Algoritmos genéticos para la anonimización de grafos

Jordi Casas-Roma
Universitat Oberta de Catalunya
jcasasr@uoc.edu

Jordi Herrera-Joancomartí
Universitat Autònoma de Barcelona
jherrera@deic.uab.cat

Vicenç Torra
IIIA-CSIC
vtorra@iiia.csic.es

***Resumen*—En los últimos años se ha producido un incremento importante en el uso de los grafos como herramientas para la representación de información. Es muy importante poder preservar la privacidad de los usuarios cuando se desea publicar parte de esta información, especialmente en el caso de las redes sociales. En este caso es imprescindible aplicar un proceso de anonimización en los datos que permita preservar la privacidad de los usuarios. En este artículo se presenta un nuevo algoritmo para la anonimización de grafos, llamado *Genetic Graph Anonymization* (GGA), basado en la modificación de aristas para preservar el modelo de k-anonimidad.**

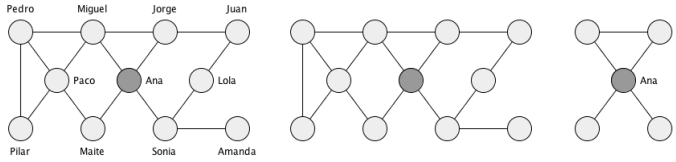
I. Introducción

En los últimos años la representación de datos en formato de grafo ha experimentado un importante auge en todos los niveles. Este formato permite representar estructuras y realidades más complejas que los tradicionales datos relacionales, que utilizan el formato de tuplas. En un formato semi-estructurado cada entidad puede presentar, al igual que los datos relacionales, una serie de atributos en formato numérico, nominal o categórico. Pero además, el formato de grafo permite representar de una forma más rica las relaciones que puedan existir entre las distintas entidades que forman en conjunto de datos. Un claro ejemplo de esta situación lo presentan las redes sociales.

En la actualidad las redes sociales se han convertido en un fenómeno muy popular, que hace que millones de usuarios de todo el mundo estén presentes en una o varias redes sociales. Independientemente de su temática u objetivo, las redes sociales presentan una gran cantidad de información muy interesante para estudios en distintos ámbitos (psicología, ciencias sociales, etc). En este sentido, la explotación de estos datos es de gran interés para científicos y empresas de todo el mundo. La problemática nace con la necesidad de preservar la privacidad de los individuos que aparecen en estas redes sociales.

Una primera aproximación a la anonimización de este tipo de datos, denominada como anonimización simple, consiste en eliminar cualquier información que permita re-identificar de forma única a un usuario dentro de los datos explotados. Un ejemplo de esta técnica se muestra en la Figura 1.

Agradecimientos: Este trabajo está parcialmente financiado por el Ministerio de Ciencia y Educación, a través de los proyectos TSI2007-65406-C03 "E-AEGIS", TIN2010-15764 "N-KHRONOUS", CONSOLIDER CSD2007-00004 "ARES" y TIN2011-27076-C03 "CO-PRIVACY".



(a) G (b) G_1 (c) G_{Sara}

Figura 1: Ejemplo de anonimización, donde G es el grafo de una red social, G_1 es el mismo grafo anonimizado y G_{Ana} es el grafo de vecindad a distancia 1 de Ana.

En la Figura 1a se puede ver un ejemplo reducido de una red social donde cada nodo representa a un individuo y cada una de las aristas representa la relación de amistad entre dos individuos. La Figura 1b proporciona la red después de un proceso de anonimización simple, en donde la topología de la red se ha mantenido inalterada, pero se han eliminado los identificadores de los nodos.

Como se puede ver en este ejemplo, un proceso de anonimización simple como el descrito puede presentar algunas debilidades, dependiendo de la información que un atacante posea de los vecinos de un nodo. Por ejemplo, si el atacante sabe que Ana tiene un total de 4 amigos y que dos de ellos, además, son amigos entre si, puede construir el grafo de vecindad a distancia 1 (*1-neighborhood*) de Ana, representado en la Figura 1c. A partir de este grafo se puede identificar de forma única a Ana dentro del grafo anonimizado, y por lo tanto, comprometer la privacidad de este usuario.

Si bien el ejemplo citado es bastante simple, nos da una idea de la dificultad que entraña el proceso de anonimización de datos expresados en formato de grafo, puesto que justamente la interrelación de los datos que permite este formato es una potente herramienta que un atacante puede utilizar para desanonimizar información protegida y comprometer información individual de los usuarios.

En el presente artículo proponemos un nuevo método, que denominamos *Genetic Graph Anonymization* (GGA), para la preservación de la privacidad en grafos basado en el modelo de k-anonimidad.

La estructura del resto del artículo es como sigue. La Sección II resume el estado del arte. En la Sección III se propone el nuevo algoritmo, denominado GGA. La Sección IV muestra los resultados empíricos de la utilización del

algoritmo. Finalmente, la Sección V concluye el artículo y propone algunas lineas futuras de investigación.

II. Estado del arte

Si bien existen distintos métodos para la protección de grafos como los que se basan en modificaciones aleatorias ([2], [3], [4], [5]) o los basados en generalización, los cuáles agrupan nodos y aristas, formando subgrafos que se anonimizan como super-nodos o super-aristas ([6], [7], [8]), en este artículo nos centramos en aquellos métodos que realizen modificaciones del grafo encaminadas a preservar el modelo de k-anonimidad.

En 2002, Sweeney presenta en [1] el concepto de k-anonimidad (*k-anonymity*) que inicialmente se diseñó para datos relacionales, pero posteriormente ha sido adaptado para trabajar con grafos. Formalmente, se define el modelo de k-anonimidad como sigue. Si $RT(A_1, \ldots, A_n)$ es una tabla formada por un conjunto de registros (tuplas) con n atributos, se dice que RT satisface el modelo de k-anonimidad si y sólo si cada una de las secuencias formadas por cualquier subconjunto de atributos QI_{RT} aparecen como mínimo k veces en $RT\,[QI_{RT}]$. A los subconjuntos de atributos QI_{RT} se les llama casi-identificadores (*quasi-identifiers*). De esta forma el modelo indica que un atacante no puede diferenciar k registros entre si. Por lo tanto, no podrá re-identificar a un individuo con una probabilidad superior a $\frac{1}{k}$.

Aplicando este modelo a los grafos, se pueden utilizar distintos conceptos como casi-identificador. Una primera opción adoptada en múltiples trabajos es utilizar el concepto de grado de los nodos como casi-identificador. De esta forma, se supone que el atacante intentará identificar nodos en el grafo original que tengan un grado único en todo el grafo, es decir, $grado(v_i) \neq grado(v_j)\ \forall j \neq i$. A partir de este conjunto de nodos con grado único, el atacante deberá buscar nodos en el grafo anonimizado con el mismo grado que los nodos identificados sobre el grafo original. Si la correspondencia es única entre ambos, el atacante habrá conseguido identificar positivamente un conjunto de nodos. En estos casos se dice que el grafo es k-anónimo en el grado, siendo k el cardinal del menor de los conjuntos de nodos del mismo grado.

En [6] Hay et al. proponen un modelo generalizado del concepto de k-anonimidad, llamado k-anonimidad de candidatos (*k-candidate anonymity*). En este modelo se define a un nodo v_i como k-anónimo de candidatos con respecto a una pregunta Q si existen, al menos, $k-1$ otros nodos en el grafo que responden positivamente a la misma pregunta Q. Formalmente, $\mid cand_Q(v_i) \mid \geq k$ donde $cand_Q(v_i) = \{v_j \in V \mid Q(v_j) = Q(v_i)\}$. Se dice que un grafo satisface la restricción de k-anonimidad de candidatos con respecto a Q si todos sus nodos son k-anónimos de candidatos respecto a Q.

Este concepto permite ampliar el modelo de k-anonimidad según el conocimiento que se presuponga que el adversario tenga de los datos, ya sea información limitada del mismo nodo (el grado del mismo), información a distancia 1 o bien información más amplia de la red. Por ejemplo, en [9] Liu y Terzi asumen un conocimiento del adversario basado en el grado de los nodos objetivos. En [10] Zhou y Pei consideran el subgrafo formado por los vecinos a distancia 1 (*1-neigborhood*) de los nodos objetivos. Y en [11] Zhou et al. consideran toda la información estructural posible alrededor del nodo objetivo y proponen un nuevo modelo llamado k-automorfismo (*k-automorphism*) para garantizar la privacidad ante ataques con este tipo de información. En este trabajo nos centraremos en la aproximación que únicamente tiene en cuenta el grado de los nodos puesto que, como veremos en los resultados de los experimentos, esta aproximación tan simple ya genera unos algoritmos con complejidades muy elevadas.

Hay et al. [2], [6] proponen un método, llamado *Vertex Refinement Queries*, para modelar el conocimiento del adversario en base a un conjunto de consultas iterativas que permiten modelar la estructura de vecindad local de un nodo del grafo. La consulta inicial, $\mathcal{H}_0(v_j)$, simplemente devuelve la etiqueta del nodo v_j; la siguiente consulta, $\mathcal{H}_1(v_j)$ devuelve el grado del nodo v_j; $\mathcal{H}_2(v_j)$ devuelve la lista de los grados de todos los nodos adyacentes del nodo v_j, y así sucesivamente. De forma general: $\mathcal{H}_i(v_j) = \{\mathcal{H}_{i-1}(v_1), \mathcal{H}_{i-1}(v_2), \ldots, \mathcal{H}_{i-1}(v_m)\}$ para $i \geq 2$ donde $v_1, \ldots, v_m$ son los nodos adyacentes a v_j. Un conjunto de candidatos para una consulta $\mathcal{H}_i$ es el conjunto de todos los nodos con el mismo valor de $\mathcal{H}_i$. Por lo tanto, la cardinalidad de un conjunto de candidatos establecido por $\mathcal{H}_i$ es el número de nodos indistinguibles en el grafo respecto a $\mathcal{H}_i$. Es relevante notar que si la cardinalidad del menor de los conjuntos de candidatos se establece en k, entonces la probabilidad de re-identificación basada en $\mathcal{H}_i$ es $\frac{1}{k}$.

II-A. k-anonimidad basada en el grado

En [9], Liu y Terzi investigan cómo modificar la estructura de un grafo $G(V, E)$ añadiendo y eliminando aristas, para que el nuevo grafo $\widetilde{G}(\widetilde{V}, \widetilde{E})$ cumpla las restricciones de k-anonimidad en el grado (*k-degree anonymity*). Esta restricción implica que para cada nodo deberán existir, al menos, otros $k-1$ nodos con el mismo grado. Este modelo preserva la privacidad ante ataques de re-identificación basados en el conocimiento del grado de los nodos objetivo. En general, cuanto mayor sea el valor de k, mayor será el grado anonimización de $\widetilde{G}$ y mayor también la pérdida de información.

Formalmente, el grafo anonimizado $\widetilde{G}(\widetilde{V}, \widetilde{E})$ a partir de las inserciones y eliminaciones de aristas en el grafo original $G(V, E)$ deberá cumplir las siguientes restricciones:

1. $\widetilde{G}$ debe ser k-anónimo en el grado
2. $V = \widetilde{V}$
3. $E \cap \widetilde{E} \approx E$

El algoritmo desarrollado consiste en dos fases:

1. La primera fase modifica la secuencia de grados del grafo original. La secuencia de grados es una secuencia de números que representan el grado de cada nodo, $d = \{grado(v_0), grado(v_1), \ldots, grado(v_n)\}$ donde $n = \mid V \mid$. A partir de la secuencia de grados del grafo original se pretende obtener una secuencia k-anónima para un valor específico de k, realizando el número

mínimo de cambios posibles en la secuencia original, ya que de esta forma se minimizan las modificaciones de aristas en el grafo. Los autores resuelven esta parte mediante técnicas de programación lineal.

2. En la segunda fase, se debe crear un grafo G_0 a partir de la secuencia k-anónima generada en la fase anterior. Luego se aplica un proceso iterativo de intercambio válido de aristas para conseguir que el nuevo grafo sea lo más parecido posible al grafo original. Los autores definen este intercambio válido de aristas como una operación entre cuatro nodos v_a, v_b, v_c y v_d de $G_i(V, E_i)$ tales que $e_{a,c}, e_{b,d} \in E_i$ y $e_{a,b}, e_{c,d} \notin E_i$ o bien $e_{a,d}, e_{b,c} \notin E_i$, donde $e_{a,b} = (v_a, v_b)$ y $G_i(V, E_i)$ representa el grafo G_0 después de i iteraciones. El objetivo es conseguir que el conjunto de aristas del nuevo grafo sea tan similar como sea posible al conjunto de aristas del grafo original ($E \cap \widetilde{E} \approx E$).

III. *Genetic Graph Anonymization (GGA)*

En la presente sección describimos nuestra propuesta para la anonimización de grafos, que denominamos *Genetic Graph Anonymization* (GGA), que se basa en algoritmos genéticos y está enfocada a obtener un grafo anonimizado que preserve el modelo de k-anonimidad.

Una descripción de nuestra propuesta a alto nivel nos permite estructurar nuestro algoritmo de anonimización en dos fases, de forma similar a la aproximación realizada por Liu y Terzi en [9]:

1. En una primera fase, a partir de la secuencia de grados de los nodos de $G(V, E)$, $d = \{d_1, \cdots, d_n\}$, se construye una nueva secuencia $\widetilde{d}$ que sea k-anónima en el grado y se minimiza la distancia entre ambas secuencias, calculada como:

$$D(d, \widetilde{d}) = \sum_{i=0}^{n} \mid \widetilde{d}_i - d_i \mid \tag{1}$$

   donde $n = \mid V \mid$.
2. En la segunda fase, se construye un nuevo grafo $\widetilde{G}(\widetilde{V}, \widetilde{E})$ en el cual $\widetilde{V} = V$, $\widetilde{E} \cap E \approx E$ y la secuencia de grados sea igual a $\widetilde{d}$.

La construcción de la secuencia de grados k-anónima determina el grado de anonimización aplicado y la distancia a la secuencia de grados del grafo original. Una secuencia óptima debe proporcionar el grado de k-anonimidad solicitado y además minimizar la distancia con la secuencia de grados del grafo original. Esta segunda condición determina, en gran medida, la utilidad de los datos.

III-A. Primera fase: obtención de la secuencia de grados k-anónima

El problema de obtener una secuencia de grados k-anónima tiene ciertas particularidades que se deben considerar:

- El número de elementos de la secuencia de grados determina el número de nodos, por lo tanto, este valor no se puede alterar.
- Dado que los valores de la secuencia de grados son los grados de nodos del grafo, estos deben tener un valor entero en el rango $[0, n-1]$, donde $n = \mid V \mid$.
- El número total de aristas del grafo es la mitad del sumatorio de la secuencia, ya que cada arista se contabiliza dos veces en la secuencia de grados. Para preservar el número de aristas, el sumatorio de la secuencia obtenida debe ser igual al sumatorio de la secuencia original.
- Los cambios realizados en la secuencia de grados se transforman en modificaciones de aristas en el grafo. Por lo tanto, es necesario realizar el número mínimo de modificaciones en la secuencia de grados (minimizando la distancia entre ambas secuencias) para obtener un grafo anonimizado con el mínimo número de modificaciones en las aristas respecto al grafo original.

Nuestra propuesta plantea para la primera fase del algoritmo de anonimización la utilización de algoritmos genéticos. Los algoritmos genéticos se basan en evolucionar una población de individuos sometiéndola a acciones aleatorias y posteriormente aplicar procesos de selección de acuerdo con algún criterio, en función del cual se decide cuáles son los individuos más adaptados, que sobreviven, y cuáles los menos aptos, que son descartados. Dichos algoritmos se componen de tres fases principales: inicialización, generación y selección de candidatos. En la primera fase, la población se inicializa a partir de los datos iniciales. En la segunda fase del algoritmo, que constituye la fase principal y se ejecuta iterativamente, los candidatos son modificados, evaluados y se aplica el proceso de selección para determinar cuáles son los individuos que pasan a la siguiente generación. En esta segunda fase, los algoritmos genéticos realizan la evolución de las poblaciones basándose en dos modificaciones elementales: la mutación y la recombinación de parejas de padres. Finalmente, en la tercera fase se selecciona el mejor candidato, que será identificado y devuelto como solución del problema.

La explicitación de esta primera fase de nuestra propuesta utilizando algoritmos genéticos queda detallada en el Algoritmo 1.

Algorithm 1 Pseudocódigo del algoritmo para obtener una secuencia k-anónima

Require: La secuencia de grados original (d) y el valor de k-anonimidad deseado (k).
Ensure: La secuencia de grados ($\widetilde{d}$) que cumple la k-anonimidad.

```
INICIALIZAR poblacion ⇐ d
k_actual ⇐ OBTENER_K poblacion
while k_actual < k do
    MUTAR poblacion
    EVALUAR nuevos candidatos
    poblacion ⇐ SELECCIONAR individuos
    k_actual ⇐ OBTENER_K individuos
end while
d̃ ⇐ SELECCIONAR mejor individuo
return d̃
```

Como se muestra en el Algoritmo 1, en nuestro caso, la población se inicializa a partir de la secuencia de grados del grafo original. Posteriormente, el bucle *while* muestra la fase de generación. En esta fase se realiza una operación de mutación básica (función MUTAR en el Algoritmo 1), que dadas las particularidades de nuestro problema, es la siguiente: sumar uno a un elemento de la secuencia de grados y restar uno a otro elemento. Esta operación representa un cambio en uno de los nodos de una arista. Por ejemplo, si a la arista $e_{0,1}$ se le modifica un nodo, se puede obtener $e_{0,2}$. Este cambio se representa en la secuencia de grados como restar uno al nodo v_1 y sumar uno al nodo v_2. Nótese que nuestro algoritmo genético no utiliza la recombinación de parejas de padres dado que esta evolución incumpliría de forma sistemática la regla que preserva el número de aristas del grafo, y por lo tanto, generaría candidatos no válidos. Consideramos que el rendimiento del algoritmo se vería afectado por la inclusión de este tipo de evolución y no se producirían mejoras en tiempo o en calidad de la solución hallada.

Cuando la generación de candidatos ha terminado, se evalúa la bondad de los candidatos obtenidos. La función de *fitness* que realiza este cálculo (función EVALUAR en el Algoritmo 1) se basa en tres parámetros:

1. El valor de k-anonimidad de la secuencia. Se debe conseguir un valor igual o superior al valor deseado.
2. La distancia entre la secuencia de grados y la secuencia de grados original. El objetivo es minimizar este valor, según describe la Ecuación 1.
3. En el caso de obtener un valor de k-anonimidad inferior al valor deseado, se considera el número de grupos de nodos que, agrupados según su grado, presentan una cardinalidad inferior al valor de k deseado. Cuando el valor de k-anonimidad sea el deseado, este parámetro vale 0.

Finalmente, para la selección de los supervivientes, es decir, los individuos que pasan a la siguiente generación, se utiliza el modelo de estado progresivo (*steady-state model*). Según este modelo, se seleccionan los peores individuos de la población actual y se reemplazan por los mejores individuos del conjunto de candidatos. Esta valoración se realiza basándose en la puntuación obtenida en la función de *fitness*.

III-B. Segunda fase: modificación del grafo original

Una vez la primera fase del algoritmo obtiene la secuencia de grados k-anónima, en la segunda fase se realizan los cambios necesarios en el grafo original para obtener un grafo k-anónimo. Las modificaciones que se hayan producido en la secuencia de grados indican los nodos que deben modificar su grado, es decir, indican los nodos en los que se deben crear y eliminar aristas.

Tal y como se muestra en el Algoritmo 2, esta fase se inicia obteniendo el vector de diferencias, $d_{difs} = d - \widetilde{d}$, que permite detectar fácilmente los nodos que deben disminuir su grado o aumentarlo. A los nodos que deben disminuir su grado se les deberán eliminar una o más aristas incidentes, mientras que a los nodos que deben aumentar su grado, se les deberán añadir nuevas aristas. Este cambio en las aristas se produce eliminando la arista $e_{p,q} \in E$ donde v_q pertenece a los nodos que deben disminuir su grado y creando la arista $e_{p,r}$ donde v_r pertenece a los nodos que deben aumentar su grado.

Algorithm 2 Pseudocódigo del algoritmo para modificar el grafo a partir de la secuencia k-anónima

Require: El grafo original $G(V, E)$, la secuencia de grados original d y la secuencia de grados k-anónima $\widetilde{d}$.
Ensure: El grafo $\widetilde{G}(V, \widetilde{E})$ donde la secuencia de grados es igual a $\widetilde{d}$ y $\widetilde{E} \cap E \approx E$.
$\widetilde{G}(V, \widetilde{E}) \Leftarrow \widetilde{G}_0(V, \widetilde{E})$
$d_{difs} = d - \widetilde{d}$
$V_{del} = \{v_i \in V \mid d_{difs}(i) < 0\}$
$V_{add} = \{v_i \in V \mid d_{difs}(i) > 0\}$
while $V_{del} \neq \emptyset$ AND $V_{add} \neq \emptyset$ **do**
  $\widetilde{E} = \widetilde{E} \setminus \{e_{p,q}\}$ donde $e_{p,q} \in E$ y $v_q \in V_{del}$
  $V_{del} = V_{del} \setminus \{v_q\}$
  $\widetilde{E} = \widetilde{E} \cup \{e_{p,r}\}$ donde $v_r \in V_{add}$
  $V_{add} = V_{add} \setminus \{v_r\}$
end while
return $\widetilde{G}$

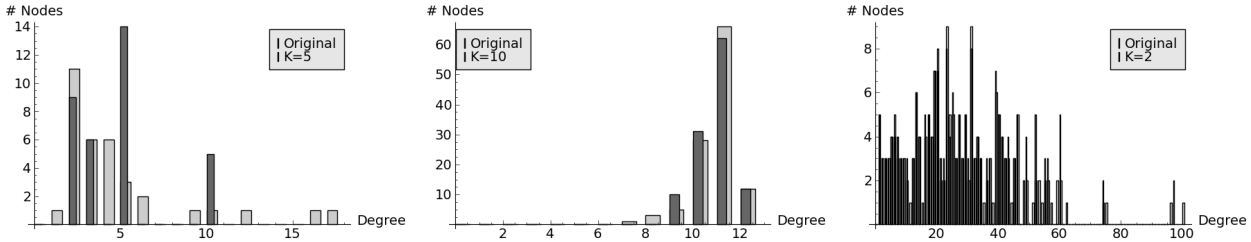
IV. Resultados experimentales

Se han utilizado tres conjuntos de datos reales para realizar los experimentos con el algoritmo *GGA*: *Zachary's Karate Club* [12], *American College Football* [13] y *Jazz Musicians* [14]. La Tabla I muestra las principales propiedades de estos grafos.

Conjunto	*Nodos*	*Aristas*	*Grado*	*Distan.*	*Diám.*
Zachary's Karate Club	34	78	4,588	2,408	5
American College	115	613	10,661	2,508	4
Jazz Musicians	198	2.742	27,697	2,235	6

Cuadro I: Propiedades de los conjuntos de datos: Número de nodos (*Nodos*), Número de aristas (*Aristas*), grado medio (*Grado*), distancia media (*Distan.*) y diámetro (*Diám.*)

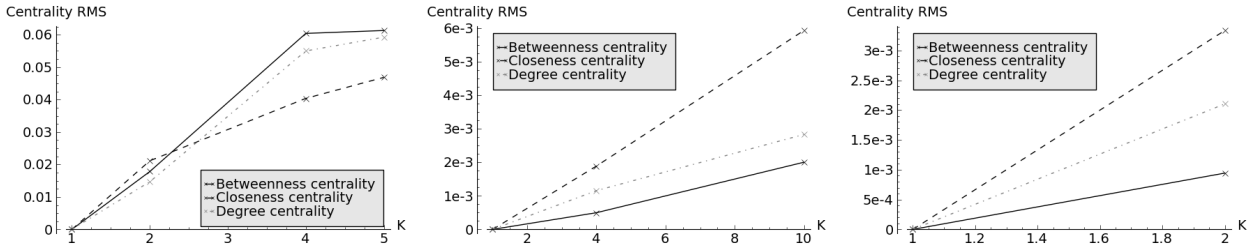
El primer conjunto de datos es un pequeño grafo de 34 nodos con un valor de k-anonimidad basado en el grado igual a 1. El algoritmo GGA permite anonimizar el grafo en valores de k iguales a 2, 4 y 5. La Figura 2a muestra el histograma de grados del grafo original y del grafo anonimizado con un valor de $k = 5$. En el histograma del grafo original se puede ver como la distribución de los grados sigue la ley de la potencia (el número de nodos decrece de forma exponencial al aumentar el grado). Para evaluar el grado de ruido introducido se analizan tres medidas relacionadas con la centralidad de los nodos del grafo. En concreto, las medidas analizadas son: *betweenness centrality* (mide la frecuencia con la que cada nodo aparece en el conjunto de caminos cortos (*shortest paths*) dentro de un grafo, *closeness centrality* (se define como la inversa de la distancia media a todos los nodos accesibles) y *degree centrality* (considera la centralidad de cada nodo asociada a su grado). La Figura 2d muestra



(a) Histograma de grados (conjunto Zachary's Karate Club)

(b) Conjunto American College

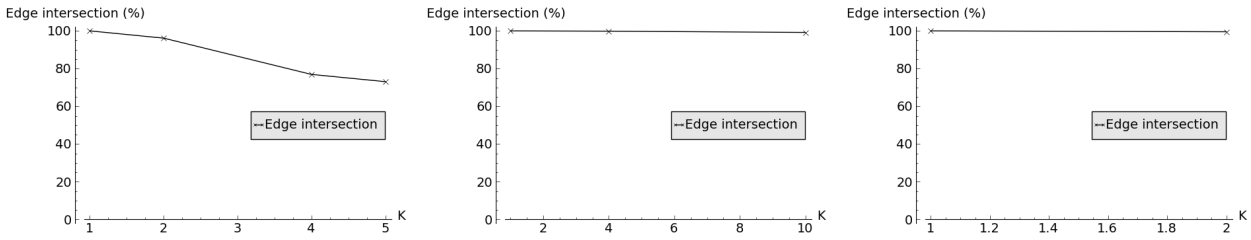
(c) Conjunto Jazz Musicians



(d) Error en las medidas de centralidad (conjunto Zachary's Karate Club)

(e) Conjunto American College

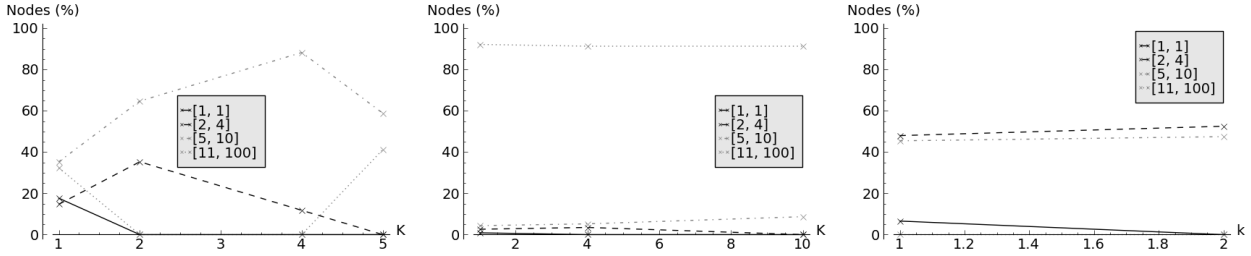
(f) Conjunto Jazz Musicians



(g) Aristas modificadas (conjunto Zachary's Karate Club)

(h) Conjunto American College

(i) Conjunto Jazz Musicians



(j) $Cand_{\mathcal{H}_1}$ (conjunto Zachary's Karate Club)

(k) Conjunto American College

(l) Conjunto Jazz Musicians

Figura 2: Resultados experimentales de los tres conjuntos de datos.

el error medio cuadrático introducido en estas medidas de centralidad según el valor de k-anonimidad conseguido. Se puede observar que el error aumenta conforme aumenta el grado de anonimización. Cabe destacar el ligero aumento que se produce entre los valores de $k = 4$ y $k = 5$. Otra medida importante para evaluar el ruido introducido en los datos es el número de aristas que se mantienen iguales entre el grafo original y los grafos anonimizados. La Figura 2g muestra este dato conforme avanza el grado de anonimización. Para un grafo con un valor de k-anonimidad igual a 2 se obtienen un 96,15 % de aristas iguales al grafo original, descendiendo hasta el 73,08 % cuando el grafo de k-anonimidad es igual a 5. Para finalizar, se muestra la evolución de $Cand_{\mathcal{H}_1}$ como información complementaria al valor de k-anonimidad conseguido en cada grafo. Esta información permite ver como evolucionan los nodos en base al riesgo de re-identificación. La Figura 2j muestra la evolución de $Cand_{\mathcal{H}_1}$ para este conjunto de datos. La linea sólida muestra el porcentaje de nodos con re-identificación directa (es decir, que tienen grado único dentro del grafo), la linea discontinua muestra los nodos con alto riesgo de re-identificación (grupos de entre 2 y 4 nodos con el mismo grado), la linea discontinua y punteada los nodos con riesgo moderado (grupos de entre 5 y 10 nodos con el mismo grado) y la linea punteada muestra los nodos con riesgo bajo

o muy bajo de re-identificación (grupos de más de 11 nodos con el mismo grado).

El segundo conjunto de datos es un grafo de 115 nodos con un valor inicial de k-anonimidad igual a 1. El algoritmo GGA consigue anonimizar el grafo en valores de k igual a 4 y 10. Observando el histograma de grados de la Figura 2b se puede ver que sólo se han modificado los nodos con grado 7 y 8 para poder obtener un grafo con un valor de k-anonimidad igual a 10. Es decir, dada la estructura de este grafo sólo se ha requerido aplicar una pequeña cantidad de modificaciones en las aristas para conseguir unos excelentes niveles de anonimización. Las medidas de centralidad, Figura 2e, muestran que se ha introducido muy poco ruido en los datos anonimizados. Cabe destacar, como se puede ver en la Figura 2h, que más del 99 % de las aristas del grafo original están presentes en ambos grafos anonimizados (99,84 % en el grafo con $k = 4$ y 99,18 % en el grafo con $k = 10$). La medida $Cand_{\mathcal{H}_1}$, Figura 2k, muestra el descenso en los grupos de re-identificación directa y alto riesgo de re-identificación hasta alcanzar el valor nulo para el grafo con valor de k-anonimidad igual a 10.

El tercer conjunto de datos es un grafo de 198 nodos con un valor inicial de k-anonimidad igual a 1. Este grafo presenta un grado medio más alto que los demás conjuntos, del orden de 27 aristas por nodo. El histograma de grados, Figura 2c, muestra la existencia de dos nodos con un grado muy superior a los demás, 96 y 100. Es importante notar que estos dos nodos presentan una re-identificación directa, pero al mismo tiempo y debido a su centralidad en el grafo, son nodos claves en la estructura de la red. Para este conjunto de datos sólo se ha conseguido anonimizar el grafo a un valor de k-anonimidad igual a 2, debido a la problemática comentada. Para poder aumentar el grado de k-anonimidad de los grafos anonimizados seria necesario modificar estos dos nodos centrales de tal forma que perderían su centralidad y, en consecuencia, se destruiría de forma muy importante la estructura de la red, dejando los datos anonimizados con una utilidad muy reducida. El grafo anonimizado con valor de k-anonimidad igual a 2 presenta un error bajo en las medidas de centralidad, Figura 2f, y un alto porcentaje de coincidencia de aristas con el grafo original (99,53 %), Figura 2i, que indica que se ha introducido poco ruido en los datos anonimizados. Para finalizar, la medida $Cand_{\mathcal{H}_1}$, Figura 2l, muestra un descenso en el grupo de re-identificación directa.

V. Conclusiones

En este artículo se ha presentado un nuevo algoritmo de anonimización de grafos, basado en la modificación de aristas para conseguir aumentar el valor de k-anonimidad del grafo anonimizado. El nuevo algoritmo, llamado *Genetic Graph Anonymization* (GGA), se basa en la modificación de la secuencia de grados mediante algoritmos genéticos. El objetivo de este proceso es conseguir una secuencia de grados que sea k-anónima en el grado y que minimice la distancia con el grafo original, para preservar al máximo la utilidad de los datos. A partir de la secuencia de grados anonimizada se aplicarán los cambios necesarios en el grafo original para que su secuencia de grados sea igual a la secuencia de grados anonimizada.

Los resultados obtenidos son favorables e indican que el algoritmo es capaz de proporcionar grafos anonimizados con un valor de k-anonimidad superior al valor inicial, y mantener, además, un nivel de ruido bajo en los datos anonimizados.

El trabajo deja una multitud de caminos abiertos para seguir con esta investigación. En primer lugar, se podría explorar la posibilidad de incluir la recombinación de padres como método de evolución. Aunque en primera instancia no parece aportar mejoras, un estudio en profundidad seria de interés para trabajos futuros. En segundo lugar, se podrían implementar otras medidas que permitieran evaluar el grado de ruido asociado al proceso de anonimización. Una posibilidad muy interesante seria considerar las propiedades espectrales de los grafos. Los valores y vectores propios de las matrices de adyacencia o de Laplace pueden ayudar a evaluar como afecta a un grafo la perturbación introducida por los procesos de anonimización. En tercer lugar, resultaría muy interesante abrir el estudio a otros tipos de grafos. El uso de grafos con pesos en las aristas o atributos en los nodos abre nuevos retos a la anonimización.

Referencias

[1] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, pp. 557–570, 2002.

[2] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, "Anonymizing social networks," Computer Science Department, University of Massachusetts Amherst, Technical Report No. 07-19, 2007.

[3] X. Ying, K. Pan, X. Wu, and L. Guo, "Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing," in *Proc. of the 3rd Workshop on Social Network Mining and Analysis*, ser. SNA-KDD '09. New York, NY, USA: ACM, 2009, pp. 10:1–10:10.

[4] L. Zhang and W. Zhang, "Edge anonymity in social network graphs," in *Proc. of the 2009 Intl. Conf. on Computational Science and Engineering - Vol. 04*. USA: IEEE Computer Society, 2009, pp. 1–8.

[5] X. Ying and X. Wu, "Randomizing Social Networks: a Spectrum Preserving Approach," in *SDM*. SIAM, 2008, pp. 739–750.

[6] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," *Proc. VLDB Endow.*, vol. 1, pp. 102–114, August 2008.

[7] A. Campan and T. M. Truta, "Data and structural k-anonymity in social networks," *Privacy, Security, and Trust in KDD*, pp. 33–54, 2009.

[8] C. C. Aggarwal and H. Wang, *Managing and Mining Graph Data*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[9] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *Proc. of the 2008 ACM SIGMOD Intl. Conf. on Management of Data*, ser. SIGMOD '08. New York, NY, USA: ACM, 2008, pp. 93–106.

[10] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *Proc. of the 2008 IEEE 24th Intl.l Conf. on Data Engineering*. USA: IEEE, 2008, pp. 506–515.

[11] L. Zou, L. Chen, and M. T. Ozsu, "k-automorphism: a general framework for privacy preserving network publication," *Proc. VLDB Endow.*, vol. 2, pp. 946–957, August 2009.

[12] W. Zachary, "Information-flow model for conflict and fission in small-groups," *J. Of Anthropological Research*, vol. 33, pp. 452—473, 1977.

[13] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, June 2002.

[14] P.Gleiser and L. Danon, "Adv. complex syst.6, 565," 2003.