

On the Diaconis-Gangolli Markov Chain for Sampling Contingency Tables with Cell-Bounded Entries

Ivona Bezáková

(Rochester Institute of Technology)

Nayantara Bhatnagar

(University of California, Berkeley)

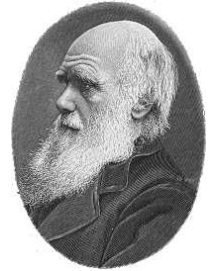
Dana Randall

(Georgia Institute of Technology)

COCOON 2009, July 14, 2009

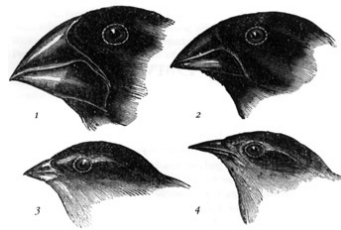
Parsing the title: Contingency Tables

A motivational example: Darwin's finches



1830's:

geological expedition
around the world

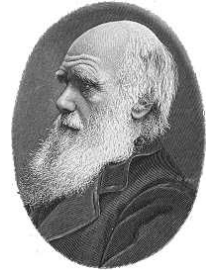




1835:

Galápagos archipelago,
Darwin observes that
individual islands are
inhabited by different
types of finches

Parsing the title: Contingency Tables

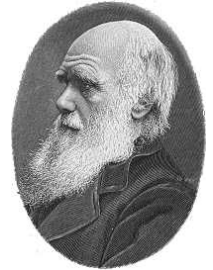
A motivational example: Darwin's finches




 	Santa Cruz	Plaza	Santa Fe	San Cristobal	Espanola	Floreana	Isabela	Ferrandina	Santiago	Rabida	Genovesa	
Small ground finch	1	1	1	1	1	1	1	1	1	1		10
Medium ground finch	1	1	1	1		1	1	1	1	1		9
Large ground finch	1						1	1	1	1	1	6
Cactus ground finch	1	1	1	1		1	1		1	1		8
Large cactus ground finch					1						1	2
Sharp-beaked ground finch								1	1		1	3
Vegetarian finch	1			1		1	1	1	1	1		7
Small tree finch	1		1	1		1	1	1	1	1		8
Medium tree finch						1						1
Large tree finch	1					1	1	1	1	1		6
Woodpecker finch	1			1			1		1			4
Mangrove finch							1	1				2
Warbler finch	1		1	1	1	1	1	1	1	1	1	10
	9	3	5	7	3	8	10	9	10	8	3	

Parsing the title: Contingency Tables

A motivational example: Darwin's finches



	Santa Cruz	Plaza	Santa Fe	San Cristobal	Espanola	Floreana	Isabela	Ferrandina	Santiago	Rabida	Genovesa	
	1	1	1	1	1	1	1	1	1	1		10
Medium ground finch	1	1	1	1		1	1	1	1	1		9
Large ground finch	1						1	1	1	1	1	6
Cactus ground finch	1	1	1	1		1	1		1	1		8
Large cactus ground finch					1						1	2
Sharp-beaked ground finch								1	1		1	3
Vegetarian finch	1			1		1	1	1	1	1		7
Small tree finch	1		1	1		1	1	1	1	1		8
Medium tree finch						1						1
Large tree finch	1					1	1	1	1	1		6
Woodpecker finch	1			1			1		1			4
Mangrove finch							1	1				2
Warbler finch	1		1	1	1	1	1	1	1	1	1	10
	9	3	5	7	3	8	10	9	10	8	3	

chance
OR
competitive pressures
?

Parsing the title: Contingency Tables

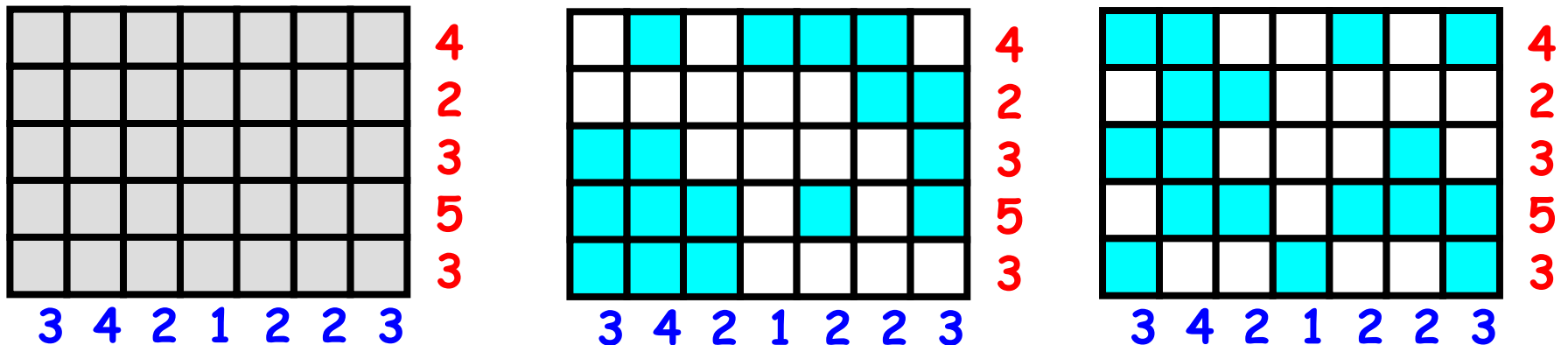
Formal definition:

Input: marginals (row/column sums)

Goal: obtain a uniformly random (binary or not) table that satisfies the given marginals

For example:

- input on the left, two possible binary tables on the right



Parsing the title: Cell-bounded Cont. Tables

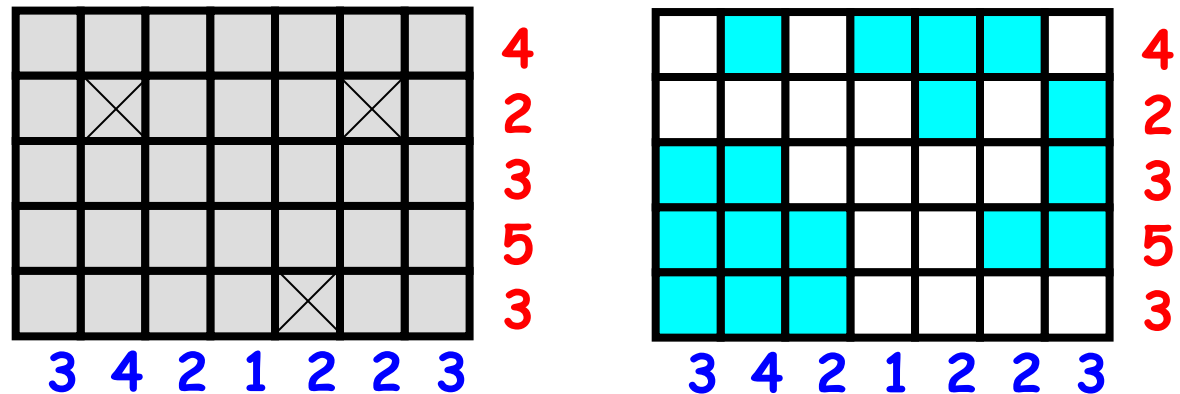
Formal definition, generalized:

Input: marginals (row/column sums) and cell-bounds (an upper-bound for every cell)

Goal: obtain a uniformly random table that satisfies the given marginals and cell-bounds

For example:

- input on the left (cell-bounds are 1, crossed cells 0), a possible table on the right



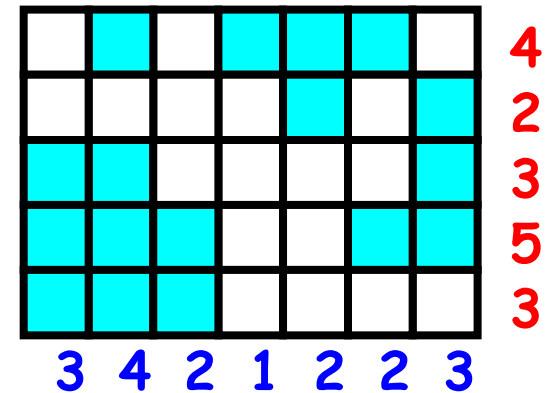
Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

- start at some valid table
- then in each step of the MC:
 - choose a random 2x2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise ignore the changes

Example:

All cell-bounds are 1.



							4
							2
							3
							5
							3
3	4	2	1	2	2	3	

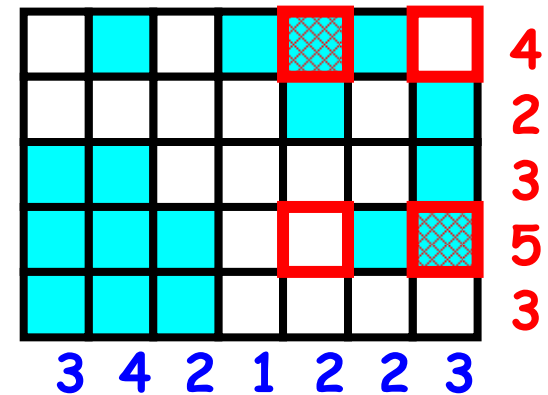
Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

- start at some valid table
- then in each step of the MC:
 - choose a random 2x2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise ignore the changes

Example:

All cell-bounds are 1.



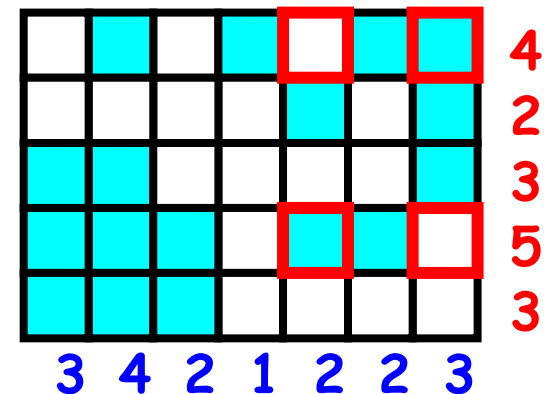
Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

- start at some valid table
- then in each step of the MC:
 - choose a random 2×2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise ignore the changes

Example:

All cell-bounds are 1.



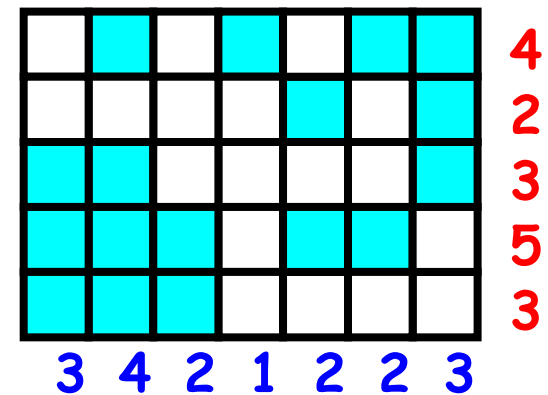
Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

- start at some valid table
- then in each step of the MC:
 - choose a random 2x2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise ignore the changes

Example:

All cell-bounds are 1.



							4
							2
							3
							5
							3
3	4	2	1	2	2	3	

OK

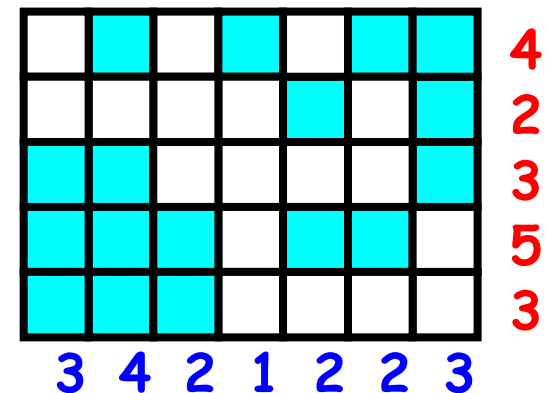
Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

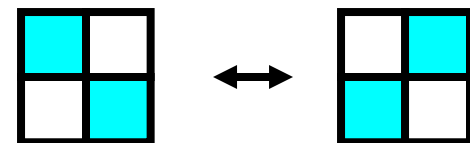
- start at some valid table
- then in each step of the MC:
 - choose a random 2x2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise ignore the changes

Example:

All cell-bounds are 1.



Schematically for cell-bounds 1:



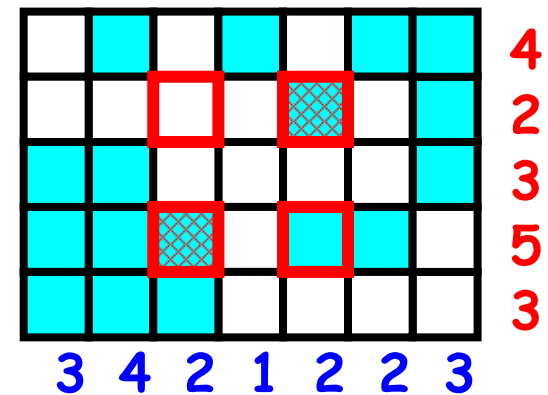
Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

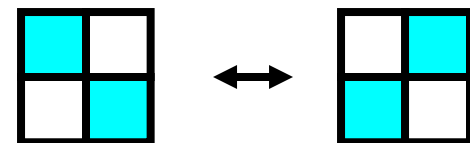
- start at some valid table
- then in each step of the MC:
 - choose a random 2x2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise ignore the changes

Example:

All cell-bounds are 1.



Schematically for cell-bounds 1:



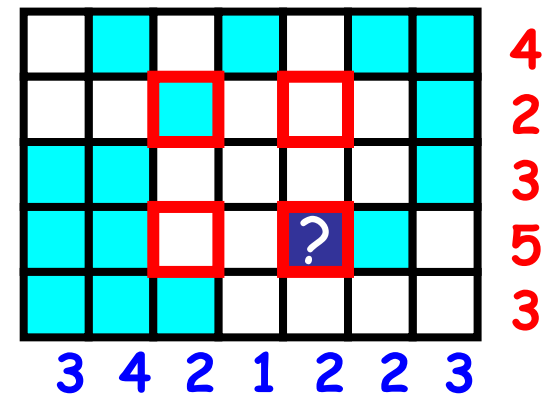
Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

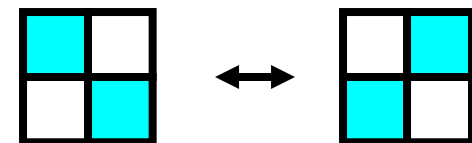
- start at some valid table
- then in each step of the MC:
 - choose a random 2x2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise ignore the changes

Example:

All cell-bounds are 1.



Schematically for cell-bounds 1:



Parsing the title: Diaconis-Gangolli MC

Diaconis-Gangolli's Markov chain ['94]:

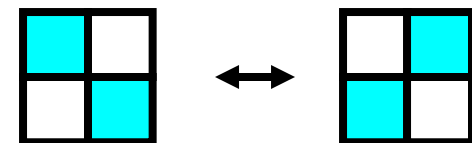
- start at some valid table
- then in each step of the MC:
 - choose a random 2x2 subtable and one of its (two) diagonals
 - try to decrease both entries by 1 and increase the other entries by 1:
if within cell-bounds, keep the changes, otherwise **ignore the changes**

Example:

All cell-bounds are 1.

							4
							2
							3
							5
							3
3	4	2	1	2	2	3	

Schematically for cell-bounds 1:



Parsing the title: Diagonis-Gangolli MC

Remarks:

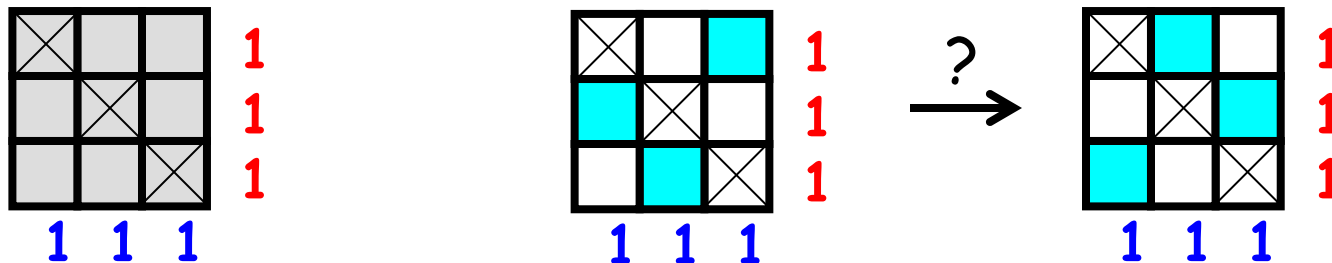
- the original chain was proposed for contingency tables (without cell-bounds)
- without cell-bounds, or with all cell-bounds 1:
the chain connects the state space (i.e., can move from every table to every other table) and the stationary distribution is uniform
- OPEN: does the chain mix rapidly? (i.e., does polynomial number of steps suffice to get a random table?)

Parsing the title: Diagonis-Gangolli MC

OPEN QUESTION: does the chain mix rapidly ?

With cell-bounds:

- the state space might not be connected, e.g.:



- if the state space is connected, does the chain mix rapidly?

This work: the answer is NO.

Remark: why cell-bounds ? One of the reasons: self-reducibility.

Related Work

Polynomial-time algorithms for:

- no cell-bounds, large marginals

[Dyer-Kannan-Mount '97, Morris '02]

- no cell-bounds, constant number of rows

[Cryan-Dyer '03, Cryan-Dyer-Goldberg-Jerrum-Martin '02, Dyer '03]

- cell-bounds, large marginals or constant number of rows

[Cryan-Dyer-Randall '05]

- all cell-bounds 1, (near-)regular marginals: rapid mixing of the Diaconis-Gangolli chain

[Kannan-Tetali-Vempala '99]

- all cell-bounds 1

[reduction from permanent: Jerrum-Sinclair-Vigoda '04, directly: Bezáková-Bhatnagar-Vigoda '07]

Related Work

Polynomial-time algorithms for:

- no cell-bounds, large marginals
[Dyer-Kannan-Mount '97, Morris '02]

- no cell-bounds, constant number of rows
[Cryan-Dyer '03, Cryan-Dyer-Goldberg-Jerrum-Martin '02, Dyer '03]

- cell-bounds, large marginals or constant number of rows
[Cryan-Dyer-Randall '05]

- all cell-bounds 1, (near-)regular marginals: rapid mixing of the Diaconis-Gangolli chain
[Kannan-Tetali-Vempala '99]

- all cell-bounds 1
[reduction from permanent: Jerrum-Sinclair-Vigoda '04,
directly: Bezáková-Bhatnagar-Vigoda '07]

This work:

cell-bounds,
slow mixing

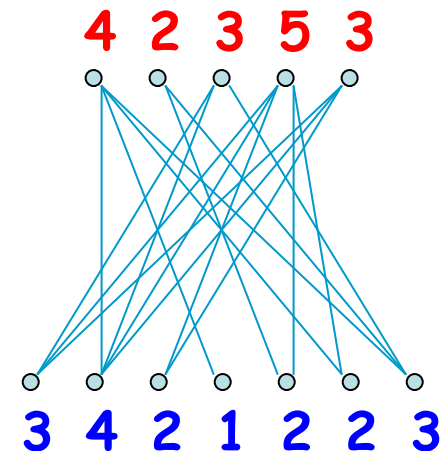
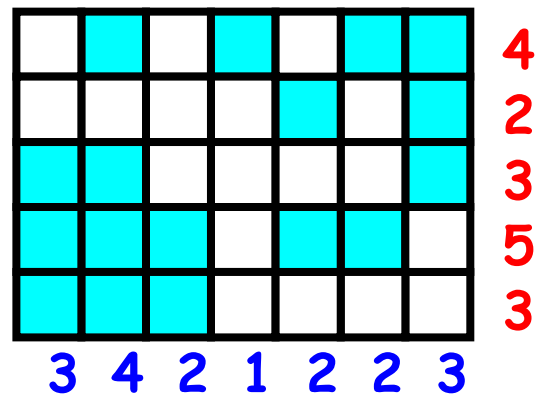
Graphs with Given Degree Sequence

Binary contingency tables = a bipartite graph problem:

Input: degrees of all vertices

Goal: sample all graphs satisfying this degree sequence

Example:



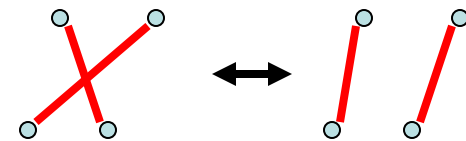
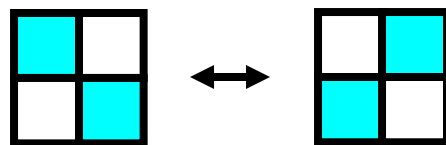
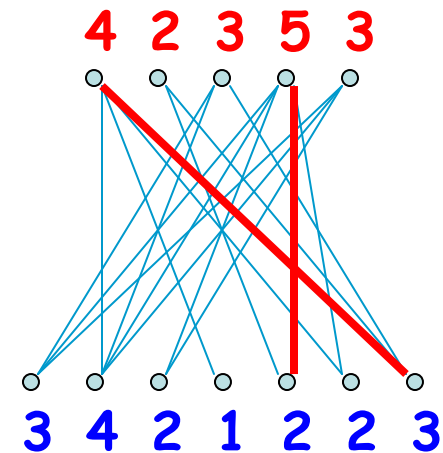
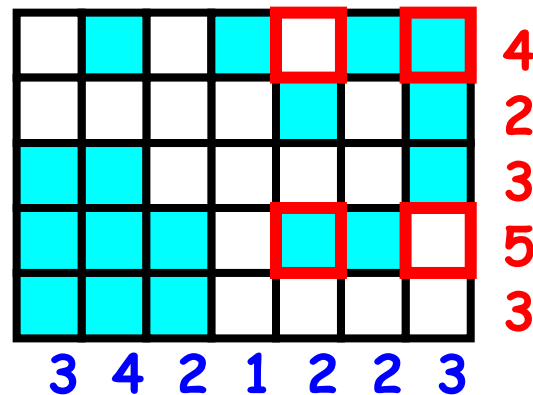
Note: cell-bounds = forbidden edges

Graphs with Given Degree Sequence

The Diaconis-Gangolli move on graphs:

- choose two edges on random, re-match the end-points (if possible, i.e., if the new edges-to-be do not already exist)

Example:



Negative Result for Dense Graphs

Thm: There exists a family of inputs where

- a) the chain connects the state space,
- b) the number of allowed edges incident to every vertex is at least $n/4$,
- c) the mixing time is at least exponential.

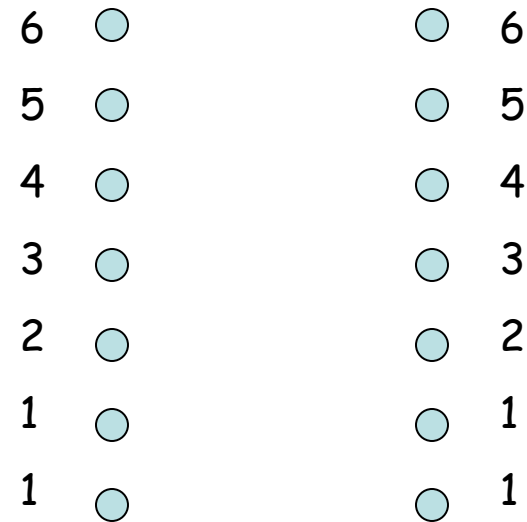
Why are dense inputs interesting ?

- non-cell-bounded tables are dense
- other sampling problems behave well for dense inputs

Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea: consider this input (no forbidden edges)

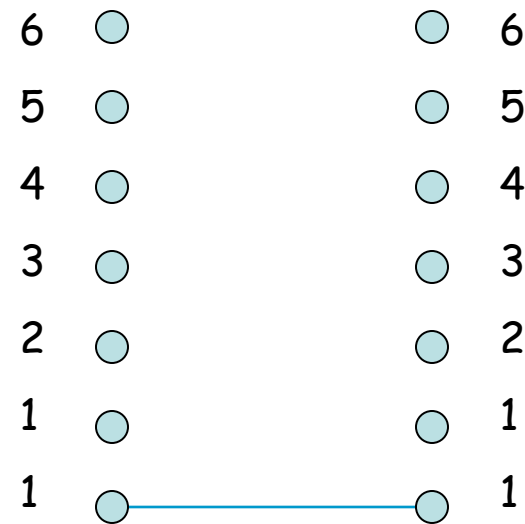


Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea: consider this input (no forbidden edges)

How many tables with 1's connected?



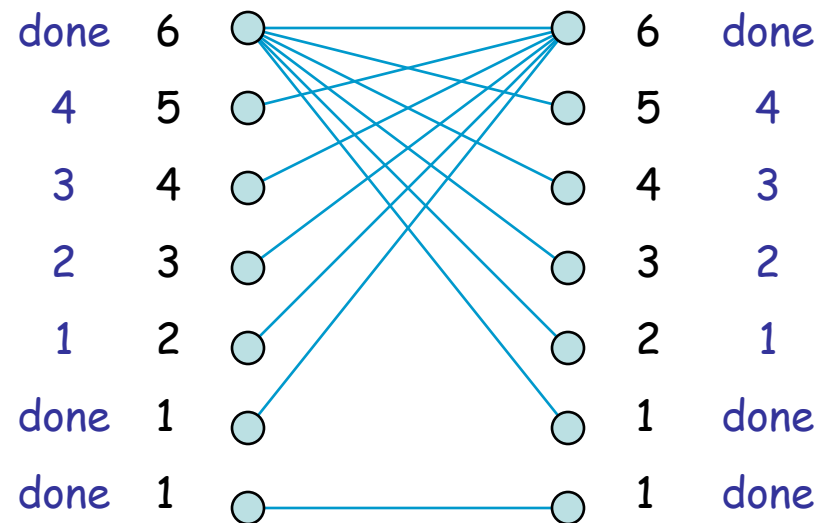
Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea: consider this input (no forbidden edges)

How many tables with 1's connected?

1



Negative Result for Dense Graphs

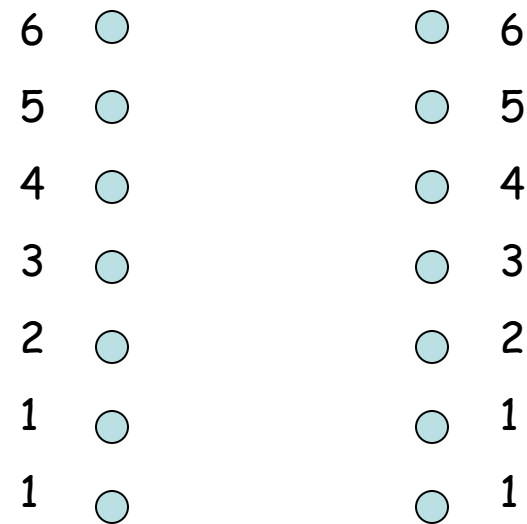
Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea: consider this input (no forbidden edges)

How many tables with 1's connected?

1

How many tables overall?



Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

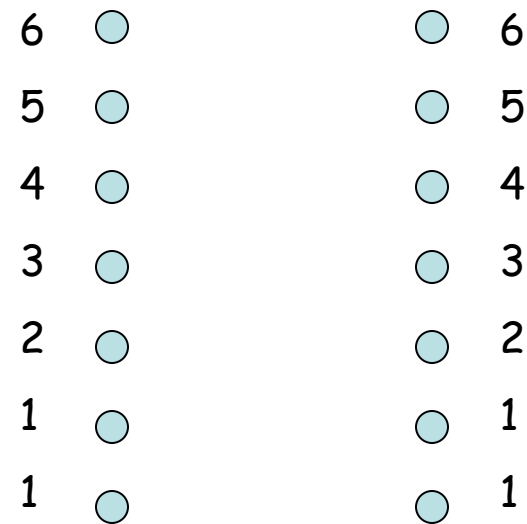
Proof idea: consider this input (no forbidden edges)

How many tables with 1's connected?

1

How many tables overall?

exponential

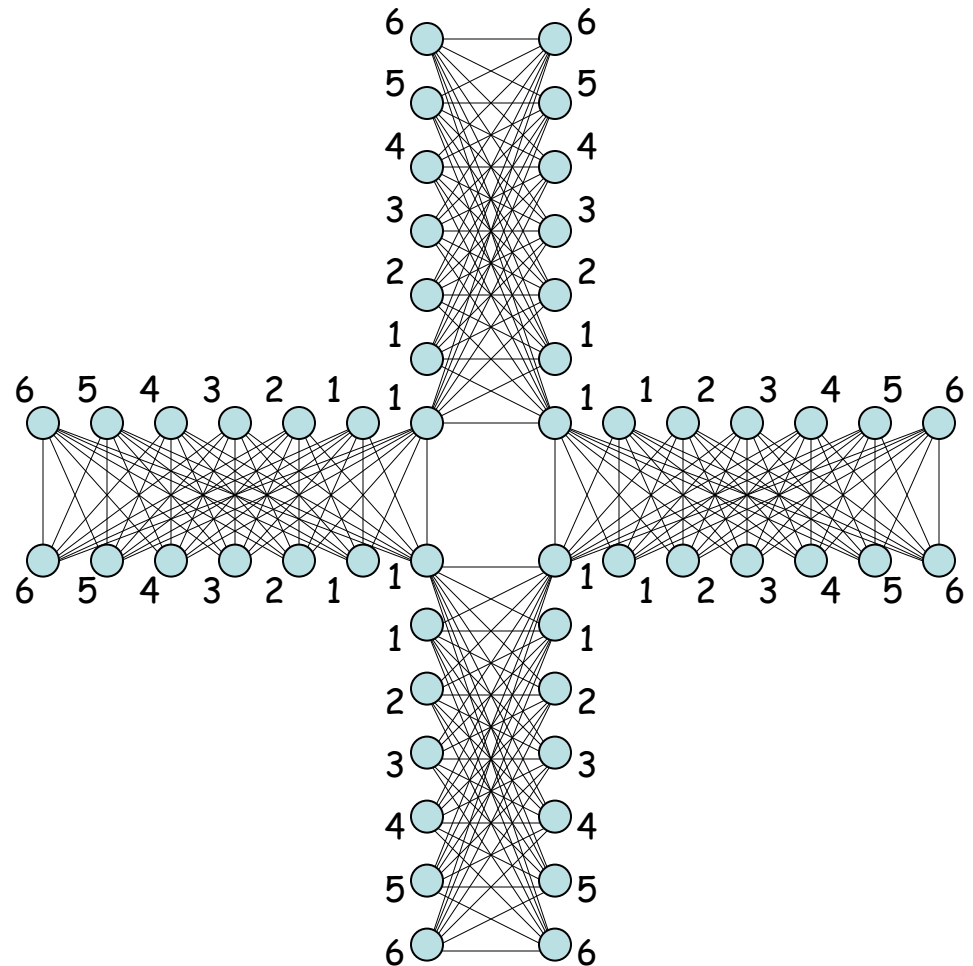


Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea:

Let's connect four copies, the shown edges are the only allowed edges (i.e., the cell-bound for the corresponding cells is 1):

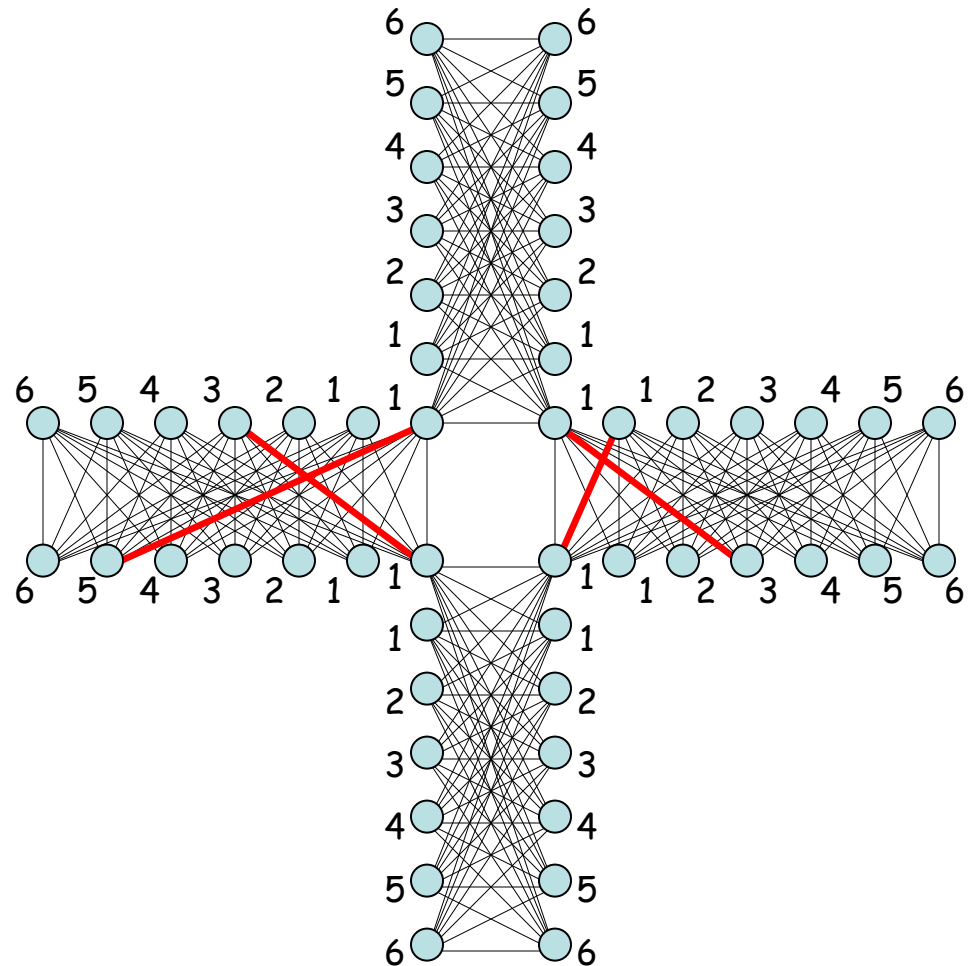


Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea:

Consider a graph with the required degree sequence, suppose the middle vertices connected as shown:

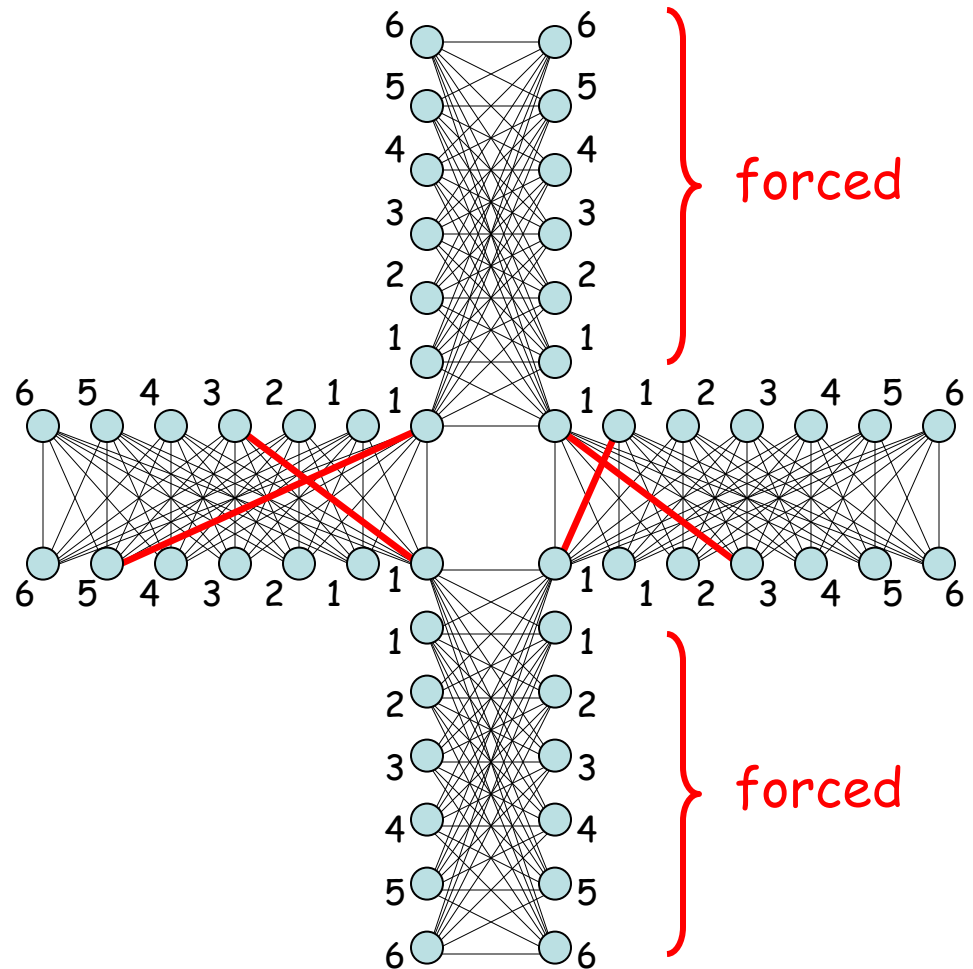


Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea:

Consider a graph with the required degree sequence, suppose the middle vertices connected as shown:



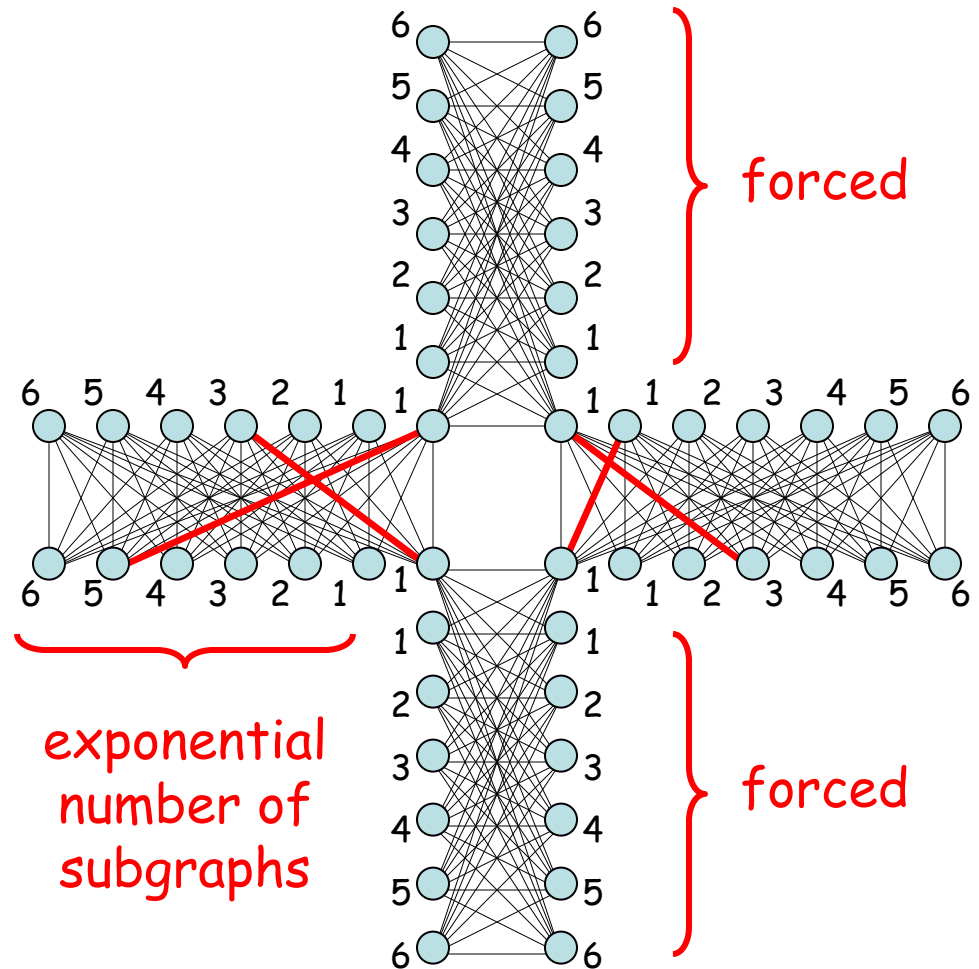
Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea:

Consider a graph with the required degree sequence, suppose the middle vertices connected as shown:

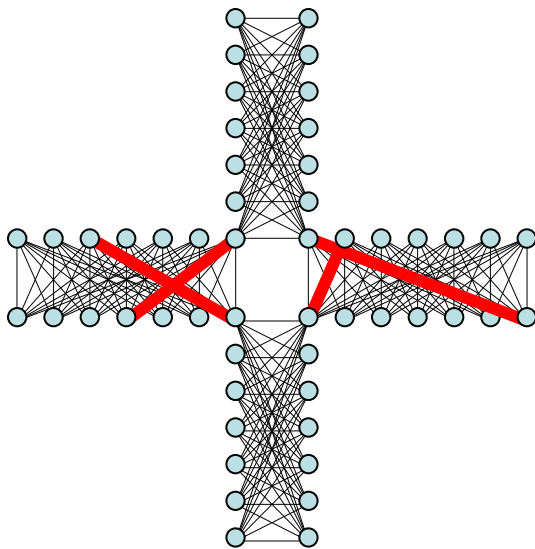
Conclusion: exponential number of "horizontal" possibilities.



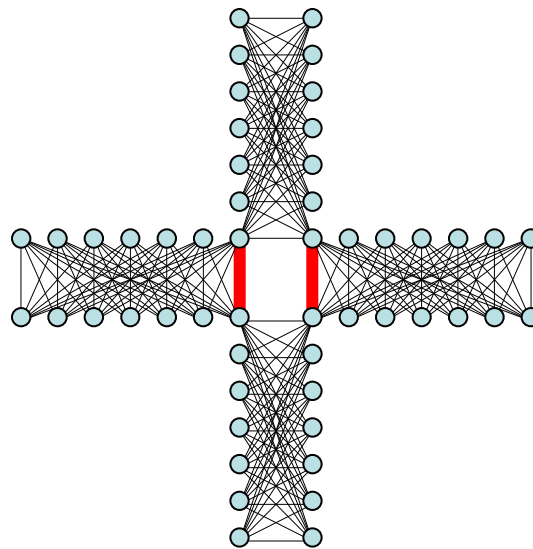
Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

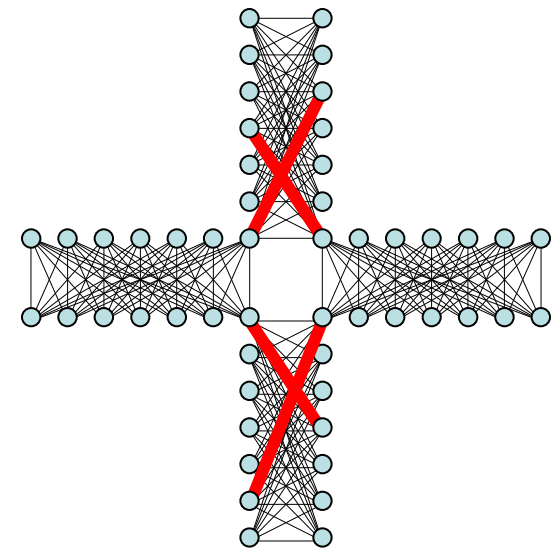
Proof idea:



exponential



one

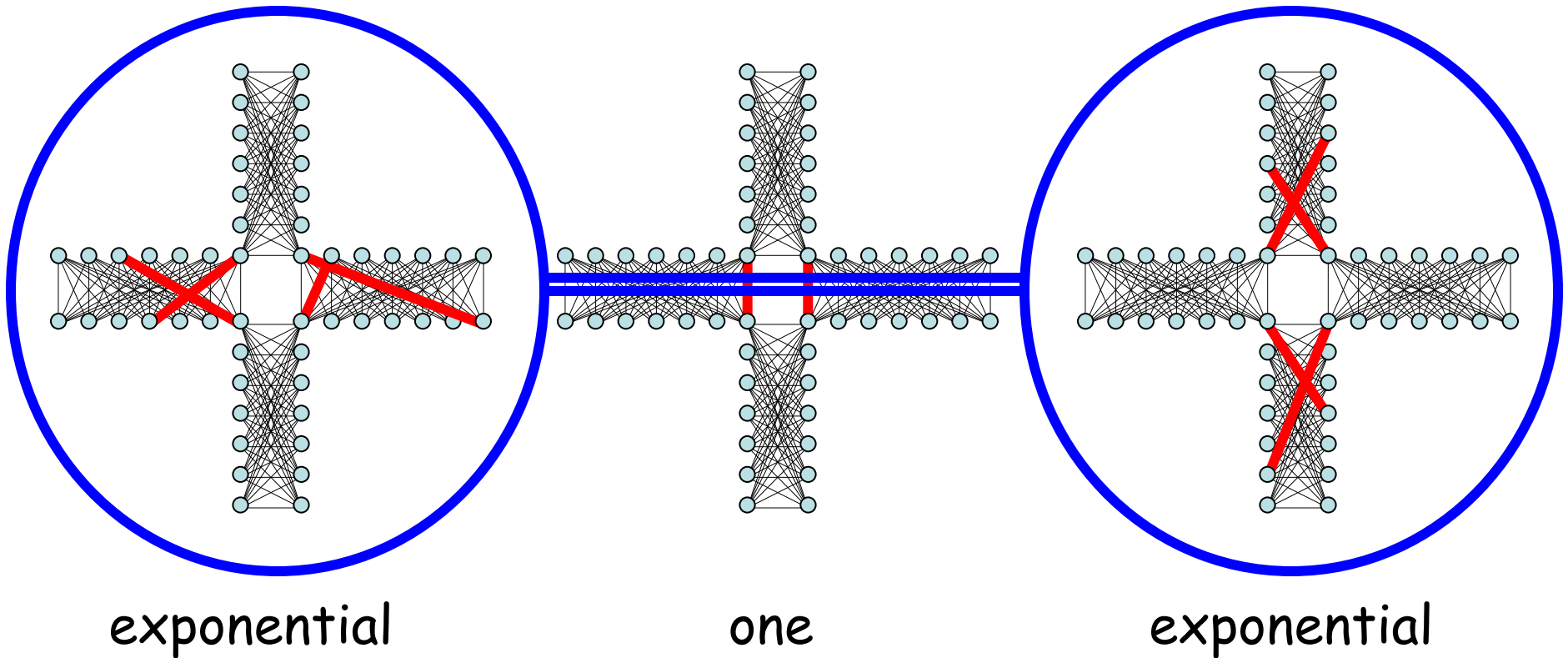


exponential

Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

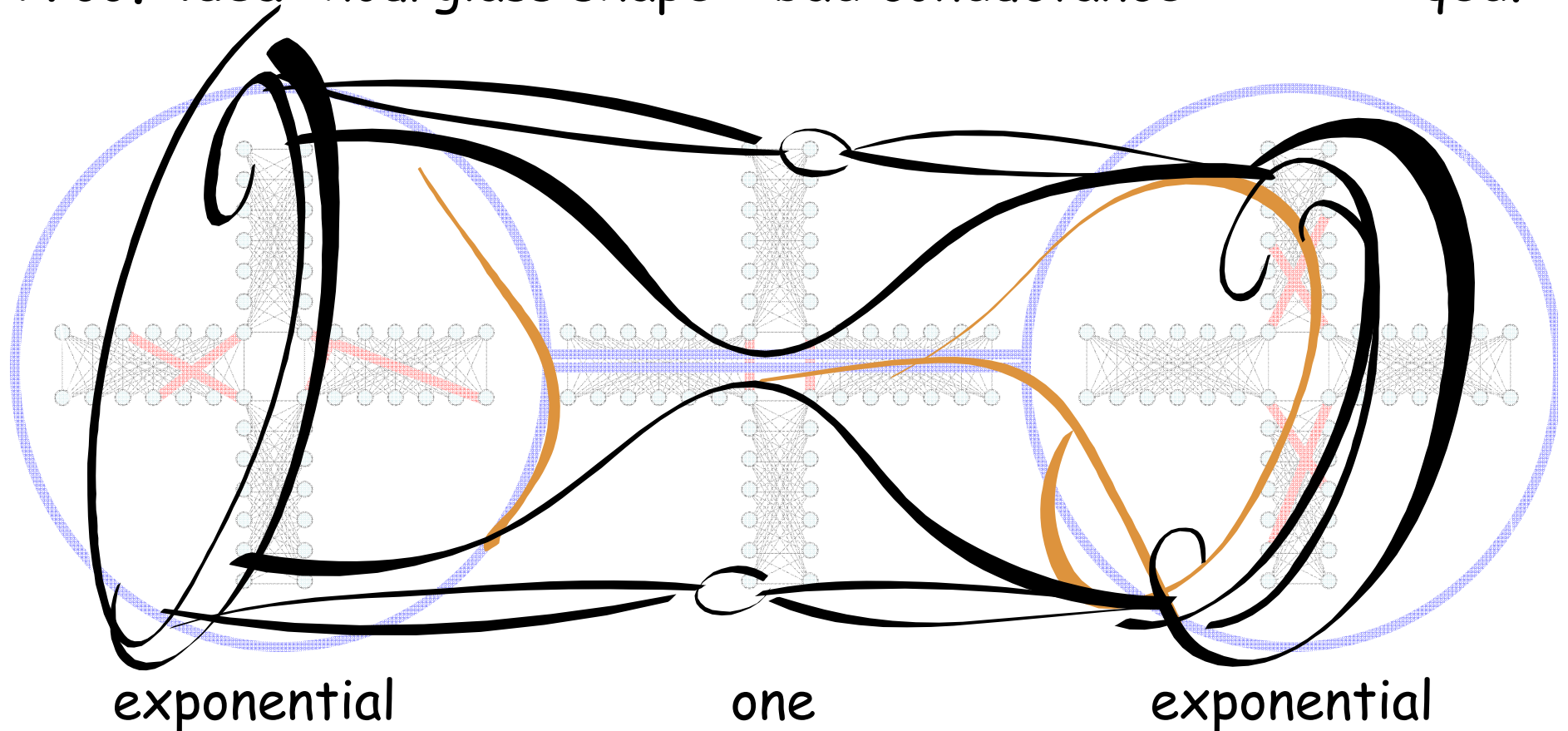
Proof idea:



Negative Result for Dense Graphs

Thm: Exponential mixing time for instances with $n/4$ -dense allowed edges.

Proof idea: hourglass shape = bad conductance qed.



Negative Result for Regular Inputs

Thm: There exists a family of inputs where

- a) the chain connects the state space,
- b) all required degrees (i.e., the marginals) are equal,
- c) the mixing time is at least exponential.

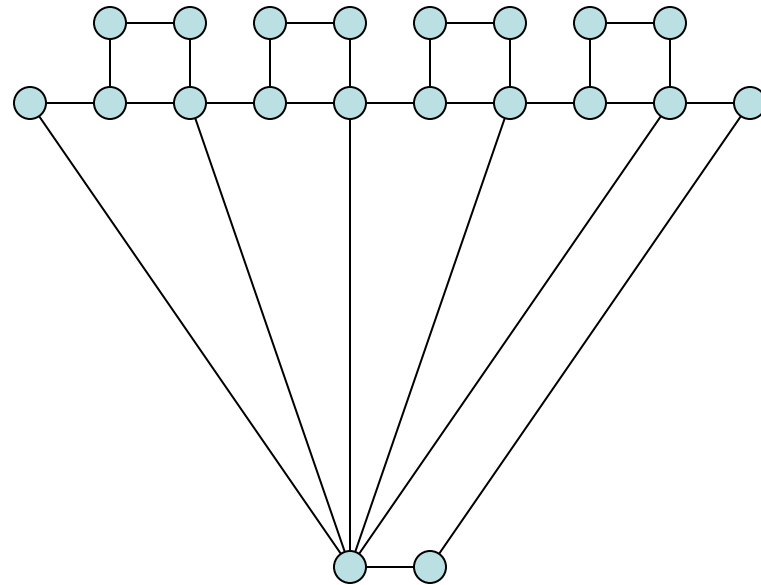
Why are regular inputs interesting ?

- if all cell-bounds are 1, the chain is known to mix rapidly for regular inputs [Kannan-Tetali-Vempala '99]

Negative Result for Regular Inputs

Thm: Exponential mixing time for regular instances.

Proof idea: consider this input (all required degrees are 1)

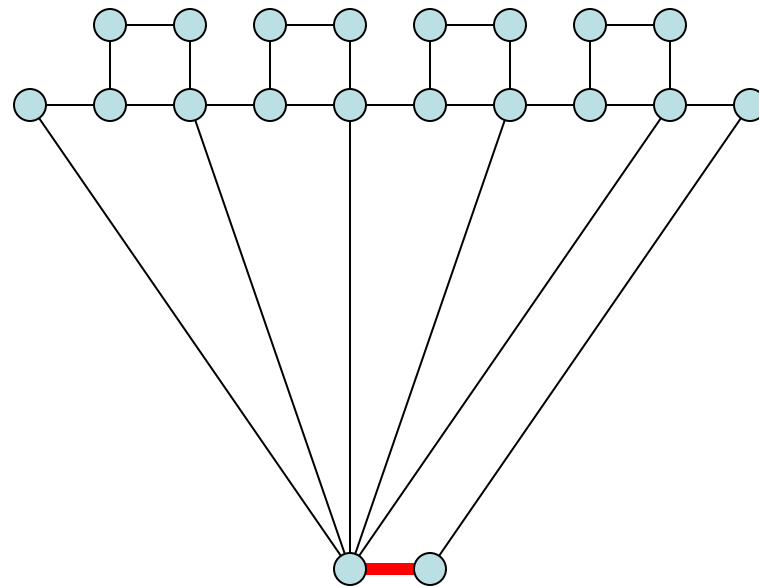


Negative Result for Regular Inputs

Thm: Exponential mixing time for regular instances.

Proof idea: consider this input (all required degrees are 1)

How many graphs contain
the bottom edge ?



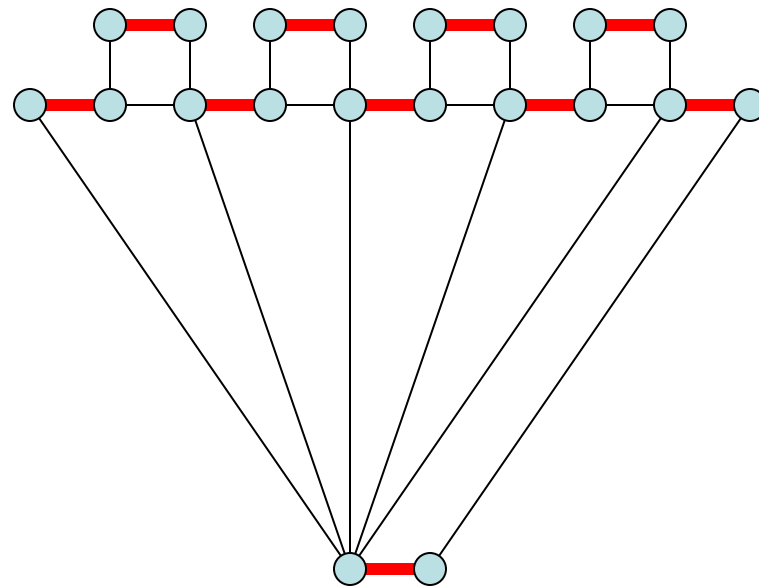
Negative Result for Regular Inputs

Thm: Exponential mixing time for regular instances.

Proof idea: consider this input (all required degrees are 1)

How many graphs contain the bottom edge?

1



Negative Result for Regular Inputs

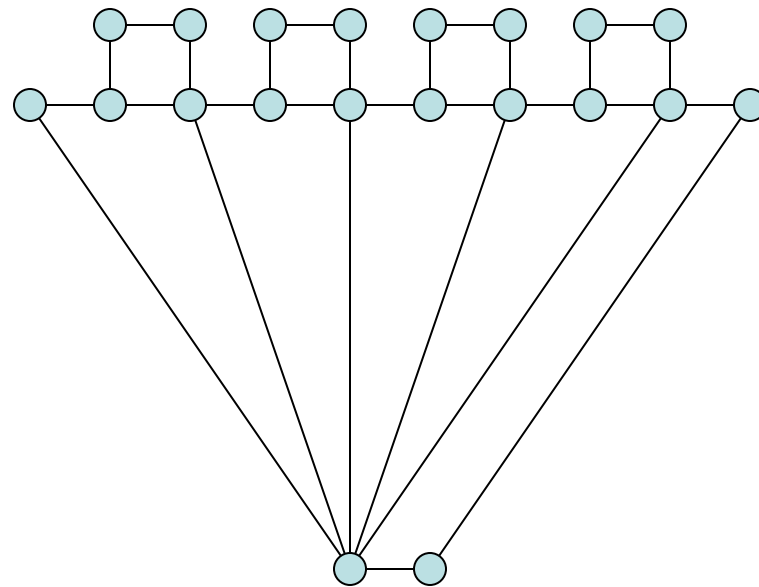
Thm: Exponential mixing time for regular instances.

Proof idea: consider this input (all required degrees are 1)

How many graphs contain
the bottom edge ?

1

How many graphs overall ?



Negative Result for Regular Inputs

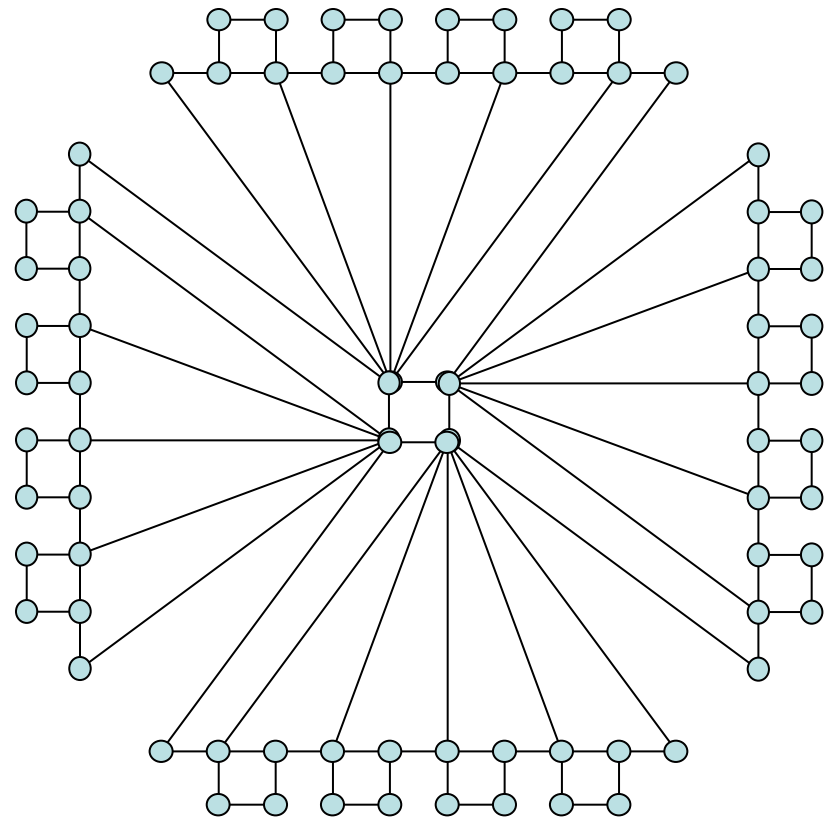
Thm: Exponential mixing time for regular instances.

Proof idea:

Let's connect four copies:

We have to verify:

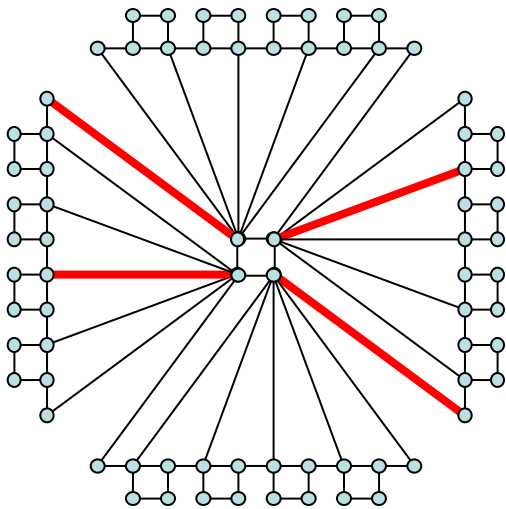
- connected state space
- exponentially small conductance



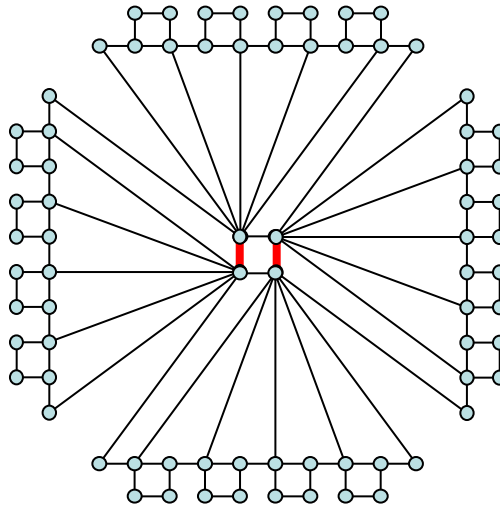
Negative Result for Regular Inputs

Thm: Exponential mixing time for regular instances.

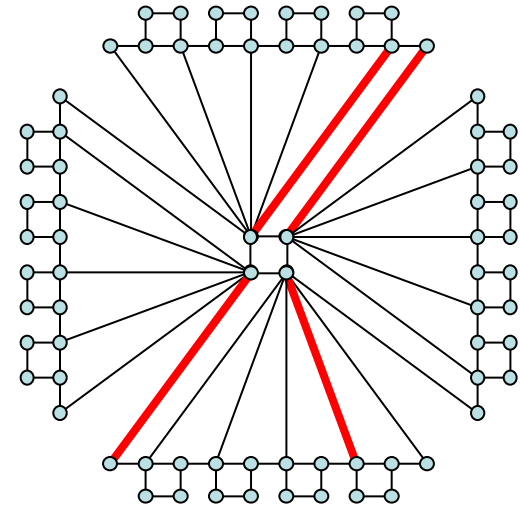
Proof idea: schematic picture of the conductance bound:



exponential



one



exponential

Open Problems

- Diaconis-Gangolli MC without cell-bounds
- Diaconis-Gangolli MC with all cell-bounds 1 (non-regular inputs)
- Other approaches to sampling contingency tables without cell-bounds
- Dense regular instances ?
- All cell-bounds nonzero ?
- The presented instances are "just-connected," are there slowly mixing instances that would be more connected ?

