On the Ensemble of Optimal Identifying Codes in a Twin-Free Graph

Iiro Honkala
University of Turku, Department of Mathematics and Statistics
20014 Turku, Finland
honkala@utu.fi

Olivier Hudry
Institut Télécom - Télécom ParisTech & CNRS - LTCI UMR 5141
46, rue Barrault, 75634 Paris Cedex 13 - France
hudry@telecom-paristech.fr

Antoine Lobstein
CNRS - LTCI UMR 5141 & Institut Télécom - Télécom ParisTech
46, rue Barrault, 75634 Paris Cedex 13 - France
lobstein@telecom-paristech.fr

May 7, 2015

Abstract

Let $G = (V, E)$ be a graph. For $v \in V$ and $r \geq 1$, we denote by $B_{G,r}(v)$ the ball of radius r and centre v. A set $C \subseteq V$ is said to be an *r-identifying code* if the sets $B_{G,r}(v) \cap C$, $v \in V$, are all nonempty and distinct. A graph G which admits an r-identifying code is called *r-twin-free*, and in this case the smallest size of an r-identifying code is denoted by $\gamma_r(G)$.

We study the ensemble of all the different *optimal* r-identifying codes C, i.e., such that $|C| = \gamma_r(G)$. We show that, given any collection $\mathcal{A}$ of k-subsets of $V_1 = \{1, 2, \ldots, n\}$, there is a positive integer m, a graph $G = (V, E)$ with $V = V_1 \cup V_2$, where $V_2 = \{n + 1, \ldots, n + m\}$, and a set $S \subseteq V_2$ such that $C \subseteq V$ is an optimal r-identifying code in G if, and only if, $C = A \cup S$ for some $A \in \mathcal{A}$. This result gives a direct connection with induced subgraphs of Johnson graphs, which are graphs with vertex set a collection of k-subsets of V_1, with edges between any two vertices sharing $k - 1$ elements.

Key Words: Graph Theory, Twin-Free Graphs, Identifiable Graphs, Identifying Codes, Johnson Graphs, Johnson Induced Subgraphs.

1 Introduction

We introduce basic definitions and notation for graphs (for which we refer to, e.g., [1] and [4]) and for identifying codes (see [8] and the bibliography at [9]).

We shall denote by $G = (V, E)$ a simple, undirected graph with vertex set V and edge set E, where an edge between $x \in V$ and $y \in V$ is indifferently denoted by $\{x, y\}$, $\{y, x\}$, xy or yx. The *order* of a graph is its number of vertices $|V|$. We shall denote a cycle of length n by $\mathcal{C}_n$.

In a connected graph G, we can define the *distance* between any two vertices x and y, denoted by $d_G(x, y)$, as the length of any shortest path between x and y. This definition can be extended to disconnected graphs, using the convention that $d_G(x, y) = +\infty$ if there is no path between x and y.

For any vertex $v \in V$ and integer $r \geq 1$, the *ball of radius* r *and centre* v, denoted by $B_{G,r}(v)$, is the set of vertices within distance r from v:

$$B_{G,r}(v) = \{x \in V : d_G(v, x) \leq r\}.$$

Two vertices x and y such that $B_{G,r}(x) = B_{G,r}(y)$ are called (G, r)*-twins*; if G has no (G, r)-twins, that is, if

$$\forall x, y \in V \text{ with } x \neq y, \;\; B_{G,r}(x) \neq B_{G,r}(y),$$

then we say that G is r*-twin-free* or r*-identifiable*. When there is no ambiguity about the graph G, we may use simply $B_r(v)$.

Whenever two vertices x and y are within distance r from each other in G, i.e., $x \in B_r(y)$ and $y \in B_r(x)$, we say that x and y r*-cover* each other. When three vertices x, y, z are such that $x \in B_r(z)$ and $y \notin B_r(z)$, we say that z r*-separates* x and y in G (note that $z = x$ is possible). A set is said to r-separate x and y in G if it contains at least one vertex which does.

A *code* C is simply a subset of V, and its elements are called *codewords*. For each vertex $v \in V$, the r*-identifying set* of v, with respect to C, is the set of codewords r-covering v, and is denoted by $I_{G,C,r}(v)$:

$$I_{G,C,r}(v) = B_{G,r}(v) \cap C.$$

We say that C is an r*-identifying* code [8] if all the sets $I_{G,C,r}(v)$, $v \in V$, are nonempty and distinct: in other words, every vertex is r-covered by at least one codeword, and every pair of vertices is r-separated by at least one codeword. Or: given the (nonempty) identifying set $I_{G,C,r}(v)$ of an unknown vertex $v \in V$, we can uniquely recover v (we also say that we r*-identify* v).

It is quite easy to observe that a graph G admits an r-identifying code if and only if G is r-twin-free; this is why r-twin-free graphs are also called r-identifiable. Also, if a vertex x is isolated or is such that for two vertices u and v, $u \neq v$, the symmetric difference $B_r(u) \Delta B_r(v)$ is reduced to the singleton $\{x\}$, then necessarily x belongs to all the r-identifying codes, and we say that x is *r-forced*.

When G is r-twin-free, we denote by $\gamma_r(G)$ the smallest cardinality of an r-identifying code in G. Any r-identifying code C such that $|C| = \gamma_r(G)$ is said to be *optimal*. The search for an optimal r-identifying code in given graphs or families of graphs is an important part of the studies devoted to identifying codes. In general, this problem is NP-hard [2].

One application of identifying codes is the following: we place ourselves in the case $r = 1$ and assume that we have to protect a museum, or any other type of premises, using smoke detectors. The museum can be viewed as a graph, where the vertices represent the rooms, and the edges, the doors between rooms. The detectors are located in some of the rooms and give the alarm whenever there is smoke in their room or in one of the adjacent rooms. If there is smoke in one room and if the detectors are located in rooms corresponding to a 1-identifying code, then, only by knowing which detectors gave the alarm, we can identify the room where there is a fire starting, or any smoke emission.

In [6], following [12] and [13] where the notion of "completely different codes" is discussed in the framework of infinite lattices, we are interested in finding graphs which have a large number of different optimal r-identifying codes. Considering again our example of the watching of a museum, this means that we want not only to use the smallest possible number of detectors, but also to have a large number of choices for their locations.

In this paper, we study the structure of the ensemble of all the optimal r-identifying codes of a graph. If $G = (V, E)$ is an r-twin-free graph, then this ensemble is trivially a collection of k-element subsets, or k-subsets, of V, for $k = \gamma_r(G)$; we denote this ensemble by $\Phi_r(G)$. Conversely, assume that $\mathcal{A}$ is a nonempty collection of some s different k-subsets $A_1, A_2, \ldots, A_s$ of $V_1 = \{1, 2, \ldots, n\}$. The question is: is there a graph G with vertex set V_1 such that $\mathcal{A}$ is equal to $\Phi_r(G)$? When $3 \leq k \leq n-3$, the answer for *almost* all collections $\mathcal{A}$ is NO; indeed, there are $2^{\binom{n}{k}}$ such collections but only $2^{\binom{n}{2}}$ different graphs (but Theorem 7 below will give an interesting example of a case when the answer is YES). However, we can ask the same question for a graph G with $n + m$ vertices, $m \geq 0$. And now the answer is YES: in Section 2, in Theorems 10 (for $r = 1$) and 12 (for all $r \geq 2$), we prove the following.

Theorem 1 *Let $1 \leq k \leq n$ and $r \geq 1$ be arbitrary. Given any nonempty collection $\mathcal{A}$ of k-subsets of $V_1 = \{1, 2, \ldots, n\}$, there is a positive integer m, a graph $G = (V, E)$ with $V = V_1 \cup V_2$, where $V_2 = \{n+1, \ldots, n+m\}$, and a set $S \subseteq V_2$ such that $C \subseteq V$ is an optimal r-identifying code in G if, and only if, $C = A \cup S$ for some $A \in \mathcal{A}$.*

So the ensemble of the optimal r-identifying codes of the graph G can be described by which k-set of vertices from V_1 we put in the code, since the other codewords (those in S) are common to all optimal codes C; now these k-sets are precisely the k-sets which belong to our target $\mathcal{A}$, and therefore the set $\Phi_r(G)$ is in some sense equivalent to $\mathcal{A}$, in the same way as the labellings of the hexagon H_1 in Figure 1(a)(1) and (a)(2) are equivalent. If, for any two k-subsets A_i and A_j in $\mathcal{A}$ we set

$$\delta(A_i, A_j) = |A_i \Delta A_j|,$$

then, setting $C_i = A_i \cup S$ and $C_j = A_j \cup S$, we can see that $\delta(C_i, C_j) = \delta(A_i, A_j)$, i.e., G is such that $\Phi_r(G)$ has exactly the same symmetric difference distribution as the arbitrary collection $\mathcal{A}$ we started from.

Of course, ideally, we would like to have a result of the type

> "... there is an integer m and a graph G of order $n + m$ such that C is an optimal r-identifying code if, and only if, $C = A$ for some $A \in \mathcal{A}$."

so that $\Phi_r(G)$ and $\mathcal{A}$ would match exactly. However, the following argument from [5] shows that we cannot in general do without the set $S \subseteq V_2$, and that having additional vertices is not enough if we do not use them as codewords.

Assume that $n \geq 2k \geq 4$ and $r \geq 1$, that $\mathcal{A}$ consists of all the k-subsets of $V_1 = \{1, 2, \ldots, n\}$, and that the optimal codes C are exactly all the k-subsets of V_1, for a graph $G = (V = V_1 \cup V_2, E)$ of order $n + m$, $m \geq 0$. For every vertex $v \in V$, the set $B_r(v) \cap V_1$ can be represented by $\beta(v)$, a binary vector of length n where the i-th bit is one if $i \in B_r(v) \cap V_1$ and zero otherwise. For any two vertices u and v in V, the sum $\beta(u) + \beta(v)$ (carried modulo 2), which gives the positions where $\beta(u)$ and $\beta(v)$ differ, must contain at least $n - k + 1$ ones: otherwise, $\beta(u)$ and $\beta(v)$ agree on at least k positions, and any code C consisting of some k of these positions does not r-separate u and v. This means that in the $(n + m) \times n$ binary array D with row-vectors $\beta(v)$, $v \in V$, the sum of any two rows has weight at least $n - k + 1$; in terms of coding theory, we say that D is a binary code of length n, size $n + m$, and minimum distance at least $n - k + 1$. Now there is a result, known as the Plotkin bound (see, e.g., [10, 2§2] or [3, Th. 12.6.4]), stating that such codes do not exist when $n \geq 2k$.

Anyway, we have established a sufficiently strong link between the ensembles of the optimal r-identifying codes of all graphs and the sets of k-subsets of n-sets to connect our investigation to the following definition from [11] and the results related to it. Before that, we recall that we say that two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are *isomorphic*, and write $G_1 \cong G_2$, if there is a bijection $\phi : V_1 \to V_2$ such that $xy \in E_1$ if, and only if, $\phi(x)\phi(y) \in E_2$ for all $x, y \in V$.

Definition 2 *Given positive integers k and n with $1 \leq k \leq n$, the* Johnson graph *$J(k, n)$ is the graph whose vertex set consists of all the k-subsets of $\{1, 2, \ldots, n\}$, with edges between two vertices sharing exactly $k - 1$ elements.*

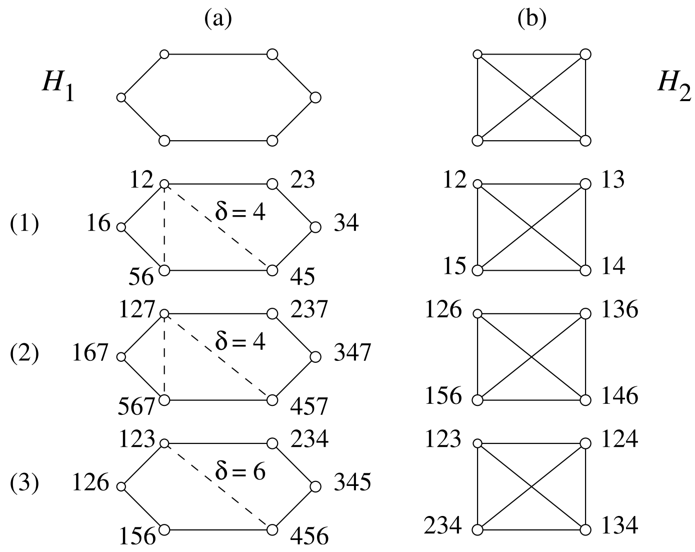

Figure 1: Examples of JIS with different labelling sets. Dotted lines are not edges; they link vertices at distance $\delta = 4$ or $\delta = 6$, according to the cases.

A graph H is isomorphic to an induced subgraph of a Johnson graph if, and only if, it is possible to assign, for some k and n, a k-subset $S_v \subseteq \{1, 2, \ldots, n\}$ to each vertex v of H in such a way that distinct vertices have distinct corresponding k-sets, and vertices v and w are neighbours if, and only if, S_v and S_w share exactly $k-1$ elements. In this case, we say that H is an induced subgraph of a Johnson graph*, or that H is a* JIS *for short.*

We denote by $\mathcal{J}$ the set of all induced subgraphs of all Johnson graphs.

Remark 3 *A JIS H can be associated to several collections of sets, possibly leading to different symmetric difference distributions, as shown in Figure 1. In (a)(1) and (a)(2), the sizes of the symmetric differences between vertices are two and four, whereas in (a)(3) they are two, four and six. In (b), the three labellings are different but have the same symmetric difference distribution, since the graph H_2 is complete and all the symmetric differences have size two.*

If we link two elements C_i and C_j in $\Phi_r(G)$ if, and only if, $\delta(C_i, C_j) = 2$, then we obtain a graph which we denote by $\mathcal{M}_r(G)$. For a given r, the set of all the graphs $\mathcal{M}_r(G)$ is denoted by $\mathcal{M}_r$, and we set $\mathcal{M} = \cup_{r \geq 1} \mathcal{M}_r$. Now, what Theorem 1 says is that every JIS belongs to $\mathcal{M}$, and even more, every JIS belongs to $\mathcal{M}_r$ for all $r \geq 1$:

Corollary 4 *Assume that H is any induced subgraph of the Johnson graph $J(k, n)$. Then, for all $r \geq 1$, there is a graph G such that $\mathcal{M}_r(G) \cong H$, and so $\mathcal{M}_r = \mathcal{J}$.* □

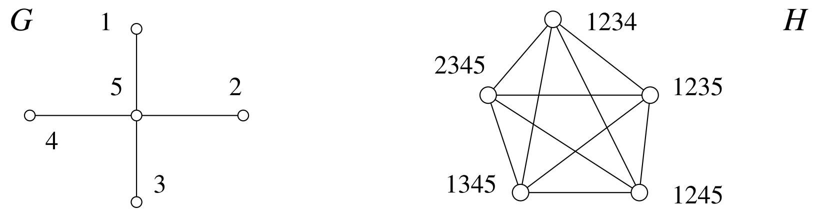

Figure 2: G is the star with five vertices, and $H = \mathcal{M}_1(G)$ is the clique K_5.

Note that we obtain similar results in the short note [7] in the slightly different context of 1-dominating and 1-locating-dominating codes, which we do not define here.

For examples of graphs which are JIS or not, we refer to [11], with a short overview in Section 2.4, but to our knowledge no classification is known.

A result about the connectivity of $\mathcal{M}(G)$ can already be useful: consider again our museum, represented by the graph G, and assume that for maintenance reasons, we wish to change the locations of the detectors and use another optimal 1-identifying code C' instead of C. If $\mathcal{M}_1(G)$ is connected, then this can be performed one detector at a time (i.e., we remove one detector and insert it in a different location), in such a way that, at each intermediate stage, the locations of the detectors always form an optimal 1-identifying code. In fact, using one spare detector, this can be perfomed in such a way that the code remains 1-identifying at all times (i.e., by always adding a detector before removing one). For instance, this is impossible if $G = \mathcal{C}_6$, the cycle with vertex set $\{0, 1, 2, 3, 4, 5\}$, since in this case $\Phi_1(G) = \{\{0, 2, 4\}, \{1, 3, 5\}\}$, and $\mathcal{M}_1(G)$ consists of two isolated vertices. In contrast, if G is a star on n vertices, then $\mathcal{M}_1(G)$ is K_n, the complete graph on n vertices (see Figure 2).

Before we proceed, we need one easy lemma which will be used implicitly in several places (and has already been used in the example of the museum, when we add one detector then remove another), one additional definition as well as some notation.

Lemma 5 *Let $G = (V, E)$ be a graph. If C is r-identifying in G, so is any $D \supseteq C$.* □

Definition 6 *The* Cartesian product *of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ having disjoint vertex sets is the graph with vertex set $V_1 \times V_2$ and edge set $\{\{x = (x_1, x_2), y = (y_1, y_2)\} : (x_1 = y_1$ and $\{x_2, y_2\} \in E_2)$ or $(\{x_1, y_1\} \in E_1$ and $x_2 = y_2)\}$.*

Finally, K_{n_1,n_2} denotes the *complete bipartite graph* with vertex set $V_1 \cup V_2$, where $V_1 = \{1, 2, \ldots, n_1\}$ and $V_2 = \{1, 2, \ldots, n_2\}$, and edge set $\{v_1 v_2 : v_1 \in V_1, v_2 \in V_2\}$.

2 The ensemble of optimal r-identifying codes

2.1 A graph with special structural properties

As discussed in the Introduction, it can be of interest to study the structure of the set $\Phi_r(G)$ of the different optimal r-identifying codes in a graph G.

The following theorem was already mentioned in the Introduction, because it gives the example of a graph G of order n such that $\Phi_1(G)$ is a collection of $\frac{n}{2}$-subsets of an n-set: G needs no additional m vertices and no subcode S.

Moreover, it answers positively and simultaneously two natural questions about $\Phi_r(G)$, at least in the case $r = 1$:

(i) is there a graph $G = (V, E)$ such that a subset C of V is an optimal r-identifying code if, and only if, $V \setminus C$ is an optimal r-identifying code?

(ii) is there a graph $G = (V, E)$ such that for every $i \in \{1, 2, \ldots, \lfloor n/2 \rfloor\}$, there are two optimal r-identifying codes C_1 and C_2 such that $|C_1 \Delta C_2| = 2i$?

Theorem 7 *Let n be an even integer, $n \geq 8$. There exists a graph $G = (V, E)$ with n vertices such that V can be partitioned into $n/2$ sets of size two, $\mathcal{V}_1$, $\mathcal{V}_2$, ..., $\mathcal{V}_{n/2}$, with the following property: $C \subseteq V$ is an optimal 1-identifying code in G if, and only if, $|C \cap \mathcal{V}_i| = 1$ for all $i = 1, 2, \ldots, n/2$.*

Remark 8 *In particular, the number of different optimal 1-identifying codes in G is $2^{n/2}$ (cf. [6] for a study on the number of optimal identifying codes in a graph, and a result better than $2^{n/2}$ for $r = 1$).*

Proof of Theorem 7. The constructions in the particular cases $n = 8$, $n = 10$ and $n = 12$ are given in Figure 3(b), (c) and (d), respectively, and are easy to check, so we go to the general case, when $n \geq 14$ is of the form $4s + 2$ or $4s + 4$, $s \geq 3$. Take s vertices a_i and s vertices b_i, $1 \leq i \leq s$, and take t vertices x_i and t vertices y_i, $1 \leq i \leq t$, where $t = s + 1$ or $t = s + 2$; let

$$V(G) = \{a_1, \ldots, a_s, b_1, \ldots, b_s, x_1, \ldots, x_t, y_1, \ldots, y_t\},$$

so that G is of order $n = 2s + 2t$. Now we link some of the vertices of V in the following way, see Figure 3(a):

(i) for $i \in \{1, 2, \ldots, s\}$, we take the four edges $a_i x_i, a_i y_i, x_i b_i, y_i b_i$;

(ii) for every $i \in \{1, 2, \ldots, s\}$, we take $x_{s+1} a_i, y_{s+1} a_i, x_{s+1} b_i, y_{s+1} b_i$;

(iii) if $t = s + 2$, we choose any two indices i_1, i_2, with $1 \leq i_1 < i_2 \leq s$, and take the edges $a_{i_1} x_{s+2}$, $a_{i_1} y_{s+2}$, $b_{i_1} x_{s+2}$, $b_{i_1} y_{s+2}$ and $a_{i_2} x_{s+2}$, $a_{i_2} y_{s+2}$, $b_{i_2} x_{s+2}$, $b_{i_2} y_{s+2}$.

Now in each pair $\{a_i, b_i\}$, $1 \leq i \leq s$, and in each pair $\{x_j, y_j\}$, $1 \leq j \leq t$ with $t = s + 1$ or $t = s + 2$, the two vertices have the same neighbours, and so at least one vertex in each pair must belong to any 1-identifying code, and at least $n/2$ codewords are needed.

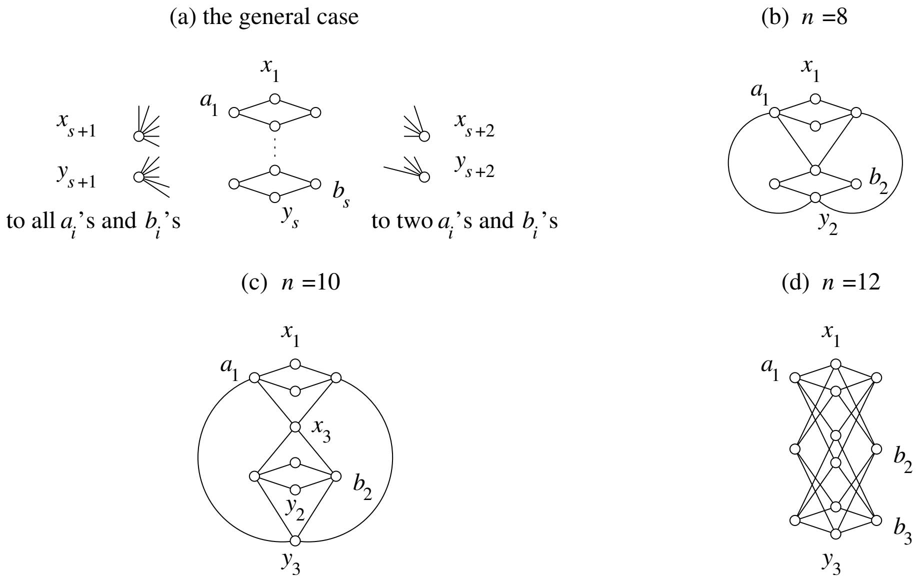

Figure 3: The constructions in the proof of Theorem 7.

Conversely, assume that C is any set of vertices containing exactly one element from each of the aforementioned pairs. We claim that C is 1-identifying.

First, every vertex is 1-covered by C: every a_i and every b_i is taken care of either by x_{s+1} or by y_{s+1}, and every x_i and y_i is taken care of by at least one vertex of type a or b.

Next, a vertex a_i or b_i is 1-covered by at least two codewords of type x, y (either x_i or y_i, and either x_{s+1} or y_{s+1}, plus maybe either x_{s+2} or y_{s+2} when they exist), whereas a vertex x_j or y_j is 1-covered by at most one codeword of type x, y (itself), so C 1-separates the vertices of type a, b from the vertices of type x, y.

If we know that the vertex we are looking for is of type a, b, we can find its index in $\{1, 2, \ldots, s\}$ simply by observing the index of the codeword (x_i or y_i) which is in its 1-identifying set, then, for this index i, we can determine whether the vertex is a_i or b_i, because exactly one of them is in the code.

Finally, if we know that the vertex we are looking for is of type x, y, we first find its index: it is i, $1 \leq i \leq s$, if only one codeword of type a, b, namely either a_i or b_i, 1-covers it, it is $s + 1$ if s codewords of type a, b 1-cover it, it is $s + 2$ if two codewords of type a, b 1-cover it (remember that $2 < s$); then, for this index, we can tell whether the vertex is x or y. $\square$

Considering any of the $2^{n/2}$ optimal 1-identifying codes as a sequence of $n/2$ 01's and 10's, we can, using the rule $01 \rightarrow 0$ and $10 \rightarrow 1$ (or vice versa), transform it into a binary vector of length $n/2$, which proves the following.

Corollary 9 *The graph G of the previous theorem is such that $\mathcal{M}_1(G)$ is isomorphic to the binary hypercube of dimension $n/2$, $n \geq 8$, n even.* $\square$

2.2 Proving the link between $\Phi_1(G)$ and JIS

As announced in the Introduction, we prove the link between $\Phi_r(G)$ and a collection of k-subsets of an n-set, first in the case $r = 1$.

Theorem 10 *Let $1 \leq k \leq n$ be an arbitrary integer, and assume that $\mathcal{A}$ is any nonempty collection of k-subsets of $V_1 = \{1, 2, \ldots, n\}$. Then there is a positive integer m, a graph $G = (V, E)$ with $V = V_1 \cup V_2$, where $V_2 = \{n+1, \ldots, n+m\}$, and a set $S \subseteq V_2$ such that $C \subseteq V$ is an optimal 1-identifying code in G if, and only if, $C = A \cup S$ for some $A \in \mathcal{A}$.*

The main idea behind Theorem 10 (and Theorem 12 as well, for $r > 1$) is first to put in G pairs of vertices, $b(A)$ and $c(A)$, for every $(k-1)$-subset A of V_1 and every k-subset A not in $\mathcal{A}$; then we choose to link them, or not to link them, to the vertices in V_1 (either directly in the case $r = 1$, or by a path for $r > 1$), in such a way that the need to separate $b(A)$ from $c(A)$ will lead to a selection of exactly k codewords in V_1, corresponding to a subset $A \in \mathcal{A}$. The aim of the rest of the graph (the elements of type e and f in Theorem 10, of type e, f, g, e', f', g' in Theorem 12) is to provide a (sufficiently large) collection of "signatures" for the whole graph: these signatures will then participate in the construction of the final identifying sets. We now give the complete, detailed proof of Theorem 10.

Proof. We begin the construction of $G = (V, E)$ by taking n vertices, a_1, a_2, …, a_n (which play the role of the vertices labelled by $1, 2, \ldots, n$ in the statement of the theorem); see Figure 4.

Corresponding to every $(k-1)$-subset A of V_1, we form two new vertices $b(A)$ and $c(A)$, which are linked together; then all the vertices a_i, $i \in A$, are linked to $b(A)$ and $c(A)$, whereas all the vertices a_i, $i \notin A$, are linked to $b(A)$ only (when $k = 1$, we have only one pair, $(b(\emptyset), c(\emptyset))$). Now we can see that when we need to 1-separate the vertices $b(A)$ and $c(A)$, if we suppose that the only vertices that can do it are the vertices of type a, then clearly the $(k-1)$-set $\{a_i : i \in A\}$ fails, but any k-subset would succeed. So, if the vertices of type a have to 1-separate every pair $(b(A), c(A))$, then we have to take at least k of these n vertices as codewords, and any choice of k vertices will do.

Next, corresponding to every k-subset B of V_1 which is not in $\mathcal{A}$, we take two new vertices $b(B)$ and $c(B)$, which are linked together; again, exactly as previously, all the vertices a_i, $i \in B$, are linked to $b(B)$ and $c(B)$, whereas all the vertices a_i, $i \notin B$, are linked to $b(B)$ only. Again, for a given B, the k-set $\{a_i : i \in B\}$ would fail to 1-separate $b(B)$ and $c(B)$, but any other k-set would do. Therefore, if the vertices of type a have to 1-separate every

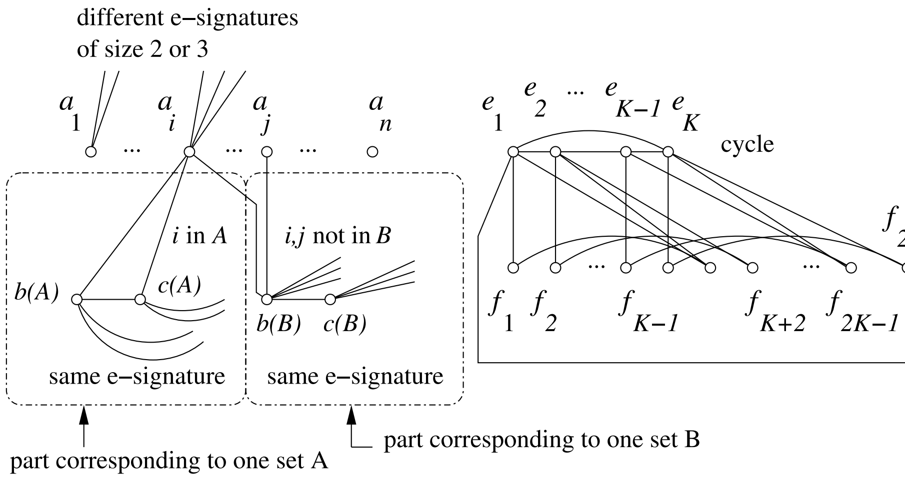

Figure 4: A global representation of the graph G in the proof of Theorem 10. A number of vertices and edges are not represented.

pair $(b(A), c(A))$ and $(b(B), c(B))$, then we have to take at least k of the n vertices of type a as codewords, and we can do it with exactly k of them if, and only if, we have chosen the elements of one of the k-sets in $\mathcal{A}$.

Now we choose the smallest integer K such that

$$\binom{K}{2} + \binom{K}{3} - 2K \geq n + \binom{n}{k-1} + \binom{n}{k} - |\mathcal{A}|, \qquad (1)$$

and add K new vertices, e_1, …, e_K, together with the edges e_1e_2, e_2e_3, …, $e_{K-1}e_K$, e_Ke_1, so that the e_i's form the cycle $\mathcal{C}_K$. We let $s = 2K$ and complete the vertex set V of G by taking s new vertices f_1, …, f_s, with the edges f_if_{i+K}, $1 \leq i \leq K$. In the sequel, for any vertex $v \in V$, we say that the e-**signature** of v is the set of those vertices e_i which 1-cover v.

We now choose the edges between the e_i's and the f_j's: they are e_if_i and e_if_{i+K}, $1 \leq i \leq K$, and $e_{i+1}f_{i+K}$, $1 \leq i \leq K-1$, and e_1f_{2K}.

Clearly, in the subgraph induced by the e_i's and f_j's (in the graph constructed so far), the optimal 1-identifying code has size K and can only be the set of all the vertices of type e: indeed, the symmetric difference of $B_1(f_i)$ and $B_1(f_{i+K})$, for $1 \leq i \leq K-1$, is the singleton $\{e_{i+1}\}$ and the symmetric difference of $B_1(f_K)$ and $B_1(f_{2K})$ is $\{e_1\}$, which means that all the vertices of type e are 1-forced; on the other hand, they suffice to constitute a 1-identifying code, since they 1-cover all the e_i's and f_j's, the e_i's have K different e-signatures of size three, the vertices f_i, $1 \leq i \leq K$, have K different e-signatures of size one, and the vertices f_i, $K+1 \leq i \leq 2K$, have K different e-signatures of size two.

Finally, each vertex of type a is assigned a different e-signature of size two or three which has not been used by the vertices of type e and f, by linking each of them to two or three e_i's in a suitable way. Similarly, for every subset A of size $k-1$, we assign both the vertices $b(A)$ and $c(A)$ the same e-signature of size two or three, in such a way that these e-signatures are different for different sets A and different from the e-signatures given to the a_i's, the e_i's and the f_i's; for every subset $B \notin \mathcal{A}$ of size k, we act similarly for $b(B)$ and $c(B)$: they receive a common signature of size two or three, different for each B and different from all the signatures previously given. All this can be done because we need K e-signatures of size three for the e_i's, K e-signatures of size two for the vertices f_i, $K+1 \leq i \leq 2K$, n e-signatures of size two or three for the a_i's, $\binom{n}{k-1}$ e-signatures of size two or three for the pairs $(b(A),\, c(A))$ corresponding to all the $(k-1)$-sets of V_1, and $\binom{n}{k} - |\mathcal{A}|$ e-signatures of size two or three for the pairs $(b(B),\, c(B))$ corresponding to the k-sets not belonging to $\mathcal{A}$, and because K satisfies inequality (1).

The construction of G is now complete; see Figure 4. What are the optimal 1-identifying codes in G? In order to 1-separate the f_j's, we need all the e_i's as codewords, and in order to 1-separate all the pairs $(b(A),\, c(A))$ and $(b(B),\, c(B))$, we need at least k codewords among the a_i's. So $\gamma_1(G) \geq K+k$. Moreover, in any 1-identifying code of size $K+k$, the indices i of the k a_i's in the code form a subset which belongs to $\mathcal{A}$.

Conversely, assume that C is any code consisting of all the e_i's and the k elements a_i, $i \in A_0$ for some $A_0 \in \mathcal{A}$. Then C is 1-identifying; indeed,

(i) all the vertices are 1-covered by C,

(ii) all the vertices f_i, $1 \leq i \leq K$, have unique e-signatures of size one,

(iii) all the a_i's, e_i's and all the vertices f_i, $K+1 \leq i \leq 2K$, have different e-signatures of size two or three, as do all the pairs $(b(A),\, c(A))$ and $(b(B),\, c(B))$;

(iv) finally, we have already observed that the very construction of the graph makes every pair of vertices $(b(A),\, c(A))$ and $(b(B),\, c(B))$ 1-separated by C. □

2.3 Proving the link between $\Phi_r(G)$ and JIS when $r \geq 1$

We now turn to the case $r \geq 1$, for the proof of which the following definition will be useful.

Definition 11 *A vertex x in an r-twin-free graph G is said to be* r-optimal *if it belongs to the intersection of all optimal r-identifying codes in G.*

Obviously, if x is r-forced, then it is r-optimal, but the converse may be false. The set of r-optimal vertices in some sense forms the identifying core of the graph.

Theorem 12 *Let $1 \leq k \leq n$ and $r \geq 1$ be arbitrary, and assume that $\mathcal{A}$ is any nonempty collection of k-subsets of $V_1 = \{1, 2, \ldots, n\}$. Then there*

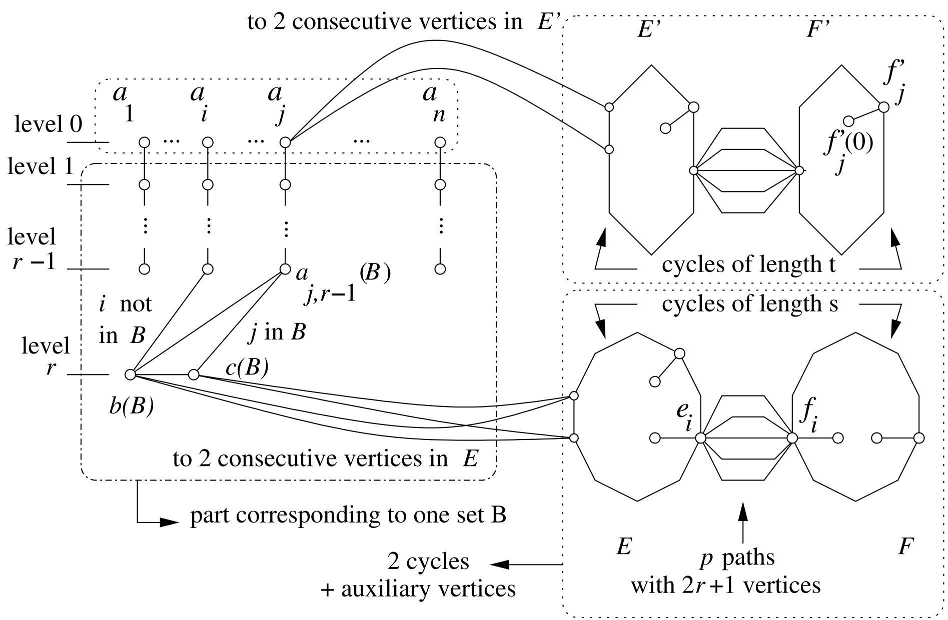

Figure 5: A global representation of the graph G in the proof of Theorem 12. A number of vertices and edges are not represented.

is a positive integer m*, a graph* $G = (V, E)$ *with* $V = V_1 \cup V_2$*, where* $V_2 = \{n+1, \ldots, n+m\}$*, and a set* $S \subseteq V_2$ *such that* $C \subseteq V$ *is an optimal* r*-identifying code in* G *if, and only if,* $C = A \cup S$ *for some* $A \in \mathcal{A}$*.*

Proof. We have already proved the case $r = 1$, so we can assume that $r \geq 2$ (this assumption is needed in Step 5).

We begin by choosing n vertices $a_1, a_2, \ldots, a_n$ (which play the role of the vertices labelled by $1, 2, \ldots, n$ in the statement of the theorem); see Figure 5. Denote by $\mathcal{B}$ the set consisting of the $(k-1)$-subsets of V_1 and the k-subsets of V_1 which do not belong to $\mathcal{A}$.

Next, for each $B \in \mathcal{B}$, we do the following. First, we take the elements of the set

$$V(B) = \{a_{i,j}(B) : 1 \leq i \leq n, 1 \leq j \leq r-1\} \cup \{b(B), c(B)\}$$

as new vertices, and for all $i = 1, 2, \ldots, n$, we take the edges

$$a_i a_{i,1}(B), a_{i,1}(B) a_{i,2}(B), \ldots, a_{i,r-2}(B) a_{i,r-1}(B).$$

Moreover, if $i \in B$, we connect $a_{i,r-1}(B)$ by an edge to both $b(B)$ and $c(B)$; otherwise, we only link $a_{i,r-1}(B)$ to $b(B)$. Finally, we link $b(B)$ and $c(B)$.

We now construct two cycles with vertex sets $E = \{e_1, e_2, \ldots, e_s\}$ and $F = \{f_1, f_2, \ldots, f_s\}$, where $s \geq 2r+2$, $s \geq |\mathcal{B}|$, and we view the indices modulo s. For each $i = 1, 2, \ldots, s$, we add a vertex $e_i(0)$ and link it to e_i, and a vertex $f_i(0)$ linked to f_i. Next, for all $i = 1, 2, \ldots, s$, we connect e_i and f_i with p paths, each path containing $2r + 1$ new vertices:

$$g_i(j,1)g_i(j,2)\ldots g_i(j,2r+1) \tag{2}$$

for $j = 1, 2, \ldots, p$: so, for all i and j,

$$e_i g_i(j,1), g_i(j,1)g_i(j,2), \ldots, g_i(j,2r)g_i(j,2r+1), g_i(j,2r+1)f_i$$

are all edges. The size of p will be specified shortly.

Next we construct a new, essentially similar structure as in the previous paragraph; namely we construct two more cycles with vertex sets $E' = \{e'_1, e'_2, \ldots, e'_t\}$ and $F' = \{f'_1, f'_2, \ldots, f'_t\}$ where $t \geq 2r + 2$, $t \geq n$, and the indices are viewed modulo t. Then, for each $i = 1, 2, \ldots, t$, we add a vertex $e'_i(0)$ and link it to e'_i, add a vertex $f'_i(0)$ linked to f'_i, and connect e'_i and f'_i with p paths (with the same p as above)

$$g'_i(j,1)g'_i(j,2)\ldots g'_i(j,2r+1) \tag{3}$$

for $j = 1, 2, \ldots, p$, so that, for all i and j,

$$e'_i g'_i(j,1), g'_i(j,1)g'_i(j,2), \ldots, g'_i(j,2r)g'_i(j,2r+1), g'_i(j,2r+1)f'_i$$

are all edges.

We choose p so that $p > 2s + 2t$. So far, we have three connected components in the current graph G: the first one will carry out the selection of the subsets $A \in \mathcal{A}$ in the same way as in the case $r = 1$; the other two will be connected to it in such a way that they "take care" of all the rest —and do not interfere with the selection process of the first part.

Finally, for every $B \in \mathcal{B}$, we connect both $b(B)$ and $c(B)$ by an edge to two consecutive vertices e_i and e_{i+1} of E in such a way that different indices i are assigned to different subsets B; this is clearly possible since we chose $s = |E| \geq |\mathcal{B}|$; similarly, for every i we link a_i to e'_i and e'_{i+1}, using the fact that $t \geq n$. This completes the construction of our graph G; see Figure 5.

Step 1. We first show that if C is any r-identifying code in G, then C must contain at least k of the vertices $a_1, a_2, \ldots, a_n$, and if it contains exactly k of them, they must form one of the sets $\{a_i : i \in A\}$ for some $A \in \mathcal{A}$.

To do this, consider one pair $(b(B), c(B))$ from $V(B)$, $B \in \mathcal{B}$. The vertices $b(B)$ and $c(B)$ can only be r-separated by an element of C which has distance r to one of them and distance $r + 1$ to the other.

Clearly, every vertex in $V(B)$ is within distance $r - 1$ from $b(B)$ and therefore within distance r from $c(B)$. So, we must exit the subgraph induced by $V(B)$. However, there are only two ways. The first possibility is to go through e_i or e_{i+1} (for the appropriate index i): but, if we can reach a vertex v in at most r steps, walking along the edges of G, by starting from $b(B)$, say, and going via e_i or e_{i+1}, then the same is true also if we start from $c(B)$, and therefore v is of no use. The second possibility is to exit via some a_i. In the sequel, we say that the vertices a_i are **on level** 0, the vertices $a_{i,j}(B)$ are **on level** j, and the vertices $b(B)$ and $c(B)$ are **on level** r; *c.f.* Figure 5. Then each edge brings us at most one level up or down, and by the time we reach any a_i having started from $b(B)$ or $c(B)$, we have already taken at least r steps. Consequently, the vertices a_i are the only ones that can r-separate $b(B)$ and $c(B)$; moreover, the vertex a_i can do this if, and only if, the edge $a_{i,r-1}(B)c(B)$ is missing, i.e., $i \notin B$.

No $(k-1)$-set D of vertices of type a can r-separate all the pairs $(b(B), c(B))$, because there is a set $B \in \mathcal{B}$ such that the elements of B are precisely the indices of the vertices of type a in D.

Therefore at least k of the n elements $a_1, a_2, \ldots, a_n$ have to be in C.

Assume now that we have in C exactly k elements of type a, and the set of indices of these codewords is $A \subseteq V_1$. If $A = B$ for some k-set $B \in \mathcal{B}$, then again, the set $\{a_i : i \in A\}$ cannot r-separate $b(B)$ and $c(B)$. Therefore $A \notin \mathcal{B}$, that is, $A \in \mathcal{A}$, and clearly the k codewords a_i, $i \in A$, can r-separate all the pairs $(b(B), c(B))$, whether $B \in \mathcal{B}$ has size k or $k-1$. This ends the first step of the proof.

Step 2. We show that for every $A \in \mathcal{A}$, the code

$$\begin{aligned} C(A) \;=\; & \{a_i : i \in A\} \cup E \cup F \cup E' \cup F' \\ & \cup \{g_i(j, r+1) : i = 1, 2, \ldots, s \text{ and } j = 1, 2, \ldots, p\} \\ & \cup \{g'_i(j, r+1) : i = 1, 2, \ldots, t \text{ and } j = 1, 2, \ldots, p\}, \end{aligned} \tag{4}$$

i.e., the code consisting of the vertices a_i, $i \in A$, all the vertices of the cycles E, F, E', F', and the middle vertices of all the paths between E and F, and E' and F', is an r-identifying code in G. Clearly, the size of $C(A)$ equals $k + (p+2)(s+t)$.

For any subset $T \subseteq C(A)$, we say that the T-**signature** of a vertex v is the set $T \cap B_{G,r}(v)$, so that two vertices with different T-signatures are r-separated by T, hence by $C(A)$. By the structure of G, we can tell the following:

- Assume that $v \in \{a_1, a_2, \ldots, a_n\}$ or that $v \in V(B)$ for some $B \in \mathcal{B}$, i.e., v is on level $j \in \{0, 1, 2, \ldots, r\}$; then the E-signature of v has size $2j$ (and the order of E is $s \geq 2r+2 \geq 2j$). By symmetry, the E'-signature of v has size $2r - 2j$ (and $t \geq 2r$). The F- and F'-signatures of v are empty.

Using this —and in particular the fact that all the signatures above have even size— we now see that $C(A)$ can uniquely r-identify all the vertices in

$$E \cup F \cup \{e_i(0), f_i(0) : 1 \leq i \leq s\},$$

as well as all the $(2r+1)sp$ vertices on the paths described in (2) (and in particular that all these vertices are r-covered by $C(A)$); indeed, this is a direct consequence of the following eight observations:

- A vertex v is one of the vertices of type g if, and only if, there is an i and a j such that $g_i(j, r+1) \in I_{G,C(A),r}(v)$; moreover, i and j are unique and identify on which path v is.
- For all i and j, the vertex $g_i(j, r+1)$ has empty E- and F-signatures.
- For all i and j, and for $h \in \{1, 2, \ldots, r\}$, the F-signature of $v = g_i(j, h)$ is empty and its E-signature has size $2r + 1 - 2h$.
- By symmetry, for all i and j, and for $h \in \{r+2, r+3, \ldots, 2r+1\}$, the E-signature of $v = g_i(j, h)$ is empty and its F-signature has size $2h - 2r - 3$.
- The E-signature of $v = e_i(0)$ has size $2r - 1$ and middle point e_i.
- The E-signature of $v = e_i$ has size $2r + 1$ and middle point e_i.
- The F-signature of $v = f_i(0)$ has size $2r - 1$ and middle point f_i.
- The F-signature of $v = f_i$ has size $2r + 1$ and middle point f_i.

In exactly the same way, we can see that all the vertices in

$$E' \cup F' \cup \{e'_i(0), f'_i(0) : 1 \leq i \leq t\},$$

as well as all the $(2r+1)tp$ vertices on the paths described in (3), are r-covered and r-identified using $C(A)$.

The vertices $a_1, a_2, \ldots, a_n$ are also r-identified by $C(A)$ (and r-covered): they are the only vertices whose E'-signature has size $2r$ and clearly the two middle points of this signature give the index of the vertex.

Finally, if $v \in V(B)$ for some $B \in \mathcal{B}$, then the two middle vertices of the E-signature of v identify the set B (and v is r-covered by $C(A)$). Then if the E-signature has size $2r$, we know that v is $b(B)$ or $c(B)$ and we have already seen that they are r-separated by some vertex $a_i \in C(A)$, because $A \in \mathcal{A}$. And if, finally, the E-signature has size $2j$ for some $j \in \{1, 2, \ldots, r-1\}$, then we know that v is on level j, i.e., $v = a_{i,j}(B)$ for some i, and it suffices to determine i. However, this can be done by looking at which are the two middle points of the E-signature of v. Therefore, $C(A)$ is r-identifying in G, which concludes Step 2.

Step 3. We next show that all the vertices in $E \cup F \cup E' \cup F'$ are r-optimal in G, i.e., assuming that C is any optimal r-identifying code in G, we show that $(E \cup F \cup E' \cup F') \subset C$.

By Step 1, at least k vertices of type a must be in C. Furthermore, as the middle points of each of the ps paths described by (2) and the pt paths described by (3) must be r-covered by some codeword, we see that at least one vertex in each of these paths must also be a codeword.

Assume now that $e_i \notin C$ for some fixed i and consider the paths $g_i(j,1)$ $g_i(j,2)\dots g_i(j,2r+1)$ for $j = 1,2,\dots,p$. The only vertices that can r-separate $g_i(j,r)$ and $g_i(j,r+1)$ are e_i and $g_i(j,2r+1)$; but our assumption is that $e_i \notin C$, and therefore $g_i(j,2r+1) \in C$ for all j. Furthermore, as $g_i(j,r)$ is not yet r-covered by any codeword, each of these p paths must contain at least two codewords. Consequently, $|C| \geq k+ps+pt+p > k+(p+2)(s+t)$ (because p has been chosen such that $p > 2s+2t$). By Step 2, C cannot be optimal. Hence e_i is r-optimal. In exactly the same way, we can see that all the vertices of F, E' and F' are r-optimal, which ends Step 3.

Step 4. We now conclude that the size of the optimal r-identifying codes in G is $k+(p+2)(s+t)$. Indeed, assume that C is an optimal r-identifying code in G. We saw in Step 1 that C must contain at least k vertices of type a, and in Step 3 that C must contain all the $2s+2t$ vertices in $E \cup F \cup E' \cup F'$. As was already noticed in Step 3, the middle points of each of the sp paths described in (2) and of the tp paths described in (3) must be r-covered by C, and therefore at least one vertex in each of these $p(s+t)$ paths is a codeword. Hence $|C| \geq k+(p+2)(s+t)$. Our claim now follows, because we saw in Step 2 that an r-identifying code with this size exists in G.

Step 5. To conclude the proof of Theorem 12, it suffices to show that every optimal r-identifying code in G is one of the codes $C(A)$ described in (4). Assume that C is any optimal r-identifying code in G. As it is optimal, it cannot contain more than k codewords of type a by Step 4, and by Step 1, it must contain at least k of these, and they must form one of the sets $\{a_i : i \in A\}$ for some $A \in \mathcal{A}$. By Step 3, $(E \cup F \cup E' \cup F') \subset C$, and by Step 4, exactly one vertex in each of the $p(s+t)$ paths described by (2) and (3) is a codeword. Consider any of them, say, the path $g_i(j,1)g_i(j,2)\dots g_i(j,2r+1)$. The code C must be able to r-separate between $e_i(0)$ and $g_i(j,1)$. Because $r \geq 2$, the vertex $e_i(0)$ cannot do it, and neither can any vertex in E; the only possibility is that $g_i(j,1)$ is r-covered by at least one codeword belonging to this particular path. In the same way, C must r-separate between $g_i(j,2r+1)$ and $f_i(0)$, which implies that $g_i(j,2r+1)$ is also r-covered by at least one codeword belonging to this path. The only way to do this with one vertex is to take the middle vertex of the path, $g_i(j,r+1)$, in the code C. As this is true for all our paths, we must have $C = C(A)$ for some $A \in \mathcal{A}$, which ends Step 5 and the proof of Theorem 12: in Step 4 we showed that $\gamma_r(G) = k+(p+2)(s+t)$, and in Step 2 that the code $C(A)$ given by (4) is r-identifying if $A \in \mathcal{A}$, and optimal; conversely, Step 5 just showed that every optimal r-identifying code in G is of the form (4) for some $A \in \mathcal{A}$. $\square$

Remark 13 *In the proofs of Theorems 10 and 12, we have not tried to minimize the number of vertices in the construction of the graph G. For* *instance, the proof of Theorem 12 could be modified so that $p = 3$ would be sufficient (instead of $p > 2s + 2t$), thus saving a large number of vertices in the construction. On the other hand, the proof would be more difficult.*

2.4 Known results on Johnson induced subgraphs

Some families of graphs are known to be JIS, some are known which are not JIS, but no characterization is available. Below, we summarize some of the results from [11].

Theorem 14 *(a) [Prop. 4] All complete graphs and all cycles are JIS;*

(b) [Prop. 5] All trees are JIS;

(c) [Prop. 6] A graph is a JIS if, and only if, all its connected components are JIS;

(d) [Prop. 7] The Cartesian product of two JIS is a JIS;

(e) [Prop. 12] Any graph obtained by removing one edge from the complete graph K_n, $n \geq 5$, is not a JIS;

(f) [Prop. 8] The complete bipartite graph $K_{2,3}$ is not a JIS. □

The graph $K_{2,3}$ can be seen as two cycles of length four sharing three vertices; if we define the graph θ_n as the graph consisting of two cycles of length n sharing $n - 1$ vertices, we have the following result from [11].

Theorem 15 *The graphs θ_4 and θ_5 are not JIS; all the graphs θ_n, $n \geq 6$, are JIS.* □

The q-ary n-dimensional hypercube is another graph which is JIS, for all $q \geq 2$ and $n \geq 1$; indeed, the q-ary words of length n in Z_q^n can be transformed into binary sequences of length qn, containing exactly n ones, applying the mapping $\phi : Z_q \to Z_2^q$, with $\phi(0) = e_1$ and $\phi(i) = e_{i+1}$ for $i \in \{1, 2, \ldots, q-1\}$, where e_i has exactly one "1" in position i, so that Z_q^n can be seen as a collection of n-subsets of $\{1, 2, \ldots, qn\}$. Theorem 7 and its corollary had already shown, in a simpler way than Theorem 10, how to construct a graph G such that $\mathcal{M}_1(G) \cong Z_2^n$, for all $n \geq 4$.

References

[1] C. BERGE: *Graphes*, Gauthier-Villars: Paris, 1983. English translation: *Graphs*, North-Holland Publishing Co.: Amsterdam, 1985.

[2] I. CHARON, O. HUDRY and A. LOBSTEIN: Minimizing the size of an identifying or locating-dominating code in a graph is NP-hard, *Theoretical Computer Science*, Vol. 290, pp. 2109–2120, 2003.

[3] G. D. COHEN, I. S. HONKALA, S. LITSYN and A. C. LOBSTEIN: *Covering Codes*, Elsevier: Amsterdam, 1997.

[4] R. DIESTEL: *Graph Theory*, Springer-Verlag: Berlin, 2005.

[5] S. GRAVIER, S. JANSON, T. LAIHONEN and S. RANTO: Graphs where every k-subset of vertices is an identifying set, *Discrete Mathematics & Theoretical Computer Science*, Vol. 16, pp. 73–88, 2014.

[6] I. HONKALA, O. HUDRY and A. LOBSTEIN: On the number of optimal identifying codes in a twin-free graph, *Discrete Applied Mathematics*, Vol. 180, pp. 111–119, 2015.

[7] I. HONKALA, O. HUDRY and A. LOBSTEIN: On the ensemble of optimal dominating and locating-dominating codes in a graph, *Information Processing Letters*, to appear.

[8] M. G. KARPOVSKY, K. CHAKRABARTY and L. B. LEVITIN: On a new class of codes for identifying vertices in graphs, *IEEE Transactions on Information Theory*, Vol. IT-44, pp. 599–611, 1998.

[9] A. LOBSTEIN: A bibliography on watching systems, identifying, locating-dominating and discriminating codes in graphs, http://perso.telecom-paristech.fr/~lobstein/debutBIBidetlocdom.pdf

[10] F. J. MACWILLIAMS and N. J. A. SLOANE: *The Theory of Error-Correcting Codes*, North-Holland: Amsterdam, 1977.

[11] R. NAIMI and J. SHAW: Induced subgraphs of Johnson graphs, *Involve*, Vol. 5, pp. 25–37, 2012.

[12] M. PELTO: On identifying and locating-dominating codes in the infinite king grid, Ph. D. Thesis, University of Turku, Finland, 133 pages, 2012.

[13] M. PELTO: A definition of uniqueness for optimal identifying and covering codes in infinite lattices, submitted, 2012.