KU LEUVEN kulak

comex
combinatorial optimization:
metaheuristics & exact methods

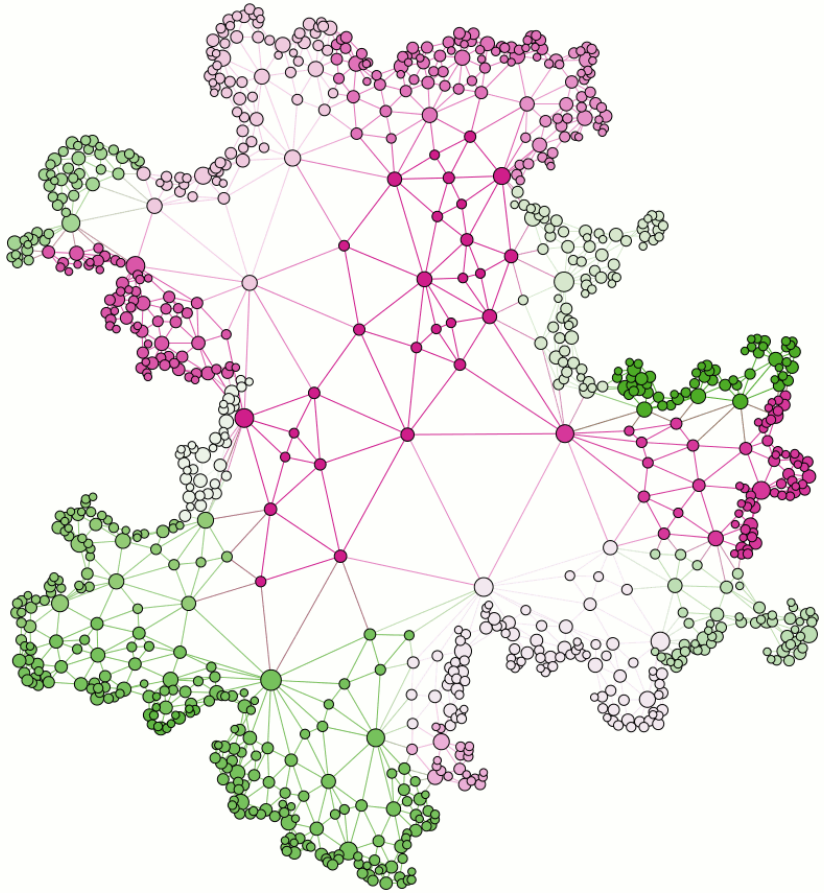
Optimization in Large Graphs: Toward a Better Future?

Pieter Leyman & Patrick De Causmaecker

pieter.leyman@kuleuven.be, patrick.decausmaecker@kuleuven.be

Introduction

- Optimization in large graphs
  - Find groups of connected individuals
  - Sociology, neurology, computer science, …
  - Large graphs: >10,000 nodes
  - Heuristic solution techniques



Source: http://ginestra-bianconi-6flt.squarespace.com

Introduction

- Shortcomings in literature
  - Vague problem definition
  - Contribution of algorithms unclear
  - Effects of data often not considered
- Barrier between different research fields
  - Operations research, computer science journals: cliques, clusters
  - Physics journals: communities (of social networks)

What is a community?

- Definition of good vs bad groups of nodes
- What characteristics are we looking for?
- Unambiguous objective function
- Theoretical framework on clusters in graphs
- Link with existing models?

What is a community?

- **Modularity: the bad**
  - Divide graph into subsets of nodes
  - High connectivity within subsets, low connectivity between subsets
  - Division sufficiently different from that in similar random graph

What is a community?

- **Modularity: the bad**
  - Divide graph into subsets of nodes
  - High connectivity within subsets, low connectivity between subsets
  - Division sufficiently different from that in similar random graph

    ➔No explicit definition of a community!

What is a community?

- **Modularity:** 3 shortcomings (Good et al., 2010)
  - Resolution limit
  - Degeneracy
  - Limiting behavior

Good B.H., de Montjoye Y.-A. & Clauset A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81: 046106.

What is a community?

- **Modularity:** 3 shortcomings (Good et al., 2010)
  - Resolution limit
    - Smaller communities may be hidden within larger ones
    - Especially problematic for hierarchical graphs
    - Comparison with similar random graph is problematic

Good B.H., de Montjoye Y.-A. & Clauset A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81: 046106.

What is a community?

- **Modularity:** 3 shortcomings (Good et al., 2010)
  - Degeneracy
    - Exponential number of high-quality solutions with an objective function value close to the optimal value
    - Partitions of solutions into subgraphs may differ greatly
    - Global optimum difficult to determine

Good B.H., de Montjoye Y.-A. & Clauset A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81: 046106.

What is a community?

- **Modularity:** 3 shortcomings (Good et al., 2010)
  - Limiting behavior
    - Maximum modularity function value depends on network size and number of communities
    - Higher value for larger and more modular networks
    - High modularity may indicate a large difference from a similar random graph rather than a sound partition

Good B.H., de Montjoye Y.-A. & Clauset A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81: 046106.

What is a community?

Modularity is commonly employed in spite of these pitfalls

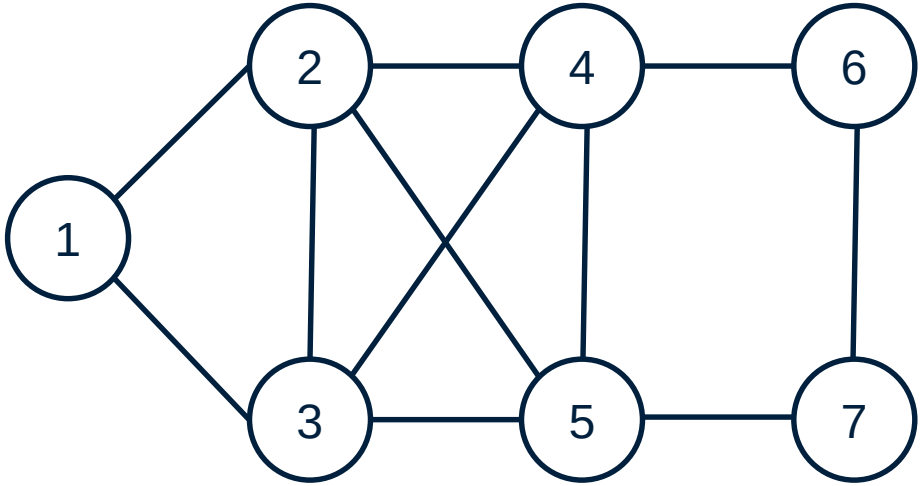
What is a community?

Modularity is commonly employed in spite of these pitfalls

What are we actually optimizing??

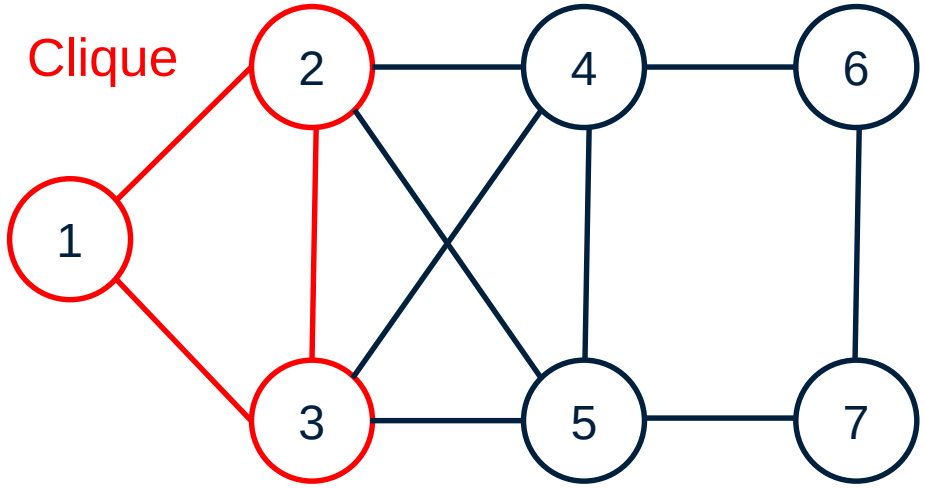
What is a community?

- **Maximal cliques: the useful**
  - Maximal clique problem



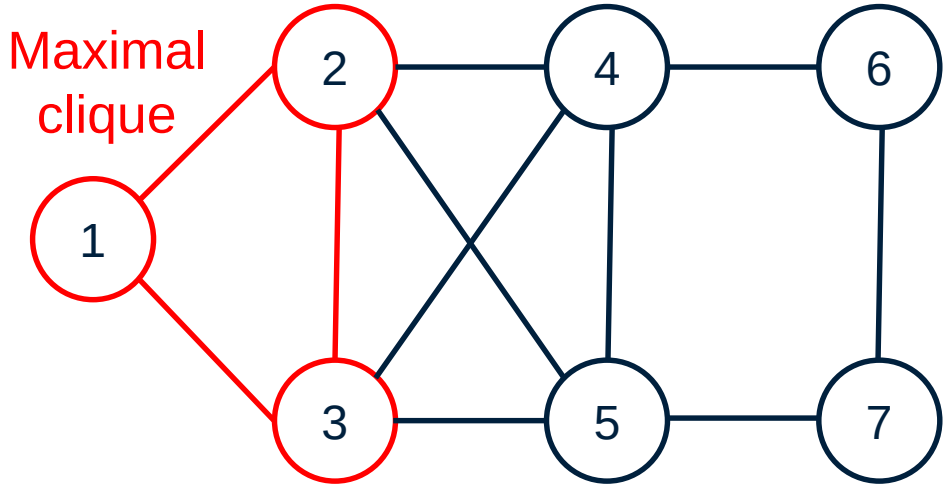
What is a community?

- **Maximal cliques: the useful**
  - Maximal clique problem



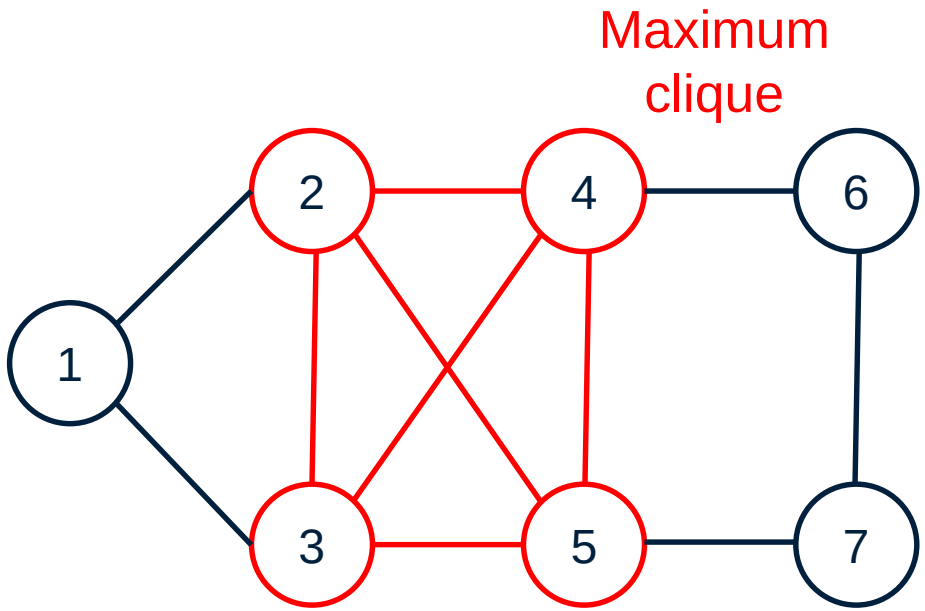
What is a community?

- **Maximal cliques: the useful**
  - Maximal clique problem



What is a community?

- **Maximal cliques: the useful**
  - Maximal clique problem



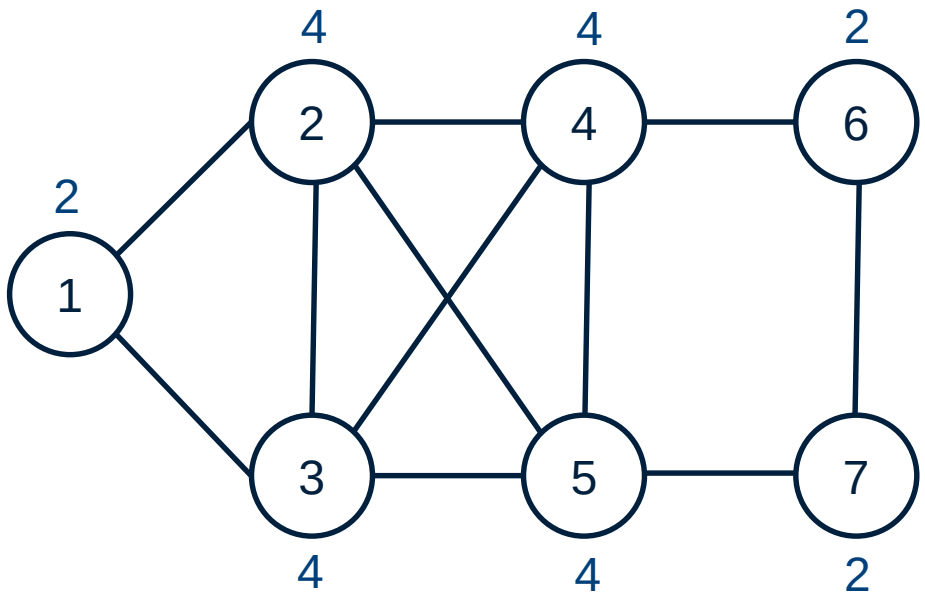
What is a community?

- **Clique relaxation: the better?**
  - Framework of Pattillo et al. (2013)
  - Focus on edge connectivity: high degree of connectedness within cliques, low degree between cliques
    - Degree of node
    - Density of subgraph/set

Pattillo J., Youssef N. & Butenko S. (2013). On clique relaxation models in network analysis. *European Journal of Operational Research*, 266: 9-18.

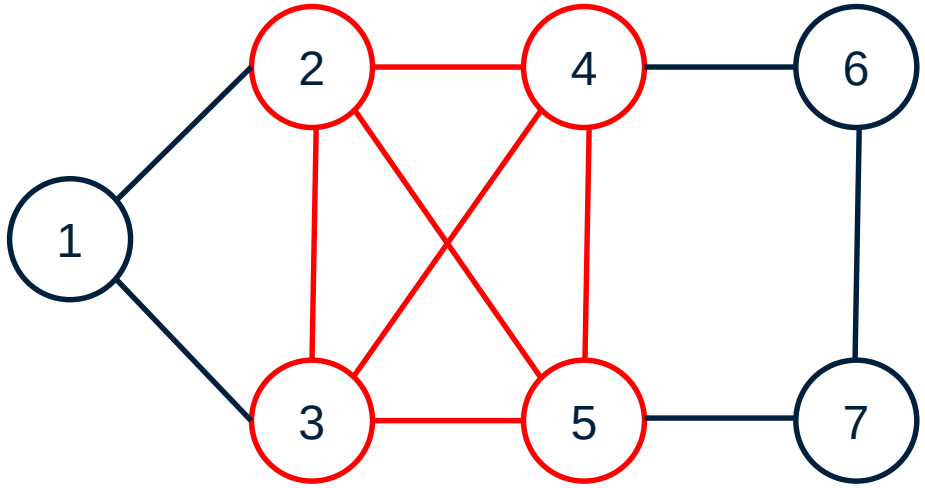
What is a community?

- **Clique relaxation: the better?**
  - Node degree: number of neighboring nodes



What is a community?

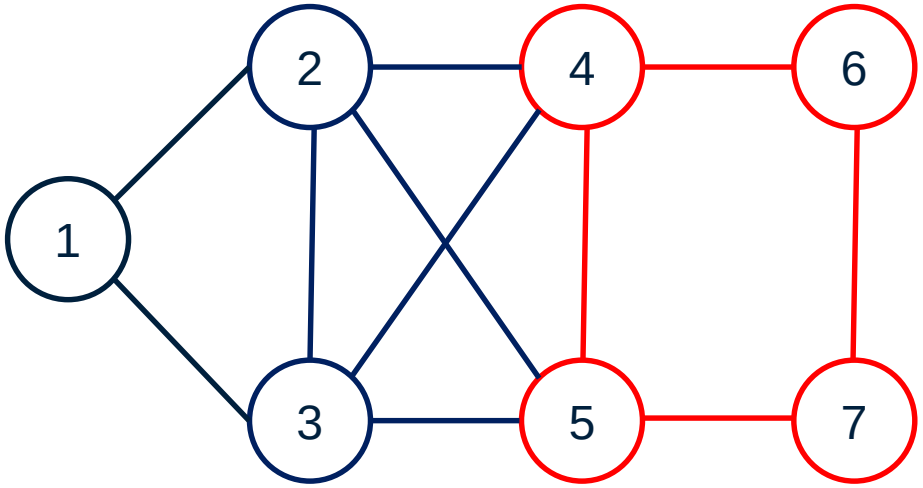
- **Clique relaxation: the better?**
  - Subgraph density: number of edges vs number of possible edges



6 possible edges vs 6 edges in subgraph ➔ Density=1

What is a community?

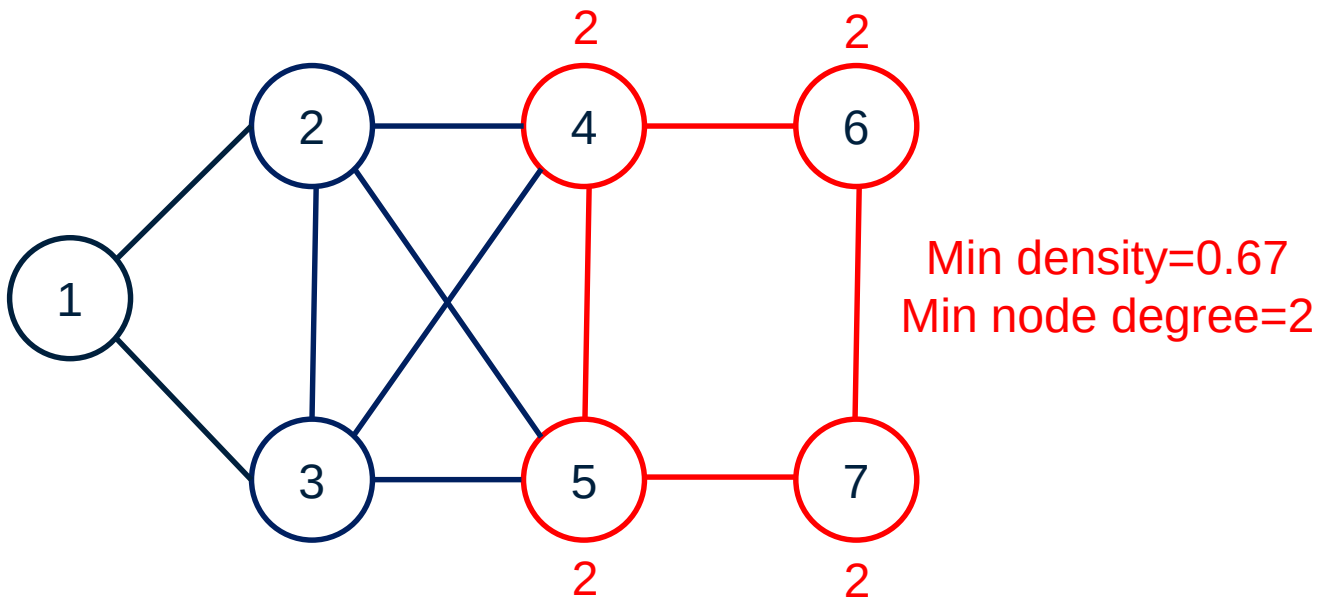
- **Clique relaxation: the better?**
  - Subgraph density: number of edges vs number of possible edges



6 possible edges vs 4 edges in subgraph ➔ Density=0.67

What is a community?

- **Clique relaxation: the better?**
  - Quasi-clique: relax density and node degree restrictions



What is a community?

- **Clique relaxation: the better?**
  - Decide on structure of subsets/communities based on parameters
    - In problem definition!!
    - Different research fields (e.g. bioinformatics, neurology)
  - Shortcomings modularity
    - Resolution limit: no bias towards larger subsets
    - Degeneracy: clear global optimum, but multiple solutions may lead to same number of subsets, albeit it with different partitions
    - Limiting behavior: no comparison with random graph

What is the impact of algorithms?

- Algorithm components
- Algorithm comparison

What is the impact of algorithms?

- **Algorithm components**
  - Unambiguous algorithm description
  - Contribution of constituting parts should be clear
    - Proper analysis of components
    - Statistical tests for validation of contribution
  - Suitability of commonly used (meta)heuristic frameworks?
    - Scale to large datasets (>10,000 nodes)?

What is the impact of algorithms?

- **Algorithm comparison**
  - Computer and code independent

    ➔Algorithm efficiency!
  - Vs existing approaches
    - No up-the-wall game!
    - Statistical tests
  - Focus on understanding, gaining insights

    ➔Hyperheuristics?

And what about data?

- Large variation in existing data, both fictitious and real-life
  - Effect of data parameters on algorithm performance?
  - Size of data?
  - Reproducibility of results?
- Often tested on only handful of networks
  - Community structure is imposed in generation or based on real-life
    - ➔ Optimum is explicitly set?!
  - Algorithms for problem **instances** rather than problems
    - ➔ Risk of overfitting

And what about data?

- Generate fictitious data with large variation in parameters
  - Derive important data parameters from current data
  - Complement current work, broader analyses
  - Test on more than a handful of networks (with known structure)
  - No knowledge of solution space
- Problem → Algorithm → Data
  - Decide on structure of subsets/communities based on parameters
    - In problem definition!!
    - Allow for application in several fields such as bioinformatics, neurology, …

Conclusions

- Problem
  - Formal framework of subsets
  - What is our optimization problem?
- Algorithm
  - Unambiguous description
  - Clear contribution
- Data
  - Test on sufficient number of networks
  - Data is unknown in advance

Optimization in Large Graphs:
Toward a Better Future?

KU LEUVEN kulak

Optimization in Large Graphs:
Toward a Better Future?

- More rigor!
- Advance state-of-the-art
- Problem → Algorithm → Data
  - Focus on all three!
  - Feedback data → algorithm: hyperheuristics, machine learning
  - Feedback data → problem: data science
  - Based on problem **instance**!

Optimization in Large Graphs:
Toward a Better Future?

- More rigor!
- Advance state-of-the-art
- Problem → Algorithm → Data
  - Focus on all three!
  - Feedback data → algorithm: hyperheuristics, machine learning
  - Feedback data → problem: data science
  - Based on problem **instance**!

KU LEUVEN kulak

comex
combinatorial optimization:
metaheuristics & exact methods

Optimization in Large Graphs: Toward a Better Future?

Pieter Leyman & Patrick De Causmaecker

pieter.leyman@kuleuven.be, patrick.decausmaecker@kuleuven.be