*R.I.T.I. nº 8 Septiembre-Diciembre 2018*

Revista Internacional TRANSPARENCIA E INTEGRIDAD

# DETECCIÓN Y PREVENCIÓN DE LAS MALAS PRÁCTICAS Y LA CORRUPCIÓN DESDE LA PERSPECTIVA DE LAS MATEMÁTICAS (*)

***J.M. Calabuig[1], H. Falciani, A. Ferrer-Sapena[1], L.M. García-Raffi[1], E. Raso[2], I. Sánchez del Toro[2] y E.A. Sánchez-Pérez[1]***

**RESUMEN**

En este artículo se hace referencia a la posible utilización de herramientas matemáticas en la prevención y detección del fraude.

La detección del fraude se plantea como uno de los mayores desafíos de las administraciones públicas. Más allá de los procesos de inspección y auditoría, la sociedad no sólo exige de sus administradores la persecución del delito sino también su prevención, para evitar, no sólo el daño económico, sino también el alto precio que en términos de carencia de servicios acaban pagando los ciudadanos, quienes previamente han cumplido con sus obligaciones contribuyendo a las arcas públicas.

**ABSTRACT**

In this article reference is made to the possible use of mathematical tools in the prevention and detection of fraud.

The detection of fraud is considered one of the biggest challenges for public administrations. Beyond the processes of inspection and auditing, society not only requires of its administrators the prosecution of the crime but also its prevention, to avoid not only the economic damage, but also the high price that in terms of lack of services end up paying citizens, who have previously fulfilled their obligations by contributing to public coffers.

## 1. INTRODUCCIÓN

La detección del fraude se plantea como uno de los mayores desafíos de las administraciones públicas. Más allá de los procesos de inspección y auditoría, la sociedad no sólo exige de sus administradores la persecución del delito sino también su prevención, para evitar, no sólo el daño económico, sino también el alto precio que en términos de carencia de servicios acaban pagando los ciudadanos, quienes previamente han cumplido con sus obligaciones contribuyendo a las arcas públicas. Es por tanto no sólo la detección sino también la prevención, uno de los caballos de batalla principales de las administraciones públicas porque posiblemente, antes de la existencia de un fraude o una práctica corrupta, existen elementos indicativos de que se puede producir ésta, indicios en forma de mala praxis (económica, administrativa, de gestión, …) o como una anomalía, un acto (económico, administrativo,…) que está fuera de la norma en algún sentido.

Grupo MADϕ
[1] Instituto Universitario de Matemática Pura y Aplicada. Universitat Politècnica de València
[2] KPI Risk, Ethics & Complience

La teoría más importante sobre el fraude es la llamada "Teoría del Triángulo del Fraude" ("Fraud Triangle Theory"), que se extiende en términos similares a la llamada Teoría del Diamante del Fraude ("Fraud Diamond Theory", Mansor 2015), y que aporta una justificación en términos psicológicos y sociológicos de las actividades fraudulentas. La Teoría del Triángulo del Fraude se basa en la existencia de una racionalización –la justificación personal de acciones deshonestas–, una presión –motivación que empuja hacia el fraude– y una oportunidad –posibilidad real de cometer una acción de fraude sin ser "pillado"–. Estos elementos permiten entender la motivación, la posibilidad y otras circunstancias sociológicas y psicológicas que inducen a la aparición de un acto de fraude en un contexto dado.

La misma práctica que supone la persecución del fraude y la corrupción, ha desarrollado nuevos mecanismos de forma acorde a los tiempos. Hemos pasado rápidamente de una era del "papel" a la era "digital" y nos encontramos rápidamente en transición hacia la era del "Big Data", en la que las empresas y administraciones manejan una gran cantidad de datos. Son muchos los factores que influyen en las nuevas formas de comisión de delitos, pero posiblemente se pueda decir sin temor a equivocarse que los métodos de análisis no han ido acordes a la cantidad de datos de la que se dispone. Parece que las empresas están respondiendo más ágilmente al desafío que esta nueva situación les plantea (el fraude no sólo se comete en la administración pública, sino también en las grandes corporaciones) lo que exige de los administradores públicos nuevas formas y métodos a la hora de perseguir y prevenir estos delitos, basados en una buena gestión de cantidades importantes de datos, así como de proveerse de las herramientas adecuadas y los equipos de personas –necesariamente interdisciplinares– para un análisis profundo de los mismos, y la búsqueda en dichos datos de aquellos patrones que marcan el camino del fraude.

Así pues, no sólo será necesario, sino esencial, que por una parte la administración disponga de una base de datos lo más completa posible de aquellas empresas y particulares con los que se relaciona económicamente en todas y cada una de sus formas (facturación, licitación,…), incluyendo no sólo los datos que recoge la propia administración en sus expedientes sino otros existentes en otras bases de datos (LibreBORME, Registro Mercantil,…). Y por otro lado, de aquellos algoritmos que permitan la detección de ciertos elementos relacionados con la mala praxis o con anomalías. La elaboración de algoritmos contiene dos elementos fundamentales. Uno de ellos, de carácter técnico y muy importante, relacionado con el tratamiento de la información y su implementación en sistemas automáticos de detección o alerta. El otro, de carácter más básico, relacionado con ideas matemáticas que nos permitan establecer acciones sobre los datos y, sobre las cuales se puedan elaborar algoritmos. En este artículo pretendemos abordar esta última faceta.

Cuando se diseñan procedimientos para el análisis de datos, en general resulta muy útil disponer de un conocimiento sobre la estructura de los mismos. Este conocimiento puede obtenerse de forma genérica, o bien a través procedimientos que caractericen dichos datos a partir de una muestra de ellos (y dentro de esta categoría se englobarían por ejemplo aquellos algoritmos basados en redes neuronales), o también estableciendo un modelo matemático a priori de los datos, su distribución, sus relaciones, etc. Resulta evidente que la eficiencia de los algoritmos que diseñemos para analizar los datos (y por lo tanto para nuestro propósito de detectar anomalías en los mismos) dependerá en gran medida de cuanto de bueno sea nuestro modelo de los datos. Sin embargo, la detección del fraude se revela como un problema más complejo. La recopilación de esos datos, no sólo se produce en las propias transacciones, sino que está dispersa en muchas ocasiones en otro tipo de documentos como puedan ser correos electrónicos, páginas webs, tablas tipo EXCEL,… Deberemos pues de disponer de herramientas que nos permitan recopilar toda una serie de datos que "gravitan" en torno a los actos

administrativos. Una administración eficiente en la prevención, detección y lucha contra el fraude debe de disponer de dichas herramientas (vía aplicaciones informáticas como pueden ser Logstash, Elasticasearch, analizadores semánticos,…).

Hay diversas técnicas matemáticas que han mostrado su utilidad en lo referente a la detección del fraude como puedan ser las técnicas basadas en el "data mining" o el aprendizaje automático ("machine learning"), entre otras (no pretendemos aquí, ni mucho menos, hacer una lista exhaustiva o erudita al respecto), pero no vamos a hacer mención de ellas en este artículo. Nuestro enfoque será un tanto distinto y el objetivo de este artículo es dar unas pinceladas generales de nuestras ideas al respecto de cómo puede ser abordado este problema.

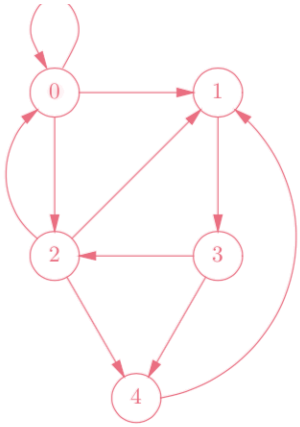
En la siguiente sección abordaremos un aspecto que nos parece importante y que hace referencia a cómo organizar los datos. En particular hablaremos de las bases de datos basadas en grafos. Ésta ha sido una solución técnica adoptada por muchas grandes compañías (como Orange, IBM, Ebay, AirBNB, entre otros) para manejar de forma eficiente grandes cantidades de datos, no solo pudiendo producir agregados de datos de estructura muy diversa, sino adoptando la potencia de la teoría de grafos para la realización de búsquedas efectivas dentro de la base.

Es evidente que ésta no es la única opción y que ello puede ser motivo de controversia entre los técnicos. Sin embargo, nosotros optaremos aquí por esta propuesta porque nos parece interesante para poder introducir otros conceptos matemáticos que creemos claves a la hora de la prevención y detección del fraude: la introducción de formas generalizadas de distancia entre datos. De esa manera, si somos capaces de definir cuando dos datos están próximos o no lo están, podremos definir herramientas de detección y alerta que sean lo suficientemente flexibles y generales como para tratar con los datos de hoy y los de mañana, entendiendo este control como una actividad dinámica. Un problema, el de la flexibilidad en la estructura de los algoritmos y en el tamaño y características de las entradas, que se hace acuciante a medida que nuestro mundo digital se complica día a día. Esa parte constituirá la Sección tercera de este artículo. Finalizaremos el mismo con unas conclusiones.

## 2. BASES DE DATOS BASADAS EN GRAFOS

Un grafo se puede considerar una relación binaria que se visualiza a través de nodos (en lo que nos atañe, datos, nombre de una empresa, caja por la que fue pagada, CIF, etc.) y las relaciones que existen entre ellos a través de líneas (aristas) o flechas si tienen dirección (arcos).

*Figura 1. Ejemplo de grafo formado por cinco nodos en el cual las relaciones entre ellos están establecidas mediante arcos*



En general los grafos permiten estructurar los datos de forma flexible y rica en las relaciones ya que podemos dotar de propiedades tanto a los nodos como a las aristas que los conectan, además de establecer relaciones de precedencia más generales que en listas y árboles, otros dos modelos muy comunes de estructuración de datos que son casos particulares de los grafos (véase por ejemplo Aho 2013). En la Figura 1 tenemos un ejemplo de grafo. En ella vemos un conjunto de nodos formado por $N = \{0,1,2,3,4\}$ y un conjunto de relaciones entre nodos dada por

$$A = \{(0,0), (0,1), (0,2), (1,3), (2,0), (2,1), (2,4), (3,2), (3,4), (4,1)\}.$$

En este caso no hay aristas (relaciones que van en los dos sentidos) sino que todos los nodos están conectados entre sí por arcos, en el sentido en que indica la flecha. Por ejemplo, del nodo 0 se puede ir al nodo 1, pero no a la inversa. Sin embargo, sí que hay un camino que relaciona el nodo 1 con el nodo 0 a través de la ruta $\{(0,1), (1,3), (3,2), (2,0)\}$. También podemos observar que los nodos $\{1,2,3\}$ forman un ciclo a través de los arcos $\{(1,3), (3,2), (2,1)\}$, que es un subconjunto del conjunto *A*, es decir, son elementos de *A*.

*Figura 2. Matriz de adyacencia del grafo de la Figura 1*

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 |
| 2 | 1 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 | 0 | 1 |
| 4 | 0 | 1 | 0 | 0 | 0 |

Cuando hablamos de un grafo nos referimos pues tanto a sus nodos como a los arcos o aristas que los conectan, es decir, a la pareja $G = (N, A)$. Pero a la hora de representar datos utilizando grafos, vamos a querer incluir otra información. Por ejemplo una factura, que puede estar representada por un nodo, puede estar en conexión con una entidad, por ejemplo una administración del estado, representada por otro nodo, siendo la relación entre ambas ”fue pagada por…”; pero puede interesarnos incluir otros datos como como la ubicación geográfica y el CIF de la empresa prestadora del servicio, la fechas de emisión y pago de la factura, el número de expediente o número de registro de la transacción, etc. Hablamos pues de estructuras más generales como hipergrafos y multigrafos, pero que en esencia hacen lo mismo: relacionan datos o conjuntos de datos a través de una o más propiedades. De aquí en adelante hablaremos de aristas y nodos en general sin entrar en muchos más detalles, puesto que no es necesario para las ideas que pretendemos desarrollar en este artículo.

En general, un grafo se puede utilizar para organizar datos de diferente naturaleza. Por ejemplo:

1) Un elemento que puede ser una factura, una transacción, una entidad, un mensaje de correo electrónico, …
2) Una serie de actores que establecen relaciones, como emisor y receptor de una factura, pagador y pagado.
3) Una dimensión temporal, que en el caso que nos ocupa puede venir dada por las fechas de publicación de un concurso, emisión de una factura, pago de una factura,…
4) Unas características o atributos que les pueden ser asignados tanto a nodos como aristas, como números identificativos, geolocalización,…

Las bases de datos basadas en grafos pertenecen a la categoría de las llamadas bases de datos NoSQL (del inglés Not Only SQL).

Pero ¿por qué son tan interesantes los grafos a la hora de estructurar datos? Los grafos tiene su origen en un insigne matemático, Leonard Euler y un conocido problema, el "problema de los puentes de Königsberg"; el trabajo que publicó al respecto se considera como el primero de Teoría de Grafos. La ciudad de Königsberg, en Prusia Oriental, está atravesada por un río, el Pregel, que se bifurca para rodear con sus brazos a la isla Kneiphof, dividiendo el terreno en cuatro regiones distintas que estaban unidas mediante siete puentes. El problema, formulado en el siglo XVIII y que resolvió Euler, consistía en encontrar un recorrido para cruzar a pie toda la ciudad, pasando sólo una vez por cada uno de los puentes, y regresando al mismo punto de inicio.

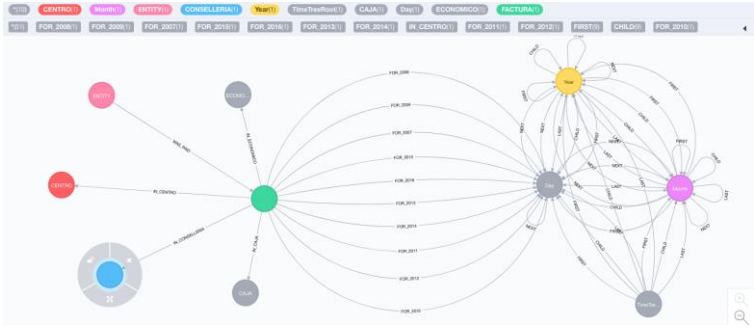
La respuesta era negativa y Euler llegó a esa conclusión utilizando grafos; en su honor, cuando la respuesta es positiva y existe tal recorrido (es decir, un recorrido que empiece y acaba en un mismo nodo y pasando una sola vez por otros nodos), se le llama *tour euleriano*. La potencialidad de los grafos se vio reforzada durante la II Guerra Mundial por su aplicación a problemas de abastecimiento y logística. La llegada de los ordenadores establecería un desarrollo rápido tanto de la teoría como de las aplicaciones.

Uno de los problemas clásicos con numerosas aplicaciones es el "problema del viajante". Su enunciado es simple: dada una lista de ciudades (nodos) y las distancias entre cada par de ellas (aristas), ¿cuál es la ruta más corta posible que visita cada ciudad exactamente una vez y al finalizar regresa a la ciudad de origen? Este problema, que se enmarca en una teoría matemática conocida como optimización combinatoria, ha sido el centro de muchos esfuerzos para desarrollar algoritmos muy eficientes que busquen recorridos por el grafo de longitud mínima (Christofides 1975).

Y es precisamente por lo que hemos traído a colación estos problemas en este artículo. Existen algoritmos muy eficientes que buscan recorridos (*tours)* en un grafo. Una base de datos orientada a grafos (BDOG) utiliza los nodos y las aristas de un grafo para representar datos y relaciones, de tal manera que la teoría de grafos puede ser usada para recorrer la base de datos.

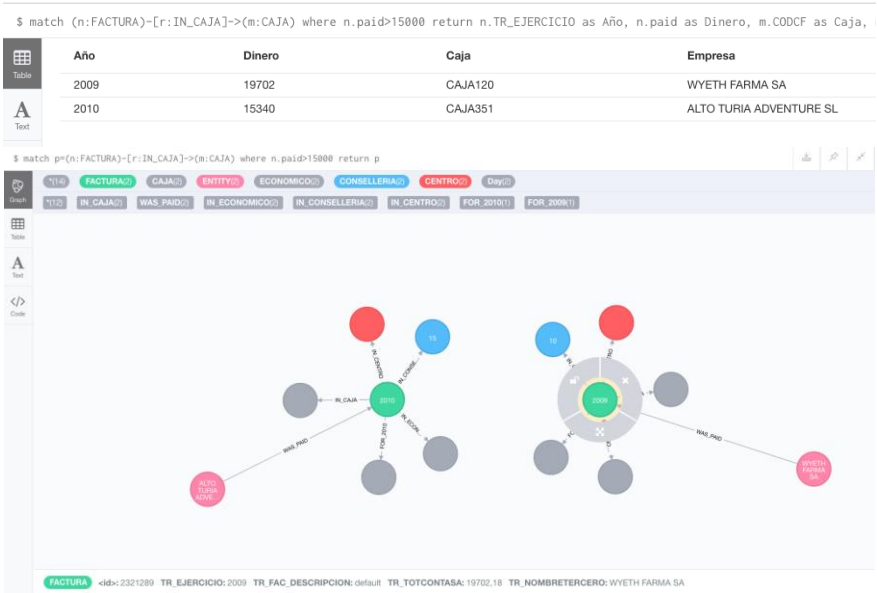
Entre las herramientas informáticas utilizadas en el análisis de las bases de datos basados en grafos está el conocido Neo4j. Este programa ha sido utilizado en la detección de las grandes tramas de corrupción (desde la lista Falciani hasta los recientes "Paradise Papers" pasando por los "Papeles de Panamá"). Terminamos esta sección presentando un ejemplo concreto de aplicación de esta herramienta a partir de los datos obtenidos en la página web de *Dades Obertes* de la *Conselleria de Transparència, Responsabilitat social, Participació i Cooperació* de la Generalitat Valenciana. La Figura 3 muestra el esquema de esta base (constituida por 3167167 nodos y 16946348 relaciones correspondientes a los datos de facturación por caja fija de la Generalitat Valenciana entre 2007 y 2016).

*Figura 3, Esquema de la base de grafos creada con Neo4j*



Neo4j tiene un lenguaje propio de búsqueda (llamado *Cypher*) que permite hacer búsquedas a la base en función de múltiples criterios. Como ejemplo en la Figura 4 se muestran los resultados obtenidos al buscar empresas cuya facturación haya superado los 15000€.

*Figura 4. Resultado obtenido al buscar empresas con facturas por un importe superior a los 15.000 € mediante Neo4j*




## 3. EL CONCEPTO DE DISTANCIA Y LAS MÉTRICAS

Resulta bastante intuitivo lo que representa una métrica. Todo el mundo entiende perfectamente lo que es la distancia entre dos puntos, al menos si se supone una geometría plana. Las matemáticas tratan de modelar y generalizar este concepto a todas luces "natural" de dos maneras. La primera, construyendo estructuras donde un punto no tiene por qué representar un punto del espacio (sino cualquier otro dato) e introduciendo una función que nos mida la distancia entre puntos/datos. Esto es ni más ni menos una métrica, una función que nos mide la distancia entre puntos/datos. Cuando utilizamos una aplicación en nuestro móvil para encontrar un lugar y le decimos que nos indique el camino, la aplicación normalmente nos calcula la distancia en metros que hay del punto donde estamos al punto de destino. Pero no es la única medida de la distancia que nos ofrece: haciendo una suposición sobre nuestra velocidad caminando o en coche, nos indica la distancia en tiempo también.

Ambas distancias están relacionadas por la velocidad, pero en principio son dos formas diferentes de medir la distancia entre dos puntos, en metros y en horas:minutos:segundos. Son por tanto dos métricas posibles. Esto nos da una idea de que dado un conjunto de puntos/datos hay más de una forma de medir la distancia entre ellos. Esto, lejos de ser ambiguo, nos da la posibilidad de una gran flexibilidad, precisamente el ingrediente que necesitamos para abordar el problema que nos ocupa, la detección del fraude.

Puesto que en este artículo abordamos las bases de datos organizadas en grafos, vamos a centrar la atención en las métricas definidas sobre grafos. Como hemos indicado anteriormente, trabajamos con versiones más generales de los grafos. En cada nodo, además de un dato puede haber propiedades. Por ejemplo, en el nodo que define una empresa, que puede venir representado por un CIF, podemos incluir su nombre, su sede social, su ubicación geográfica, y propiedades semejantes. De la misma forma, las aristas que indican relaciones pueden estar dotadas de propiedades. Por ejemplo, si dos nodos del grafo está relacionados por, en el caso de contratación, "fue licitado a", donde uno de los nodos sería una administración y el otro una

empresa beneficiaria, dicha arista o relación puede contener información de fechas, número de expediente, etc.

Algunos de esos datos, que corresponden a dimensiones adicionales del problema, nos permiten desde el punto de vista matemático definir distancias entre puntos. Por ejemplo, dos empresas beneficiarias de contratos licitados por la misma administración podrían tener números CIF distintos y sin embargo, tener la misma ubicación geográfica (en el sentido de las mismas coordenadas cartográficas, o si se quiere expresar de otra manera, que ambas empresas ocupan el mismo local). Siguiendo con nuestro planteamiento, en la dimensión correspondiente a la ubicación geográfica, a pesar de ser dos empresas distintas, sin embargo la "distancia" entre ellas es nula. De la misma forma, como hemos dicho, también podemos dotar a las aristas o relaciones de propiedades. Podemos por ejemplo medir distancia entre la fecha de tramitación de una factura y la fecha de pago a la empresa proveedora del servicio. En principio para empresas que proveen el mismo servicio bajo las mismas condiciones de contratación, es esperable que las distancias temporales en términos de fechas de tramitación y pago sigan una determinada distribución (pongamos por caso una distribución gaussiana o exponencial). Midiendo estas distancias se podrían detectar anomalías en los procesos de caja o licitación.

Las funciones matemáticas a través de las cuales podemos definir este tipo de distancias son muy variadas (Deza et al 2009). Sin embargo, encontrar aquellas que son especialmente sensibles a ciertas anomalías en los datos requiere un estudio cuidadoso.

## 4. CONCLUSIONES

En este artículo hemos tratado de mostrar cómo las matemáticas pueden aportar nuevas herramientas para la prevención y detección del fraude. En particular, nuestra propuesta se basa en dos pilares fundamentales: organizar los datos en unas estructuras matemáticas bien conocidas como son los grafos, y aplicar en ellos el concepto de distancia. Aunque de forma más genérica, la idea subyacente es el concepto de proximidad que corresponde a una disciplina de las matemáticas conocida como topología. A esta disciplina, la topología, no le son ajenos otros problemas de proximidad a parte de la proximidad entre bases de datos (un problema clásico en Theoretical Computer Science), como puedan ser la proximidad entre puntos geográficos o pixeles en una imagen, etc.

Nuestra aproximación al problema reside en definir distancias sobre la base de datos que sean sensibles a determinados comportamientos de los datos que se desvían de cierta normalidad, o en otras palabras, a la búsqueda de patrones indicativos de una posible estrategia fraudulenta. Evidentemente, y desde un punto de vista matemático, existirá una probabilidad no nula de que comportamientos normales sean detectados como anómalos, pero evidentemente nuestro punto de vista es que ésa es parte de la labor de inspección y la revisión de casos, y que una herramienta como ésta debe ayudar a una mayor y mejor vigilancia de las cuentas públicas.

Queremos destacar que en nuestro planteamiento descartamos hacer hipótesis a priori sobre la estructura de los datos y el modelo que los representa. Consideramos que puede resultar peligroso y puede dar lugar a esquemas rígidos de búsqueda (como aquellos basados en cuestionarios fijos), que den lugar a que el fraude pase inadvertido, produciéndose además el efecto negativo de una cantidad importante de falsas alertas. Por lo tanto, los procesos de selección de datos y de elección de las funciones matemáticas y los criterios a la hora de determinar posibles situaciones de fraude o anomalías en los datos deben de ser, no sólo hechas con cautela, sino que también los criterios deben de surgir de los propios datos y no de planteamientos a priori. En este caso, además, la información previa sobre datos de situaciones fraudulentas puede ser de gran ayuda a la hora de decidir qué procedimientos y de qué forma deben ser implantados.

Para finalizar diremos que en la era del *Big Data* se hace imprescindible en la lucha contra el fraude la formación de equipos multidisciplinares formados por economistas, abogados, informáticos,…, y desde luego matemáticos. El mundo y su complejidad son matemáticos.

## 5. BIBLIOGRAFÍA


Abdullahi, R.; Mansor, N.; Nuhu, M.S. (2015): *Fraud Triangle Theory and Fraud Diamond Theory. Understanding the Convergent and Divergent for Future Research*. European Journal of Business and Management. Vol. 7, nº 28, 30-37.

Aho, A.V. (2013): Ullman J.D. *Foundations of computer science*.

Christofides, N. (1975): *Graph theory: An algorithmic approach*. Academic Press Inc.

Deza, M.M.; Deza, E. (2009): *Encyclopedia of distances*. Springer, Berlin Heidelberg.



(*) La presentación del contenido de este artículo ha recibido financiación a través de la convocatoria pública de la Generalitat Valenciana (DOGV 8064, de 16-6-2017) relativa al Sistema de alertas rápidas en la lucha contra la corrupción.