A Topology-Based Clustering Algorithm for Analysis Very Large Biological Networks

Xiaowei Xu
Existing Algorithms

Cut-based algorithms
- Min-max cut, Ding et al., 2001
- Normalized cut, Shi & Malik, 2000

Modularity-based algorithms
- Newman & Girvan, 2004
- Clauset et al., 2004

Find clusters, but not hubs and outliers

Maximize intra-cluster connection and minimize inter-cluster connections
Topology-based Clustering

- **Vertex structure:** $\Gamma(v) = \{w \in V \mid (v, w) \in E\} \cup \{v\}$
- **Structural similarity:** $\sigma(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{|\Gamma(v)| |\Gamma(w)|}}$
- **$\varepsilon$-Neighborhood:** $N_\varepsilon(v) = \{w \in \Gamma(v) \mid \sigma(v, w) \geq \varepsilon\}$
- **Core:** $CORE_{\varepsilon, \mu}(v) \equiv |N_\varepsilon(v)| \geq \mu$
- **Direct structure reachable:** $DirRECH_{\varepsilon, \mu}(v, w) \equiv CORE_{\varepsilon, \mu}(v) \land w \in N_\varepsilon(v)$
- **Structure reachable:** transitive closure of direct structure reachability
- **Structure connected:** $CONNECT_{\varepsilon, \mu}(v, w) \equiv \exists u \in V : RECH_{\varepsilon, \mu}(u, v) \land RECH_{\varepsilon, \mu}(u, w)$. 
Topology-based clustering

- **Topology-based cluster C**
  - Connectivity: \( \forall v, w \in C : \text{CONNECT}_{\varepsilon, \mu} (v, w) \)
  - Maximality: \( \forall v, w \in V : v \in C \land \text{REACH}_{\varepsilon, \mu} (v, w) \Rightarrow w \in C \)

- **Hubs:**
  - Not belong to any cluster
  - Bridge to at least two clusters

- **Outliers:**
  - Not belong to any cluster
  - Connect to at most one cluster
MPHGO: Mining Partitions and Hubs in Graphs with Outliers

- Complexity: $O(n)$

ALGORITHM MPHGO($G=\langle V, E \rangle, \varepsilon, \mu$)
// all vertices in $V$ are labeled as unclassified;
for each unclassified vertex $v \in V$ do
// STEP 1. check whether $v$ is a core;
if $\text{CORE}_{\varepsilon, \mu}(v)$ then
// STEP 2.1. if $v$ is a core, a new partition is expanded;
genenerate new partition ID;
insert all $x \in N_v(v)$ into queue $Q$;
while $Q \neq 0$ do
  $y =$ first vertex in $Q$;
  $R = \{x \in V \mid \text{DirRECH}_{v,x}(y, x)\}$;
  for each $x \in R$ do
    if $x$ is unclassified or non-member then
      assign current partition ID to $x$;
      if $x$ is unclassified then
        insert $x$ into queue $Q$;
        remove $y$ from $Q$;
    else
      // STEP 2.2. if $v$ is not a core, it is labeled as non-member
      label $v$ as non-member;
  end for.
// STEP 3. further classifies non-members
for each non-member vertex $v$ do
  if $\exists x, y \in \Gamma(v) \land x.\text{partitionID} \neq y.\text{partitionID}$ then
    label $v$ as hub
  else
    label $v$ as outlier;
end for.
end MPHGO.

Xiaowei Xu (Ph.D.), UALR
Running Time Comparison

![Running Time Comparison Diagram]

- MPHGO
- FastModularity

Xiaowei Xu (Ph.D.), UALR
Find Conferences of College Football Teams
Clustering Political Books
Conclusion

- We propose a novel topology based network clustering algorithm:
  - It is fast: $O(n)$
  - It can find clusters, as well as hubs and outliers

- We are looking forward to the applications to large biological networks