# Some results on more flexible versions of Graph Motif

Romeo Rizzi[1] Florian Sikora[2]

[1]Universitá di Udine – Italy
[2]Lehrstuhl Bioinformatik Jena – Germany

florian.sikora@uni-jena.de

CSR 2012

# Outline

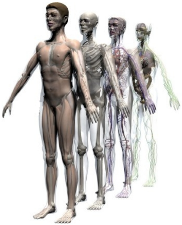
# Outline

Introduction

Graph Motif : Querying motifs without topology

Approximate motifs

# Motivations

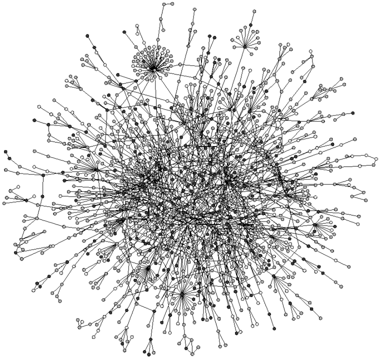


25000 30000 45000

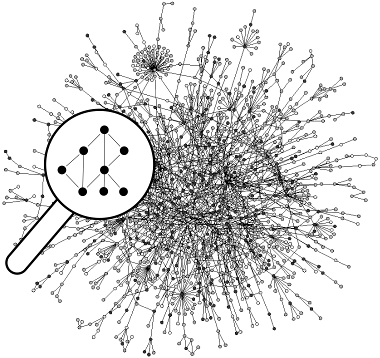
- Human complexity $\not\Leftrightarrow$ # of genes ?

# Proteins network

- Proteins can (physically) interact with other proteins (PPI).
- Biologically obtained... with noise !



# Proteins network



- Use a **graph** representation:
  - Proteins are nodes.
  - Interactions are edges.

# Issues

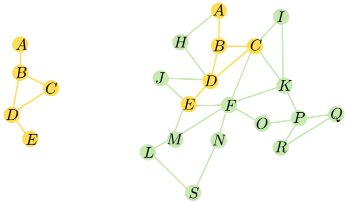
- New techniques: increase the available data [Sharan et Ideker 2006].
  - 2001: some hundreds.
  - 2006: thousands.
- Lot of databases:
  - BIND,
  - DIP,
  - KEGG,
  - MINT,
  - ...

# Issues

- New techniques: increase the available data [Sharan et Ideker 2006].
  - 2001: some hundreds.
  - 2006: thousands.
- Lot of databases:
  - BIND,
  - DIP,
  - KEGG,
  - MINT,
  - ...
- Computational solutions are needed.

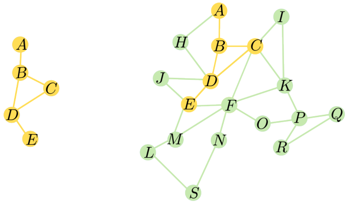
# Motif search

- Goal: find a subnetwork with both labels and topology of a given motif.



# Motif search

- Goal: find a subnetwork with both labels and topology of a given motif.




- Look for motifs to retrieve some known functions.
- **Deduce** information from well known species to less known species.

# Outline

Introduction

Graph Motif : Querying motifs without topology

Approximate motifs

# Graph Motif

- Large part of the literature deals with motif provided with a topology.

# Graph Motif

- Large part of the literature deals with motif provided with a topology.
- Fact : biological data are **very noisy** [Edwards et al. 2002]:
  - Missing informations (false negatives). About 50%.
  - Erroneous informations (false positives). About 50%.
- Topology of the motif can be unknown *a priori*.
- Different functions can have a same topology.

# Graph Motif

- Large part of the literature deals with motif provided with a topology.
- Fact : biological data are **very noisy** [Edwards et al. 2002]:
  - Missing informations (false negatives). About 50%.
  - Erroneous informations (false positives). About 50%.
- Topology of the motif can be unknown *a priori*.
- Different functions can have a same topology.
- Topology can be irrelevant.

# Graph Motif [Lacroix et al. 2006]

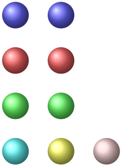
- Each network node is colored by its “function”.
- Motif is a (multi) set of colors
- Does the motif appears as a connected subgraph of the network ?

# Graph Motif [Lacroix et al. 2006]

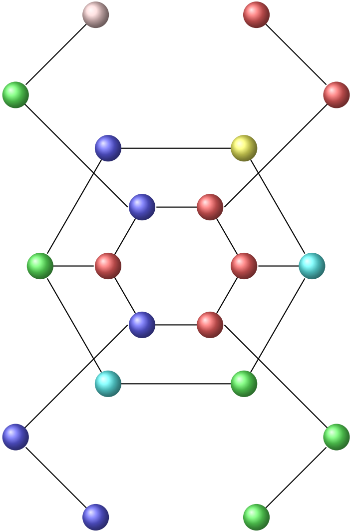
- Each network node is colored by its "function".
- Motif is a (multi) set of colors
- Does the motif appears as a connected subgraph of the network ?
- Topology is only the connectivity of the solution.

# Graph Motif – A toy example

$M$ :

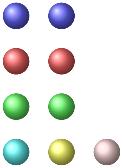


$G$ :

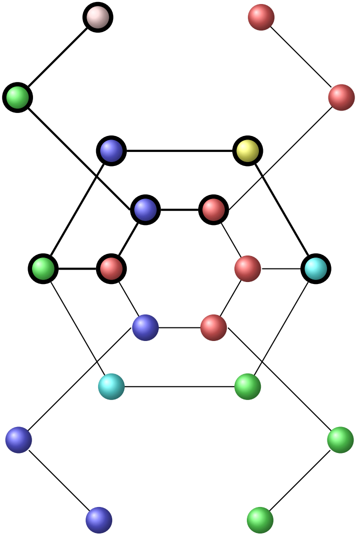


# Graph Motif – A toy example

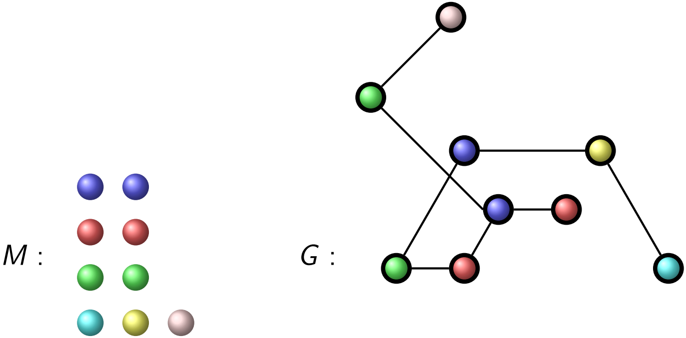
$M$ :



$G$ :



# Graph Motif – A toy example



# Graph Motif

- Applied to different type of biological networks.
  - Initially for metabolic networks [Lacroix et al. 2006].
  - Useful for PPI networks [Bruckner et al. 2009].
- But also for social networks [Betzler et al. 2008, S. 2011].

# Graph Motif – Complexity

- Problem is NP-complete.
  - Even on strong conditions (tree of maximum degree 3, tree of depth 2...) [Fellows et al. 2007, Ambalath et al. 2010]

# Graph Motif – Complexity

- Problem is NP-complete.
  - Even on strong conditions (tree of maximum degree 3, tree of depth 2...) [Fellows et al. 2007, Ambalath et al. 2010]
- Must cope with hardness:
  - Some FPT algorithms.
    - $k$ is the size of the solution.
    - $\mathcal{O}^*(2^k)$ for colorful motifs,
    - $\mathcal{O}^*(4^k)$ for multiset motifs [Guillemot and S. 2010].
    - Hard if the parameter is the number of different colors.

# Graph Motif – Complexity

- Problem is NP-complete.
  - Even on strong conditions (tree of maximum degree 3, tree of depth 2...) [Fellows et al. 2007, Ambalath et al. 2010]
- Must cope with hardness:
  - Some FPT algorithms.
    - $k$ is the size of the solution.
    - $\mathcal{O}^*(2^k)$ for colorful motifs,
    - $\mathcal{O}^*(4^k)$ for multiset motifs [Guillemot and S. 2010].
    - Hard if the parameter is the number of different colors.
  - Approximation.

# Outline

Introduction

Graph Motif : Querying motifs without topology

**Approximate motifs**

# Variants

- Experimental data, **noisy**.
- Ask for an exact occurrence is likely to fail.

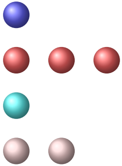
- Must allow insertions, deletions...: optimisation problems!

# A variant for Graph Motif : Minimum Substitutions

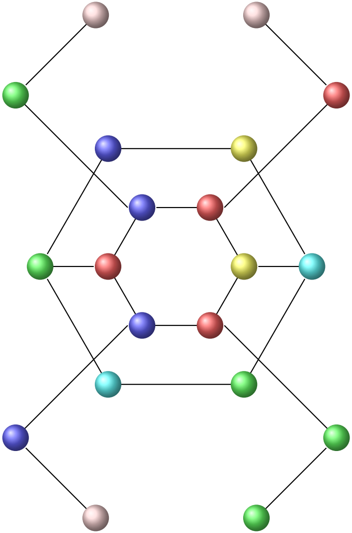
- A variant: MINIMUM SUBSTITUTIONS [DONDI ET AL. 2011].
- Find an occurrence with a maximum number of colors from the motif, but with the same size.
- **Substitution** of motif colors by new colors.

# Minimum Substitutions: Toy example

$M$ :

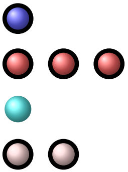


$G$ :

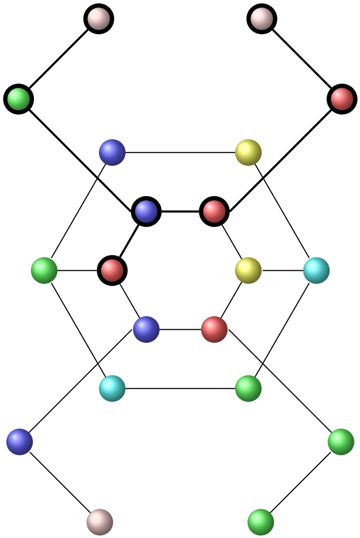


# Minimum Substitutions: Toy example

$M$ :

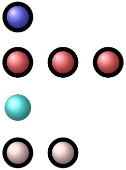


$G$ :

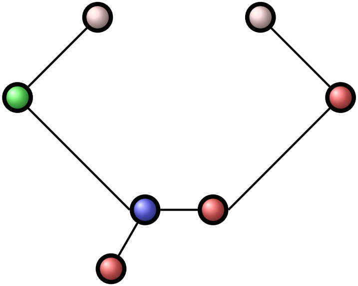


# Minimum Substitutions: Toy example

$M$ :



$G$ :



# A variant for Graph Motif : Minimum Substitutions

- NP-hard even if $G$ is a tree of maximum degree 4 where each color occurs at most twice but FPT [Dondi et al. 2011].
- Result: there is no approximation ratio within $c \log |V|$, unless $\mathrm{P} = \mathrm{NP}$, even if the motif is colorful (at most one occurrence of each color) and $G$ is a depth 2 tree.
- Reduction from Set Cover.

# Construction

$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$

# Construction

$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$

# Construction

$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$

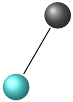
# Construction

$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$

$M$ $G$

# Construction

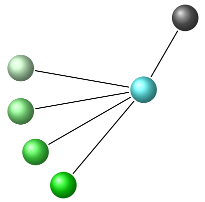
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\},\{x_2, x_3\},\{x_2\}\}$$



$M$ $G$

# Construction

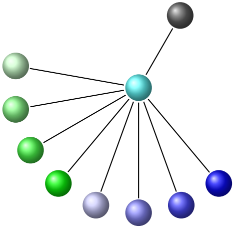
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$



$M$ $G$

# Construction

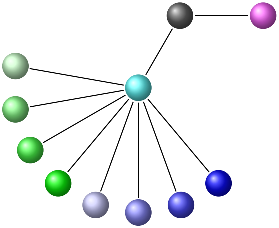
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$



$M$ $G$

# Construction

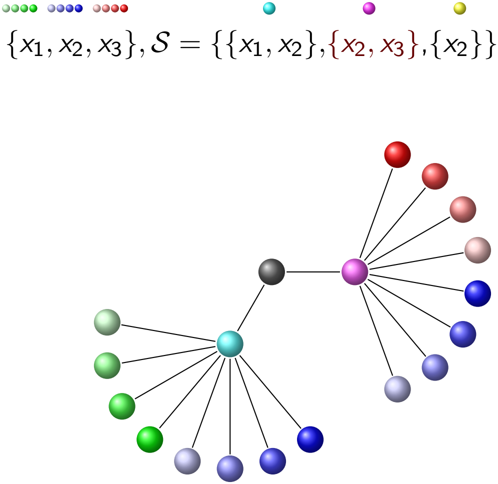
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$



$M$ $G$

# Construction

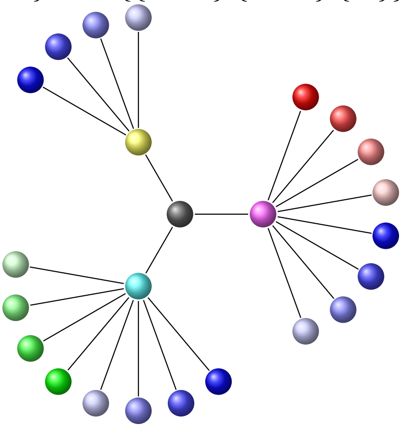
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$



$M$ $G$

# Construction

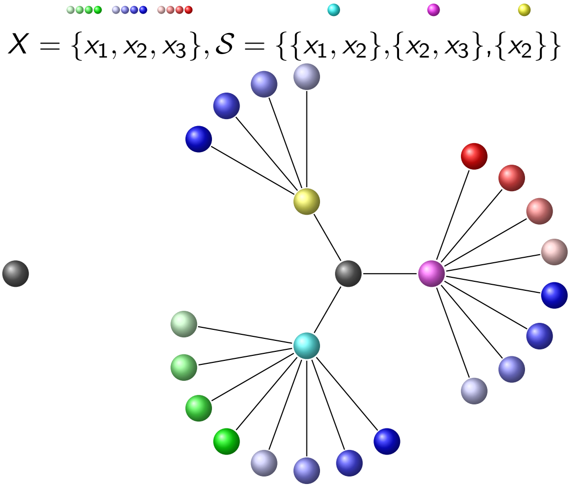
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\},\{x_2, x_3\},\{x_2\}\}$$



$M$ $G$

# Construction

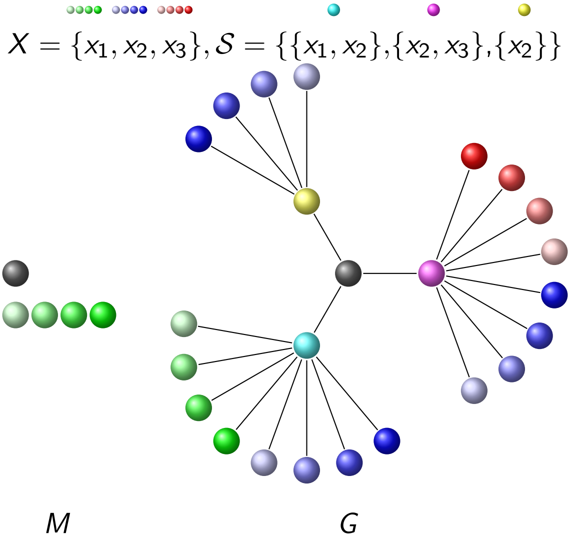
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$



$M$ $G$

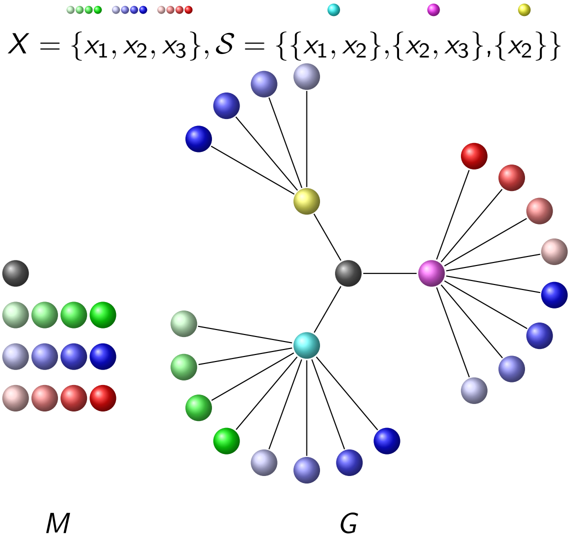
# Construction

$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$



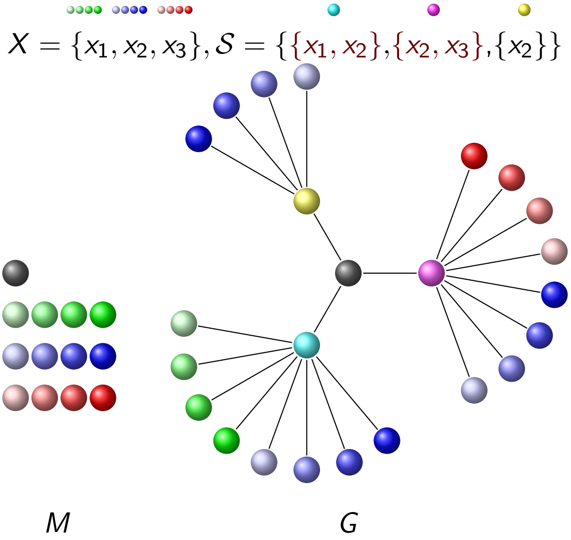
# Construction

$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$



# Construction

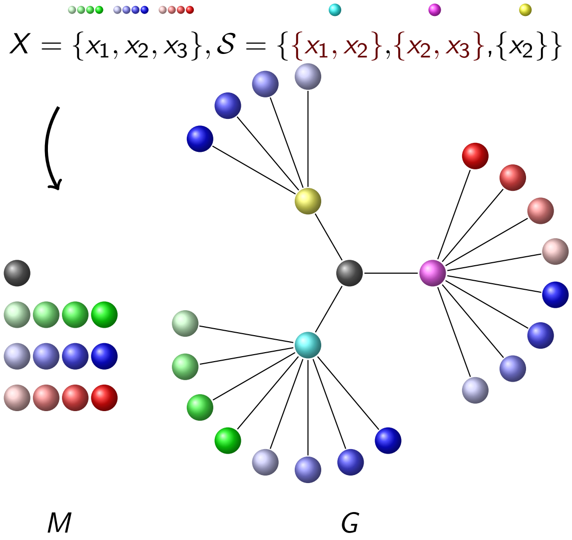
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$




$|\mathcal{S}'|$ = number of substitutions

# Construction

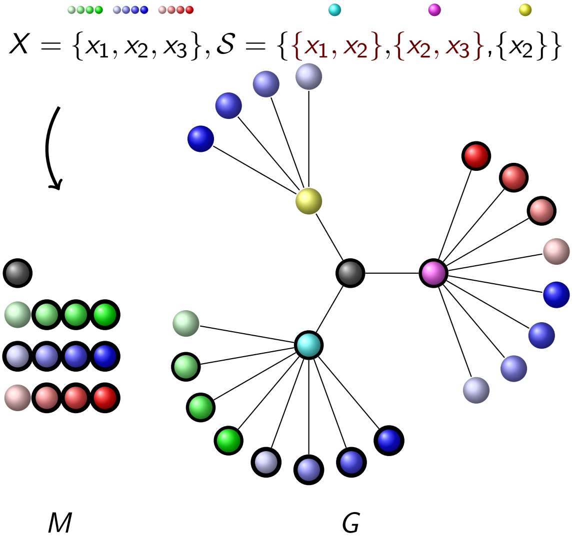
$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$




$$|\mathcal{S}'| = \text{number of substitutions}$$

# Construction

$$X = \{x_1, x_2, x_3\}, \mathcal{S} = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_2\}\}$$




$|\mathcal{S}'|$ = number of substitutions

# Result

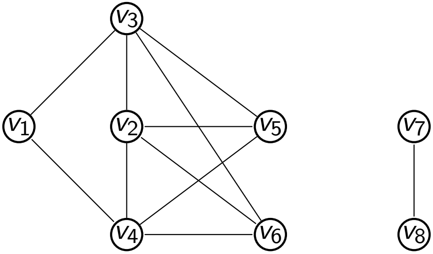
- There is no approximation ratio within $c \log |X|$ for MINIMUM SET COVER (unless P = NP) [RAZ ET SAFRA 1997].

# Result

- There is no approximation ratio within $c \log |X|$ for MINIMUM SET COVER (unless P $=$ NP) [RAZ ET SAFRA 1997].
- There is no approximation ratio within $c \log |V|$ for MINIMUM SUBSTITUTIONS (unless P $=$ NP).

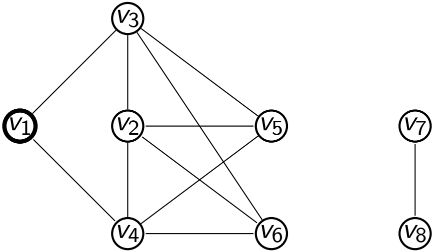
# Using graph modules

- Introduction of the (well-known) notion of **graph modules** into Graph Motif .



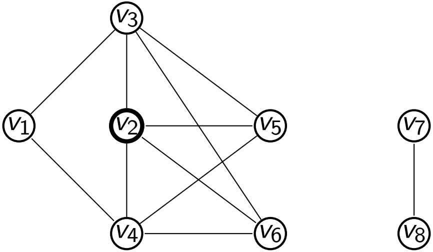
# Using graph modules

- Introduction of the (well-known) notion of **graph modules** into Graph Motif .



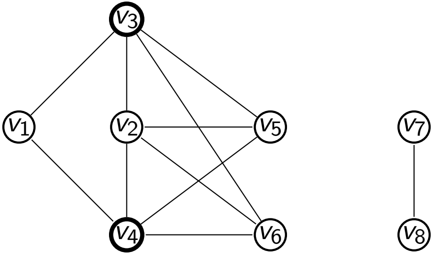
# Using graph modules

- Introduction of the (well-known) notion of **graph modules** into Graph Motif .



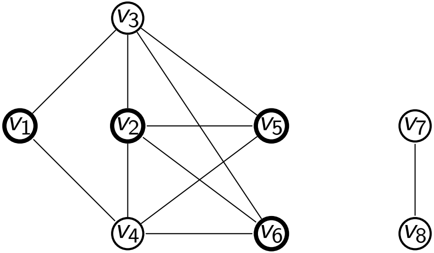
# Using graph modules

- Introduction of the (well-known) notion of **graph modules** into Graph Motif .



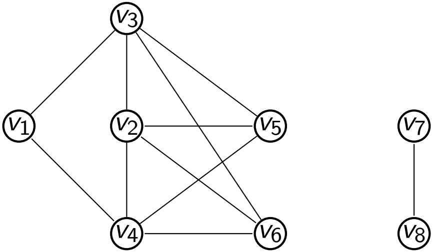
# Using graph modules

- Introduction of the (well-known) notion of **graph modules** into GRAPH MOTIF .



# Using graph modules

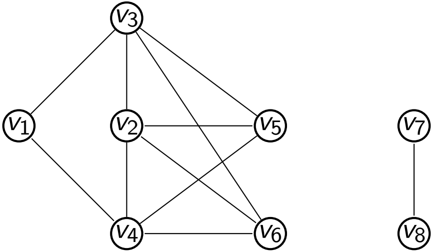
- Introduction of the (well-known) notion of **graph modules** into Graph Motif .




- NP-complete

# Using graph modules

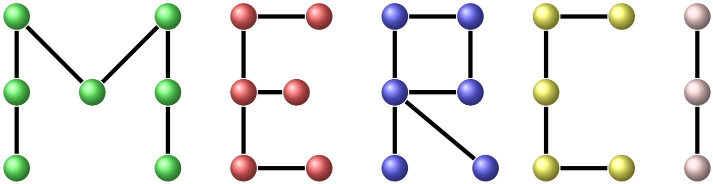
- Introduction of the (well-known) notion of **graph modules** into Graph Motif .




- NP-complete
- Exponential number of modules...
- ... but "generators" can be store in a linear tree.
- Using this to have FPT algorithms.

# Other result in the paper

- MAX MOTIF is hard to approximate.


MERCI