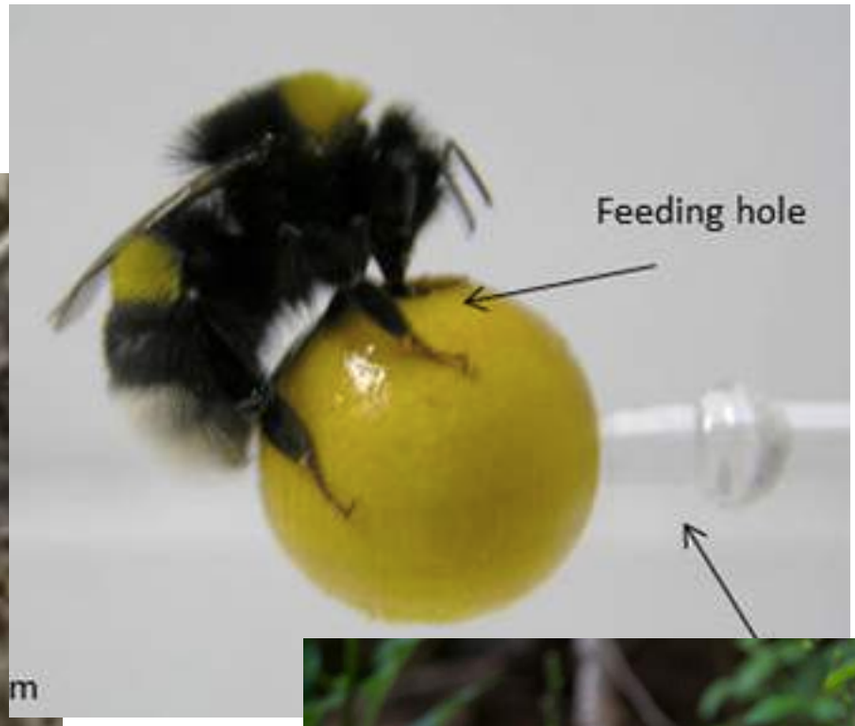
COLLÈGE DE FRANCE
1530

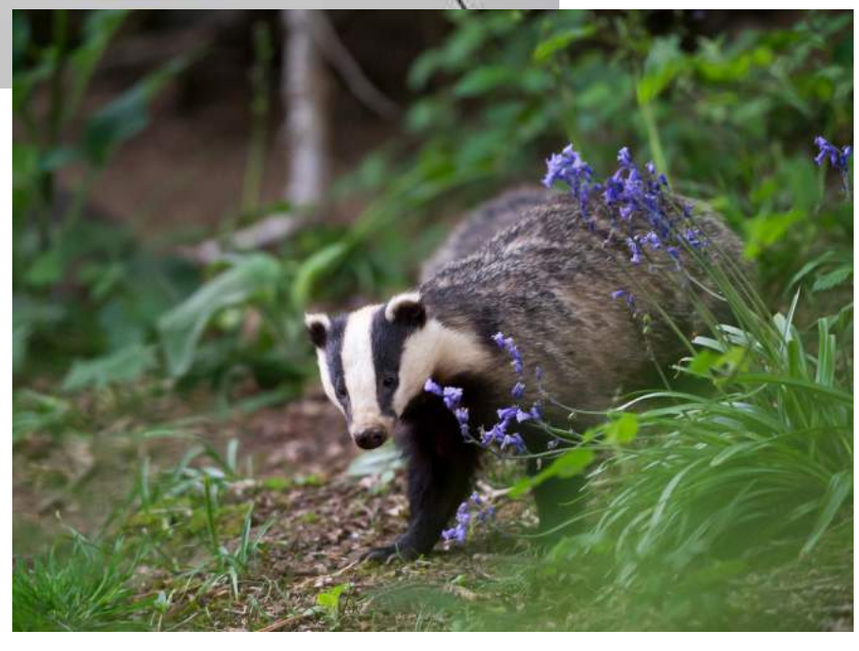
Algorithmes par marches aléatoires

Claire Mathieu

Une marche aléatoire est un algorithme de recherche dans un graphe

Dans la nature : recherche de nourriture





Peut-on en tirer des algorithmes efficaces ?

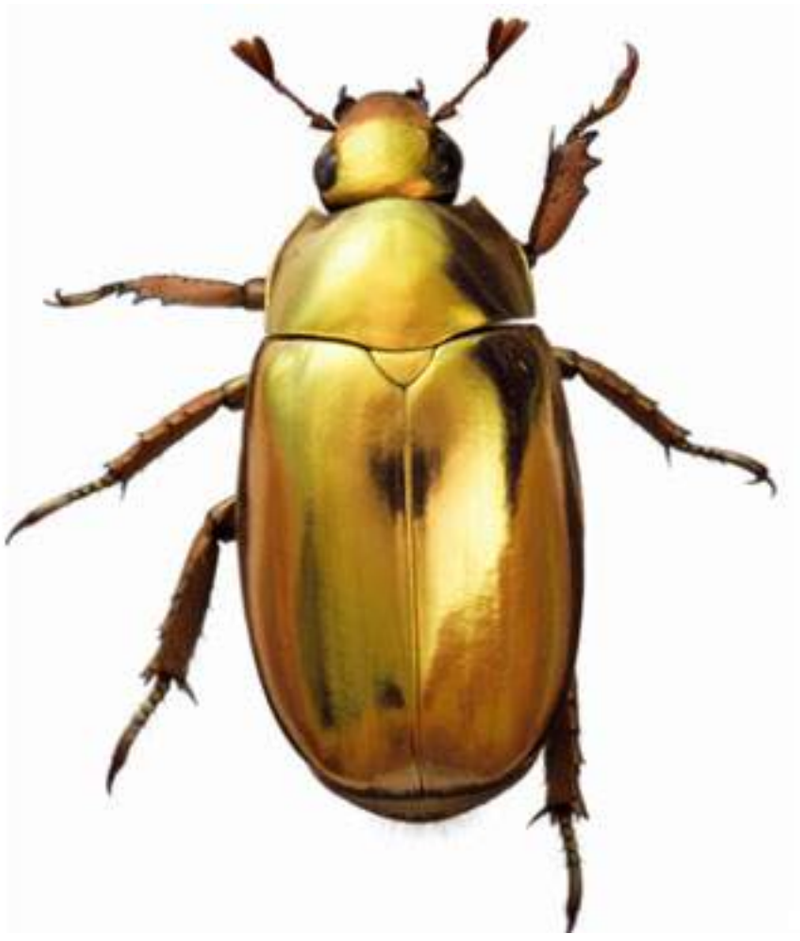
Décrypter un message

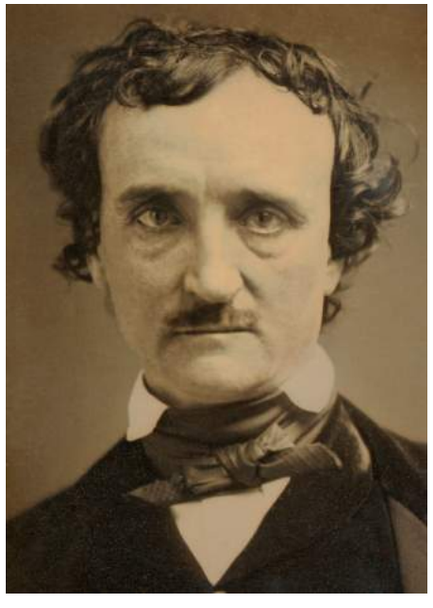
Colorier un graphe

Estimer la taille d’un réseau social

Décrypter un message

Une fiction : le scarabée d'or





Texte chiffré

53‡‡†305))6*;4826)4‡.)4‡);806*;48‡8
¶60))85;1‡(;:‡*8†83(88)5*†;46(;88*96
?;8)‡(;485);5*†2:*‡(;4956*2(5*—4)8
¶8*;4069285);)6†8)4‡‡;1(‡9;48081;8:8‡
1;48†85;4)485†528806*81(‡9;48;(88;4
(‡?34;48)4‡;161;:188;‡?;

Pour le déchiffrer : (1)

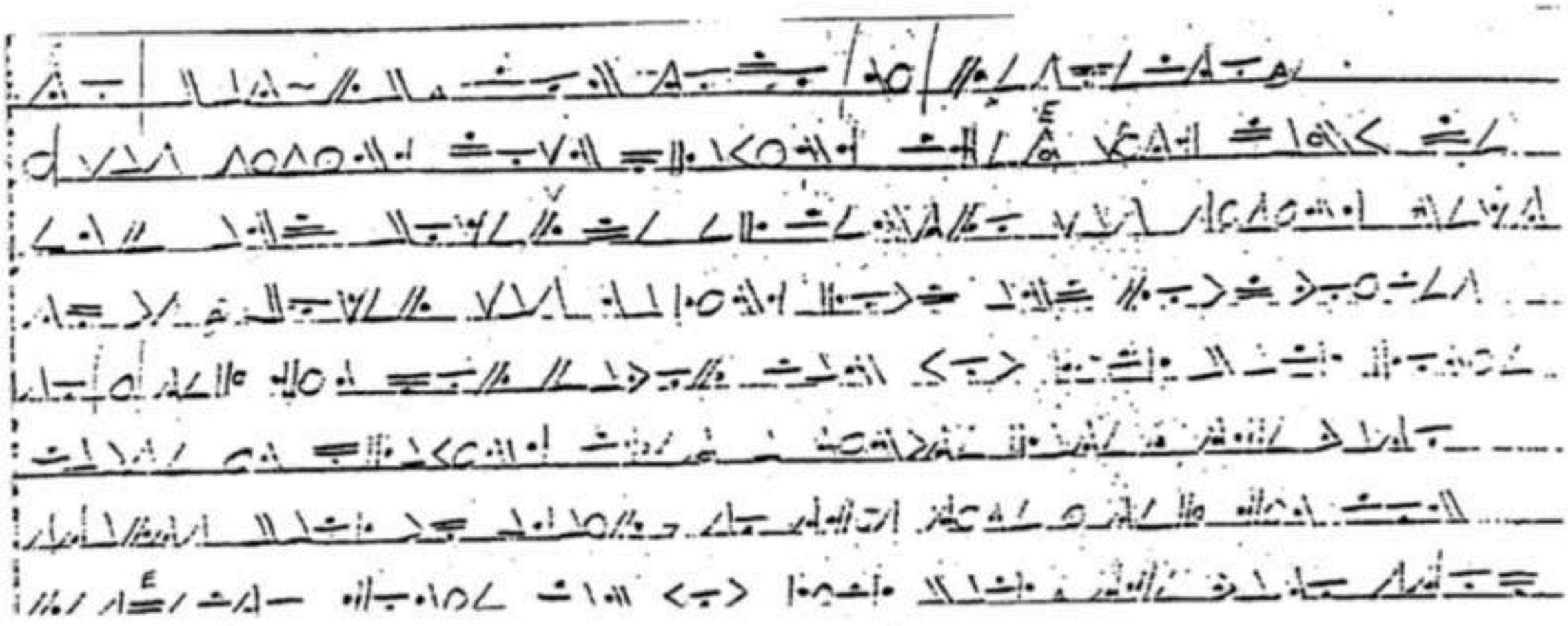
La lettre la plus fréquente en anglais est le “e”
Le symbole le plus fréquent dans le texte est le “8”
Donc le “8” code un “e”

Pour le déchiffrer : (2)

Le mot le plus fréquent en anglais est le mot “the”
La suite de 3 symboles la plus fréquente dans le texte finissant par 8 est “;48”
Donc “;48” code “the”

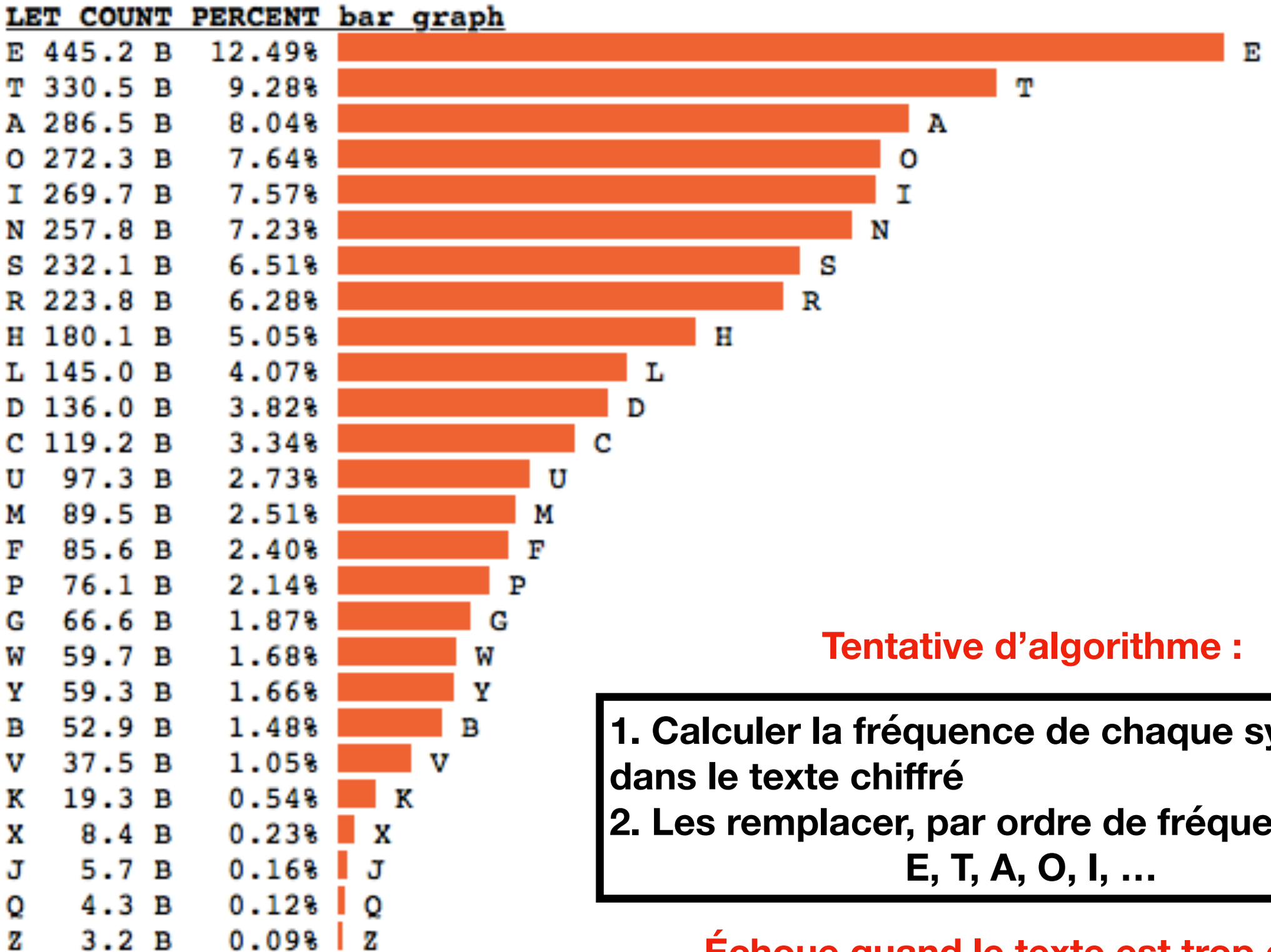
Un cas réel : communication entre prisonniers

Texte fourni par un psychologue d'une prison de Californie



Comment le déchiffrer ?

Fréquences des lettres en anglais



Tentative d’algorithme :

1. Calculer la fréquence de chaque symbole dans le texte chiffré
2. Les remplacer, par ordre de fréquence, par :
E, T, A, O, I, …

Échoue quand le texte est trop court

Fréquences des paires de lettres en anglais

BI	COUNT	PERCENT	bar graph
TH	100.3 B	(3.56%)	TH
HE	86.7 B	(3.07%)	HE
IN	68.6 B	(2.43%)	IN
ER	57.8 B	(2.05%)	ER
AN	56.0 B	(1.99%)	AN
RE	52.3 B	(1.85%)	RE
ON	49.6 B	(1.76%)	ON
AT	41.9 B	(1.49%)	AT
EN	41.0 B	(1.45%)	EN
ND	38.1 B	(1.35%)	ND
TI	37.9 B	(1.34%)	TI
ES	37.8 B	(1.34%)	ES
OR	36.0 B	(1.28%)	OR
TE	34.0 B	(1.20%)	TE
OF	33.1 B	(1.17%)	OF
ED	32.9 B	(1.17%)	ED
IS	31.8 B	(1.13%)	IS
IT	31.7 B	(1.12%)	IT
AL	30.7 B	(1.09%)	AL
AR	30.3 B	(1.07%)	AR
ST	29.7 B	(1.05%)	ST
TO	29.4 B	(1.04%)	TO
NT	29.4 B	(1.04%)	NT
NG	26.9 B	(0.95%)	NG
SE	26.3 B	(0.93%)	SE
HA	26.1 B	(0.93%)	HA
AS	24.6 B	(0.87%)	AS
OU	24.5 B	(0.87%)	OU
IO	23.5 B	(0.83%)	IO

Après avoir lu un T, la lettre suivante est

H avec probabilité 38%

I avec probabilité 14%

Après avoir lu un I, la lettre suivante est

T avec probabilité 15%

etc.

Après avoir lu un T, la lettre suivante est

H avec probabilité 38%

I avec probabilité 14%

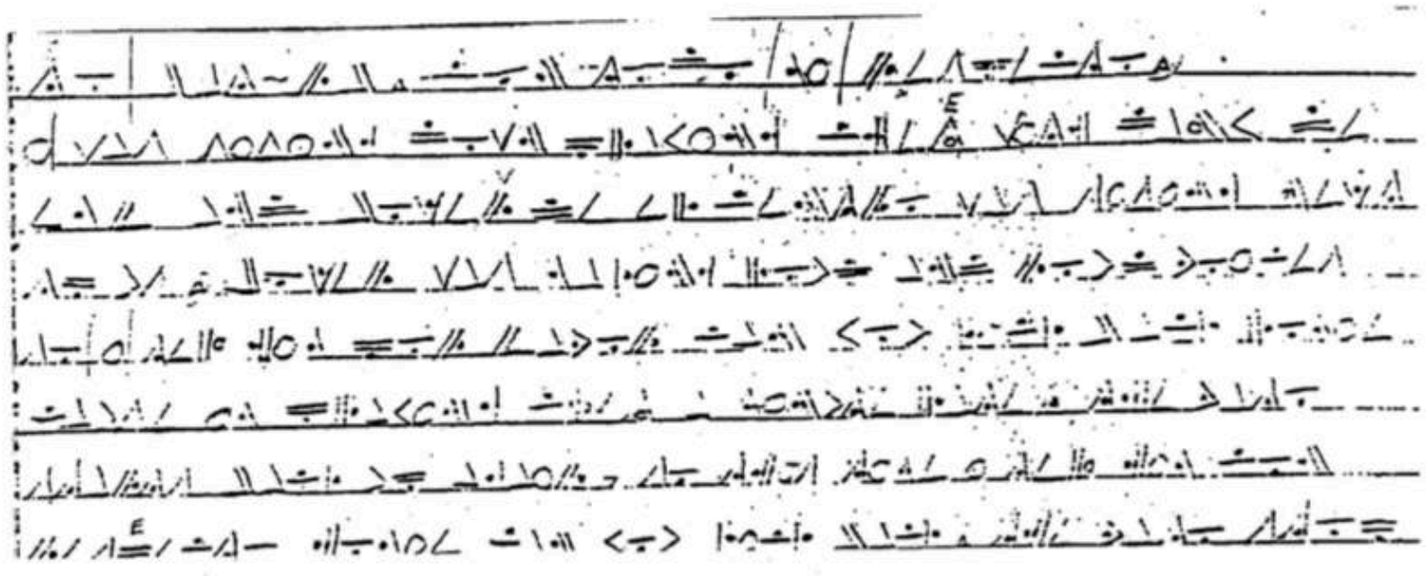
Après avoir lu un I, la lettre suivante est

T avec probabilité 15%

etc.

Qualité d'une suite de lettres :

$$Q(TITH\ldots) = \underset{\textbf{TI}}{0.14} \times \underset{\textbf{IT}}{0.15} \times \underset{\textbf{TH}}{0.38} \times \cdots$$

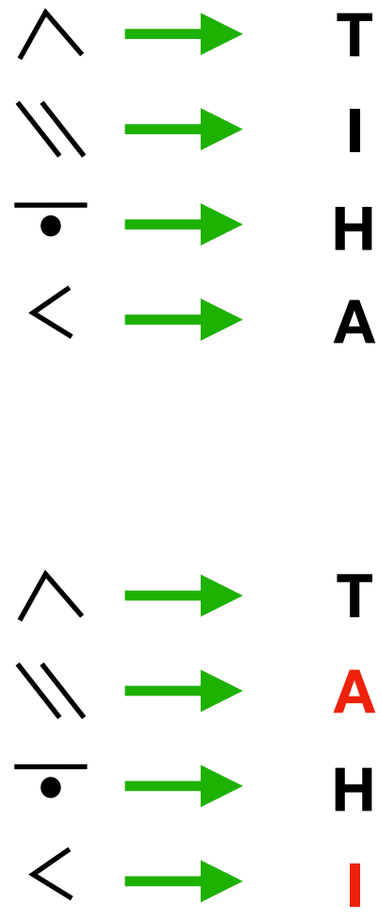


Algorithme pour déchiffrer

faire une correspondance arbitraire entre symboles et lettres de l'alphabet
remplacer les symboles par les lettres
calculer la qualité q du texte produit

répéter suffisamment de fois
  choisir 2 lettres au hasard et les échanger dans la correspondance
  calculer la qualité q' du nouveau texte
  si elle est plus grande,
    on prend cette nouvelle correspondance
  sinon,
    avec probabilité q'/q on prend la nouvelle correspondance
    avec probabilité 1-q'/q on garde l'ancienne correspondance

donner le texte ainsi déchiffré



Est-ce un algorithme de déchiffrement raisonnable ?

Test

sur un texte connu : une page de Hamlet
artificiellement chiffrée
en faisant une correspondance aléatoire
entre lettres et des symboles
Le résultat est-il correct ?

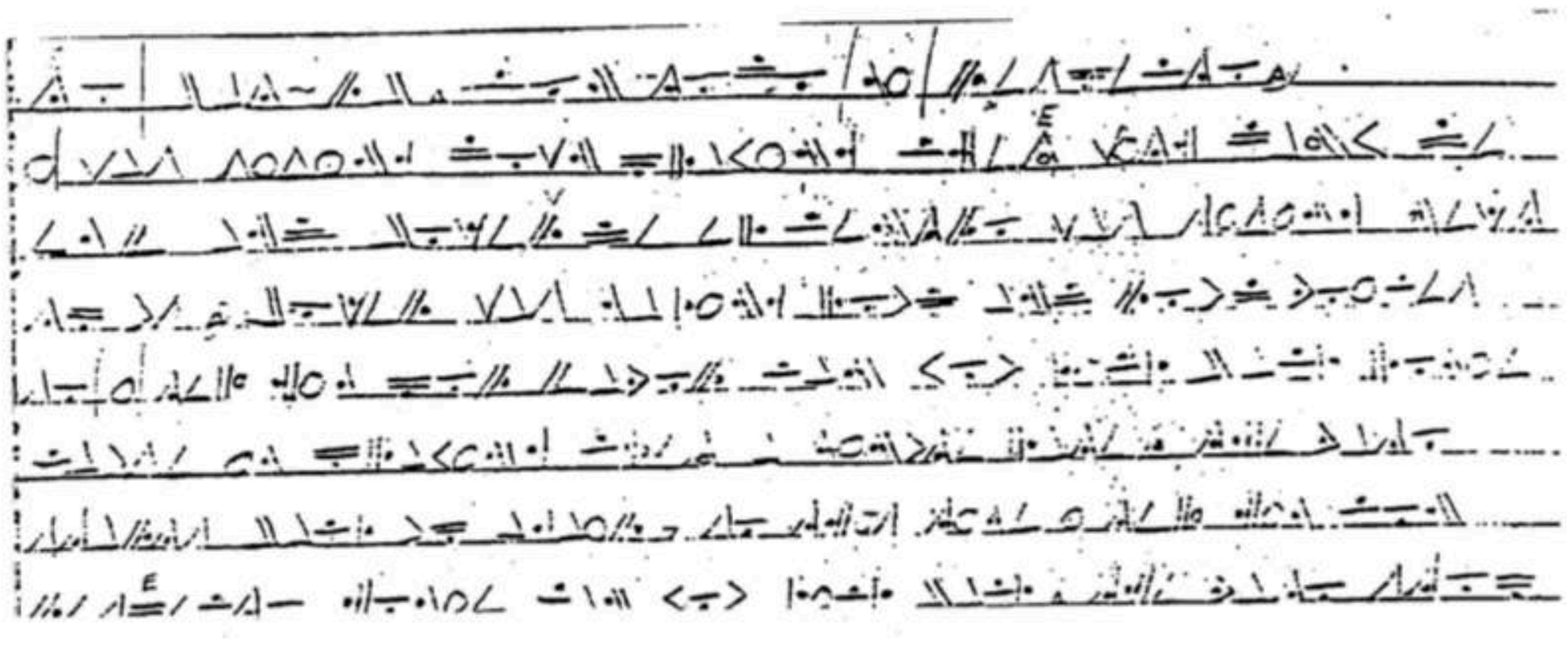
ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS
FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END

Test

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

correspondance correcte après 2000 répétitions !

Résultat sur le texte des prisonniers après 2000 itérations



to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and

ça marche !

Pourquoi l'algorithme marche-t-il ?

LET	COUNT	PERCENT	bar graph
E	445.2 B	12.49%	E
T	330.5 B	9.28%	T
A	286.5 B	8.04%	A
O	272.3 B	7.64%	O
I	269.7 B	7.57%	I
N	257.8 B	7.23%	N
S	232.1 B	6.51%	S
R	223.8 B	6.28%	R
H	180.1 B	5.05%	H
L	145.0 B	4.07%	L
D	136.0 B	3.82%	D
C	119.2 B	3.34%	C
U	97.3 B	2.73%	U
M	89.5 B	2.51%	M
F	85.6 B	2.40%	F
P	76.1 B	2.14%	P
G	66.6 B	1.87%	G
W	59.7 B	1.68%	W
Y	59.3 B	1.66%	Y
B	52.9 B	1.48%	B
V	37.5 B	1.05%	V
K	19.3 B	0.54%	K
X	8.4 B	0.23%	X
J	5.7 B	0.16%	J
Q	4.3 B	0.12%	Q
Z	3.2 B	0.09%	Z

BI	COUNT	PERCENT	bar graph
TH	100.3 B	(3.56%)	TH
HE	86.7 B	(3.07%)	HE
IN	68.6 B	(2.43%)	IN
ER	57.8 B	(2.05%)	ER
AN	56.0 B	(1.99%)	AN
RE	52.3 B	(1.85%)	RE
ON	49.6 B	(1.76%)	ON
AT	41.9 B	(1.49%)	AT
EN	41.0 B	(1.45%)	EN
ND	38.1 B	(1.35%)	ND
TI	37.9 B	(1.34%)	TI
ES	37.8 B	(1.34%)	ES
OR	36.0 B	(1.28%)	OR
TE	34.0 B	(1.20%)	TE
OF	33.1 B	(1.17%)	OF
ED	32.9 B	(1.17%)	ED
IS	31.8 B	(1.13%)	IS
IT	31.7 B	(1.12%)	IT
AL	30.7 B	(1.09%)	AL
AR	30.3 B	(1.07%)	AR
ST	29.7 B	(1.05%)	ST
TO	29.4 B	(1.04%)	TO
NT	29.4 B	(1.04%)	NT
NG	26.9 B	(0.95%)	NG
SE	26.3 B	(0.93%)	SE
HA	26.1 B	(0.93%)	HA
AS	24.6 B	(0.87%)	AS
OU	24.5 B	(0.87%)	OU
IO	23.5 B	(0.83%)	IO

Beaucoup plus d'information dans les fréquences des paires de lettres que dans les fréquences des lettres

Beaucoup plus facile de distinguer les lettres les unes des autres quand on connaît leurs "amies" (celles qui sont à côté)

Inutile de chercher beaucoup plus loin : cela suffit !

Avec un simple échange de 2 lettres dans la correspondance, peu d'itérations suffisent !

Algorithme générique (Métropolis)

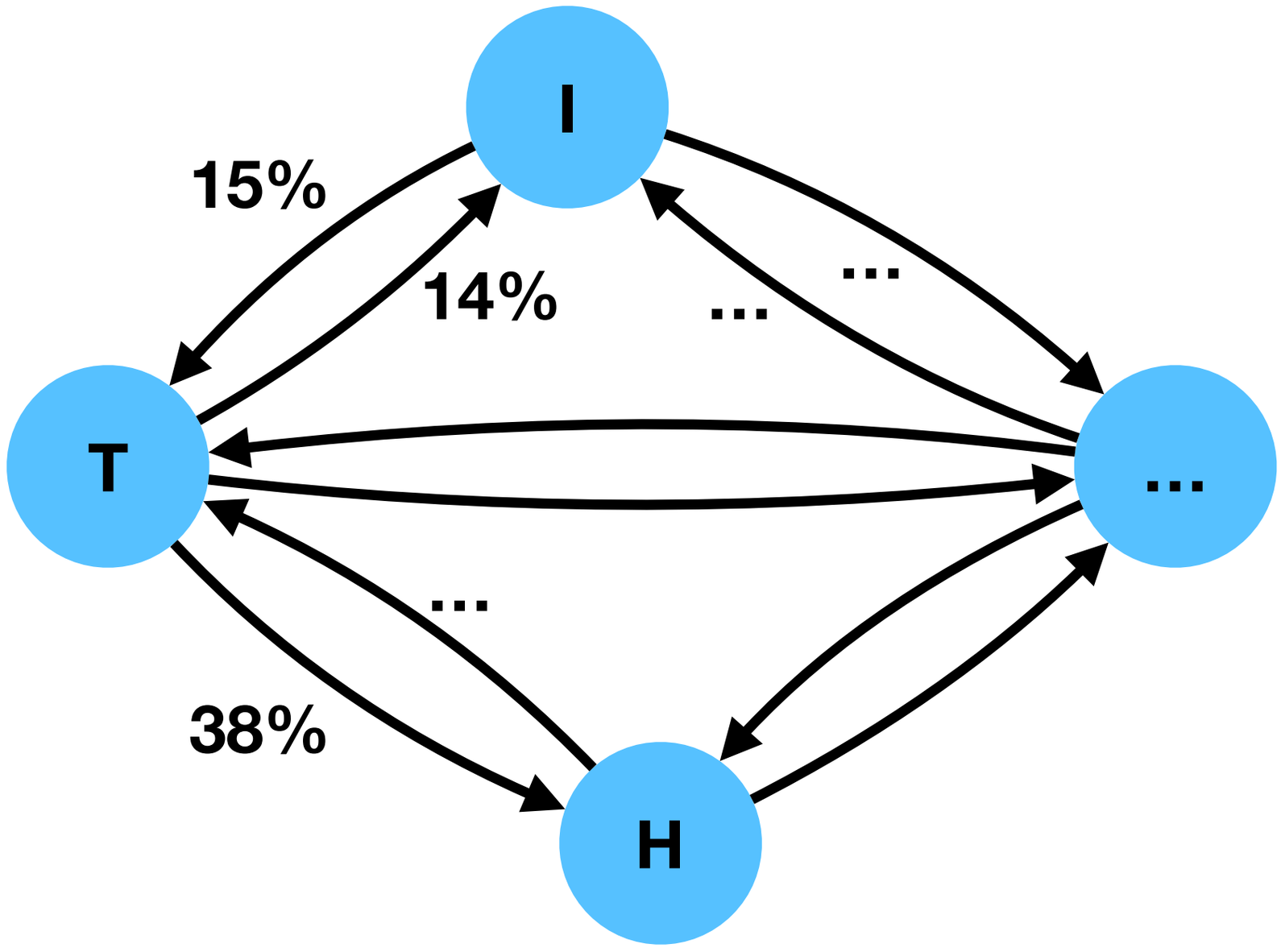
prendre une solution initiale arbitraire, calculer sa qualité q

répéter suffisamment de fois
  faire une modification locale de la solution, calculer sa qualité q'
  si elle est plus grande,
    on prend cette nouvelle solution
  sinon,
    avec probabilité q'/q on prend la nouvelle solution
    avec probabilité 1-q'/q on garde l'ancienne solution

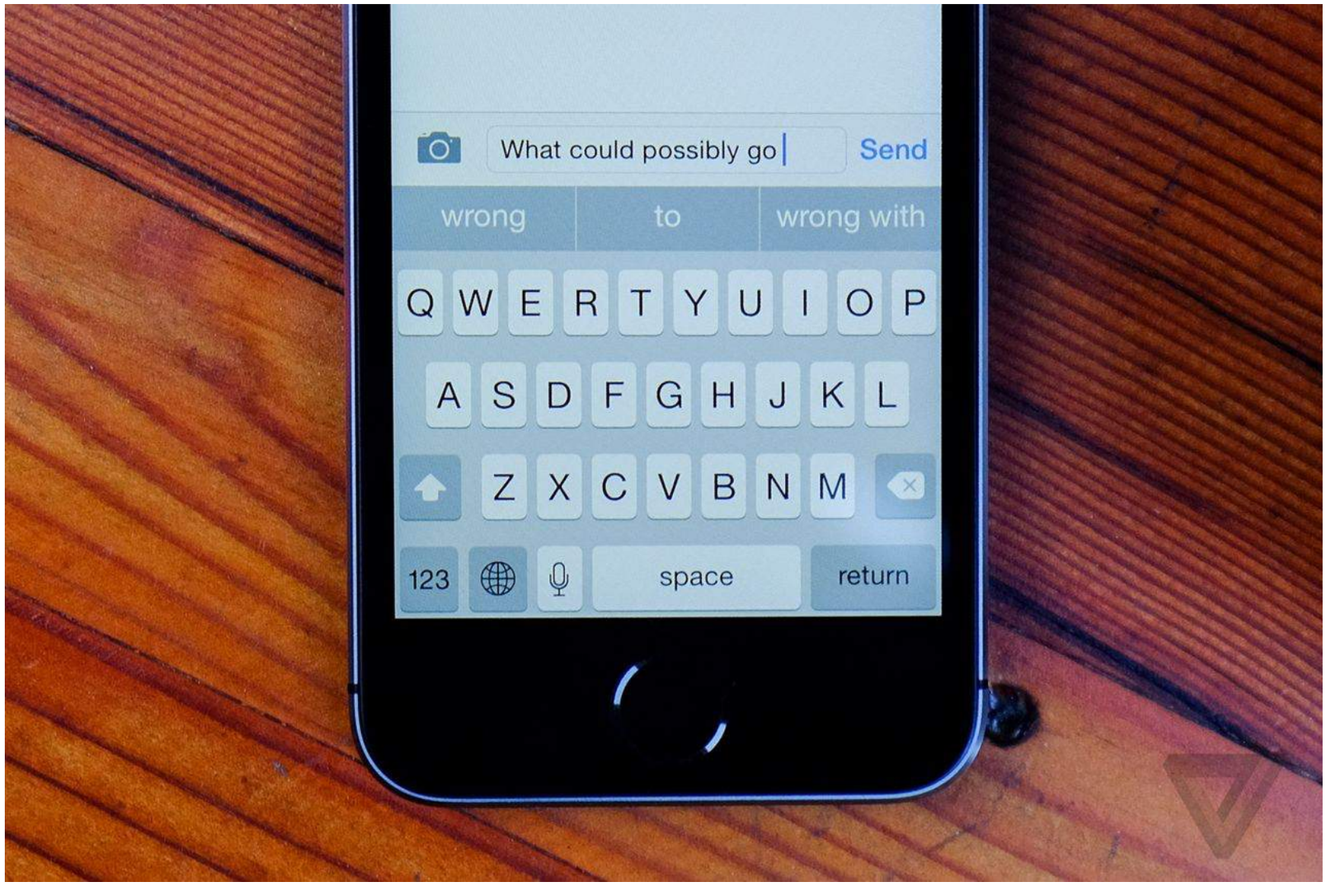
donner la solution ainsi calculée

Les difficultés en pratique : définir
"qualité"
"modification locale"
"suffisamment de fois"

Réseau “social” des lettres de l'alphabet



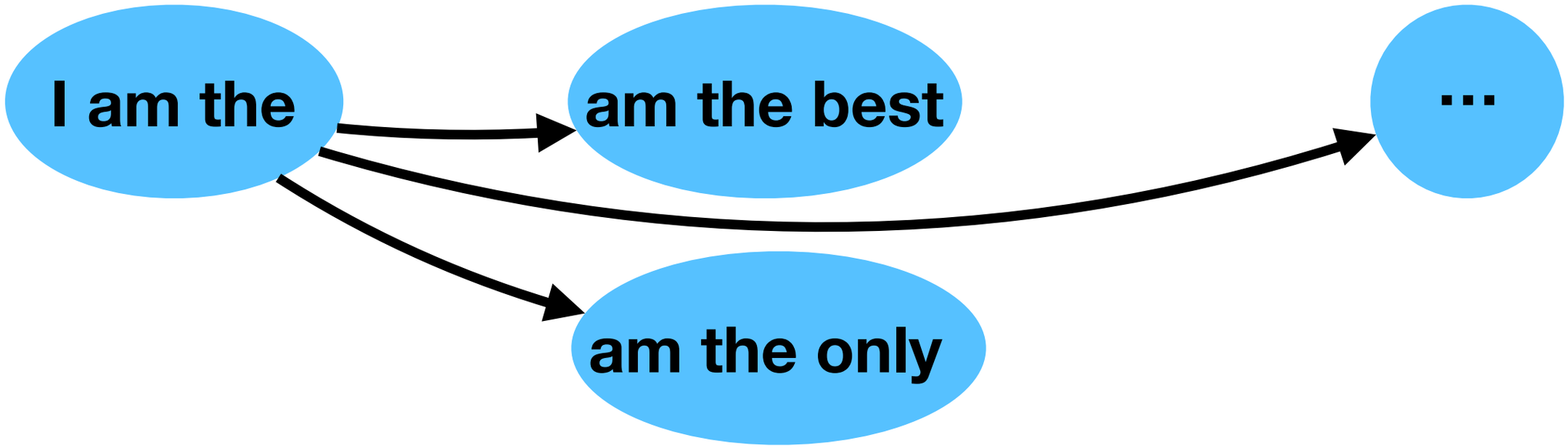
Application au quotidien : auto-complétion des mots



https://filiph.github.io/markov/

Application ludique : pastiche de tweet

À partir de l'historique des tweets de Donald Trump
calculer les fréquences de chaque suite de 3 mots
définir une chaîne de Markov
une marche aléatoire génère un tweet
"dans le style de Trump"



Donald J. Trump
@realDonaldTrump

Follow

#TrumpVlog Trouble in paradise. When I renovated Wollman Rink under budget, ahead of schedule. When others drop out, I was always for it!

4:38 - 9 Jan 2018

13K 37K

Donald J. Trump
@realDonaldTrump

Follow

Make the Boston killer applying today for a ridiculous $1, 000 for NOTHING

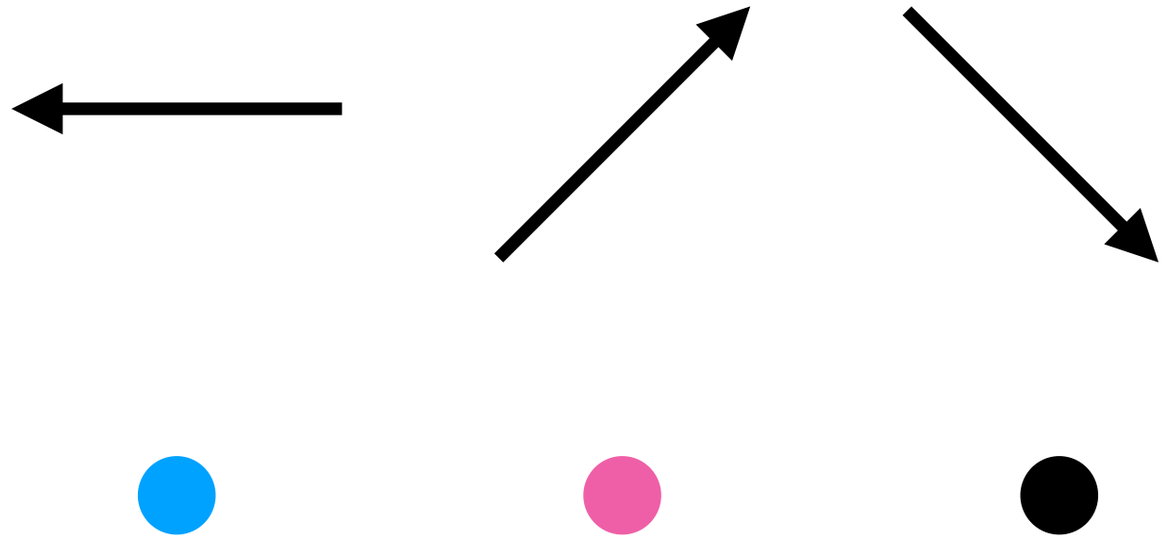
4:38 - 9 Jan 2018

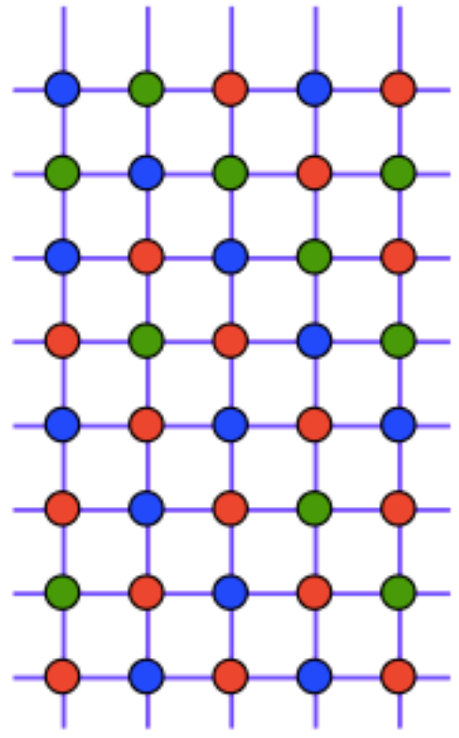
13K 37K

Colorier un graphe

Physique statistique : modèle de Potts

Spins dans un réseau : q=3 états des spins



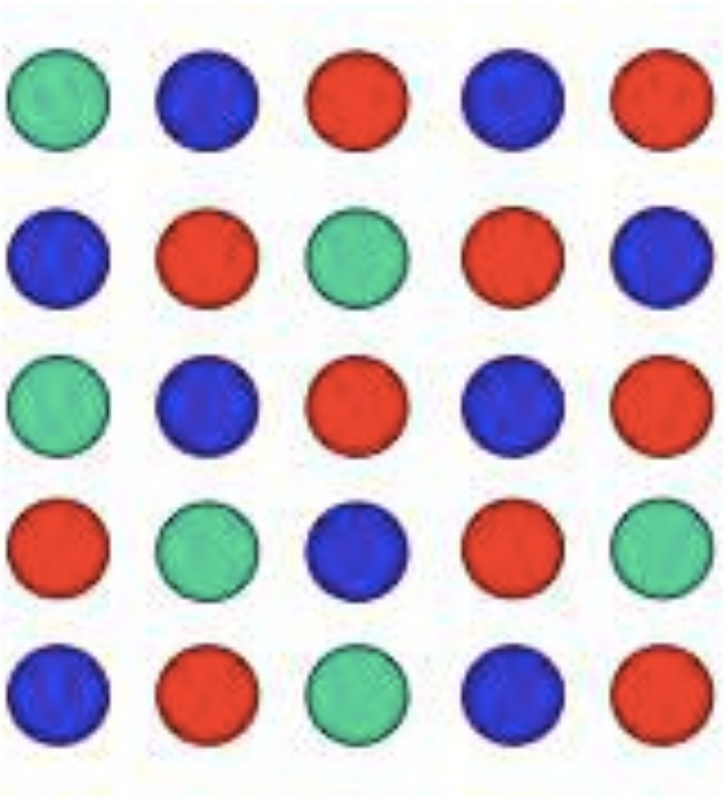


Interactions entre voisins
Des spins voisins sont dans des états différents
(“hard-core”)

Coloriage d’un graphe avec q=3 couleurs

À quoi ressemble un coloriage aléatoire ?

Comment construire un coloriage aléatoire ?



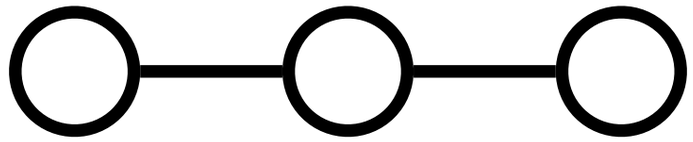
prendre un coloriage initial arbitraire

répéter suffisamment de fois
  choisir un sommet et une couleur au hasard
  si on en a le droit,
    donner à ce sommet la nouvelle couleur

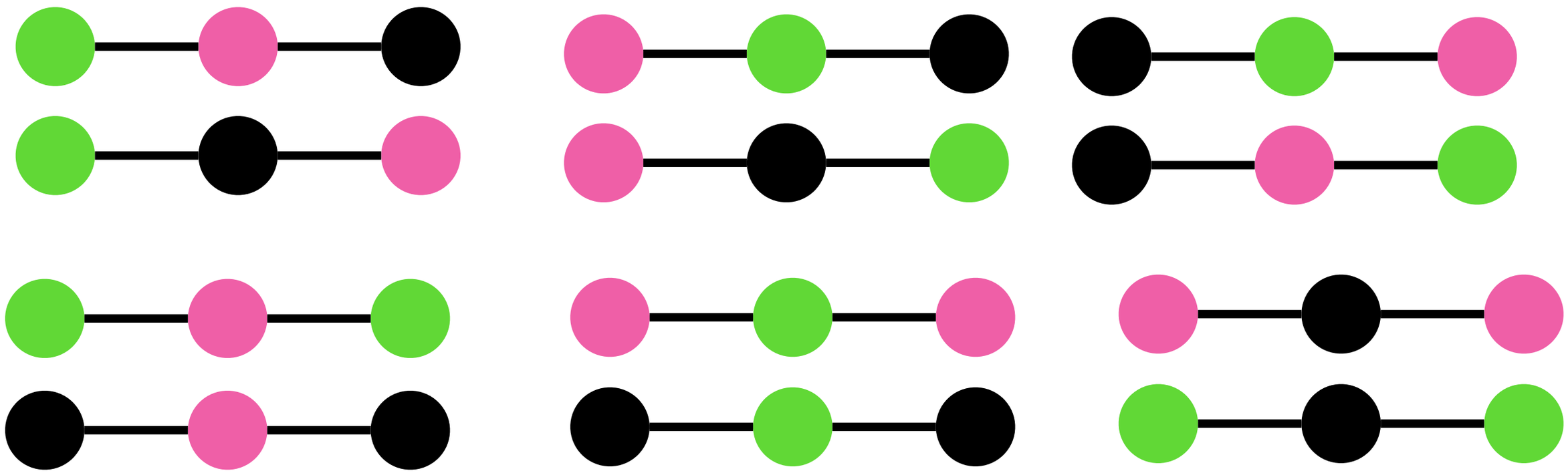
résultat : le coloriage ainsi calculé

Au bout d'un temps infini (distribution stationnaire) :
tous les coloriages donnés en résultat ont la même probabilité
distribution uniforme

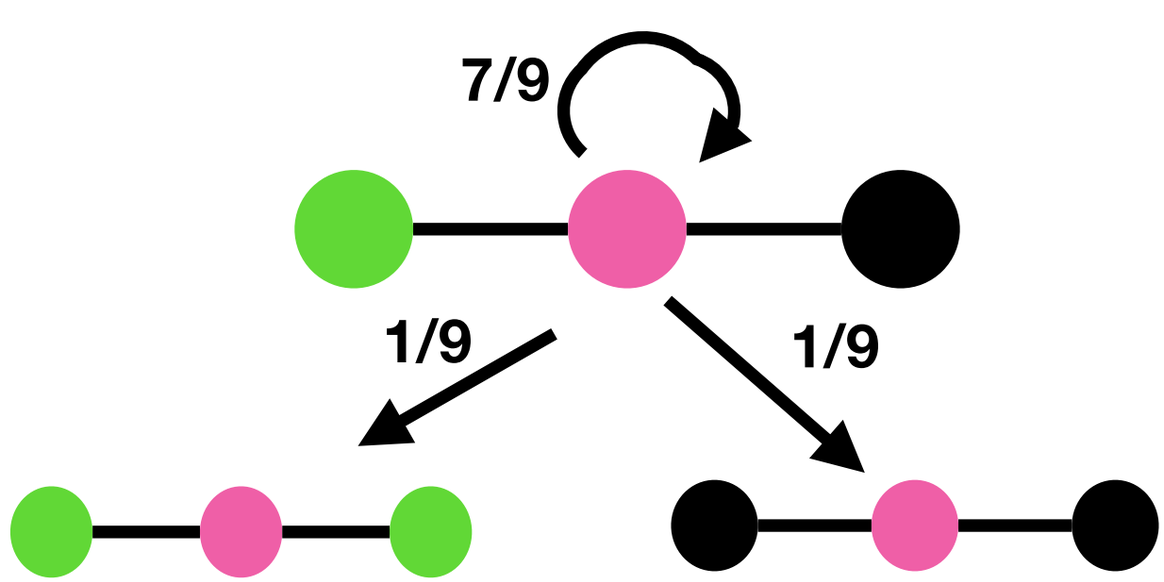
Le graphe :

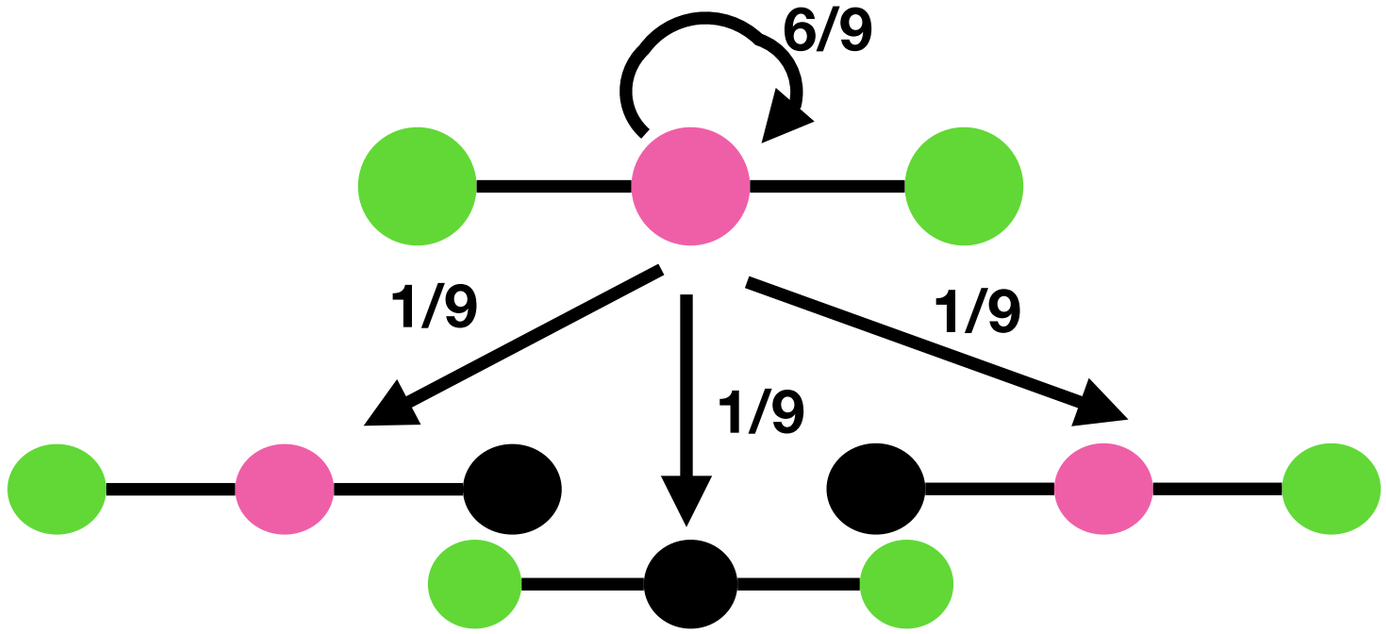


Les 3-coloriages du graphe :

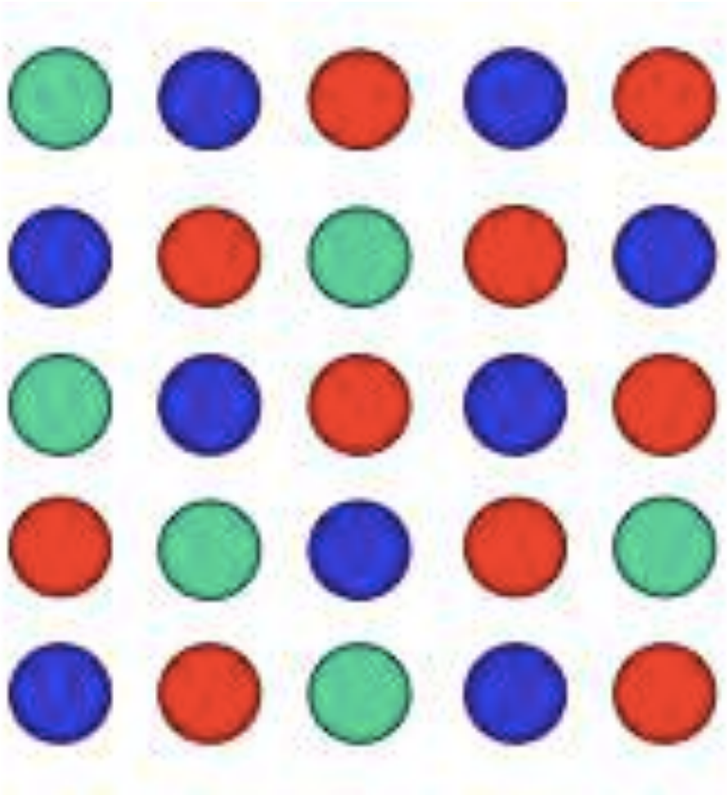


Les transitions entre 3-coloriages :





Comment construire un coloriage aléatoire ?



prendre un coloriage initial arbitraire

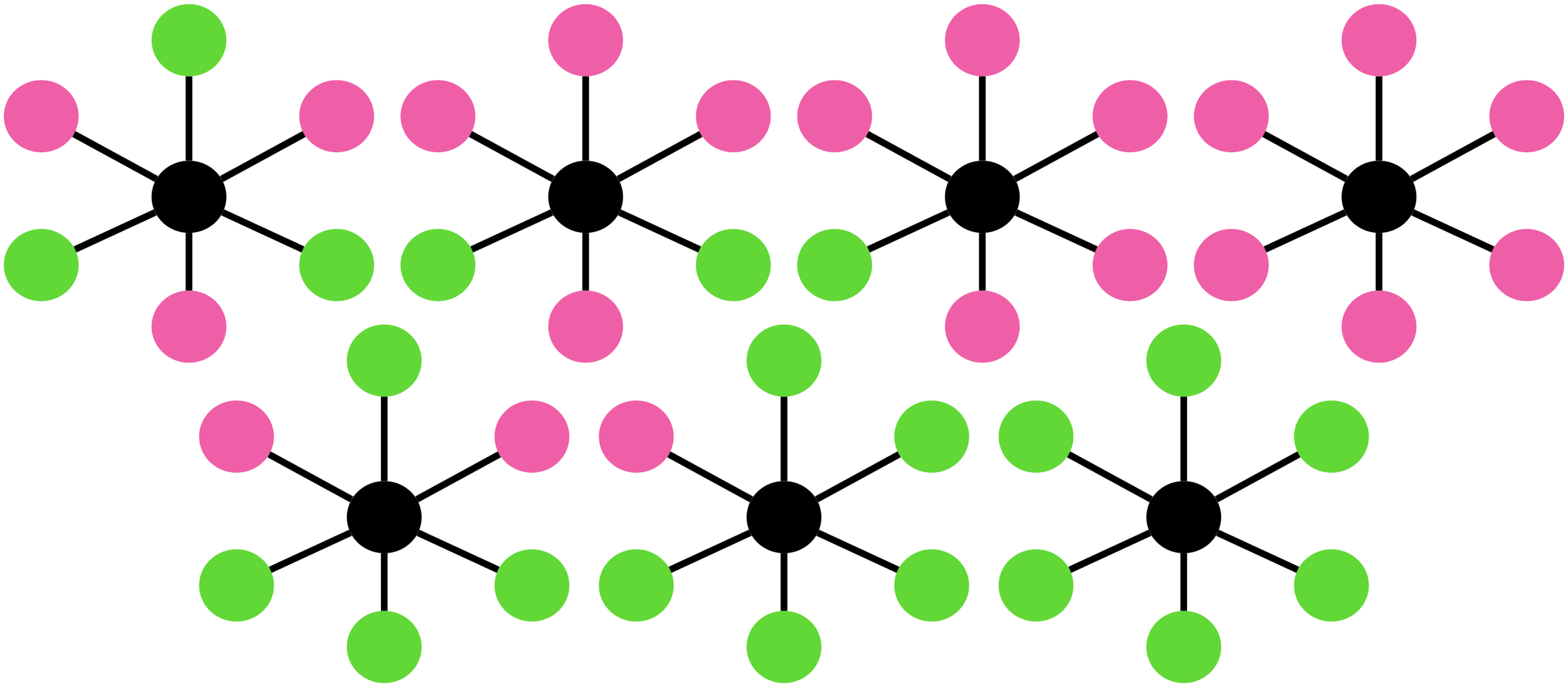
répéter suffisamment de fois
  choisir un sommet et une couleur au hasard
  si on en a le droit,
    donner à ce sommet la nouvelle couleur

résultat : le coloriage ainsi calculé

Au bout d'un certain temps (fini) :
tous les coloriages donnés en résultat ont presque la même probabilité
distribution quasi-uniforme

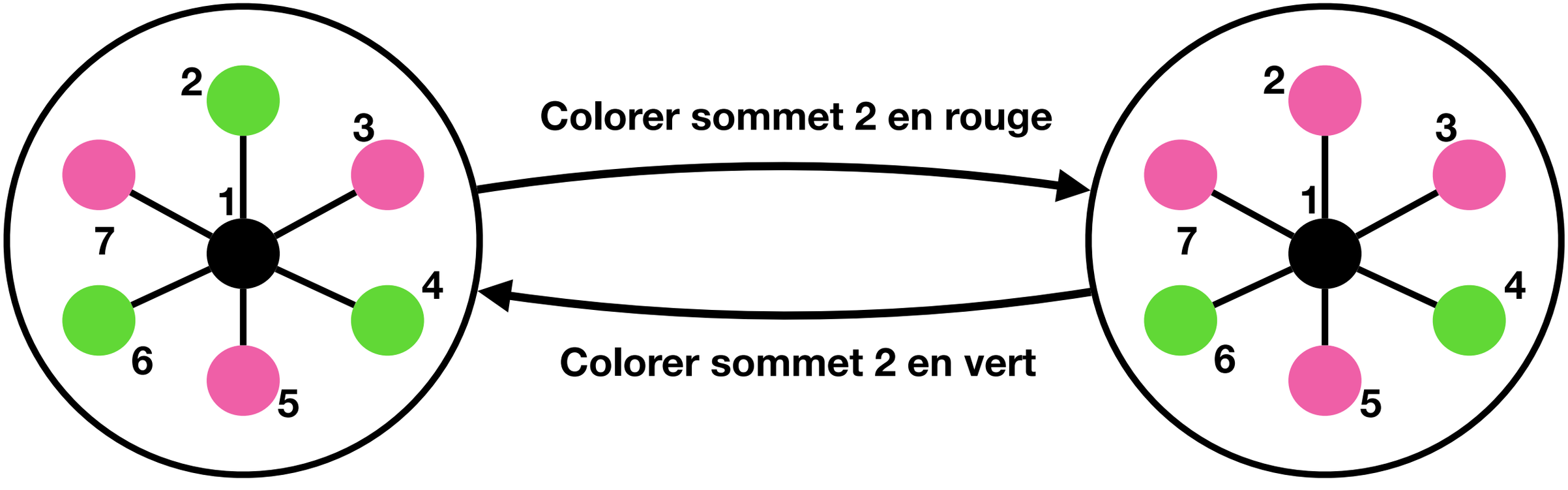
Temps de convergence ?

Un cas désespéré : le graphe étoile



Quelques configurations

Le graphe des configurations



sommet : une configuration

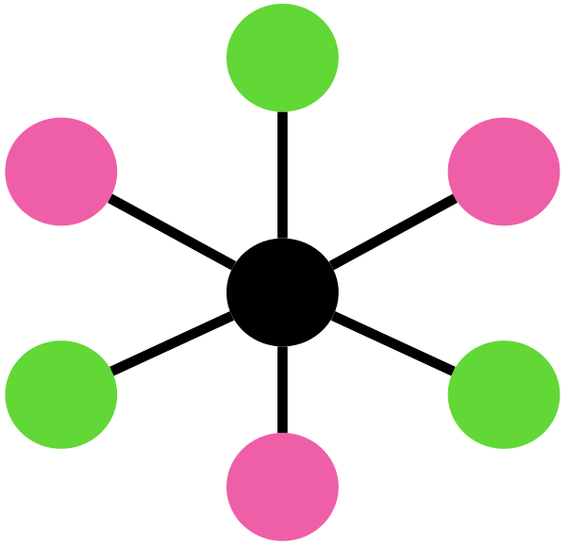
prendre un coloriage initial arbitraire

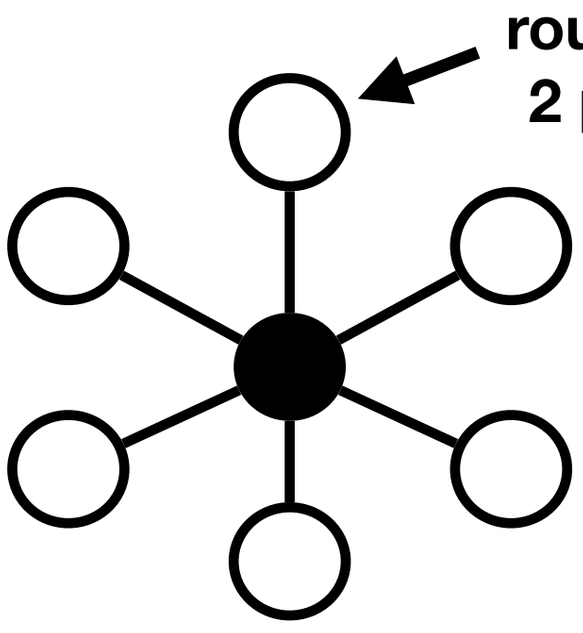
répéter suffisamment de fois
  choisir un sommet et une couleur au hasard
  si on en a le droit,
    donner à ce sommet la nouvelle couleur

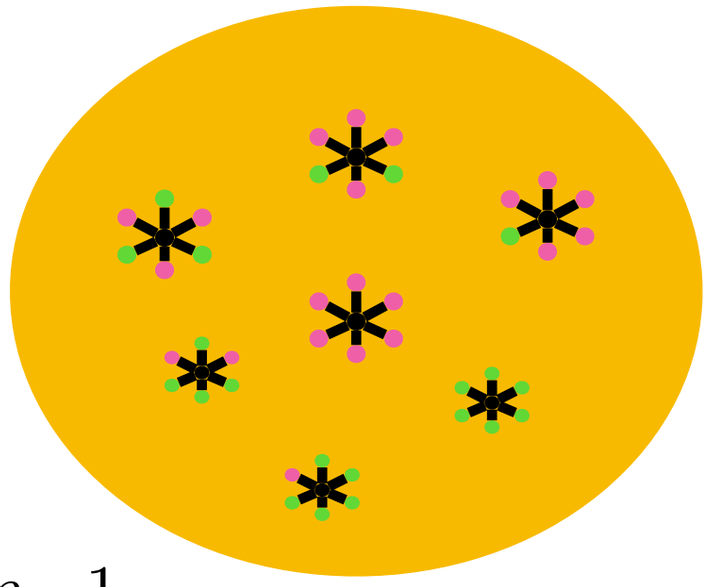
résultat : le coloriage ainsi calculé

Marche aléatoire sur le graphe des configurations

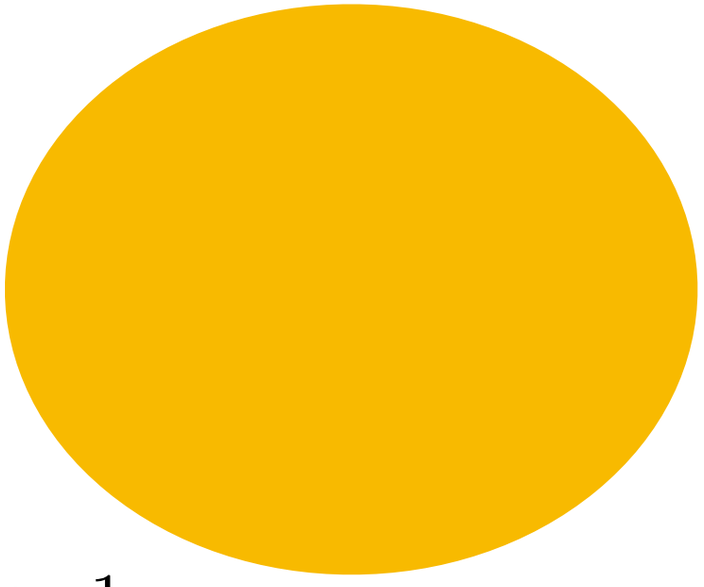
Étude du graphe des configurations



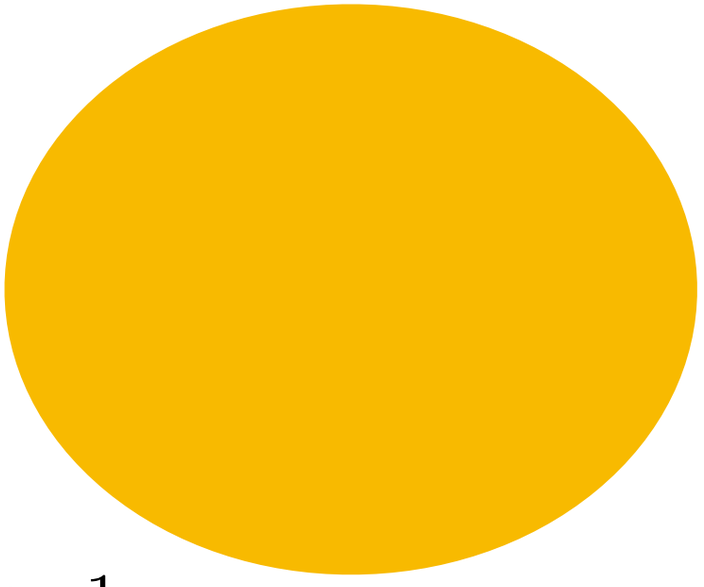




2^{n-1} configurations avec sommet central noir

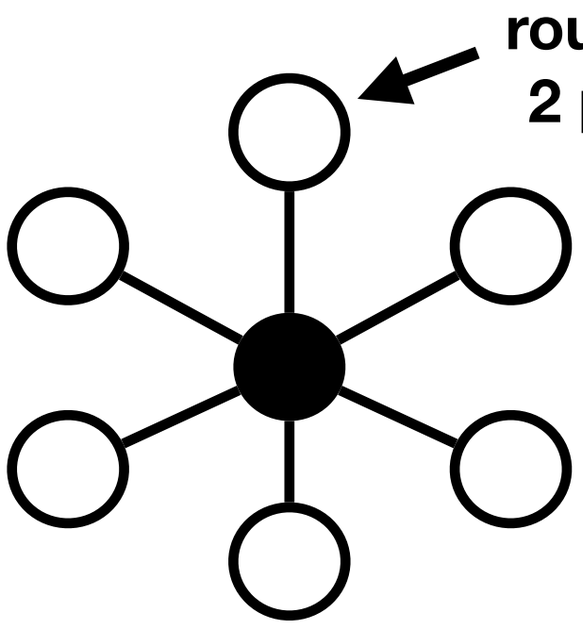


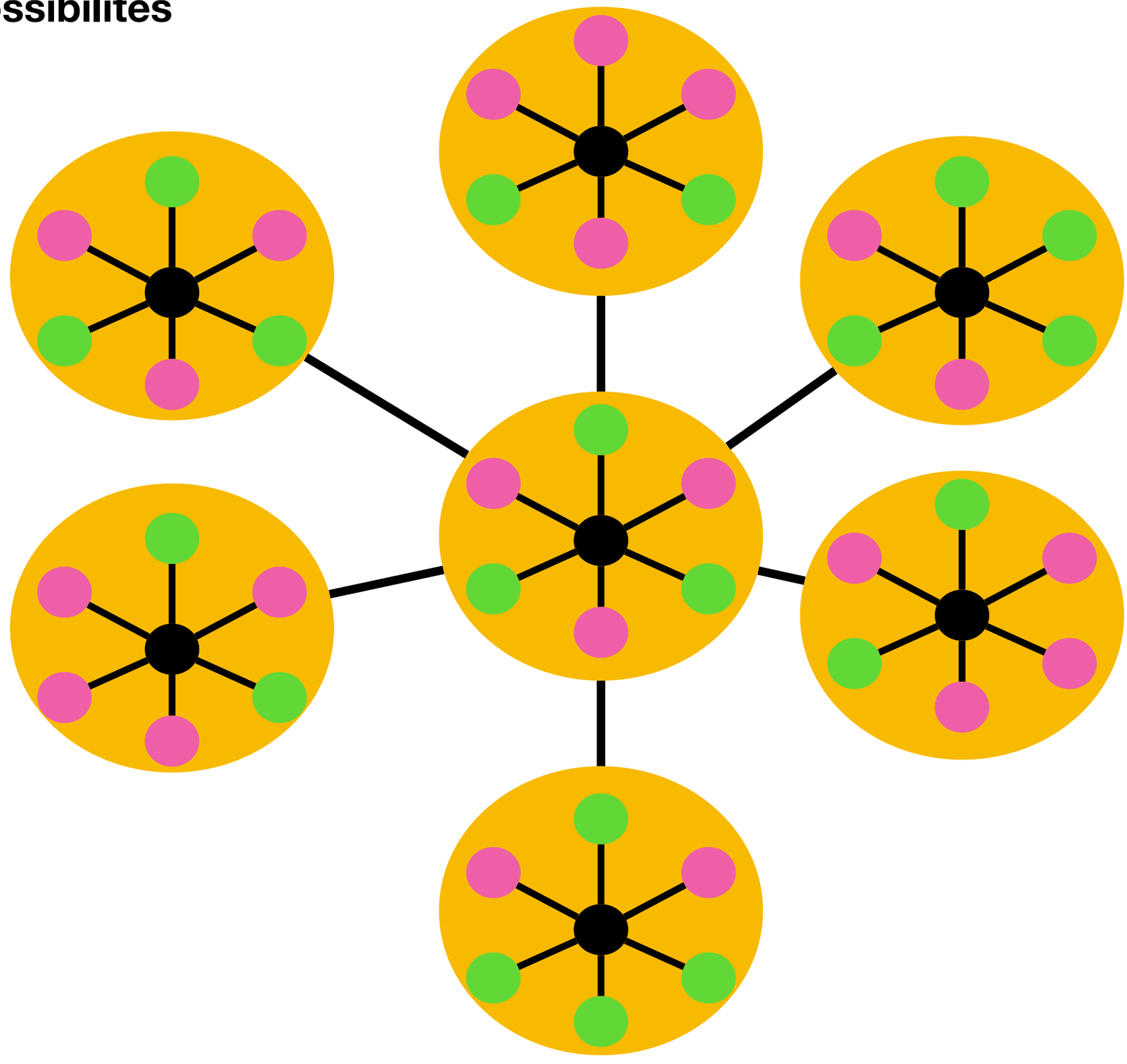
2^{n-1} configurations avec sommet central rouge



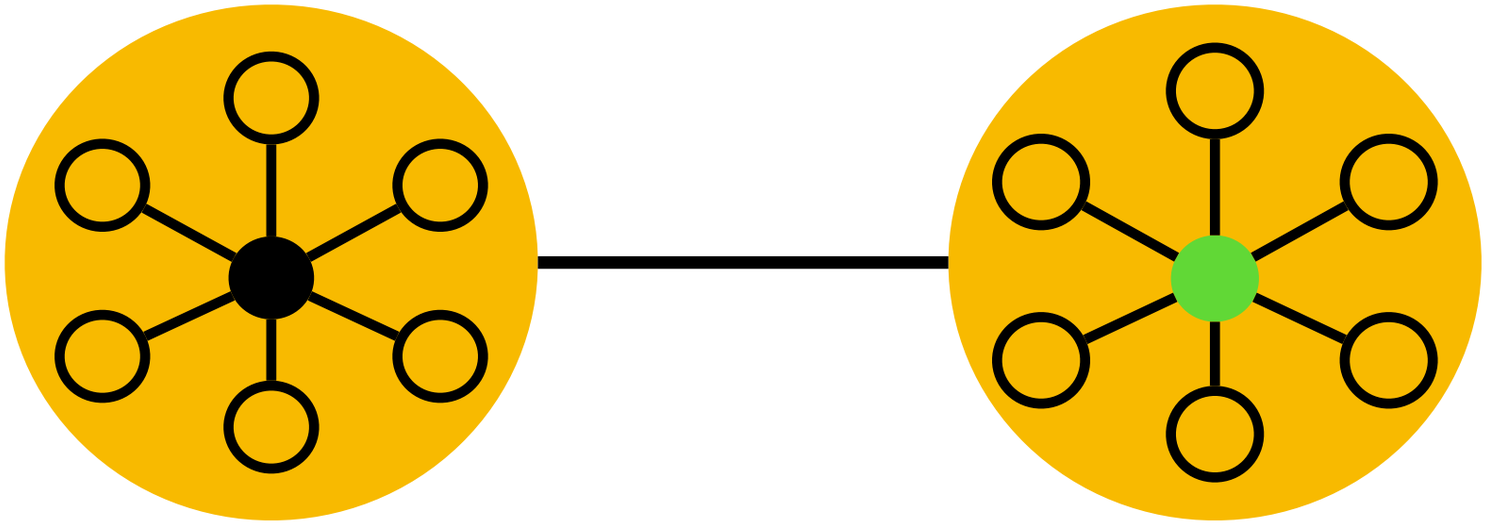
2^{n-1} configurations avec sommet central vert

Transitions entre configurations avec sommet central noir

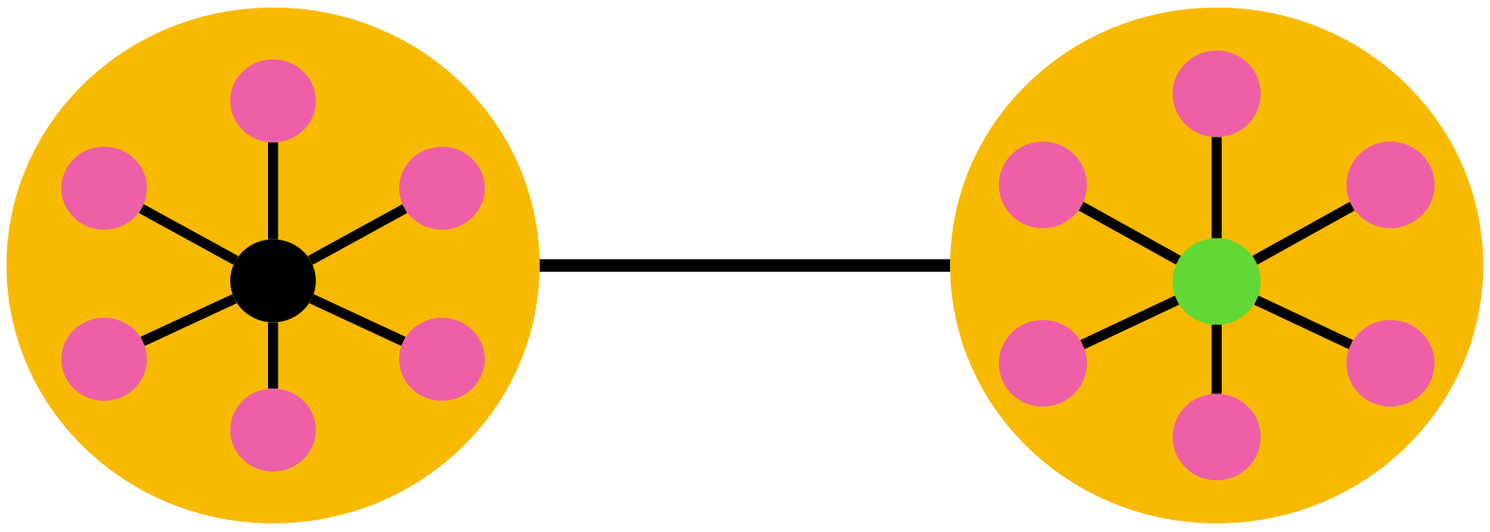




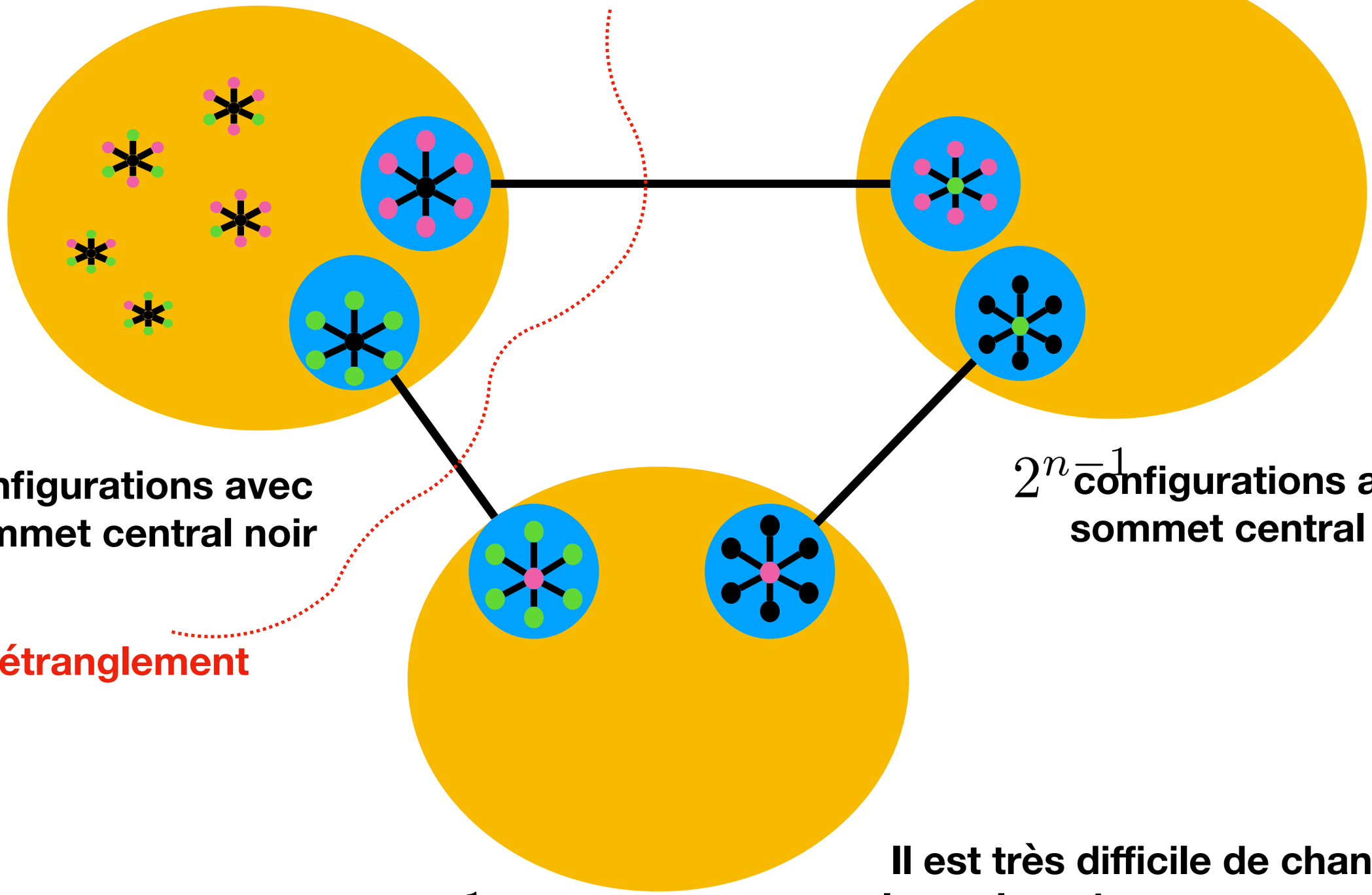
Transitions sommet central noir - sommet central vert



N’est possible que pour une seule configuration !



Graphe des configurations



2^{n-1} configurations avec sommet central noir

goulot d'étranglement

2^{n-1} configurations avec sommet central vert

2^{n-1} configurations avec sommet central rouge

Il est très difficile de changer la couleur du sommet central: pour avoir une distribution quasi-uniforme, il faut attendre un temps exponentiel…

prendre un coloriage initial arbitraire

répéter suffisamment de fois
  choisir un sommet et une couleur au hasard
  si on en a le droit,
    donner à ce sommet la nouvelle couleur

résultat : le coloriage ainsi calculé

“ Suffisamment de fois” = un temps exponentiel,
pour le graphe-étoile

“ Suffisamment de fois” = un temps polynomial,
pour la grille

Pour comprendre combien de temps prend
l’algorithme de marche aléatoire,
il faut comprendre le graphe

Le temps de convergence dépend
de la géométrie du graphe des configurations

prendre un coloriage initial arbitraire

répéter suffisamment de fois
  choisir un sommet et une couleur au hasard
  si on en a le droit,
    donner à ce sommet la nouvelle couleur

résultat : le coloriage ainsi calculé

Problème physique : modèle de Potts
Algorithme d'échantillonage aléatoire
Analyse de convergence : géométrie des graphes

Physique (statistique), informatique (algorithmique), mathématique (géométrie) :
tout est lié

Taille d’un réseau social

Que savons-nous de Facebook, Twitter, Linkedin, etc.?

Le graphe est grand et il évolue. Nous n'en voyons qu'une petite fraction. Que pouvons-nous faire ?

Nous connaissons notre degré. Nous avons accès à nos "amis" (voisins).

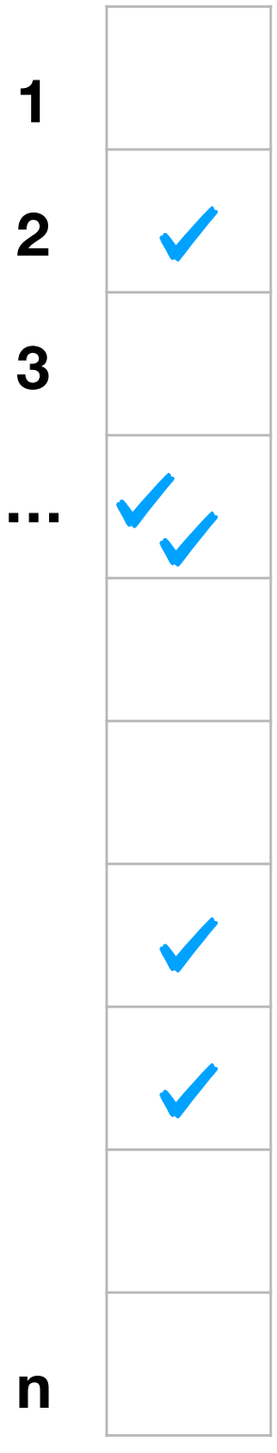
Et de là aux voisins de nos voisins, etc. Nous pouvons explorer le graphe à partir d'un voisin choisi au hasard : marche aléatoire dans le réseau. Qu'est-ce que cela peut nous apprendre sur le graphe ?

Une question de base : combien y a-t-il de sommets ?

Qu: Au bout d'un temps suffisamment long, où est la marche aléatoire ?
Rép: à un sommet tiré au hasard avec probabilité proportionnelle à son degré.

Algorithme On-ne-sait-comment

Pour i=1,2,… :
Faire une marche aléatoire suffisamment longue.
Soit x(i) le sommet où la marche aboutit.
De la distribution des x(i) et de leurs degrés,
déduire, on ne sait comment, une estimation de la taille du graphe.



Si tous les sommets ont le même degré

x(i) est un sommet aléatoire uniforme

paradoxe des anniversaires :

après $\sqrt{n}$ tirages on a une collision :

x(k)=x(i) pour i<k.

Idée : soit k le premier tirage donnant une collision.

Alors n vaut environ k^2

Hypercube : $n = 2^m$ sommets

Si les degrés sont divers

Faire k marches aléatoires
Coll = nombre de collisions
Deg = somme des degrés
Inv = somme des inverses des degrés
n = E(Deg)E(Inv)/(2E(Coll)) * (r-1)/r

Algorithme

Pour i=1,2,…, k :
Faire une marche aléatoire suffisamment longue.
Soit x(i) le sommet où la marche aboutit.
De la distribution des x(i) et de leurs degrés,
déduire une estimation de la taille du graphe comme suit :
Coll = nombre de collisions
Deg = somme des degrés
Inv = somme des inverses des degrés
Résultat = Deg * Inv/(2 * Coll) * (k-1)/k

Peut-on mieux faire ?

Oui si on utilise les sommets intermédiaires visités pendant la marche aléatoire

Algorithme si tous les sommets ont le même degré

Pour i=1,2,…, k :

Faire deux marches aléatoires x(i), y(i) suffisamment longues (longueur T).

Soit x(i,t) le sommet où la marche x(i) se trouve à l'instant t.

Déduire une estimation de la taille du graphe comme suit :

Coll(i) = nombre de collisions entre x(i,t) et y(i,s) pour tous les instants t,s

Résultat = $T^2/(\sum_i \mathrm{Coll}(i)/k)$

Pour comprendre le graphe réseau social, il faut comprendre combien de temps prend l'algorithme de marche aléatoire

Conclusion

Une marche aléatoire est un algorithme de recherche dans un graphe par modifications locales aléatoires

Lien intime entre

- **temps de convergence de marche aléatoire sur un graphe, et**
- **structure de ce graphe**

Ils s'éclairent l'un l'autre.