clique, then $a = a_1$ and, by Step 1, $a = b + c + 1$; in this case, the statement of Theorem 2 is satisfied: one component of H (H itself!) contains $b + c + 1$ vertices of A, all the vertices of $B \cup C$, while the other components (there is none!) are reduced to one vertex of A. If H has at least two components, using the results of Step 2 and the relation $a_1 \leqslant b + c + 1$ got in Step 1, we get successively: $a_1 = b + c$ or $a_1 = b + c + 1$; H has $a - a_1 + 1$ components, that is $a - (b + c) + 1$ or $a - (b + c)$ components; in the first (resp. second) case, one component of H contains all the vertices of B, all the vertices of C and $(b + c)$ (resp. $b + c + 1$) vertices of A while the other components of H are reduced to one vertex of A.

Conversely, still for $a > b + c$, consider an equivalence graph H with $a - (b + c) + 1$ (resp. $a - (b + c)$) components, with one component containing all the vertices of B, all the vertices of C and $b + c$ (resp. $b + c + 1$) vertices of A, the other cliques of H being reduced to one vertex of A. From Theorem 1, H is at minimum distance from $K_{a,b+c}$. As the graph G is between $K_{a,b+c}$ and H, Lemma 3 involves that H is at minimum distance from G. □

Corollary 3. *Let $G = K_{a,b,c} = (X = A \cup B \cup C, U)$ be a complete tripartite graph with $|A| = a$, $|B| = b$, $|C| = c$, $a \geqslant b$ and $a \geqslant c$. Then*:

- *if $a \leqslant b + c$, $z(G) = (a^2 + b^2 + c^2 - a - b - c)/2$;*
- *if $a \geqslant b + c + 1$, $z(G) = ab + ac - bc - b - c$.*

Proof. It follows directly from Theorem 2. □

6. Generalization to complete *k*-partite graphs

The following theorem is a generalization of Theorem 2 to complete k-partite graphs for any $k \geqslant 3$. But we may notice that Theorem 3 is not a generalization of the second part of Theorem 1 when applied to complete bipartite graphs.

Theorem 3. *Let k be greater than or equal to 3 and let $G = K_{a_1,a_2,\ldots,a_k} = (X = A_1 \cup \cdots \cup A_k, U)$ be a complete k-partite graph with $|A_i| = a_i$ for $1 \leqslant i \leqslant k$ and $a_1 \geqslant a_i$ for $2 \leqslant i \leqslant k$. Then*:

- *if $a_1 \leqslant \sum_{i=2}^{k} a_i$, the complete graph $H = K_{a_1+a_2+\cdots+a_k}$ defined on X is at minimum distance from G, and it is the unique graph at minimum distance from G;*
- *if $a_1 \geqslant \sum_{i=2}^{k} a_i + 1$, an equivalence graph H is at minimum distance from G if and only if it contains $a_1 - \sum_{i=2}^{k} a_i + 1$ or $a_1 - \sum_{i=2}^{k} a_i$ cliques and, in the first (resp. second) case, there is one clique containing simultaneously all the vertices of $\bigcup_{i=2}^{k} A_i$ as well as $\sum_{i=2}^{k} a_i$ (resp. $\sum_{i=2}^{k} a_i + 1$) vertices of A_1, while each other clique of H is reduced to one vertex of A_1.*

Proof. The proof is by induction on k. For $k = 3$, the statement of Theorem 3 is exactly the same as the one of Theorem 2.

Suppose that k is greater than 3 and that the statement of Theorem 3 is true for $k - 1$. We may assume without loss of generality that a_2 is greater than or equal to a_k. Then consider the complete $(k-1)$-partite graph $G' = (X = A_1 \cup \cdots \cup A_{k-2} \cup A', U')$ with $A' = A_{k-1} \cup A_k$. Notice that G and G' have the same vertex set and that all the edges of G' are edges of G. Set $a' = |A'| = a_{k-1} + a_k$. Then we get $a' \leqslant a_1 + a_2 \leqslant \sum_{i=1}^{k-2} a_i$: A' is not bigger than the union of the other cliques.

First case: $a_1 \leqslant \sum_{i=2}^{k} a_i$.

If $a' \geqslant a_1$, A' is the biggest clique of G'; then the induction hypothesis implies that $K_{a_1+a_2+\cdots+a_{k-2}+a'} = K_{a_1+a_2+\cdots+a_k}$ is the only graph at minimum distance from G'; as G is between G' and $K_{a_1+a_2+\cdots+a_k}$, Lemma 4 gives the expected result for G. Similarly, if $a' < a_1$, A_1 is the biggest clique of G' but is not bigger than the union of the other cliques; then the same argument as before allows us to conclude.

Second case: $a_1 \geqslant \sum_{i=2}^{k} a_i + 1$.

Then we have $a_1 \geqslant a'$. The induction hypothesis implies that G is between G' and any graph at minimum distance from G'. So, here again by Lemma 4, the set of Zahn equivalence graphs of G is the same as the set of Zahn equivalence graphs of G'. Hence the result. □

Next theorem gives the value of Zahn index of complete k-partite graph for $k \geqslant 2$ (thus it is a generalization of Corollaries 1 and 3).

Theorem 4. *For $k \geqslant 2$, let $G = K_{a_1,a_2,\ldots,a_k}$ be a complete k-partite graph with $a_1 \geqslant a_i$ for $1 \leqslant i \leqslant k$. Then*:

- *if $a_1 \leqslant \sum_{i=2}^{k} a_i$, $z(G) = \frac{1}{2}\sum_{i=1}^{k} a_i(a_i - 1)$;*
- *if $a_1 \geqslant \sum_{i=2}^{k} a_i + 1$, $z(G) = \frac{1}{2}\left[\sum_{i=1}^{k} a_i(a_i - 1) - (2a_1 - n)(2a_1 - n - 1)\right]$,*

where $n = a_1 + \cdots + a_k$ is the order of G.

Proof. For $k = 2$, Corollary 1 gives the result. For $k \geqslant 3$, we consider two cases, as for Theorem 3.

First case: $a_1 \leqslant \sum_{i=2}^{k} a_i$.

Theorem 3 shows that, in this case, there is only one equivalence graph at minimum distance from G, which is the complete graph. To change the k-partite graph G into the complete graph defined on the same number of vertices, it is necessary to add all the missing edges in each part of G; hence the result:

$$z(G) = \frac{1}{2}\sum_{i=1}^{k} a_i(a_i - 1).$$

Second case: $a_1 \geqslant \sum_{i=2}^{k} a_i + 1$.

Let A'_1 denote a part of A_1 having $\sum_{i=2}^{k} a_i = n - a_1$ vertices. According to Theorem 3, $z(G)$ is the distance between G and an equivalence graph H with $a_1 - \sum_{i=2}^{k} a_i + 1$ cliques of which one is induced by $A'_1 \cup \bigcup_{i=2}^{k} A_i$, while each other clique of H is reduced to one vertex of A_1. To change G into H, we have to add all the edges with both extremities in A'_1 or in A_i for $2 \leqslant i \leqslant k$, and we have to remove all the edges between $A_1 - A'_1$ and the parts $A_2, \ldots, A_k$. As we have $|A_1 - A'_1| = a_1 - (n - a_1) = 2a_1 - n$, we get

$$\begin{aligned} z(G) &= \frac{1}{2}(n - a_1)(n - a_1 - 1) + \frac{1}{2}\sum_{i=2}^{k} a_i(a_i - 1) + (2a_1 - n)(n - a_1) \\ &= \frac{1}{2}\left[\sum_{i=1}^{k} a_i(a_i - 1) - (2a_1 - n)(2a_1 - n - 1)\right]. \end{aligned}$$

This completes the proof. □

To conclude this paper, we state a last corollary:

Corollary 4. *Let G a graph containing $K_{a_1,a_2,\ldots,a_k}$ as a partial graph, for some $k \geqslant 3$ and with $a_1 \geqslant a_i$ for $2 \leqslant i \leqslant k$ and $a_1 \leqslant \sum_{i=2}^{k} a_i + 1$. Then $K_{a_1+a_2+\cdots+a_k}$ is a Zahn equivalence graph of G and $z(G) = n(n-1)/2 - m$, where $n = \sum_{i=1}^{k} a_i$ denotes the number of vertices of G and m its number of edges. Moreover, if $a_1 \leqslant \sum_{i=2}^{k} a_i$, then $K_{a_1+a_2+\cdots+a_k}$ is the only Zahn equivalence graph of G.*

Proof. Notice that G is between $K_{a_1,a_2,\ldots,a_k}$ and $K_{a_1+a_2+\cdots+a_k}$. For $a_1 \leqslant \sum_{i=2}^{k} a_i$, we know from Theorem 3 that $K_{a_1+a_2+\cdots+a_k}$ is the unique Zahn equivalence graph of $K_{a_1,a_2,\ldots,a_k}$; in this case, Lemma 4 gives the expected result. For $a_1 = \sum_{i=2}^{k} a_i + 1$, by Theorem 3, $K_{a_1+a_2+\cdots+a_k}$ is still a Zahn equivalence graph of $K_{a_1,a_2,\ldots,a_k}$ (though it is not the only one); hence the result, by Lemma 3. □

Notice that it is possible (see [5]) to determine, for any fixed $k \geqslant 3$ and any fixed integers $a_1, a_2, \ldots, a_k$, whether a graph contains $K_{a_1,a_2,\ldots,a_k}$ as a partial graph in polynomial time. Then, for these graphs, Corollary 4 shows that it is easy to compute their Zahn index.

7. Conclusion

As a conclusion, we may summarize the previous results.

Though the computation of Zahn index is NP-hard in general, it becomes polynomial for bipartite graphs, because of the relation $z(G) = m - \nu$, where G is any bipartite graph, m the number of edges of G and ν the maximum cardinality of a matching of G. Anyway, the Zahn equivalence graphs of such a graph G may be very numerous and their structures may be very different (see the remarks of Section 4).

For complete k-partite graphs $G = (X, U)$ with $k \geqslant 3$, Sections 5 and 6 provide explicit formulas between $z(G)$ and the cardinalities of the sets of the k-partition of X. We also give characterizations of the equivalence graphs at minimum distance from G, showing sometimes that the optimal solution is unique.

Some of these results are extended to graphs containing a complete balanced or nearly balanced bipartite graph or a complete k-partite graph ($k \geqslant 3$) as a partial graph, leading in some cases to the uniqueness of the optimal solution of

such a graph. These graphs constitute also a family of graphs for which the computation of Zahn index and of a Zahn equivalence graph is polynomial.

References

[1] M. Barbut, B. Monjardet, Ordre et classification, algèbre et combinatoire, tomes I et II, Hachette, Paris, 1970.
[2] J.-P. Barthélemy, B. Leclerc, The Median Procedure for Partitions, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 19, 1995, pp. 3–34.
[3] J.-P. Barthélemy, B. Monjardet, The median procedure in cluster analysis and social choice theory, Math. Soc. Sci. 1 (1981) 235–267.
[4] S. Chah, Classification of heterogenous data: micro computers, Paper Presented at the Third International Symposium on Data Analysis, Brussels, Belgium, 1985.
[5] I. Charon, O. Hudry, Integer partition and search of partial graphs which are complete multipartite graphs, in preparation.
[6] J.A. Hartigan, Clustering Algorithms, Wiley, New York, 1975.
[7] M. Krivanek, J. Moravek, NP-hard problems in hierarchical-tree clustering, Acta Inform. 23 (1986) 311–323.
[8] L. Lovasz, M.D. Plummer, Matching Theory, Annals of Discrete Mathematics, vol. 29, North Holland, Amsterdam, 1986.
[9] J.-F. Marcotorchino, Agrégation des similarités en classification automatique, thèse de doctorat, université Paris VI, Paris, 1981.
[10] B. Mirkin, Group Choice, Winston, Washington, 1979.
[11] O. Opitz, M. Schader, Analyse qualitativen Daten: Einführung und Übersicht, OR Spektrum 6 (1984) 67–83.
[12] H. Späth, Cluster-Formation und Analyse: Theorie, FORTRAN-Programme und Beispiele, R. Oldenbourg Verlag, München, Germany, 1983.
[13] I. Tomescu, La réduction d'un graphe à une réunion de cliques, Discrete Math. 10 (1974) 173–179.
[14] UNO, Resolutions and Decisions Adopted by the General Assembly during the First Part of its Thirty-ninth Session (from 18 September to 18 December 1984), 1985, pp. 412–419.
[15] G. Vescia, Descriptive classification of cetacea: whales, porpoises and dolphins, in: J.-F. Marcotorchino, J.-M. Proth, J. Janssen (Eds.), Data Analysis in Real Life Environment: Ins and Outs of Solving Problems, Elsevier Science Publishers, North Holland, 1985, pp. 7–24.
[16] Y. Wakabayashi, Aggregation of binary relations: algorithmic and polyhedral investigations, PhD Thesis, Augsbourg, 1986.
[17] C.T. Zahn, Approximating symmetric relations by equivalence relations, SIAM J. Appl. Math. 12 (1964) 840–847.