Hierarchical versus Flat Communities
Daichi Kuroda, Maximilien Dreveton, Matthias Grossglauser, Patrick Thiran
INDY (Information and Network Dynamics), EPFL

Background & Motivation

Community Detection
- Networks are commonly used to represent datasets of pairwise interactions (e.g., human interactions [1], spatial networks, transportation networks).
- Most datasets have community structure → Community Detection.

Hierarchical Community Detection
Community structure is often hierarchical.

Proposed Definition of the Hierarchical Tree

Definition [Hierarchical Tree] A rooted tree \( T \) is a hierarchical tree for the primitive communities \( b = \{b_1, b_2, \ldots, b_k\} \) w.r.t. a similarity function \( s \) if any vertex \( t \) on \( T \) satisfies:

\[
\forall i, j \in [k] s.t. d_{\text{HC}}(b_i, b_j) > d_{\text{HC}}(b_i, b_0) \Rightarrow s(b_i, t) > s(b_0, t),
\]

where \( d_{\text{HC}}(b_i, b_j) \) is the tree distance from the root to the least common ancestor of \( b_i, b_j \),
- If primitive communities \( b_i, b_j \) belong to a super community, \( b_i, b_j \) is more similar w.r.t. \( s \) to \( b_0 \) than to any primitive community \( b_k \) which does not belong to the super-community.
- Maximum-vertices hierarchical tree is the most informative.
- A star graph, i.e., a tree where all leaves are directly connected to the root, always satisfies the condition \( \rightarrow \) this can be regarded as a flat community.

Theorem (Uniqueness) The maximum-vertices hierarchical tree for a graph \( G \) with primitive communities \( b \) w.r.t. a similarity function \( s \) is unique.

Numerical Results

Bottom-up [2]:
- first identity the primitive communities \( b = \{b_1, b_2, \ldots, b_k\} \) and then repeatedly merge the communities using a linkage method.
- Recall of the super communities is dependent on the accuracy of recovering \( b \).

Algorithm 1

Algorithm 1 recovers the maximum-vertices hierarchical tree.

Finding Hierarchy in Practice

Definition (Approximately Hierarchical Tree) A rooted tree \( T \) is approximately hierarchical tree for the primitive communities \( b = \{b_1, \ldots, b_k\} \) w.r.t. a similarity function \( s \) if any vertex \( t \) on \( T \) satisfies:

\[
\frac{\sum_{b_i, b_j \in E_T} s(b_i, t) \cdot s(b_j, t)}{\sum_{b_i, b_j \in E_T} s(b_i, b_j)} \geq 1 - \delta. \tag{2}
\]

Motivation
- In reality, you normally only have access to noisy observation \( t \).
- Want to avoid finding spurious levels \( \rightarrow \) introduce \( \epsilon \).
- Want to have some buffer to be resistant to some noise or outlier \( \rightarrow \) introduce \( \delta \).

Finding the approximately hierarchical trees is done by using Algorithm 1, but instead of using condition \( \epsilon \), use condition \( \delta \).

Contributions
1. Introduced a natural and concrete definition of the hierarchical trees,
   - uniqueness of maximum-vertices hierarchical tree:
2. Proposed an algorithm to discover the maximum-vertices hierarchical trees and established a guarantee of it.

Motivation
- Want to avoid finding spurious levels
- Want to have some buffer to be resistant to some noise or outlier

Reference