Ariel Salgado,* Mariella Fumagalli,** Analía González Simonetto,** Alejandra Ibañez,** Patricia Bernardi,** Carlos Somigliana,** Mercedes Salado Puerto,** Inés Caridi*

Cohesión de las estructuras resultantes en redes sociales: estudio de caso sobre la desaparición de personas en la provincia de Tucumán

Cohesion of the resulting structures in social networks: case study on the disappearance of people in the province of Tucumán

Abstract | In this work, we propose a quality characterization of the structures arising in social networks, in terms of the homogeneity among the individuals belonging to those structures or groups. The proposed techniques are intended to be used in sparse networks, in which isolated groups of agents (clusters) emerge, but can be directly implemented for more dense networks, with communities are detected by any method. The group quality is quantified in terms of their inner similarity or homogeneity, based on a small set of TRUE/FALSE (belongs or not belongs to the category) categorical variables, known for all elements in the network, which were not used previously in the network definition, and also they are not necessarily disjoint. The methodology is to compare each group with what could be obtained if the connections were randomly assigned, to quantify each category's presence in that group, but also taking into account that categories should not mix too much if they do not mix in the whole population. The specific application of these methodologies aims to take advantage of the network characterization, helping to make priority rankings in the investigation of missing people, because it gives connections, bonds, paths and nodes that to help to understand the operational framework of the last Argentinian civic and military dictatorship, in the Tucumán province.
Keywords | networks, clusterization, inner similarity, external validation.

Resumen | En este trabajo se propone un método para caracterizar la calidad de las estructuras que surgen en redes de origen social, en términos de la homogeneidad de los agentes

Recibido: 6 de enero de 2020. Aceptado: 22 de septiembre de 2020.
*Instituto de Cálculo, Universidad de Buenos Aires y Conicet, Argentina.
**Equipo Argentino de Antropología Forense.
Correos electrónicos de contacto: inescaridi@yahoo.com.ar | eaaf@eaaf.org

Salgado, Ariel, Mariella Fumagalli, Analía González Simonetto, Alejandra Ibañez, Patricia Bernardi, Carlos Somigliana, Mercedes Salado Puerto, Inés Caridi. «Cohesión de las estructuras resultantes en redes sociales: estudio de caso sobre la desaparición de personas en la provincia de Tucumán.» *Interdisciplina* 9, nº 23 (enero–abril 2021): 83-96.
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

Volumen 9, número 23, (83-96), enero–abril 2021
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

que integran esas estructuras o grupos. Las técnicas propuestas fueron pensadas para redes con pocas conexiones (*sparse networks*) en las que pueden determinarse grupos conectados entre sí y aislados del resto (los *clusters* de la red), aunque son inmediatamente extendibles a redes más densas, donde los grupos se determinan mediante métodos de detección de comunidades. Los grupos se evalúan de acuerdo con su similaridad u homogeneidad interna, con base en un conjunto pequeño de variables categóricas conocidas para todos los agentes de la red, del tipo TRUE/FALSE (pertenece o no a la categoría), que no fueron usadas en la definición de la red y que no necesariamente son disjuntas. Comparando cada grupo con lo esperado si las conexiones hubieran surgido por azar, podemos evaluar el grado en que la presencia de una categoría dentro de este, difiere de la aleatoriedad. Sin embargo, grupos con presencia fuerte de dos categorías que no se vinculan entre sí en la población completa no son considerados como aceptables. La aplicación específica de esta metodología busca caracterizar las relaciones relevantes en un sistema de individuos mediante redes y conocer las estructuras emergentes para ordenar prioridades en la investigación de personas desaparecidas, ya que brinda conexiones, vínculos, recorridos y nodos, que pueden facilitar la comprensión del contexto y modo en el que operó la última dictadura militar argentina, en la provincia de Tucumán.

Palabras clave | redes, *clusterización*, similaridad interna, validación externa.

Introducción

EN EL PROBLEMA ABSTRACTO, partimos de una red G conformada por un conjunto de N nodos, que representarán a los individuos, y un conjunto de L conexiones entre ellos. En nuestro caso de interés, hay muchas menos conexiones en la red que el total posible ($L << N(N-1)/2$) por lo que la red es *sparse*. Además, la red es del tipo no pesada ni dirigida y su estructura está compuesta por varios *clusters* (grupos de individuos relacionados entre sí y aislados del resto) o se pueden determinar distintas comunidades en ella. Nos enfocaremos en la versión *sparse* principalmente, dado que se corresponde con nuestro caso de estudio. Sin importar cómo fueron definidas las conexiones de esa red, ya sea con base en información explícita de las relaciones entre los individuos o relaciones inferidas a partir de la información de los nodos, existe otra información que no fue usada para construir la red, y que puede servir para caracterizar la calidad de los *clusters* encontrados en términos de la similaridad entre sus integrantes respecto de este set de variables. Las medidas utilizadas en este sentido obedecen al análisis de *cluster* para medir la similaridad interna mediante coeficientes de "información mutua puntual", los cuales miden la correlación entre variables, la distribución hipergeométrica para muestras relativamente pequeñas y un test de significancia estadístico.

El estudio de grafos como herramienta matemática tuvo un gran impulso a partir de los trabajos de Erdos y Renyi (1960), con la teoría de grafos aleatorios,

aunque recién en trabajos posteriores (Watts y Strogatz 1999; Barabási y Albert 1999) se ha empezado a comprender cuál es la estructura de las redes reales. Las redes como herramienta para caracterizar la conectividad del mundo social de forma empírica han sido utilizadas desde el trabajo ya histórico sobre los grados de separación entre individuos (Milgram 1967), la caracterización de grupos de trabajo, como las redes de colaboración científica (Newman 2001), o del posicionamiento político de una sociedad (Boutyline y Willer 2017). Un paso más allá en la caracterización incluye el empleo de redes para, por ejemplo, estimar el número de personas desaparecidas en un terremoto (Bernard *et al.* 1991). La detección de comunidades dentro de las redes es un problema largamente discutido en la literatura (Javed *et al.* 2018). Se han definido múltiples magnitudes que permiten medir cuán fuertemente relacionado está un grupo entre sí con respecto al resto. La más famosa de ellas es la modularidad, que mide el exceso de conexiones entre los integrantes respecto a lo que se esperaría si se hubiesen formado "al azar" (siguiendo algún modelo de lo que el azar representa) (Newman 2006). En el caso de que la red sea de tipo *sparse*, este problema se simplifica, pues consideramos como grupos a las componentes disjuntas de la red.

Habiendo definido los grupos de la red (*clusters* o comunidades), resta contestar la pregunta de si estos grupos están captando algún fenómeno subyacente (grupos de mejores amigos en comunidades o, como en este caso, personas con destino común en redes de personas desaparecidas, entre otros). Cuantificar la calidad de los grupos es un tema de difícil acuerdo en la literatura de redes complejas. Las denominadas medidas internas (como la modularidad) califican los *clusters* en términos de su estructura topológica (Almeida *et al.* 2011), basándose en la fuerza de la conexión entre sus nodos en comparación con el resto. Si este tipo de medidas reproducen (o deberían reproducir) el efecto de fenómenos subyacentes, es aún un tema de discusión. En nuestro caso, el foco no está puesto en la estructura topológica del grupo, sino en la composición del mismo en términos de las características de los individuos que incluye. Al no haber considerado previamente la información que se usará para validar esos *clusters*, las medidas que la usan se denominan externas. Dado que la información disponible para la calificación depende fuertemente del problema concreto, han sido empleadas múltiples medidas con esta finalidad. Dom (2002) y Wu *et al.* (2009) emplean medidas basadas en teoría de la información entre los *clusters* y las categorías de interés. Por otra parte, Xiong y Li (2013) presentan exhaustivamente múltiples medidas, ejemplificadas con la evaluación de *clusters* generados por el método *k-means*, y discutiendo métodos basados en teoría de la información, entre otros.

Entre las hipótesis que sustentan este trabajo, señalamos que estos tipos de hechos, relativos a la desaparición forzosa de ciudadanos por el propio Estado, no se presentan como acciones aisladas sino que diferentes acciones tienen la

Volumen 9, número 23, (83-96), enero–abril 2021
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

probabilidad de guardar relaciones entre sí. Por consiguiente, la formalización de la red entre individuos y la determinación de las estructuras que emergen de esa red pueden ayudar a pensar nuevas hipótesis de trabajo para quienes investigan los hechos. Además, la identificación y caracterización de los *clusters* de la red ofrecen herramientas para ordenar prioridades en la investigación de personas desaparecidas en el contexto de la última dictadura cívico militar en Argentina, en tanto que brinda conexiones, vínculos, recorridos y nodos en el lamentable proceso de la desaparición de las personas.

Cabe señalar que esta red no es explícita, pues las relaciones relevantes, excepto algunas, no se conocen de antemano sino que se infieren a partir de la información de las personas hasta el momento del secuestro. De tal modo, en este trabajo se caracterizan las estructuras resultantes a partir de una red generada para el conjunto de datos de la provincia de Tucumán, Argentina. Por último, se expondrá la hipótesis principal de este trabajo: que los grupos de personas cuyos secuestros estuvieron muy relacionados entre sí pudieron haber seguido un mismo destino de cautiverio y muerte. Por esta razón, caracterizar los grupos puede ayudar a priorizar las búsquedas durante los trabajos de investigación.

Presentación del caso

Durante la última dictadura militar en Argentina (1976-1983),[1] se establecieron varios circuitos de Centros Clandestinos de Detención (CCD) en diferentes lugares de todo el país, donde las personas desaparecidas fueron detenidas ilegalmente sin ningún tipo de garantías constitucionales, siendo luego la mayoría de ellas asesinada. Incluso hoy, el destino final de la mayor parte de las personas desaparecidas sigue siendo desconocido. Desde 1985, el Equipo Argentino de Antropología Forense (EAAF) utiliza un enfoque multidisciplinario para la investigación y documentación científica de violaciones a los derechos humanos para recuperar e identificar los restos de los desaparecidos, no solo de la dictadura Argentina sino también de otros contextos.[2] Para desarrollar este trabajo, el EAAF recopila datos provenientes de fuentes de información muy variadas, buscando construir hipótesis de identidad que luego son evaluadas en combinación con evidencia genética, en lo que se conoce como *investigación preliminar*. Esta información proviene de fuentes tan variadas como entrevistas a familiares y perso-

1 Para más información, buscar los informes anuales EAAF, años 2005, 2006, 2007 y Especial EAAF-ILLID 2008.
2 El EAAF es una organización no gubernamental sin fines de lucro, que ha estado trabajando en la identificación de los restos de personas desaparecidas en más de 50 países desde 1985. http://www.eaaf.org/

nas liberadas, informes judiciales, notas periodísticas de la época y muchas otras. La identificación de personas en este tipo de contextos masivos es un proceso que comienza con la organización y el análisis de los datos preliminares para guiar las búsquedas, con el fin de construir hipótesis de identidad que luego se evalúan con evidencia genética.

En un trabajo previo se combinaron redes complejas con técnicas estadísticas de validación para determinar un conjunto de reglas o condiciones adecuadas para definir las conexiones entre personas que fueron desaparecidas en la provincia de Tucumán, Argentina (Caridi *et al.* 2011). El propósito fue formalizar una red en la que las personas representan los nodos o puntos de la red, y las conexiones entre ellas se establecen con base en información de tipo geográfica, temporal y política conocida de dichas personas. Por ejemplo, si dos personas fueron secuestradas en lugares cercanos y en momentos cercanos, y además integraban una misma organización política, entonces se establece una conexión entre ellas, porque posiblemente sus secuestros estuvieron relacionados. En esta etapa los hechos fueron estudiados con reglas de este estilo, que combinan la información de las personas con diferentes parámetros. Uno de los objetivos de ese trabajo consistió en lograr que esas conexiones fueran consistentes con la información que era conocida por los investigadores en el mismo contexto de Tucumán respecto a los llamados "grupos de referencia", que son grupos de personas que se sabía que estaban relacionados entre sí, conocimiento que tenían los investigadores del lugar, construido con base en diversas fuentes de información más el propio trabajo de investigación. De modo que, una regla aceptable no debería romper estos grupos de referencia ni pegarlos entre sí.

Entre las reglas aceptables se determinaron las mejores, como aquellas que daban lugar a los mejores grupos en términos de la información acerca del destino de cautiverio de algunas personas (doce CCD que operaron en dicha región). Estos resultados permitieron detectar los grupos más vinculados entre sí y sugerir un Centro Clandestino de Detención (CCD) para algunos de ellos como posible destino de cautiverio de todos los integrantes del grupo.

A partir de la red que resulta de la utilización de las mejores reglas encontradas, se detectaron los grupos de individuos que conforman los llamados *clusters* de la red. Aunque dichas reglas se validaron con información que no fue usada para construir la red, se trata de reglas homogéneas que emplean la misma definición para conectar a dos personas durante todo el periodo y para todas las zonas geográficas. Una definición de la red que pueda adaptarse a las heterogeneidades del problema podría ser más adecuada para representar un fenómeno que necesariamente es heterogéneo, tanto geográficamente, al incluir zonas urbanas y rurales, como temporalmente, al abarcar largos periodos de tiempo, en este caso de más de 4 años.

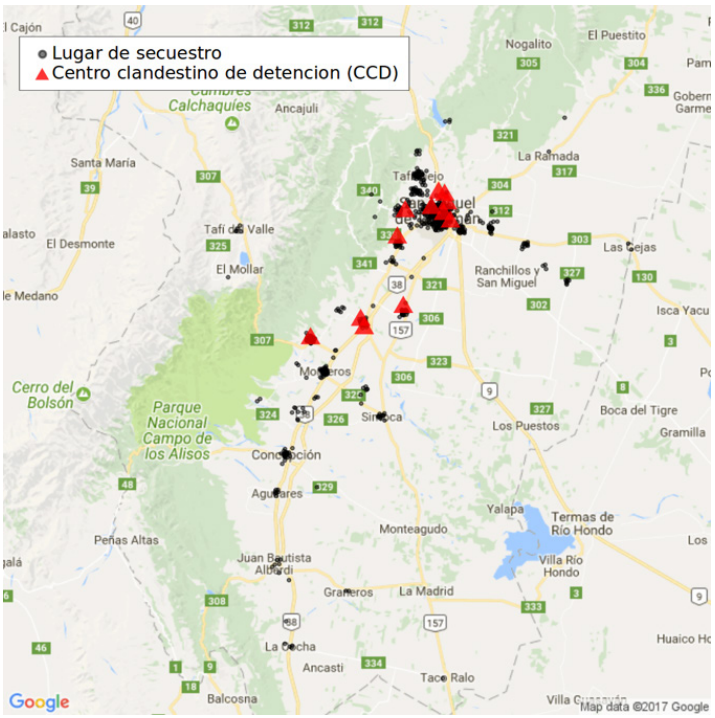
Volumen 9, número 23, (83-96), enero–abril 2021
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

Un requisito necesario para avanzar en el desarrollo de reglas heterogéneas consiste en evaluar la calidad de los resultados que se obtienen de la aplicación de cada regla, en particular, caracterizar los *clusters* que resultan de la red, con base en información que no fue usada para definir la red. Un mecanismo que permita calificar la calidad del *cluster*, en términos de la información disponible y más adelante también de la estructura misma de los *clusters*, puede ayudar a ordenar los resultados en términos de su calidad. También podría ayudar a entender si un *cluster* resultaría mejor si se dividiera en dos de menor tamaño o, por el contrario, si dos de ellos pudieran ser coalescentes entre sí (o sea, que pudieran unirse), permitiendo adaptar la definición de las conexiones a las diferentes condiciones geográfico-temporales, así como en relación con otras variables.

Definición de la red y determinación de las estructuras emergentes

Como ejemplo, dado un conjunto de *N* personas, una regla conecta a dos personas si se cumplen alguna de las condiciones:

Figura 1. Mapa de la provincia de Tucumán, Argentina.



Nota: Los triángulos rojos representan los Centros Clandestinos de Detención que operaron en la provincia, y los puntos negros los eventos de secuestro de personas.
Fuente: Elaboración propia.

a) tienen la misma militancia política, y un código postal cercano (menos de 17 km), y fecha de secuestro que no difiere en más de 7 días (≤ 7);
b) al menos uno tiene militancia política desconocida, y ambos tienen código postal cercano (menos de 17 km) y fecha de secuestro que no difiere en más de 5 días (≤ 5).

Una vez generadas las conexiones entre las distintas personas, se identifican los *clusters* de la red. Esto equivale a etiquetar cada nodo de la red (individuo) con el número del *cluster* al que pertenece. A partir de la regla del ejemplo, estos pueden estar integrados por personas de una misma filiación política como no (que tenían militancia en una misma organización o no), y tener una extensión geográfica y temporal cambiante. Dada la variabilidad del resultado, es útil contar con un mecanismo que permita calificar la calidad del *cluster* en términos de la información disponible y de la estructura de la red.

En la siguiente sección se explica cómo evaluamos la calidad de estos *clusters* en términos de su *similaridad interna*. Los indicadores elegidos para evaluar esta similaridad es ocupacional (trabajo, ocupación) y educacional de las personas, y que no fue usada para definir las conexiones entre las de la red. Si bien la motivación de este trabajo fue aplicarlo al problema argentino, el método es fácilmente generalizable a otras opciones categóricas. Asimismo, el método es directamente aplicable a problemas en los que se busca calificar comunidades en redes densas, en vez de *clusters* en redes de tipo *sparse*, ya que no depende de cómo se hayan determinado los grupos. Se categorizó la información ocupacional en cuatro variables relacionadas con las particularidades del contexto estudiado: trabajadores del ferrocarril (*fc*), trabajadores en ingenios azucareros (*ing*) —categoría que a su vez se puede subdividir por ingenio—, participantes del ambiente universitario (*au*) (incluyendo desde estudiantes hasta personas que trabajaban en la universidad) y empleados públicos (*ep*) (incluyendo docentes del ambiente público, fuerzas de seguridad, administrativos, municipales, legisladores y otros). Las cuatro categorías no son necesariamente excluyentes. Por ejemplo, participantes del ambiente universitario y empleados públicos tienen múltiples integrantes en común, debido al carácter público de las universidades. El resultado de esta categorización es que a cada persona se le asigna una etiqueta para cada categoría, TRUE /FALSE, que señala si pertenece o no a esa categoría.

Presencia de categorías en los clusters

Una vez determinados los *clusters* de la red, los caracterizamos en términos de incluir personas similares respecto al atributo ocupacional. Para eso, nos interesa cuantificar si cierta categoría *c* de las variables ocupacionales tiene o no una presencia fuerte en un dado *cluster*. Supongamos que de los *N* individuos

Volumen 9, número 23, (83-96), enero–abril 2021
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

totales, N_c es la cantidad de personas de la población total que se incluyen en esa categoría. Para medir si la presencia de cierta categoría es fuerte en un *cluster* dado, se procede a comparar el *cluster* observado con lo esperable si este se hubiese formado tomando al azar a los individuos. La comparación con el azar la realizaremos a través de una prueba de hipótesis estadístico. Llamaremos x_{qc} al número de personas en la categoría *c* presentes en el cluster *q*. Asimismo, n_q es el tamaño del *cluster q*. Bajo la hipótesis nula (el *cluster* fue formado completamente al azar), la probabilidad de obtener un *cluster* con x_{qc} personas de la categoría observada sigue una distribución del tipo hipergeométrica, coherente para muestras relativamente pequeñas. A diferencia de una distribución binomial, la distribución hipergeométrica es una distribución discreta que modela el número de eventos en una muestra fija cuando el número de elementos total de la población es conocido. Cada elemento de la muestra puede ser un evento o no, no existiendo remplazo, por lo tanto, la probabilidad de que un elemento sea seleccionado aumenta con cada ensayo. El p valor (*pv*) es la probabilidad de que, bajo la hipótesis nula, se observe un número igual o mayor que x_{qc} de individuos en la categoría *c*. Consideramos que la categoría *c* está fuertemente presente en el *cluster q* si *pv* es menor a cierto valor límite α (denominado en la literatura estadística *significancia estadística*, la cual consiste en un contraste de hipótesis destinada a obtener un valor p inferior a α a fin de poder rechazar la hipótesis nula). Para el conjunto de datos estudiado, elegimos este valor de la significancia para garantizar que ningún *cluster* tenga una presencia fuerte de dos categorías negativamente asociadas (esto es, que a nivel global no se superpongan). A continuación describiremos este proceso.

Relación entre categorías

Las categorías ocupacionales no son disjuntas. La pertenencia de una persona a más de una categoría puede deberse tanto a que su profesión pertenece a más de una categoría como a que tiene más de una profesión. En la tabla 1, pueden verse los tamaños de las intersecciones entre categorías. Es importante tener en cuenta esta información, pues nuestro objetivo es evaluar si la presencia de una categoría es fuerte en un *cluster*, pero considerando que no deberían mezclarse (demasiado) categorías que no se vinculan entre sí, es decir, categorías que no se superponen no deberían estar fuertemente presentes a la vez, teniendo como referencia la población global.

Para cuantificar la superposición entre categorías, empleamos la información mutua puntual (pmi) entre pares de categorías calculadas sobre la base de datos completa. El pmi se define según la ecuación:

$$pmi(x,y) = \frac{P(x,y)}{P(x)P(y)}$$

Tabla 1. Intersección entre las distintas categorías en la población general.

	au	*fc*	*ep*	*ing*
au	806	3	156	8
fc		52	33	0
ep			457	1
ing				96

Nota: La sigla *au* corresponde a ambiente universitario, *fc* a ferrocarril, *ep* a empleados públicos e *ing* a ingenios azucareros. La diagonal indica el número de personas en cada categoría. Las celdas por encima de la diagonal muestran la cantidad de casos compartidos. Al ser los valores de la tabla simétricos solo se representa la diagonal superior.
Fuente: Elaboración propia.

Siendo $P(x,y)$ la fracción de agentes en las categorías x e y, $P(x)$ la fracción de agentes en la categoría x. La base del logaritmo elegida cambia la escala de los valores, pero no el ordenamiento relativo entre los mismos, por lo que no es relevante para el análisis. El pmi mide la correlación entre las variables x e y. Su valor es positivo si $P(x,y) > P(x)P(y)$ (es decir, la probabilidad es mayor que si fueran independientes), y negativo si $P(x,y) < P(x)P(y)$.

Considerando las dos sentencias: x = "la persona pertenece a la categoría c_1", y = "la persona pertenece a la categoría c_2", empleando el *pmi* podemos ver qué categorías tienen una asociación superior a la esperada por puro azar. Si consideramos $c_1 = c_2$, entonces $pmi(x, x) = -\log P(x)$. Mientras más bajo es este valor, mayor es la probabilidad de pertenecer a la categoría c_i, es decir, más común es la categoría en cuestión. Valores positivos de *pmi* indican que si la persona pertenece a la categoría c_1, las posibilidades de que pertenezca a la categoría c_2 aumentan (y viceversa), respecto a lo que resultaría si pertenecer a ambas categorías fueran hechos independientes. De esta forma, tenemos una medida de qué pares de categorías están correlacionadas positivamente y qué pares negativamente. Consideraremos aceptable que un *cluster* tenga presencia fuerte

Tabla 2. Valor del pmi entre las distintas categorías con base en la información completa de la base de datos.

pmi	*au*	*fc*	*ep*	*ing*
au	1.88	−2.33	0.33	−1.79
fc	−	5.84	2.05	$-\infty$
ep	+	*	2.70	−3.88
ing	−			4.95

Nota: En la diagonal inferior se indica si están correlacionadas positiva o negativamente. La sigla *au* corresponde a ambiente universitario, *fc* a ferrocarril, *ep* a empleados públicos e *ing* a ingenios azucareros.
Fuente: Elaboración propia.

Volumen 9, número 23, (83-96), enero–abril 2021
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

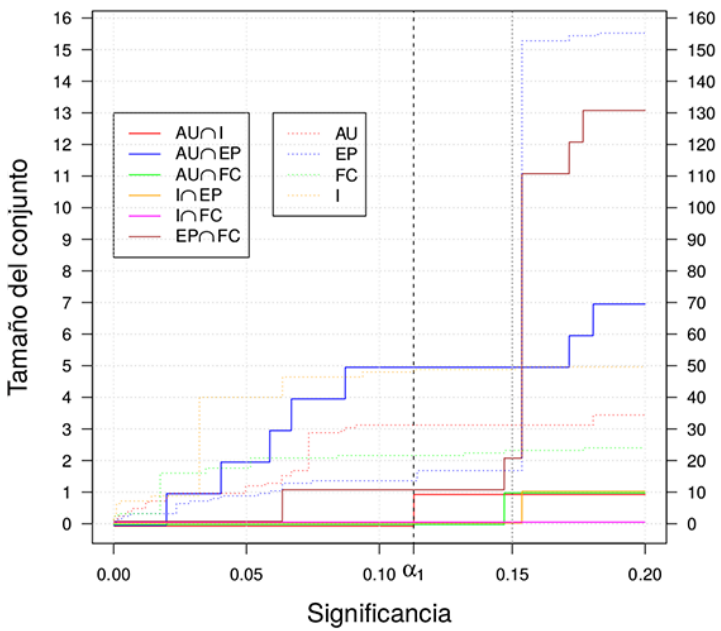
de más de una categoría a la vez solo si esas categorías están positivamente correlacionadas con base en la información completa de la población.

Calidad de un cluster

La propuesta es entonces calificar como "bueno" a un *cluster* que tiene una fuerte presencia de una categoría de la variable ocupacional, sin tener también fuerte presencia en otras categorías correlacionadas negativamente. Como mencionamos, la presencia es fuerte si p, valor asociado a la categoría, es menor a cierta significancia α fijada previamente. Consideramos la totalidad de los *clusters* y evaluamos la cantidad de los buenos de cada categoría en función de la significancia (si *pv* es menor a cierto valor límite α) α. En la figura 2, graficamos esta relación, así como la cantidad de *clusters* que son buenos por tener presencia fuerte en un par de categorías a la vez.

El valor de α a partir del cual consideramos el p valor suficientemente bajo para ser considerado un *cluster* bueno debe ser fijado externamente. Partiendo

Figura 2. Número de *clusters* buenos en función de la significancia α.



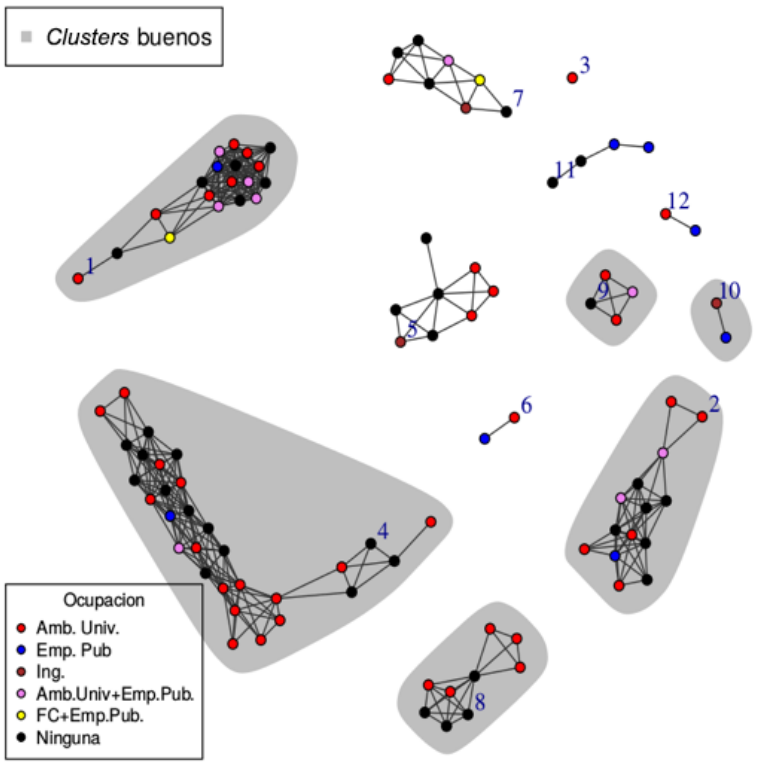
Nota: El eje izquierdo corresponde a las intersecciones y el derecho a las categorías por separado. El valor α_1 indica el punto a partir del cual comienza a haber *clusters* buenos en categorías correlacionadas negativamente. AU: Ambiente universitario, EP: Empleo público, I: Ingenio, FC: Ferrocarril.
Fuente: Elaboración propia.

de la figura 2, y teniendo en cuenta las correlaciones de la tabla 2, elegimos como significancia $\alpha_1 = 0.112$. Este valor de corte asegura que obtengamos el número máximo posible de *clusters* buenos, sin que ninguno de ellos tenga fuerte presencia en categorías anti-correlacionadas.

Resultados

A partir de la categorización de cada *cluster* con base en su composición ocupacional separamos los *clusters* generados por la regla en buenos o malos. Consideramos buenos aquellos *clusters* cuyo *pv* es menor a $\alpha_1 = 0.112$, el valor límite para evitar mezclar categorías correlacionadas negativamente. Asignamos un orden de calidad, según el *pv* más bajo asignado al *cluster*. El mejor *cluster* es aquel que tiene el *pv* más bajo. Observando la figura 3, podemos ver un ejemplo de 12

Figura 3. Ejemplo de distintos *clusters* obtenidos para la regla tomada como ejemplo.



Nota: En gris se resaltan los *clusters* que son buenos en términos de su composición ocupacional.
Fuente: Elaboración propia.

Volumen 9, número 23, (83-96), enero–abril 2021
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

Tabla 3. Categorías presentes en los *clusters* de la figura 3.

Cluster	*AU*	*EP*	FC	ING	Calidad	Orden
1	✔	✔	✘	✘	Bueno	2
2	✔	✘	✘	✘	Bueno	3
3	✘	✘	✘	✘	Malo	-
4	✔	✘	✘	✘	Bueno	1
5	✘	✘	✘	✘	Malo	-
6	✘	✘	✘	✘	Malo	-
7	✘	✘	✘	✘	Malo	-
8	✔	✘	✘	✘	Bueno	6
9	✔	✘	✘	✘	Bueno	4
10	✘	✘	✘	✔	Bueno	5
11	✘	✘	✘	✘	Malo	-
12	✘	✘	✘	✘	Malo	-

Nota: El tilde (palomita) indica que el *cluster* es bueno y la cruz que es malo (en una categoría dada). El orden relativo se basa en los p valores. A menor p valor, un orden más bajo. La sigla *au* corresponde a ambiente universitario, *fc* a ferrocarril, *ep* a empleados públicos e *ing* a ingenios azucareros.
Fuente: Elaboración propia.

clusters con su clasificación. Hay *clusters* buenos tanto grandes como chicos, y con distinta estructura interna (ya que esta no se manifiesta en el análisis de ninguna forma). El criterio del *pv* tiene en cuenta tanto el número de personas en una categoría (en el *cluster* y a nivel global), como el tamaño del *cluster* y el de la población total. Por esta razón, categorías con más integrantes (como ambiente universitario) necesitan un número mayor de integrantes en el *cluster* para satisfacer el criterio. Por otro lado, *clusters* pequeños son más susceptibles a fluctuaciones (efecto considerado naturalmente por el pv). En *clusters* más grandes el pv se vuelve más sensible, haciendo que diferencias más pequeñas se vuelvan más relevantes. Los *clusters* que no agrupan significativamente ninguna categoría son considerados malos, pues no cumplen el objetivo de reunir personas probablemente asociadas. El *cluster* 8 conglomera a más del 50% de personas del ambiente universitario, siendo consistente su clasificación con ello. Algo similar ocurre en el 9, 1, 2 y 4 que tienen composiciones más mezcladas, pero aun así predomina el ambiente universitario lo suficiente como para resultar significativo en cantidad. En el resto de los grupos las cantidades de cada ocupación son pequeñas y por eso no son signficativos. El orden de importancia de los *clusters* está dado por cuán significativo es cada uno. Que el 4 sea el más significativo está directamente asociado con su tamaño, pues al aumentar el tamaño, desvia-

ciones más pequeñas (en proporción) se vuelven más importantes. El resto del ordenamiento es consistente con este efecto, excepto por el *cluster* 8, el cual pierde importancia en el orden al tener un número alto de integrantes en una categoría de por sí numerosa.

Conclusiones

Como se ha podido observar, este trabajo tiene una implicancia en dos niveles diferentes, por una parte, en el ámbito de los derechos humanos constituye una herramienta eficaz para el trabajo de investigación y esclarecimiento en relación con los desaparecidos en la última dictadura militar argentina. O sea, las estimaciones de las estructuras de los vínculos, a partir de determinas reglas o condiciones, puede servir para analizar la probabilidad del recorrido que podrían haber seguido las personas desaparecidas.

Por otra parte, en términos metodológicos, dentro del análisis de redes sociales, representa una manera sencilla de comparar la calidad de grupos encontrados en esta red, contrastándolos con la situación de asociarlos azarosamente. En este sentido, el método propuesto tiene la ventaja de ser fácilmente interpretable en los términos de las poblaciones de cada una de las categorías elegidas. Una vez seleccionados los grupos importantes, los grupos más llamativos pueden ser empleados como punta de lanza para siguientes investigaciones por parte de los expertos, con base en personas ya identificadas que forman parte del mismo grupo.

A su vez, estas pruebas permitieron identificar potenciales problemas de la metodología empleada. De modo que el ensayo ha dejado un serie de interrogante autocrítico sobre el procedimiento, por ejemplo: ¿tiene sentido considerar relevante un grupo que contiene un único individuo en cierta categoría?

En el futuro, sería interesante diseñar métodos que tengan en cuenta además la estructura interna de cada *cluster*, y que puedan nutrirse de este método para decidir si conviene "pegar" o "cortar" grupos con base en la composición de los grupos previos y los resultantes (así como de otros factores geotemporales). Esto permitiría adaptar una regla global homogénea a heterogeneidades temporales y geográficas, entre otras, sin necesidad de definir una regla que se adapte a cada caso. ID

Referencias

Almeida, Hélio, Dorgival Guedes, Wagner Meira y Mohammed J. Zak. 2011. Is there a best quality metric for graph clusters? *Joint European conference on machine learning and knowledge discovery in* databases. Berlin, Heidelberg: Springer, 44-59.

Volumen 9, número 23, (83-96), enero–abril 2021
DOI: https://doi.org/10.22201/ceiich.24485705e.2021.23.77347

Barabási, Albert-László y Albert Réka. 1999. Emergence of scaling in random networks. *Science*, 286(5439): 509-512. https://doi.org/10.1126/science.286.5439.509

Bernard, H. Russell, Eugene C. Johnsen, Peter D. Killworth y Scott Robinson. 1991. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social science research*, 20(2): 109-121.

Boutyline, Andrei y Robb Willer. 2017. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, 38(3): 551-569.

Caridi, Inés, Claudio O. Dorso, Pablo Gallo y Carlos Somigliana. 2011. A framework to approach problems of forensic anthropology using complex networks. *Physica A: Statistical Mechanics and its Applications*, 390(9): 1662-1676.

Dom, Byron E. 2002. An information-theoretic external cluster-validity measure. En *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 137-145.

Erdős, Paul y Alfréd Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1): 17-60, 1960.

Javed, Muhammad. A., Muhammad Shahzad Younis, Siddique Latif, Junaid Qadir y Adeel Baaig. 2018. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, 108: 87–111.

Milgram, Stanley. 1967. The small world problem. *Psychology Today*, 2(1): 60-67.

Newman, Mark E. 2001. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review* E, 64(1): 016131.

Newman, Mark E. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (23): 8577-8582.

Watts, Duncan J. y Steven H. Strogatz. 1999. Small worlds: The dynamics of networks between order and randomness. *Nature*, 393-440.

Wu, Junje, Jian Chen, Hui Xiong y Ming Xie. 2009. External validation measures for K-means clustering: A data distribution perspective. *Expert Systems with Applications*, 36(3): 6050-6061.

Xiong, Hui y Zhongmou Li. 2013. Clustering validation measures. En Charu C. Aggarwal y Chandan K. Reddy (eds.), *Data clustering. Algorithms and applications*. Londres: Taylor and Francis Group, 571-602.