Local Search of Communities in Large Graphs

Wanyun Cui, Fudan University
Yanghua Xiao, Fudan University
Haixun Wang, Microsoft Research Asia
Wei Wang, Fudan University
Outline

- Motivations
- Model
- Algorithms
- Experiments
- Contributions
Outline

- Motivations
- Model
- Algorithms
- Experiments
- Contributions
Complex network

- Complex network is everywhere.
Complex network

- Complex network is everywhere.

**Internet**
Complex network

- Complex network is everywhere.

Protein Network
Complex network

- Complex network is everywhere.
Community structures

- Complex network is everywhere.
- Most real life networks have community structures.
  - The graph can be divided into different groups
    - the vertices within each group are closely connected
    - the vertices between different groups are sparsely connected

Social Network  Internet  Protein Network
Revealing community structures

- Two possible ways
  - community detection
  - community search
Community detection vs. community search

- Community detection: divides the entire network to find communities

A coauthorship network. Each color represents a community.
Community detection vs. community search

- Community detection: divides the entire network to find communities
  - Too costly
  - Unfriendly to dynamic changing graphs
Community detection vs. community search

- Community detection: find the community of a specific vertex.
Community search: applications

- Friend recommendation on Facebook
- Advertising on social networks
- Infectious disease control
- Semantic expansion
Outline

- Motivations
- Model
- Algorithms
- Experiments
- Contributions
Community search

- **CSM**: community search with *maximality* constraint
  - Finding the community with highest goodness
- **CST**: community search with *threshold* constraint
  - Finding one community with $goodness \geq threshold$
CST: why not the best?

- Community search with threshold constraint: flexible in controlling the community’s structure.

- Application: infectious disease control
  - a less contagious disease: a small and closely related community that a few should be found.
  - a highly contagious disease: a loose community structure that more people can be found
Model: minimum degree

- **Goal:**
  - to measure the “closeness” of the community.

- **Definition:**
  - The goodness of a vertex set (community) $H$, is the minimum degree of $H$’s induced subgraph.
Model: minimum degree (example)

- Goodness({a}) = 0
Model: minimum degree (example)

- Goodness(\{a, e\}) = 1
Model: minimum degree (example)

- Goodness(\{a,b,c,d,e\}) = 3
Model: minimum degree (example)

- Goodness(\{a,b,c,d,e,f\})=1
- The minimum degree stop us from picking f.

- Challenge:
  - Non-monotonicity
Model: problem definition

- CSM: community search with **maximality** constraint
  - Finding the community that its minimum degree is maximized

- CST(k): community search with **threshold** constraint
  - Finding one community with minimum degree $\geq k$
Model: Relationship between CST and CSM

- The best community of v has minimum degree m
- CSM is equal to the CST(m)
Model: Relationship between CST and CSM

- The best community of v has minimum degree m
- CSM is equal to the CST(m)
Model: Relationship between CST and CSM

- From CST to CSM: we can enumerate k starting from 0, to find the last non-empty solution CST(k) for CSM.
Example: Relationship between CST and CSM

- Query node: a
- CST(0): \{a\}
- CST(1): \{a,b\}
- CST(2): \{a,b,c,d\}
- CST(3): \{a,b,c,d,e\}
- CST(4): no solution
Example: Relationship between CST and CSM

- **Query node:** a
- **CST(0):** \{a\}
- **CST(1):** \{a, b\}
- **CST(2):** \{a, b, c, d\}
- **CST(3):** \{a, b, c, d, e\}
- **CST(4):** no solution
Model: Relationship between CST and CSM

- A CSM framework: use a binary search procedure for CST(k) to find CSM
  - Run CST for \( \log_2 \deg(v) \) times
  - \( \deg(v) \) is the degree of query node \( v \)
Outline

- Motivations
- Model
- **Algorithms**
- Experiments
- Contributions
Algorithm to CST

- Framework
  - Complexity analysis
- Optimization
  - Candidate selection
CST: framework

- Framework: Iteratively adding new nodes from the local graph.
CST: framework

- Framework: Iteratively adding new nodes from the local graph.
- $K=3$, query vertex = a
CST: framework

- Framework: Iteratively adding new nodes from the local graph.
- \( K=3 \), query vertex = a
CST: framework

- Framework: Iteratively adding new nodes from the local graph.
- K=3, query vertex=a
CST: framework

- Framework: Iteratively adding new nodes from the local graph.
- K=3, query vertex=a
CST: framework

- Framework: Iteratively adding new nodes from the local graph.
- K=3, query vertex=a
CST: framework

- Framework: Iteratively adding new nodes from the local graph.
  - $K=3$, query vertex = a
- Until these nodes form a valid community.
CST: candidate selection

- How to pick a new vertex from the current local graph?
**CST: candidate selection**

- How to pick a new vertex from the current local graph?
CST: candidate selection

- Ig: largest increment of goodness
- li: largest number of incidence
CST: candidate selection

- li
  - f: 0
  - d: 1

- lg
  - f: 0
  - d: 1
Algorithm to CSM: framework

- CSM can be solved using CST(k), by the binary search on k
- CSM1: trade quality for performance by tuning parameter
- CSM2: optimal solution.
Outline

- Motivations
- Model
- Algorithms
- **Experiments**
- Contributions
Experiments

- Setup
- Dataset

- AMD Athlon™X2 2GHz
- 2GB memory
Experiments

- Setup

- Dataset

| Dataset    | |V|   | |E|    |
|------------|-------------------|
| DBLP       | 481K              | 1.72M |
| Berkeley   | 654K              | 6.58M |
| Youtube    | 1.1M              | 3.0M  |
| Livejournal| 4.0M              | 34.7M |
Experiments: case studies

- Reasonable community:
  - Senses are highly related to “pot”
  - Hypernym: vessel, container
  - Adjective: containerful
  - Related entities: dish, bowl
Experiments: performance for CST

- local search performs better than global search in most cases
Experiments: performance for CST

- Similar results can be found in other networks.
Experiments: performance for CST

- $l_i$ is better than $l_g$
Experiments: performance of CSM

- CSM2 performs the best
Experiments on Synthetic Networks

- Scalability
- Sensitive to community structure
Synthetic Networks: scalability

(a) CST

(b) CSM
Synthetic Networks: Sensitive to community structure

- Local search is sensitive to the clearness of community structures in the network.

![Graphs showing synthetic networks sensitivity](image)
Contributions

- Problem definition
- Model
- Algorithms
- Extensive experiments
Thank you!