How to sample connected $K$-partitions of a graph

Marina Meila
mmp@stat.washington.edu

August 2, 2018

Abstract

A connected undirected graph $G = (V,E)$ is given. This paper presents an algorithm that samples (non-uniformly) a $K$ partition $U_1, \ldots, U_K$ of the graph nodes $V$, such that the subgraph induced by each $U_k$, with $k = 1 : K$, is connected. Moreover, the probability induced by the algorithm over the set $C_K$ of all such partitions is obtained in closed form.

1 Problem and notation

A connected undirected graph $G = (V,E)$ with $|V| = n$ is given. A connected $K$-partition of $G$ denotes a partition of $V$ into $K$ clusters $U_1, \ldots, U_K$, such that the subgraph of $G$ induced by each $U_k$, with $k = 1 : K$, is connected. Here $K$ is considered fixed and may be omitted for brevity.

A connected partition is denoted by $C$, and the set of all connected $K$ partitions of $G$ is denoted by $C_K$. Counting $|C_K|$ is known to be hard in general [1].

Denote by $T$ a spanning tree of $G$, and by $T$ the set of all spanning trees of $G$. The spanning trees of a simple undirected graph can be counted by Tutte’s Matrix Tree Theorem [2]. This theorem extends to multigraphs with no self loops. Let $t(G) = |T|$, and $t(S)$ the number of spanning trees in the subgraph of $G$ induced by $S \subset V$. The Matrix Tree Theorem states that $t(G) = \det(L(G)^*)$ where $L(G) = D(G) - A(G)$ the diagonal degree matrix minus the adjacency matrix of $G$ (i.e. the unnormalized Laplacian of graph $G$), and $L^*$ is a minor of matrix $L$, i.e $L$ with the $i$-th row and column removed, for some arbitrary $i$. Note that $t(G)$ is 0 if $G$ is not connected and that $\det L = 0$ always, as the rows of $L$ sum to 0.

2 An algorithm for sampling from $C_K$

The following algorithm samples connected $K$-partitions, non-uniformly.

Algorithm SampleConnectedPartition($K, G$)

1. Sample a spanning tree $T \in T$ uniformly at random.

2. Remove $K - 1$ edges from $T$ uniformly at random without replacement.

Return the connected components $U_{1:K}$ of $T$ obtained in Step 2.

Proof sketch: it is obvious that each $U_k$ is connected. Step 1 can be performed for example by assigning the edges random weights and computing the minimum spanning tree with these weights.

We say that a spanning tree $T \in T$ is compatible with a partition $C \in C_K$ iff $C$ can be obtained from $T$ by removing $K - 1$ edges.
3 Analysis. Probability induced by SampleConnectedPartition on $C_K$

The question now is: what is the probability of obtaining a given partition $U_{1:K}$ by the SampleConnectedPartition algorithm?

We first explain the idea for $K = 2$; in this case we remove a single edge from $T$. Let $S \subset V$ ($S$ represents $U_1$ or $U_2$). Denote by $\partial S$ the edges between $S$ and $V \setminus S$. Any spanning tree $T$ must intersect $\partial S$ (otherwise $T$ would not be connected). If $|T \cap \partial S| > 1$, no edge removal will produce the partition $C = (S, V \setminus S)$. But if $|T \cap \partial S| = 1$, then w.p. $1/(n-1)$ the partition is obtained, namely when the single edge in $T \cap \partial S$ is deleted from $T$.

For a fixed $S$, let the event $\mathcal{T}_S = \{|T \cap \partial S| = 1\} \subset \mathcal{T}$. Note that fixing $S$ in this case amounts to fixing the partition $C$.

Any $T$ in $\mathcal{T}_S$ contains a spanning tree of $S$, a spanning tree of $V \setminus S$, and one edge from $\partial S$. Hence,

$$|\mathcal{T}_S| = t(S)t(V \setminus S)|\partial S|$$

and

$$P(C) = \frac{P(\mathcal{T}_S)}{n-1} = \frac{t(S)t(V \setminus S)|\partial S|}{(n-1)t(G)}$$

Now, let’s consider the general case of a $K$ partition $C = (U_1, \ldots, U_K)$. Each $T$ that is compatible with $C$ must contain a spanning tree $T_k$ of the subgraph induced by $U_k$, for each $k = 1 : K$. Furthermore, these trees must be connected by edges between two clusters $U_k, U_{k'}$, ensuring that no loops are formed. In other words, to complete $\cup_{1:k} T_k$ to a spanning tree $T$ of $G$ that is compatible with $C$, we contract each $U_k$ to a single node; all the edges between $U_k$ and $U_{k'}$ are now between the two nodes representing $U_k$ and $U_{k'}$. Hence, we obtain a multigraph $M(G, C)$ with $K$ nodes. Any spanning tree of $M(G, C)$ completes $\cup_{1:k} T_k$ to a spanning tree of $G$.

The number of spanning trees in the multigraph $M(G, C)$ is obtained again by the Matrix Tree Theorem, where each edge has a weight equal to its multiplicity.

Once we have a $T$ compatible with $C$, we need to remove the set of $K-1$ edges connecting the clusters $U_{1:K}$, out of $\binom{n-1}{K-1}$ possible edge removals. Hence,

$$P(U_{1:K}) = \frac{t(M(G, U_{1:K}))\prod_{k=1}^{K} t(U_k)}{\binom{n-1}{K-1}t(G)}$$

This analysis also shows that SampleConnectedPartition samples every connected partition of $G$ with non-zero probability.

4 An example

Let the graph $G$ with $n = 10$ be defined by the following adjacency matrix $A$.

|     | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|-----|----|----|----|----|----|----|----|----|----|----|
| 1   | 0  | 1  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| 2   | 1  | 0  | 1  | 1  | 1  | 0  | 0  | 0  | 0  | 0  |
| 3   | 1  | 1  | 0  | 1  | 0  | 0  | 0  | 0  | 1  | 0  |
| 4   | 1  | 1  | 1  | 0  | 0  | 1  | 0  | 0  | 0  | 0  |
| 5   | 0  | 1  | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  |
| 6   | 0  | 0  | 0  | 1  | 1  | 0  | 1  | 0  | 0  | 1  |
| 7   | 0  | 0  | 0  | 0  | 1  | 1  | 0  | 1  | 0  | 0  |
| 8   | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 1  | 1  |
| 9   | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 1  | 0  | 1  |
| 10  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 1  | 0  |
The node degrees are

\[
\begin{array}{cccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
3 & 4 & 4 & 4 & 3 & 4 & 3 & 3 & 3 & 3 \\
\end{array}
\]

and the Laplacian matrix is

\[
\begin{array}{cccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
1 & 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\
2 & -1 & 4 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\
3 & -1 & -1 & 4 & -1 & 0 & 0 & 0 & -1 & 0 \\
4 & -1 & -1 & -1 & 4 & 0 & -1 & 0 & 0 & 0 \\
5 & 0 & -1 & 0 & 0 & 3 & -1 & -1 & 0 & 0 \\
6 & 0 & 0 & 0 & -1 & -1 & 4 & -1 & 0 & 0 \\
7 & 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 & 0 \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 3 & -1 \\
9 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 3 \\
10 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & -1 \\
\end{array}
\]

Let the $K = 3$ clusters be $U_1 = \{1, 2, 3, 4\} , U_2 = \{5, 6, 7\}, U_3 = \{8, 9, 10\}$. Then,

\[ t(A) = \det(L_{1:9,1:9}) = 4,546 \quad t(U_1) = 16, \quad t(U_2) = t(U_3) = 3 \quad (4) \]

and

\[ M(G) = \begin{bmatrix}
0 & 2 & 1 \\
2 & 0 & 2 \\
1 & 2 & 0
\end{bmatrix} \quad t(M(G)) = 8 \quad \binom{n-1}{K-1} = 36 \quad (5) \]

Hence, the probability of the partition $(U_1, U_2, U_3)$ is equal to $\frac{16 \times 3 \times 3 \times 8}{36 \times 4546} = 0.0070$.

**Acknowledgement**

This problem was suggested by a question from Steve K.

**References**

[1] A. Vince Counting connected sets and connected partitions of a graph Australasian Journal of Combinatorics, 67, 2017.

[2] D. M. West An introduction to graph theory Prentice Hall, 2001.