Distances of Centroid Sets in a Graph-Based Construction for Information Security Applications

J. Abawajy, A.V. Kelarev, M. Miller and J. Ryan

Abstract. The aim of this paper is to prove that, for every balanced digraph, in every incidence semiring over a semifield, each centroid set J of the largest distance also has the largest weight, and the distance of J is equal to its weight. This result is surprising and unexpected, because examples show that distances of arbitrary centroid sets in incidence semirings may be strictly less than their weights. The investigation of the distances of centroid sets in incidence semirings of digraphs has been motivated by the information security applications of centroid sets.

Mathematics Subject Classification (2010). Primary 68T05; Secondary 68T10.

Keywords. Centroid sets, directed graphs, distances.

1. Introduction

The investigation of the centroid sets in ring and semiring constructions based on digraphs has been motivated by the information security applications of the centroid sets (cf., [1, 15, 17, 35]). These applications rely on the sets of centroids with large distances (cf. [7, 27, 28, 30]). It is well known and easy to verify that in every ring the distance of each centroid set is equal to its weight (cf. [29, 22]). The notion of a weight is conceptually simpler than the distance, and so all previous studies in the literature have so far only dealt with the weights of the centroid sets. Likewise, the investigation of centroid sets in more general semiring constructions has initially focused on the weights. However, in semiring constructions the value of a distance may be strictly less than the weight and it is precisely the distance of a centroid set that is crucial for applications (see Section 3 for more details). Therefore it is important to consider the distances of centroid sets and compare them to the weights of the sets.

This is the first article devoted to the distances of the centroid sets in semiring constructions. Incidence semirings of directed graphs are well known and have good relations to other constructions. They have been considered previously with respect to other problems, for example, in [5, 6, 18].

We tackle the distances of the centroid sets in the incidence semirings of digraphs. Our main theorem establishes that in every incidence semiring of a balanced digraph over a semifield, each centroid set J of the largest distance also has the largest weight at the same time, and the distance of J is equal to its weight (see Theorem 4.1 in Section 4). This result is surprising and unexpected, because in general the distances of centroid sets in incidence semirings may be strictly less than their weights (see Example 4).

2. Preliminaries

We use standard notions and terminology and refer the readers to the books [19, 22, 23, 32] and papers [9, 16, 27] for more detailed discussions. This section contains concise prerequisites required for the main theorem and its proof. Here the word 'digraph' means a finite directed graph without multiple parallel edges but possibly with loops. Let us also refer to the recent survey [2] and articles [20, 21], where the readers can find additional background information. Throughout this paper, the set of all positive integers is denoted by $\mathbb{N}$, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. As explained in [5], in the investigation of incidence semirings of digraphs it is convenient to include semirings without identity elements into consideration. This means that we use the following definition of a semiring.

Definition 2.1. A *semiring* is a set R with two binary operations, addition $+$ and multiplication $\cdot$, such that the following conditions hold:

(S1) $(R, +)$ is a commutative semigroup with zero 0,
(S2) $(R, \cdot)$ is a semigroup,
(S3) multiplication distributes over addition,
(S4) zero 0 annihilates R, i.e., $0 \cdot R = R \cdot 0 = 0$.

A *semifield* is a semiring F such that the set of all nonzero elements of F forms a group with respect to multiplication. An example of a semifield is the *Boolean semiring*, which is the set $\mathbb{B} = \{0, 1\}$ with operations $+ = \max$ and $\cdot = \min$. The identity element of a semifield F is denoted by 1 or 1_F. The notion of an incidence semiring is a generalization of an incidence ring (cf. [22, §3.15], [26] and [32]).

Definition 2.2. Let $D = (V, E)$ be a digraph, and let R be a semiring. The *incidence semiring* of D over R is the set consisting of zero 0 and all finite sums $\sum_{i=1}^{n} r_i(g_i, h_i)$, where $n \geq 1$, $r_i \in R$, $(g_i, h_i) \in E$, equipped with the standard addition and multiplication defined by the distributive law and the rule

$$(g_1, h_1) \cdot (g_2, h_2) = \begin{cases} (g_1, h_2) & \text{if } h_1 = g_2 \text{ and } (g_1, h_2) \in E, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

for all $(g_1, h_1), (g_2, h_2) \in E$. Empty sums are assumed to be equal to zero. The incidence semiring of D over R is denoted by $I_D(R)$.

Every element x in $I_D(R)$ has a unique representation as a sum $x = \sum_{e \in E} x_e e$, where $x_e \in R$ for all $e \in E$, and where only a finite number of the coefficients x_e are nonzero. If e is an edge in E and 1 is the identity element of R, then the element $1e$ of $I_D(R)$ can be also written down using the notation $e = 1e$. A digraph $D = (V, E)$ is said to be *balanced* if, for all $g_1, g_2, g_3, g_4 \in V$ with (g_1, g_2), (g_2, g_3), (g_3, g_4), $(g_1, g_4) \in E$, the following equivalence holds:

$$(g_1, g_3) \in E \Leftrightarrow (g_2, g_4) \in E,$$

see [22, §3.15].

Definition 2.3. For each balanced digraph D and every semiring R, any set of elements $g_1, \ldots, g_k \in I_D(R)^1 = I_D(R) \cup \{1\}$ generates the *centroid set* $C(g_1, \ldots, g_k)$ consisting of all sums of these elements and their multiples, i.e.,

$$C(g_1, \ldots, g_k) = \\ = \left\{ \sum_{j=1}^{m_1} \ell_{1,j} g_1 r_{1,j} + \cdots + \sum_{j=1}^{m_k} \ell_{k,j} g_k r_{k,j} \;\middle|\; \begin{array}{l} \ell_{i,j}, r_{i,j} \in I_D(R)^1 \\ m_1, \ldots, m_k \in \mathbb{N} \end{array} \right\}. \tag{2.2}$$

For a pair of elements $x, y \in I_D(R)$, where $x = \sum_{e\in E} x_e e$ and $y = \sum_{e\in E} y_e e$, the distance from x to y is denoted by $\mathrm{d}(x, y)$ and is defined as the number of edges e in E such that $x_e \neq y_e$. The *distance* of a subset S of $I_D(R)$ is the minimum nonzero distance $\mathrm{d}(x, y)$ between elements $x, y \in S$. The *weight* of an element $x = \sum_{e\in E} x_e e \in I_D(R)$ is denoted by $\mathrm{wt}(x)$ and is defined as the number of nonzero coefficients x_e in the sum. The *weight* of a subset S of $I_D(R)$ is the minimum weight of a nonzero element in S. Let us refer to [22, 27, 29] for more details and the definitions of other standard terms not defined in the paper.

Our examples use the concept of a semigroup semiring (cf. [22, §3.2] and [23, Chapter 10]).

Definition 2.4. Let R be a semiring, and let S be a semigroup. The *semigroup semiring* is the set

$$R[S] = \left\{ \sum_{i=1}^{n} r_i s_i \;\middle|\; n \in \mathbb{N}, r_i \in R, s_i \in S \right\} \tag{2.3}$$

equipped with addition and multiplication defined by the associative and distributive laws and the rules $r_1s_1 + r_2s_1 = (r_1 + r_2)s_1$ and $(r_1s_1)(r_2s_2) = (r_1r_2)(s_1s_2)$, for all $r_1, r_2 \in R$, $s_1, s_2 \in S$.

An element x of a semigroup is called an *idempotent* if $x = x^2$. A commutative semigroup entirely consisting of idempotents is called a *semilattice*. Every semilattice is a partially ordered set with respect to the natural order defined by $x \leq y \Leftrightarrow xy = x$. This makes it possible to represent and define semilattices using diagrams.

3. Motivation

Efficient classifiers play crucial roles in information security (cf. [1, 15, 24, 30, 35]), as well as other application areas (cf. [4, 3, 25, 33]). Ring and semiring constructions can be used in order to generate convenient sets of centroids for centroid-based classifiers and to design combined multiclass classifiers capable of correcting the errors of individual initial classifiers. Accordingly, centroid sets in semiring constructions defined by graphs can be applied in the following two ways.

First, centroid sets are valuable for the design of centroid-based classifiers. They conduct the classification process as illustrated in Figure 1. After selecting appropriate attributes, all instances of data are representated as sequences of a fixed number m of attributes $(a_1, \ldots, a_m)$, where $a_i \in F$ and F can be regarded as a semifield. The set of all sequences of length m with entries in F is denoted by F^m. Every centroid-based classifier selects special elements $c_1, \ldots, c_k$ in F^m. These special elements are called *centroids* (see, for example, [7, 27, 28, 30]). Each centroid c_i, where $i = 1, \ldots, k$, defines its own class $N(c_i)$ that consists of all vectors v such that c_i is the nearest centroid of v. This means that every vector is assigned to the class of its nearest centroid.

Second, centroid sets are often used in information security for analysis of data to combine initial binary classifiers (see, for example, [17, 35]). Recall that a classifier is said to be *binary* if it divides all data into two classes. Initially, several individual binary classifiers are trained on a data set. Then they are combined into one unified classification scheme with several classes as depicted in Figure 2. This method is often recommended for various applications, because of its effectiveness and ability to correct errors of the individual initial classifiers (cf. [34], Section 7.5).

If the number of individual binary classifiers is equal to m and the set of their outputs can be endowed with operations to turn it into a semifield, then every collection of outputs of the initial classifiers can be represented as a vector $(o_1, \ldots, o_m) \in F^m$. To define the unified multiclass classifier a set of centroids $c_1, \ldots, c_k$ is chosen in F^m. The class $N(c_i)$ of the centroid c_i is defined as the set of all output vectors that have c_i a nearest vector among all the centroids selected. This procedure of defining unified multiclass classifiers by combining initial binary classifiers is also often used in information security.

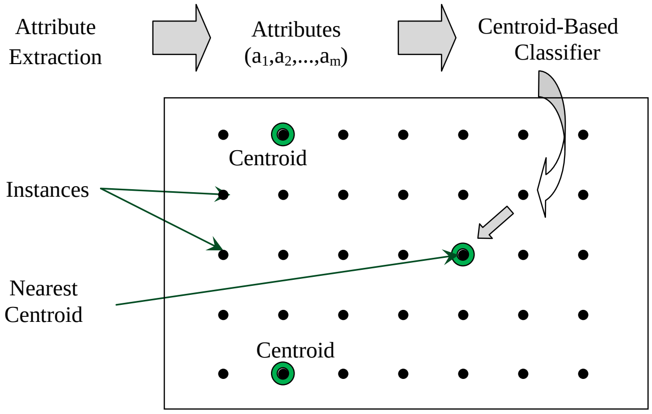

FIGURE 1. Centroid-based classifier.

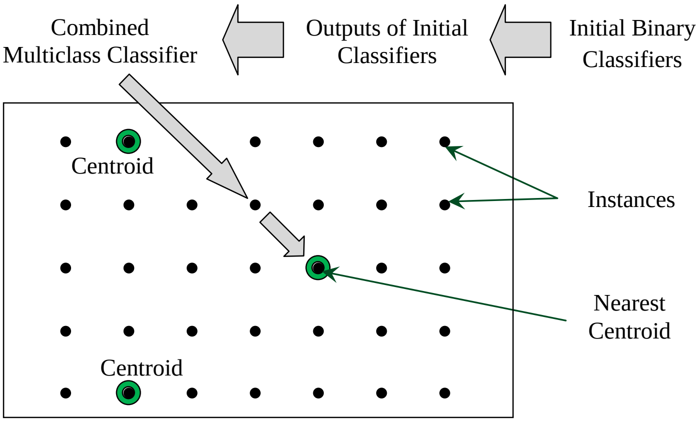
Combined Multiclass Classifier
Outputs of Initial Classifiers
Initial Binary Classifiers
Centroid
Instances
Nearest Centroid
Centroid

FIGURE 2. Combined multiclass classifier.

4. Main Result

Our main theorem shows that, in every incidence semiring over a semifield, each centroid set J of the largest distance has the largest possible weight, and the distance of J is equal to its weight.

Theorem 4.1 (Main Theorem). *Let D be a balanced digraph, F a semifield, and let $I_D(F)$ be an incidence semiring. If J is a centroid set with the largest distance among the distances of all centroid sets in $I_D(F)$, then J also has the largest weight among the weights of all centroid sets in $I_D(F)$ and the distance of J is equal to its weight.*

An example of a centroid set J in an incidence semiring such that the distance $\mathrm{d}(J)$ is strictly less than the weight $\mathrm{wt}(J)$ is shown in Figure 3.

Example. Let n be a positive integer, F a semifield, and let $D_0 = (V_0, E_0)$ be the digraph with the set $V_0 = \{a_0, b_0, c_0\}$ of vertices and the set $E_0 = \{(a_0, b_0), (a_0, c_0)\}$ of edges. For a positive integer n and $i = 1, \ldots, n$, let $D_i = (V_i, E_i)$ be the digraph with the set $V_i = \{a_i, b_i\}$ of vertices and the set $E_i = \{(a_i, b_i)\}$ of edges. Let $D = (V, E) = D_0 \cup D_1 \cup \cdots \cup D_n$ be the union of these digraphs, i.e., $V = V_0 \cup V_1 \cup \cdots \cup V_n$ and $E = E_0 \cup E_1 \cup \cdots \cup E_n$. Put $x = \sum_{i=0}^{n}(a_i, b_i)$ and $y = (a_0, c_0) + \sum_{i=1}^{n}(a_i, b_i)$. Then it is easily seen that the centroid set $C(x, y)$ coincides with the set $\{fx \mid f \in F\} \cup \{fy \mid f \in F\}$. Therefore $\mathrm{d}(C(x, y)) = 1$ and $\mathrm{wt}(C(x, y)) = n + 1$.

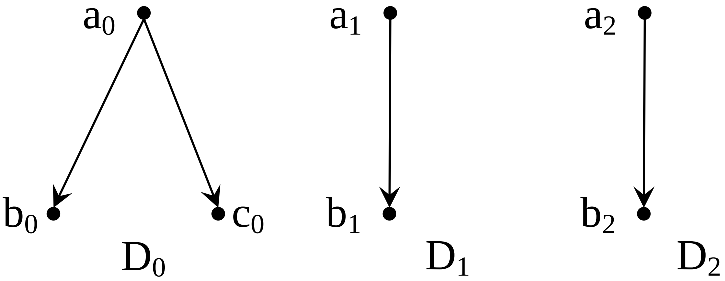

FIGURE 3. Digraphs D_0, D_1, D_2 such that $I_{D_0\times D_1\times D_2}(F)$ has a centroid set J with $\mathrm{d}(J) < \mathrm{wt}(J)$.

A semiring R is said to be *zero-divisor-free*, if $xy = 0$ implies $x = 0$ or $y = 0$, for any $x, y \in R$. Clearly, every semifield is zero-divisor-free. On the other hand, there exist zero-divisor-free semirings which are not semifields.

Example. Let $S = \{s_0, s_1\}$ be a semilattice such that $s_0s_1 = s_0$, as shown in Figure 4. Consider the semigroup semiring $\mathbb{B}[S]$ over the Boolean semiring $\mathbb{B} = \{0, 1\}$. It is routine to verify that $R = \mathbb{B}[S]$ is zero-divisor-free. However, $(1s_0)(1s_1) = 1s_0$ in $\mathbb{B}[S]$. Therefore R is not a semifield.

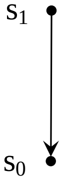

FIGURE 4. The semilattice $S = \{s_0, s_1\}$ of Example 4 as a partially ordered set.

Our next example shows that it is impossible to generalize our main theorem by weakening the hypothesis that F is a semifield and replacing it with the requirement that F be a zero-divisor-free semiring with identity element. This example uses the notation $\mathcal{G}_L$ defined in the next section, and relies on Lemma 5.3 also proved in the next section. We keep the proofs, technical lemmas and definitions in the next section, but present the next example here, since it is essential for understanding the role of conditions included in our main theorem.

Example. Let $D = (V, E)$ be the digraph with the set $V = \{u_1, u_2, v, w\}$ of vertices and the set $E = \{(u_1, v), (u_2, v), (u_1, w), (u_2, w), (v, w)\}$ of edges, and let F be the semiring of Example 1. Consider the element $x = s_1(u_1, v) + s_0(u_2, v)$ in $I_D(F)$. It is straightforward to verify that x belongs to $\mathcal{G}_L$, and so Lemma 5.3 tells us that $\mathrm{d}(C(x)) = \mathrm{wt}(x)$. However, $xs_1(v, w) = s_1(u_1, w) + s_0(u_2, w)$ and $xs_0(v, w) = s_0(u_1, w) + s_0(u_2, w)$ imply that $\mathrm{d}(C(x)) = 1 < wt(x)$. This example shows that it is impossible to generalize our main theorem by weakening the hypothesis that F is a semifield and replacing it with the requirement that F be a zero-divisor-free semiring with identity element.

It would be interesting to study the distances of centroid sets in the incidence semirings of various important special classes of graphs. For example, Moore graphs are important and have been investigated by many authors. Recall that a *Moore graph* or digraph is a graph or digraph that meets the Moore bound or directed Moore bound, respectively. Let us refer to the survey [31] and articles [8, 10, 11, 12, 14] for more information pertaining to the Moore graphs.

Problem 1. For each positive integer n, find the largest possible integer m such that there exists a Moore graph D with n vertices, a semifield and a centroid set J in $I_D(F)$ such that $\mathrm{d}(C) = m$.

5. Proof of the main theorem

Let us begin with an easy lemma, included for completeness.

Lemma 5.1. *If C is a centroid set in a semiring, then* $\mathrm{d}(C) \leq \mathrm{wt}(C)$.

Proof. Take x_1, x_2 in C such that $\mathrm{d}(C) = \mathrm{d}(x_1, x_2)$. Then $\mathrm{d}(x_1, x_2)$ is the minimal nonzero distance between a pair of elements in $C(x)$. Let x be an element of C such that $\mathrm{wt}(x) = \mathrm{wt}(C)$. Then $\mathrm{wt}(x) = \mathrm{d}(x, 0)$. Therefore $\mathrm{d}(x_1, x_2) \leq \mathrm{wt}(x) = \mathrm{wt}(C)$. □

The following definitions are required for the proof of the main theorem. An element of $I_D(F)$ is said to be *homogeneous* if it belongs to the union $\cup_{e\in E} Fe$. Following [6], for any vertex $v \in V$, we define the sets of vertices

$$\mathrm{In}(v) = \{w \in V \mid (w, v) \in E\},$$
$$\mathrm{Out}(v) = \{w \in V \mid (v, w) \in E\}.$$

Let us define the sets of edges

$$\begin{aligned} E_{\text{in}} &= \{(u, v) \in E \mid \exists w \in V : (w, u), (w, v) \in E\}, \\ E_{\text{out}} &= \{(u, v) \in E \mid \exists w \in V : (u, w), (v, w) \in E\}. \end{aligned}$$

For any pair of elements $x = \sum_{e\in E} x_e e$ and $y = \sum_{e\in E} y_e e$ in $I_D(F)$, put

$$\text{diff}(x, y) = \{e \in E \mid x_e \neq y_e\}, \tag{5.1}$$

$$\text{supp}(x) = \{e \in E \mid x_e \neq 0\}. \tag{5.2}$$

For a positive integer $m \in \mathbb{N}$, denote by $\mathcal{P}_m$ the set of all pairs (U, v), where $v \in V$ and $U \subseteq \mathrm{In}(v)$ are such that $|U| = m$, $(u, v) \notin E_{\text{in}}$ for all $u \in U$, and the intersection $\mathrm{Out}(v) \cap \mathrm{Out}(u)$ is equal to the same set for all vertices u in U. Let W_L be the largest positive integer such that $\mathcal{P}_{W_L}$ is not empty, or zero if such integers do not exist. Denote by $\mathcal{G}_L$ the set of all elements $x = \sum_{u\in U} x_u(u, v)$, for all pairs $(U, v) \in \mathcal{P}_{W_L}$.

For the proof we need [6, Theorem 4.1], which gives a description of all centroid sets of the largest weight. For convenience of the readers we include it as a separate lemma.

Lemma 5.2. ([6, Theorem 4.1]) *Let R be a zero-divisor-free semiring with identity element, D a balanced graph, and let $C = C(g_1, \ldots, g_k)$ be an ideal with the largest weight in $I_D(R)$. Then the following conditions hold:*

(i) $\mathrm{wt}(C) = \max\{1, W_L, W_R, W_Z\}$*;*
(ii) *C contains an element of weight $\mathrm{wt}(C)$ belonging to the union of $\mathcal{G}_Z$, $\mathcal{G}_L$ and $\mathcal{G}_R$;*
(iii) $\mathrm{wt}(C(r)) = \mathrm{wt}(r) = W_Z$*, for all $r \in \mathcal{G}_Z$;*
(iv) $\mathrm{wt}(C(r)) = \mathrm{wt}(r) = W_L$*, for all $r \in \mathcal{G}_L$;*
(v) $\mathrm{wt}(C(r)) = \mathrm{wt}(r) = W_R$*, for all $r \in \mathcal{G}_R$.*

Lemma 5.3. *If $x \in \mathcal{G}_L$, then* $\mathrm{d}(C(x)) = \mathrm{wt}(x)$.

Proof. The definition of the set $\mathcal{G}_L$ tells us that there exists a pair $(U, v) \in \mathcal{P}_{W_L}$ such that $|U| = W_L$ and

$$x = \sum_{u\in U} x_u(u, v) \in I_D(F), \tag{5.3}$$

where $0 \neq x_u \in F$. Therefore $\mathrm{wt}(x) = |U| = W_L$.

To prove that $\mathrm{d}(C(x)) = \mathrm{wt}(x)$, pick any pair of elements $x^{(1)}, x^{(2)}$ in $C(x)$ such that $\mathrm{d}(C(x)) = \mathrm{d}(x^{(1)}, x^{(2)})$. Then $\mathrm{d}(x^{(1)}, x^{(2)})$ is nonzero and is the minimal distance between a pair of elements that belong to $C(x)$. Therefore $\mathrm{d}(x^{(1)}, x^{(2)}) \leq \mathrm{wt}(x)$, because $\mathrm{wt}(x) = \mathrm{d}(0, x)$ and $0, x \in C(x)$. It remains to verify that the reversed inequality $\mathrm{d}(x^{(1)}, x^{(2)}) \geq \mathrm{wt}(x)$ holds.

For $i = 1, 2$, it follows from (2.2) that $x^{(i)}$ can be represented in the form

$$x^{(i)} = \sum_{j=1}^{k^{(i)}} \ell_j^{(i)} x r_j^{(i)}, \tag{5.4}$$

for some $k^{(i)} \in \mathbb{N}_0$ and $\ell_j^{(i)}, r_j^{(i)} \in I_D(F) \cup \{1\}$. We may assume that only nonzero summands $\ell_j^{(i)} x r_j^{(i)}$ have been included in the representation (5.4) for $x^{(i)}$. Since x was chosen in $\mathcal{G}_L$, we have $(U, v) \in \mathcal{P}_{W_L}$. Therefore U is a subset of $\mathrm{In}(v)$ such that $(u, v) \notin E_{\mathrm{in}}$, for all $u \in U$. It follows from (2.1) that $I_D(F)(u, v) = 0$, for all $u \in U$. Therefore $I_D(F)x = 0$. Hence $\ell_j^{(i)} x = 0$ for all $\ell_j^{(i)} \in I_D(F)$. Since only nonzero terms have been included in (5.4), further we may assume that all the $\ell_j^{(i)}$ are equal to 1 in the expression (5.4) for $x^{(i)}$. This means that

$$x^{(i)} = \sum_{j=1}^{k^{(i)}} x r_j^{(i)}, \tag{5.5}$$

where $k^{(i)} \in \mathbb{N}_0$ and $r_j^{(i)} \in I_D(F) \cup \{1\}$. We can collect all similar terms with coefficients $r_j^{(i)}$ equal to 1 into one summand, and assume that from the very beginning $x^{(i)}$ has been rewritten in the form

$$x^{(i)} = m^{(i)} x + \sum_{j=1}^{k^{(i)}} x r_j^{(i)}, \tag{5.6}$$

where $k^{(i)}, m^{(i)} \in \mathbb{N}_0$ and $r_j^{(i)} \in I_D(F)$.

Every element $r_j^{(i)} \in I_D(F) = \bigoplus_{e \in E} Fe$ can be represented as a sum

$$r_j^{(i)} = \sum_{e \in E} (r_j^{(i)})_e e, \tag{5.7}$$

where $(r_j^{(i)})_e \in F$ and where only a finite number of the coefficients $(r_j^{(i)})_e$ are nonzero. Now we substitute all expressions (5.7) for $r_j^{(i)}$ in (5.6) and apply the distributive law. This rewrites $x^{(i)}$ in the form similar to (5.6) with all the general elements $r_j^{(i)}$ replaced by homogeneous elements $(r_j^{(i)})_e e$. To simplify notation we may assume that from the very beginning all the elements $r_j^{(i)}$ in (5.6) have been chosen homogeneous, that is $r_j^{(i)} = f_j^{(i)}(u_j^{(i)}, v_j^{(i)})$ for some $f_j^{(i)} \in F$, $(u_j^{(i)}, v_j^{(i)}) \in E$.

Since we have already assumed that $x r_j^{(i)} \neq 0$ in (5.6), it follows from (5.3) and (2.1) that $u_j^{(i)} = v$, for all $j = 1, \dots, k^{(i)}$. Besides, $(U, v) \in \mathcal{P}_{W_L}$ implies that the intersection $\mathrm{Out}(v) \cap \mathrm{Out}(u)$ is equal to one and the same set T for all u in U. Given that $x r_j^{(i)} \neq 0$, there exists $w \in U$ such that $(w, v)(v, v_j^{(i)}) \neq 0$. Hence $(w, v)(v, v_j^{(i)}) = (w, v_j^{(i)}) \in E$ by (2.1). In particular, $v_j^{(i)} \in \mathrm{Out}(v) \cap \mathrm{Out}(w) = T$. Therefore $v_j^{(i)} \in \mathrm{Out}(v) \cap \mathrm{Out}(u) = T$ for all $u \in U$. By (2.1), we get $(u, v)(v, v_j^{(i)}) = (w, v_j^{(i)}) \in E$, for all $u \in U$. Therefore, substituting the expression (5.3) for x in (5.6) and applying (2.1), we get

$$x^{(i)} = m^{(i)} \sum_{u \in U} x_u (u, v) + \sum_{j=1}^{k^{(i)}} \left(\sum_{u \in U} x_u f_j^{(i)} (u, v_j^{(i)}) \right), \tag{5.8}$$

where $m^{(i)}, k^{(i)} \in \mathbb{N}_0$, $0 \neq x_u \in F$, $v_j^{(i)} \in V$ and $0 \neq f_j^{(i)} \in F$. Next, we show that two more simplifications can be made to the expression (5.8).

First, suppose that $v_{j_1}^{(i)} = v_{j_2}^{(i)}$, for some $1 \leq j_1 < j_2 \leq k^{(i)}$. Then the equality

$$x_u f_{j_1}^{(i)}(u, v_{j_1}^{(i)}) + x_u f_{j_2}^{(i)}(u, v_{j_2}^{(i)}) = x_u (f_{j_1}^{(i)} + f_{j_2}^{(i)})(u, v_{j_1}^{(i)})$$

implies

$$\left(\sum_{u \in U} x_u f_{j_1}^{(i)}(u, v_{j_1}^{(i)})\right) + \left(\sum_{u \in U} x_u f_{j_2}^{(i)}(u, v_{j_2}^{(i)})\right) = \left(\sum_{u \in U} x_u \left(f_{j_1}^{(i)} + f_{j_2}^{(i)}\right)(u, v_{j_1}^{(i)})\right).$$

Applying this we can combine the likely terms in (5.8). Therefore to simplify notation we may assume that in the sum (5.8) from the very beginning all the $v_j^{(i)}$ are pairwise distinct.

Second, suppose that $v_j^{(i)} = v$, for some $1 \leq j \leq k^{(i)}$. Then we get $(m^{(i)} 1_F + f_j^{(i)}) \in F$ and

$$m^{(i)} x_u(u, v) + x_u f_j^{(i)}(u, v_j^{(i)}) = (m^{(i)} 1_F + f_j^{(i)}) x_u(u, v_j^{(i)}).$$

Therefore

$$m^{(i)} \sum_{u \in U} x_u(u, v) + \left(\sum_{u \in U} x_u f_j^{(i)}(u, v_j^{(i)})\right) = \left(m^{(i)} 1_F + f_j^{(i)}\right) \left(\sum_{u \in U} x_u(u, v_j^{(i)})\right).$$

Applying this to combine similar terms in (5.8), we may assume that $m^{(i)} = 0$ whenever there exists j such that $v_j^{(i)} = v$.

In order to make it easier to compare $x^{(1)}$ and $x^{(2)}$, we are going to rewrite (5.8) in another form. For $i = 1, 2$, put $V^{(i)} = \{v_1^{(i)}, \ldots, v_{k^{(i)}}^{(i)}\}$. Set $\overline{V} = V^{(1)} \cup V^{(2)}$. Denote the elements of $\overline{V}$ by $v_1, \ldots, v_k$ so that $\overline{V} = \{v_1, \ldots, v_k\}$. For $i = 1, 2$ and $j = 1, \ldots, k$, if there exists ℓ such that $v_j = v_\ell^{(i)}$, then we put $h_j^{(i)} = f_\ell^{(i)}$. If there does not exist ℓ such that $v_j = v_\ell^{(i)}$, then we put $h_j^{(i)} = 0$. Using these new coefficients $h_j^{(i)}$, we can rewrite (5.8) as follows

$$x^{(i)} = m^{(i)} \sum_{u \in U} x_u(u, v) + \sum_{j=1}^{k} \left(\sum_{u \in U} x_u h_j^{(i)}(u, v_j)\right), \tag{5.9}$$

where $m^{(i)}, k \in \mathbb{N}_0$, $0 \neq x_u \in F$, $v_j \in \overline{V}$, $h_j^{(i)} \in F$, and where all elements $v_1, \ldots, v_k$ are pairwise distinct.

By the choice of the pair $x^{(1)}, x^{(2)}$, we have $x^{(1)} \neq x^{(2)}$. Therefore there exists an element (u_0, v_0) in $\mathrm{diff}(x^{(1)}, x^{(2)})$. Hence (5.9) implies

$$\mathrm{diff}(x^{(1)}, x^{(2)}) \subseteq \mathrm{supp}(x^{(1)}) \cup \mathrm{supp}(x^{(2)}) \subseteq \cup\{(u, w) \mid u \in U, w \in \{v\} \cup \overline{V}\}.$$

Therefore $u_0 \in U$ and $v_0 \in \{v\} \cup \overline{V}$. Consider two possible cases.

Case 1. $v_0 \in \overline{V}$. Without loss of generality we may assume that $v_0 = v_1$ and $v_1 \in V^{(1)}$, because $\overline{V} = V^{(1)} \cup V^{(2)}$. Then $h_1^{(1)} \neq 0$. Clearly, the edge (u_0, v_0) occurs in the expressions (5.9) for $x^{(1)}$ and $x^{(2)}$ with coefficients $x_{u_0} h_1^{(1)}$ and $x_{u_0} h_1^{(2)}$, respectively. It follows from $(u_0, v_0) \in \mathrm{diff}(x^{(1)}, x^{(2)})$ that $h_1^{(1)} \neq h_2^{(1)}$. Given that F is a semifield, we get $x_u h_1^{(1)} \neq x_u h_2^{(1)}$, for all $u \in U$. (Notice that in this step of the proof we have to use the fact that F is a semifield. It is impossible to weaken this condition by requiring that F be a zero-divisor-free semiring with identity element. Indeed, there exist zero-divisor-free semirings with elements $x_{u_0}, x_{u_1}, h_1^{(1)}, h_1^{(2)}$ such that $x_{u_0} h_1^{(1)} \neq x_{u_0} h_1^{(2)}$ but $x_{u_1} h_1^{(1)} = x_{u_1} h_1^{(2)}$, see Example 4 in Section 4.) Hence

$$\{(u, v_1) \mid u \in U\} \subseteq \mathrm{diff}\left(\sum_{u \in U} x_u(u, v_1) h_1^{(1)}, \sum_{u \in U} x_u(u, v_1) h_1^{(2)}\right), \tag{5.10}$$

because F is a semifield. Now, we look at two subcases.

Subcase 1.1. $v_0 \neq v$. Then expression (5.9) does not have any other summands that involve edges ending in v_1, with the exception of summands already listed in (5.10). Therefore (5.10) yields us

$$\{(u, v_1) \mid u \in U\} \subseteq \text{diff}(x^{(1)}, x^{(2)}).$$

Thus, $|\,\text{diff}(x^{(1)}, x^{(2)})| \geq |U| = \text{wt}(x)$, as required.

Subcase 1.2. $v_0 = v$. Then $v = v_1$, and so we can simplify (5.9) as follows. Note that $m^{(i)}1_F + h_1^{(i)} \in F$. Therefore $m^{(i)}x_u(u,v) + x_u h_1^{(i)}(u, v_1) = x_u(m^{(i)}1_F + h_1^{(i)})(u,v)$ implies that

$$m^{(i)} \sum_{u \in U} x_u(u,v) + \sum_{u \in U} x_u h_1^{(i)}(u, v_1) = \sum_{u \in U} x_u(m^{(i)}1_F + h_1^{(i)})(u,v).$$

Using this equality we can replace $m^{(i)}$ by 0 and at the same time replace $h_1^{(i)}$ by $(m^{(i)}1_F + x_u h_1^{(i)})$ in (5.9). In order to keep notation simple, this transformation allows us to assume that from the very beginning $m^{(i)} = 0$ in (5.9). This implies that expression (5.9) does not have any other nonzero summands involving edges which end in v_1, with the exception of summands already listed in (5.10). Therefore (5.10) yields us $\{(u, v_1) \mid u \in U\} \subseteq \text{diff}(x^{(1)}, x^{(2)})$, and so $|\,\text{diff}(x^{(1)}, x^{(2)})| \geq |U| = \text{wt}(x)$, in this subcase, too.

Case 2. $v_0 = v \notin V$. Then $m^{(1)}x_{u_0} \neq m^{(1)}x_{u_0}$. Therefore $m^{(1)}1_F \neq m^{(1)}1_F$ in F. It follows that $m^{(1)}x_u \neq m^{(1)}x_u$ for each $u \in U$, because F is a semifield. Hence

$$\{(u,v) \mid u \in U\} \subseteq \text{diff}\left(m^{(1)} \sum_{u \in U} x_u(u,v), m^{(2)} \sum_{u \in U} x_u(u,v) \right). \tag{5.11}$$

In this case other edges ending in v do not occur in any other summands from (5.9). Therefore (5.11) implies

$$\{(u,v) \mid u \in U\} \subseteq \text{diff}(x^{(1)}, x^{(2)}).$$

Hence $|\,\text{diff}(x^{(1)}, x^{(2)})| \geq |U| = \text{wt}(x)$ in this case, too.

Thus, $|\,\text{diff}(x^{(1)}, x^{(2)})| \geq \text{wt}(x)$ in all possible cases. This completes the proof. □

Example 4 demonstrates that the proof of Lemma 5.3 has to use the hypothesis that F is a semifield.

For a positive integer m, let $\mathcal{Q}_m$ denote the set of all pairs (v, U), where $v \in V$ and $U \subseteq \text{Out}(v)$ are such that $|U| = m$, $(v,u) \notin E_{\text{out}}$ for all $u \in U$, and the intersection $\text{In}(v) \cap \text{In}(u)$ is equal to one and the same set for all vertices u in U. Denote by W_R the largest positive integer such that $\mathcal{Q}_m$ is not empty, or zero if such positive integers do not exist. Let $\mathcal{G}_R$ be the set of all elements $x = \sum_{u \in U} x_u(v,u) \in I_{D_{\text{out}}}(F)$, where $0 \neq x_u \in F$ and $(v, U) \in \mathcal{Q}_{W_R}$.

Lemma 5.4. *If* $x \in \mathcal{G}_R$, *then* $\text{d}(C(x)) = \text{wt}(x)$.

Proof. The proof is dual to that of Lemma 5.3, and so we omit it. □

Let $\mathcal{G}_Z$ be the set of all elements $x = \sum_{(u,v) \in E \setminus (E_{\text{out}} \cup E_{\text{in}})} x_{u,v}(u,v) \in I_D(F)$ such that $0 \neq x_{u,v} \in F$ for all $(u,v) \in E \setminus (E_{\text{out}} \cup E_{\text{in}})$. Put $W_Z = |E \setminus (E_{\text{out}} \cup E_{\text{in}})|$.

Lemma 5.5. *If* $x \in \mathcal{G}_Z$, *then* $\text{d}(C(x)) = \text{wt}(x)$.

Proof. The definition of the set $\mathcal{G}_Z$ shows that

$$x = \sum_{(u,v) \in E \setminus (E_{\text{out}} \cup E_{\text{in}})} x_{u,v}(u,v) \in I_D(F) \tag{5.12}$$

where $0 \neq x_{u,v} \in F$ for all $(u,v) \in E \setminus (E_{\text{out}} \cup E_{\text{in}})$. Therefore $\text{wt}(x) = |E \setminus (E_{\text{out}} \cup E_{\text{in}})|$.

To prove that $\text{d}(C(x)) = \text{wt}(x)$, choose $x^{(1)}, x^{(2)}$ in $C(x)$ such that $\text{d}(C(x)) = \text{d}(x^{(1)}, x^{(2)})$. Then $\text{d}(x^{(1)}, x^{(2)})$ is the minimal nonzero distance between a pair of elements in $C(x)$.

For $i = 1, 2$, it follows from (2.2) that $x^{(i)}$ can be represented in the form

$$x^{(i)} = \sum_{j=1}^{k^{(i)}} \ell_j^{(i)} x r_j^{(i)} = \sum_{j=1}^{k^{(i)}} \left(\sum_{(u,v)\in E\setminus(E_{\rm out}\cup E_{\rm in})} \ell_j^{(i)} x_{u,v}(u,v) r_j^{(i)} \right) \tag{5.13}$$

for some $k^{(i)} \in \mathbb{N}_0$ and $\ell_j^{(i)}, r_j^{(i)} \in I_D(F) \cup \{1\}$. We may assume that only nonzero summands $\ell_j^{(i)} x r_j^{(i)}$ have been included in (5.13). For any edge (u, v) in (5.13), we have $(u, v) \notin E_{\rm in}$, and so (2.1) implies that $I_D(F)(u, v) = 0$. Hence $\ell_j^{(i)}(u, v) = 0$ for all $\ell_j^{(i)} \in I_D(F)$. Since only nonzero terms have been included in (5.13), further we may assume that all the $\ell_j^{(i)}$ are equal to 1 in the expression (5.13) for $x^{(i)}$. Similarly, $(u, v) \notin E_{\rm out}$ and (2.1) yield us that $(u, v)I_D(F) = 0$. Hence $(u, v)r_j^{(i)} = 0$ for all $r_j^{(i)} \in I_D(F)$. Therefore we may also assume that all the $r_j^{(i)}$ in (5.13) are equal to 1. It follows that $x^{(i)} = m^{(i)}x$, for some $m^{(i)} \in \mathbb{N}$. Since $x^{(1)} \neq x^{(2)}$, we get $m^{(1)}1_F \neq m^{(2)}1_F$, where 1_F is the identity element of F. Therefore $m^{(1)}1_F x_u \neq m^{(2)}1_F x_u$, for each $x_u \in F$, because F is a semifield. Therefore $\mathrm{d}(x^{(1)}, x^{(2)}) = E \setminus (E_{\rm out} \cup E_{\rm in}) = \mathrm{wt}(x)$, as required. □

Proof of Theorem 4.1. Choose a pair of elements $x, y \in J$ such that $\mathrm{d}(x, y) = \mathrm{d}(J)$. Let C be a centroid set in $I_D(F)$ such that the weight of K is the largest one among the weights of all centroid sets in $I_D(F)$. Condition (i) of Lemma 5.2 tells us that $\mathrm{wt}(C) = \max\{1, W_Z, W_L, W_R\}$. Therefore the following cases are possible.

Case 1. $\mathrm{wt}(C) = 1$. Since $\mathrm{wt}(J) \leq \mathrm{wt}(C)$, we get $\mathrm{wt}(J) = 1$. Lemma 5.1 implies that $\mathrm{d}(J) = 1$. Hence $\mathrm{wt}(J) = \mathrm{d}(J)$, as required.

Case 2. $\mathrm{wt}(C) = W_R$. By the definition of W_R, there exists an element $x \in \mathcal{G}_R$ satisfying $\mathrm{wt}(x) = W_R$. Lemma 5.3 tells us that $\mathrm{d}(C(x)) = \mathrm{wt}(x)$. The maximality of $\mathrm{wt}(C)$ implies that $\mathrm{wt}(J) \leq \mathrm{wt}(C)$. Lemma 5.1 shows that the reversed inequality $\mathrm{d}(J) \leq \mathrm{wt}(J)$ holds. Thus $\mathrm{d}(J) = \mathrm{wt}(J) = \mathrm{wt}(C)$ in this case, too.

Case 3. $\mathrm{wt}(C) = W_L$. By the definition of W_L, there exists an element $x \in \mathcal{G}_L$ satisfying $\mathrm{wt}(x) = W_L$. Lemma 5.4 says that $\mathrm{d}(C(x)) = \mathrm{wt}(x)$. Since $\mathrm{wt}(C)$ is the largest weight among the weights of all centroid sets, we get $\mathrm{d}(J) = \mathrm{wt}(C) > \mathrm{wt}(J)$. Lemma 5.1 establishes the reversed inequality. Hence $\mathrm{d}(J) = \mathrm{wt}(J) = \mathrm{wt}(C)$, as claimed.

Case 4. $\mathrm{wt}(C) = W_Z$. By the definition of W_Z, there exists an element $x \in \mathcal{G}_Z$ satisfying $\mathrm{wt}(x) = W_Z$. Lemma 5.5 tells us that $\mathrm{d}(C(x)) = \mathrm{wt}(x)$. The maximality of $\mathrm{wt}(C)$ implies that $\mathrm{d}(J) = \mathrm{wt}(C) > \mathrm{wt}(J)$. The reversed inequality holds by Lemma 5.1. Therefore $\mathrm{d}(J) = \mathrm{wt}(J) = \mathrm{wt}(C)$, again.

Thus, we see that $\mathrm{d}(J) = \mathrm{wt}(J) = \mathrm{wt}(C)$ in all possible cases. This completes the proof. □

References

[1] J. Abawajy and A. V. Kelarev, *A multi-tier ensemble construction of classifiers for phishing email detection and filtering*, Cyberspace Safety and Security, CSS 2012, Lecture Notes in Computer Science, **7672** (2012), 48–56.

[2] J. Abawajy, A. V. Kelarev and M. Chowdhury, *Power graphs: a survey*, Electronic J. Graph Theory and Applications, **1** (2013) (2), 125–147.

[3] J. H. Abawajy, A. V. Kelarev and M. Chowdhury, *Multistage approach for clustering and classification of ECG data*, Computer Methods and Programs in Biomedicine, **112** (2013), 720-730.

[4] J. Abawajy, A. Kelarev, M. Chowdhury, A. Stranieri, H. F. Jelinek, *Predicting cardiac autonomic neuropathy category for diabetic data with missing values*, Computers in Biology and Medicine, **43** (2013), 1328–1333.

[5] J. Abawajy, A. V. Kelarev, M. Miller, J. Ryan, *Incidence semirings of graphs and visible bases*, Bulletin of the Australian Mathematical Society, 2014, DOI:10.1017/S000497271300083X.

[6] J. Abawajy, A. V. Kelarev, J. L. Yearwood, C. Turville, *A data mining application of the incidence semirings*, Houston J. Math., **39** (2013) (4), 1083–1093.

[7] J. Abawajy, A. V. Kelarev and J. Zeleznikow, *Optimization of classification and clustering systems based on Munn semirings*, Semigroup Forum, DOI: 10.1007/s00233-013-9488-5.

[8] E. T. Baskoro, Y. M. Cholily and M. Miller, *Structure of selfrepeat cycles in almost Moore digraphs with selfrepeats and diameter 3*, Bulletin of the Institute of Combinatorics and its Applications, **46** (2006), 99-109.

[9] M. Baca, S. Jendrol, M. Miller and J. Ryan, *On irregular total labellings*, Discrete Mathematics, **307** (2007), 1378–1388.

[10] E. T. Baskoro, Y. M. Cholily and M. Miller, *Enumeration of vertex orders of almost Moore digraphs with selfrepeats*, Discrete Mathematics, **308** (2008), 123–128.

[11] E. T. Baskoro, M. Miller and J. Plesnik, *On the structure of digraphs with order close to the Moore bound*, Graphs and Combinatorics, **14** (1998), 109–119.

[12] E. T. Baskoro, M. Miller and J. Plesnik, *Further results on almost Moore digraphs*, Ars Combinatoria, **56** (2000), 43–63.

[13] E. T. Baskoro, M. Miller, J. Širán and M. Sutton, *Complete characterization of almost Moore digraphs of degree three*, Journal of Graph Theory, **48** (2005), 112–126.

[14] E. T. Baskoro, M. Miller, J. Širán and M. Sutton, *Complete characterization of almost Moore digraphs of degree three*, Journal of Graph Theory, **48** (2005), 112–126.

[15] G. Beliakov, J. Yearwood and A. Kelarev, *Application of rank correlation, clustering and classification in information security*, Journal of Networks, **7** (2012), 935–945.

[16] E. Dahlhaus, P. Horak, M. Miller and J. Ryan, *The train marshalling problem*, Discrete Applied Mathematics, **103** (2000), 41–54.

[17] R. Dazeley, J. L. Yearwood, B. H. Kang and A. V. Kelarev, *Consensus clustering and supervised classification for profiling phishing emails in internet commerce security*, Lecture Notes in Computer Science, **6232/2011**(2010), 235–246.

[18] D. Y. Gao, A. V. Kelarev and J. L. Yearwood, *Optimization of matrix semirings for classification systems*, Bull. Aust. Math. Soc., **84** (2011), 492–503.

[19] J. S. Golan, *Semirings and Their Applications*, Kluwer Academic Publishers, Dordrecht, 1999.

[20] A. V. Kelarev, *On undirected Cayley graphs*, Australasian Journal of Combinatorics, **25** (2002), 73–78.

[21] A. V. Kelarev, *Labelled Cayley graphs and minimal automata*, Australasian J. Combinatorics, **30** (2004), 95–101.

[22] A. V. Kelarev, *Ring Constructions and Applications*, World Scientific, River Edge, NJ, 2002.

[23] A. V. Kelarev, *Graph Algebras and Automata*, Marcel Dekker, New York, 2003.

[24] A. Kelarev, S. Brown, P. Watters, X.-W. Wu, and R. Dazeley, *Establishing reasoning communities of security experts for internet commerce security*, In: “Technologies for Supporting Reasoning Communities and Collaborative Decision Making: Cooperative Approaches”, IGI Global, 2011, pp. 380–396.

[25] A. Kelarev, B. Kang and D. Steane, *Clustering algorithms for ITS sequence data with alignment metrics*, Lect. Notes Artificial Intelligence, **4304** (2006),1027–1031.

[26] A. V. Kelarev and D. S. Passman, *A description of incidence rings of group automata*, Contemporary Mathematics, **456** (2008), 27–33.

[27] A. Kelarev, J. Ryan and J. Yearwood, *Cayley graphs as classifiers for data mining: The influence of asymmetries*, Discrete Math., **309** (2009), 5360–5369.

[28] A. V. Kelarev, J. L. Yearwood and M. A. Mammadov, *A formula for multiple classifiers in data mining based on Brandt semigroups*, Semigroup Forum, **78** (2009), 293–309.

[29] A. V. Kelarev, J. L. Yearwood and P. W. Vamplew, *A polynomial ring construction for classification of data*, Bull. Aust. Math. Soc., **79** (2009), 213–225.

[30] A. V. Kelarev, J. L. Yearwood, P. Watters, X. W. Wu, J. H. Abawajy and L. Pan, *Internet security applications of the Munn rings*, Semigroup Forum, **81** (2010), 162–171.

[31] M. Miller and J. Širán, *Moore graphs and beyond: A survey of the degree/diameter problem*, Electronic J. Combinatorics, Dynamic Survey, DS20, 2013, 92pp.

[32] E. Spiegel and C. J. O'Donnell, *Incidence Algebras*, Marcel Dekker, New York, 1997.

[33] A. Stranieri, J. Abawajy, A. Kelarev, S. Huda, M. Chowdhury, H. F. Jelinek, *An approach for Ewing test selection to support the clinical assessment of cardiac autonomic neuropathy*, Artificial Intelligence in Medicine, **58** (2013), 185–193.

[34] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Edition, Elsevier/Morgan Kaufman, Amsterdam, 2010.

[35] J. Yearwood, D. Webb, L. Ma, P. Vamplew, B. Ofoghi and A. Kelarev, *Applying clustering and ensemble clustering approaches to phishing profiling*, Data Mining and Analytics 2009, Proc. 8th Australasian Data Mining Conference: AusDM 2009, (1-4 December 2009, Melbourne, Australia) CRPIT, Vol. 101, pp. 25–34.

Acknowledgment

The authors are grateful to two referees for thorough corrections that have helped to improve the text of this article.

J. Abawajy
School of Information Technology
Deakin University
221 Burwood, Melbourne
Victoria 3125, Australia
e-mail: jemal.abawajy@deakin.edu.au

A.V. Kelarev
School of Information Technology
Deakin University
221 Burwood, Melbourne
Victoria 3125, Australia, and
School of Mathematical and Physical Sciences
University of Newcastle, University Drive
Callaghan, NSW 2308, Australia
Victoria 3125, Australia
e-mail: andreikelarev-deakinuniversity@deakin.edu.au

M. Miller
CARMA Priority Research Centre
School of Mathematical and Physical Sciences
University of Newcastle, University Drive
Callaghan, NSW 2308, Australia, and
Department of Mathematics
University of West Bohemia
Pilsen, Czech Republic
e-mail: Mirka.Miller@newcastle.edu.au

J. Ryan
School of Electrical Engineering and Computer Science
University of Newcastle, University Drive
Callaghan, NSW 2308, Australia
e-mail: Joe.Ryan@newcastle.edu.au