Optimality of Spectral Sparsification in Distributed Systems with Application to Data Clustering

Fabricio Mendoza Granada

Núcleo de Investigación y Desarrollo Tecnológico
Facultad Politécnica - Univervisad Nacional de Asunción
fabromendoza95@gmail.com

Resumen

La técnica de dispersión espectral es utilizada en una variedad de aplicaciones tales como resolución de sistemas Laplacianos y búsqueda de multicortes en un grafo mediante sus propiedades espectrales. Esencialmente, aproxima el espectro de un grafo por un factor constante reponderando algunas de sus aristas y eliminando otras. En este trabajo, demostramos que la dispersión espectral funciona bajo la suposición de que los datos de entrada están distribuidos en diferentes sitios. Esencialmente, demostramos que si G es un grafo con sus aristas distribuidas alrededor de s sitios entonces existe un grafo $H = (V, \cup_{i=1}^{s} \hat{E}_i)$ tal que H es un dispersor espectral de G y $\hat{E}_i$ es el conjunto de aristas reponderadas en el sitio i.

La idea de datos que se solapan entre diferentes sitios nos ha inspirado a estudiar modelos de comunicación, los cuales funcionan como herramientas teóricas para estudiar algoritmos que trabajan con datos distribuidos. En particular, nos hemos enfocado en el modelo Number-On-Forehead, el cual es una poderosa herramienta con aplicaciones teóricas en complejidad de circuitos. Nuestro segundo resultado es un protocolo de comunicación que aproxima el multicorte de un grafo dado en el modelo Number-On-Forehead bajo la suposición de que la familia de aristas tiene intersección única.

1. Motivación

El rápido incremento de disponibilidad de datos provenientes de múltiples fuentes y la necesidad imperiosa de procesarlos utilizando una cantidad óptima de recursos motivan este trabajo. Esta tesis se enfoca en el estudio de la bien conocida técnica de aproximación de grafos llamada *dispersión espectral*. Pretende contribuir algunos resultados teóricos acerca de dispersión espectral en sistemas distribuidos donde no todos los datos están disponibles para cada sitio. Además, los datos puede que aparezcan repetidos múltiples veces, por ende las propiedades espectrales del grafo resultante deberían ser consistentes con las del grafo original.

2. Dispersión Espectral

La dispersión espectral es una técnica de aproximación de grafos basada en la aproximación de las matrices Laplacianas de dichos grafos. La matriz Laplaciana de un grafo de define como

$$L_G = D_G - A_G,$$

donde D_G y A_G son las matrices de grado ponderado y adyacencia ponderada respectivas de G. Esencialmente, la técnica de dispersión espectral toma un grafo G como entrada y construye un grafo H con la siguiente características

1. H tiene menos aristas que G,
2. El espectro de G y H son próximos por un factor constante.

En este caso consideramos el espectro de un grafo G como el conjunto de autovalores de su Laplaciano correspondiente. Formalmente, se dice que H es un dispersor ϵ-espectral de G si y solo si

$$(1-\epsilon)x^T\mathcal{L}_G x \leq x^T\mathcal{L}_H x \leq (1+\epsilon)x^T\mathcal{L}_G x$$

donde $\mathcal{L}_G$ y $\mathcal{L}_H$ son los Laplacianos normalizados de G y H respectivamente y $x \in \mathbb{R}^n$ con $n = |V|$.

3. Partición por Cardinalidad de Solapamiento

La *partición por cardinalidad de solapamiento* es una manera de particionar un conjunto $A = \bigcup_{i=1}^{t} A_i$ con respecto al número de veces que cada elemento $a \in A$ aparece en la familia $\{A_i\}_{i \leq t}$ donde $t \in \mathbf{N}$ y $A_i \subseteq A$. Un ejemplo se puede observar en la siguiente figura

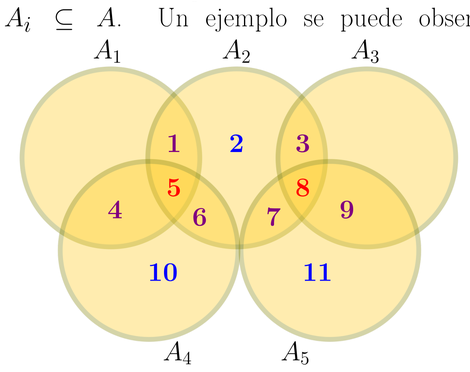

Figura 1: *En este ejemplo podemos observar que el conjunto* $\{1, \ldots, 11\}$ *puede particionarse en* $\{\{2, 10, 11\}, \{1, 3, 4, 6, 7, 9\}, \{5, 8\}\}$. *Notar que cada elemento de un mismo conjunto en la partición aparece el mismo número de veces en la familia de conjuntos. Por ejemplo, 5 y 8 aparecen 3 veces. Este número de ocurrencias es llamado* cardinalidad de solapamiento.

La contribución principal de este trabajo es la demostración de que la técnica de dispersión espectral funciona cuando los datos están alojados en diferentes sitios. El siguiente teorema expresa nuestro resultado principal

Teorema 1 *Sean* $(1 = c_1 < c_2 < \cdots < c_k)$ *las cardinalidades de solapamiento sobre la familia* $\{E_i\}_{i \leq t}$ *con* $\{E'_{c_j}\}_{j \leq k}$ *su partición por cardinalidad de solapamiento asociada y* $L_{G_1}, \ldots, L_{G_t}$ *los Laplacianos de* $G_1, \ldots, G_t$. *Si* $H_i = (V, D_i, h_i)$ *es un dispersor* ϵ*-espectral de* G_i, *entonces* $H = (V, \bigcup_{i=1}^{t} D_i, h)$ *es un dispersor* ϵ'*-espectral de* G *donde* $h(e) = \frac{\sum_{i=1}^{t} h_i(e)}{c_1 c_k}$ *y* $\epsilon' \geq 1 - \frac{1-\epsilon}{c_k}$.

El Teorema 1 dice que la unión de dispersores espectrales de un grafo dado es un dispersor espectral de ese grafo. Esto se puede observar en la siguiente figura

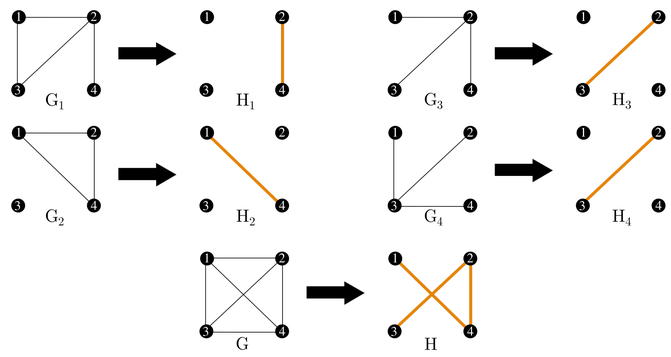

Figura 2: *En la parte de arriba podemos observar una familia de subgrafos de* K_4 *con sus respectivas aproximaciones espectrales a la derecha de cada uno. En la parte de abajo se muestra lo que el Teorema 1 clama, esto es, la unión de dispersores espectrales es un dispersor espectral de la unión de subgrafos.*

4. Aplicaciones en el Problema de Agrupamiento

El agrupamiento es una técnica de aprendizaje no supervisado que puede ser visto como un proceso de multicorte en el contexto de grafos. El resultado de esta sección hace uso del Teorema 1 para aplicar la técnica de agrupamiento en un contexto distribuido. Para aplicar nuestro resultado principal utilizamos un modelo de solapamiento de datos que nos permita reducir el costo de comunicación en el ambiente distribuido. El mismo está representado por la estructura combinatorial llamada *girasol* o *Sistema-*Δ. Un *girasol* o *Sistema-*Δ es una familia de conjuntos $\mathcal{A} = \{A_1, \ldots, A_t\}$ donde $A_i \subseteq [n]$ y $(A_i \cap A_j) = \bigcap_{k=1}^{t} A_j = K$ para todo $i \neq j$ y cualquier natural n. Para medir la cantidad de recursos utilizados para computar un multicorte en un sistema distribuido estudiamos *complejidad de la comunicación*. La complejidad de la comunicación estudia la cantidad de bits que deben ser compartidos por varios computadores en un sistema distribuido para poder realizar una tarea. Existen varios modelos de comunicación y en este trabajo estudiamos el modelo Number-On-Forehead (NOF), en cual cada sitio tiene acceso a la información de los demás pero no el suyo. Finalmente, el multicorte de un grafo puede ser aproximado por una técnica llamada *agrupamiento espectral* que utiliza el espectro de un grafo para encontrar una partición óptima de sus vértices. La técnica de agrupamiento espectral consiste en los siguientes pasos generales

1. Construir el Laplaciano del grafo G, L_G,
2. Computar el Laplaciano normalizado $\mathcal{L}_G = D_G^{-1/2} L_G D_G^{-1/2}$,
3. Hallar los autovalores de $\mathcal{L}_G$,
4. Tomar los autovalores en orden no descendente $\lambda_0 = 0 \leq \lambda_1 \leq \cdots \leq \lambda_n$
5. Construir la matriz X con los autovectores correspondientes a los k primeros autovalores como columnas
6. Aplicar k-means sobre las filas de X

El resultado principal de esta sección permite que cada sitio compute agrupamiento espectral sobre una aproximación del grafo de datos subyacente. El mismo se expresa en el siguiente teorema

Teorema 2 *Sea* $\{E_i\}_{i \leq s}$ *un Sistema-*Δ *débil con cada* $|E_k| = \ell$ *para* $k = 1, 2, \ldots, s$, *y suponga que* $s \geq \ell^2 - \ell + 3$. *Existe un protocolo de comunicación tal que luego de dos rondas de comunicación cada sitio conoce un dispersor* ϵ*-espectral del grafo completo* G *con un costo de comunicación* $O\left(\log\left(\frac{n}{\epsilon^2}\sqrt{1-\delta}\right)\right)$.

Luego de aplicar el protocolo resultante del Teorema 2 cada sitio puede computar el multicorte de G con el algoritmo de agrupamiento espectral.

5. Conclusiones

- En este trabajo hemos demostrado que la unión de dispersores espectrales de subgrafos de un grafo dado G es un dispersor espectral de G,
- Dimos una computación exacta del factor de aproximación espectral ϵ' para la unión de dispersores espectrales y,
- Presentamos una aplicación de la unión de dispersores espectrales para computar el agrupamiento en un problema con datos distribuidos y solapados.

Articulos Publicados

[1] Mendoza-Granada, F., & Villagra, M. (2020). A Distributed Algorithm for Spectral Sparsification of Graphs with Applications to Data Clustering. In: *Proceedings of the 18th Cologne-Twente Workshop on Graphs and Combinatorial Optimization (CTW)*

[2] Mendoza-Granada, F., & Villagra, M. (2019, July). Number-On-Forehead Communication Complexity of Data Clustering with Sunflowers. In Anais do IV Encontro de Teoria da Computação. SBC.

[3] Granada, F. A. M., & Villagra, M. (2018). Distributed spectral clustering on the coordinator model. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, 6(2).

[4] Granada, F. A. M., Mercado, S., & Villagra, M. (2018). Deterministic Graph Spectral Sparsification. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, 6(2).

CONACYT

PROCIENCIA
PROGRAMA PARAGUAYO PARA EL DESARROLLO DE LA CIENCIA Y TECNOLOGÍA