Complexity and Polynomial-Time Approximation Algorithms around the Scaffolding Problem

Annie Chateau – Rodolphe Giroudeau
LIRMM, Montpellier

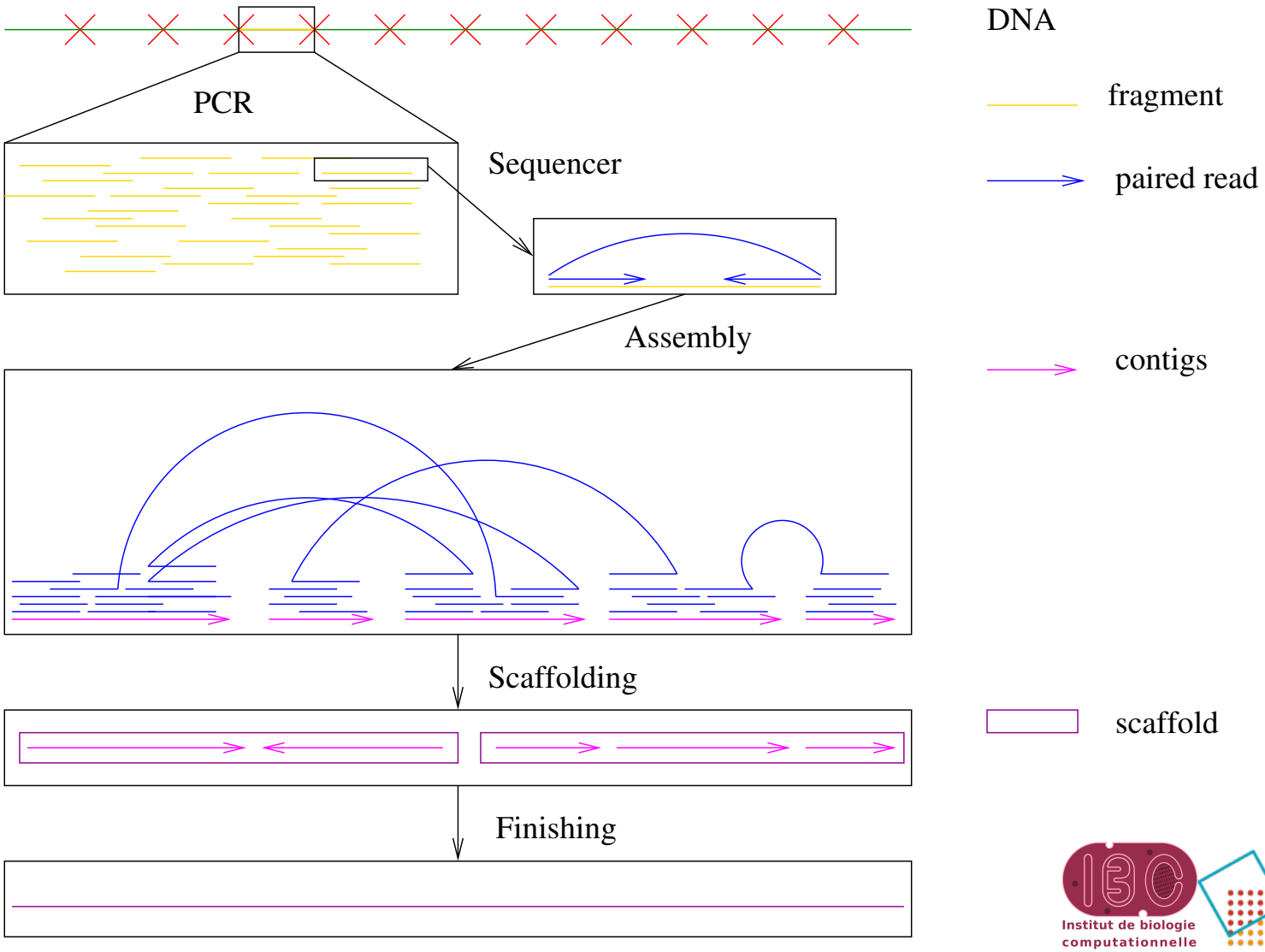
AlCoB 2014 – Tarragona, July 1-3

Institut de biologie computationnelle
Laboratoire Informatique Robotique Microélectronique Montpellier

Sommaire

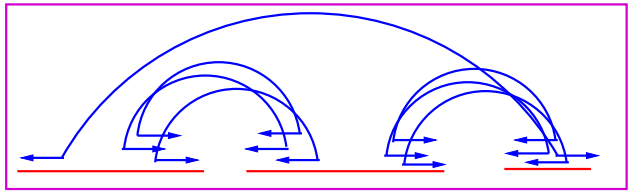
Institut de biologie computationnelle
Laboratoire Informatique Robotique Microélectronique Montpellier

From the molecule to genomic sequence



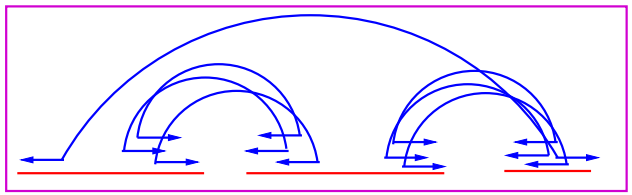
Generation of the scaffold graph

- Paired reads are often used for scaffolding



Generation of the scaffold graph

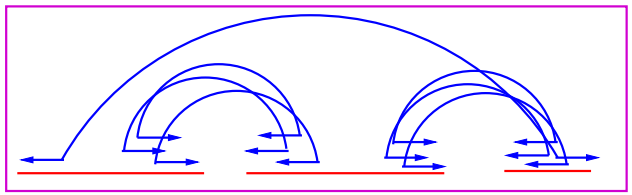
- Paired reads are often used for scaffolding



- **Mapping** of the reads on the contigs and compute the number of pairs of reads supporting each hypothesis
  $\Rightarrow$ the scaffold graph

Generation of the scaffold graph

- Paired reads are often used for scaffolding

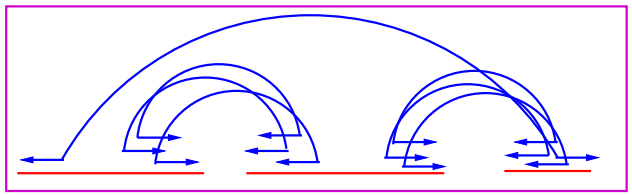


- **Mapping** of the reads on the contigs and compute the number of pairs of reads supporting each hypothesis
  $\Rightarrow$ the scaffold graph

- Nodes = contigs extremities

Generation of the scaffold graph

- Paired reads are often used for scaffolding



- **Mapping** of the reads on the contigs and compute the number of pairs of reads supporting each hypothesis
  $\Rightarrow$ the scaffold graph

- Nodes = contigs extremities
- Edges = $\begin{cases} \text{contigs} \\ \text{stories} \end{cases}$ (weight = number of pairs of reads)

Constrained genomic structure

We consider a given genomic structure $(\sigma_p, \sigma_c) \in \mathbb{N} \times \mathbb{N} \backslash \{(0,0)\}$, where :

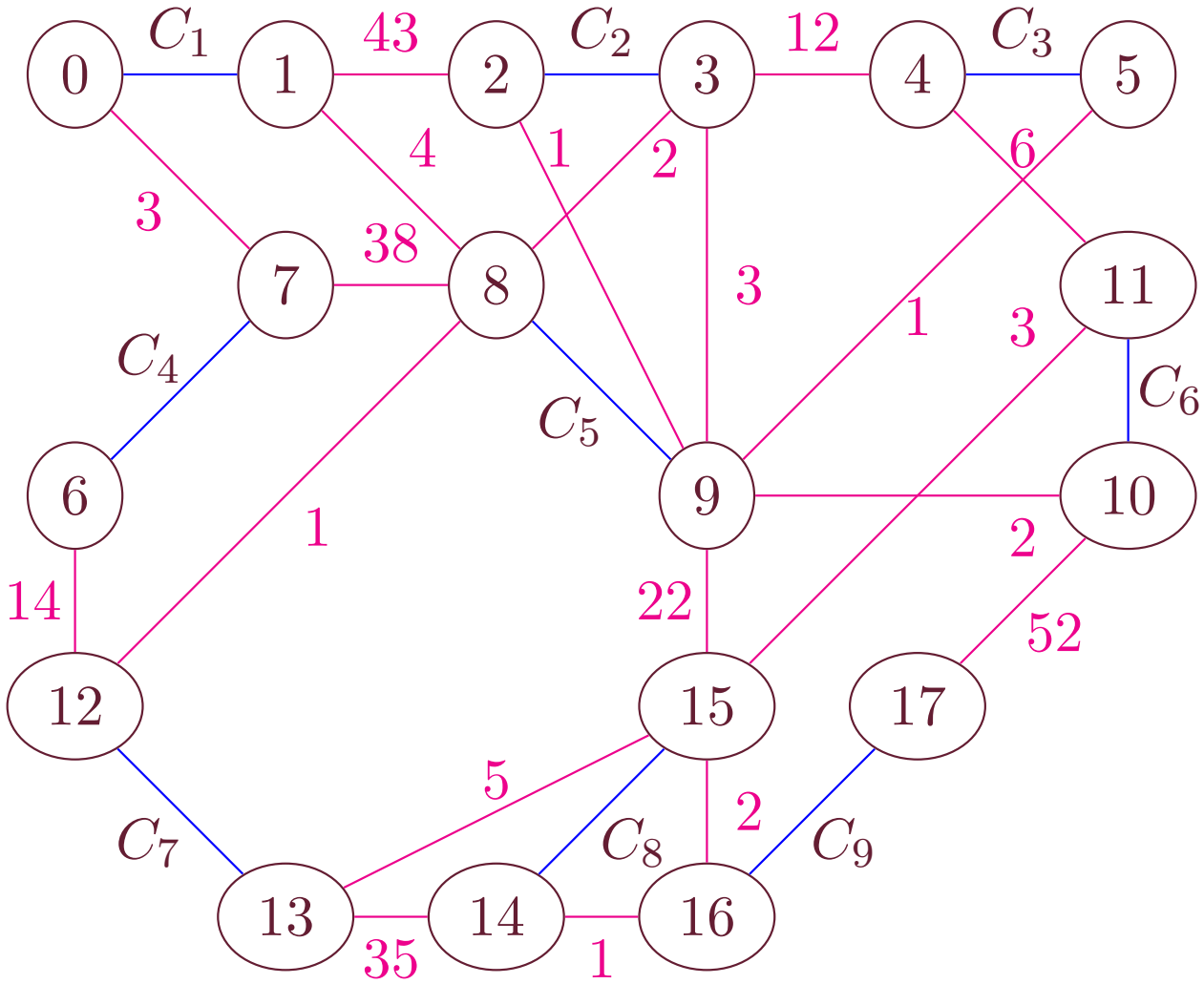
- σ_p is the number of linear chromosomes (paths)
- σ_c is the number of circular chromosomes (cycles)

What is the best possible way to scaffold according to this constrained structure ?

Institut de biologie computationnelle

Laboratoire Informatique Robotique Microélectronique Montpellier

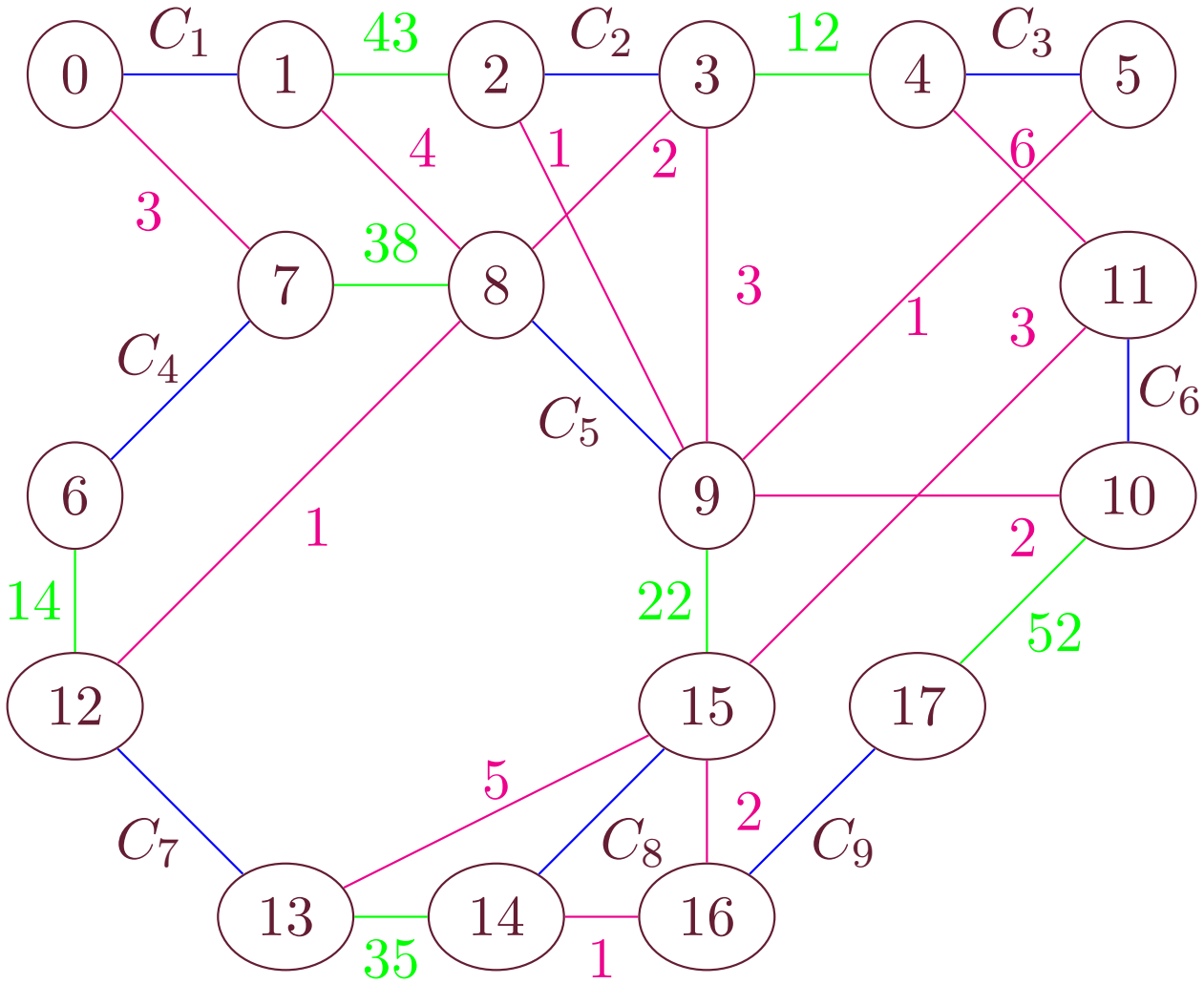
The scaffold graph



$n = 9$
$\sigma_p = 2$
$\sigma_c = 1$
18 vertices
9 contig edges
20 weighted edges

Institut de biologie computationnelle
Laboratoire Informatique Robotique Microélectronique Montpellier

The scaffold graph



$n = 9$
$\sigma_p = 2$
$\sigma_c = 1$
18 vertices
9 contig edges
20 weighted edges

Institut de biologie computationnelle
Laboratoire Informatique Robotique Microélectronique Montpellier

The (σ_p, σ_c)-Scaffold Problems

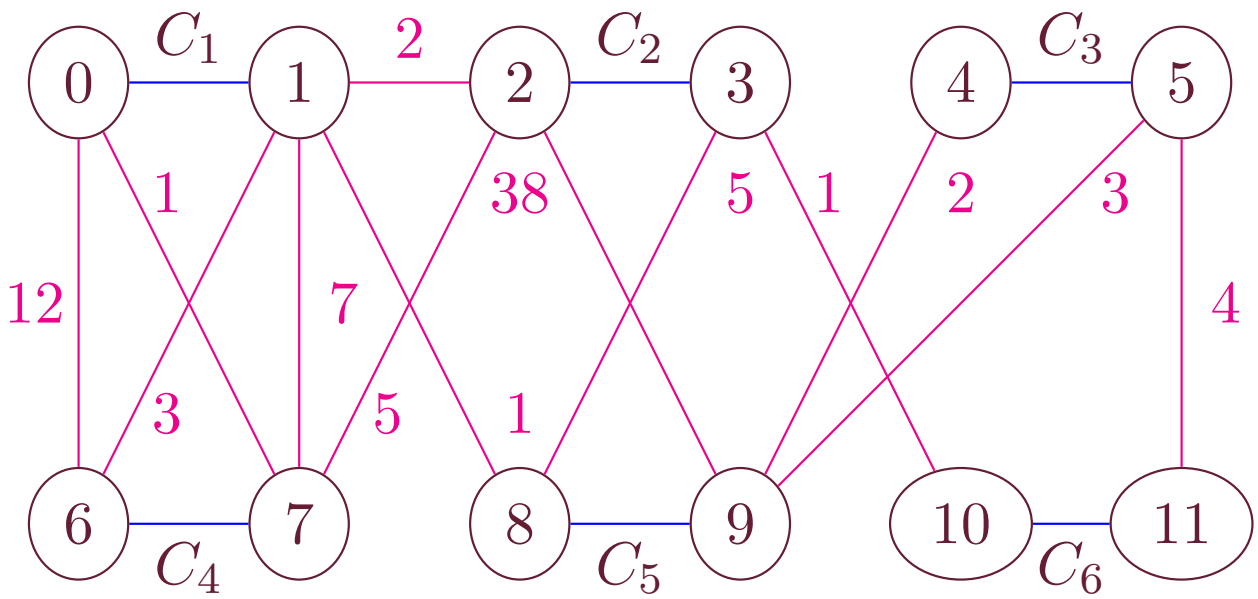
(Decision) $(\sigma_p, \sigma_c)-$Scaffold Problem (SP) :
Instance : $G = (V, E)$ graph with $2n$ vertices. Let M^* be a perfect matching of G, and $(\sigma_p, \sigma_c) \in \mathbb{N} \times \mathbb{N} \backslash \{(0, 0)\}$.
Question : Does it exist a vertex disjoint collection of exactly σ_p alternating-paths and σ_c alternating-cycles, covering the vertices of G ?

(Optimization)Min/Max-$(\sigma_p, \sigma_c)-$Scaffold Problem :
Instance : $G = (V, E, w)$ graph with $2n$ vertices. Let M^* be a perfect matching of G, and $(\sigma_p, \sigma_c) \in \mathbb{N} \times \mathbb{N} \backslash \{(0, 0)\}$.
Question : Find a vertex disjoint collection of exactly σ_p alternating-paths and σ_c alternating-cycles, covering the vertices of G, and of minimal (resp. maximal) total weight.

Sommaire

Polynomial case

Theorem
When the size n of the matching M^* satisfies $n = \sigma_p + 2\sigma_c$, the problem (σ_p, σ_c)-SP (resp. Min/Max-(σ_p, σ_c)−SP) is polynomial.

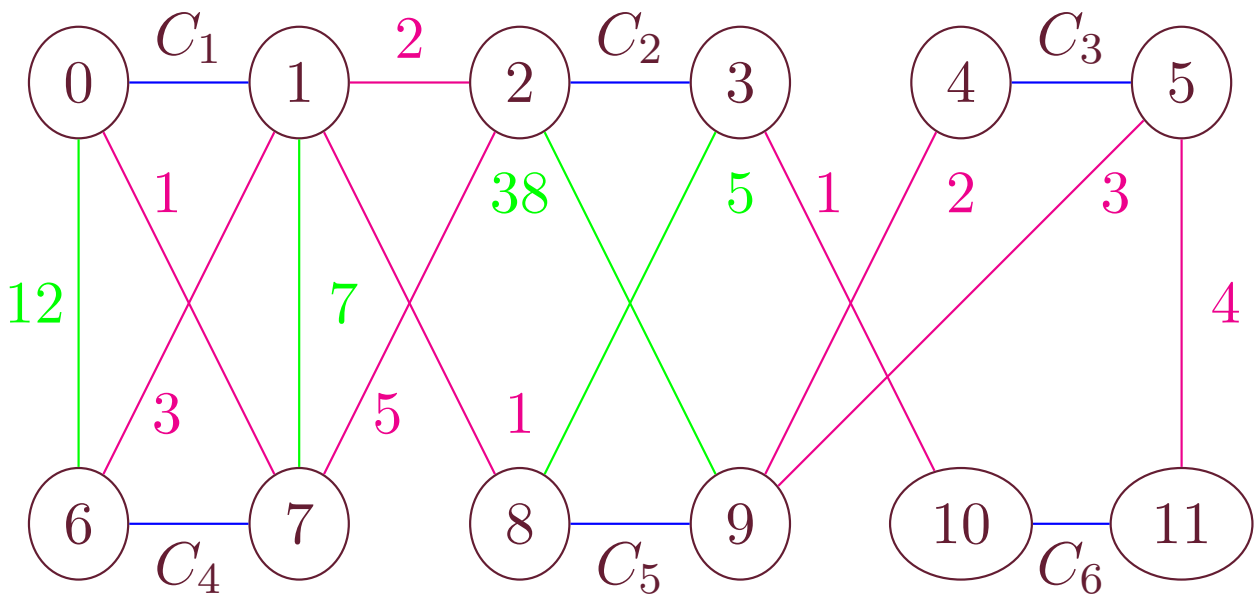


$n = 6$
$\sigma_p = 2$
$\sigma_c = 2$

Institut de biologie computationnelle
Laboratoire Informatique Robotique Microélectronique Montpellier

Polynomial case

Theorem
When the size n of the matching M^* satisfies $n = \sigma_p + 2\sigma_c$, the problem (σ_p, σ_c)-SP (resp. Min/Max-(σ_p, σ_c)−SP) is polynomial.



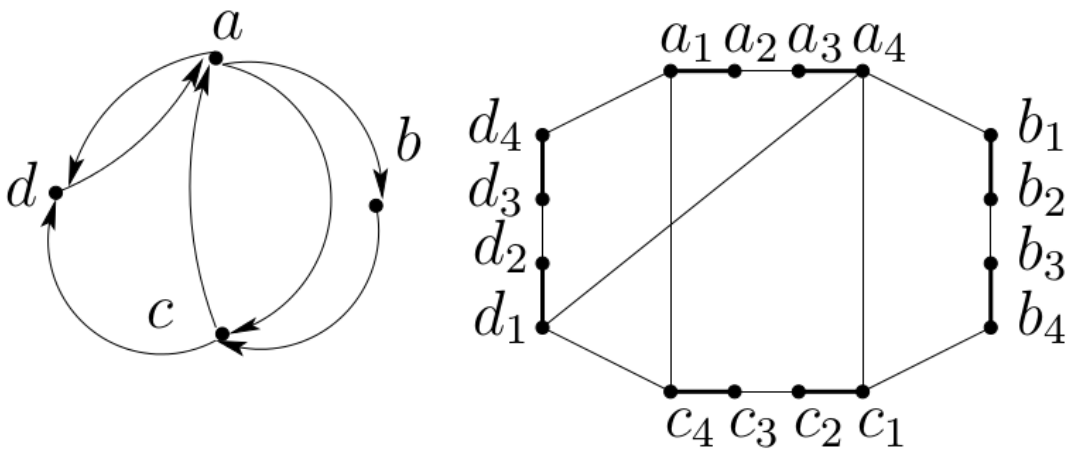
$n = 6$
$\sigma_p = 2$
$\sigma_c = 2$

Institut de biologie computationnelle
Laboratoire Informatique Robotique Microélectronique Montpellier

NP-completeness

Theorem
The problem $(\sigma_p, \sigma_c)-$SP is $\mathcal{NP}-$complete, even if the graph is bipartite.

Idea : reduction from Directed Hamiltonian Cycle



Sommaire

Approximability

Approximation algorithm $\mathcal{A}$ with guaranteed ratio $\rho > 1$:

- $S_{\mathcal{A}}$ the solution given by $\mathcal{A}$

Approximability

Approximation algorithm $\mathcal{A}$ with guaranteed ratio $\rho > 1$:

- $S_{\mathcal{A}}$ the solution given by $\mathcal{A}$
- S_{opt} the optimal solution

Approximability

Approximation algorithm $\mathcal{A}$ with guaranteed ratio $\rho > 1$:

- $S_{\mathcal{A}}$ the solution given by $\mathcal{A}$
- S_{opt} the optimal solution
- $w(S_{\mathcal{A}}) \leq w(S_{opt}) \leq w(S_{\mathcal{A}}) \times \rho$ (maximization)

Non-approximability results for the Min problem

Theorem

The problem Min$-(\sigma_p, \sigma_c)-$SP *is non-approximable, unless* $\mathcal{P} = \mathcal{NP}$*, even if the graph is bipartite.*

Polynomial-time approximation algorithms for the Max problem

- Greedy algorithm : ratio 3
- Perfect matching based algorithm : ratio 3 in the general case, ratio 2 in the case $(0, 1)$

Greedy algorithm

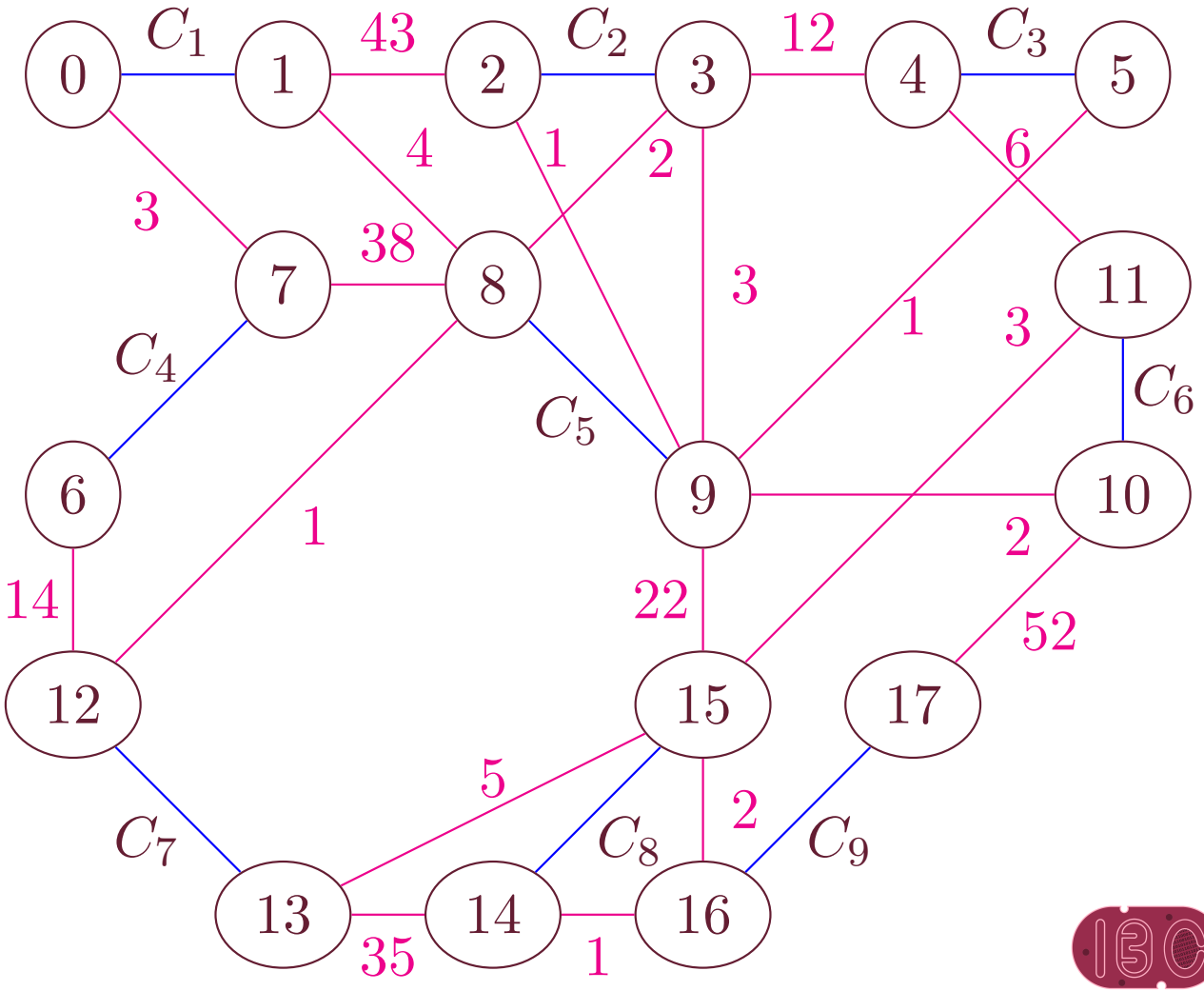
- Sort the edges by decreasing weight
- Add the edges in that order, if it is compatible with the desired structure
- When an edge is chosen, eliminate the contiguous edges
- Until we obtain exactly σ_p paths and σ_c cycles

Theorem

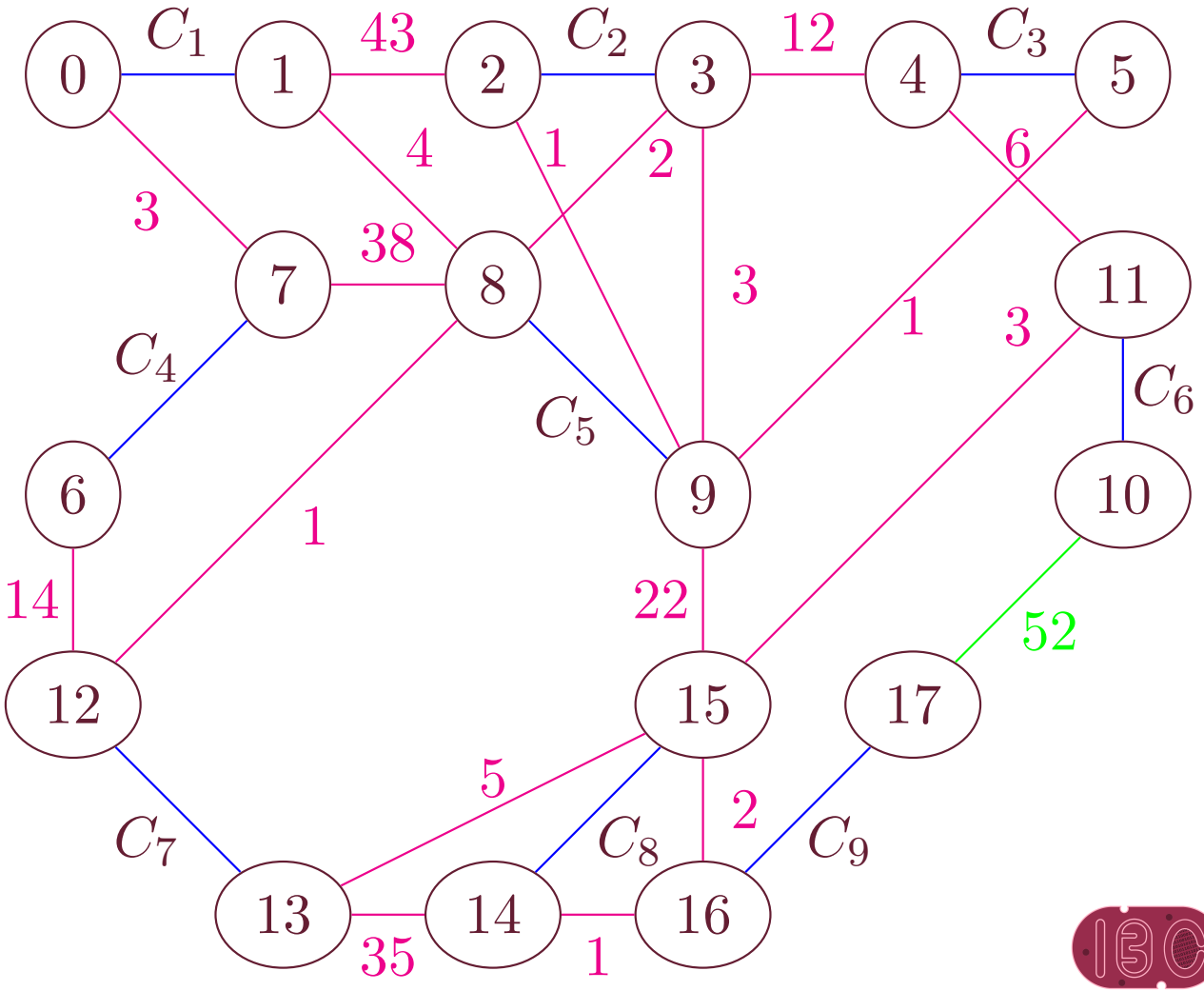
This algorithm gives a solution for the Max-(σ_p, σ_c)−scaffold problem *in complete graphs with non-negative weight, with a tight approximation rate of* 3 *when* $n \geq 2(\sigma_p + 2\sigma_c)$*, and a time complexity* $\mathcal{O}(n^2 \log n)$.

Institut de biologie computationnelle
Laboratoire Informatique Robotique Microélectronique Montpellier

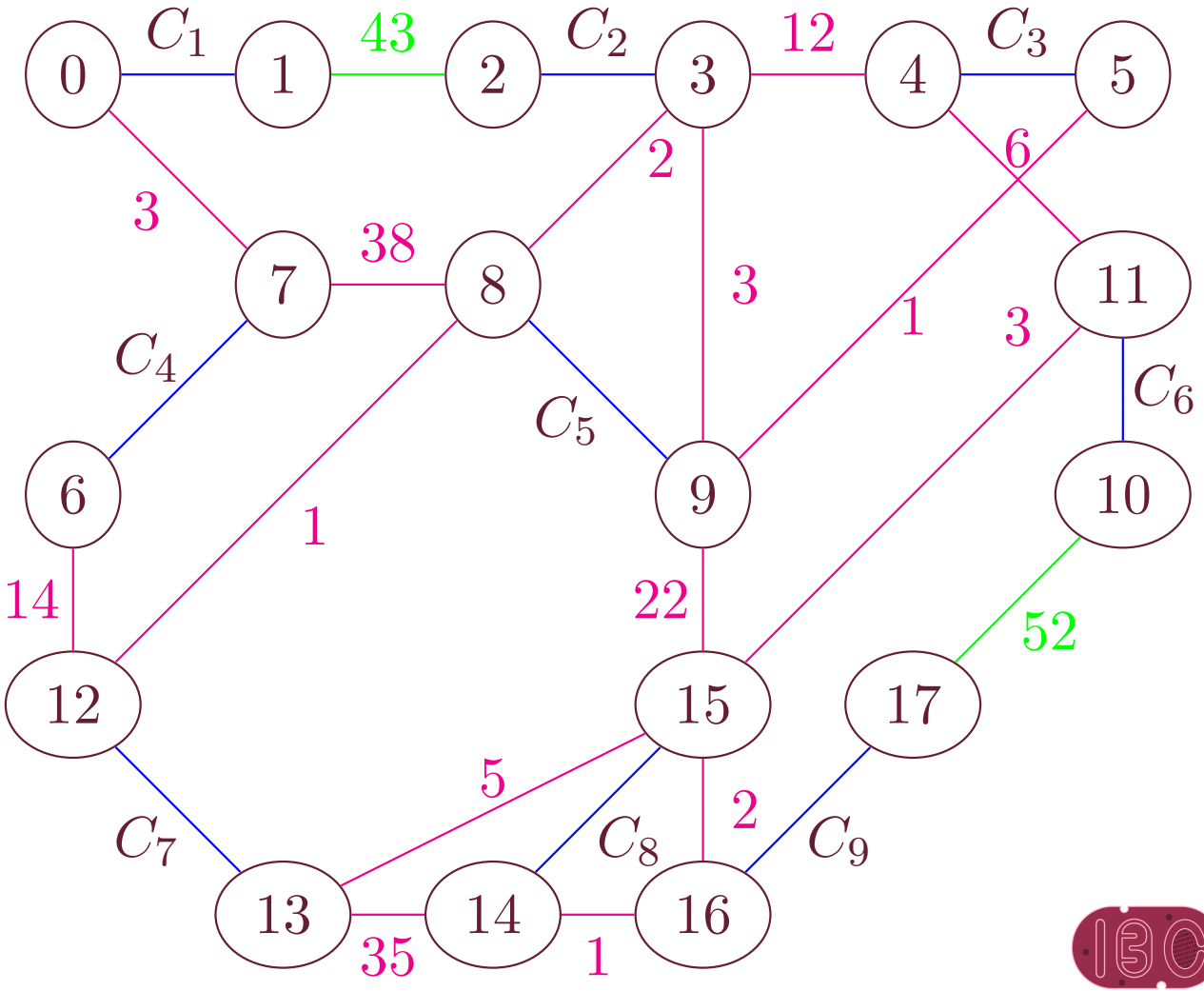
Greedy algorithm



Greedy algorithm



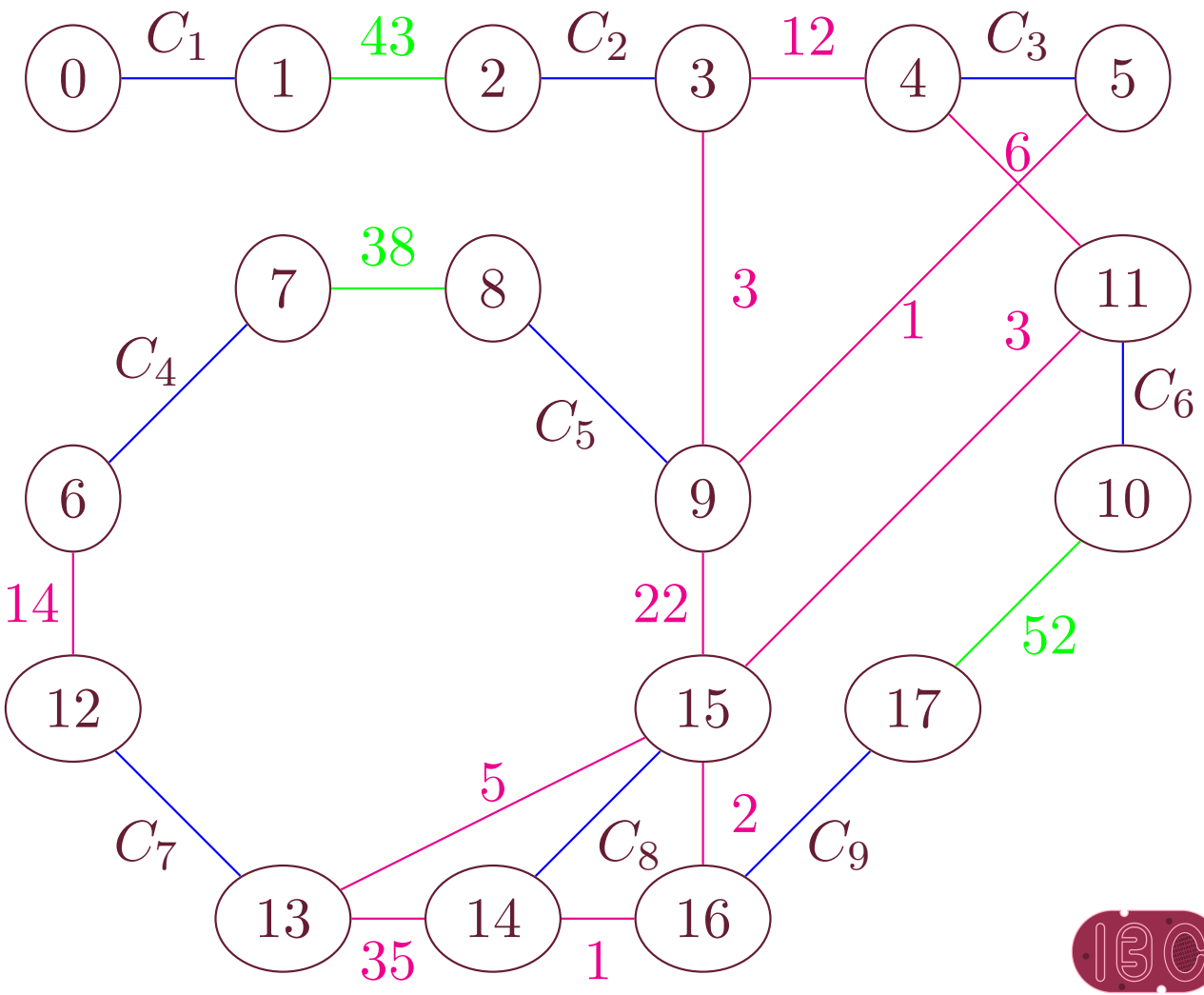
Greedy algorithm



Institut de biologie computationnelle

Laboratoire Informatique Robotique Microélectronique Montpellier

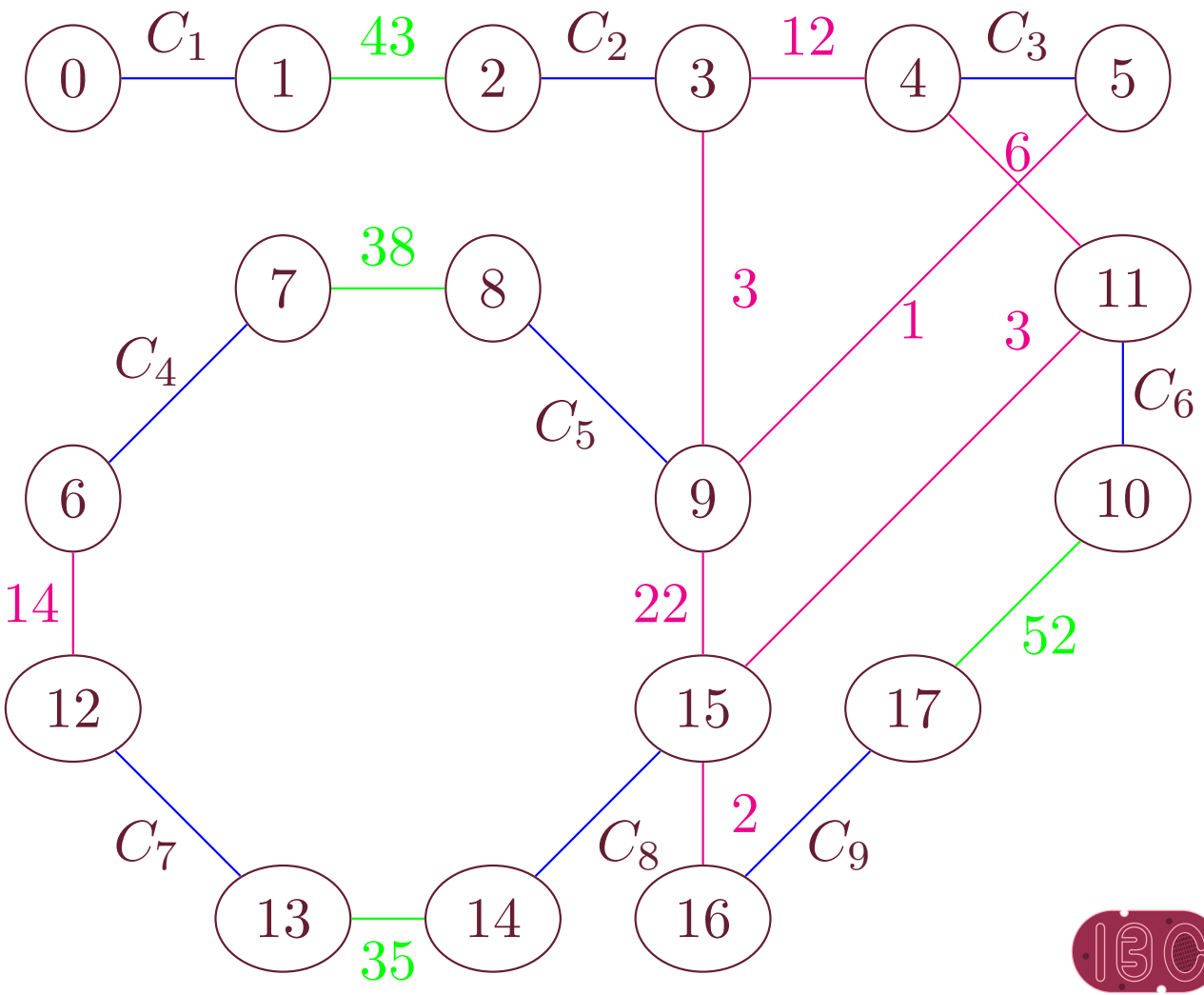
Greedy algorithm



Institut de biologie computationnelle

Laboratoire Informatique Robotique Microélectronique Montpellier

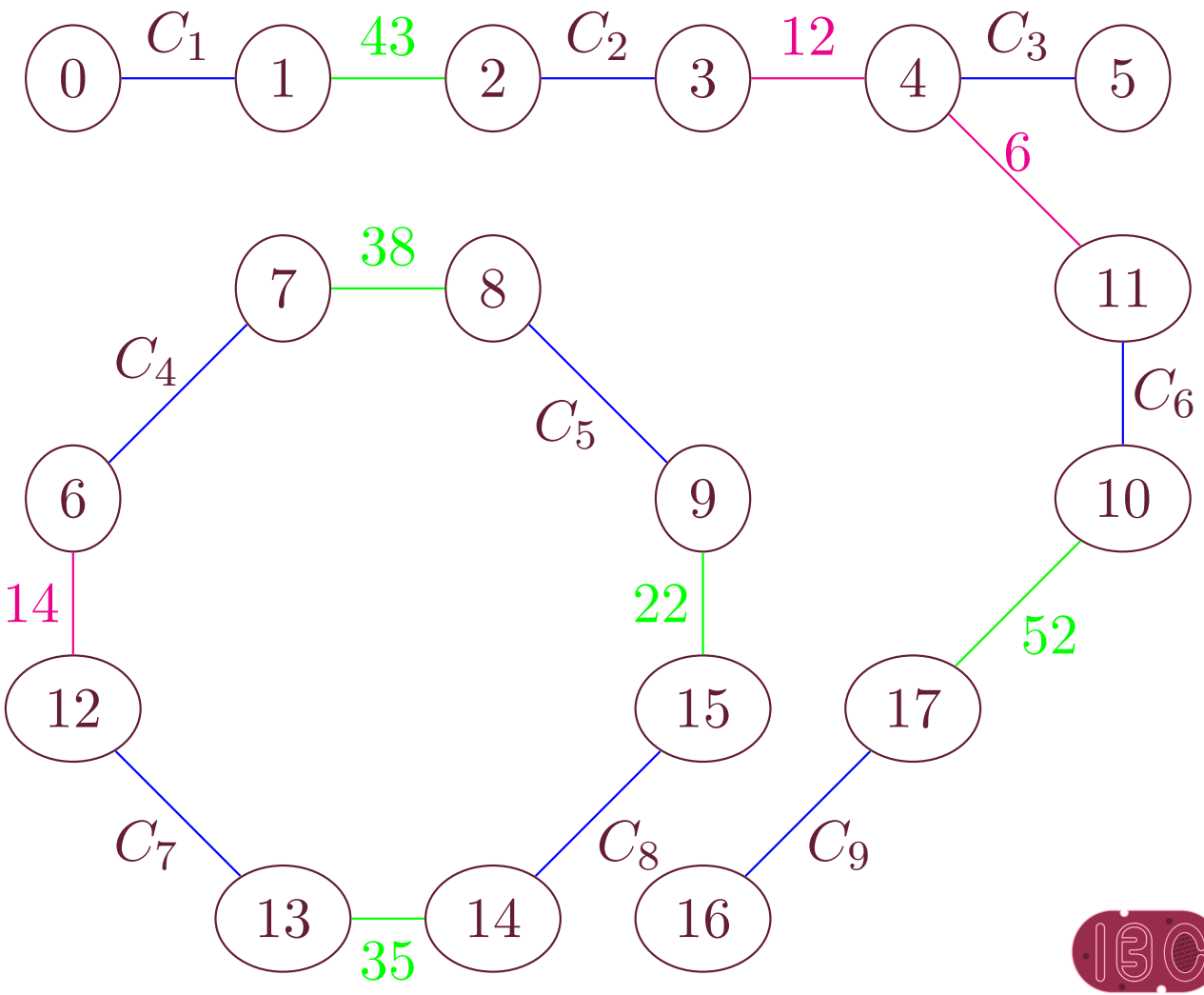
Greedy algorithm



Institut de biologie computationnelle

Laboratoire Informatique Robotique Microélectronique Montpellier

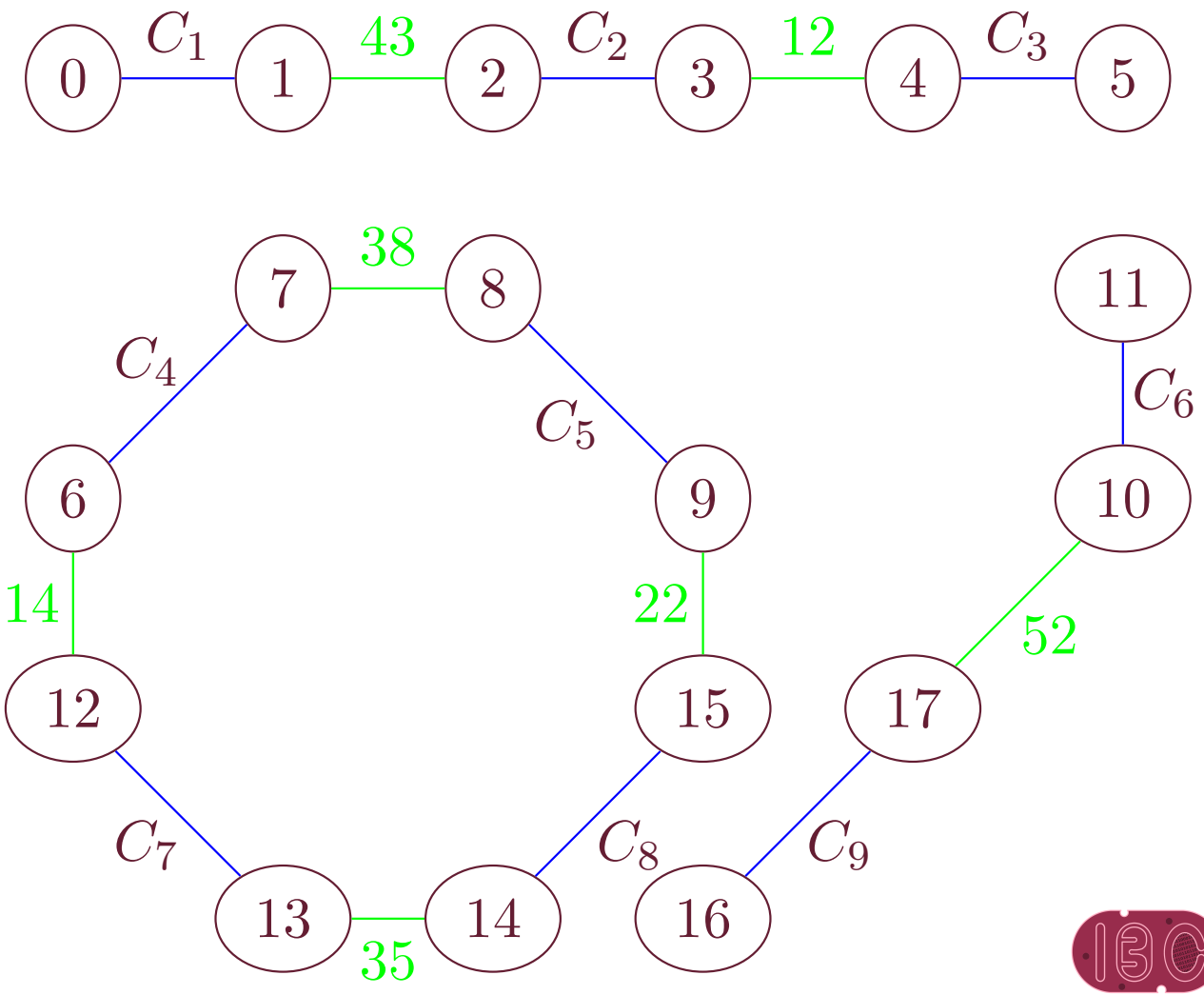
Greedy algorithm



Institut de biologie computationnelle

Laboratoire Informatique Robotique Microélectronique Montpellier

Greedy algorithm

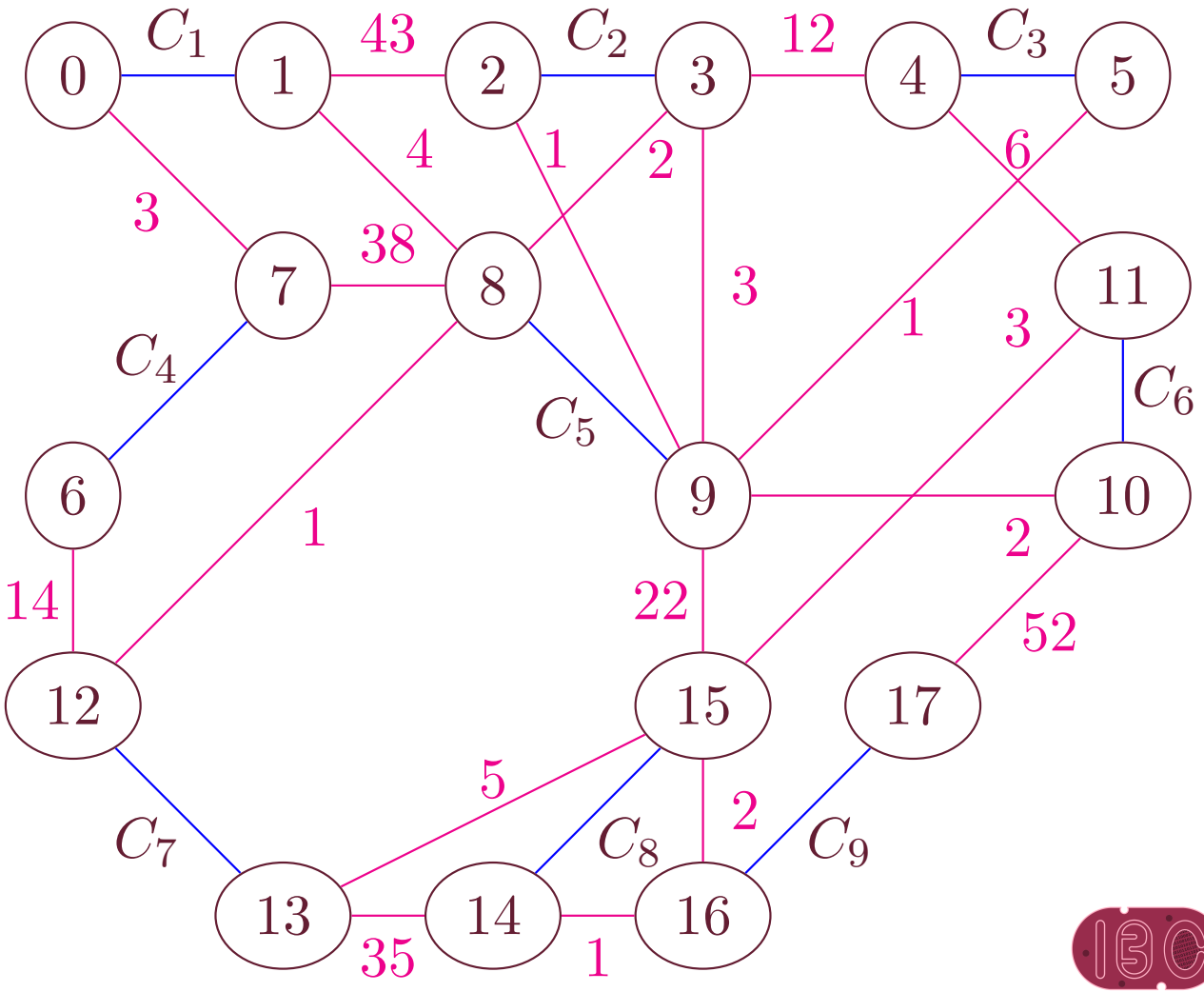


Perfect matching algorithm

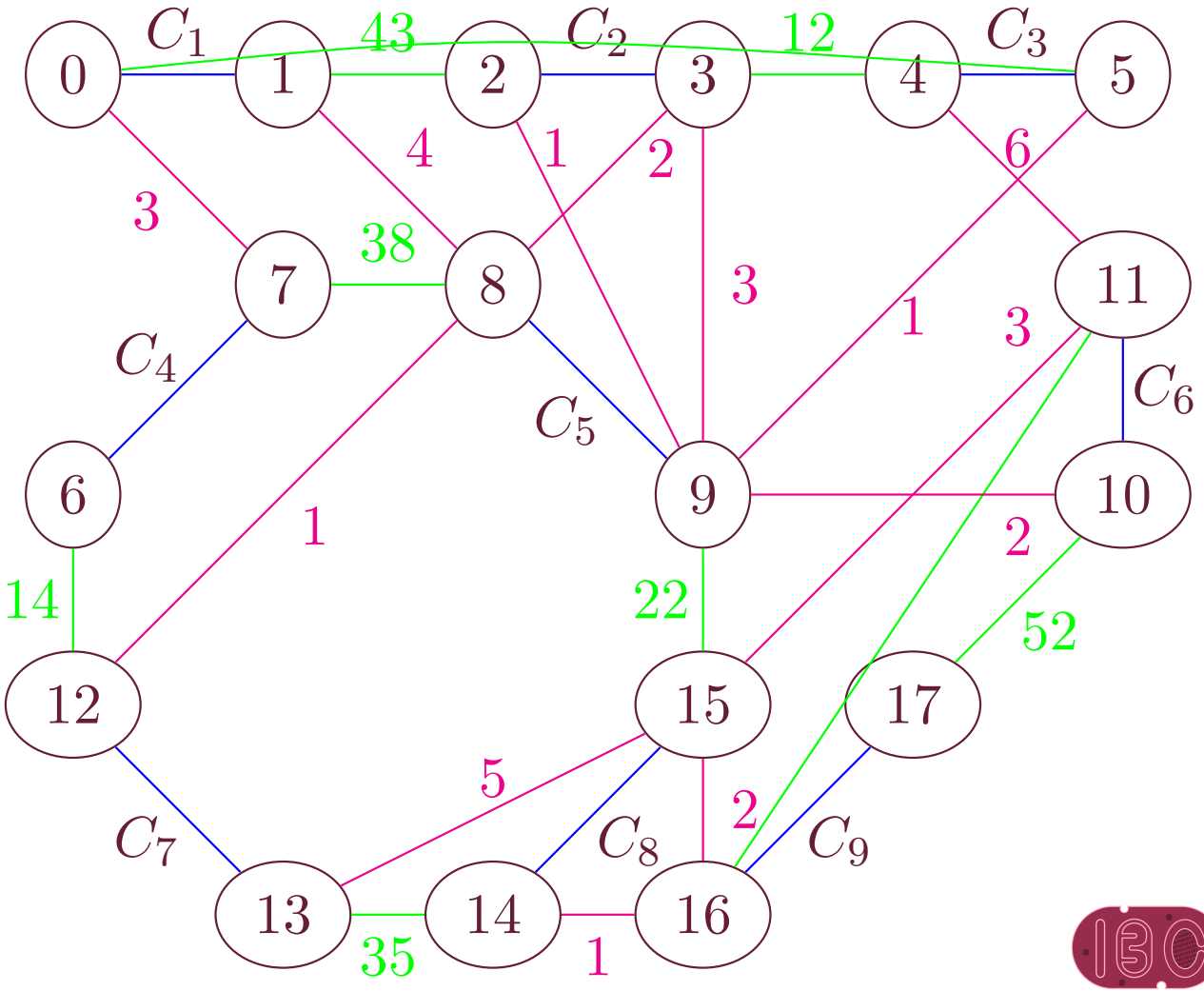
- First compute a perfect matching of maximal weight $\rightarrow$ cycle cover of the graph
- Select a subset of "removable edges" that can be temporarily erased
- Reconnect the paths to form σ_p paths and σ_c cycles

Theorem
When $n \geq 2(\sigma_p + 2\sigma_c)$, this algorithm provides a solution for the Max$-(\sigma_p, \sigma_c)-$SP *in complete graphs with non-negative weights, with a tight approximation ratio of three and a time complexity $\mathcal{O}(n^3)$.*

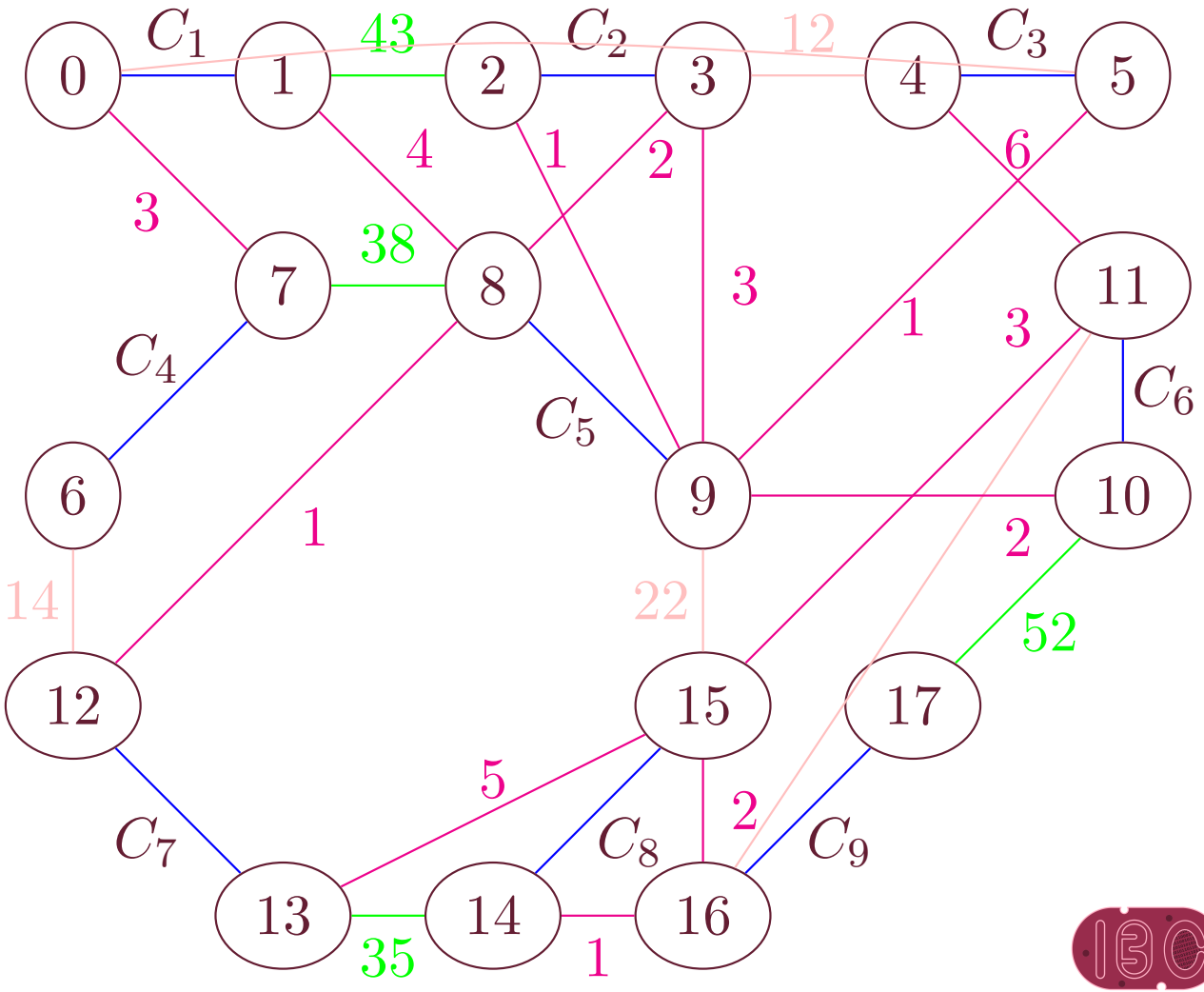
Perfect matching algorithm



Perfect matching algorithm



Perfect matching algorithm



Sommaire

Conclusion

Problems	Complexity	Approximability	
	Decision	Min	Max
$(\sigma_p, \sigma_c)-$SP	$\mathcal{NP}$-Complete	Non-approximable	Ratio 3
$(0,1)-$SP	$\mathcal{NP}$-Complete	Non-approximable	Ratio 2

Completing the model

- Multiplicities on the contigs $m : \mathcal{C} \to \mathbb{N}$. Problem : find an optimal vertex cover (maximal weight) such that each contig edge $e \in \mathcal{C}$ is used exactly $m(e)$ times (or interval)

Completing the model

- Multiplicities on the contigs $m : \mathcal{C} \to \mathbb{N}$. Problem : find an optimal vertex cover (maximal weight) such that each contig edge $e \in \mathcal{C}$ is used exactly $m(e)$ times (or interval)

- Multicriteria Average quality score on the stories $\Rightarrow$ two criteria to optimize.

Completing the model

- Multiplicities on the contigs $m : \mathcal{C} \to \mathbb{N}$. Problem : find an optimal vertex cover (maximal weight) such that each contig edge $e \in \mathcal{C}$ is used exactly $m(e)$ times (or interval)
- Multicriteria Average quality score on the stories $\Rightarrow$ two criteria to optimize.
- Integrating insert sizes especially in the case where multiple sources are used (e.g. PacBio + Illumina reads)

Approximability

- Lower bounds ?

Approximability

- Lower bounds ?
- Other polynomial-time algorithms with a better ratio ?

Approximability

- Lower bounds ?
- Other polynomial-time algorithms with a better ratio ?
- Discuss the ratio according to n vs. $2\sigma_c + \sigma_p$

Approximability

- Lower bounds ?
- Other polynomial-time algorithms with a better ratio ?
- Discuss the ratio according to n vs. $2\sigma_c + \sigma_p$
- Randomized algorithms ?

Biological Validation

- Need to compare to other methods (Implementation and experiments in progress...)

Biological Validation

- Need to compare to other methods (Implementation and experiments in progress...)
- Difficulties to find a good benchmark and good measures

Hunt *et al. Genome Biology* 2014, **15**:R42
http://genomebiology.com/2014/15/3/R42

Genome Biology

RESEARCH Open Access

A comprehensive evaluation of assembly scaffolding tools

Martin Hunt[1*], Chris Newbold[2,1], Matthew Berriman[1] and Thomas D Otto[1]

Institut de biologie computationnelle

Laboratoire Informatique Robotique Microélectronique Montpellier

That's all folks

- Questions ?

That's all folks

- Questions ?
- Answers ?