GRAPH DISTANCES FOR DETERMINING INTER-ENTITIES RELATIONS: A TOPOLOGICAL APPROACH TO FRAUD DETECTION *(WORK IN PROGRESS)*

J.M. CALABUIG, H. FALCIANI, A. FERRER-SAPENA, L.M. GARCÍA-RAFFI AND E.A. SÁNCHEZ-PÉREZ

Abstract. Given a set Ω and a proximity function $\phi : \Omega \times \Omega \to \mathbb{R}^+$, we define a new metric for Ω by considering a path distance in the global graph Ω, in which all the points are considered to be directly connected in a full graph. We analyze the properties of such a metric, and several procedures for defining the initial proximity matrix $(\phi(a,b))_{(a,b)\in\Omega\times\Omega}$. Our motivation has its roots in the current interest in finding effective algorithms for detecting and classifying relations among elements of a social network. For example, the analysis of a set of companies working for a given public administration or other figures in which fraud detection automatic systems are needed. Using this formalism, we state our main idea regarding fraud detection, that essentially says that fraud can be detected because it produces a meaningful local change of density in the metric space defined in this way.

1. Introduction

The great increase of network conexions due to the broad use of internet has opened the door to a new way of social organization, that allows to generate solid and powerful structures to commit economic fraud. In parallel, the development of the same technical tools that permit to stablish these criminal networks allow to create new procedures to detect them. Indeed, fraud detection is a current hot topic appearing daily in the news, and this produces a high demand of theoretical and practical mathematical instruments for fighting against fraud. Some theoretical developments coming from social sciences have been presented since the mid-twentieth century: the most poverful approach from this point of view seems to be the so called Fraud Triangle theory, that have show to be useful also in applications (see for example [3, 7, 8, 14, 15]). Our methodology, however, is based on the mathematical analysis of fraud.

The aim of this paper is to explain a new topological framework for understanding and detecting the processes of fraud. The big ammount of

2010 *Mathematics Subject Classification.* 54A10,05C12.

Key words and phrases. Graph distance, fraud detection, quasi-pseudo-metric, concentration of mass.

information that the new technologies bring into the scene have changed the way a scientist can understand the fraud as a mathematical phenomenon: invoices, emails, company registers, provide highly meaningful information that may help the analyst to detect evidences of fraud. The extraordinarily large set of data that accompanies any fraud process makes necessary to change the usual analysis tools, traditionally based on the lawyers study of related documents. New ways of understanding and informatic tools are clearly needed, and the theoretical development of the associated mathematical models must grow together. Therefore, our idea is to propose a new model based on a topological graph approach to the analysis of networks.

Several mathematical theories have been already applied to fraud detection, involving quite different approaches: game theory, statistical analysis, graph theory,... (see for example [1, 9, 11, 13, 17, 18]). One of the more succesful has been shown to be the graph based analytical approach, which has already given some programs for fraud detection. In this paper we propose a new technique for defining quasi-pseudo-metrics for complete graph structures. The vertices/nodes are the elements that must be analyzed: persons, entities, companies, invoices, emails... Starting with a graph with edges among vertices having a finite set of properties, we establish a way for defining a family of quasi-pseudo-metrics for translating the graph to a topological space. We will call such a struture a "topological graph", and the topology will be constructed using quasi-pseudo-metrics (see for example [6, 10] for the basics). Once we can define neighbours of vertices, we use the topological properties to characterize the relevant elements of the space, that must become the main objects of the antifraud analysis. Besides the topological space, we need an additive set function acting in the class of all subsets of the original set of nodes —a measure— for helping to measure the "size" (given in terms of number of elements, weighted means, or similar mathematical features) of the neighbours of the nodes. Together, both tools (metric and measure) allow to define the fundamental object of our model: the density of the family of neighbours of a given node.

The abstract main supporting idea of our model is that the fraud can be detected by searchig for *unusual concentration of mass phenomena* in a specifically defined topological graph. It can be established broadly in the following terms: *the "map of density" of a graph should follow an easy-to-recognize pattern. If no previous information on the pattern is available, then the hypothesis must be that the relevant vertices —the ones that must focus the attention of the antifraud analysts— are the ones in which there is an anomalous density distribution. In other where the uniform density distribution is assumed as reference pattern. Small local densities as well as big local densities should indicate a "hot node" in terms of corruption, and would allow to classify the different schemes of fraud.*

In this article we firstly present the mathematical structure, showing at each step examples that would help the reader to follow the development of the model. The main results will be shown in the central part of the paper.

2. Preliminaries

Let us introduce some technical formal concepts. We use standard mathematical notation. We will write $\mathbb{R}^+$ for the set of non-negative real numbers. A quasi-pseudo-metric on a set Ω ([6, 10]) is a function $d : \Omega \times \Omega \to \mathbb{R}^+$ satisfying that for $a, b, c \in \Omega$,

(1) $d(a, b) = 0$ if $a = b$, and
(2) $d(a, b) \leq d(a, c) + d(c, b)$.

Such a function is enough for defining a topology by means of the basis of neighborhoods that is given by the open balls. If $\varepsilon > 0$, we define the ball of radius ε and center in $a \in \Omega$ as

$$B_\varepsilon(a) := \Big\{ b \in \Omega : d(a, b) < \varepsilon \Big\}.$$

Note that this topology is in fact given by the countable basis of neighborhoods provided by the balls $B_{1/n}(x) = \{y \in X : d(x, y) \leq 1/n\}$, $n \in \mathbb{N}$. The resulting metrical/topological structure (Ω, d) is called a quasi-pseudo-metric space.

If the function d is symmetric, that is, $d(a, b) = d(b, a)$, then it is called a pseudo-metric. If d separates points —that is, if $d(a, b) = 0$ only in the case that $a = b$— but it is not necessarily symmetric, then it is called a quasi-metric. Finally, if both requirements hold —symmetry and separation—, d is called a metric (or a distance). In this case, the topology generated by the balls is Hausdorff. These notions have been already used in several applied contexts, as for example for the design of semantic computational tools ([12, 16]) or the analysis of complexity measures in theoretical computer science ([4, 5]).

3. Mathematical structures for detection of fraud in public administration and business.

In this section we introduce the general framework for understanding the fraud processes into a mathematical structure. Let Ω be a set of objects of the same class related to the representation of individuals of a system. Typically, this set is composed by vectors containing information of different type, each class in each coordinate. A vector v in this class (belonging to a subset Ω of a vector space V) is univocally associated to an individual: for example, the set Ω may be composed by invoices of a given year paid by a public administration; each vector may be given by the atributes of the invoice, for example, First coordinate= date of payment, Second coordinate= total amount paid, Third coordinate= name of the company, that is,

$$v = \big(\text{date of payment, total amount paid, name of the company}\big).$$

Let us consider now a quasi-pseudo-metric d in the set Ω. The explanation of different systematic procedures for defining it will be given in the next section. In the model it may represent the proximity of different elements

of Ω among them, and the definition must make sense for measuring the economic activity (or other kind of relevant activities) related to the process that is being analyzed. For instance, in the previous example a reasonable distance will be given by the following function. Let $v = (x_1, x_2, x_3)$, $w = (y_1, y_2, y_3) \in \Omega$. We define

$$d(v, w) = d_1(x_1, y_1) + d_2(x_2, y_2) + d_3(x_3, y_3),$$

where $d_1(x_1, y_1) = |x_1 - y_1|$, $d_2(x_2, y_2) = |x_2 - y_2|$ and $d_3(x_3, y_3) = 0$ if the invoice v was paid to the same company that the invoice w, and $d_3(x_3, y_3) = 1$ otherwise. This clearly defines a distance.

Let us explain other example with some details.

Example 3.1. *The set of objects Ω is defined by companies involved in providing services to the public administration in a given year. Each of them is represented by a vector defined by*

- *First coordinate= total amount paid to the company (in K Euros).*
- *Second coordinate= number of services provided by the company.*
- *Third coordinate= geographical location of the company (first coordinate of the position vector).*
- *Fourth coordinate= geographical location of the company (second coordinate of the position vector).*

This set would be considered a sufficient system, in the sense that it would contain enough information for detecting an anomalous behavior. We identify each company with its representing vector, that is, Ω is a subset of $\mathbb{R}^4$. We have to measure the distance among the elements that are considered here. The first obvious choice is to measure the Euclidean distance among vectors, that is if $v_1, v_2 \in \Omega$,

$$d(v_1, v_2) = \left\| v_1 - v_2 \right\|_2,$$

where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^4$. However, this option provides an information that only allows to compare companies among them, and grouping them by similarity of activity and location. A priori, it does not seem to be useful for fraud detection.

A more subtle option would be the following. Consider the seminorms

$$p_E(x_1, x_2, x_3, x_4) = \left\| (x_1, x_2, 0, 0) \right\|_2, \quad v = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4,$$

and

$$p_L(x_1, x_2, x_3, x_4) = \left\| (0, 0, x_3, x_4) \right\|_2, \quad v = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4.$$

Both of them are seminorms, and so the formulas $d_E(v_1, v_2) = p_E(v_1 - v_2)$ and $d_L(v_1, v_2) = p_L(v_1 - v_2)$ define pseudometrics ($d(v_1, v_2) = 0$, does not necessarily imply $v_1 = v_2$). The first one allows grouping companies by similar economic activity —that is, a small neigbourhood of a company/vector v contains companies with similar economic relation with the public administration. Also, a big value of $p_E(v)$ in comparison with the values of p_E of

other companies indicates a big economical activity, that would be an indication either of fraud or risk of fraud. The second one —d_L— would be used for detecting changes of names of the same company for hiding an unusual recruitment with the public administration of a single company.

Let us define now two more structures. Consider the σ-algebra $\mathcal{B}$ of Borel sets of (Ω, d) —typically, Ω will be a finite set and $\mathcal{B}$ will be the class 2^Ω of all the subsets of Ω—. Consider a Borel measure $\mu : \mathcal{B} \to \mathbb{R}^+$. On the other hand, consider also a function $\psi : \Omega \times \mathbb{R}^+ \to \mathbb{R}^+$ that is increasing with respect to the second variable. It will be considered as a radial weight associated to the radius of the balls for the metric topology.

Definition 3.2. *Let $F(\mathbb{R}^+, \mathbb{R}^+)$ be the set of real non-negative functions acting in the positive real numbers. We define the density function $\mathcal{F}$ as the function-valued map*

$$\mathcal{F} : \Omega \times \mathbb{R}^+ \to F(\mathbb{R}^+, \mathbb{R}^+)$$

given by

$$(a, \varepsilon) \mapsto \mathcal{F}(a, \varepsilon) = f_a(\varepsilon) := \frac{\mu(B_\varepsilon(a))}{\psi(a, \varepsilon)}.$$

Remark 3.3. *Let us explain a —in a sense canonical— example of this notion. Consider a finite set of companies $\Omega = \Omega_0$ in the setting of Example 3.1. Take $\mu(\cdot) = |\cdot|$ to be the counting measure on the σ-algebra of all finite subsets 2^{Ω_0}, and $\psi(a, \varepsilon) = \varepsilon^4$ for all $a \in \Omega_0$ —the power 4 for representing the magnitude of a hypervolume in a space of 4-dimensions—. The metric d is the one defined in the first part of this example. In this case,*

$$\mathcal{F}(a, \varepsilon) = f_a(\varepsilon) = \frac{|B_\varepsilon(a)|}{\varepsilon^4} = \frac{1}{\varepsilon^4} \times \Big(\textit{number of companies in } \{b \in \Omega : \|b - a\|_2 < \varepsilon\} \Big).$$

This formula is clearly defining a density-type parameter: it is given by a ratio among "number of things" in a given volume of the space and the "size" of such volume.

We are prepared now to define the main concept of this paper.

Definition 3.4. *Let $r > 0$. We define the concentration of mass (out of a neighbourhood of the element a of size r), or the local density around a, as the function $C_r : \Omega \to \mathbb{R}^+ \cup +\infty$ given by*

$$C_r(a) = \int_r^{+\infty} f_a(\varepsilon)\, d\nu(\varepsilon), \quad a \in \Omega,$$

where ν is (another countably additive) Borel measure on $(0, \infty)$.

For ν, we are thinking in a Dirac's delta of a given value $\varepsilon_0 > 0$, or Lebesgue measure $d\varepsilon$. Note that the requirement $r > 0$ is impossed to assure the convergence of the integral, at least in the canonical case explained in Remark 3.3. In the standard finite case, if d is a distance, it can be taken as

the minimum of all the pairwise distances in the set Ω not being 0, assuring in this way that $B_r(a)$ contains just an element for any $a \in \Omega$.

The central methodological idea of the present paper is that fraud detection may be considered as a systematic procedure for finding "outliers" in a quasi-pseudo-metric space. Indeed, fraud can be modelled as a *concentration of mass phenomenon: that is, elements $a \in \Omega$ are associated to processes that are "suspicious of fraud" if $C_r(a)$ has an unexpected value —that is, either "too high" or "too low" when comparing with the mean value—.* Each of these deviations can be interpreted in different terms, providing diverse figures of fraud.

It must be taken into account that special elements in the system may have high values of C_r and this situation can be considered as "normal": for instance, if there is only one company providing a given service; or, the name of the responsible of the public administration would appear in all the invoices.

Remark 3.5. *Although the way of measuring local density given in Definition 3.4 seems to be the most adequate to the original problem, other ways of measuring this magnitude would make sense. For instance, for the discrete case we can compute the supremum of the size of the balls r for which the ball contains only its center a, that is*

$$r_{\max}(a) := \sup\{r > 0 : |B_r(a)| = 1\},$$

that coincides with the minimum distance to the closer element of the space, that is

$$r_{\max}(a) = \min\{d(a, b) : b \in \Omega,\, b \neq a\}.$$

Note that in this case, a big value of $r_{\max}$ means small density.

Remark 3.6. *In the examples in this section it has been used the Euclidean norm in the finite dimensional spaces for constructing the underlying topological structure. This way of measuring the distances is easy and provides directly a metric in the set Ω. However, this is not the best option in general, and an alternate method for defining metric structures is required. The reason is that often the indexes that are naturally used for indicating the distance among elements of Ω are not metrics; in fact, they are not quasi-pseudo-metrics. Let us explain this relevant point with an example.*

Suppose that Ω is a set of person in a social net, and we have a function ϕ that "measures" the "level of familiarity" among the elements of Ω in the following way: $\phi(a, b) = 1$ if a an b are close friends, $\phi(a, b) = 2$ if a an b are friends but they meet occasionally, $\phi(a, b) = 3$ if a an b are just acquaintances, and $\phi(a, b) = 4$ if a an b never met. It may clearly happen that a is a close friend of b, b is a close friend of c, but a and c are only acquaintances; that is

$$3 = \phi(a, c) > \phi(a, b) + \phi(b, c) = 1 + 1,$$

and so the triangular inequality does not hold. This means that ϕ is not a quasi-pseudo-metric, but a natural function for measuring social distances.

We will solve this problem by defining a general rule for generation of quasi-pseudo-distances by means of the notion of proximity function, that will be introduced in the next section. As we will see there, the function ϕ above is a canonical example of such a proximity function.

4. The general scheme for the use of graph quasi-pseudo-metrics for fraud detenction

We are interested in defining a general procedure for analyzing relations inside a set Ω defined by "entities" (including persons, companies,...) using the information appearing in text documents, considering that as sets of emails, contracts, invoices and so. In a broad sense, the method follows the next steps.

1) Detection and definition of a non-ambiguous set of entities for starting the analysis. For doing this, the analyst must choose it, and a specific setting must be performed for a fixed kind of fraud. Authomatic processes can also be used: for example, semantic parsing techniques provided by the Stanford group could be applied as well as neural networks for training the searching system.

2) Definition of the matrix associated to a *proximity function.* This is a function $\phi : \Omega \times \Omega \to \mathbb{R}^+$ that describes by means of a non-negative real number a relation among the entity a and the entity b, both of them in Ω, which represent how far the individuals — "entities"— are connected as elements of the network concerning the economical/administrative activities. A small value of $\phi(a, b)$ means that both a and b can be often found as parts of the same activity/business; a big one, that there is not such a relation. Although the function is supposed to be bounded (typically, by 1), it is not assumed to be a distance. However, it may be assumed to be symmetric and $\phi(a, b) = 0$ if and only if $a = b$, and so it only fails subadditivity for being a metric; such functions are sometimes called semimetrics.

3) Definition of a distance on the set Ω by using a "sub-additive gauge" for ϕ, that is, a new function $d : \Omega \times \Omega \to \mathbb{R}^+$ that satisfies that
   a) it is a metric,
   b) and for all $a, b \in \Omega$, $d(a, b) \leq \phi(a, b)$.

   Of course, for this to be true we need a proximity function ϕ that is symmetric and separates points. In particular, $d(a, b) = 0$ if and only if $a = b$.

   We will explain later on how to define explicitly such a function d given a function ϕ. In fact, the method that we propose is the

main contribution of the present work, and has been performed in a specific way for solving the problem that we explained above and we originally faced.

4.1. **The sub-additive gauge of a proximity function ϕ.** For the construction of such a gauge, given a function ϕ with the requirements explained above we use a path-distance-like definition by considering a path distance in the global graph Ω, in which all the vertices are assumed to be connected —a complete graph—. We analyze the properties of such a metric, and several procedures for defining the initial proximity matrix $(\phi(a,b))_{(a,b)\in\Omega\times\Omega}$.

In Section 15.1 in [2, p.276], a weighted path metric for a connected graph is defined as follows. If e is an edge of the graph, write $w(e)$ for the value of a positive weight; w is so assumed to be a (real positive) function acting in the set of edges of the graph. The path distance d_G among to vertices a and b of the graph is given by

$$d_G(a,b) := \inf\Big\{ \sum_{e_i\in P} w(e_i)\Big\},$$

where the infimum is computed over all paths $P = \{e_i : i \in I_P\}$ that allow to go from a to b.

We are interested in a construction that is similar to (but not equal to) a weighted path metric defined on the set of all the vertices of a connected graph. In our case all couples of elements of the set are assumed to be directly connected by an edge, that is, the graph is complete. Consider a non-increasing sequence $W := (W_i)_{i=1}^{\infty}$ of positive real numbers, all of them less or equal to one. Given two points $a, b \in \Omega$, we define

(4.1)

$$d_\phi(a,b) = \inf\Big\{W_1\phi(a,b),\, \inf\Big\{W_2\big(\phi(a,c)+\phi(c,b)\big) : a\neq c\neq b,\, c\in\Omega\Big\}, ...$$

$$..., \inf\Big\{W_n\big(\phi(a,c_1)+\sum_{i=1}^{n-2}\phi(c_i,c_{i+1})+\phi(c_{n-1},b)\big),\, a, c_1, ..., c_{n-1}, b \,\textit{pairwise}\neq\Big\}, ...\Big\}.$$

Lemma 4.1. *The function $d_\phi(a,b)$ defined as above is a metric on Ω.*

We will use the particular case given by the weights sequence $W = (1/i)_{i=1}^{\infty}$, and so the distance function is defined by

$$d_\phi(a,b) = \inf\Big\{\phi(a,b),\ \inf\Big\{\frac{\phi(a,c)+\phi(c,b)}{2} : a\neq c\neq b\Big\}, \tag{4.2}$$

$$\inf\Big\{\frac{\phi(a,c_1)+\phi(c_1,c_2)+\phi(c_2,b)}{3} : a\neq c_1\neq b,\, a\neq c_2\neq b,\, c_1\neq c_2\neq b\Big\},$$

$$..., \inf\Big\{\frac{\phi(a,c_1)+\sum_{i=1}^{n-2}\phi(c_i,c_{i+1})+\phi(c_{n-1},b)}{n},\, a, c_1, ..., c_{n-1}, b \ \textit{pairwise}\neq\Big\}...\Big\}.$$

Suppose now that the set Ω is finite, $|\Omega| = n \in \mathbb{N}$. Then we can represent ϕ by means of the matrix of its range, that is,

$$\Phi = \begin{bmatrix} \phi(a_1, a_1) & \cdots & \phi(a_1, a_n) \\ \vdots & \ddots & \vdots \\ \phi(a_1, a_n) & \cdots & \phi(a_n, a_n) \end{bmatrix} = \begin{bmatrix} 0 & \cdots & \phi(a_1, a_n) \\ \vdots & \ddots & \vdots \\ \phi(a_1, a_n) & \cdots & 0 \end{bmatrix}.$$

We will call the matrix Φ the proximity matrix associated to ϕ.

Example 4.2. *Let us give some examples of proximity matrices.*

1) *The first easy example is given by the metric defined in Example 3.1. In this case, the proximity function is just the Euclidean metric; that is, $\phi = d$. Consequently, the corresponding proximity matrix Φ is a metric matrix.*

2) *Let us show two examples of such construction that are not defined as in Example 3.1. For the first one, consider Ω to be a group of individuals that are involved in a business, and the only information we have about it is written in a set M of documents. We want to perform an analysis of the influence of the individuals in Ω in the business. In order to do this and as a first approximation, we consider the following proximity function.*

   Given $a, b \in \Omega$, take the number of times $M_{a,b}$ that a appears together with b in a document. Define

$$\phi_M(a, b) = \frac{M - M_{a,b}}{M}, \quad a, b \in \Omega.$$

   Another step is needed to clean the matrix in case there are two different individuals in Ω such that they coincide in all the documents. In this case, they must be considered just as only one vertex of the corresponding complete graph. Note also that $M_{a,b} = 1$ indicates that a and b are not appearing together in any document. However, this does not mean that the distance among them has necessarily the maximum value. The reason is that it may happen that a appear in a document with c, and c with b. Using an adequate formula for d_ϕ —for example the one given by equation (4.2) with weights $W_i = 1/i$ as in the particular case given above—, we can easily see that $d_{\phi_M}(a, b) < 1$.

3) *Let us show now a different way of defining a proximity function for the same problem. Let $N = |\Omega|$ and assume that there are M documents. Take the $N \times M$-matrix C of all the counts $C(a, m)$ of the times that the individual a appears in document m. Normalize all the vectors appearing in the rows and compute $A = C \cdot C^T$. It is an $N \times N$-matrix giving the "cosinus" between elements of Ω. If the element $A(a, b)$ is near to one, this means that they appear in almost the same documents; if it is near to 0, it means that they are not appearing together.*

Take the $N \times N$-matrix $\mathbb{I}_{N\times N}$ in which all the coefficients are equal to 1, and compute Φ as

$$\Phi = \mathbb{I}_{N\times N} - A.$$

It gives a different proximity matrix. Actually, this construction is the one that we will consider as standard, and will be developed with some detail in the next section. As we will show there, it can be interesting to combine different metrics, some/all of them defined by proximity functions.

4.2. **Proximity functions defined by means of correlation matrices: the standard model.** Let us fix a canonical version of the formulae explained in the previous parts of this section. It follows the lines of Example 4.2, 3).

A. Take a set of N entities Ω and a set of M properties —quantifiable by means of positive real numbers— associated to each element $a \in \Omega$. Construct the set of N vectors v_a each of them containing the numerical value of the properties of a fixed $a \in \Omega$.

B. Take the matrix C defined in a way that each row is such a vector v_a after normalization, that is $v_a/\|v_a\|_2$ (we use the Euclidean norm for normalizing).

C. Consider the correlation matrix $A = C \cdot C^T$ and take as proximity matrix $\Phi = \mathbb{I}_{N\times N} - A$. Note that it is symmetric.

D. Use formula (4.2) for defining the pseudo-metric d_ϕ.

E. The final distance for performing the analysis is given by the formula

$$d(a,b) = k \cdot \frac{\|v_a - v_b\|_2}{\max\{\|v_c\|_2 : c \in \Omega\}} + d_\phi(a,b), \quad a,b \in \Omega.$$

Here, $k > 0$ is a parameter for balancing both components of the distance. The first one allows to measure the size of the vectors, for detecting the case that one of its values has unexpected values (for example, a big ammount of money appearing in any coordinate of v_a). The second one provides information about the coincidence of coordinates, measuring it using the "cosinus distance".

Let us explain a complete example using this method.

Example 4.3. *Consider 4 companies, a_i, $i = 1, ..., 4$, which have been hired by a public administration (PA) for doing similar services. We are interested in analyzing if there is any irregular behavior in any of them in 2017. We will show two problems and the models that correspond to each of them. We only have information regarding total amount of money that PA paid to each of them in 2017 and the number of contracts with each company.*

(1) *Suppose that we want to analyze if the total amount of money x_i, $i = 1, ..., 4$, got by each company a_i is either equally distributed among all the companies or we can find different patterns regarding that to divide the companies in two groups. Let us use the procedure explained above. The "vector of properties" v_i for each company a_i contains just a coordinate, x_i. The values (in thousands of euros) are $x_1 = 4$, $x_2 = 2$, $x_3 = 2$, and $x_4 = 1$. The "Euclidean part" of the pseudo-distance is then given by*

$$d_E(a_i, a_j) := |x_i - x_j| / \max\{4, 2, 1\} = |x_i - x_j|/4, \quad i, j = 1, ..., 4.$$

The part of the pseudo-metric given by the correlation matrix is given (after normalization) by the trivial formula

$$\mathbb{I} - A = \mathbb{I} - C \cdot C^T = \mathbb{I} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus, the final pseudo-metric contains only the Euclidean component, and is represented by the matrix

$$d = d_E = \begin{bmatrix} 0 & 1/2 & 1/2 & 3/4 \\ 1/2 & 0 & 0 & 1/4 \\ 1/2 & 0 & 0 & 1/4 \\ 3/4 & 1/4 & 1/4 & 0 \end{bmatrix}.$$

This pseudo-metric allow to separate the set of the four companies in two disjoint balls; indeed, for example for $\varepsilon = 3/8$, we have

$$B_{3/8}(a_1) = \{a_1\}, \quad \textit{and} \quad B_{3/8}(a_2) = \{a_2, a_3, a_4\}.$$

The local density in both companies, computed as the ratio among the number of elements in each ball and the radius of the —one dimensional— balls give the values for $\varepsilon = 3/8$,

$$Density_{3/8}(a_1) = |B_{3/8}(a_1)|/(3/8) = 8/3$$

and

$$\begin{aligned} Density_{3/8}(a_2) &= |B_{3/8}(a_2)|/(3/8) = 8 \\ &= Density_{3/8}(a_3) = Density_{3/8}(a_4). \end{aligned}$$

Therefore, it can be easily seen that there is a concentration of mass around a_2, and a_1 is surrounded by an area of low density. In this sense, it can be established that a_1 is an isolated point in terms of density, so it is suspicious of receiving an special treatment from PA. Of course, this fits with the fact that a_1 got the biggest amount of money in the contracts among all companies, and the difference with the other ones seems to be meaningful.

This example is very easy, and present it just for showing how the formalism works for almost trivial cases. No analyst needs to use this procedure for obtaining this conclusion.

(2) *Suppose now that we want to analyze a different aspect of the same problem, and we include in the investigation the number of contracts of each of the companies with PA in 2017 given the total amounts of money presented in (1). Now we consider two properties —two-coordinates vectors— for each company: the first coordinate is the amount of money in (1), and the second one if the number of contracts. We have the following values:* $a_1 = (4,3)$, $a_2 = (2,1)$, $a_3 = (2,2)$, *and* $a_4 = (1,1)$. *For the aim of simplicity, we identify the companies* a_i *with its two-coordinates property vectors* (x_i, y_i), $i, j = 1, ..., 4$.

As in the previous case, we have that the Eclidean part of the distance is given by the Euclidean norm divided by the maximum of the norms, that is, taking into account that

$$\|a_1\| = 5, \quad \|a_2\| = \sqrt{5}, \quad \|a_3\| = 2\sqrt{2}, \quad \|a_4\| = \sqrt{2},$$

we get

$$d_E(a_i, a_j) = \|(x_i, y_i) - (x_j, y_j)\|_2 / \max\{\|a_i\|_2\} = \frac{\|(x_i, y_i) - (x_j, y_j)\|_2}{5}.$$

This gives the metric matrix

$$D_E = \begin{bmatrix} 0 & \frac{2\sqrt{2}}{5} & \frac{\sqrt{5}}{5} & \frac{\sqrt{13}}{5} \\ \frac{2\sqrt{2}}{5} & 0 & \frac{1}{5} & \frac{1}{5} \\ \frac{\sqrt{5}}{5} & \frac{1}{5} & 0 & \frac{\sqrt{2}}{5} \\ \frac{\sqrt{13}}{5} & \frac{1}{5} & \frac{\sqrt{2}}{5} & 0 \end{bmatrix} \sim \begin{bmatrix} 0 & 0.566 & 0.447 & 0.721 \\ 0.566 & 0 & 0.2 & 0.2 \\ 0.447 & 0.2 & 0 & 0.283 \\ 0.721 & 0.2 & 0.283 & 0 \end{bmatrix}.$$

On the other hand, the proximity matrix given by the correlation matrix is in this case meaningful. Indeed,

$$\mathbb{I} - A = \mathbb{I} - C \cdot C^T$$

$$= \mathbb{I} - \begin{bmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} \frac{4}{5} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{3}{5} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \sim \begin{bmatrix} 0 & 0.016 & 0.010 & 0.010 \\ 0.016 & 0 & 0.051 & 0.051 \\ 0.01 & 0.051 & 0 & 0 \\ 0.01 & 0.051 & 0 & 0 \end{bmatrix}.$$

This is not a pseudo-metric matrix: note for example tbat

$$0.051 = \phi(a_2, a_3) > \phi(a_2, a_1) + \phi(a_1, a_3) = 0.016 + 0.010.$$

In order to provide a pseudo-metric d_ϕ *preserving as much as possible the size of the coefficients of the original proximity matrix, we use (4.1) with all weights equal to one, that is* $W_i = 1$, $i = 1, ..., 4$. *We obtain the pseudo-metric matrix*

$$d_\phi \sim \begin{bmatrix} 0 & 0.016 & 0.010 & 0.010 \\ 0.016 & 0 & 0.026 & 0.026 \\ 0.01 & 0.026 & 0 & 0 \\ 0.01 & 0.026 & 0 & 0 \end{bmatrix}.$$

The final distance matrix is then given by

$$D = \lambda\, D_E + d_\phi.$$

This can be used for the analysis in the same way that was made in (1). However, if we look at the two matrizes as separately , we get more information about the problem.

(i) *Using d_E, we find again a similar conclusion that the one we get in (1): the first company is the only element in the ball of radius $\varepsilon = 0.4$. However, a ball of the same size $\varepsilon = 0.4$ centered in a_2 contains the rest of the elements, a_2, a_3 and a_4. The same argument that was used in (1) using $Density_{0.4}$ provides the same conclusion that in (1).*

(ii) *The second matrix —associated to $d\phi$— centers the attention in other element. In this case, the ball $B_{0.015}(a_2)$ only contains a_2. However, the ball $B_{0.015}(a_1)$ contains a_1, a_3 and a_4. The density arround a_2 is then smaller than density around a_1, a_3 and a_4. This means that a_2 would be suspicious of getting a special treatment, or at least that its hiring pattern is not the same. Note that this pseudo-metric measures the proportion between amount of money and number of contracts. The result shows that the company a_2 is not following the same proportion, what means that the money associated to each contract is different. This may be just by chance, but also would indicate that there is someone interested in manimulating the standard hiring procedure, and so it would be suspicious of fraud.*

5. Final remarks: application to detection of irregular behavior of elements in a network

In this section and to finish the paper, we give some open ideas for applying the ideas developed in the paper. We can consider the following problems to solve as application of our metric structure.

- The first and canonical one: *given an entity $a \in \Omega$, find the rest of the elements of Ω that are near* (distance less than $\varepsilon > 0$). This is the first step of the neighbourhood analysis that allow to compute a density map for searching anomalous behaviours. But is also provides a primary information, providing the entities that are close to a given one a with respect to the criterium used for the construction of the proximity function.
- *Degree of dependence of the "graph distance" on a single element $a \in \Omega$*: this is the norm of the difference of the submatrix D_a that is obtained by eliminating the row and column associated to a in the distance matrix D, and the distance matrix $D(-a)$ that is computed when the set considered is $\Omega \setminus \{a\}$ instead of Ω. If the value is small, this means that the element a is not relevant for the graph, it is not

really connected or it is not giving easy paths for other entities to be connected.

- *Optimization*: given a point $a \in \Omega$ and a subset $S \subset \Omega$, find the element(s) b in S such that $d_\phi(a, b)$ attains its minimum.
- *A singular-values-type method for determining the classes of equivalence of entities in the space having the same behavior*, in the sense that they appear in the same documents. We use the matrix A defined in Example 4.2, 3). Consider the individuals a_1 to a_n and suppose they are appearing in the same documents, and they are the only ones appearing in these documents. Then we can write the vectors of the matrix A corresponding to these individuals as

$$1/\sqrt{n}\,(1, 1, ..., 1, 0, ...0),$$

where the coefficient equal to 1 appear in the n first positions. On the other hand, the other individuals have coefficients that are all of them 0 in the first n positions (check that, this is a consequence of the construction of A based in the fact that they are appearing in disjoint documents). When the corresponding submatrix is diagonalized, we obtain an eigenvalue that is not zero and other one that is 0, that has multiplicity $n-1$. Therefore, there is only one document-appearing behavior, the rest only repeat the behavior of the first individual. Of course, we rarely are going to find this pure behavior, and so we use the ideas of the singular vales method for giving the "almost zero" version.

For doing this, compute the eigenvalues of the matrix $\{\lambda_i : 1 \leq i \leq m\}$. Fix $\varepsilon > 0$, and take the subspace S_ε generated by the eigenvectors associated to the eigenvalues $\lambda_i < \varepsilon$. Write the equation $A = U^T \Delta U$ (U is the matrix of change of basis) and compute the vectors $v_a = (0, \cdots, 1, \cdots 0)$ representing the elements $a \in \Omega$ that satisfy that Uv_a is in S_ε. This is the set that can be eliminated from the original set Ω, since they have an equivalent behavior that any of the ones for which $\lambda_i \geq \varepsilon$.

References

[1] Bolton R.J and and Hand D.J. "Unsupervised Profiling Methods for Fraud Detection". (unpublished, available in Google Scholar).

[2] M.M. Deza, and E. Deza. "Encyclopedia of distances." Springer, Berlin Heidelberg, 2009.

[3] Jack Dorminey, A. Scott Fleming, Mary-Jo Kranacher,Richard A. Riley, Jr. "The Evolution of Fraud Theory." Issues in Accounting Education: May (2012), Vol. 27, No. 2, pp. 555–579.

[4] García-Raffi, L. M., S. Romaguera, and E. A. Sánchez-Pérez. "Sequence spaces and asymmetric norms in the theory of computational complexity." Mathematical and computer modelling 36.1-2 (2002): 1-11.

[5] Garca-Raffi, L. M., S. Romaguera, and M. P. Schellekens. "Applications of the complexity space to the general probabilistic divide and conquer algorithms." Journal of Mathematical Analysis and Applications 348.1 (2008): 346-355.

[6] Künzi, Hans-Peter A. "Quasi-uniform spaceseleven years later." In Topology Proc, vol. 18, pp. 143-171. 1993.
[7] Mansor, "Fraud Triangle Theory and Fraud Diamond Theory. Understanding the Convergent and Divergent For Future Research."
[8] Theodore J. Mock, Rajendra P. Srivastava, and Arnold M. Wright. "Fraud Risk Assessment Using the Fraud Risk Model as a Decision Aid". Journal of Emerging Technologies in Accounting. Spring (2017), Vol. 14, No. 1, pp. 37–56.
[9] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature." Decision Support Systems, (2011) 50(3), 559-569. Some concrete techniques and applications of data mining procedures are shown.
[10] Reilly, Ivan L., P. V. Subrahmanyam, and M. K. Vamanamurthy. "Cauchy sequences in quasi-pseudo-metric spaces." Monatshefte fr Mathematik 93.2 (1982): 127-140.
[11] Richhariya, P. and Singh P.K. "A Survey on Financial Fraud Detection Methodologies." International Journal of Computer Applications (2012) 45, 22 pp.975 – 1007.
[12] Romaguera, Salvador, Michel P. Schellekens, and Oscar Valero. "The complexity space of partial functions: a connection between complexity analysis and denotational semantics." International Journal of Computer Mathematics 88.9 (2011): 1819-1829.
[13] Szárnyas, G., Koovár, Z., Salánki, A., Varró, D. "Towards the Characterization of Realistic Models: Evaluation of Multidisciplinary Graph Metrics". In Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems (2016), pp. 87–94.
[14] Gregory M. Trompeter, Tina D. Carpenter, Naman Desai, Keith L. Jones, and Richard A. Riley, Jr. "A Synthesis of Fraud-Related Research". AUDITING: A Journal of Practice and Theory (2013), Vol. 32, No. Supplement 1, pp. 287–321.
[15] Gregory M. Trompeter, Tina D. Carpenter, Keith L. Jones, and Richard A. Riley, Jr. "Insights for Research and Practice: What We Learn about Fraud from Other Disciplines". Accounting Horizons December 2014, Vol. 28, No. 4, pp. 769-804.
[16] Valero, Oscar, Jesus Rodriguez-Lopez, and Salvador Romaguera. "Denotational semantics for programming languages, balanced quasi-metrics and fixed points (SICMMSE2006)." International Journal of Computer Mathematics 85.03-04 (2008): 623-630.
[17] T. Jeffrey Wilks and Mark F. Zimbelman. "Using Game Theory and Strategic Reasoning Concepts to Prevent and Detect Fraud". Accounting Horizons (2004), Vol. 18, No. 3, pp. 173–184.
[18] Zhao, J., Lau, R. Y., Zhang, W., Zhang, K., Chen, X., Tang, D. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision Support Systems 86 (2016): 109–121.

E-mail address: jmcalabu@upv.es, anfersa@upv.es, lmgarcia@upv.es, easancpe@upv.es