

Complejidad computacional de un caso especial del cálculo de grupos en las redes de interacción

Roberto Alonso, Raúl Monroy y Luis A. Trejo

Departamento de Ciencias Computacionales
Instituto Tecnológico y de Estudios Superiores de Monterrey,
Campus Estado de México, México
{A01013374,raulm,ltrejo}@itesm.mx

Resumen En este trabajo proponemos una aproximación para realizar una demostración de la complejidad computacional del problema del Cálculo de Grupos. El Cálculo de Grupos es un problema del modelo social conocido como Redes de Interacción [5] que resulta ser complejo. Mediante reducciones de Karp realizamos una demostración de un caso especial de nuestro problema que consideramos es esencial para lograr una demostración completa del problema del Cálculo de Grupos. Si se determina la complejidad computacional del Cálculo de Grupos, nos aseguramos de que el problema es realmente complejo y sólo puede ser resuelto parcialmente mediante heurísticas.

1. Introducción

En el modelo Redes de Interacción [5] existe un problema llamado *problema del Cálculo de Grupos*, este cálculo de estructuras sociales es necesario para entender mejor el comportamiento de un grupo de usuarios que hacen uso de recursos. Este conocimiento a su vez será útil para la detección de ataques de Denegación de Servicio Distribuidos (DDoS por sus siglas en inglés) dirigidos a servidores DNS.

Para entender mejor los problemas, la teoría de la complejidad clasifica los problemas en función de su dificultad para resolver estos. Típicamente la clase P representa aquellos problemas que pueden ser resueltos en tiempo polinomial, mientras que la clase NP representa aquellos problemas que sólo pueden ser resueltos con una máquina no-determinística de Turing (NDTM por sus siglas en inglés). En este trabajo mostramos que un caso especial del Cálculo de Grupos (**SGC**) es un problema NP-Completo. El paso presentado en este trabajo será esencial para demostrar que la versión generalizada del problema es también NP-Completo.

Para demostrar que el Cálculo de Grupos (SGC) en su forma generalizada es NP-Completo requerimos de hacer una reducción en tiempo polinomial de un problema NP-Completo a nuestro problema. Además es necesario demostrar que una respuesta en el problema NP-Completo existe *sii* hay una respuesta en nuestro problema.

El trabajo está dividido de la siguiente manera: En la sección Redes de Interacción y Complejidad del cálculo de grupos abordamos el tema del modelo social y mostramos el cálculo de la complejidad temporal para calcular todas las posibles estructuras sociales. De forma general presentamos la reducción de Karp, en la sección 3, que consiste en una técnica para realizar de forma simple demostraciones de problemas nuevos. En la sección 4 mostramos cómo se realizó la reducción del problema del *Hitting Set* al caso especial del SGC. En las conclusiones, en la sección 5, presentamos algunas ideas de como puede usarse la reducción de este trabajo para demostrar que el problema general de decisión del SGC es NP-Completo.

2. Redes de interacción y complejidad del cálculo de grupos

La hipótesis manejada en las Redes de Interacción, se basa en el uso de perfiles sociales derivados de la caracterización de usuarios de red que hacen uso del servicio DNS, y en el cual se observan las interacciones que tienen estos usuarios con un servidor DNS. Un **usuario** es aquella entidad en una red que solicita resolución DNS. Por lo tanto, consideraremos que todas las peticiones realizadas a través de un dispositivo de red, sin importar si es realizada por un software o un usuario de computadora, son peticiones de un mismo **usuario** si y sólo si, provienen de la misma IP.

El objetivo principal de este enfoque, es el de caracterizar los perfiles sociales de los usuarios, ya que esto nos permitirá observar comportamientos de uso de recursos DNS. Es decir, estaremos definiendo características de los usuarios que nos permitirán distinguir lo que denominaremos como normalidad social y anormalidad social. Por comodidad este modelo es nombrado **Redes de Interacción** y será representado como un grafo nombrado *grafo de conectividad*.

Al obtener parte de la caracterización social, podemos esperar relaciones entre los dominios que visita un usuario y otro. Esto quiere decir que nuestros hábitos de navegación a través de Internet son similares o iguales a los de otra persona. Por ejemplo, el dominio `www.ask.com` sería una de las últimas opciones como buscador a comparación de `www.google.com`. Por esa razón, sabemos que existen patrones de consultas comunes realizadas a un servidor DNS recursivo, estas consultas comunes entre más de un usuario se denomina **grupo**.

2.1. Complejidad temporal del cálculo de grupos

Un **grupo** es un conjunto de **k** usuarios que tienen en común la consulta de **n** recursos. La complejidad temporal para determinar todos los posibles grupos en una Red de Interacción está dada por la siguiente fórmula:

$$\sum_{k=2}^{N \leq r} C(r, k)$$

donde r es el número de recursos únicos en un conjunto de peticiones, y N es el tamaño del grupo más grande asumiendo que conocemos ese tamaño o que el algoritmo continúa corriendo hasta que no encuentre un grupo.

Encontrar todos los grupos posibles dado un *grafo de conectividad* es un problema intratable y lo llamaremos problema **SGC general**. El problema generalizado del SGC requiere de grandes recursos computacionales y puede ser considerado como un problema de enumeración en el que calcular y escribir la respuesta es igual de complejo para ciertas instancias.

Desde el punto de vista de la teoría de la complejidad computacional, debemos acotar la versión del SGC general a una versión de decisión que pueda ser resuelta con una NDTM, esta acotación será nombrada **SGC- δ** . Los problemas de decisión son aquellos que pueden ser contestados con la respuesta 'si' en caso de existir una respuesta válida y 'no' en caso de que sea inexistente una respuesta válida para el problema. Demostrar que la versión de decisión del problema del Cálculo de Grupos es uno de los pasos para comprobar que el problema **SGC general** es complejo.

En este trabajo mostramos que un caso especial del Cálculo de Grupos es NP-Completo. Esta demostración podría ser usada para demostrar que **SGC- δ** es NP-Completo y posteriormente que el problema **SGC general** está en la clase #P.

3. Determinar la complejidad computacional de problemas poco conocidos

La teoría de la complejidad clasifica los problemas en función de la dificultad para resolverlos. Asumiendo que $P \neq NP$, los problemas en P son aquellos tratables mientras que los NP son la clase de problemas que sólo pueden ser resueltos por una Máquina de Turing No-Determinista.

El problema del cálculo de grupos no pertenece a la clase P dado que la cantidad de recursos computacionales necesarios para resolverlo crece exponencialmente con respecto al número de dominios consultados por los usuarios. Sin embargo, es insuficiente decir que el problema es NP-Completo sólo porque se sospecha no pertenece a P por tanto, usaremos una Reducción de Karp [1,2] para demostrar que el problema del Cálculo de Grupos es NP-Completo.

Una Reducción de Karp es una de las formas para demostrar que un problema pertenece a la clase NP. Está basada en el trabajo que realizó Cook-Levin acerca de la primera demostración de un problema NP. Karp establece por transitividad que si un problema cuya clase es desconocida puede ser mapeado de un problema NP-Completo como SAT, entonces hay un prueba de que el problema desconocido (π) pertenece a la clase NP. Es decir,

$$NP \leq_p SAT \leq_p \pi$$

por transitividad,

$$NP \leq_p \pi$$

Una reducción de Karp es uno de los métodos preferidos para demostrar si un nuevo problema es NP-Completo. Para realizar una reducción de Karp se deben seguir los siguientes pasos:

- Demuestra que el problema está en NP.
- Selecciona un problema NP-Completo.
- Realiza una reducción en tiempo polinomial del problema NP-Completo al problema a ser demostrado.

Una reducción simplemente se refiere al mapeo de un problema a otro problema. La reducción siempre se tiene que hacer del problema conocido NP al problema No-conocido.

4. Reduciendo el problema de Hitting Set a un caso especial del problema del cálculo de grupos

La versión generalizada del problema del Cálculo de Grupos (**SGC**) pretende determinar todos los posibles grupos dado un grafo de conectividad. Sin embargo, para realizar la demostración de que un problema es NP-Completo, requerimos de hacer una acotación de tal forma que sea posible definir **SGC** como un problema de decisión.

4.1. Problema de decisión del SGC y el problema del Hitting Set (HS)

Por comodidad, la versión de decisión del problema **SGC**, será nombrada **SGC- δ** y está definida de la siguiente forma:

Dados un grafo de conectividad $G = (V, E)$, dos conjuntos disjuntos $I = \{i_1, i_2, i_3, \dots, i_k\}$ y $D = \{d_1, d_2, d_3, \dots, d_l\}$ representando los nodos $V = I \cup D$ y las consultas de los usuarios i_k representada por los arcos, ¿existe un grupo de tamaño 2?

El problema anterior es la definición del problema **SGC- δ** que es de interés para las Redes de Interacción. Sin embargo, vamos a utilizar una aproximación similar a la presentada en [3,4] para hacer la demostración de ese problema. Dicha aproximación consiste en demostrar un caso especial del problema general **SGC** para utilizarla en el problema de interés. Para lograr lo anterior definimos un nuevo problema **SGC** nombrado **SGC- σ** de la siguiente forma:

Dado un grafo de conectividad $G = (V, E)$, ¿existe una interacción completa? Es decir, en el grafo de conectividad, ¿todos los usuarios forman parte de al menos un grupo de tamaño dos ?

Proponemos realizar una reducción de Karp utilizando el problema de Hitting Set a un caso particular del problema del Cálculo de Grupos (**SGC**). Hitting Set (**HS**) es un problema NP-Completo definido de la siguiente forma:

Dado un conjunto finito S , una colección C de subconjuntos de S , un número positivo $K \leq |S|$, ¿existe un subconjunto $S' \subseteq S$ tal que S' contiene al menos un elemento de cada subconjunto de C y $|S'| \leq K$?

Teorema 1 (SGC- σ es NP completo).

Demostración. El problema está en NP porque una NTDM puede resolver el problema en tiempo no-determinístico y verificar la respuesta en tiempo polinomial. Para hacer la transformación $\mathbf{HS} \leq_p \mathbf{SGC}\text{-}\sigma$, debemos de construir un grafo de conectividad $G(V,E)$ tal que existe una interacción completa si y sólo si existe un HS en S.

Podemos transformar \mathbf{HS} a $\mathbf{SGC}\text{-}\sigma$ de la siguiente forma. Sea $S = \{s_1, s_2, s_3, \dots, s_n\}$, sea $\mathcal{C} = \{C_1, C_2, C_3, \dots, C_m\}$ una colección de subconjuntos de S, y sea $K \leq |S|$ una instancia de \mathbf{HS}

Por cada elemento s_i vamos a construir un componente nombrado subgrafo social. El subgrafo social se construye de la siguiente forma:

1. Agrega un usuario y dos recursos que estén siendo consultados por este usuario. (Fig. 1).

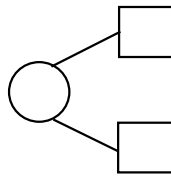


Figura 1. Usuario consultando dos recursos.

2. Si existe una colección C_j que contenga s_i agregar un usuario etiquetado como C_j conectado a los recursos agregados recientemente. (Fig. 2).

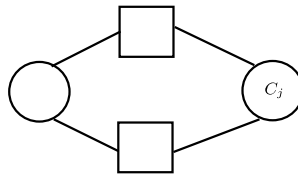


Figura 2. Usuario C_j conectado a dos recursos.

3. Agregar un usuario conectado a uno de los recursos que no tenga grado mayor a 2 y agregar un recurso más conectado al nuevo usuario. (Fig. 3).
4. En caso de existir otro C_j que contenga s_i repetir los pasos 2 y 3.
5. Si ya no existen más C_j que contengan s_i encerrar el componente en una caja etiquetada s_i . (Fig. 4).

Cada colección C_j corresponde a un usuario i_j en el problema $\mathbf{SGC}\text{-}\sigma$. Por simplicidad, el usuario está representado como un círculo y etiquetado en el grafo que se construye como *usuario* C_j , este usuario hace consultas a recursos 'dummy'. Dado que nos interesa la formación de una interacción completa, no es relevante con que recursos se esté formando esa relación.

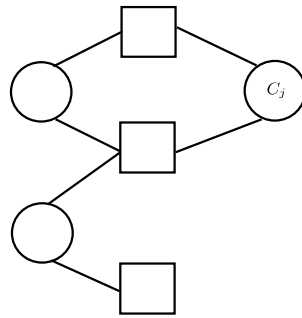


Figura 3. Usuario nuevo encadenado al componente.

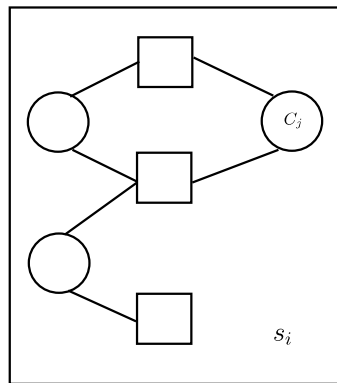


Figura 4. Usuario nuevo encadenado al componente.

Ejemplo de transformación dada una instancia del HS. Sea $S = \{s_1, s_2, s_3, s_4\}$ y $C_1 = \{s_1, s_2, s_3\}$, $C_2 = \{s_2, s_4\}$, $C_3 = \{s_1, s_2, s_4\}$ y $K=2$ una instancia del problema HS. Siguiendo la transformación obtenemos un grafo que puede contener una interacción completa. (Fig. 5).

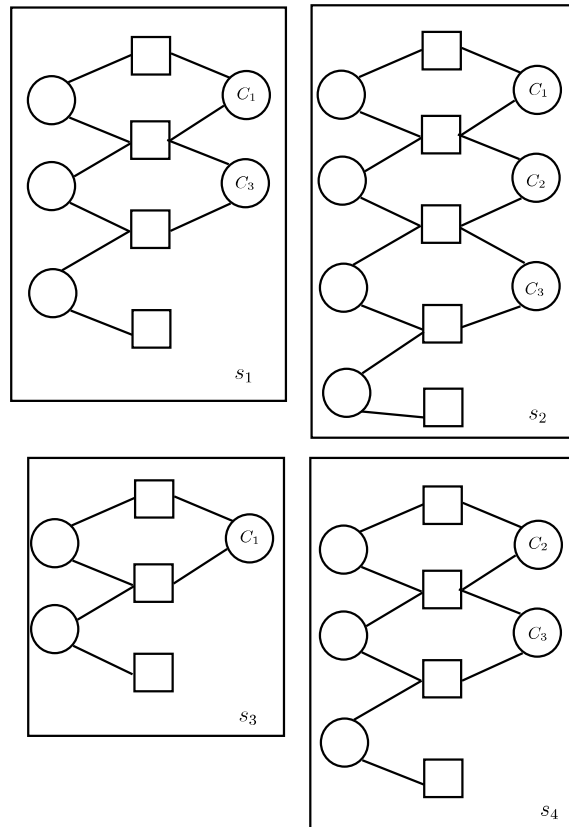


Figura 5. Grafo resultante de la instancia $S = \{s_1, s_2, s_3, s_4\}$ y $C_1 = \{s_1, s_2, s_3\}$, $C_2 = \{s_2, s_4\}$, $C_3 = \{s_1, s_2, s_4\}$.

Una vez transformado el problema, es necesario demostrar que existe una solución en el problema NP-Completo si y sólo si existe una solución en el problema nuevo. Es decir, si existe un HS en la instancia presentada, debe existir una interacción completa en el grafo y si no hay una respuesta válida en la instancia de HS tampoco se debe encontrar una interacción completa en el grafo. Recordando el ejemplo anterior, una respuesta del problema HS válida sería $S' = \{s_1, s_4\}$ ya que contiene al menos un elemento de cada C_j y además $|S'| = 2$.

Para determinar si es posible encontrar una interacción completa seleccionamos los componentes correspondientes en el grafo. La unión de ambos componentes $\{s_1\} \cup \{s_2\}$ es el grafo que debe contener la interacción completa (Fig. 6).

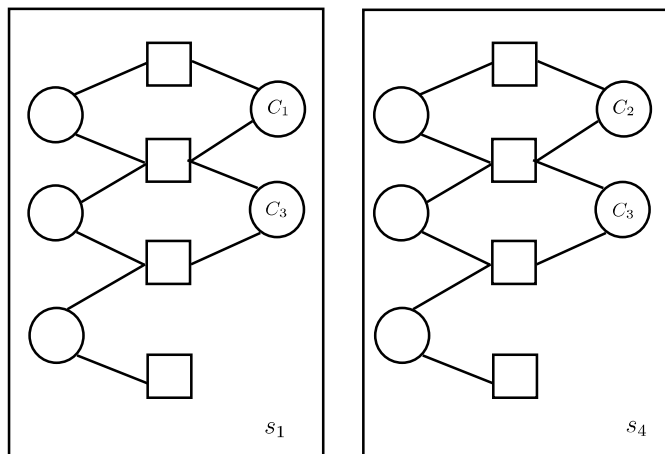


Figura 6. Resultado de la unión de s_1 y s_4 .

Podemos notar que al seleccionar esos dos componentes, estamos encontrando que todos los usuarios intervienen en la formación de al menos un grupo de tamaño dos por lo que es una respuesta válida en el problema **SGC- σ** .

Si seleccionamos $S' = \{s_1, s_3\}$ como respuesta del problema HS podemos notar que *no* es una respuesta válida. Seleccionando los componentes correspondientes en el grafo de conectividad (Fig. 7), podemos notar que no todos los usuarios están formando al menos un grupo.

Por lo tanto, queda demostrado que $\text{HS} \leq_p \text{SGC-}\sigma$ y en consecuencia $\text{NP} \leq_p \text{SGC-}\sigma$. ■

La importancia de la reducción en el caso especial radica en que tenemos un problema NP-Completo comprobado que está dentro del dominio de nuestro problema y por tanto, puede ser utilizado para realizar la demostración de $\text{SGC-}\delta$. Es decir, $\text{SGC-}\sigma \leq_p \text{SGC-}\delta$ es una aproximación para demostrar que encontrar un grupo de tamaño dos es complejo.

5. Conclusiones y trabajo futuro

Las redes de interacción es un modelo temporal que se construye a partir del análisis de datos sociales. Está basado en la caracterización de los usuarios y los recursos. Los usuarios y los recursos tienden a formar grupos con respecto a las preferencias que tienen estos usuario. Calcular estos grupos es esencial para la

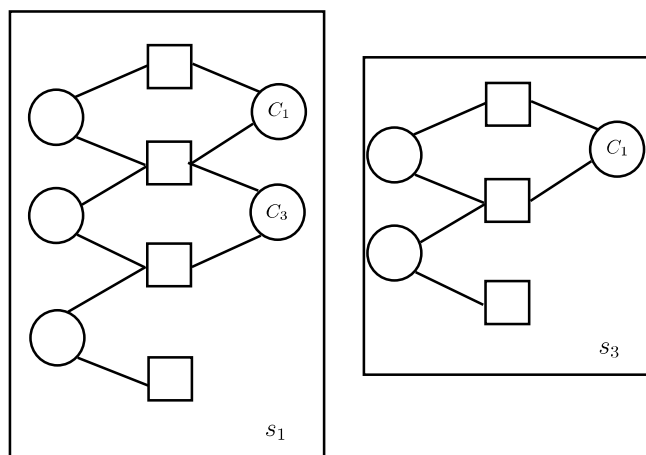


Figura 7. Resultado de la unión de s_1 y s_3 .

detección de anomalías sociales y en última instancia la detección temprana de ataques DDoS a servidores DNS.

El cálculo de grupos es complejo dado que no existe un algoritmo polinomial capaz de resolver el problema de forma correcta y completa para cualquier instancia del problema. Por tanto, se requiere de diferentes métodos para aproximar una solución completa. Podemos aplicar diferentes métodos para aproximar el cálculo de estos grupos, algunos relacionados con segmentación de imágenes o manipulación de matrices.

Queda demostrado que el caso especial del problema SGC está en la clase NP y es NP-Completo. Esta demostración puede ser usada para continuar la reducción al problema SGC- δ ya que ambos problemas se encuentran en el mismo dominio. Podemos utilizar la reducción SGC- σ simplemente copiando la instancia del problema. Si existe una interacción completa es evidente que existe un grupo de tamaño 2, aunque si existe un grupo de tamaño 2 no implica la existencia de una interacción completa.

Los resultados obtenidos en este trabajo son prometedores. A pesar de no ser el problema que buscamos exactamente es una aproximación y un paso inicial para demostrar que la versión SGC- δ es NP-Completo. Más aún, puede darnos una idea de cómo hacer la demostración de la versión general del SGC en el que se puede preguntar si existen varios grupos de diferentes tamaños, lo cual se sospecha también es un problema complejo.

Al demostrar la versión generalizada del SGC podemos concluir que la versión de optimización de nuestro problema es NP-Hard y la versión de enumeración pertenece a la clase #P. Este conocimiento nos ayudará a entender mejor el problema del Cálculo de Grupos y proponer diferentes heurísticas que nos ayuden a resolver el problema de manera eficiente aunque incompleta.

Referencias

1. Karp, R.M.: Reducibility Among Combinatorial Problems, Complexity of Computer Computations, Miller, R.E and Thatcher, J.W. editors, pp. 85-103 (1972)
2. Garey M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman and Company (1991)
3. Papadimitriou, Ch.: The NP-Completeness of the bandwidth minimization problem, Springer Wien, pp. 263-270 (1976)
4. Blanche, G., Karpinski, M., Wirtgen, J.: On Approximation Intractability of the Bandwidth Problem (1997)
5. Alonso, R., Vázquez, J., Trejo, L., Monroy, R., Sánchez, E.: How Social Networks can Help to Detect DDoS Attacks on DNS Servers, Proceedings of the 3rd MICAI Workshop on Computer Security in 8th Mexican International Conference on Artificial Intelligence, WSEC09-MICAI '09 (2009)