Community detection

Maximilien Danisch

LIP6 – CNRS and Université Pierre et Marie Curie

first_name.last_name@lip6.fr

Outline

Community detection

Goal: Identify automatically relevant groups of nodes.

Applications:

- Understand the structure of a network
- Detect specific communities (web pages, proteins, ...)
- Help visualization
- Improvement information retrieval (search engines, recommendation, ...)

What is a community?

Set of nodes that share something:

- Affiliation (friends, colleagues, club, ...)
- Similar interests (tagging systems, ...)
- Similar contents (movies, books, products, web pages, ...)
- ...

What is the connexion with the network structure?

What is a community?

Set of nodes that share something:

- Affiliation (friends, colleagues, club, ...)
- Similar interests (tagging systems, ...)
- Similar contents (movies, books, products, web pages, ...)
- ...

What is the connexion with the network structure?

What is a community?

Set of nodes that share something:

- Affiliation (friends, colleagues, club, ...)
- Similar interests (tagging systems, ...)
- Similar contents (movies, books, products, web pages, ...)
- ...

What is the connexion with the network structure?

More densely connected inside than outside

Find a single community (intuition)

Structural approaches: cohesive subgraphs

Exercise: suggest a relevant definition of a community.
Example: a clique

Optimization approach: quality function

Quality function: quantitatively evaluate the quality of a set of nodes as a community.
Exercise: suggest a relevant quality function.

Outline

One graph partitioning method: Label propagation

Near linear time algorithm to detect community structures in large-scale networks - *Raghavan et al.*

- **Step 1:** give a unique label to each node in the network
- **Step 2:** Arrange the nodes in the network in a random order
- **Step 3:** for each node in the network in this random order: set its label to a label occurring with the highest frequency among its neighbours (if it is not already the case)
- **Step 4:** go to 2 as long as there exists a node with a label that does not have the highest frequency among its neighbours.

To shuffle in a clean way:
`https://en.wikipedia.org/wiki/Fisher-Yates_shuffle`

One graph partitioning method: Label propagation

Exercise: why such an algorithm should lead to relevant groups?

Exercise: Which data structure should be used to implement this algorithm efficiently?

One graph partitioning method: Label propagation

Exercise: why such an algorithm should lead to relevant groups?

- Densely connected groups should reach a common label.
- When such a consensus group is created it should expand until being stopped by other equivalent consensus groups.

Exercise: Which data structure should be used to implement this algorithm efficiently?

Outline

Validation: the case of a partition

Comparing the performance of several algorithms:

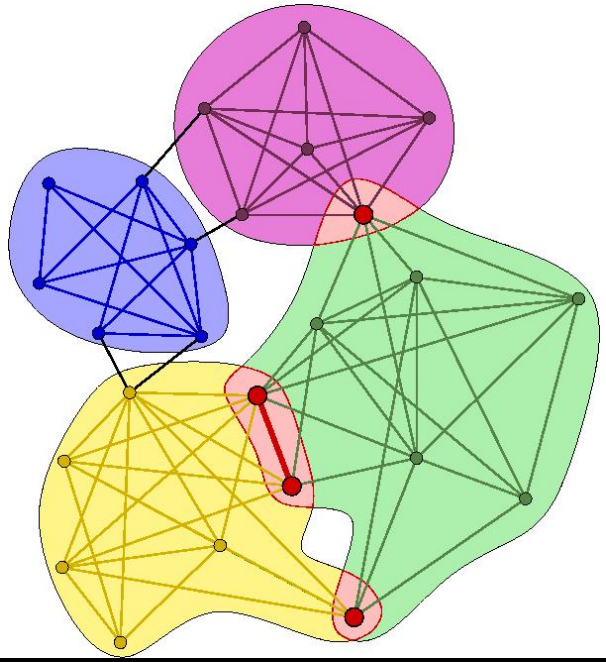
- Using synthetic graphs with a known community structure
  - **Exercise:** suggest such a graph
  - LFR bechmark: link
- Using a metric evaluating how similar the found community structure is to the ground-truth one
  - **Exercise:** Suggest such a metric
  - Adjusted Rand Index (ARI): Wikipedia paper
  - Normalized Mutual Information (NMI)
  - ...

Outline

One overlapping method: k-clique percolation

Definition: Two k-cliques are considered adjacent if they share $k - 1$ nodes.

Definition: A community is defined as the maximal union of k-cliques that can be reached from each other through a series of adjacent k-cliques.



Exercise: how can we find all "communities" efficiently for $k = 3$?