Community detection in networks

Santo Fortunato

Aalto University
Outline

1) Introduction
2) Global optimization techniques: limits
3) Local techniques: OSLOM
4) Testing algorithms
5) Summary
Networks

Protein-protein interaction networks

Network: simplest representation of a complex system
Networks

Social networks
Networks

Internet
Important features of a system and its dynamics from purely structural information
Community structure

Communities: sets of tightly connected nodes

- People with common interests
- Scholars working on the same field
- Proteins with equal/similar functions
- Papers on the same/related topics
- ...
Community detection

Theoretical reasons

- Organization
- Node features
- Node classification
- Missing links
Community detection

Graph visualization
Community detection
Practical reasons: recommendation systems
Community detection

Practical reasons: recommendation systems
Community detection

Practical reasons: recommendation systems
Community detection

Practical reasons: unknown protein functions
Community detection

Practical reasons: unknown protein functions
Community detection

Practical reasons: unknown protein functions
Difficult problem!
Difficult problem

Ill-defined problem:
• What is a community/partition?
• What is a good community/partition?

Complications:
• Link directions
• Link weights
• Overlapping communities
• Hierarchical structure
Global optimization

Principle:
• Function $Q(P)$ that assigns a score to each partition
• Best partition of the network $\rightarrow$ partition corresponding to the maximum/minimum of $Q(P)$

Problems:
• Good partition does not imply good clusters
• Answer depends on the whole graph $\rightarrow$ it changes if one considers portions of it or if it is incomplete
Global optimization

Principle:
• Function $Q(P)$ that assigns a score to each partition
• Best partition of the network $\rightarrow$ partition corresponding to the maximum/minimum of $Q(P)$

Problems:
• Good partition does not imply good clusters
• Answer depends on the whole graph $\rightarrow$ it changes if one considers portions of it or if it is incomplete
Modularity optimization

\[ Q = \frac{1}{m} \sum_{c=1}^{n_c} \left( l_c - \frac{d_c^2}{4m} \right) \]


**Goal:** find the maximum of Q over all possible network partitions

**Problem:** NP-complete (Brandes et al., 2007)!
Resolution limit

\[ Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[ l_c - \frac{1}{4} \left( \frac{d_c}{\sqrt{m}} \right)^2 \right] \]

modularity’s scale

Consequences

\[ d_c < \sqrt{m} \]

Resolution limit

Subgraph 1, degree $k_1$

Subgraph 2, degree $k_2$

Expected number of edges between the two subgraphs in modularity’s null model:

$$m \left( 2 \cdot \frac{k_1}{2m} \cdot \frac{k_2}{2m} \right) = \frac{k_1 k_2}{2m}$$

if $k_1 = k_2 = d_c$ → $\frac{d_c^2}{2m}$
Resolution limit

Question: What is the origin of the resolution limit?

Answer: global null model is unrealistic!

Question: What is the origin of the resolution limit?

Answer: global null model is unrealistic!

Multi-resolution methods?

\[ Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[ l_c - \gamma \frac{d_c^2}{4m} \right] \]

Double trouble:
1) Small clusters are merged
2) Large clusters are split

Hard to find values of resolution parameter that eliminate both problems!

Local optimization

Principle:
• Communities are local structures
• Local exploration of the network, involving the subgraph and its neighborhood

Advantages:
• Conceptual advantage: communities are “local”
• Absence of global scales -> no resolution limit
• One can analyze only parts of the network
Local optimization

Implementation:
• Function $Q(C)$ that assigns a score to each subgraph
• Best cluster $\rightarrow$ cluster corresponding to the maximum/minimum of $Q(C)$ over the set of subgraphs including a seed node

Example: Local Fitness Method (LFM)

Local optimization: OSLOM

Basics:
- LFM with fitness expressing the statistical significance of a cluster with respect to random fluctuations
- Statistical significance evaluated with Order Statistics

First multifunctional method:
- Link direction
- Link weight
- Overlapping clusters
- Hierarchy

Local optimization: OSLOM

Welcome to OSLOM's Web page

OSLOM stands for Order Statistics Local Optimization Method and is a clustering algorithm designed for networks.

Download the code (beta version 2.4, last update: September, 2011)

The package contains the source code and the instructions to compile and run the program. You will also get a simple script which we implemented to visualize the clusters found by OSLOM. This script writes a pajek file which in turn can be processed by pajek or gephi.

This is an example of how the visualization looks like.

http://www.oslom.org/
Testing clustering algorithms

**Question:** how to test clustering algorithms?

**Answer:** checking whether they are able to recover the known community structure of benchmark graphs

**Planted l-partition model** (Condon & Karp, 1999)

**Ingredients:**
1) \( p = \) probability that vertices of the same cluster are joined
2) \( q = \) probability that vertices of different clusters are joined

**Principle:** if \( p > q \) the groups are communities
The LFR benchmark

Realistic feature: power law distributions of degree and community size

# Testing clustering algorithms

## A comparative analysis

<table>
<thead>
<tr>
<th>Author</th>
<th>Label</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girvan &amp; Newman</td>
<td>GN</td>
<td>$O(nm^2)$</td>
</tr>
<tr>
<td>Clauset et al.</td>
<td>Clauset et al.</td>
<td>$O(n \log^2 n)$</td>
</tr>
<tr>
<td>Blondel et al.</td>
<td>Blondel et al.</td>
<td>$O(m)$</td>
</tr>
<tr>
<td>Guimerà et al.</td>
<td>Sim. Ann.</td>
<td>parameter dependent</td>
</tr>
<tr>
<td>Radicchi et al.</td>
<td>Radicchi et al.</td>
<td>$O(m^4/n^2)$</td>
</tr>
<tr>
<td>Palla et al.</td>
<td>Cfinder</td>
<td>$O(\exp(n))$</td>
</tr>
<tr>
<td>Van Dongen</td>
<td>MCL</td>
<td>$O(nk^2)$, $k &lt; n$ parameter</td>
</tr>
<tr>
<td>Rosvall &amp; Bergstrom</td>
<td>Infomod</td>
<td>parameter dependent</td>
</tr>
<tr>
<td>Rosvall &amp; Bergstrom</td>
<td>Infomap</td>
<td>$O(m)$</td>
</tr>
<tr>
<td>Donetti &amp; Muñoz</td>
<td>DM</td>
<td>$O(n^3)$</td>
</tr>
<tr>
<td>Newman &amp; Leicht</td>
<td>EM</td>
<td>parameter dependent</td>
</tr>
<tr>
<td>Ronhovde &amp; Nussinov</td>
<td>RN</td>
<td>$O(n^\beta)$, $\beta \sim 1$</td>
</tr>
</tbody>
</table>

Testing clustering algorithms

A comparative analysis

LFR benchmark graphs
Testing clustering algorithms

A comparative analysis

Random graphs: no clusters!
Testing clustering algorithms

Limits of artificial benchmarks:
• Relationships with real community structure unclear
• Risk of creating algorithms performing well on the benchmarks and not so well on real networks

Solution: real networks with ground truth classification?


Warning: classification must be reliable!
Dynamic clustering

Typical approach:
• Find the clusters for each snapshot with a static method
• Associate clusters of different snapshots

Limit: partitions independent of the history of the system

Alternative: exploiting the structural information of different snapshots -> cluster stability
Dynamic clustering

Consensus clustering

**Goal:** finding median partition of network

**Steps:**
1) Compute partition of each snapshot with static algorithm
2) Compute consensus partition for sequences of $k$ consecutive snapshots

Resulting partitions more accurate and stable!

NetCom Analyzer

NetCom Analyzer
COMMUNITY detection in complex NETWORKs

Test And Share Your Algorithm
Suggest Relevant Publications
Find The Clusters In Your Data

NetCom Analyzer is the first portal entirely dedicated to the analysis of community structure in networks. You can test your own algorithms, share them with the other users, and/or analyze your own datasets with the methods available in the library. You may also suggest relevant publications about community structure in networks and publish new networked datasets with built-in communities.

Algorithms
FRINGE
Camilo Palazuelos, Marta Zorrilla

Clique Percolation Method
G. Palla, I. Derenyi, I. Farkas and T. Vicsek

Louvain Method
Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre

Edge Clustering Algorithm
Filippo Radicchi

Publications
FRINGE: a new approach to the detection of overlapping communities in graphs
Camilo Palazuelos, Marta Zorrilla

The map equation
Martin Rosvall, Daniel Andrist, and Carl T. Bergstrom

Maps of random walks on complex networks reveal community structure
M. Rosvall and C. T. Bergstrom

Finding statistically significant communities in networks

Datasets
Zachary karate club
Vertices are members of a karate club in the United States, who were monitored during a period of three years. Edges connect members who had social interactions outside the club. W. W. Zachary, J. Antropol Res., 33, 452 (1977)

Dolphin social network
Vertices of the network are dolphins and two dolphins are connected if they were seen together more often than expected by chance. D. Lusseau, Proc. Royal Soc. London B, 270, 513-516 (2003)

http://www.netcom-analyzer.org/
Summary of the talk

1) Global optimization methods have important limits: local optimization looks more natural and promising

2) Validation:
   a) artificial benchmarks useful, not 100% reliable
   b) real networks with ground truth information
Summary of the field

1) What is a community? No unique answer! Definition is system- and problem-dependent

2) Magic method? No such thing! Domain dependent methods?

3) Low complexity techniques (down to linear!)

4) Versatile methods: directed networks, weighted networks, overlapping communities, hierarchy

5) Attention on validation

6) Constraints: a (new) method should
   a) not split cliques
   b) not merge cliques, if well-separated
   c) not find communities in random graphs

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly

Community detection in graphs
Volume 486, Issues 3-5, February 2010, Pages 75-174
Fortunato, S.

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. We will attempt a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks. © 2009 Elsevier B.V.
Acknowledgements

Alex Arenas
Marc Barthélémy
Alberto Fernández
Sergio Gómez

Janos Kertész
Mikko Kivelä
Andrea Lancichinetti
Vito Latora

Massimo Marchiori
Filippo Radicchi
José J. Ramasco
Jari Saramäki