Muestreo Probabilístico en Gráficas

Sergio Rajsbaum* Carlos Velarde †

(Mayo 16,1997)

Resumen

Se revisan trabajos recientes de David Karger donde se explora el muestreo como una herramienta para resolver problemas en gráficas y se prueba que la muestra, o *esqueleto* resultante de muestrear aleatoriamente las aristas de una gráfica aproxima los valores de *todos* los cortes de la gráfica original. Esto es útil para disminuir la complejidad de algoritmos de conexidad conocidos.

1 Introducción

En estadística se entiende por *muestra* una fracción de una población seleccionada según un critrerio determinado, con el fin de estudiar sobre ella ciertas características, válidas para sacar conclusiones sobre las de la población total. En el muestreo son importantes los procedimientos para determinar el tamaño de la muestra, de acuerdo a que tan precisamente se requiere que ésta represente las características de interés.

El muestreo es importante porque ahorra tiempo y da la posibilidad de hacer una revisión minuciosa de los componentes de la muestra. De hecho, mientras mayor sea la precisión requerida, la muestra será mayor y, por lo tanto, el tiempo ahorrado será menor. Éstas son algunas razones por las que el uso del muestreo es tan común, por ejemplo en encuestas, en control de calidad o cuando se toma una muestra de sangre para analizar su composición.

Aquí veremos una forma novedosa de muestreo en una *gráfica* G —estructura matemática constituída por un conjunto de *vértices* V y un conjunto de parejas de vértices E llamadas *aristas* (Figura 1). Las gráficas sirven para modelar gran variedad de situaciones. Por ejemplo, la red Internet se puede representar mediante una gráfica, donde los vértices corresponden a computadoras y las aristas a líneas de intercomunicación.

*Instituto de Matemáticas, U.N.A.M., México, rajsbaum@servidor.unam.mx

†Unidad de Cómputo para la Investigación, D.G.S.C.A., U.N.A.M., México, velarde@servidor.unam.mx

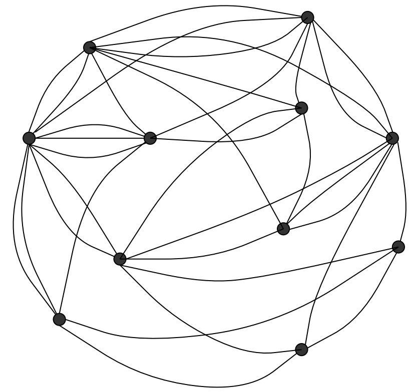

Figura 1: *Ilustración de una gráfica: los puntos representan vértices y los arcos aristas.*

2 Muestreo en Gráficas para Conexidad

Un *corte* en una gráfica G es una partición del conjunto de vértices en dos conjuntos no vacíos; el *valor* del corte es el número de aristas con un vértice en cada conjunto del corte. Por ejemplo, en la gráfica de la figura 1, el corte que separa los cinco vértices de la izquierda de los seis vértices de la derecha vale 15.

Los cortes son importantes porque pueden usarse para definir y calcular otras propiedades de la gráfica. Así, la *conexidad* de G se define como el valor del corte mínimo. Dicho en términos de la gráfica de computadoras antes mencionada, la conexidad nos da el mínimo número de aristas que deben fallar para que dos computadoras queden sin camino de intercomunicación entre ellas. Es claro que la conexidad de una gráfica finita se puede determinar probando todas las maneras de partirla en dos, contando en cada caso cuántas aristas cruzan de una parte hacia la otra y al final tomando el mínimo de todos estos números. Pero este proceso es prohibitivamente tardado (de tiempo exponencial), inclusive para gráficas mucho más pequeñas que Internet. Afortunadamente, existen algoritmos eficientes (de tiempo polinomial) para calcular el valor del corte mínimo; no obstante, para gráficas muy grandes, como puede ser la de Internet, sería deseable volverlos aún más eficientes.

En trabajos recientes, Karger [6, 3, 4, 5] explora el muestreo como una técnica para resolver de manera mas eficiente problemas de optimización en gráficas. La idea es tomar una muestra aleatoria de la gráfica, cuidando que sea representativa de las propiedades de conexidad de la gráfica dada. Sobre la muestra —gráfica con menos aristas que la original— es posible correr cualquier algoritmo de conexidad, confiando en que se obtendrán los mismos resultados de conexidad, pero mas rápido, ya que el tiempo de ejecución del algoritmo está en función del tamaño de la gráfica. En sus trabajos, Karger prueba que si la gráfica muestreada es suficientemente grande, entonces con muy alta probabilidad la muestra preserva las propiedades de conexidad, y por lo tanto, con muy alta probabilidad el algoritmo al correrse sobre la muestra responderá correctamente.

El ejemplo de Internet es uno de tantos problemas de la vida real que se pueden modelar como un problema de conexidad en gráficas. Las mismas técnicas se pueden aplicar para resolver problemas de flujo en redes, como los asociados a un sistema de distribución de bienes através de una red de carreteras. En este caso, los vértices de la gráfica corresponden a lugares de producción o consumo de bienes y las aristas representan carreteras por donde éstos se han de transportar. Además, a cada carretera le corresponde una capacidad de flujo máximo, digamos en toneladas por día. Interesa calcular la máxima cantidad de bienes que se puede enviar, por día, de un sitio cualquiera de la red a otro.

3 El Resultado Principal

El trabajo de Karger explora el muestreo como una herramienta para resolver problemas en gráficas. Prueba que la muestra, o *esqueleto* que resulta de muestrear aleatoriamente aristas de la gráfica, manteniendo los mismos vértices, aproxima los valores de *todos* los cortes de la gráfica original. Como consecuencia importante se obtiene una disminución de la complejidad de los algoritmos de conexidad conocidos.

A continuación describimos más formalmente el resultado principal de muestreo para conexidad obtenido por Karger [5]. Sea $G = (V, E)$ una gráfica. Sea $\hat{G}$ la gráfica que se obtiene al asignar a cada arista e de E una probabilidad de elegirla p_e. $\hat{G}$ es una gráfica con *pesos* (o *capacidades*) y el valor de un corte en $\hat{G}$ —en contraste con el valor de un corte en G— es igual a la suma de los pesos de las aristas que ligan los dos conjuntos del corte. Sea $\hat{c}$ el valor del corte mínimo de $\hat{G}$. Si E' es el conjunto de aristas de G seleccionadas (con probabilidad p_e) en un experimento (aleatorio), a la gráfica *aleatoria* $G' = (V, E')$ se le llama *esqueleto de* G. El teorema principal dice que, siempre y cuando $\hat{c}$ sea suficientemente grande, cada corte en el esqueleto tendrá un valor aproximadamente igual a su *valor esperado*, es decir, aproximadamente igual al valor del mismo corte en $\hat{G}$.

Cuando las probabilidades de muestreo de las aristas se toman todas iguales a p, se usa $G(p)$ para designar al esqueleto y se tiene el siguiente corolario. Sea G una gráfica de n vértices y corte mínimo c y tomemos $p = 2(d + 2)(\ln n)/\epsilon^2 c$, entonces:

> *La probabilidad de que algún corte en $G(p)$ tenga un valor más de $(1 + \epsilon)$ o menos de $(1 - \epsilon)$ veces su valor esperado es $O(1/n^d)$.*

La página HTTP://THEORY.LCS.MIT.EDU/~KARGER/PAPERS.HTML de David Karger contiene los artículos [6, 3, 4, 5]. Para información mas general de algoritmos en gráficas se pueden consultar los textos [1, 2].

Referencias

[1] S. Even, *Graph Algorithms,* Computer Science Press, 1979.

[2] Cormen, C. Leiserson, R. Rivest, *Introduction to Algorithms,* MIT Press, 1990.

[3] D. Karger, "Using Random Sampling to Find Maximum Flows in Uncapacitated Undirected Graphs," Proc. de 29th ACM Symposium on the Theory of Computing (STOC), 1997, p. 240–249.

[4] D. Karger, "Experimental Study of Minimum Cut Algorithms," SODA 1996.

[5] D. Karger, "Random Sampling in Cut, Flow, and Network Design Problems conference version," Proc. de 26th ACM Symposium on the Theory of Computing (STOC), 1994, p. 648–657.

[6] D. Karger, "Random Sampling in Graph Optimization Problems," tesis de doctorado, Universidad de Stanford, octubre 1994.