Incompatibility Graphs in Data Mining

Bruno Simeone[1], Endre Boros[2], Federica Ricca[1], and Vincenzo Spinelli[3]

[1] Dip. di Scienze Statistiche, Sapienza, Università di Roma,
Piazzale Aldo Moro 5, 00185, Rome, Italy
federica.ricca@uniroma1.it
[2] RUTCOR - Rutgers Center for Operations Research, Rutgers University,
640 Bartholomew Road, Piscataway, NJ 08854-8003
boros@rutcor.rutgers.edu
[3] ISTAT - Istituto Nazionale di Statistica,
Via Tuscolana, 1788, 00173, Rome, Italy
vispinel@istat.it

Summary. In this paper, we investigate a new class of graphs, called "Incompatibility Graphs" which arise from Box Clustering. Besides their importance for the applications in data mining, these graphs have an intrinsic interest from a theoretical viewpoint, since they generalize some important classes of graphs, namely, chordal and weakly chordal graphs. The special structure of the Incompatibility Graphs can be exploited to efficiently solve some key-problems related to Box Clustering, such as the "Maximum Box" and the "Minimum Covering by Boxes" problems. In fact, we show that these two problems can be formulated as a vertex packing and a vertex coloring one, respectively, in an Incompatibility Graph, and that one can solve in polynomial time the former and, for two important subclasses of instances, also the latter.

Key words: box clustering, forbidden graphs, vertex packing, vertex coloring, graph recognition.

1 Introduction

In this paper, we introduce Incompatibility Graphs (IGs), a class of graphs that arises in the *Box Clustering* (*BC*) approach to the supervised classification of data. Box Clustering was introduced in [12] and it can be viewed as an offspring of a more general methodology, called *Logical Analysis of Data* (*LAD*) (see, for example, [2, 3, 4, 7, 11]). Unlike *LAD*, besides binary data, *BC* is also able to deal with numerical and ordinal data. The input of a *BC* problem is a *training* data set, consisting of a finite set of points in a d-dimensional space, which are classified either as *positive* or *negative* according to the value of the given classification variable. A box (i.e., a d-dimensional closed interval) is called *positive* (or *negative*) if it includes some positive (resp. negative) observations, but does not include any negative (resp. positive)

one. Positive and negative boxes will also be called *homogeneous*. The output of a *BC* model is a set of homogeneous boxes, which are used to predict the class of the observations belonging to a *testing* data set (see, [18]).
In this paper we address two key-problems related to *BC*, namely, the Maximum Box (*MB*) and the Minimum Covering by Boxes (*MCB*), see [8, 12]. In order to solve these problems, we introduce Incompatibility Graphs that, for a given set of observations in the d-dimensional space, represent the structural relations of homogeneity between pairs of points. These graphs have shown to be of help in the solution of both the two above mentioned *BC* problems, but they also showed to have an autonomous interest from a theoretical viewpoint.

The broad structure of the paper is as follows. In Section 2, we recall some basic notions and definitions in *BC* and we formally state problems *MB* and *MCB*. In Section 3, we give the general definition of IGs and we show that any arbitrary graph is an IG in a space of sufficiently large dimension. By restricting the dimension of the space, we obtain a much richer structure. From Section 4 on, we focus on the two dimensional case, and we show that, in this case, IGs feature strong structural properties. In particular, they are shown to be a generalization of weakly chordal graphs. In Section 6, we also show that IGs have small radius. In section 5 we introduce a subclass of IGs, called "polarized incompatibility graphs", having a structure similar to IGs, but characterized by some special feature related to the relative position of points in $\mathbb{R}^2$. In Section 7, we show that in the two dimensional case *MB* and *MCB* can be formulated as a vertex packing and a vertex coloring problem, respectively, on an IG; the special structure of IGs makes the former one, and, in many instances, also the latter one, solvable in polynomial time. In Section 8, we summarize our main conclusions, pointing out some directions for future research. Finally, in the Appendix, we provide a mixed integer linear programming model for the recognition of IGs.

2 Box Clustering

As previously mentioned, *BC* has the capability to deal directly with numerical, ordinal and binary variables. Since a binary variable can be thought to take values in $\{0,1\}$ and an ordinal variable in the standard set $\{1,\dots,p\}$ for some positive integer p, we may regard an observation, w.l.o.g., as a point in the real d-dimensional space $\mathbb{R}^d$. Suppose that a set S of (positive and negative) observations is given in $\mathbb{R}^d$ and that $x \in \mathbb{R}^d, x = (x_1,\dots,x_i,\dots,x_d)$, is the generic vector of an observation in the data set. Let $l,u \in \mathbb{R}^d$ be such that $l_i \le u_i, i = 1,\dots,d$. A *box* I (or *hyper-rectangle*) in $\mathbb{R}^d$ is defined as follows:

$$I(l,u) = \{x \in \mathbb{R}^d : l_i \le x_i \le u_i, i = 1,\dots,d\}. \tag{1}$$

Given a finite set $S \subset \mathbb{R}^d$ the *box-closure* of S is the smallest box (i.e., the intersection of all boxes) containing all points in S (see [12]). Let $L_i = \min_{x\in S}\{x_i\}$ and $U_i = \max_{x\in S}\{x_i\}$, $i = 1,\dots,d$. The box-closure of S is given by $[S] = I(L,U)$. Notice that the two vectors L and U defines two *bounding points* in $\mathbb{R}^d$ that univocally determine

the box $I(L,U)$. For the sake of simplicity, in the following the same notation B will be adopted to denote both a box and the set of points included in a box.
For any finite set S of points, the box-closure of S can be also seen as the intersection of all boxes containing S, that is:

$$[S] = \bigcap_{\substack{l,u \in \mathbb{R}^d : l \leq u \\ I(l,u) \supseteq S}} I(l,u). \tag{2}$$

Suppose that $S \subset \mathbb{R}^d$ is the set of the n points in $\mathbb{R}^d$ representing a *BC* data set. Let P and N denote the two finite non empty subsets of S corresponding to the positive and the negative points, respectively, so that one has $P \cap N = \emptyset$ and $S = P \cup N$. Let us denote by $\mathscr{B}$ a set of boxes in $\mathbb{R}^d$ with $|\mathscr{B}| = m$, i.e., $\mathscr{B} = \{B_1, \ldots, B_m\}$.

Definition 1. $\mathscr{B}$ is a covering for P if (a) every point $x \in P$ is included in some $B_i \in \mathscr{B}$; (b) every box B_i includes at least one point $x \in P$. In other words, $S \subseteq \bigcup_{i=1}^{m} B_i$ and $S \cap B_i \neq \emptyset$ for every $i = 1, \ldots, m$.

Definition 2. $\mathscr{B}$ is a homogeneous set of boxes for P if for every box $B_i \in \mathscr{B}$ one has $N \cap B_i = \emptyset$ and $P \cap B_i \neq \emptyset$.

If $\mathscr{B}$ is homogeneous, we can univocally assign a positive label to each box.

Definition 3. A *box system* for P is a covering by homogeneous boxes for P.

Similar definitions hold for the set of negative points N. On the basis of the above definitions, different *BC* models were provided in the literature in order to study different data analysis problems [19, 20]. One of the main *BC* problems is *Minimum Covering by Boxes*, which can be formulated as follows:

- *Given a set $S = P \cup N$ of points in $\mathbb{R}^d$, find a box system $\mathscr{B}^*$ for P (or N) such that the number of boxes in $\mathscr{B}^*$ is a minimum.*

The *Maximum Box* problem can be stated as follows:

- *Given a set $S = P \cup N$ of points in $\mathbb{R}^d$, find a positive (resp. negative) box that contains the largest number of points in P (resp. N).*

Without loss of generality, in the rest of the paper we will refer to the case of systems of positive boxes. The case of systems of negative boxes is similar.

3 Incompatibility Graphs

Given two finite and disjoint sets of points $P,N \subseteq \mathbb{R}^d$, we associate to them a graph $G = G_{P,N}$ with vertex set $V(G) = P$ and such that two vertices $u, v \in P$ are connected in G by an edge if $[u,v] \cap N \neq \emptyset$. We call $G_{P,N}$ the *d-incompatibility graph* (d-IG) of P and N.

Definition 4. *(General definition of IG)*
A graph $G=(V,E)$ is a d-IG, if there are two finite and disjoint point sets $P,N\subseteq\mathbb{R}^d$ such that $G_{P,N}$ is equal to G. The pair (P,N) will be called a *d-embedding* of G.

We shall say that $y\in N$ *fathers*, or *triggers*, edge $(a,b)\in E$, if $a,b\in P$ and $y\in[a,b]$. Notice that the edge $(a,b)\in E$ to be triggered by more than one point $y\in N$.

Theorem 1. *(Universality of IGs)*
If $G=(V,E)$ is an arbitrary graph then there always exist $d\geq 1$ and $P,N\subseteq\mathbb{R}^d$ such that $G_{P,N}$ is isomorphic to G, i.e., it is a d-IG.

Proof. Let us choose $d=|V|$, and define the set of positive points $P=\{e_i\}_{i=1,\ldots,d}$, where $e_i=(0,\ldots,1,\ldots,0)$. If $(e_i,e_j)\in E$, then we consider the negative point in N: $q_{ij}=e_i+e_j=(0,\ldots,1,\ldots,1,\ldots,0)$. We check that $G=G_{P,N}$ is d-IG. We have $P\cap N=\emptyset$. Moreover, for every $e_i,e_j\in P$: such that (e_i,e_j) is an edge of G, one has: $[e_i,e_j]\cap N=\{q_{ij}\}$. □

The above theorem states that every graph G can be seen as a d-IG, but this is not true in every dimension d (see Section 4). After Theorem 1, we will refer to a d-IGsimply by IG when this does not cause any confusion. The following two properties hold.

Property 1. (IG hereditary property)
G is a d-IG iff every induced subgraph of G is a d-IG.

Property 2. (IG monotonic dimension property)
If G is d-IG then G is h-IG for every $h\geq d$.

On the one hand, in view of Property 1, we are interested in investigating the minimal (w.r.t. the vertex-set inclusion) *non*-IGs, i.e. those graphs that are not IGs but all their induced subgraphs are. On the other hand, after Property 2, we are interested in defining the minimum space dimension where the graph G can be embedded as an IG. Generally, this is no easy task, but there are some trivial classes of d-IGs. We use the standard notation P_r, C_r, K_r to denote paths, cycles and cliques with r vertices; and $K_{r,s}$ to denote complete bipartite graph with r and s vertices, respectively, on the two sides.

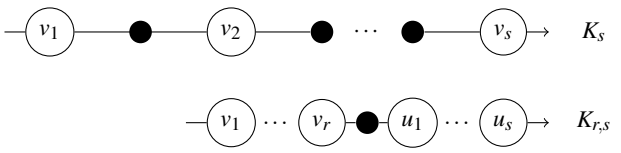

Fig. 1. Embeddings for K_s and $K_{r,s}$ in $\mathbb{R}$.

Proposition 1. *(Complete graphs)*
For every $r,s,d\geq 1$: the graphs K_r and $K_{r,s}$ are d-IGs.

Proof. Figure 1 shows examples of embedding for the two graphs as 1-IGs and, hence, d-IGs for any $d \geq 1$ by Property 2. □

In this paper, we use a very specific notion of "general position" which is stated in the following definition.

Definition 5. A finite and non-empty set $S \subseteq \mathbb{R}^d$ is in *general position* if no pair of points lie on the same hyperplane parallel to a coordinated one.

In Figure 2-(a) the points $\{i, j, k, r\}$ are in general position in $\mathbb{R}^2$, but the points $\{i, j, k, r, s\}$ in Figure 2-(b) are not, in fact, the points s and k are aligned on the same vertical line. Let us consider C_6 and its two embeddings shown in Figure 3. The embedding (a) has four negative points, while (b) has only two, and its positive points are not in general position. Let us first observe that if G is d-IG, then we may always assume that the points of $S = P \cup N$ are in general position.

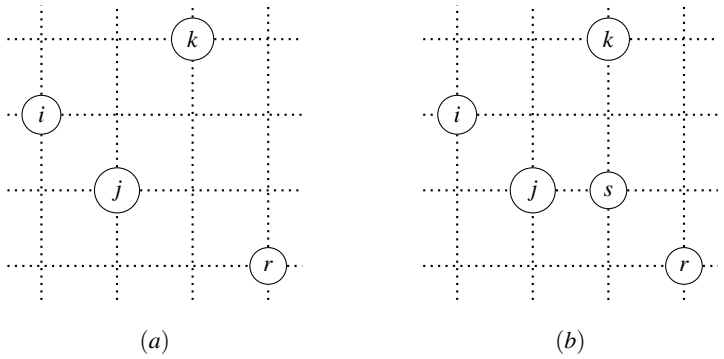

Fig. 2. Embeddings in the plane.

Theorem 2. *(Embeddings and general position)*
If G is a d-IG, then it is isomorphic to some $G_{P,N}$ ($G \sim G_{P,N}$), such that $P \cap N = \emptyset$ and the points of $P \cup N$ are in general position.

Proof. Let us assume, for simplicity, that $d = 2$. Let $G \sim G_{P',N'}$ for some point sets with $P', N' \subseteq \mathbb{R}^2$, and let us choose constants $\varepsilon,\ \delta$ such that $0 < 2\varepsilon < \delta$ and

$$2\delta < \min\{d([a,b],y) \mid a,b \in P',\ a \neq b,\ (a,b) \notin E(G), y \in N'\},$$

where $d(a,b)$ denotes the Euclidean distance of points a and b, and for point sets $d(A,B)$ is defined by $d(A,B) = \min_{a \in A,\ b \in B} d(a,b)$. Let us note that all quantities on the right hand side of the above inequality are positive, and hence such positive ε and δ exist. Let us denote by S the set of unit vectors in the plane, and let

$$S' \ = \ S \setminus \left(\{(1,0),(-1,0),(0,1),(0,-1)\} \cup \left\{ \frac{a-b}{\|a-b\|} \,\middle|\, a,b \in P',\ n \neq b \right\} \right).$$

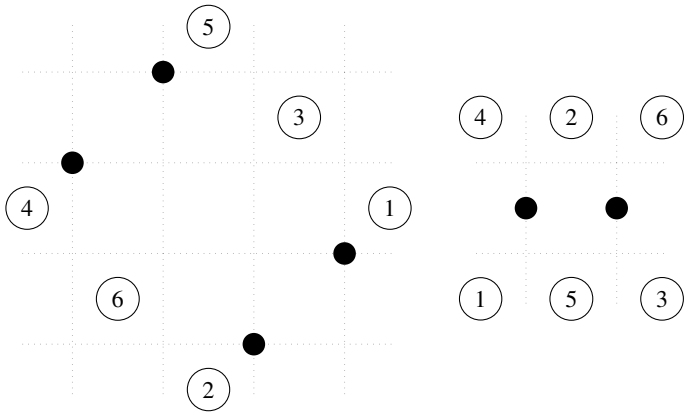

Fig. 3. Two embeddings of C_6 as IG.

Choose vectors $v_a \in S'$ for $a \in P'$ such that $v_a \neq v_b$ whenever $a \neq b$, and set

$$P \;=\; \{a+\varepsilon v_a \mid a \in P'\}.$$

Let us further denote by $D(y)$ the open disk of radius δ around point $y \in N'$, i.e.,

$$D(y) \;=\; \{z \mid d(y,z) < \delta\}.$$

Let us note on the one hand that for all $a,b \in P'$, $(a,b) \notin E(G)$ and $y \in N'$ we must have $d(D(y),[a+\varepsilon v_a, b+\varepsilon v_b]) > 2\delta - (\delta+\varepsilon) > \varepsilon > 0$, i.e., $D(y) \cap [a+\varepsilon v_a, b+\varepsilon v_b] = \emptyset$. On the other hand, for all $a,b \in P'$ and for all $x \in [a,b] \setminus [a+\varepsilon v_a, b+\varepsilon v_b]$ we have $d(x,[a+\varepsilon v_a, b+\varepsilon v_b]) \leq \varepsilon$, and consequently, for all $a,b \in P'$ and $y \in N' \cap [a,b]$ we must have $D(y) \cap int([a+\varepsilon v_a, b+\varepsilon v_b]) \neq \emptyset$. Since this intersection is a nonempty open set, we can choose a $z(y,a,b) \in D(y) \cap int([a+\varepsilon v_a, b+\varepsilon v_b]) \neq \emptyset$ for all such tuples such that no two of these vectors are on the same horizontal or vertical line, and none of them are on the same horizontal or vertical line as a point in P. Let

$$N \;=\; \{z(y,a,b) \mid a,b \in P',\; a \neq b, y \in N' \cap [a,b]\}.$$

Due to the above properties, we can conclude that $G_{P',N'} \sim G_{P,N}$. The proof for arbitrary dimension d is based on a similar perturbation technique. □

At first sight, IGs seem to be a geometrical concept. The following theorem provides a purely order-theoretic characterization.

Theorem 3. *$G = (V,E)$ is a d-incompatibility graph iff there are two finite sets P, N, with $|P| = |V|$, $P \cap N = \emptyset$, and d linear orders $\prec_i, i = 1,\ldots,d$ on $S = P \cup N$ such that $(a,b) \in E$ iff there is some $q \in N$ for which*

$$(a \wedge_i b) \prec_i q \prec_i (a \vee_i b), \quad i = 1, \ldots, d \tag{3}$$

where $(a \wedge_i b)$ *and* $(a \vee_i b)$ *denote the minimum and the maximum, respectively, of* a *and* b *in the linear order* $\prec_i, i = 1, \ldots, d$.

Proof. Only if). Let (P,N) be a d-embedding of G such that the points in S are in general position, and define for $u, v \in S$

$$u \prec_i v \text{ iff } u_i < v_i \tag{4}$$

Then (3) holds.
If). Let $s = |S|$. For each $u \in S$, let $\mathbf{r}(u) \in \{1, \ldots, s\}^d$ be the vector whose i-th component $r_i(u)$ is the rank of u in the linear order $\prec_i, i = 1, \ldots, d$ (so the least element of S in $\prec_i$ has rank 1 and the greatest one has rank s). Finally, set $P^* = \{\ \mathbf{r}(u) : u \in P\}$ and $N^* = \{\mathbf{r}(u) : u \in N\}$. Notice that

$$(a \wedge_i b) \prec_i q \prec_i (a \vee_i b) \text{ iff } \min\{r_i(a), r_i(b)\} < r_i(q) < \max\{r_i(a), r_i(b)\},\ i = 1, \ldots, d.$$

Therefore, G is isomorphic to G_{P^*,N^*} and thus it is a d-IG. □

4 Incompatibility Graphs in the plane

In this section we shall focus on the case of $d = 2$. To simplify terminology, we shall call a graph $G = (V,E)$ an *incompatibility graph* if it is a 2-incompatibility graph. A 2-embedding of G will be called a *plane embedding* or simply an *embedding* of G. By Property 1 in Section 3, any induced subgraph of an incompatibility graph is also an IG. Thus, incompatibility graphs can be characterized by a (possibly infinite) list of forbidden subgraphs. Given $x = (x_1, x_2) \in \mathbb{R}^2$, let us define its (open) orthants as follows:

$$NE(x) = \left\{ (y_1,y_2) \middle| \begin{array}{l} x_1 < y_1 \\ x_2 < y_2 \end{array} \right\} \qquad SE(x) = \left\{ (y_1,y_2) \middle| \begin{array}{l} x_1 < y_1 \\ x_2 > y_2 \end{array} \right\}$$

$$NW(x) = \left\{ (y_1,y_2) \middle| \begin{array}{l} x_1 > y_1 \\ x_2 < y_2 \end{array} \right\} \qquad SW(x) = \left\{ (y_1,y_2) \middle| \begin{array}{l} x_1 > y_1 \\ x_2 > y_2 \end{array} \right\}$$

Since we can assume w.l.o.g. that all point sets realizing IGs are in general position, see Theorem 2, for any two points $x, y \in P$, $x \neq y$, we have exactly one of the containments $y \in NE(x)$, $y \in NW(x)$, $y \in SW(x)$, or $y \in SE(x)$. Let us then call a pair $x, y \in \mathbb{R}^2$ a *monotone pair*, if either $x \in NE(y)$ or $y \in NE(x)$, and call it a *saddle pair* otherwise. Note that for a saddle pair we must have either $x \in NW(y)$ or $y \in NW(x)$. Two pairs are *coherent* if they are either both monotone or both saddle. Clearly, whenever y belongs to an orthant of x, then x belongs to the opposite type of orthant of y. For example, $x \in NE(y)$ iff $y \in SW(x)$.
Let us now go back to the structure of IGs. Note first that if $y \in N$, then P is partitioned into four sets by the four orthants of y, and the two sets of points lying within opposite orthants are the vertices of a complete bipartite subgraph of $G_{P,N}$, not necessarily induced since other points of N may trigger more edges.

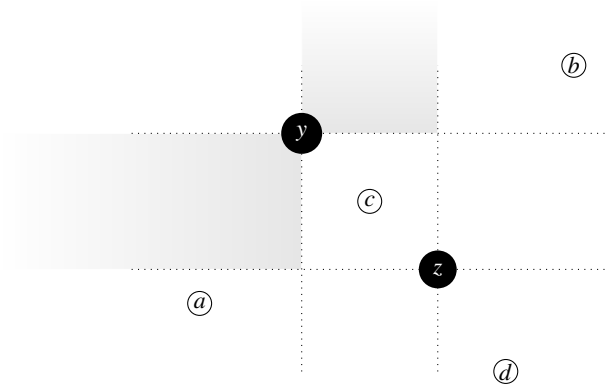

Fig. 4. A possible configuration in Lemma 1.

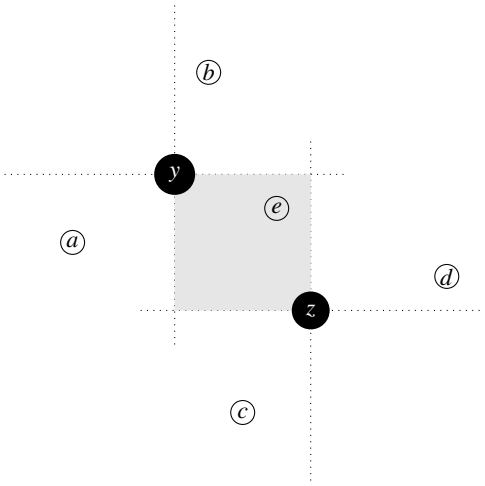

Fig. 5. Configuration corresponding to Lemma 2.

Lemma 1. *If (a,b) and (c,d) are two edges of an induced $2K_2$ of the IG $G = G_{P,N}$ such that a,b is a monotone pair and c,d is a saddle pair, then $[a,b] \cap [c,d] \cap N \neq \emptyset$, or in other words, there is a single point in N that triggers both edges (a,b) and (c,d). Consequently, any other point $e \in P$ must be connected by an edge in G to at least one of $\{a,b,c,d\}$.*

Proof. Let us denote by $y,z \in N$ the points triggering the edges (a,b) and (c,d), respectively. After some rotations and re-labeling, if necessary, we can assume that $a \in SW(y)$, $b \in NE(y)$, $c \in NW(z)$, $d \in SE(z)$ and $y \in NW(z)$, see Figure 4. Note that c may belong to any of the orthants of y. Nevertheless. since (a,d) and (b,d) are not

edges of G, $a,b \notin NW(z)$ follows, and consequently, $a \in SW(z)$ and $b \in NE(z)$ are implied. Thus, $z \in [a,b] \cap [c,d] \cap N$, as claimed. Therefore, for any other point $e \in P$, $z \in N$ will trigger an edge between e and one of $\{a,b,c,d\}$, no matter where e is on the plane. □

Lemma 2. *If (a,b) is an edge triggered by $y \in N$ and (c,d) is an edge triggered by $z \in N$ forming an induced $2K_2$ of the IG $G = G_{P,N}$ such that the pairs a,b and c,d are coherent, and $e \in P$ is a vertex which is not connected to any of the points $\{a,b,c,d\}$, then the points $\{a,b,e\}$ belong to the same orthant of z ($NW(z)$ or $SE(z)$), the points $\{c,d,e\}$ belong to the same orthant of y ($NW(y)$ or $SE(y)$), and in particular $e \in [y,z]$.*

Proof. After appropriate rotations and/or relabeling we can assume w.l.o.g. that $a \in SW(y)$, $b \in NE(y)$, $c \in SW(z)$, $d \in NE(z)$ and $y \in NW(z)$. Since there is no edge between $\{a,b\}$ and $\{c,d,e\}$ we must have $\{c,d,e\} \subset NW(y) \cup SE(y)$. If $\{c,d,e\}$ were not all in the same orthant of y, then y would trigger an edge between these vertices, which then could only be (c,d). But then y would also trigger an edge between e and one of $\{a,b,c,d\}$, a contradiction, showing that $\{c,d,e\}$ all belong to the same orthant of y. Then z must also belong to this orthant, and hence we have $\{c,d,e\} \subset SE(y)$ implied. Analogous arguments show that we must have $\{a,b,e\} \subseteq NW(z)$. Thus, $e \in NW(z) \cap SE(y) = [y,z]$ is implied (see Figure 5). □

Theorem 4. *An incompatibility graph cannot have $3K_2$ as an induced subgraph.*

Proof. Assume indirectly that $(a_i,a_i') \in E$, $i = 1,2,3$ form an induced $3K_2$ in an incompatibility graph $G = G_{P,N}$, and let us denote by $y_i \in N$ the corresponding points inducing these edges, respectively. Let us first note that the pairs $\{a_i,a_i'\}$ must all be coherent. Otherwise, if e.g., $\{a_1,a_1'\}$ and $\{a_2,a_2'\}$ are not coherent, then by Lemma 1 we can assume that $y_1 = y_2$, and then all four orthants of y_1 will contain a point of $\{a_1,a_1',a_2,a_2'\}$. Thus, y_1 would trigger an edge between a_3 and at least one of $\{a_1,a_1',a_2,a_2'\}$, contradicting the assumption that no such edge exists. Thus, after appropriate rotations we can assume that all pairs $\{a_i,a_i'\}$, $i = 1,2,3$ are monotone. Let us apply now Lemma 2 for all 5-tuples obtained by deleting one of a_i or a_i', $i = 1,2,3$. It follows that $\{a_i,a_i'\} \subseteq [y_j,y_k]$ for all $i \neq j \neq k \neq i$, and thus $y_i \in [y_j,y_k]$ is implied for all choices of $\{i,j,k\} = \{1,2,3\}$. This leads to a contradiction, since if, e.g., y_1 and y_2 are the closest in Euclidean distance, then $y_3 \notin [y_1,y_2]$. □

Theorem 5. *An incompatibility graph cannot have C_7 as an induced subgraph.*

Proof. Assume indirectly that $\{a_0,a_1,\ldots,a_6\}$ forms an induced C_7 of $G = G_{P,N}$, where $(a_i,a_{i+1}) \in E$ and indices are meant mod 7 here and in the sequel. Assume first that the pairs $\{a_i,a_{i+1}\}$ are not coherent. Then we must have pairs $\{a_i,a_{i+1}\}$ and $\{a_{i+3},a_{i+4}\}$ that are not coherent. Then, by Lemma 1 we have a point $y \in N$ inducing both edges (a_i,a_{i+1}) and (a_{i+3},a_{i+4}). We can always assume w.l.o.g. that $i = 0$, and $a_0 \in SW(y)$, $a_1 \in NE(y)$, $a_3 \in SE(y)$ and $a_4 \in NW(y)$. Since a_6 is not connected to $\{a_1,a_3,a_4\}$, we must have $a_6 \in NE(y)$, and analogously $a_5 \in SE(y)$. Furthermore, since a_2 is not connected to $\{a_0,a_4\}$, vertex a_2 must lie west of y. But

then, no matter where a_2 is, y will trigger an edge between a_2 and one of $\{a_1, a_3\}$. This contradiction proves that all edges of this C_7 configuration must be of the same orientation, say all of them are monotone. Then, every edge $(a_i, a_{i+1}) \in E$ moves from west to east or east to west (and south to north or north to south at the same time). Since $(a_i, a_{i+2}) \notin E$ these moves must alternate as we go around this C_7. Since C_7 has an odd number of edges, this leads to a contradiction, proving the claim. □

Theorem 6. *An incompatibility graph cannot have C_8 as an induced subgraph.*

Proof. Assume indirectly that $\{a_0, a_1, \ldots, a_7\}$ forms an induced C_8 of $G = G_{P,N}$, where $(a_i, a_{i+1}) \in E$ and indices are meant mod 8 here and in the sequel. Analogously to the proof of Theorem 4, we can conclude that all edges of this C_8 configuration must be coherent, say all of them are monotone. Let us denote by $y_i \in N$ the point inducing edge $(a_i, a_{i+1}) \in E$. Since $a_i, a_{i+1}, a_{i+3}, a_{i+4}$ and a_{i+6} form a $2K_2$ plus an isolated vertex, we can apply Lemma 2, and conclude that $a_{i+6} \in [y_i, y_{i+3}]$, and that the pairs $\{y_i, y_{i+3}\}$ are saddle. Thus, for every index i we can conclude that there is a point y_j more to the east than a_i. On the other hand, since every edge of this C_8 is monotone, one of $\{a_j, a_{j+1}\}$ is more to the east than y_j, for all indices $j = 0, \ldots, 7$. Thus, the eastmost point in this subconfiguration can be neither an a_i nor an y_j, leading to a contradiction which proves our claim. □

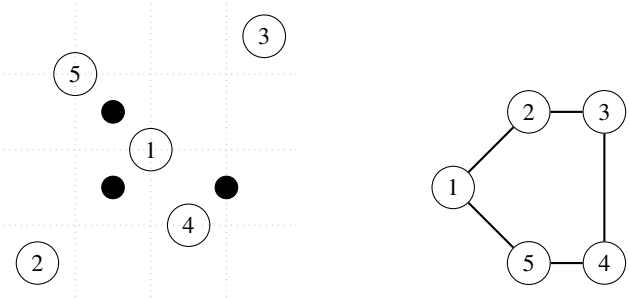

Fig. 6. C_5 as IG.

We notice here that C_5 and C_6 are IGs: in Figure 6 one can see an embedding for C_5, while Figure 3 shows two embeddings for C_6.

Corollary 1. *An incompatibility graph cannot have C_n, $n \geq 7$, as an induced subgraph.*

Proof. If $n = 7, 8$ the statement is true by Theorems 5 and 6, respectively; if $n \geq 9$ then $3K_2$ is an induced subgraph of C_n and the corollary is true by Theorem 4. □

After the above result, IGs can be considered as a generalization of weakly chordal graphs [10], i.e., graphs with forbidden C_n for each $n > 4$, by Corollary 1.

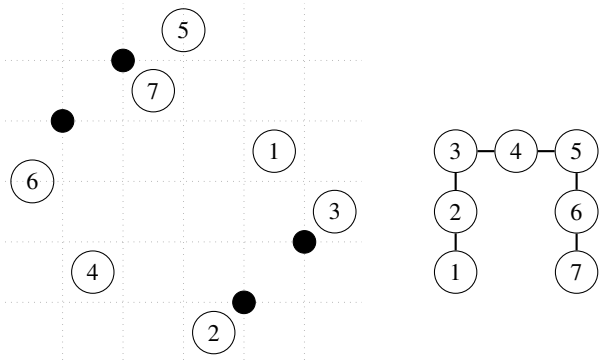

Fig. 7. P_7 as IG.

In Figure 7 one can see an embedding for P_7 as an IG, but the following result states that IGs cannot have long paths as induced subgraphs.

Corollary 2. *An incompatibility graph cannot have P_n, $n \geq 8$, as an induced subgraph.*

Proof. If $n \geq 8$ then $3K_2$ is an induced subgraph of P_n and the corollary is true after Theorem 4. □

Remark. If a graph is a d-IG we can substitute each vertex by a stable set and still obtain a d-IG.

Theorem 7. *If $G \cup K_1$ is IG then $G \cup nK_1$ is IG for every $n \geq 1$.*

Proof. Let $K_1 = \{x\}$. For any $y \in N$, let $\Omega(x,y)$ be the orthant of y containing x. Since vertex x is isolated, the orthant of y opposite to $\Omega(x,y)$ does not contain any point in P. Let $\Omega = \Omega(x) \equiv \bigcap_{y \in N} \Omega(x,y)$. Then Ω is either an orthant or an open rectangle. Place $n-1$ positive points in Ω. Each such point x' must be an isolated vertex, since for each $y \in N$ $x' \in \Omega(x,y)$ and thus x' cannot be adjacent to any vertex of G, else also x would. □

We realized that, in addition to the forbidden subgraphs reported so far, there exist a great many others. Relying on a concise (polynomial-sized) mixed integer linear programming formulation of the recognition problem for *non*-IGs (see Appendix A), we have written a computer program for the generation of all (vertex-wise) minimal non-IGs with n vertices. We have run this program[4] for $n \in \{1,2,\ldots,10\}$, and the results are shown in Table 1.

[4] The procedure is based on some functions of the *Nauty* environment (http://cs.anu.edu.au/~bdm/nauty). In particular, we used the package of programs called "*gtools*" distributed along with *Nauty* to provide efficient processing of files of graphs (see, [17, 22]).

Table 1. Summary of minimal non-IGs.

n	Graphs with n vertices	non-IGs		minimal non-IGs	
1	1	0		0	
2	2	0		0	
3	4	0		0	
4	11	0		0	
5	34	0		0	
6	156	1	0.6%	1	100.0%
7	1,044	14	1.3%	4	28.6%
8	12,346	454	3.7%	76	16.7%
9	274,668	31,767	11.6%	2,956	9.3%
10	12,005,168	3,632,681	30.3%	102,292	2.8%

The third column shows the total number of non-IGs for the graphs with n vertices, and the fourth one the percentage with respect to the total number of graphs having n vertices. Notice that $3K_2$ is the only non-IGwith 6 vertices. The fifth and sixth columns show the number of minimal non-IGs and their percentage w.r.t. non-IGs, respectively. From this table it is clear that, when n increases, the number of non-IGs rapidly increases, while the percentage of minimal non-IGs rapidly decreases. Hence we did not go beyond $n = 10$ vertices.

Even if, for obvious reasons, we cannot present all the minimal non-IGs of Table 1, in Figure 8 we show the four minimal non-IG graphs with $n = 7$.

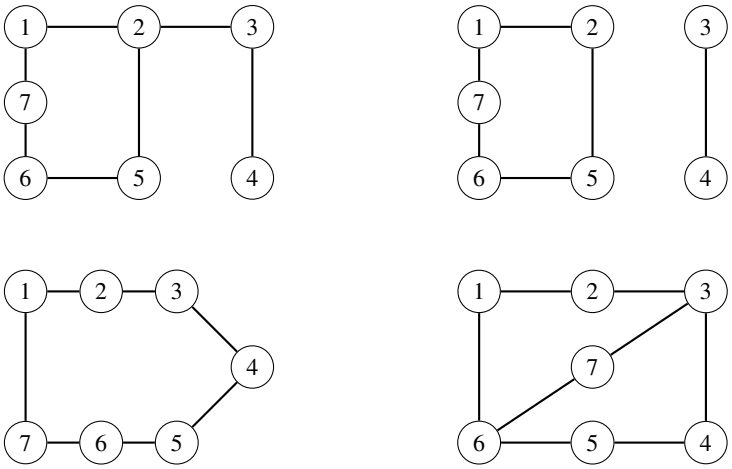

Fig. 8. Minimal non-IGs with $p = 7$.

5 Polarized Incompatibility Graphs in the plane

In this section, we introduce a subclass of graphs, called "polarized" IGs, whose definition is based on the notions of monotone and saddle pairs introduced in Section 4.

Definition 6. An incompatibility graph $G = (V,E)$, is *polarized* if it has an embedding where the edges are either all monotone or all saddle.

An example is given by the graph P_7 which is a polarized IG since the embedding provided in Figure 7 corresponds only to monotone edges.

Proposition 2. *Any polarized IG is a comparability graph.*

Proof. In any polarized IG the vertex adjacency relation is a partial order in P. As a matter of fact, for a given embedding of G, one has

$$(x,y) \in E \Leftrightarrow ([x,y] \cap N \neq \emptyset) \wedge ((y \in NE(x)) \vee (x \in NE(y)). \tag{5}$$

Since $((y \in NE(x)) \vee (x \in NE(y))$ is a partial order, the statement follows. □

As a consequence of Proposition 2, polarized IGs cannot have all induced subgraphs that are forbidden for comparability graphs. For example, *net graphs* and C_5 - for which an embedding as an IG exists - are forbidden for polarized IGs (see the famous Gallai's characterization theorem of comparability graphs [9]).

Another consequence of the above result is that the vertex coloring problem on a polarized IG can be solved in polynomial time. Actually, after Proposition 2, the following result holds.

Corollary 3. *Polarized IGs can be colored in polynomial time.*

Proof. All comparability graphs can be colored in polynomial time, see [10].

□

Theorem 8. *(NW Theorem)*
If $G = (V,E)$ is a polarized IG and i,p,q,r are four vertices of G for which the following conditions hold: (a) $\{(p,q),(q,r),(p,r) \in E\}$, and (b) $\{(i,p),(i,q),(i,r) \notin E\}$, see Figure 9-(a). Then in any embedding of G one must have, for some $h \in \{p,q,r\}$, either $h \in NW(i)$ or $i \in NW(h)$.

Proof. In view of the symmetry of relation (a), we may always assume that in the embedding of G, one of two conditions holds: (1) $q \in NE(p)$ and $r \notin NE(q)$, or (2) $q \notin NE(p)$ and $r \in NE(q)$. So p,q,r are embedded as in Figure 9-(b). Suppose that neither of the relations (c) holds. So i must necessarily belong to one of the four areas labeled A, B, C, D in Figure 9-(b). But then i must be adjacent to at least one of the vertices p,q,r, and this is a contradiction to (b). □

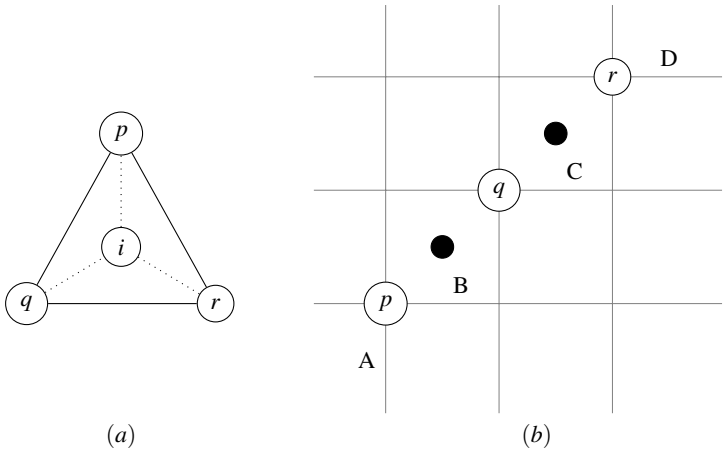

Fig. 9. NW Theorem.

After Proposition 2, we know that polarized IGs are comparability graphs. Figure 10 shows the relation between polarized IGs, IGs and the well known classes of graphs given by Comparability, Co-comparability and Permutation graphs.

In Table 2 we report the number of minimal non-IGs which are comparability graphs. A small number of such graphs could help in the recognition of polarized IGs. Actually, the table shows that the number of comparability minimal non-IGs is quite large already for small values of n, thus suggesting that it might not be an easy task to recognize polarized IGs by checking all such graphs one by one.

Table 2. Summary of minimal non-IGs.

n	comparability minimal non-IGs
$1-5$	0
6	1
7	0
8	10
9	159
10	865

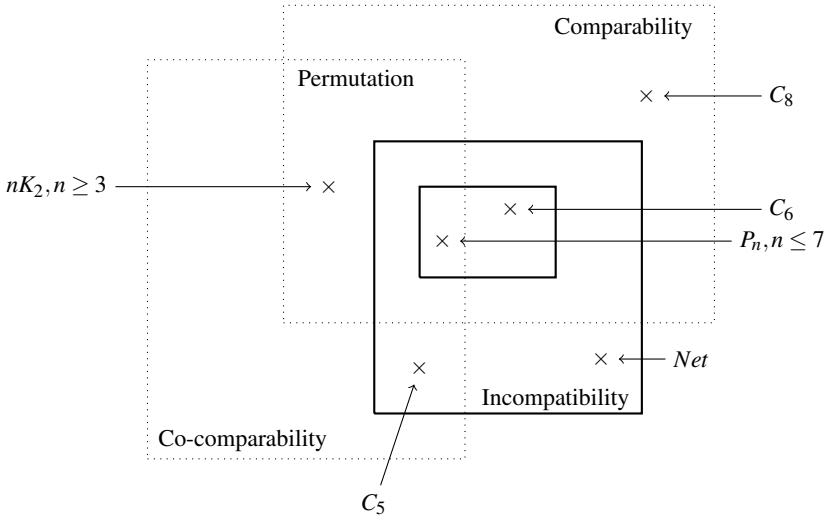

Fig. 10. Graph classes. The inner box represents polarized IGs.

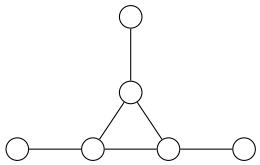

Fig. 11. A net graph.

6 Radius of 2-Incompatibility Graphs

In this section, we investigate the radius of IGs. The results related to the absence of long induced paths in an IG suggest that we can find small upper bounds on the radius of connected IGs.

Given a graph $G = (V,E)$ we denote by $d(x,y)$ the distance of vertices $x,y \in V$, that is the minimum number of edges on an x-y path in G. Let $r_G(x) = \max_{y \in V} d(x,y)$ and we define $r(G) = \min_{x \in V} r_G(x)$ as the *radius* of G. Note that $d(x,x) = 0$ by this definition, furthermore, if G is disconnected, then $r(G) = +\infty$; while in a connected graph G all $d(x,y)$ values are finite, and hence $r(G)$ is finite.

Our main result in this section is that 2-IG-s cannot have a large radius. In fact, we prove a more general result, which may be of some interest on its own.

For stating our results formally, we need a few more notations. For a vertex $x \in V$ and a subset $S \subseteq V$ of the vertices we denote by $d(x,S) = \min_{y \in S} d(x,y)$ the distance of

x from S. Given a graph H, we say that G is $H-free$, if it does not have an induced subgraph isomorphic to H.

Lemma 3. *Assume that $G=(V,E)$ is $3K_2$-free, and $X \subseteq V$ induces a subgraph containing an induced $2K_2$. Then, if $(u,v) \in E$ is an edge between vertices $u,v \in V \setminus X$ outside of X, we must have $d(u,X)+d(v,X) \leq 3$. In particular, the sets $S_1 = \{u \in V \mid d(u,X)=1\}$ and $S_2 = \{u \in V \mid d(u,X)=2\}$ cover all vertices in $V \setminus X$ and S_2 is stable.*

Proof. Let $e=(u,v)$ and assume indirectly that $d(u,X)+d(v,X) > 3$. Then, since $e \in E$, we must have both $d(u,X) \geq 2$ and $d(v,X) \geq 2$. Consequently, edge e with the two edges of an induced $2K_2$ within X would induce a $3K_2$, contradicting our assumption. □

Let us note next that the following claim follows by the definitions.

Lemma 4. *If $r(G) \geq 4$ for a connected graph $G=(V,E)$, then for every vertex $u \in V$ we have another vertex $v \in V$ such that $d(u,v)=4$.*

Proof. This is because a subpath of a shortest path is also a shortest path, thus if there is a vertex w from a finite distance $d \geq 4$ from u, then on a shortest path from u to w we must have a vertex v exactly at distance 4. □

The above simple claim implies that in a connected G with $r(G) \geq 4$ all vertices are endpoints of an induced P_5. We shall make repeated use of this basic observation in the sequel.

Theorem 9. *If a connected graph $G=(V,E)$ is $3K_2$-free and has $r(G) \geq 4$, then it either has a C_6 or a C_8 as an induced subgraph.*

Proof. Let us choose an arbitrary point $u=u_0$ in V and consider the induced P_5 starting at this vertex $A=\{u_0,u_1,u_2,u_3,u_4\}$. Then, by Lemma 3 we have all points of G within distance 2 from A, since it contains and induced $2K_2 = \{(u_0,u_1),(u_3,u_4)\}$. If all vertices of V are within distance 2 from the set $\{u_1,u_2,u_3\}$, then all are within distance 3 from u_2, contradicting our assumption that $r(G) \geq 4$. Thus, we must have $(u_5,u_6) \in E$ such that u_5 is connected only to the end point(s) of the path A, and u_6 is not connected to u_i, $i<5$. Without any loss of generality, we can assume that $(u_4,u_5) \in E$, and thus $B=\{u_2,u_3,u_4,u_5,u_6\}$ is an induced P_5. If $(u_0,u_5) \in E$, then G contains the $C_6 = \{u_0,u_1,u_2,u_3,u_4,u_5\}$, and we are done. Otherwise, we can assume that $C=A \cup B = \{u_0,u_1,u_2,u_3,u_4,u_5,u_6\}$ is an induced P_7. Let us now consider the point $v \in V$ at distance 4 from u_3. Since the set C contains an induced $2K_2$, Lemma 3 is applicable, and thus all points in V are within distance 2 from C. It follows that v must be at distance 2 from $\{u_1,u_5\}$. Without any loss of generality, we can assume that $d(u_5,v)=2$, and let $\{u_5,u_7,u_8=v\}$ denote a shortest path of length 2 between u_5 and $u_8=v$. Since $d(u_3,u_8)=4$ we cannot have any edge between u_7 and $\{u_2,u_3,u_4\}$. If $u_6=u_7$, the edges $T=\{(u_0,u_1),(u_3,u_4),(u_6,u_8)\}$ form either an induced $3K_2$, contradicting our assumptions (edge (u_0,u_8) does not exist), or $\{u_0,u_1,u_2,u_3,u_4,u_5,u_6,u_8\}$ forms an induced C_8 (edge (u_0,u_8) exists), as claimed.

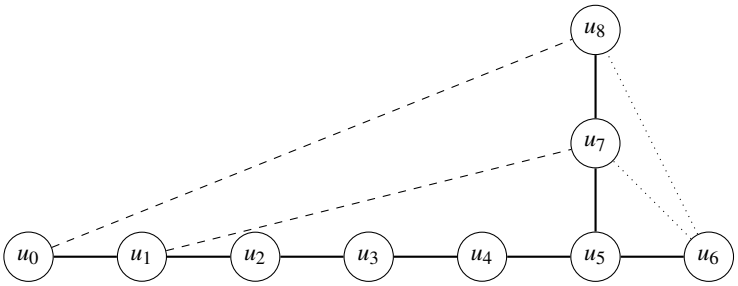

Fig. 12. Set D in proof of Theorem 9 solid lines show edges, dotted lines show possible edges, while dashed lines show possible edges some or all of which must be there. This subgraph is the result of the assumptions that G does not contain an induced $3K_2$, and has $d(u_0,u_4)=4$, $d(u_2,u_6)=4$ and $d(u_3,u_8)=4$.

Finally, if $u_7 \neq u_6$, in the set $D=\{u_0,u_1,u_2,u_3,u_4,u_5,u_7,u_8\}$ we can have only edges between the sets $\{u_0,u_1\}$ and $\{u_7,u_8\}$. Note that we cannot have $(u_0,u_7)\in E$, because of $d(u_0,u_4)=4$, and we cannot have $(u_1,u_8)\in E$ because of $d(u_3,u_8)=4$. On the other hand, we must have one or both of (u_0,u_8) and (u_1,u_7) as edges of G, since otherwise $\{(u_0,u_1),(u_3,u_4),(u_7,u_8)\}$ forms an induced $3K_2$.

Thus, in the set $D=\{u_0,u_1,u_2,u_3,u_4,u_5,u_7,u_8\}$ we have the edges as indicated on Figure 12 by solid lines, and potential edges as indicated by dotted or dashed lines, and in fact we must have one or both of the dashed $\{u_0,u_1\}$ and $\{u_7,u_8\}$, as edges in G. In either case, the set D includes an induced C_6 or an induced C_8, as claimed. □

Theorem 10. *If $G=(V,E)$ is a connected $\{3K_2,C_7\}$-free graph that has an induced C_6, then it has $r(G)\leq 3$.*

Proof. Let $C=\{u_0,u_1,u_2,u_3,u_4,u_5\}$ be an induced C_6 in G, as shown in Figure 13, and assume indirectly that $r(G)\geq 4$. Since C contains induced $2K_2$ subgraphs, Lemma 3 implies that all vertices of G are within distance 2 from C, moreover, since there are 3 induced $2K_2$ graphs in C, vertices at distance 2 from C must be of distance 2 from at least 2 different, non-opposite vertices of C.

Let us then consider an induced P_5 starting at vertex $u_i\in C$, which exist by Lemma 4, and denote by (v_i,w_i) its last edge, i.e., $d(u_i,w_i)=4$, for $i=0,1,...,6$.
It is easy to see that essentially there are only two possibilities to realize $d(u_i,w_i)=4$. Look at Figure 13 to illustrate these cases for $i=0$. Either we have $u_3=v_0$ and then there are no other edges in the subgraph induced by $C\cup\{w_0\}$ (Type I configuration), or $d(v_0,\{u_2,u_3,u_4\})=1$ (Type II configuration), and the subgraph induced by $C\cup\{v_0,w_0\}$ has only the solid edges shown in Figure 13 plus at least two out of the three dotted edges, that is, at least two out of (u_2,v_0), (u_3,v_0) and (u_4,v_0) must be edges of G.

We shall now consider all possibilities for the type of configuration that can occur for vertices u_0, u_2 and u_4. Due to circular symmetry, it is enough to consider the following four cases.

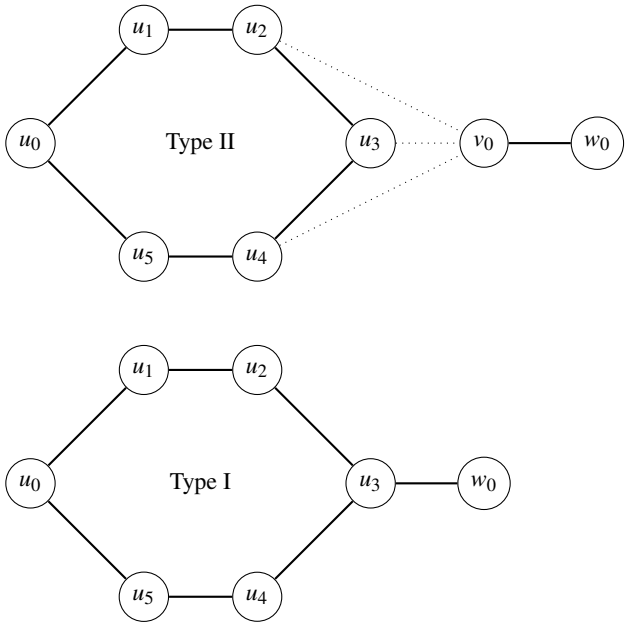

Fig. 13. The two possible configuration to realize $d(u_0, w_0) = 4$. In the case of the top, type II figure we must have at least two out of the three dotted lines as edges of the graph. There are no other edges, than indicated, in these induced subgraphs.

1. All three of $\{u_0, u_2, u_4\}$ are of type I (see Figure 14).

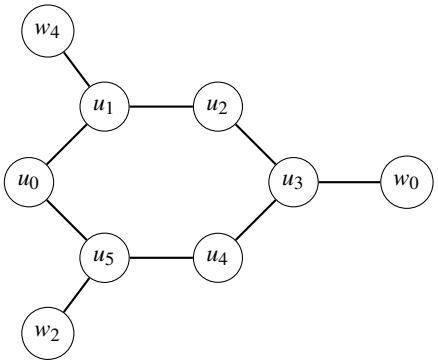

Fig. 14. Case: u_0, u_2 and u_4 are all of type I.

In this case none of $\{w_0, w_2, w_4\}$ are connected to any other (than the ones indicated in Figure 14) vertices of C. Furthermore, w_i and w_j for $i \neq j$ cannot be

connected due to the $d(u_i, w_i) = 4$, $i = 0, 2, 4$ conditions. For instance, an edge between w_0 and w_2 would imply that $d(u_0, w_0) = 3$, contradicting our assumption that it is 4. Thus, the graph on Figure 14 is an induced subgraph of G. Since it contains the $3K_2 = \{(u_3, w_0), (u_1, w_4), (u_5, w_2)\}$, this contradicts our assumptions about G.

2. Vertex u_0 is of type II, while vertices u_2 and u_4 are of type I (see Figure 15).

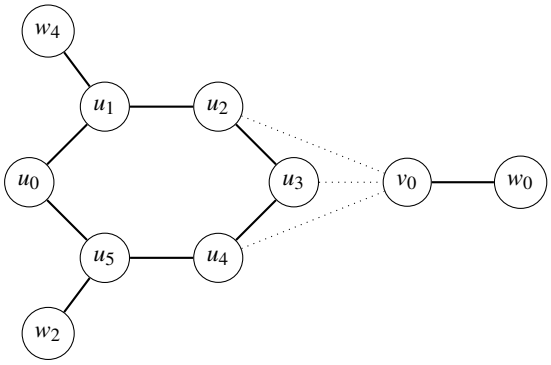

Fig. 15. Case: u_0 is of type II, while u_2 and u_4 are of type I.

In this case the only possible additional edges in the subgraph of G induced by the vertices in Figure 15 could be edges between the sets $\{w_2, w_4\}$ and $\{v_0, w_0\}$. Since at least two out of the three dotted edges must be in $E(G)$, we have $d(u_2, v_0) \leq 2$ and $d(u_4, v_0) \leq 2$. Consequently we cannot have edges (w_2, v_0) or (w_4, v_0) in G because of $d(u_2, w_2) = d(u_4, w_4) = 4$. Furthermore, we also have $d(u_0, w_2) = d(u_0, w_4) = 2$, and thus any edge between w_0 and $\{w_2, w_4\}$ would make $d(u_0, w_0) < 4$. Consequently, Figure 15 represents an induced subgraph of G, implying that it must contain the induced $3K_2 = \{(u_1, w_4), (u_5, w_2), (v_0, w_0)\}$, a contradiction with our assumptions on G.

3. Vertices u_0 and u_2 are of type II, while vertex u_4 is of type I (see Figure 16).
In this case we should focus on the fact that G does not contain an induced $3K_2$. Thus, we must have some edges between the sets w_4, and the sets $\{v_2, w_2\}$ and $\{v_0, w_0\}$. Similarly to the previous case, we can argue that w_4 cannot be connected to any of the other four vertices. Thus we must have an edge, since G is $3K_2$-free, between $\{v_2, w_2\}$ and $\{v_0, w_0\}$. It is easy to see that we cannot have an edge between v_i and w_j for $\{i, j\} = \{0, 2\}$. For instance, the existence of the edge (v_0, w_2) would imply $d(u_2, w_2) \leq 3$ contradicting our assumption that $d(u_2, w_2) = 4$. Finally, vertices w_0 and w_2 are of distance 2 from C and hence by Lemma 3, there is no edge between them. Thus, the only possibility is that $(v_0, v_2) \in E(G)$. Therefore, $d(u_0, w_0) = 4$ implies that (u_0, v_2) is not an edge of

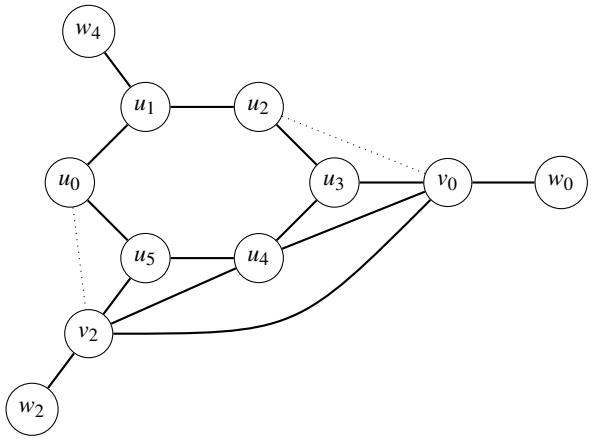

Fig. 16. Case: u_0 and u_2 are of type II, while u_4 is of type I.

G. Similarly, $d(u_2, w_2) = 4$ implies that (u_2, v_0) is not en edge of G. Thus, the graph shown in Figure 16 is an induced subgraph of G, in which the solid lines represent the edges. Hence, the set $Q = \{u_0, u_1, u_2, u_3, v_0, v_2, u_5\}$ is an induced C_7 in G, contradicting our assumptions.

4. All the vertices u_0, u_2, and u_4 are of type II. In this case we can repeat the arguments of the previous case and conclude that at least one of the (v_i, v_j), $i \neq j$ pairs must be an edge in G. For example, if $(v_0, v_2) \in E$ we can again conclude that the set Q is an induced C_7 in G, contradicting our assumptions.

Since in all cases we obtained a contradiction, our indirect assumption about the radius must be wrong, proving the statement. □

Corollary 4. *If $G = (V, E)$ is a connected $\{3K_2, C_7, C_8\}$-free graph then it has $r(G) \leq 3$.*

Proof. Assume indirectly that $r(G) \geq 4$. Then, by Theorem 9 G must have an induced C_6, which by Theorem 10 implies $r(G) \leq 3$. This contradiction proves our claim. □

Corollary 5. *A connected 2-IG has radius at most 3.*

Proof. By Theorems 4, 5 and 6, a 2-IG is $\{3K_2, C_7, C_8\}$-free, thus the claim follows by Corollary 4. □

7 Box Systems, Vertex Packing and Vertex Coloring

In this section, we analyze the relations between homogeneous boxes in a d-dimensional space and the *vertex packings* (or *stable sets*) of an incompatibility graph

G, that is, subsets of pairwise non-adjacent vertices of G. A *vertex coloring* of a graph G is defined as a partition of its vertices into vertex packings (each vertex packing of the partition corresponds to a different color). Two well known combinatorial problems on G are the following: i) find a maximum (cardinality) vertex packing in G; ii) find a minimum vertex coloring, that is, a partition of the vertex-set of G into the minimum possible number of vertex packings (colors).

The following Carathéodory-type theorem (proved in [5]), and the implied corollaries, provide results that can be exploited to derive equivalences in $\mathbb{R}^2$ between the solutions of the *BC* problems formulated in Section 2 (i.e., Maximum Box and Minimum Covering by Boxes) and the solutions of the maximum vertex packing and minimum vertex coloring problems, respectively, on the corresponding $G_{P,N}$. Unfortunately, in $\mathbb{R}^d$ these equivalences do not hold, but we can still establish relations between the two pairs of problems.

Theorem 11. *Let $X \subseteq \mathbb{R}^d$ be a finite set of cardinality $m \geq d > 1$. Then for any point $x \in [X]$, there exists a subset $Y \subset X$ of size at most d such that $x \in [Y]$.*

Then, two fundamental corollaries follows.

Corollary 6. *Let $X \subseteq \mathbb{R}^d$ be a finite set of cardinality $m \geq d > 1$, then*

$$[X] = \bigcup_{\substack{Y \subseteq X \\ |Y| = d}} [Y]$$

The above result implies in particular that for $d = 2$ and a finite set $\emptyset \neq S \subseteq \mathbb{R}^d$ one has:

$$d = 2 \Rightarrow [S] = \bigcup_{s_1, s_2 \in S} [s_1, s_2] \tag{6}$$

Let us note that in $d > 2$ dimension the box hulls of pairs of points may not cover the box hull of S, that is,

$$d > 2 \Rightarrow [S] \supseteq \bigcup_{s_1, s_2 \in S} [s_1, s_2]$$

Example 1. Let us consider $S = \{x_1, x_2, x_3\} \subset \mathbb{R}^3$, where $x_1 = (0,1,1)$, $x_2 = (1, \frac{1}{2}, \frac{1}{2})$, and $x_3 = (\frac{1}{2}, 0, 0)$. One can check that

$$[S] = [\{x_1, x_2, x_3\}] \neq [\{x_1, x_2\}] \cup [\{x_1, x_3\}] \cup [\{x_2, x_3\}].$$

We are now able to state the following results for the 2-dimensional case.

Theorem 12. *Let $S = P \cup N \subset \mathbb{R}^2$. A box is homogeneous iff its points correspond to a stable set of $G_{P,N}$.*

Proof. The proof follows from the identity 6. □

Remark. In $\mathbb{R}^2$, a maximum box on S defines a maximum stable set in $G_{P,N}$, and vice versa. Thus, the maximum box problem on S is equivalent to the maximum stable set problem in $G_{P,N}$. In other words, the number of points in a maximum box is the *stability number* of $G_{P,N}$ (i.e., the cardinality of a maximum stable set in $G_{P,N}$).

Corollary 7. *In $\mathbb{R}^2$, the minimum covering by boxes problem on S is equivalent to the minimum vertex coloring in $G_{P,N}$. In other words, the number of boxes in a minimum covering by boxes is equal to the chromatic number of $G_{P,N}$ (i.e., the minimum number of vertex packings, or colors, necessary to cover the vertices of $G_{P,N}$).*

Proof. Given a minimum covering by boxes in $\mathbb{R}^2$, a coloring of the corresponding $G_{P,N}$ can be obtained by assigning to each point the color of the box it belongs to. When a point is contained in more than one box, one can choose arbitrarily one of the colors of such boxes. Notice that in this process each color must be assigned to at least one vertex, else the covering would not be minimum. On the other hand, given a coloring of $G_{P,N}$, one obtains a box system by considering the box closures of the subsets of vertices with the same color. It must be noticed that in this way one may get only a *covering* of the vertices of $G_{P,N}$ by vertex packings, since box closures of disjoint sets of colored vertices of $G_{P,N}$ may have some points of P in common, see, Figure 17. Since the number of boxes matches the number of colors both ways, the statement follows. □

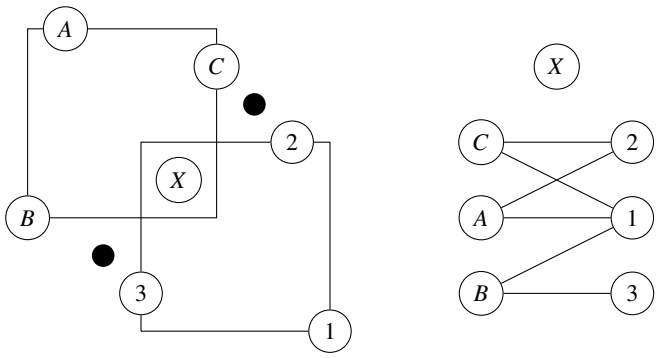

Fig. 17. Left: the groups of points A, B, C and 1,2,3 form two boxes in the box system. Right: the groups of vertices A,B,C,X and 1,2,3,X form two vertex packings in $G_{P,N}$. Point X belongs to both boxes, and it is covered by both vertex packings.

The nice result is that in $\mathbb{R}^2$ the vertex packing problem on $G_{P,N}$ can be solved in polynomial time; the same is true for the vertex coloring problem (at least) when $G_{P,N}$ is non trivially disconnected.

Theorem 13. *If G is a 2-IG, a maximum stable set in G can be found in polynomial time.*

Proof. The proof follows from a result by Balas and Yu [1] which bounds the number of maximal stable sets in a graph without large induced matchings and from another result by Johnson and Yannakakis [15] stating that all stable sets of a graph can be generated in time linear in the number of its maximal stable sets. □

Now we focus our attention to colorings of 2-IGs.

Lemma 5. *Every IG embedding in $\mathbb{R}^2$ of an induced C_5 must admit both a monotone edge and a saddle one.*

Proof. Suppose, to the contrary, that all edges are monotone, say. Since C_5 is an odd cycle, there must exist along it a pair of incident edges (u,v) and (v,w) such that $v \in NE(u)$ and $w \in NE(v)$. Then there would exist also an edge (u,w) by the transitivity of the NE relation and the fact that the rectangle $[\{u,w\}]$ contains at least one negative point. But the existence of (u,w) contradicts the assumption that C_5 is induced. A similar proof holds when all the edges of C_5 are saddle. □

We shall say that a connected component of a graph is *nontrivial* if it contains at least one edge. A graph is *nontrivially disconnected* if it has at least two nontrivial connected components.

Theorem 14. *If an IG in $\mathbb{R}^2$ is nontrivially disconnected, then it has exactly two non trivial connected components and they are both weakly chordal.*

Proof. The absence of induced $3K_2$s implies that an IG has at most two nontrivial components. Thus, we have exactly two components and they must be both IGs for the hereditary property. Being IGs, they cannot have induced C_k, $k \geq 7$. Furthermore, neither of them may have an induced $2K_2$, since, otherwise, a $3K_2$ would be implied and this means that induced C_6 are forbidden. Finally, they cannot have an induced C_5, since $C_5 \cup K_2$ is forbidden in an IG for the following reason: by Lemma 5 there would exist in C_5 an edge e such that e and $f = K_2$ are not coherent; but then there would be a vertex of C_5 that is not adjacent to e, and obviously not adjacent to f, contradicting Lemma 1. □

Corollary 8. *If $G_{P,N}$ in $\mathbb{R}^2$ has at least two, and hence exactly two, nontrivial connected components, then it can be recognized and colored in polynomial time.*

Proof. Follows from Theorem 14 and from the well-known result that weakly chordal graphs can be recognized and colored in polynomial time, see [14]. □

Although not as strong as in the 2-dimensional case, some results can be stated also for the d-dimensional case. In $\mathbb{R}^d$, it is still true (by construction) that the points of a homogeneous box correspond to a vertex packing in the related IG, but the reverse is not true. In fact, a vertex packing of the given $G_{P,N}$ may not result in a homogeneous box, see Figure 17. However, the stability number of $G_{P,N}$ is always an *upper bound* on the cardinality of a maximum homogeneous box. On the other hand,

for the same reason, the chromatic number of $G_{P,N}$ is a *lower bound* on the cardinality of a minimum covering by boxes of S. These two properties might be exploited in the design of heuristic or exact algorithms for the two above *BC* problems in an arbitrary dimension.

8 Conclusions

In this paper, we have introduced and investigated a class of graphs, dubbed Incompatibility Graphs, arising in the context of *Box Clustering*.

Two fundamental problems *Box Clustering* is faced with are: (i) find a positive box containing the largest number of points (*maximum box* problem); (ii) find the smallest number of positive boxes covering all the positive points in the data set (*minimum covering by boxes* problem). Incompatibility graphs are of help in the solution of both problems in general, but their use becomes particularly attractive in the two-dimensional case, for two main reasons:

- the maximum box problem and the minimum covering by boxes problem can be formulated as a vertex packing and a vertex coloring in the corresponding 2-*incompatibility graph*, respectively.
- 2-*incompatibility graph*s, unlike general d-incompatibility ones, have a very special structure, which allows one to solve the vertex packing problem - and, in two significant subclasses, also the vertex coloring one - in polynomial time.

The core of the paper is devoted to the analysis of such structure. (2-)incompatibility graphs are shown to have no large induced matchings, and therefore no large induced paths or cycles. Actually, they cannot have any cycle of length greater than 6. Hence they may be viewed as a generalization of chordal and weakly chordal graphs. Besides, they have many other forbidden subgraphs. Dwelling on the fact that the recognition problem for this class of graphs (whose complexity is open) can be formulated as a polynomial-sized mixed integer linear program, we have generated, with the help of an appropriate computer program, the $105,329$ forbidden subgraphs with at most 10 vertices. The number and the variety of these graphs makes us believe that the recognition of an incompatibility graph by forbidden subgraphs might be a quite challenging task. Perhaps the property that, when the graph is connected, its radius is small might be of help here.

On the other hand, we have shown that, when an incompatibility graph is non trivially disconnected, then it admits exactly two nontrivial connected components, and they are both weakly chordal. Thus, in this case, the vertex coloring problem can be solved in polynomial time. This nice feature is shared by those IGs that admit an embedding where all edges have a SW-NE direction (the complexity is still open in the remaining cases). Instead, the vertex packing problem in an IG is *always* solvable in polynomial time on the grounds of the results provided in [1, 15].

Furthermore, we have pointed out that the notion of incompatibility graph - at first sight, a geometrical one - is purely order-theoretical. This result may be the starting point of further fruitful investigations.

Appendix

A - Windrose *MILP* Model for the Recognition of IGs

The input graph is $G = (P,E)$, with $n = |P|$ nodes and $m = |E|$ edges. W.l.o.g., we may assume that $|N| = m$. In the sequel, the unordered pair ij is identified with the ordered pair (i, j), where $i < j$. Furthermore we can always consider

$$P = \{1,\ldots,n\}$$

$$N = \{n+1,\ldots,n+m\}.$$

The binary variables of the model are defined as: for every $i, j \in P \cup N$

$$\mathrm{NE}_{ij} = 1 \Leftrightarrow \textit{ j is located NE of i}$$

$$\mathrm{NW}_{ij} = 1 \Leftrightarrow \textit{ j is located NW of i}$$

The nonnegative real variables are u^h_{ij} and v^h_{ij} where $(i, j) \in E$ and $h \in N$. In any optimal solution of the *MILP*, u^h_{ij} and v^h_{ij} will take only values 0 or 1.
Let us now consider the constraints:

- *The relations NE and NW are transitive* - for every $i, j, k \in P \cup N$, and $i \neq j \neq k$

$$\mathrm{NE}_{ij} + \mathrm{NE}_{jk} - \mathrm{NE}_{ik} \leq 1$$
$$\mathrm{NW}_{ij} + \mathrm{NW}_{jk} - \mathrm{NW}_{ik} \leq 1$$

- *The points are in general position* - for every $i \in P$, $j \in P \cup N$, and $i \neq j$

$$\mathrm{NE}_{ij} + \mathrm{NW}_{ij} + \mathrm{NE}_{ji} + \mathrm{NW}_{ji} = 1.$$

  This means that j must be in one of the four orthants w.r.t. i.
- *Constraints related to edges* - for every $(i, j) \in E$, $h \in N$

$$u^h_{ij} \leq \mathrm{NE}_{ih} + \mathrm{NE}_{ji}$$
$$u^h_{ij} \leq \mathrm{NE}_{hj} + \mathrm{NE}_{ji}$$
$$u^h_{ij} \leq \mathrm{NE}_{jh} + \mathrm{NE}_{ij}$$
$$u^h_{ij} \leq \mathrm{NE}_{hi} + \mathrm{NE}_{ij}$$

$$v^h_{ij} \leq \mathrm{NW}_{ih} + \mathrm{NW}_{ji}$$
$$v^h_{ij} \leq \mathrm{NW}_{hj} + \mathrm{NW}_{ji}$$
$$v^h_{ij} \leq \mathrm{NW}_{jh} + \mathrm{NW}_{ij}$$
$$v^h_{ij} \leq \mathrm{NW}_{hi} + \mathrm{NW}_{ij}$$

$$\sum_{h \in N} (u^h_{ij} + v^h_{ij}) \geq 1$$

  In any optimal solution at least one between u^h_{ij} and v^h_{ij} must take the value 1, but not both of them, and thus a negative point h is the box $]i, j[$. This means that the last constraint is always satisfied at the optimal value, and, therefore, it can be ignored.

- *Constraints related to non-edges* - for every $(i,j) \notin E$, $i \neq j$, $h \in N$

$$\begin{aligned} &\mathrm{NE}_{ij} + \mathrm{NE}_{ji} \leq 1 \\ &\mathrm{NE}_{ih} + \mathrm{NE}_{hj} + \mathrm{NE}_{ij} \leq 2 \\ &\mathrm{NE}_{jh} + \mathrm{NE}_{hi} + \mathrm{NE}_{ji} \leq 2 \end{aligned}$$

$$\begin{aligned} &\mathrm{NW}_{ij} + \mathrm{NW}_{ji} \leq 1 \\ &\mathrm{NW}_{ih} + \mathrm{NW}_{hj} + \mathrm{NW}_{ij} \leq 2 \\ &\mathrm{NW}_{jh} + \mathrm{NW}_{hi} + \mathrm{NW}_{ji} \leq 2 \end{aligned}$$

Let us now consider the objective function:

$$\max \sum_{(i,j)\in E} \sum_{h\in N} (u_{ij}^h + v_{ij}^h)$$

The objective function is chosen so that in any optimal solution one has: $u_{ij}^h = 1$ or $v_{ij}^h = 1$.

In the above *MILP* model (a) the number of variables is $O\left(n^2 + m^2\right)$, and (b) the number of constraints is $O\left((n+m)^3\right)$. The validity of such model rests upon Theorem 3 and the following result.

Proposition 3. *The two posets on P defined by the relations NE_{ij} and NW_{ij} have dimension* 2*.*

Proof. For convenience, define for all the pairs $i, j \in P$ the new variables SW and SE as follows

$$\begin{aligned} \mathrm{SW}_{ij} &= \mathrm{NE}_{ji} \\ \mathrm{SE}_{ij} &= \mathrm{NW}_{ji} \end{aligned}$$

and the generic position constraint for every $i, j \in P$ transforms into

$$\mathrm{NE}_{ij} + \mathrm{NW}_{ij} + \mathrm{SE}_{ij} + \mathrm{SW}_{ij} = 1$$

Let us define for every $i, j \in P$

$$\begin{aligned} i \prec_1 j &\Leftrightarrow (\mathrm{NE}_{ij} = 1) \vee (\mathrm{SE}_{ij} = 1) \\ i \prec_2 j &\Leftrightarrow (\mathrm{NW}_{ij} = 1) \vee (\mathrm{SW}_{ij} = 1) \end{aligned}$$

Then, one has that completeness holds for $\prec_1$ and $\prec_2$:

$$\begin{aligned} (i \prec_1 j) &\Leftrightarrow \neg(j \prec_1 i) \\ (i \prec_2 j) &\Leftrightarrow \neg(j \prec_2 i) \end{aligned}$$

Furthermore, from the transitivity constraints we obtain

$$\max\{\mathrm{NE}_{ij} + \mathrm{NE}_{jk} - \mathrm{NE}_{ik}, \mathrm{NE}_{ji} + \mathrm{NE}_{kj} - \mathrm{NE}_{ki}\} \leq 1$$

and taking into account the definition of SE and the generic position, one has

$$\max\{\mathrm{NE}_{ij}+\mathrm{NE}_{jk}-\mathrm{NE}_{ik},\mathrm{SE}_{ij}+\mathrm{SE}_{jk}-\mathrm{SE}_{ik}\}\leq 1$$

and this proves the transitivity of $\prec_1$; similarly one proves the transitivity of $\prec_2$. In conclusion, $\prec_1$ and $\prec_2$ are linear orders, hence, the NE and the NW relations have dimension 2. □

References

1. E. BALAS, C.S. YU. On graphs with polynomially solvable maximum-weight clique problem, *Networks*, 19: 247-253, 1989.
2. T.BONATES, P.L. HAMMER. Logical Analysis of Data: from combinatorial optimization to medical applications, *Annals of Operations Research*, 148: 203-225, 2006.
3. E. BOROS, T. IBARAKI, K. MAKINO. Boolean Analysis of Incomplete Examples, *Rutgers Research Report* 7-1996.
4. E. BOROS, P.L. HAMMER, T. IBARAKI, A. KOGAN, E. MAYORAZ, AND I. MUCHNIK. An implementation of logical analysis of data, *IEEE Transactions on Knowledge and Data Engineering*, 12: 292-306, 2000.
5. E. BOROS, T. HORIYAMA, T. IBARAKI, K. MAKINO, M. YAGIURA. Finding Essential Attributes from Binary Data, *Annals of Mathematics and Artificial Intelligence*, 39: 223-257, 2003.
6. A. BRANDSTDT, V.B. LE, J.P. SPINRAD. *Graph Classes: A Survey*. SIAM Monographs on Discrete Mathematics and Applications, ISBN 978-0-89871-432-6, 1999.
7. Y. CRAMA, P.L. HAMMER, T. IBARAKI. Cause-effect relationship and partially defined Boolean functions, *Annals of Operational Research*, 16: 299-325, 1988.
8. J. ECKSTEIN, P.L. HAMMER, Y. LIU, M. NEDJAK, B. SIMEONE. Computational Optimization and its Applications, *Computational Optimization and its Applications*, 23: 285-298, 2002.
9. T. GALLAI. Transitiv Orientierbare Graphen, *Acta Mathematica Academiae Scientiarum Hungaricae*, 18: 25-66, 1967.
10. M.C. GOLUMBIC. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York, 1980.
11. P.L. HAMMER. Partially defined Boolean functions and cause-effect relationship, *Lecture at the International Conference on Multi-Attribute Decision Making Via Or-Based Expert Systems*, University of Passau Germany, April 1986.
12. P.L. HAMMER, Y. LIU, S. SZEDMÁK, B. SIMEONE. Saturated systems of homogeneous boxes and the logical analysis of numerical data, *Discrete Applied Mathematics*, 144: 103-109, 2004.
13. F. HARARY. *Graph Theory*, AddisonWesley, Reading 1969.
14. R.B. HAYWARD, J.P. SPINRAD, R. SRITHARAN. *Improved algorithms for weakly chordal graphs*. ACM Trans. Algorithms, Volume 3, Number 2, ISSN 1549-6325. ACM, New York, NY, USA. 2007.
15. D.S. JOHNSON, M. YANNAKAKIS. *On generating all maximal independent sets*. Information Processing Letters, 27: 119-123, 1988.
16. A. KANEKO, M. KANO. Discrete Geometry on Red and Blue Points in the Plane - A Survey, in *Discrete and Computational Geometry*, The Goodman-Pollack Festschrift, Springer, 2003 (pp. 551-570).
17. B.D. MCKAY. Practical graph isomorphism. *Congressus Numeratium*, 30: 45-87, 1981.
18. T. MITCHELL. *Machine learning*. Mc Grow-Hill, New York, 1997.

19. B. SIMEONE, M. MARAVALLE, F. RICCA, V. SPINELLI. Logic Mining of non-logic data: some extensions of box clustering, 21*st European Conference on Operational Research (EURO XXI)*, Reykjavik, Iceland, July 2006.
20. B. SIMEONE, V. SPINELLI. The optimization problem framework for box clustering approach in logic mining, 22*nd European Conference on Operational Research (EURO XXII)*, Prague, July 2007.
21. P. VAN'T HOF AND D. PAULUSMA. A new characterization of P_6-free graphs, *Technical Report*, Department of Computer Science, Durham University, 2008.
22. B.D. MCKAY. *nauty User's Guide (Version* 2.4*)*. Department of Computer Science. Australian National University Canberra ACT 0200, Australia.