Finding Approximate and Constrained Motifs in Graphs

Riccardo Dondi [1] **Guillaume Fertin** [2]
Stéphane Vialette [3]

[1]Università di Bergamo, [2]Université de Nantes, [3]Université Paris-Est

CPM 2011, June 29, 2011

Outline

Outline

Network Querying Problem

Network analysis (protein-protein interaction, metabolic networks):

- given a query Q (a small network), and a network N
- identify subnetworks of N similar to Q

Approach relies on precise information on the query network $\longrightarrow$ information is often missing

GRAPH MOTIF

[Lacroix, Fernandes, Sagot, TCBB 2006; Bruckner et al, JCB 2010] introduced queries that do not rely on the conservation of the topology:

- network: a vertex-colored graph G
- Query: a motif $\mathcal{M}$(a multiset of colors)
- identify a subgraph of G similar to $\mathcal{M}$

GRAPH MOTIF - Combinatorial Problem

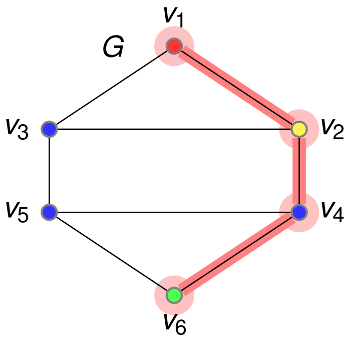
Problem (GRAPH MOTIF)

Input: *a vertex-colored graph G and a colored motif $\mathcal{M}$ (a multiset of colors).*
Output: *a connected component $G' = (V', E')$ of G so that $\mathcal{C}(V') = \mathcal{M}$.*

G' is an occurrence of $\mathcal{M}$ in G

When $\mathcal{M}$ is a set $\rightarrow \mathcal{M}$ is colorful

GRAPH MOTIF - Combinatorial Problem



$\mathcal{M} = \{red, yellow, blue, green\}$

An occurrence of $\mathcal{M}$: $\{v_1, v_2, v_4, v_6\}$

GRAPH MOTIF - Computational Complexity

GRAPH MOTIF is **NP**-complete:

- even if $\mathcal{M}$ is colorful and the graph G is a tree and each colors has at most 3 occurrences in G [Fellows et al, JCSS 2011]
- even if $\mathcal{M}$ consists of two colors, and the graph is bipartite with maximum degree 4 [Fellows et al, JCSS 2011]

GRAPH MOTIF - Parameterized Complexity

Previous results:

- GRAPH MOTIF is in FPT, when parameterized by $k = |\mathcal{M}|$ [Fellows et al, JCSS 2011; Betzler et al, TCBB 2011; Guillemot, Sikora, MFCS 2010]

In [Betzler et al, TCBB 2011] *colorful recoloring* technique:

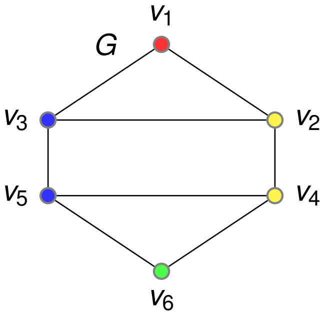
Lemma (Betzler et al, TCBB 2011)

Given a motif $\mathcal{M}$, the number of trials to achieve a colorful recoloring of $\mathcal{M}$ with an error probability of ε is $|\ln(\varepsilon)| \cdot O(e^{|\mathcal{M}|})$.

Outline

1. Introduction
   - Biological Motivations
   - GRAPH MOTIF
2. Extensions of GRAPH MOTIF
3. MIN-SUB and MIN-ADD
   - Computational Complexity
   - Parameterized Complexity
4. CGM
   - CGM - Parameterized Complexity
5. Conclusion

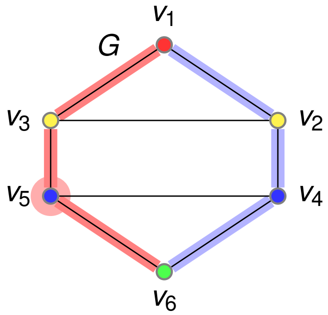
Extensions of GRAPH MOTIF



$\mathcal{M} = \{red, yellow, blue, green\}$

Measurement errors: the GRAPH MOTIF problem too stringent

Extensions of GRAPH MOTIF



$\mathcal{M} = \{red, yellow, blue, green\}$

Previous knowledge on the structure of motif: mandatory vertices

Extensions of GRAPH MOTIF

Variants GRAPH MOTIF:

- MIN-CC, MAXIMUM MOTIF [`Dondi, Fertin, Vialette, JDA 2011`]
- approximate occurrence of $\mathcal{M}$: substitutions (MIN-SUB), additions (MIN-ADD) [`Bruckner et al, JCB 2010`]
- exact occurrence of $\mathcal{M}$ containing mandatory vertices: CGM

MIN-SUB and MIN-ADD

Problem (MIN-SUB)

Input: *a vertex-colored graph G and a colored motif* $\mathcal{M}$.
Output: *a connected subgraph* $G' = (V', E')$ *of G, such that* $C(V')$ *can be obtained with at most p substitutions from* $\mathcal{M}$.

Problem (MIN-ADD)

Input: *a vertex-colored graph G and a colored motif* $\mathcal{M}$.
Output: *a connected subgraph* $G' = (V', E')$ *of G, such that* $C(V') \supseteq \mathcal{M}$ *and* $|C(V') \setminus \mathcal{M}| \leq p$.

CGM

Problem (CGM)

Input: *a vertex-colored graph* $G = (V, E)$*, a set of mandatory vertices* $V_M \subseteq V$ *and a colored motif* $\mathcal{M}$.
Output: *a connected subgraph* $G' = (V', E')$ *of G, such that* $C(V') = \mathcal{M}$ *and* $V_M \subseteq V'$.

Optinal occurrences=$\mathcal{M} \setminus c(V_M)$

Outline

MIN-SUB and MIN-ADD - Computational Complexity

Theorem

MIN-SUB *and* MIN-ADD *are* **NP***-hard, even when the graph is a tree* T *of maximum degree* 4, $\mathcal{M}$ *is colorful, each color has at most* 2 *occurrences in* T.

Proof.

Reductions from Minimum Vertex Cover on cubic graphs. □

MIN-SUB and MIN-ADD- Parameterized Complexity

MIN-SUB and MIN-ADD are **NP**-hard when the size of the solution is a constant

MIN-ADD is in **FPT** when parameterized by $|\mathcal{M}|$ [Guillemot, Sikora, MFCS 2010; Bruckner et al, JCB 2010]

Lemma

MIN-SUB *is in* **FPT** *when parameterized by* $|\mathcal{M}|$.

MIN-SUB and MIN-ADD- Parameterized Complexity

Lemma

MIN-SUB *is in* **FPT** *when parameterized by* $|\mathcal{M}|$.

Proof.

Colorful case:

- Visit the vertices of a connected component of G, allowing to visit some vertices more than once.
- Each color visited more than once $\rightarrow$ substitution.

General case:
Combination with colorful recoloring. □

Outline

CGM - FPT Algorithm

Theorem

The CGM *problem is in FPT, when parameterized by the number of optional occurrences k and by the treewidth δ of the input graph.*

Proof.

FPT algorithm for the colorful case:

- nice tree decomposition of the input graph
- each bag of the nice tree decomposition has size at most $\delta + 1$

□

CGM - FPT Algorithm

FPT algorithm for the colorful case:

- each subset X_i' of a bag X_i is partitioned in at most δ connected components: at most $\delta^{\delta+1}$ possible partitions
- each partition must be feasible

Definition

A partition P of $X_i' \subseteq X_i$ is *feasible* when

1. $X_i \cap V_M \subseteq X_i'$;
2. for each pair of vertices $u, v \in X_i'$, $c(u) \neq c(v)$;
3. each set of P is a maximal connected component of $G[X_i']$.

CGM - Hardness of Parameterization

Theorem

The CGM *problem, parameterized by the number of optional occurrences, is* $W[2]$*-hard, even when the input graph is of diameter* 2*.*

Proof.

Parameterized preserving reduction from MIN SET COVER. □

Outline

1. Introduction
   - Biological Motivations
   - GRAPH MOTIF
2. Extensions of GRAPH MOTIF
3. MIN-SUB and MIN-ADD
   - Computational Complexity
   - Parameterized Complexity
4. CGM
   - CGM - Parameterized Complexity
5. Conclusion

Future Directions

Future directions:

- More efficient fixed-parameter algorithms for MIN-SUB and MIN-ADD
- Efficient heuristics for MIN-SUB, MIN-ADD, CGM
- New variants of GRAPH MOTIF

Conclusion

Conclusion:

- GRAPH MOTIF and biological motivations
- Extensions of GRAPH MOTIF $\Rightarrow$ MIN-SUB, MIN-ADD, CGM
- Hardness results for MIN-SUB and MIN-ADD
- MIN-SUB is in FPT
- CGM is in FPT for graphs of bounded treewidth
- CGM is W[2]-hard for graphs of maximum diameter 2
- Future directions