Original Syntenic Distance between Genomes

A Complete Characterization of Maximal Instances

Cedric Chauve

LaCIM, Université du Québec à Montréal, CANADA

Guillaume Fertin

IRIN, Université de Nantes, FRANCE

Goal : inference of genome rearrangement.

Determining evolutionary distances between species.

When the gene order within chromosomes is known, e.g. :

- reversal distance
- transposition distance

Sometimes, the gene order within chromosomes is not known.

$\Rightarrow$ **Original Synteny** : Model introduced by Ferretti, Nadeau and Sankoff - CPM 1996.

Genome $\mathcal{G}_1 \Rightarrow$: set of chromosomes

Chromosome $\Rightarrow$: set of genes

A genome $\mathcal{G}_1$ can be written as follows

$$\mathcal{G}_1 = \{S_1, S_2 \ldots S_m\}$$

where each S_i is a set of genes taken in $\{g_1, g_2 \ldots g_N\}$.

Crucial Hypothesis : no gene appears <u>more than once</u> in $\mathcal{G}_1$!

$\Rightarrow \mathcal{G}_1$ is a partition of $\{g_1, g_2 \ldots g_N\}$.

Set of genes=$\{a, b, c, d, e, f, g, h\}$

$\mathcal{G}_1$=

$\{a, b, c\}$ (Chromosome 1)

$\{d, e\}$ (Chromosome 2)

$\{f, g, h\}$ (Chromosome 3)

$\mathcal{G}_2$=

$\{a, d, e\}$ (Chromosome 1)

$\{b, g, h\}$ (Chromosome 2)

$\{c, f\}$ (Chromosome 3)

Goal : compare two genomes $\mathcal{G}_1$ and $\mathcal{G}_2$ over the same set of genes $\{g_1, g_2 \ldots g_N\}$.

$\Rightarrow$ compute the minimum number of operations that are necessary to go from $\mathcal{G}_1$ to $\mathcal{G}_2$.

3 operations are allowed : fusion, fission, translocation.

(1) Fusion : two sets of genes A and B merge to give C.

Example : Fusion($\{a, b\}, \{h, i\}$) $\Rightarrow \{a, b, h, i\}$

(2) Fission : a set of genes C is splitted into two sets A and B (fission = “opposite” of fusion).

Example : Fission($\{b, e, f, g, h\}$)$\Rightarrow$ $\{b, f, h\}\{e, g\}$

(3) Translocation : two sets of genes A and B exchange some subsets of their genes.

Example : Translocation($\{a, b, c\}, \{d, f\}$)$\Rightarrow$ $\{a, c\}\{b, d, f\}$

Original Syntenic Distance between 2 genomes $\mathcal{G}_1$ and $\mathcal{G}_2$:

- denoted by $d(\mathcal{G}_1, \mathcal{G}_2)$
- smallest number of fusions, fissions and translocations that are necessary to go from $\mathcal{G}_1$ to $\mathcal{G}_2$.

Only defined for two genomes $\mathcal{G}_1$ and $\mathcal{G}_2$ over the same **support**, that is the same set of genes.

Computing the syntenic distance between 2 genomes is **NP-hard**.

Approximation algorithms : ratio 2

- Ferretti et al. - 1996
- DasGupta et al. - 1998
- Liben-Nowell - 2001

Here, we look at pairs of genomes $\mathcal{G}_1$ and $\mathcal{G}_2$ that are at **maximal distance** (this distance is called **diameter**).

For any pair of genomes

- $\mathcal{G}_1$ over m chromosomes and
- $\mathcal{G}_2$ over n chromosomes

We can modify this instance into an instance

- ($\mathcal{G}'_1$,$\mathcal{G}'_2$)
- such that $\mathcal{G}_2$ is transformed in $\mathcal{G}'_2 = \{1\}, \{2\} \ldots \{n\}$

$\Rightarrow$ The syntenic distance is conserved.

$\mathcal{G}_1 = \{a, b, c\}, \{d, e\}, \{f, g, h\}$

$\mathcal{G}_2 = \{a, d\}, \{b, g, h\}, \{c, f\}, \{e\}$

transformed into

$\mathcal{G}'_1 = \{1, 2, 3\}, \{1, 4\}, \{2, 3\}$

$\mathcal{G}'_2 = \{1\}, \{2\}, \{3\}, \{4\}$

(i) $[n]$ represents the set $\{1, 2 \ldots n\}$.

(ii) (m, n)**-instance** S $(m \leq n)$

- instance **of size** m
- over $[n]$ = set of m subsets $(S_1, S_2 \ldots S_m)$ of $[n]$

(iii) n**-instance** when $m = n$. This instance is then said to be **square**.

[Kleinberg, Liben-Nowell - 2000] : for any n-instance ($n \geq 4$), the diameter $D(n) = 2n - 4$.

- $\mathcal{G}_1$ = n copies of set $[n]$
- $\mathcal{G}_2$ = $\bar{\mathcal{G}}_n$ = $\{1\}, \{2\} \ldots \{n\}$

[Pisanti,Sagot-2002] : for any (m, n)-instance ($m, n \geq 4$), the diameter $D(m, n) = m + n - 4$.

$\Rightarrow$ does not depend on the number of genes N.

Characterizing Instances at Distance = Diameter : Why ?

Approximation algorithm : ratio 2 $\Rightarrow$ asks for a better solution.

Our hope : better understanding of the combinatorial nature of the problem.

Diameter is a natural thing to look at when you look at a distance between objects (cf. [Kleinberg, Liben-Nowell - 2000]).

Besides, this would give a solution to 2 open problems from Kleinberg and Liben-Nowell.

Open Problems in [Kleinberg, Liben-Nowell - 2000] :

1. "Given genomes $\mathcal{G}_1$ and $\mathcal{G}_2$, each with n chromosomes, what is the computational complexity of determining whether $d(\mathcal{G}_1, \mathcal{G}_2) = 2n - 4$?"
2. "It would be interesting to [...] try characterizing those n-chromosome genomes $\mathcal{G}$ for which $d(\mathcal{G}, \bar{\mathcal{G}}_n) = 2n - 4$".

Answers in this talk.

Generalization to (m, n)-instances.

For any instance S, there always exists a sequence that :

- solves it optimally and
- begins with **fusions**, followed by **translocations**, then by **fissions**.

Definition : an instance $\mathcal{G} = \{S_1, S_2 \ldots S_m\}$ **dominates** $\mathcal{G}' = \{S'_1, S'_2 \ldots S'_m\}$ if $S'_i \subseteq S_i$ for all $1 \leq i \leq m$.

Monotonicity : if an instance S dominates an instance S', then $d(S, \bar{\mathcal{G}}_n) \geq d(S', \bar{\mathcal{G}}_n)$.

Incomplete Gossiping : we can simulate a sequence of translocations by a sequence of calls in the framework of the “Incomplete Gossiping” problem.

But what is this “Incomplete Gossiping” problem ?

Problem linked to communications ; appeared in the 50s.

A set of people $\{1, 2 \ldots n\}$, each of which holds a piece of information (info $\{i\}$ for person i).

Communications between them in the **telephone mode** :

- any call involves two persons only
- two people who communicate exchange all they know

Goal : everyone knows everything.

Very wide literature, many variants.

One of the first problems :

- no restriction (anyone can call anyone else)
- **minimum number of calls** so that everyone knows everything ?

Answer : $2n - 4$ for any $n \geq 4$ (to be compared to the n-instances, for which $D(n) = 2n - 4$).

Incomplete Gossiping : two people choose to exchange only a subset of what they know.

Example : $Gos(\{\underline{2}, 4, 5\}, \{\underline{1}, \underline{3}, 7\}) = \{1, 2, 3, 4, 5\}, \{1, 2, 3, 7\}$
$\Rightarrow$ "Opposite" of translocation.

Property : if it possible to solve a gossiping instance with k **calls**, then we can solve the equivalent compact n-instance S with k **translocations**.
$\Rightarrow d(S, \bar{\mathcal{G}}_n) \leq k$.

$\Rightarrow$ **Incomplete Gossiping** will be central in the following.

Any (m, n)-instance S can be seen as a binary matrix M_S (m rows, n columns) :

- $M_S[i, j] = 1$ if the set S_i contains j
- $M_S[i, j] = 0$ otherwise.

Example : $S = \{\{1, 2, 3\}, \{1, 2\}, \{3, 4\}, \{2\}\}$

$$M_S = \begin{matrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{matrix}$$

Two instances S and T are said to be **equivalent** iff M_T can be obtained from M_S by a permutation of its rows and columns.
In that case, $d(S, \bar{\mathcal{G}}_n) = d(T, \bar{\mathcal{G}}_n)$
Example : M_S and M_T below are equivalent.

$$M_S = \begin{matrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{matrix} \qquad M_T = \begin{matrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{matrix}$$

Notion of **dual instance** of S (S of size m over $[n]$) $\Rightarrow$ instance T of size n over $[m]$ (transposition of the matrix).
$\Rightarrow$ Conservation of Syntenic distance.

n

1. Square instances :

   (a) determine necessary conditions.
   Intuitive idea : instances have to be "sufficiently complicated" (matrix full of 1s).

   (b) show that these conditions are sufficient.

2. Generalize the result to (m, n)-instances.

Lemma 1 : Let $n \geq 5$ and $\mathcal{S} = \mathcal{S}_1, \ldots, \mathcal{S}_n$ an n-instance. If there exists $i \in [n]$ such that $|\mathcal{S}_i| \leq n-3$, then $\mathcal{S}$ is not maximal.

$\Rightarrow$ In matrix M_S, there cannot exist a row with three or more 0s.

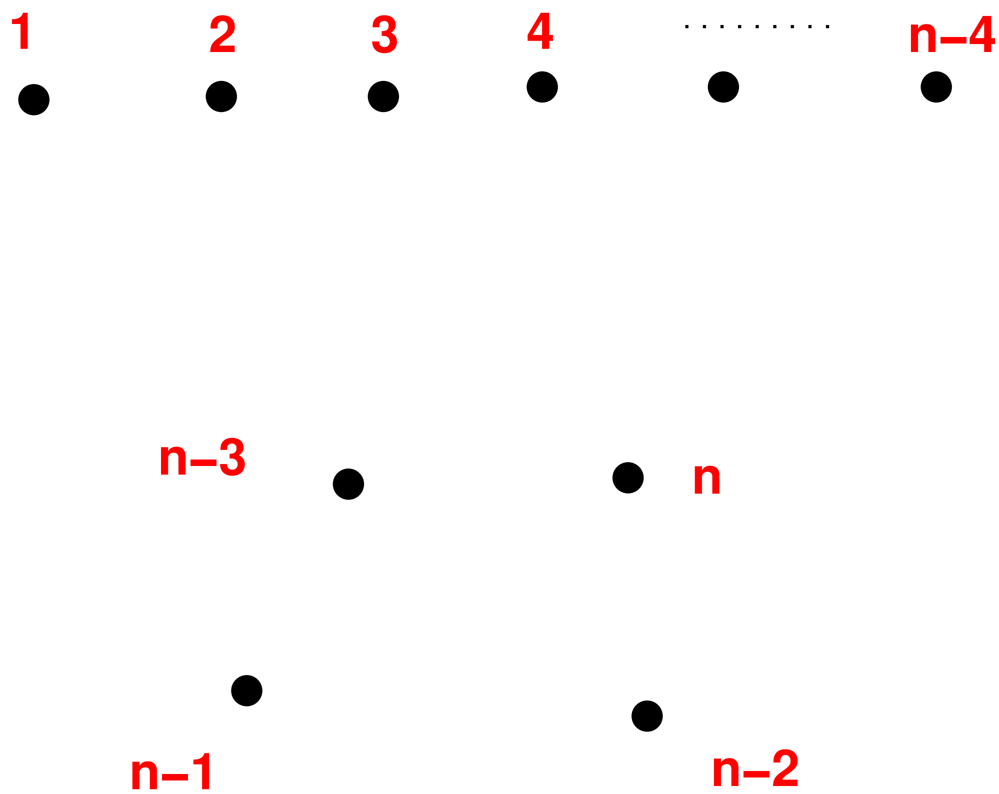
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	0	0	0

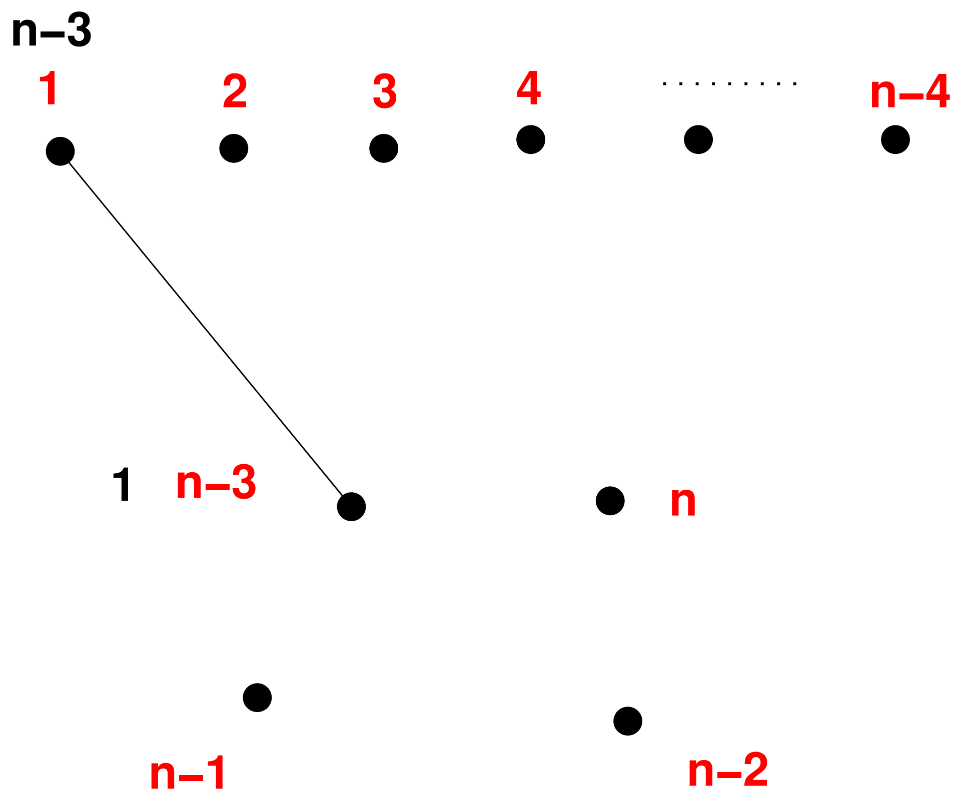
$\Rightarrow$ Proof by equivalence with the (incomplete) gossiping problem.

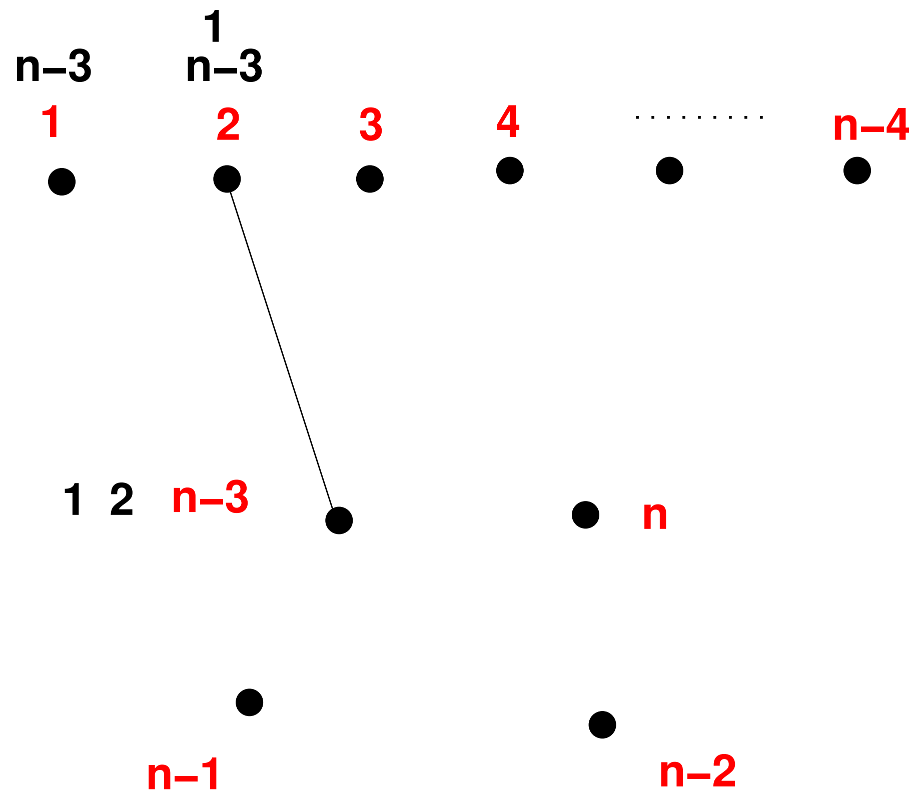
Among the n-instances that satisfy the constrains of Lemma 1, we take the instance S that dominates all the others.

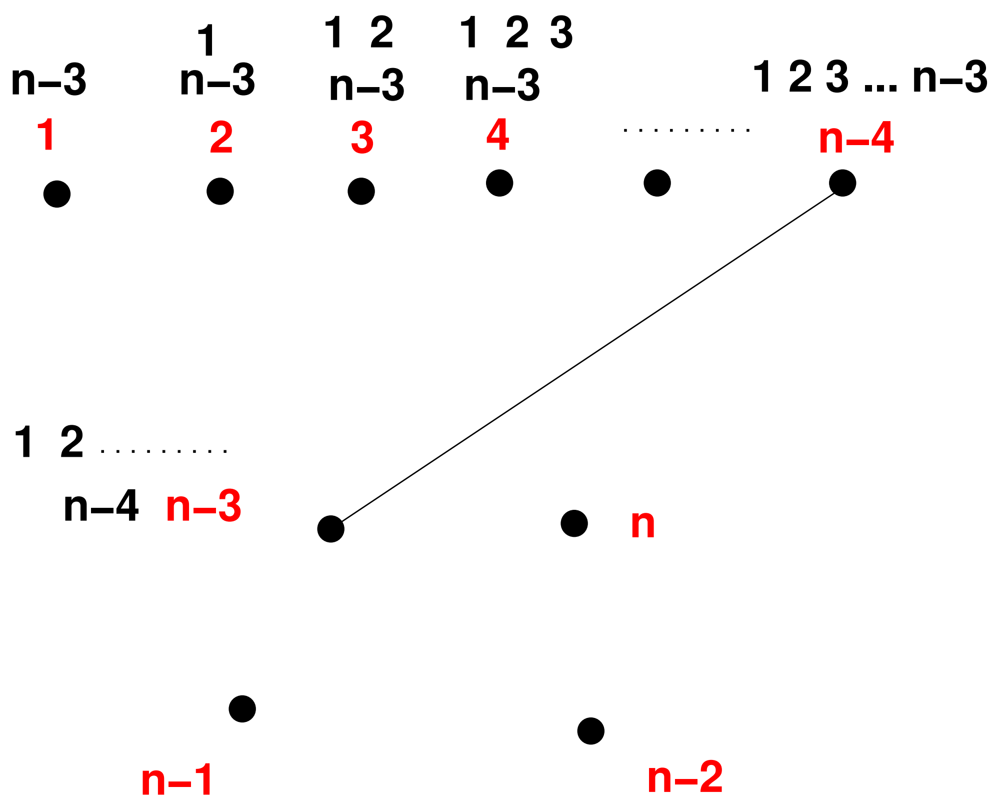
We show that we can solve S with $2n-5$ translocations, since with $2n-5$ calls we can solve the equivalent gossiping problem.

1. For all $1 \leq i \leq n-4$, i calls $n-3$
2. $n-2$ calls $n-3$
3. $n-1$ calls n
4. $n-3$ calls $n-1$ ($\Rightarrow n-3$ and $n-1$ know $[n]$)
5. $n-2$ calls n ($\Rightarrow n-2$ and n know $[n]$)
6. For all $1 \leq j \leq n-5$, j calls n

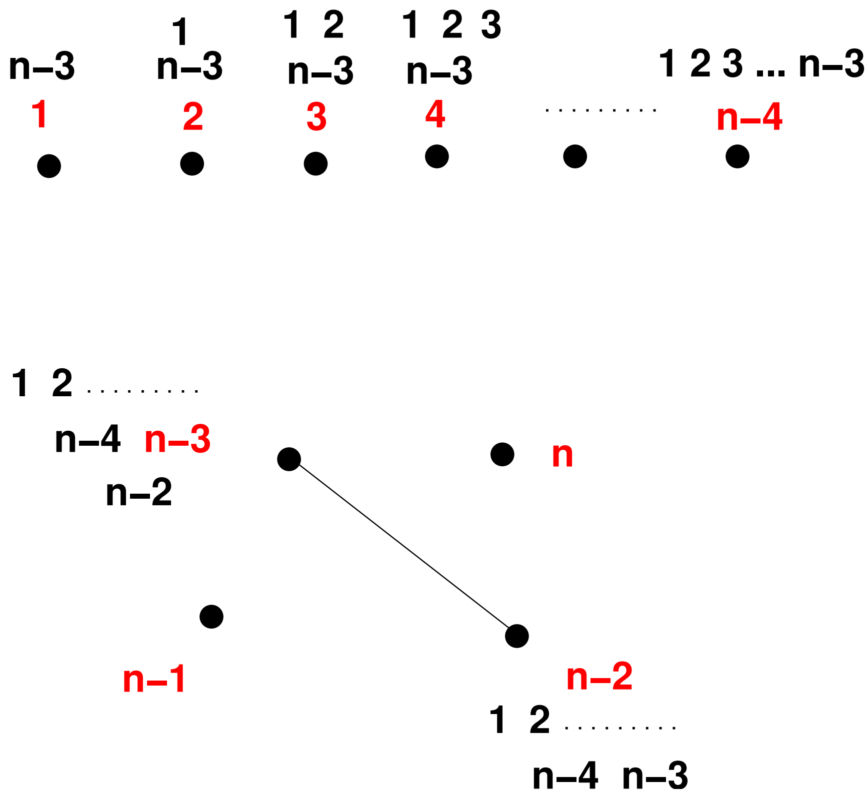

1
2
3
4
. . . .
n−4
n−3
n
n−1
n−2

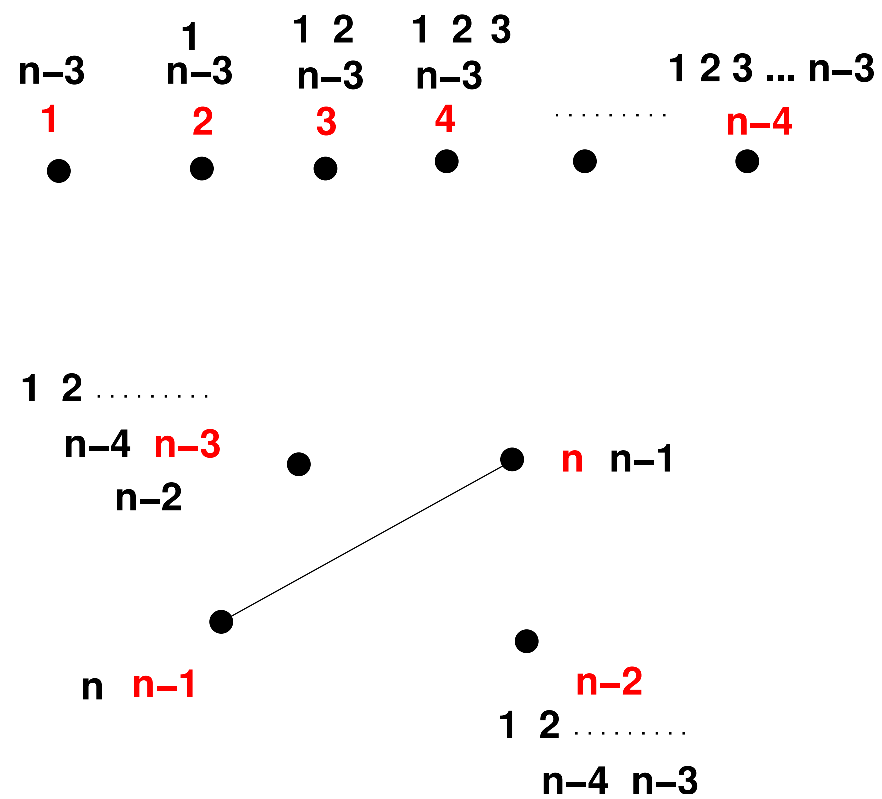

n−3
1
2
3
4
. . . .
n−4
1
n−3
n
n−1
n−2

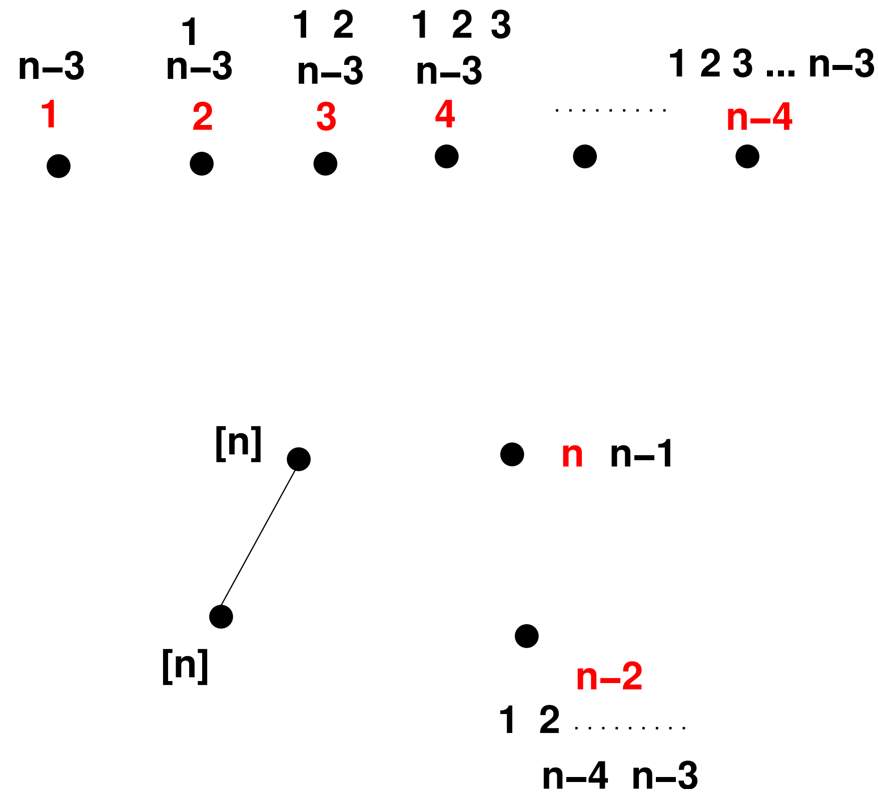

1
n–3
n–3
1
2
3
4
n–4
1 2
n–3
n
n–1
n–2

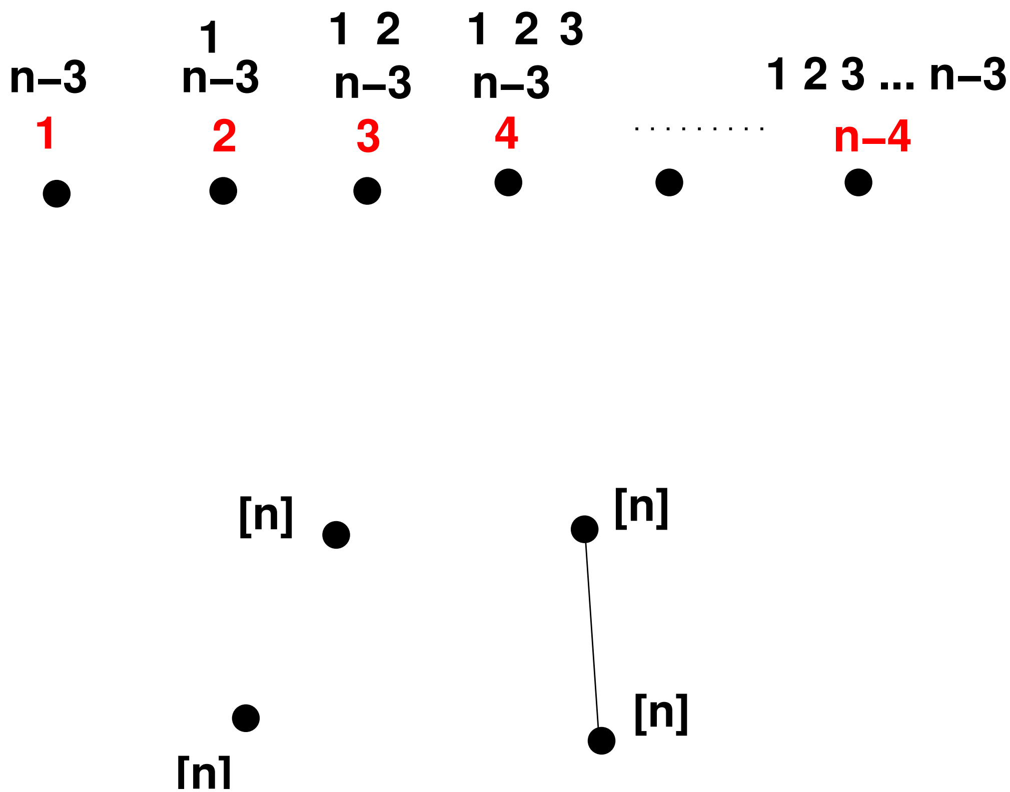


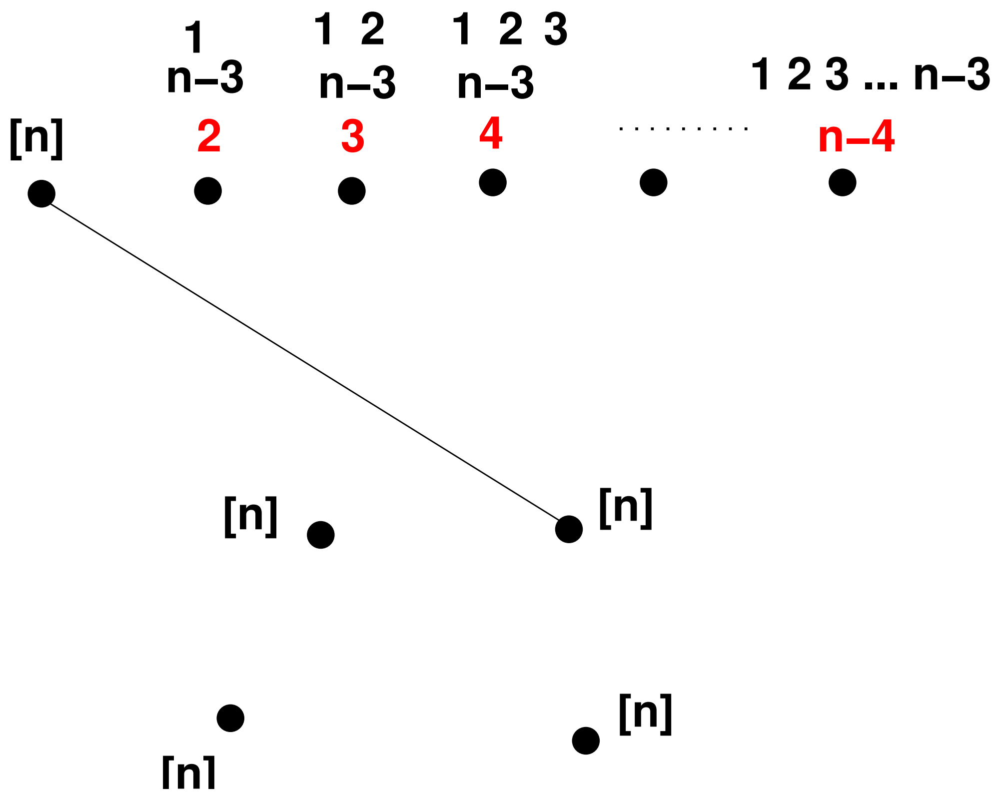
1
1 2
1 2 3
n–3
n–3
n–3
n–3
1 2 3 ... n–3
1
2
3
4
n–4
1 2
n–4 n–3
n
n–1
n–2

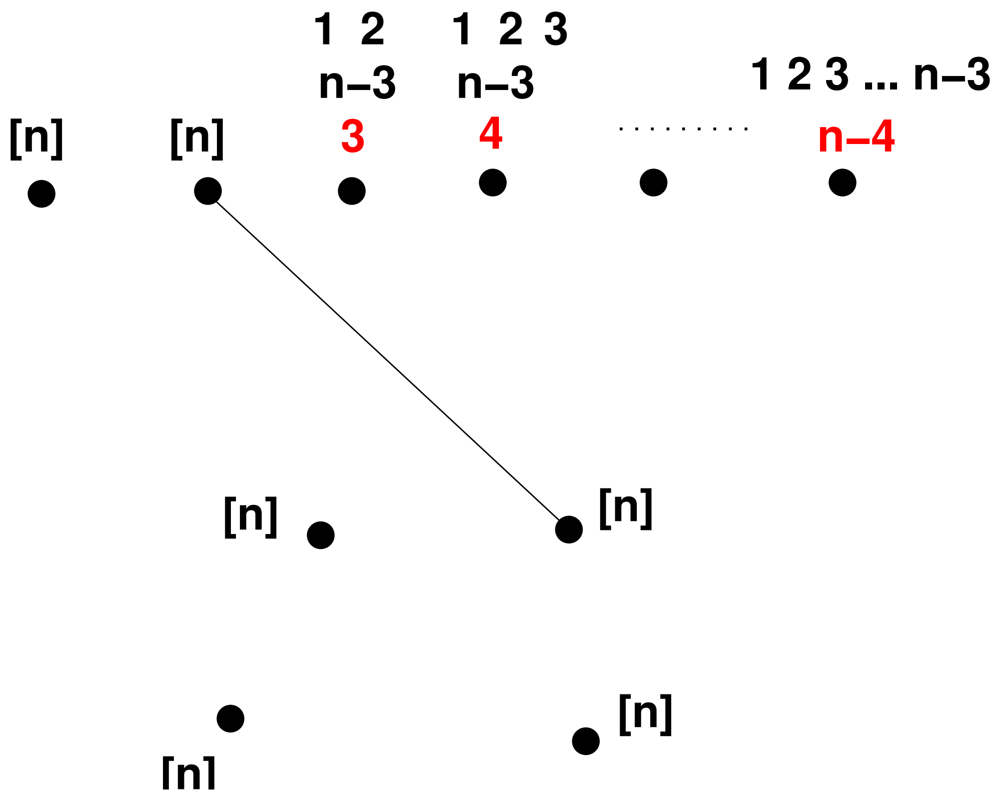

1
n−3
1
n−3
2
1 2
n−3
3
1 2 3
n−3
4
1 2 3 ... n−3
n−4
1 2
n−4 n−3
n−2
n
n−1
n−2
1 2
n−4 n−3

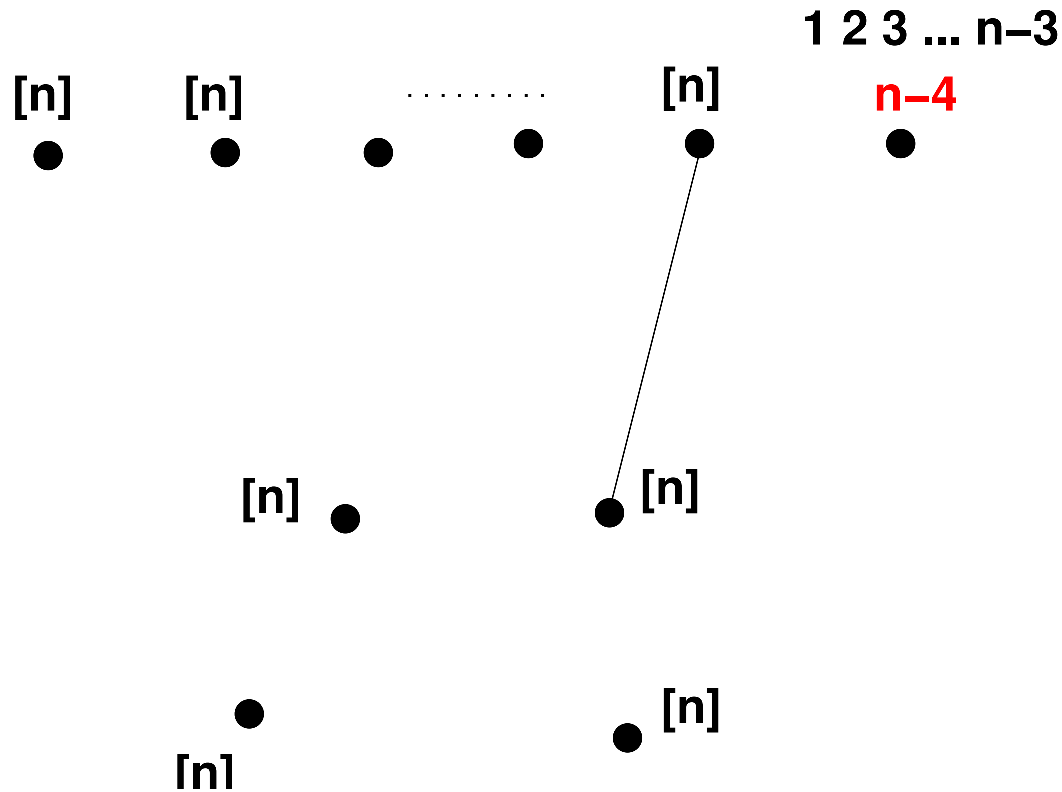

n–3
1
1
n–3
2
1 2
n–3
3
1 2 3
n–3
4
1 2 3 ... n–3
n–4
1 2
n–4 n–3
n–2
n n–1
n n–1
n–2
1 2
n–4 n–3


n−3
1
1
n−3
2
1 2
n−3
3
1 2 3
n−3
4
1 2 3 ... n−3
n−4
[n]
[n]
n n−1
n−2
1 2
n−4 n−3


1
1 2
1 2 3
n–3
n–3
n–3
n–3
1 2 3 ... n–3
1
2
3
4
n–4
[n]
[n]
[n]
[n]


1
n–3
1 2
n–3
1 2 3
n–3
1 2 3 ... n–3
[n]
2
3
4
n–4
[n]
[n]
[n]
[n]


1 2
n–3
1 2 3
n–3
1 2 3 ... n–3
[n]
[n]
3
4
n–4
[n]
[n]
[n]
[n]


1 2 3 ... n–3
n–4
[n]
[n]
[n]
[n]
[n]
[n]
[n]

Consequences :

1. $n-4$ knows $[n]/\{n-2, n-1, n\}$
2. the others know $[n]$
3. $2n-5$ calls overall

$\Rightarrow D(S) \leq 2n-5$.

$\Rightarrow$ any n-instance S' satisfying the constraints of Lemma 1 is such that $D(S') \leq D(S) \leq 2n-5$.

Lemma 2 : For any $n \geq 4$, let $\mathcal{S} = \mathcal{S}_1, \ldots, \mathcal{S}_n$ be an n-instance. If there exists $i, j \in [n]$ and $x, y \in [n]$ ($i \neq j$ and $x \neq y$) such that $x \notin \mathcal{S}_i$, $y \notin \mathcal{S}_i$ and $x \notin \mathcal{S}_j$, then $\mathcal{S}$ is not maximal.

$\Rightarrow$ In matrix M, if a row contains two 0s at indices x and y, there cannot be another 0 in column x.

$$\begin{array}{cccccc}
X & X & X & X & X & X \\
X & X & X & X & X & X \\
X & X & X & X & X & X \\
X & X & X & X & X & X \\
X & X & X & X & X & 0 \\
X & X & X & X & 0 & 0
\end{array}$$

$\Rightarrow$ Proof by equivalence with the (incomplete) gossiping problem.

Lemma 3 : For any $n \geq 5$, let $\mathcal{S} = \mathcal{S}_1, \ldots, \mathcal{S}_n$ an n-instance. If there exists $i, j, k \in [n]$ ($i \neq j \neq k$) and $x \in [n]$ such that $|\mathcal{S}_i| \leq n - 2$, $x \notin \mathcal{S}_j$ and $x \notin \mathcal{S}_k$, then $\mathcal{S}$ is not maximal.

$\Rightarrow$ In matrix M, if a row contains two 0s, then no column can contain two 0s.

$$\begin{matrix} X & X & X & X & X & X \\ X & X & X & X & X & X \\ X & X & X & X & X & X \\ X & X & X & X & X & 0 \\ X & X & X & X & X & 0 \\ X & X & X & 0 & 0 & X \end{matrix}$$

$\Rightarrow$ Proof by equivalence with the (incomplete) gossiping problem.

Lemma 4 : For any $n \geq 6$, let $\mathcal{S} = \mathcal{S}_1, \ldots, \mathcal{S}_n$ an n-instance. If there exists $i, j \in [n]$ ($i \neq j$) such that $|\mathcal{S}_i| \leq n-2$ and $|\mathcal{S}_j| \leq n-2$, then $\mathcal{S}$ is not maximal.

$\Rightarrow$ In matrix M, if a row contains two 0s, no other row can contain two 0s.

X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	0	0	X	X
X	X	X	X	0	0

$\Rightarrow$ Proof by equivalence with the (incomplete) gossiping problem.

Candidate Matrices for maximal n-instances ($n \geq 6$) :

(i) $\mathcal{A}(n)$ or

(ii) $\mathcal{B}(n)$ or

(iii) dual($\mathcal{B}(n)$) or

(iv) any instance that dominates cases (i),(ii) or (iii).

Example with $n = 6$

$$\mathcal{A}(6) = \begin{matrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \end{matrix} \qquad \mathcal{B}(6) = \begin{matrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{matrix}$$

(a) Show that solving $\mathcal{A}(n)$, $\mathcal{B}(n)$ and any instance that dominates them **using translocations only** costs $2n - 4$.

(b) Show that the use of fusions and fissions does not decrease the overall number of mutations needed to solve $\mathcal{A}(n)$ or $\mathcal{B}(n)$.

(a) Solving $\mathcal{A}(n)$ and $\mathcal{B}(n)$ **using translocations only** costs $2n-4$.

Lemma 5 : No instance dominating $\mathcal{A}(n)$ can be solved in less than $2n-4$ translocations.

Proof :

(1) If S = n copies of $[n]$, $D(S)=2n-4 \Rightarrow$ Kleinberg and Liben-Nowell.

(2) In the other cases : equivalence with gossiping.
There exists in S an S_i such that $|S_i|=n-1$.
Suppose $S_i=[n]/j$.

Example : $i = 4$, $j = 6$.

$$\begin{array}{l}
S_1 = 1\ 1\ 1\ 1\ 1\ 1 \\
S_2 = 1\ 1\ 1\ 1\ 1\ 1 \\
S_3 = 1\ 1\ 1\ 1\ 1\ 1 \\
S_4 = 1\ 1\ 1\ 1\ 1\ 0 \\
S_5 = 1\ 1\ 1\ 1\ 1\ 1 \\
S_6 = 1\ 1\ 1\ 1\ 1\ 1
\end{array}$$

In S_i, $\forall\ 1 \leq k \neq j \leq n$, the information k must be known :

- either directly
- or indirectly

$\Rightarrow n - 2$ calls, where j is never involved.

Example : $i = 4$, $j = 6$.

$$\begin{array}{l}
S_1 = \color{green}{1}\ 1\ 1\ 1\ 1\ 1 \\
S_2 = 1\ \color{green}{1}\ 1\ 1\ 1\ 1 \\
S_3 = 1\ 1\ \color{green}{1}\ 1\ 1\ 1 \\
S_4 = 1\ 1\ 1\ \color{green}{1}\ 1\ \color{red}{0} \\
S_5 = 1\ 1\ 1\ 1\ \color{green}{1}\ 1 \\
S_6 = 1\ 1\ 1\ 1\ 1\ \color{green}{1}
\end{array}$$

But j must also be transmitted to the $n-2$ other sets

$\Rightarrow n-2$ calls at least, in which j is always involved.

$\Rightarrow 2n-4$ calls overall.

Lemma 6 : No instance dominating $\mathcal{B}(n)$ can be solved in less than $2n-4$ translocations.

Proof : Similar to Lemma 5 (only more complicated).

(b) Show that the use of fusions and fissions does not decrease the overall number of mutations needed to solve $\mathcal{A}(n)$ or $\mathcal{B}(n)$.

$\Rightarrow$ By contradiction : if there existed at least one fusion, then there should be at least one fission (because the instance is square).

We know we can do fusions first, then translocations and finally fissions.

Between 1st fusion and last fission, we can transform the instance into an $(n-1)$-instance T.

We show that T always dominates either $\mathcal{A}(n-1)$ or $\mathcal{B}(n-1)$.

$\Rightarrow$ If there were $k \geq 1$ fusions (hence $k \geq 1$ fissions), then there would be at least $2(n-k)-4$ translocations.

$\Rightarrow$ Always $k+2(n-k)-4+k=2n-4$ mutations at least.

Theorem : For any $n \geq 6$, $\mathcal{A}(n)$, $\mathcal{B}(n)$, their dual instances and all the instances that dominate them form the set of maximal n-instances.

Corollary 1 : Any maximal n-instance can be solved by translocations only.

Corollary 2 : We can decide in $O(n^2)$ if an n-instance is maximal.

Remark : Small cases ($n \leq 5$) solved as well.

Theorem : For all $n \geq 6$ and $n > m$, $\mathcal{C}(m, n)$, $\mathcal{D}(m, n)$ and all the instances that dominate them form the set of maximal (m, n)-instances.

Example $n = 9, m = 6$.

$$\mathcal{C}(6,9) = \begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{matrix} \qquad \mathcal{D}(6,9) = \begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{matrix}$$

Corollary 3 : We can decide in polynomial time in n and m if an (m, n)-instance is maximal.

Remark : Small cases solved as well.

Answer to two open problems by Kleinberg and Liben-Nowell about the n-instances.

Generalization to (m, n)-instances.

Interesting equivalence between translocation and incomplete gossiping.

- Is it possible to improve the approximation ratio of 2 for the computation of the syntenic distance ?

- Maximal square instances can be solved by translocations only. We know instances for which it is not true.

$\Rightarrow$ Can we characterize those instances for which an optimal solution is given by translocations only ?