

# MAPSI — cours 6 : Chaîne de Markov

Nicolas Thome  
Transparents de Vincent Guigue  
[nicolas.thome@isir.upmc.fr](mailto:nicolas.thome@isir.upmc.fr)

LIP6 / ISIR – Sorbonne Université, France

- Les problèmes traités jusqu'ici :

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ \vdots & & & \\ X_{N1} & X_{N2} & \cdots & X_{Nd} \end{bmatrix}, \text{ et parfois : } Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

- Chaque individu  $\mathbf{x} = [x_1, x_2, \dots, x_d]$  est un vecteur i.i.d.
- **Les séquences** ne rentrent pas dans ce cadre

# Traitement des séquences

## Tâches :

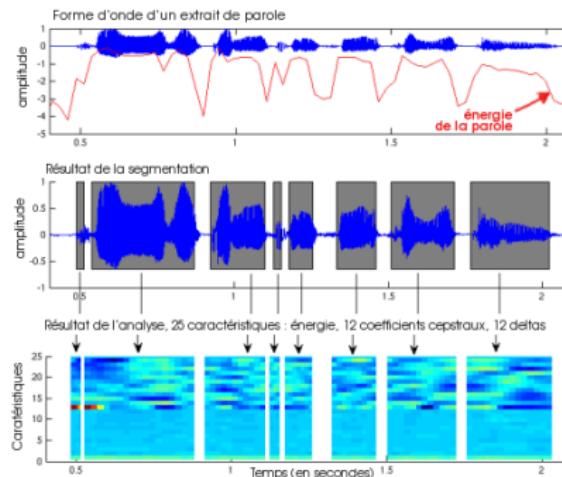
- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion
- Reconnaissance de chaîne de caractères



# Traitement des séquences

Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion
- Reconnaissance de paroles



# Traitement des séquences

## Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion
- Reconnaissance de mouvements



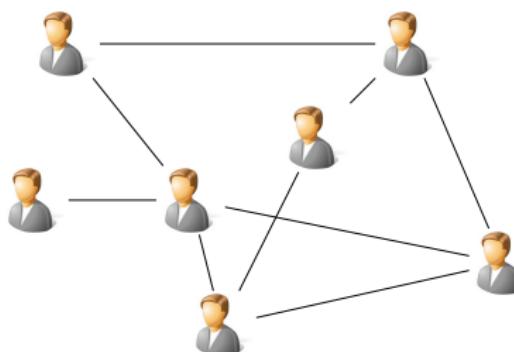
# Traitement des séquences

## Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Génération de séquences
- Modélisation de la diffusion
- Génération de mouvements

## Tâches :

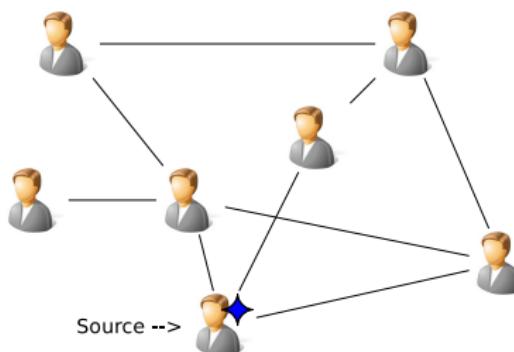
- Classification / Clustering
- Etiquetage / Segmentation
- Diffusion dans les graphes
- Génération de séquences
- Modélisation de la diffusion



# Traitement des séquences

## Tâches :

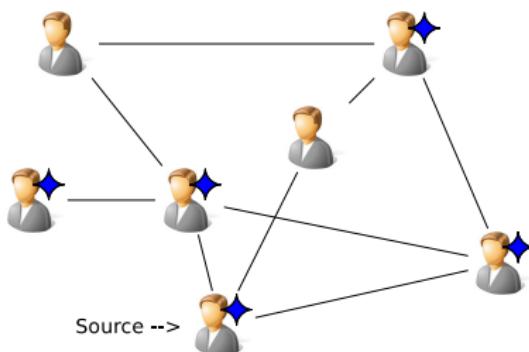
- Classification / Clustering
- Etiquetage / Segmentation
- Diffusion dans les graphes
- Génération de séquences
- Modélisation de la diffusion



# Traitement des séquences

## Tâches :

- Classification / Clustering
- Etiquetage / Segmentation
- Diffusion dans les graphes
- Génération de séquences
- Modélisation de la diffusion



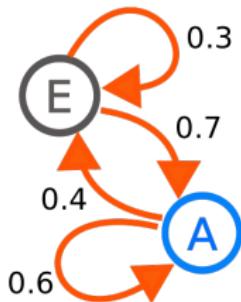
# Approche générative

## Problème

- Méthodes standards de classification/clustering  $\Rightarrow$  transition difficile vers les **données de taille variable**
- $\Rightarrow$  modèles génératifs de séquences (de taille variable)

Approche vectorielle :  $X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ \vdots & & & \\ X_{N1} & X_{N2} & \cdots & X_{Nd} \end{bmatrix}$

Approche séquentielle :



# Modèles génératifs

Nicolas Thome  
Transparents de Vincent Guigue  
[nicolas.thome@isir.upmc.fr](mailto:nicolas.thome@isir.upmc.fr)

LIP6 / ISIR – Sorbonne Université, France

# Rappel sur les modèles génératifs

- ➊ Choix d'une modélisation des données :  $p(\mathbf{x}|\theta)$
- ➋ Apprentissage = trouver  $\theta$
- ➌ Application possible : décision bayesienne

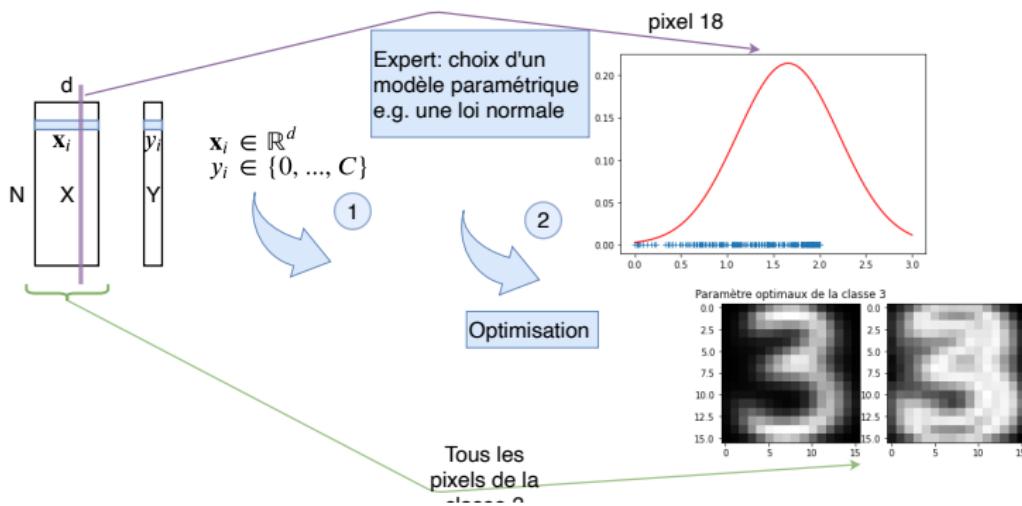
$$r(\mathbf{x}) = \arg \max_k p(\theta_k | \mathbf{x}) = \frac{p(\mathbf{x} | \theta_k) p(\theta_k)}{p(\mathbf{x})}$$

- ➍ Application bis : génération de  $\tilde{\mathbf{x}} \sim \mathcal{D}(\theta_k)$

Apprentissage d'un modèle génératif  $\Leftrightarrow$  Estimation de densité

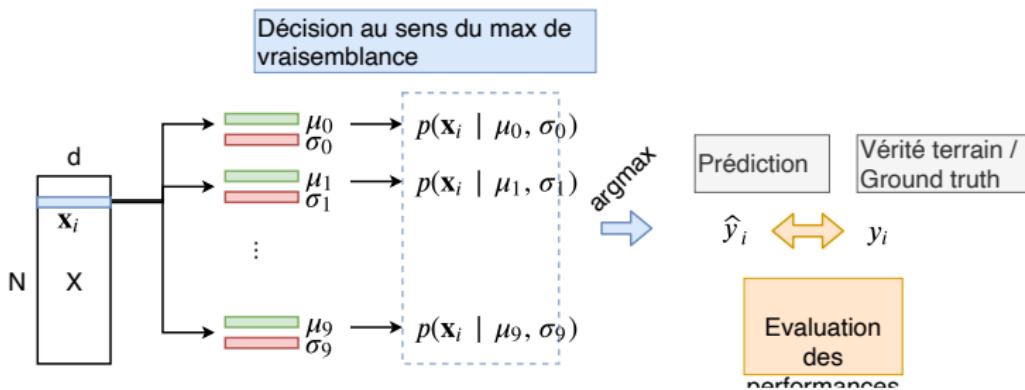
- ➎ Estimer  $\theta_k$  = estimer une densité de probabilité d'une classe
- ➏ Hypothèse (forte) : les  $\theta_k$  sont supposés indépendants
- ➐ Techniques d'estimation des  $\theta_k$

## 1 Schéma général du maximum de vraisemblance :



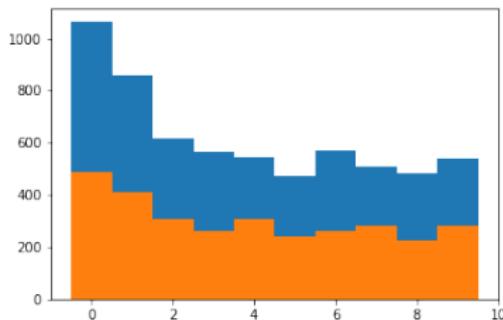
- 2 Inférence au sens du max de vraisemblance
- 3 Prise en compte des a priori
- 4 Génération

- 1 Schéma général du maximum de vraisemblance :
- 2 Inférence au sens du max de vraisemblance



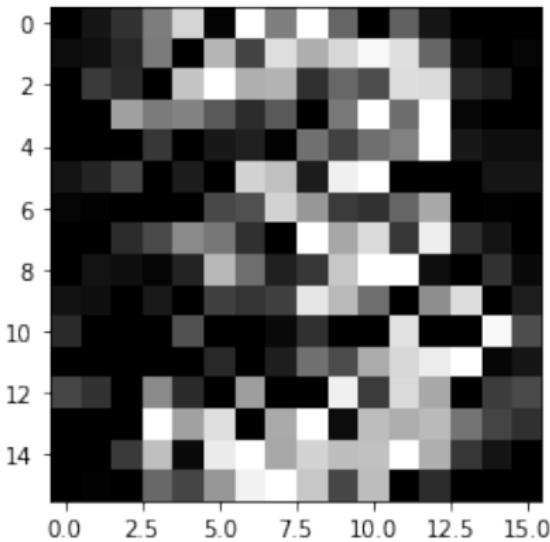
- 3 Prise en compte des a priori
- 4 Génération

- ① Schéma général du maximum de vraisemblance :
- ② Inférence au sens du max de vraisemblance
- ③ Prise en compte des a priori



- ④ Génération

- ➊ Schéma général du maximum de vraisemblance :
- ➋ Inférence au sens du max de vraisemblance
- ➌ Prise en compte des a priori
- ➍ Génération



# Maximum de vraisemblance (données vec.)

- $D = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  exemples supposés générés par  $p(\mathbf{x}|\theta)$
- Adéquation entre les données et le modèle
  - Notion de vraisemblance des observations
  - Hyp : les données sont indépendantes

$$\mathcal{L}(D, \theta) = p(D|\theta) = \prod_{i=1}^N p(\mathbf{x}_i|\theta)$$

- Optimisation :

$$\theta^* = \arg \max_{\theta} (\mathcal{L}(D, \theta)) = \arg \max_{\theta} (\log \mathcal{L}(D, \theta))$$

- Résolution :
  - Analytique :  $\frac{\partial \mathcal{L}(D, \theta)}{\partial \theta} = 0$
  - Approchée : EM, gradient...

# Traitement des séquences

- Modèle du jour = **modèle de Markov**
- Entrée =  $\mathbf{x} = \{x_1, \dots, x_T\}$ 
  - Séquence d'observations, taille variable
  - Dépendance entre  $x_t$  et  $x_{t+1}$
  - Hypothèse i.i.d. entre les  $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle  $\lambda = \{\Pi, A\}$  [spécifique]
    - 1 modèle par classe [classique]
  - Catégorisation (non-supervisée)
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

# Traitement des séquences

- Modèle du jour = **modèle de Markov**
- Entrée =  $\mathbf{x} = \{x_1, \dots, x_T\}$ 
  - Séquence d'observations, taille variable
  - Dépendance entre  $x_t$  et  $x_{t+1}$
  - Hypothèse i.i.d. entre les  $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle  $\lambda = \{\Pi, A\}$  [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance :  $\log \mathcal{L} = \sum_i p(\mathbf{x}_i | \lambda)$  [classique]
  - Catégorisation (non-supervisée)
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

# Traitement des séquences

- Modèle du jour = **modèle de Markov**
- Entrée =  $\mathbf{x} = \{x_1, \dots, x_T\}$ 
  - Séquence d'observations, taille variable
  - Dépendance entre  $x_t$  et  $x_{t+1}$
  - Hypothèse i.i.d. entre les  $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle  $\lambda = \{\Pi, A\}$  [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance :  $\log \mathcal{L} = \sum_i p(\mathbf{x}_i | \lambda)$  [classique]
    - Apprentissage :  $\lambda^* = \text{Argmax}_{\lambda} \log \mathcal{L}$  [classique]
  - Catégorisation (non-supervisée)
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

# Traitement des séquences

- Modèle du jour = **modèle de Markov**
- Entrée =  $\mathbf{x} = \{x_1, \dots, x_T\}$ 
  - Séquence d'observations, taille variable
  - Dépendance entre  $x_t$  et  $x_{t+1}$
  - Hypothèse i.i.d. entre les  $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle  $\lambda = \{\Pi, A\}$  [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance :  $\log \mathcal{L} = \sum_i p(\mathbf{x}_i | \lambda)$  [classique]
    - Apprentissage :  $\lambda^* = \text{Argmax}_{\lambda} \log \mathcal{L}$  [classique]
    - Inférence :  $c^* = \text{Argmax}_c p(\mathbf{x}_i | \lambda_c^*)$  [classique]
  - Catégorisation (non-supervisée)
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

# Traitement des séquences

- Modèle du jour = **modèle de Markov**
- Entrée =  $\mathbf{x} = \{x_1, \dots, x_T\}$ 
  - Séquence d'observations, taille variable
  - Dépendance entre  $x_t$  et  $x_{t+1}$
  - Hypothèse i.i.d. entre les  $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle  $\lambda = \{\Pi, A\}$  [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance :  $\log \mathcal{L} = \sum_i p(\mathbf{x}_i | \lambda)$  [classique]
    - Apprentissage :  $\lambda^* = \text{Argmax}_{\lambda} \log \mathcal{L}$  [classique]
    - Inférence :  $c^* = \text{Argmax}_c p(\mathbf{x}_i | \lambda_c^*)$  [classique]
  - Catégorisation (non-supervisée)
    - ... idem + bouclage EM [classique]
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...

# Traitement des séquences

- Modèle du jour = **modèle de Markov**
- Entrée =  $\mathbf{x} = \{x_1, \dots, x_T\}$ 
  - Séquence d'observations, taille variable
  - Dépendance entre  $x_t$  et  $x_{t+1}$
  - Hypothèse i.i.d. entre les  $\mathbf{x}_i$
- Problématiques
  - Classification (supervisée)
    - Modèle  $\lambda = \{\Pi, A\}$  [spécifique]
    - 1 modèle par classe [classique]
    - Vraisemblance :  $\log \mathcal{L} = \sum_i p(\mathbf{x}_i | \lambda)$  [classique]
    - Apprentissage :  $\lambda^* = \text{Argmax}_{\lambda} \log \mathcal{L}$  [classique]
    - Inférence :  $c^* = \text{Argmax}_c p(\mathbf{x}_i | \lambda_c^*)$  [classique]
  - Catégorisation (non-supervisée)
    - ... idem + bouclage EM [classique]
  - Décodage : traduction, reconnaissance d'écriture, décodage génome, ...
    - ... prochaine séance, modèle plus compliqué

- Outil pour faire de la prévision dans des **espaces discrets**
- Chaîne de Markov d'ordre  $k$ 
  - Séquence de variables aléatoires  $\mathbf{x} = (x_1, \dots, x_T)$  qui prend ses valeurs dans un ensemble fini d'états  $Q = (q_1, \dots, q_N)$  et qui vérifie les propriétés dites de Markov :
    - Horizon de taille  $k$  :
  - Pour simplifier les notations, on se limite dans la suite à des chaînes d'ordre 1.
- Exemple : météo sur un an (soleil, nuage, pluie)
  - $Q = [\text{So}, \text{Nu}, \text{Pl}]$
  - $\mathbf{x} = [x_1 = \text{Nu}, x_2 = \text{So}, \dots, x_{365} = \text{Pl}]$

Soit un exemple très simple :

- $Q = [\textcolor{orange}{So}, \textcolor{gray}{Nu}, \textcolor{blue}{Pi}]$
- $\mathbf{x} = [x_1 = \textcolor{gray}{Nu}, x_2 = \textcolor{orange}{So}, \dots, x_{365} = \textcolor{blue}{Pi}]$
- Combien de paramètres pour une chaîne d'ordre 1 ?

Soit un exemple très simple :

- $Q = [\textcolor{orange}{So}, \textcolor{gray}{Nu}, \textcolor{blue}{Pi}]$
- $\mathbf{x} = [x_1 = \textcolor{gray}{Nu}, x_2 = \textcolor{orange}{So}, \dots, x_{365} = \textcolor{blue}{Pi}]$
- Combien de paramètres pour une chaîne d'ordre 1 ?
- Combien de paramètres pour une chaîne d'ordre 2 ?

9 (+3)

27 (+3)

- Une chaîne de Markov d'ordre 1 est entièrement spécifiée par la donnée :

- d'une matrice de transition

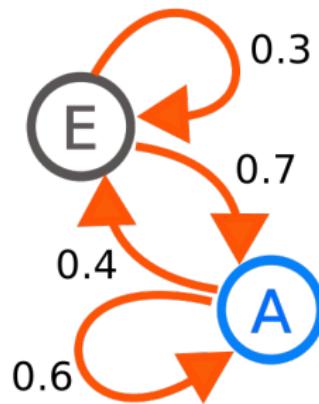
$$A = [a_{ij} = p(x_{t+1} = q_j | x_t = q_i)]$$

- et des probabilités initiales :

$$\Pi = [\pi_i = p(x_1 = q_i)]$$

- Probabilité d'une séquence

$$p(\mathbf{x}|\lambda) = p(x_1, \dots, x_T|\lambda)$$



# Hypothèse markovienne d'ordre 1

Décomposition du calcul

$$\begin{aligned} p(\mathbf{x}|\lambda) &= p(x_1, \dots, x_T|\lambda) \\ &= p(x_T|x_1, \dots, x_{T-1}, \lambda) \times p(x_1, \dots, x_{T-1}|\lambda) \\ &= p(x_T|x_1, \dots, x_{T-1}, \lambda) \times p(x_{T-1}|x_1, \dots, x_{T-2}, \lambda) \dots \\ &\quad \times p(x_1, \dots, x_{T-2}|\lambda) \\ &= \prod_{t=2}^T p(x_t|x_1, \dots, x_{t-1}, \lambda) p(x_1|\lambda) \end{aligned}$$

Après hypothèse d'ordre 1 :

$$\begin{aligned} p(\mathbf{x}|\lambda) &= \prod_{t=2}^T p(x_t|x_1, \dots, x_{t-1}, \lambda) p(x_1|\lambda) = \prod_{t=2}^T p(x_t|x_{t-1}, \lambda) p(x_1|\lambda) \\ &= \pi_{x_1} \prod_{t=2}^T a_{x_{t-1}, x_t} \end{aligned}$$

- Matrice de transition  $\mathbf{A}$  de taille  $N \times N$ 
  - Lignes  $\mathbf{x}_t$ , colonnes  $\mathbf{x}_{t+1}$
  - Matrice stochastique, chaque ligne somme à 1 :  $\sum_{j=1}^N a_{ij} = 1$

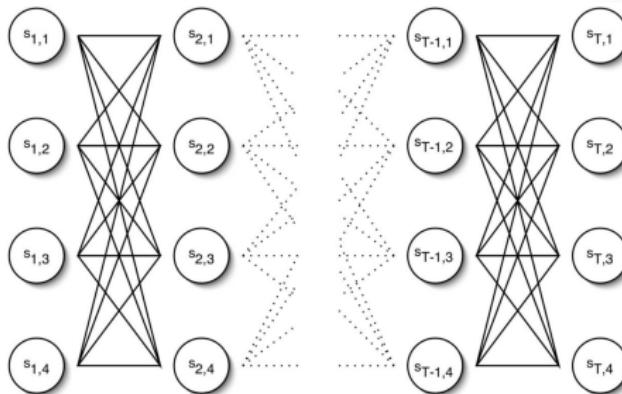
- Matrice de transition  $\mathbf{A}$  de taille  $N \times N$ 
  - Lignes  $\mathbf{x}_t$ , colonnes  $\mathbf{x}_{t+1}$
  - Matrice stochastique, chaque ligne somme à 1
- $\mathbf{p}_t := p(x_t = q_i | x_1, \dots, x_{t-1}, \lambda), \forall i \in \{1; N\}$ 
  - $\mathbf{p}_t$  vecteur ligne, taille  $1 \times N$
- $\mathbf{p}_{t+1} := p(x_{t+1} = q_i | x_1, \dots, x_t, \lambda), \forall i \in \{1; N\}$ 
  - $\mathbf{p}_{t+1}$  vecteur ligne, taille  $1 \times N$

# Représentation matricielle

- Matrice de transition  $\mathbf{A}$  de taille  $N \times N$
- $\mathbf{p}_t / \mathbf{p}_{t+1}$  vecteurs de probabilité à  $t / t + 1$
- Treillis pour calculer le vecteur de probabilité  $\mathbf{p}_{t+1} \forall j \in \{1; N\}$

$$\bullet p(x_{t+1} = q_j) = \sum_{i=1}^N p(x_t = q_i)p(x_{t+1} = q_j | x_t = q_i) = \sum_{i=1}^N p(x_t = q_i) a_{ij}$$

$$\mathbf{p}_{t+1} = \mathbf{p}_t \times \mathbf{A}$$



# Représentation matricielle : exemple

Il fait soleil... Quel temps fera-t-il dans  $N$  jours ? (distribution de probabilités)

Il fait soleil... Quel temps fera-t-il dans  $N$  jours ? (distribution de probabilités)

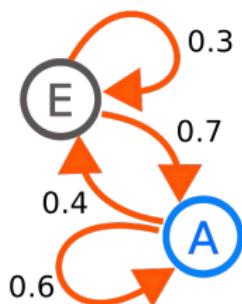
- ➊ Je rentre sur la ligne *soleil*...  $x_0 = S$
- ➋ A  $t = 1$ ,  $\{a_{x\cdot}\}$  me donne la distribution des probabilités des états  
 $p(x_1 = q_i) = a_{S,i} \Rightarrow \mathbf{p}_1 = \{a_{S,i}\}_{i \in \{1;N\}}$

Il fait soleil... Quel temps fera-t-il dans  $N$  jours ? (distribution de probabilités)

- ➊ Je rentre sur la ligne *soleil*...  $x_0 = S$
- ➋ A  $t = 1$ ,  $\{a_{x_i}\}$  me donne la distribution des probabilités des états  
 $p(x_1 = q_i) = a_{S,i} \Rightarrow \mathbf{p}_1 = \{a_{S,i}\}_{i \in \{1;N\}}$
- ➌ Ensuite on itère :
  - $\mathbf{p}_2 = \mathbf{p}_1 \times \mathbf{A}$
  - $\mathbf{p}_N = \mathbf{p}_1 \times \mathbf{A}^{N-1}$

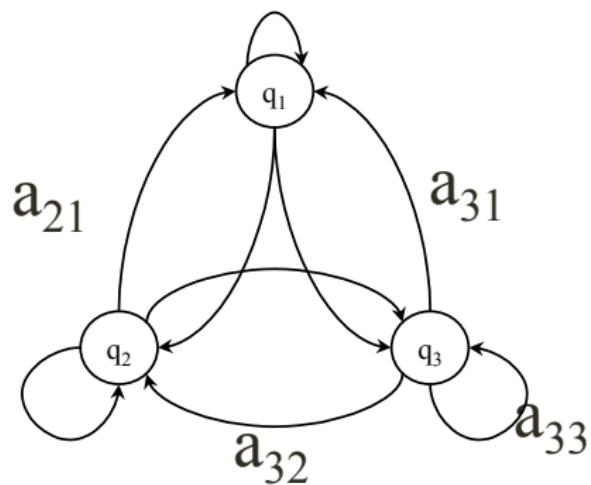
# Représentation graphique

Automate basique :



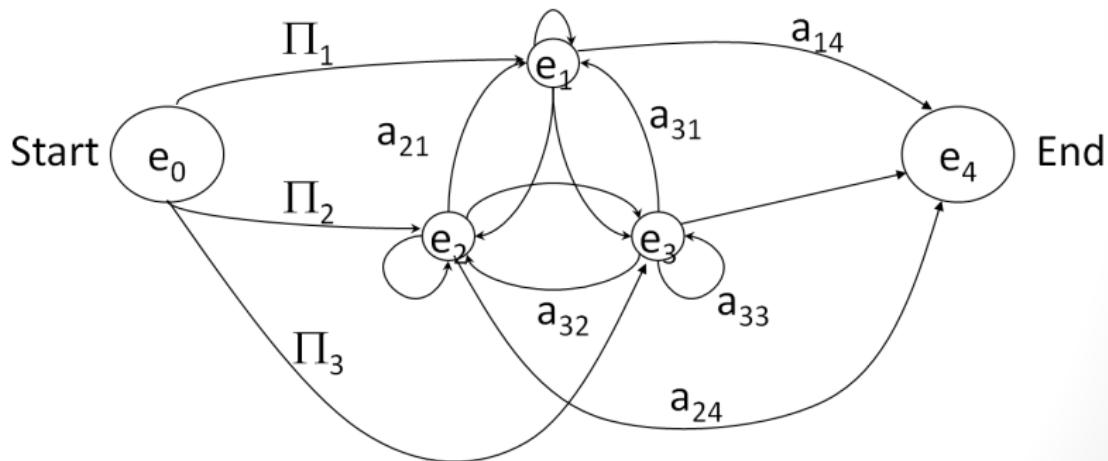
# Représentation graphique

Comment introduire la notion de commencement ?



# Représentation graphique

Avec des noeuds identifiés de début/fin



---

**Algorithm 1:** Génération d'une séquence  $\mathbf{x}$ 

**Data:**  $A, \Pi$

**Result:**  $\mathbf{x}$

$\mathbf{x} \leftarrow [];$

Tirer  $x_1$  en fonction de  $\Pi$ ;

$x_t \leftarrow x_1, t \leftarrow 1;$

$\mathbf{x} \leftarrow [\mathbf{x}, x_{courant}]$ ;

**while**  $x_t$  n'est pas l'état final **do**

$x_{t+1} \leftarrow$  tirage selon  $(A(x_t, :))$ ;

$t \leftarrow t + 1$ ;

---

- Plusieurs variantes dans la clause du *while*
- Comment effectuer un tirage selon une loi de probabilité discrète ?

Soit la loi :

$A$	1	2	3
$P(A)$	0.3	0.2	0.5

Comment effectuer un tirage selon  $P(A)$  ?

- ➊ Faire la somme cumulée de la loi

$A$	1	2	3
<i>cumsum</i>	0.3	0.5	1

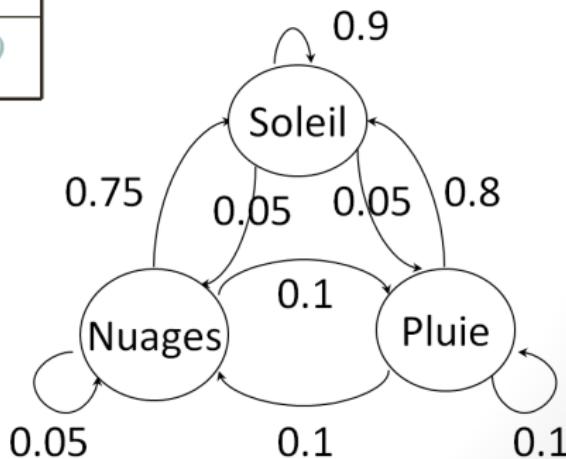
- ➋ Tirer un nombre  $x$  entre 0 et 1 selon la loi uniforme
- ➌ Initialiser  $vx = 1$
- ➍ Tant que  $cumsum[vx] < x$ 
  - ➎  $vx++$

# Génération (sur un exemple)

- $q_1 = \text{Pluie}$ ,  $q_2 = \text{Nuages}$ ,  $q_3 = \text{Soleil}$

- $A =$

0.1	0.1	0.8
0.1	0.15	0.75
0.05	0.05	0.9



PSSSSSSSSSSSSNSSSSSSSSSSNSSSSSSNSSSSNSS

SSSSSSSSSSSSSSPSSSSSNPSSSSSPNSSSSSNNSSS

SSSSPSSSSSSSSSSPSSSSSSSSSSSSSSPSSSSSSSSSS

- Quelle est la probabilité d'observer une séquence de soleil de longueur  $d$ ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

- Quelle est la probabilité d'observer une séquence de soleil de longueur  $d$ ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la probabilité d'observer une séquence de soleil de longueur  $d$  (étant donné que l'on se trouve au premier jour de soleil) ?

*Loi géométrique*

Notons la longueur de la sous-séquence de soleil  $D_S$ ,

$$P(D_S = d) = a_{ss}^{d-1}(1 - a_{ss})$$

- Quelle est la probabilité d'observer une séquence de soleil de longueur  $d$ ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la longueur moyenne d'une séquence de soleil ?

*Espérance de la loi géométrique*

$$E[D_S] = \sum_{d=1}^{\infty} d a_{ss}^{d-1} (1 - a_{ss}) = \frac{1}{1 - a_{ss}}$$

- Quelle est la probabilité d'observer une séquence de soleil de longueur  $d$ ? [en se trouvant au premier jour de soleil]
- Quelle est la durée moyenne d'une séquence consécutive de soleil ?

Quelle est la longueur moyenne d'une séquence de soleil ?

*Espérance de la loi géométrique*

$$E[D_S] = \sum_{d=1}^{\infty} d a_{ss}^{d-1} (1 - a_{ss}) = \frac{1}{1 - a_{ss}}$$

Sketch of proof (wikipedia) avec  $k = d - 1$  et  $p = 1 - a_{ss}$  :

$$\begin{aligned} E(Y) &= \sum_{k=0}^{\infty} (1-p)^k p \cdot (k+1) \\ &= p \sum_{k=0}^{\infty} (1-p)^k (k+1) \\ &= p \left[ \frac{d}{dp} \left( -(1-p) \sum_{k=0}^{\infty} (1-p)^k \right) \right] \\ &= -p \frac{d}{dp} \frac{1-p}{p} = \frac{1}{p}. \end{aligned}$$

- **Stationnarité** : existe-t-il une mesure stationnaire  $\mu$  t ?

$$\mu = \mu A$$

- $\mu$  = **pondération stationnaire** (inchangée après une transition)
- si  $\forall i, \mu_i \geq 0, \sum_i \mu_i = 1$  :  $\mu$  est alors une **distribution stationnaire**
- si  $A$  est **irréductible**,  $\mu$  est unique et :  $\mu = \text{distribution moyenne des états}$

$$\mu = \mu \mathbf{A}$$

- $\mathbf{A}$  matrice stochastique, vecteur  $(1, \dots, 1)$  vecteur propre avec valeur propre 1
- Stationnarité :  $\mu^T = \mathbf{A}^T \mu^T$
- Si une mesure stationnaire existe, solution donnée par le vecteur propre de  $\mathbf{A}^T$  de valeur propre 1

## CM irreducible finie

- CM irréductible : partant de chaque état, la probabilité d'y revenir est non nulle
- CM irréductible finie : on y revient en un nombre moyen d'étapes fini
- ie : le graphe est fortement connexe, pas d'état final/absorbant

## Périodicité

- Un état est dit périodique de période  $k$  ( $k > 1$ ), si on ne peut y revenir (après l'avoir quitté) qu'en un nombre d'étapes multiples de  $k$ .
- La période d'une CM est définie comme le PGCD de la période de tous ses états.
- La période d'une CM est égale au PGCD de la longueur de tous les circuits (élémentaires) du graphe associé.
- Une CM est dite *apériodique* si sa période est égale à 1.

Un processus ergodique est un processus stochastique pour lequel les statistiques peuvent être approchées par l'étude d'une seule réalisation suffisamment longue.

- Soit  $\pi$  une distribution de probabilité sur les états.
- Une chaîne de Markov est ergodique si  $\pi_n$  converge, indépendamment de  $\pi_0$ .
- $\pi_n$  converge alors vers  $\pi^*$ , la distribution stationnaire.

## Théorème

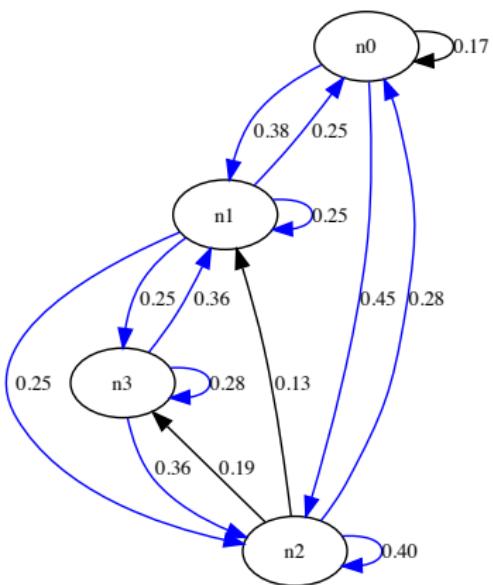
Les chaînes **irréductible** et **apériodique** sont **ergodiques**

- Des sous-séries d'observations sont-elles récurrentes dans une CM ?

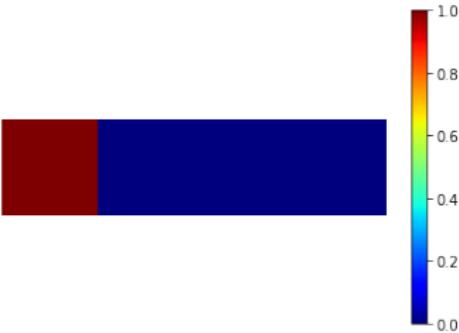
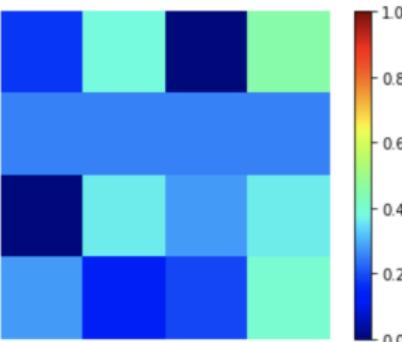
## Périodicité

- Un état est dit périodique de période  $k$  ( $k > 1$ ), si on ne peut y revenir (après l'avoir quitté) qu'en un nombre d'étapes multiples de  $k$ .
- La période d'une CM est définie comme le PGCD de la période de tous ses états.
- La période d'une CM est égale au PGCD de la longueur de tous les circuits (élémentaires) du graphe associé.
- Une CM est dite *apériodique* si sa période est égale à 1.

# Exemples & discussion (1)

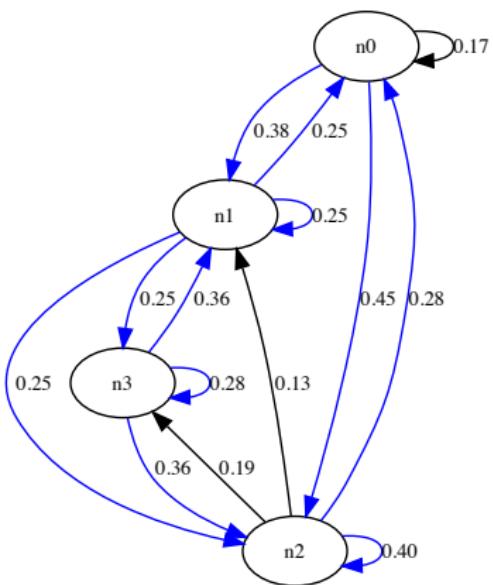


Matrice de transition :

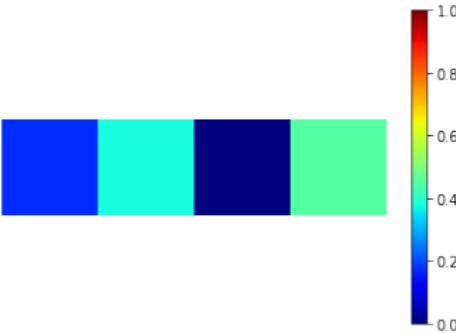
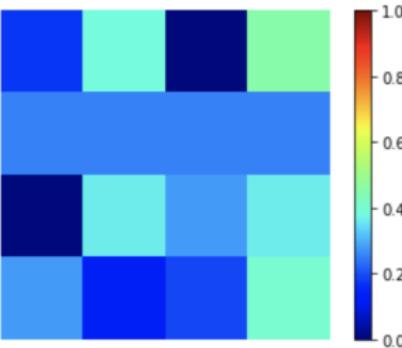


$$t = 0$$

# Exemples & discussion (1)

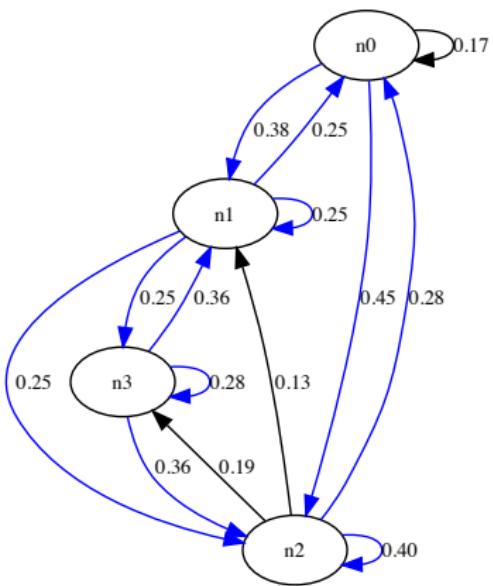


Matrice de transition :

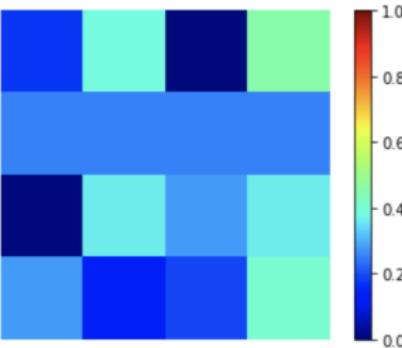


$t = 1$

# Exemples & discussion (1)

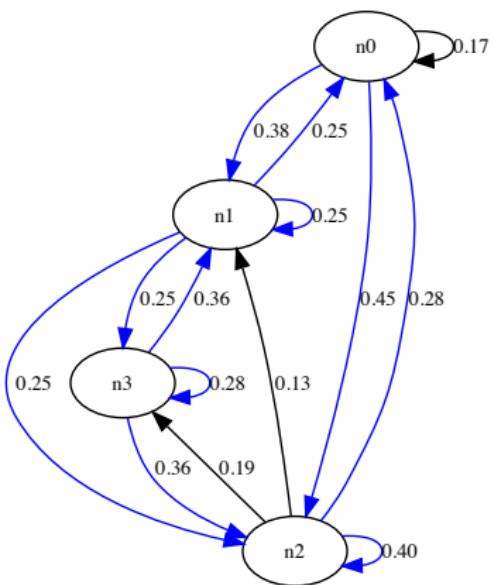


Matrice de transition :

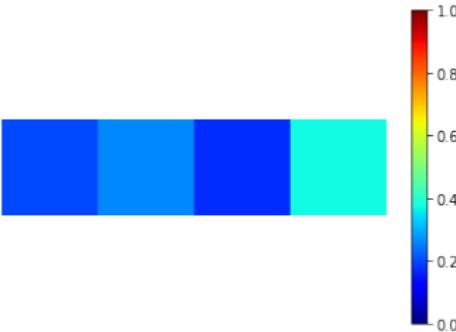
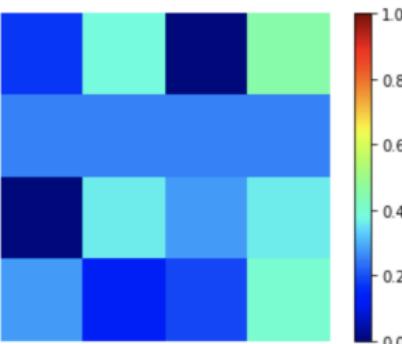


$t = 2$

# Exemples & discussion (1)

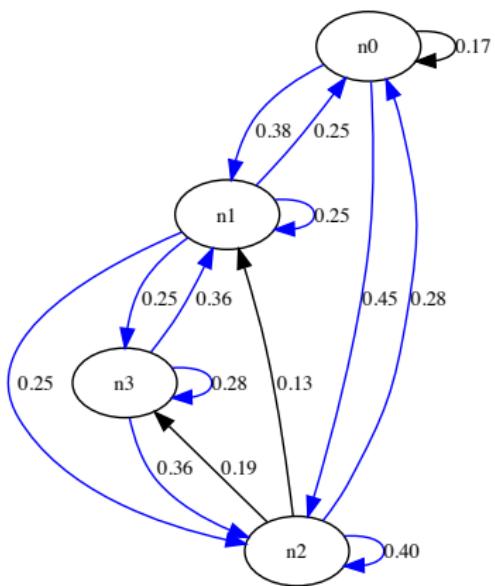


Matrice de transition :

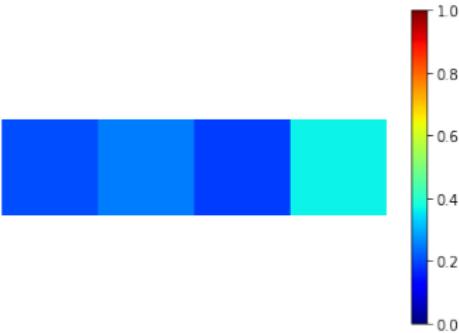
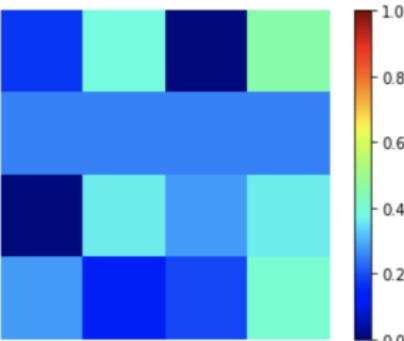


$t = 3$

# Exemples & discussion (1)

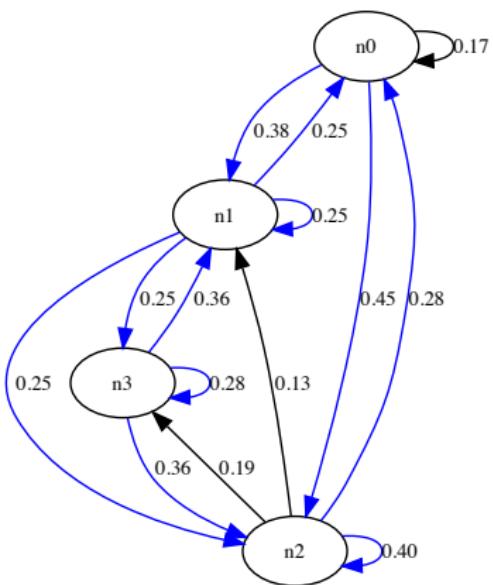


Matrice de transition :

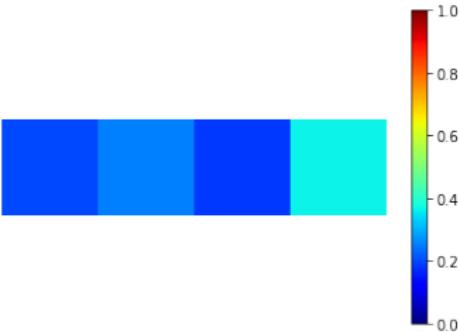
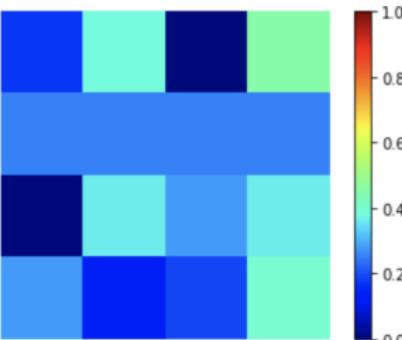


$t = 4$

# Exemples & discussion (1)



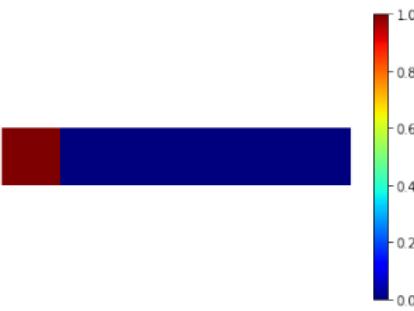
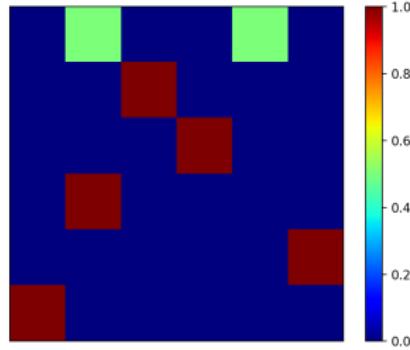
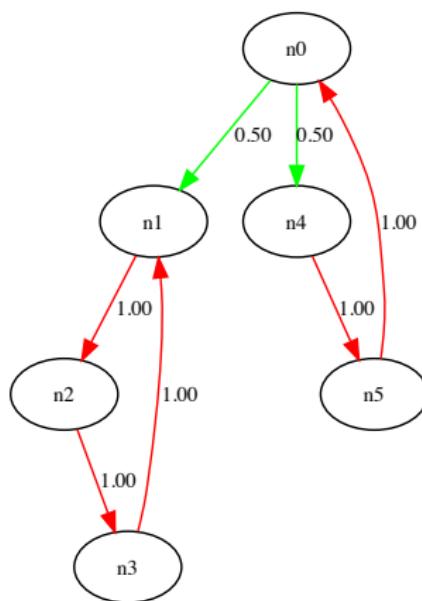
Matrice de transition :



$t = 5$

# Exemples & discussion (2)

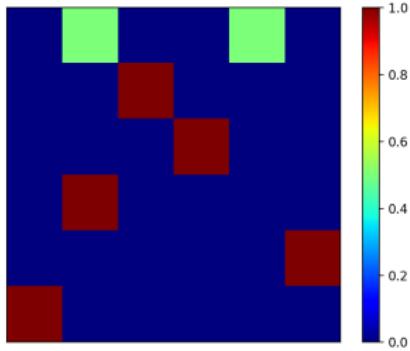
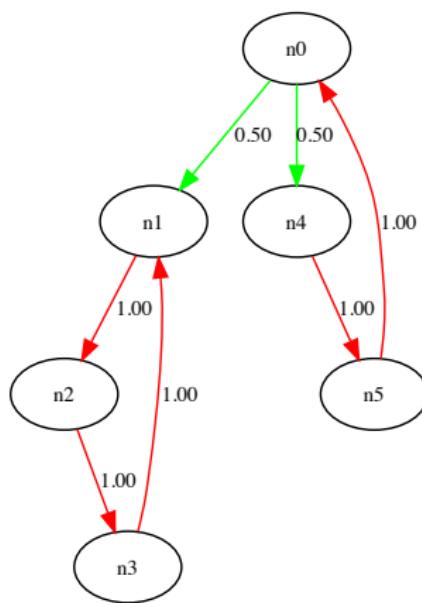
Matrice de transition :



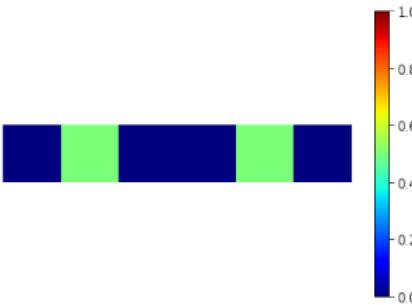
$t = 0$

# Exemples & discussion (2)

Matrice de transition :

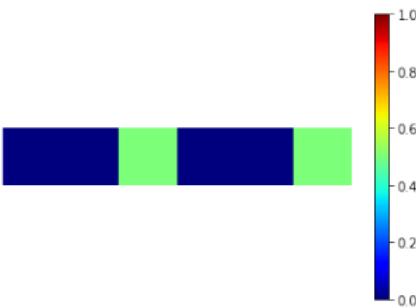
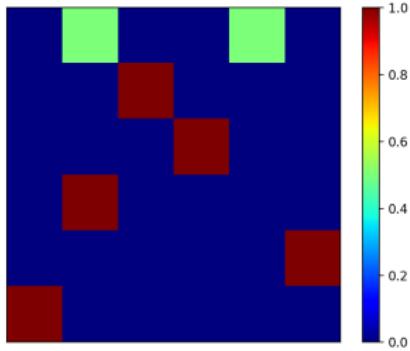
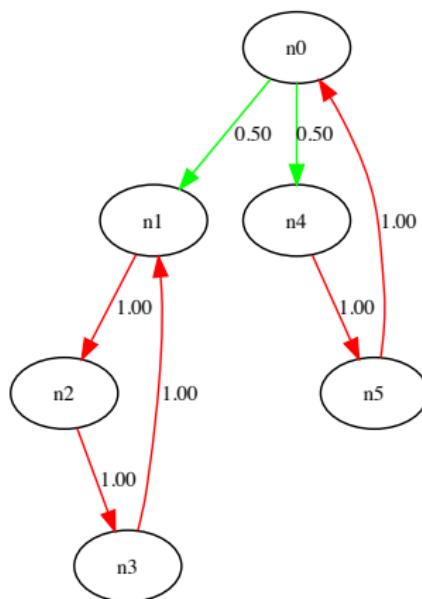


$t = 1$



# Exemples & discussion (2)

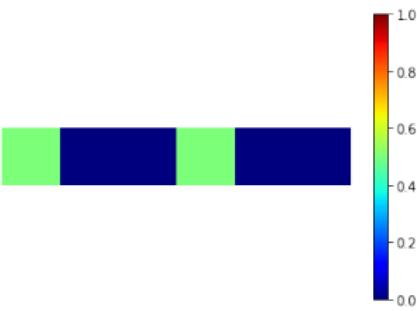
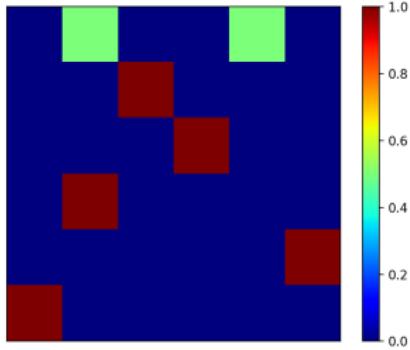
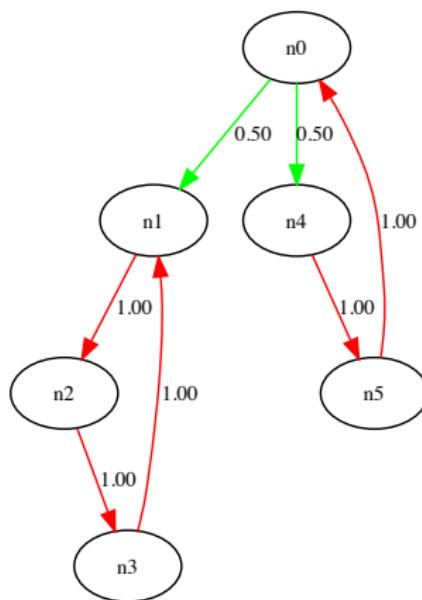
Matrice de transition :



$t = 2$

# Exemples & discussion (2)

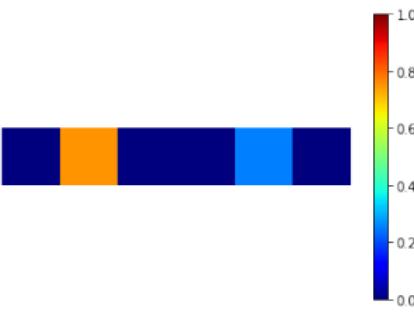
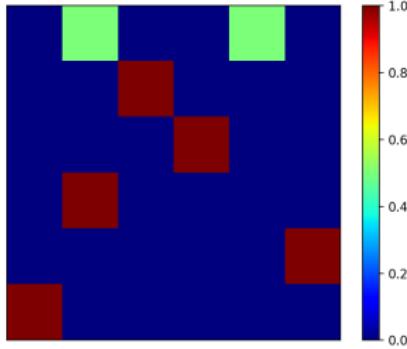
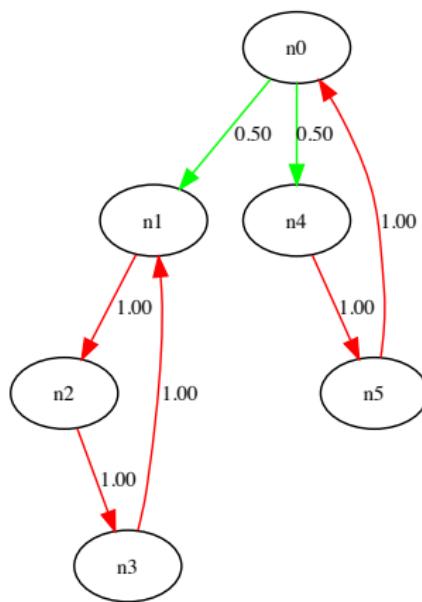
Matrice de transition :



$t = 3$

# Exemples & discussion (2)

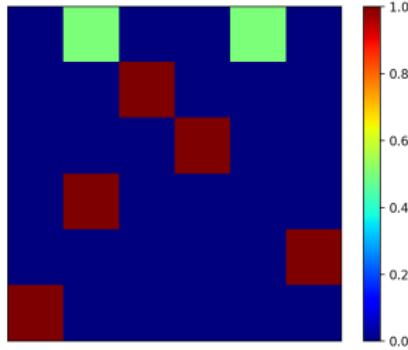
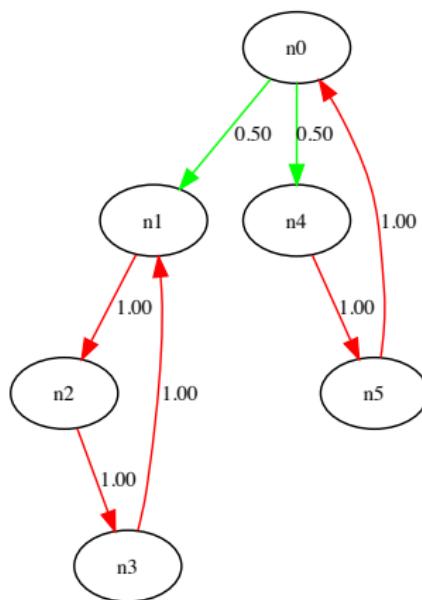
Matrice de transition :



$t = 4$

# Exemples & discussion (2)

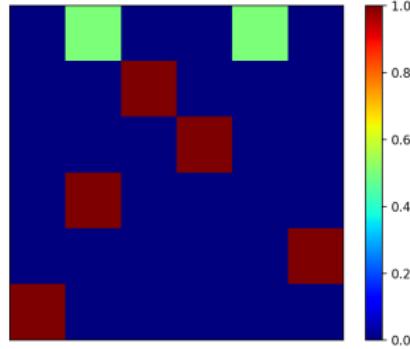
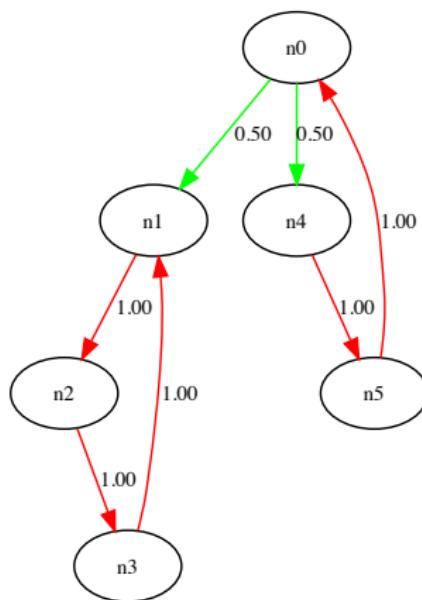
Matrice de transition :



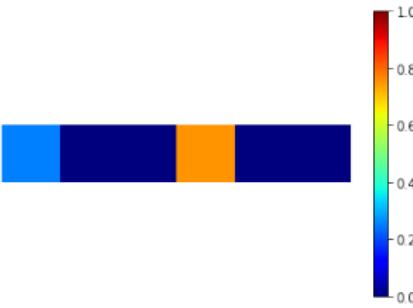
$t = 5$

# Exemples & discussion (2)

Matrice de transition :

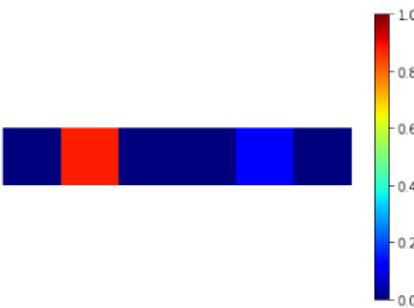
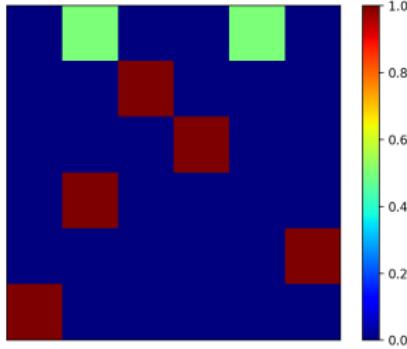
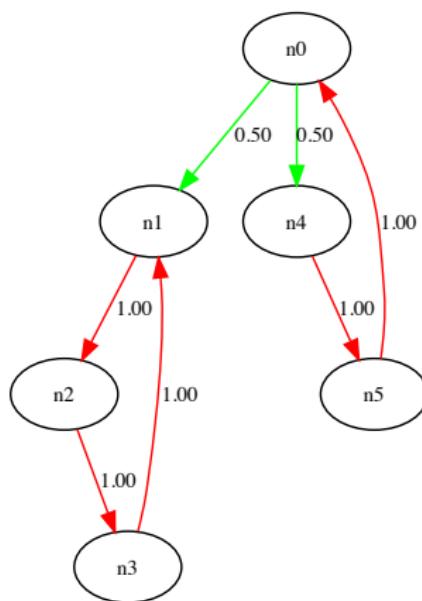


$t = 6$



# Exemples & discussion (2)

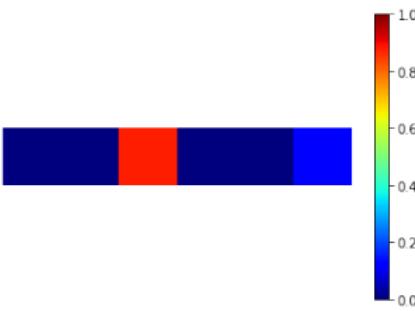
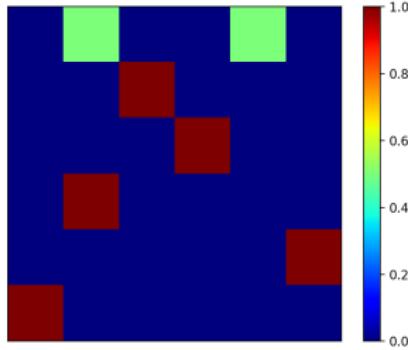
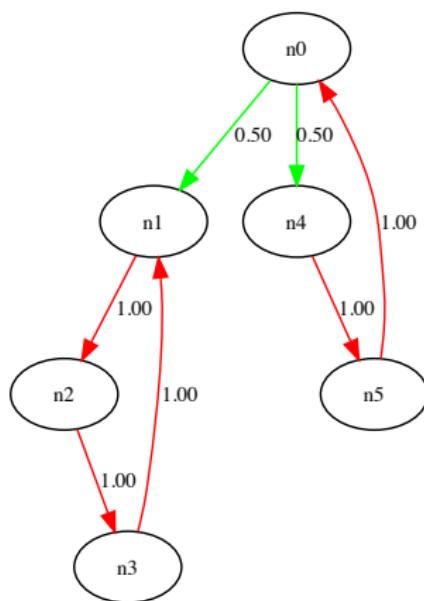
Matrice de transition :



$t = 7$

# Exemples & discussion (2)

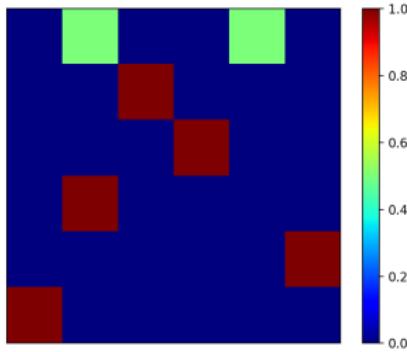
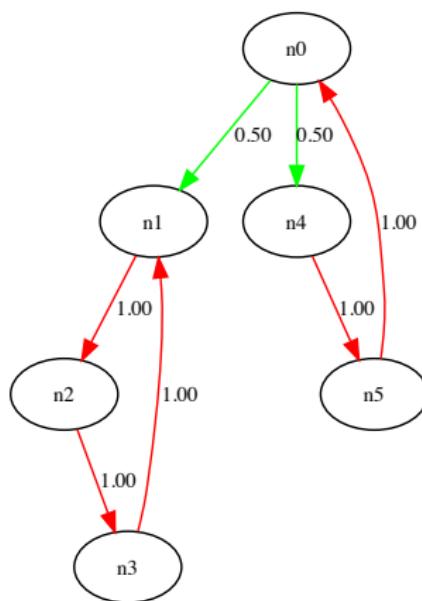
Matrice de transition :



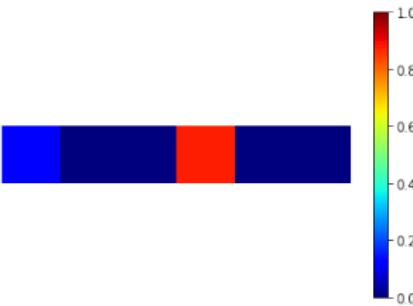
$t = 8$

# Exemples & discussion (2)

Matrice de transition :

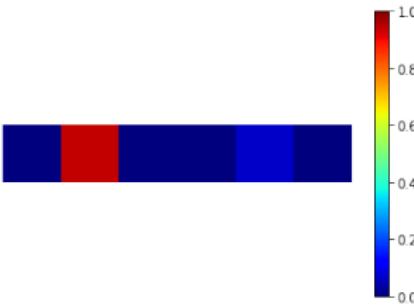
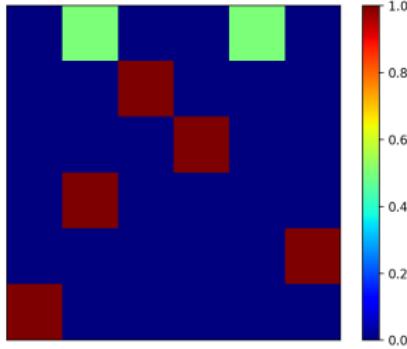
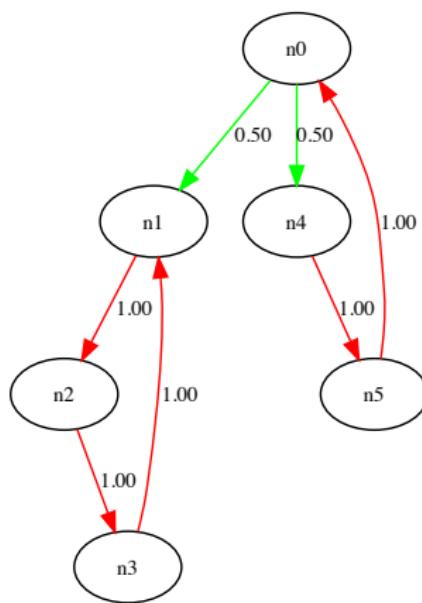


$t = 9$



# Exemples & discussion (2)

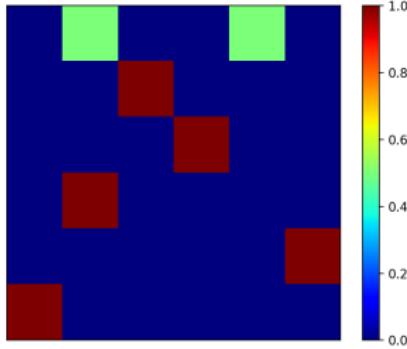
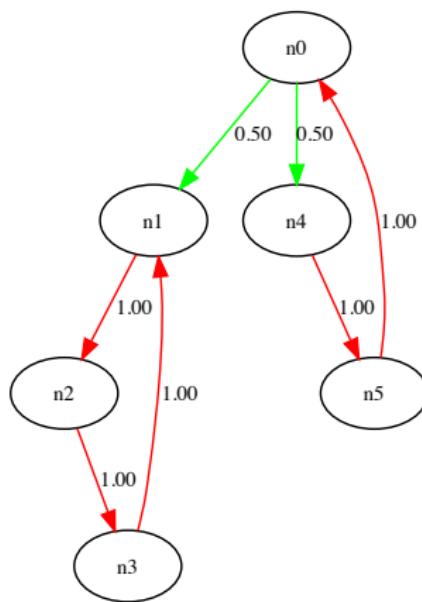
Matrice de transition :



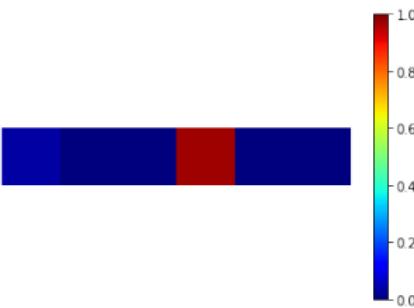
$t = 10$

# Exemples & discussion (2)

Matrice de transition :

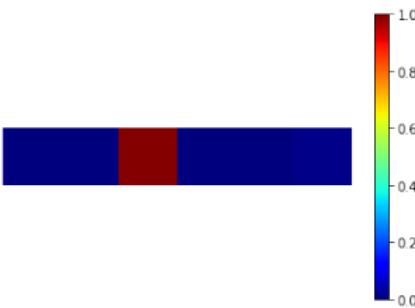
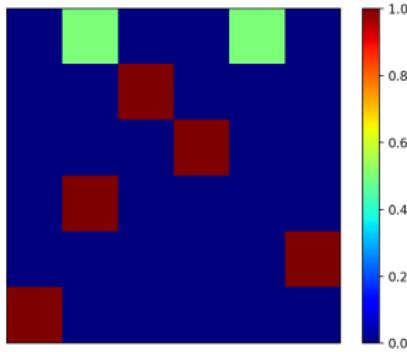
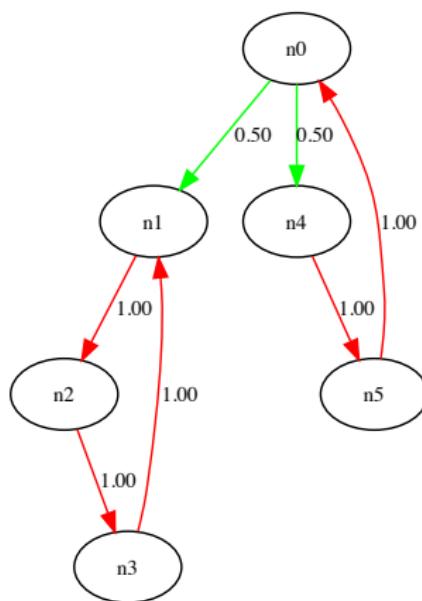


$t = 15$



# Exemples & discussion (2)

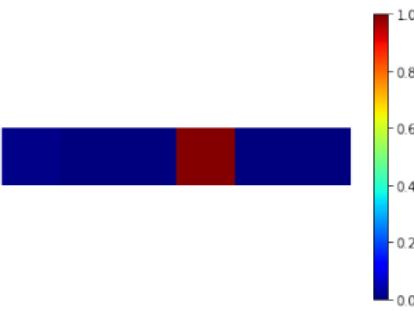
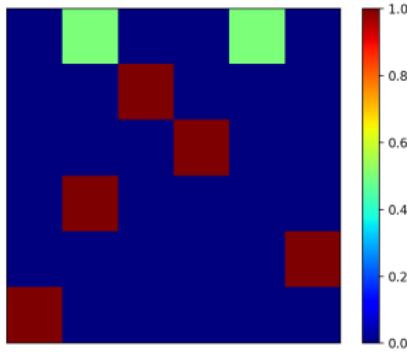
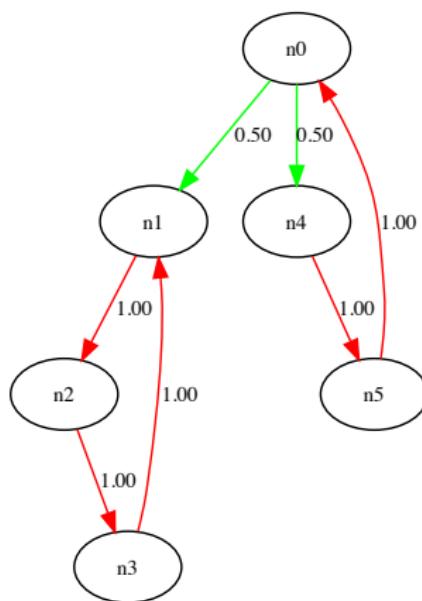
Matrice de transition :



$t = 20$

# Exemples & discussion (2)

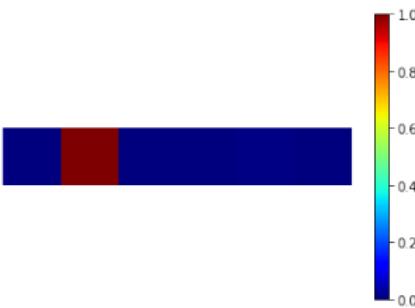
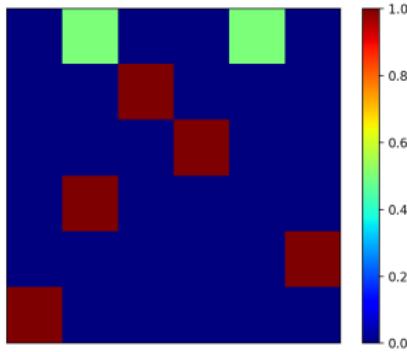
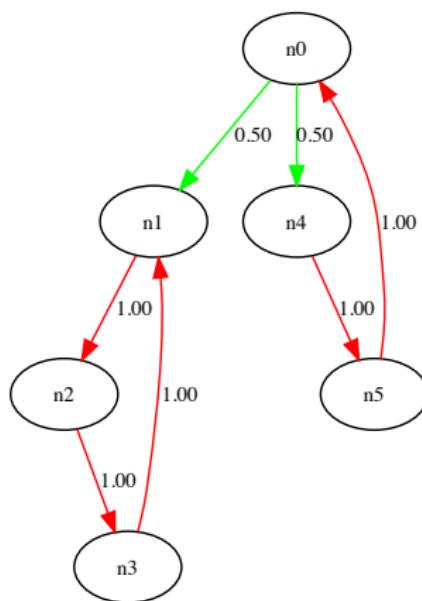
Matrice de transition :



$t = 30$

# Exemples & discussion (2)

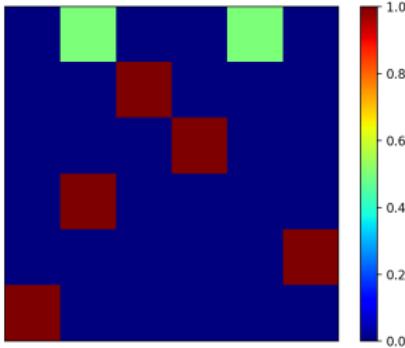
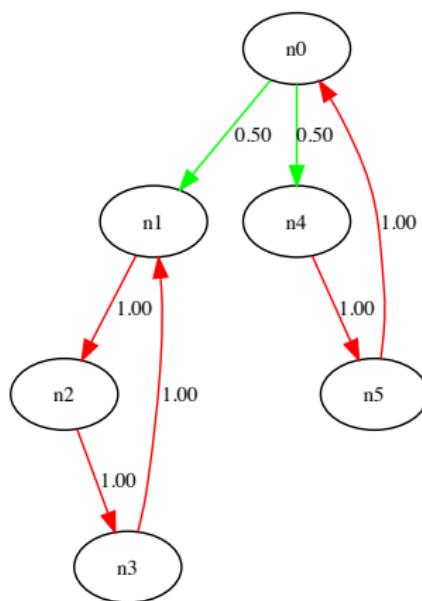
Matrice de transition :



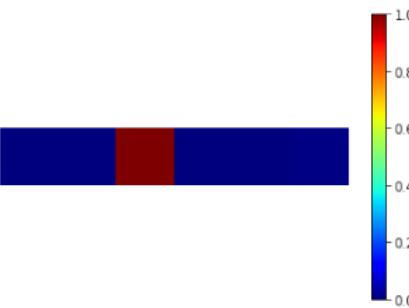
$t = 50$

# Exemples & discussion (2)

Matrice de transition :

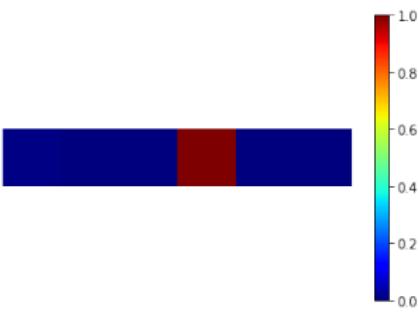
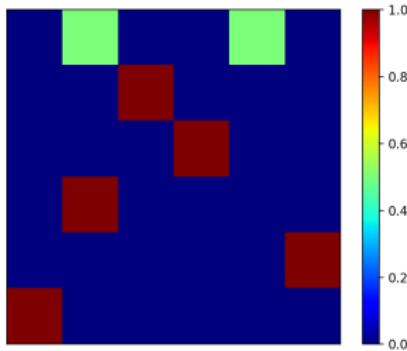
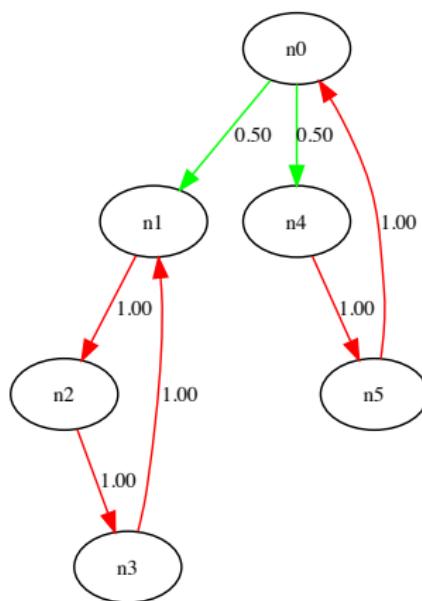


$t = 50$



# Exemples & discussion (2)

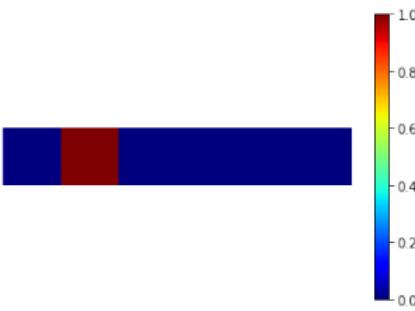
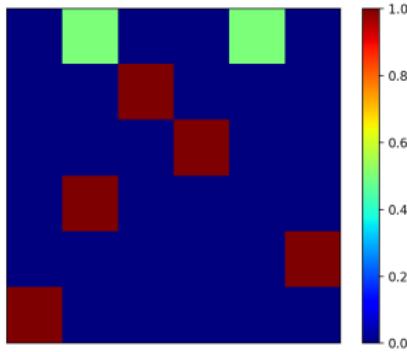
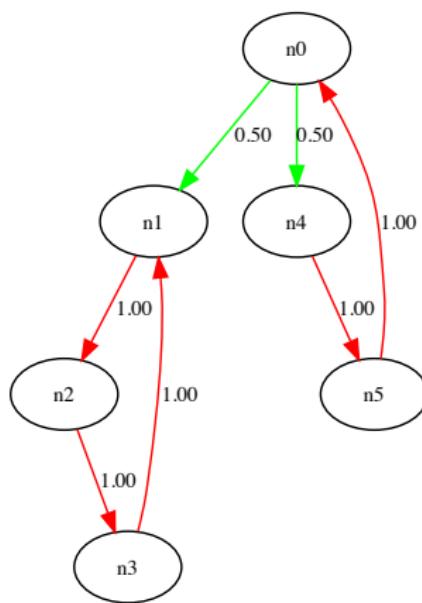
Matrice de transition :



$t = 60$

# Exemples & discussion (2)

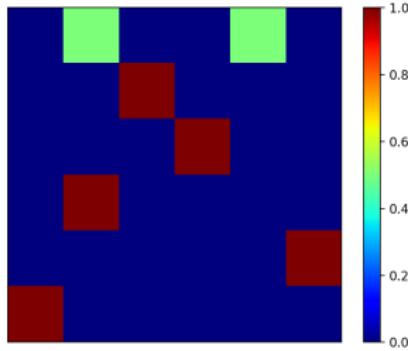
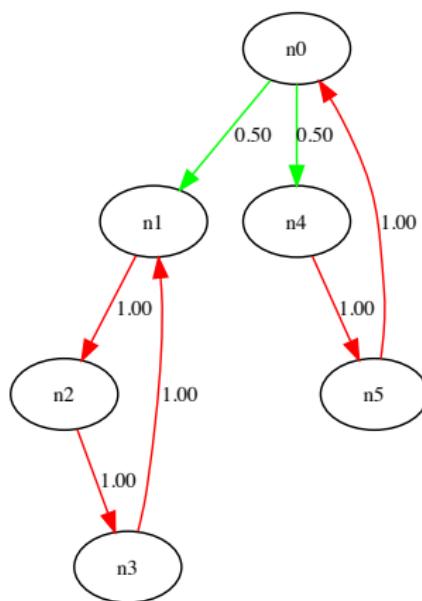
Matrice de transition :



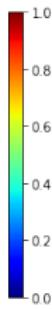
$t = 70$

# Exemples & discussion (2)

Matrice de transition :

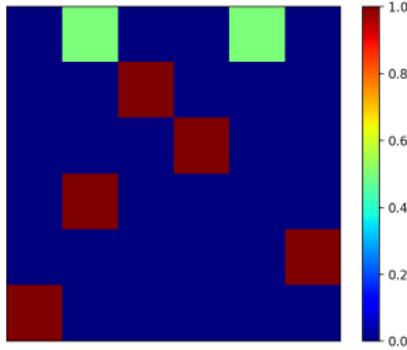
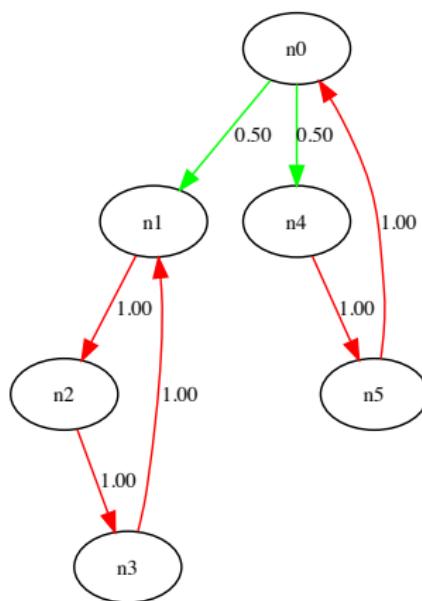


$t = 80$



# Exemples & discussion (2)

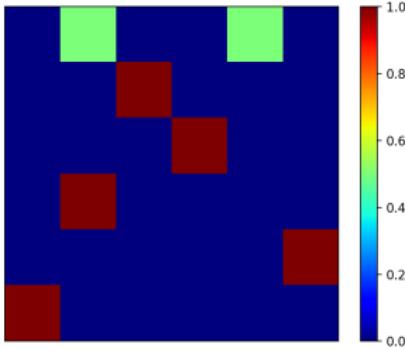
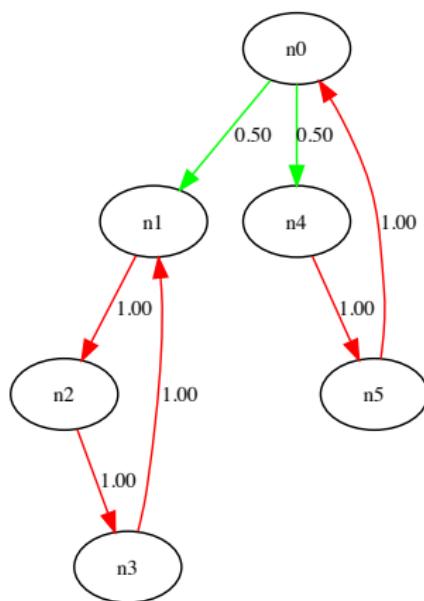
Matrice de transition :



$t = 90$

# Exemples & discussion (2)

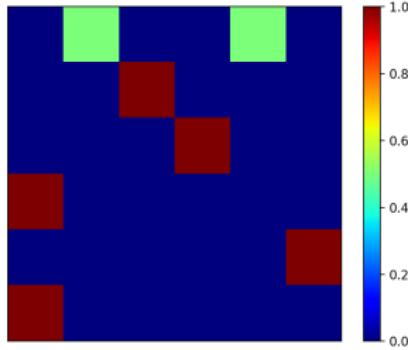
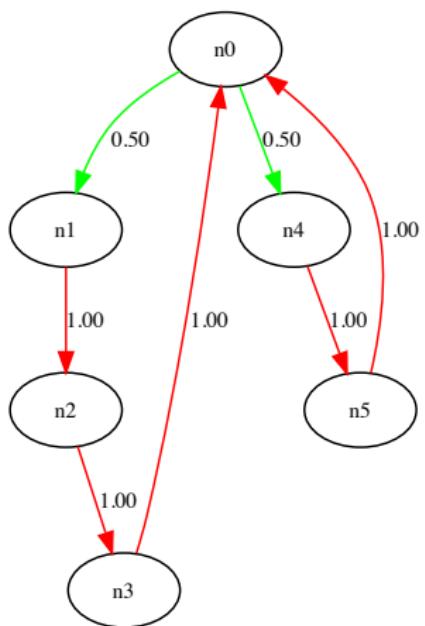
Matrice de transition :



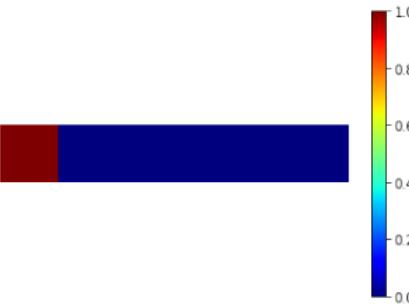
$t = 100$

# Exemples & discussion (3)

Matrice de transition :

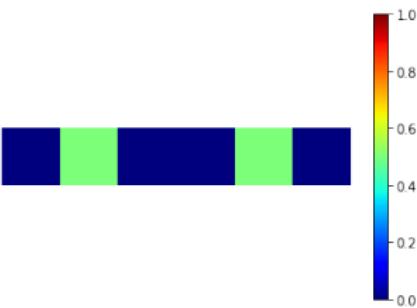
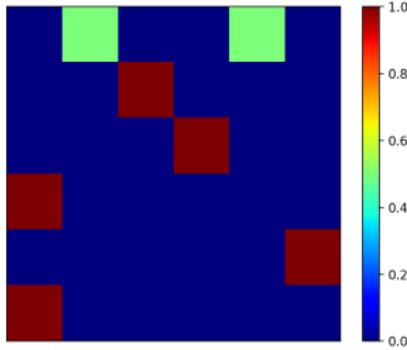
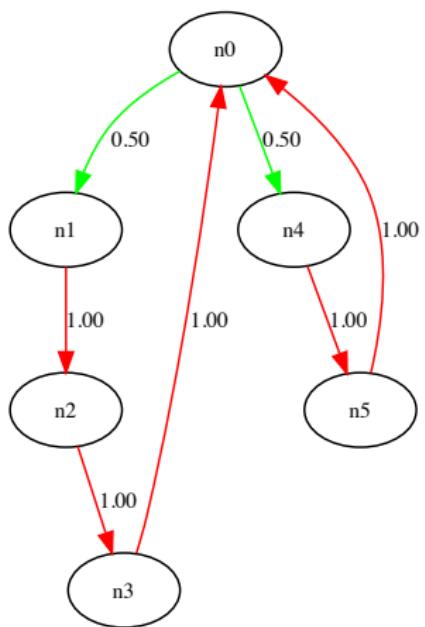


$t = 0$



# Exemples & discussion (3)

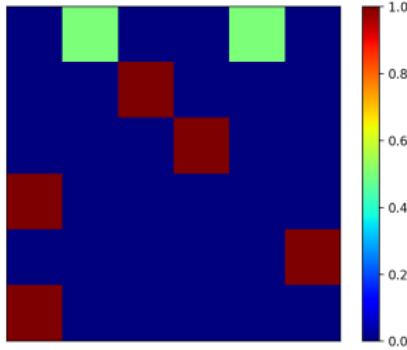
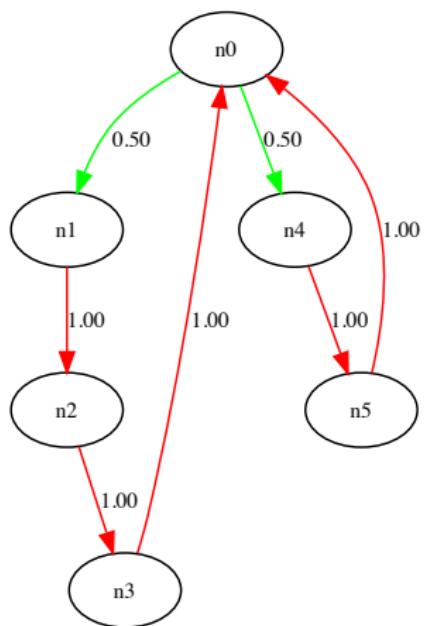
Matrice de transition :



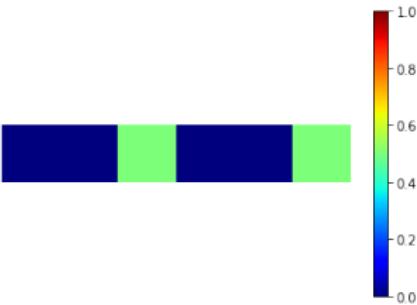
$t = 1$

# Exemples & discussion (3)

Matrice de transition :

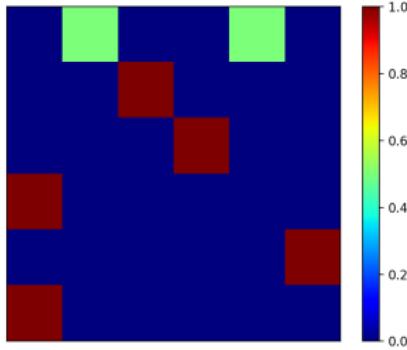
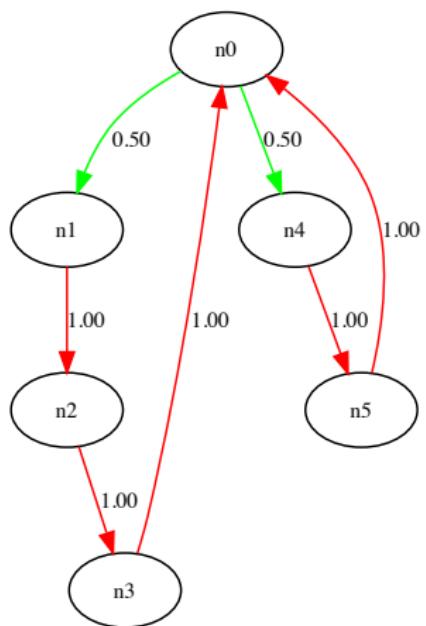


$t = 2$

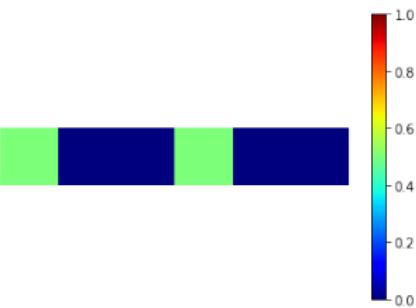


# Exemples & discussion (3)

Matrice de transition :

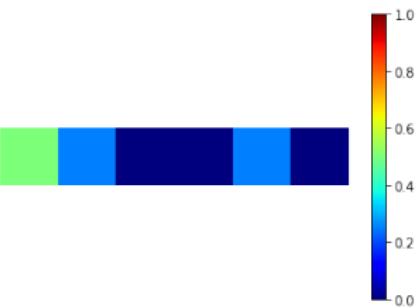
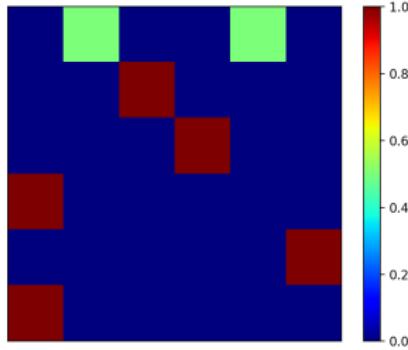
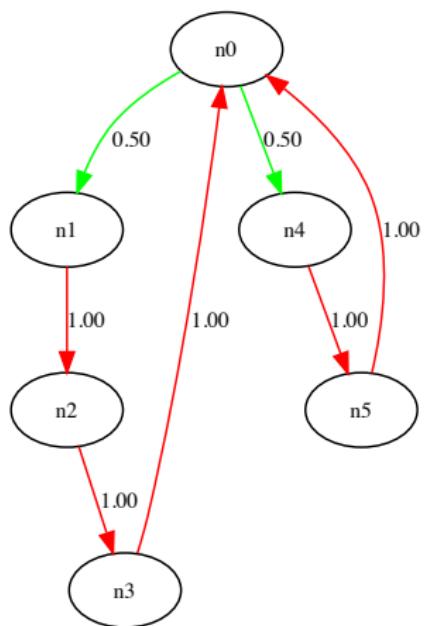


$t = 3$



# Exemples & discussion (3)

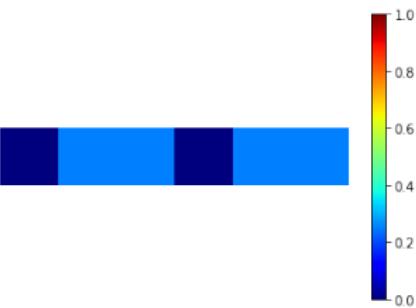
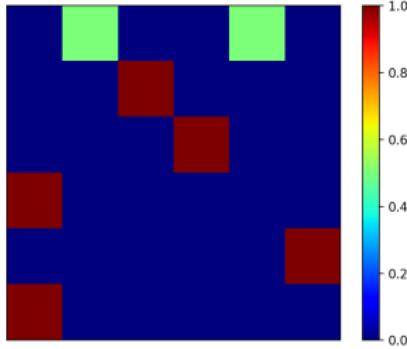
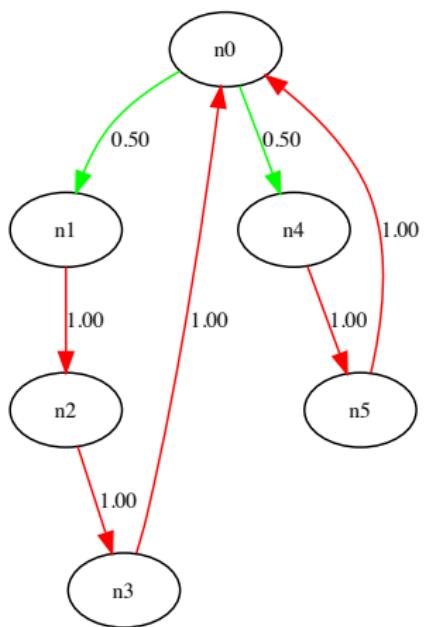
Matrice de transition :



$t = 4$

# Exemples & discussion (3)

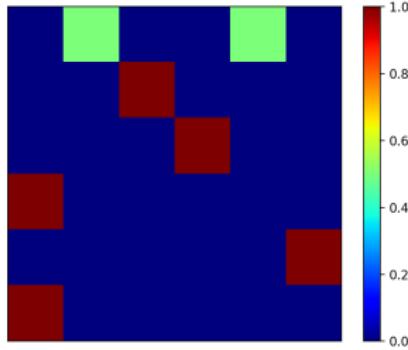
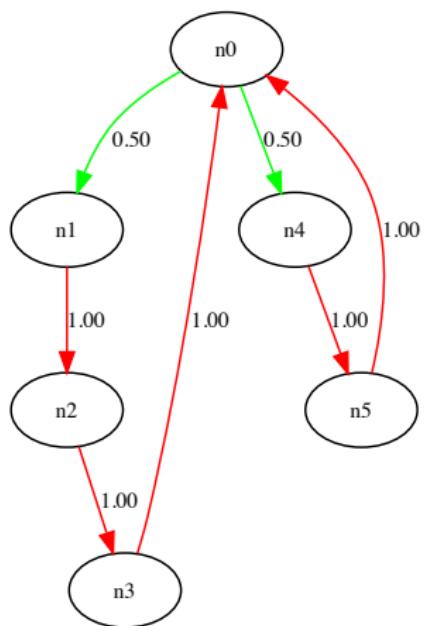
Matrice de transition :



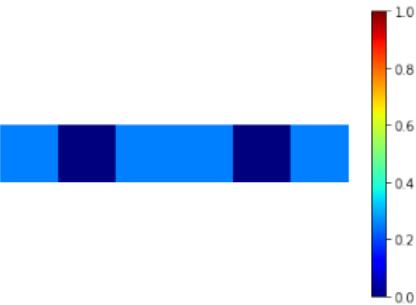
$t = 5$

# Exemples & discussion (3)

Matrice de transition :

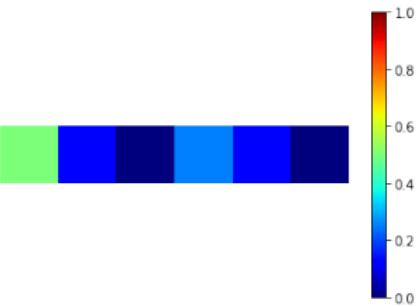
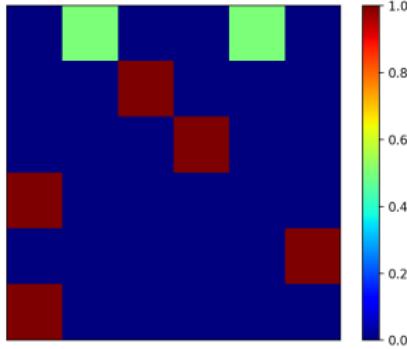
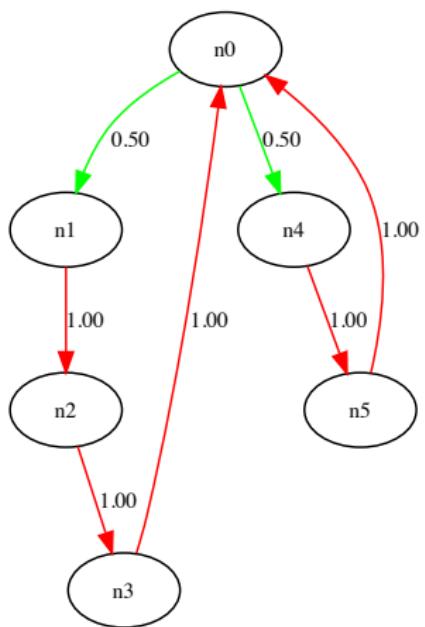


$t = 6$



# Exemples & discussion (3)

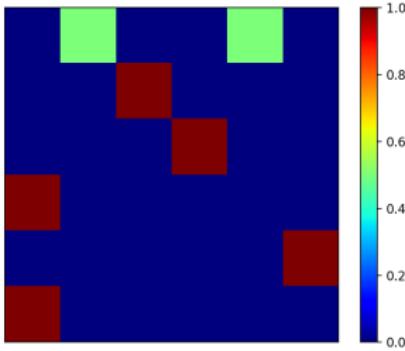
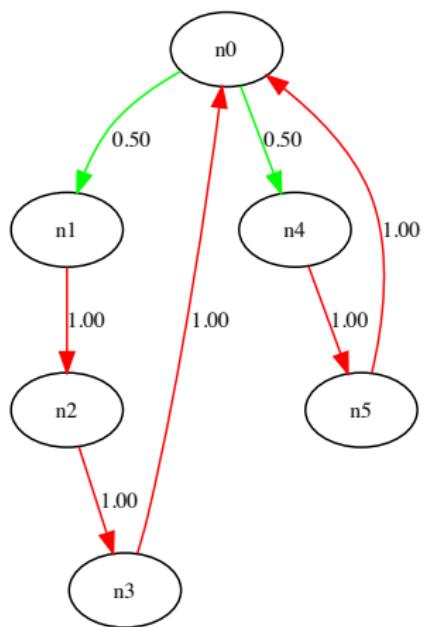
Matrice de transition :



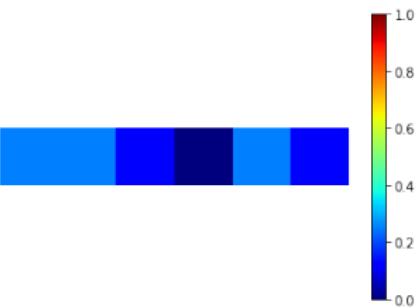
$t = 7$

# Exemples & discussion (3)

Matrice de transition :

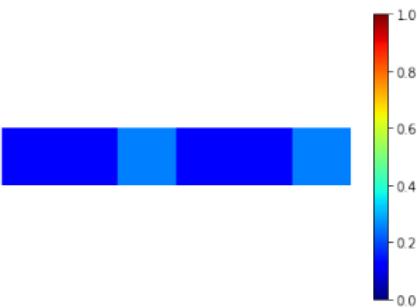
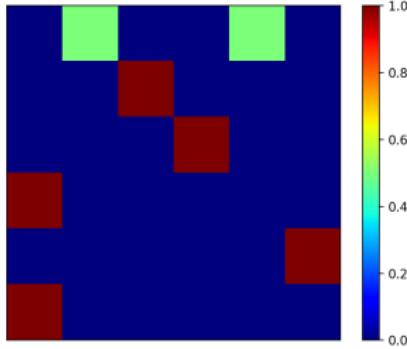
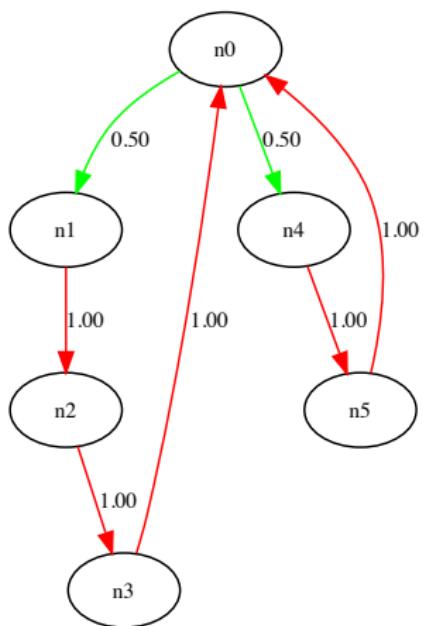


$t = 8$



# Exemples & discussion (3)

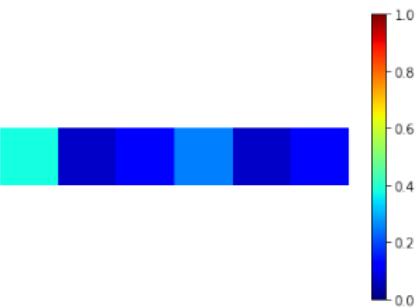
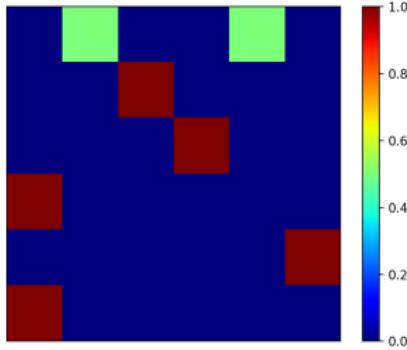
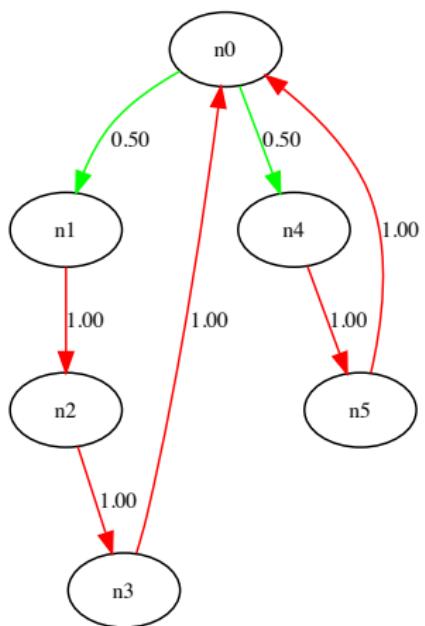
Matrice de transition :



$t = 9$

# Exemples & discussion (3)

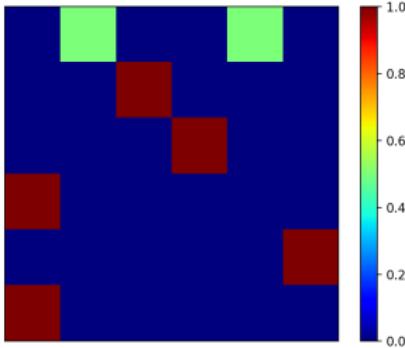
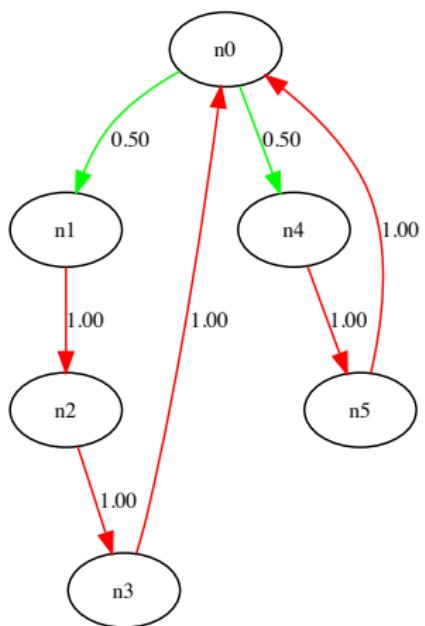
Matrice de transition :



$t = 10$

# Exemples & discussion (3)

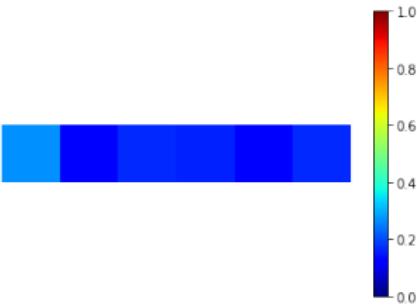
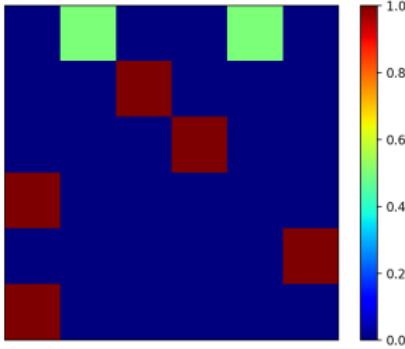
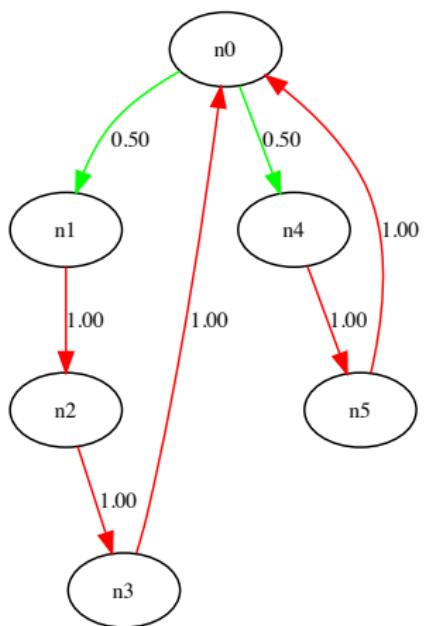
Matrice de transition :



$t = 15$

# Exemples & discussion (3)

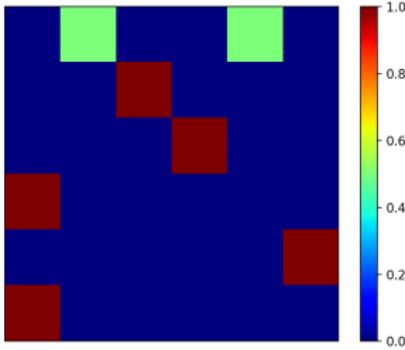
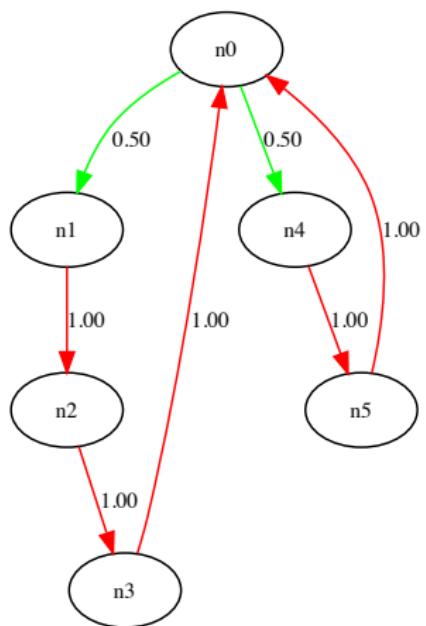
Matrice de transition :



$t = 20$

# Exemples & discussion (3)

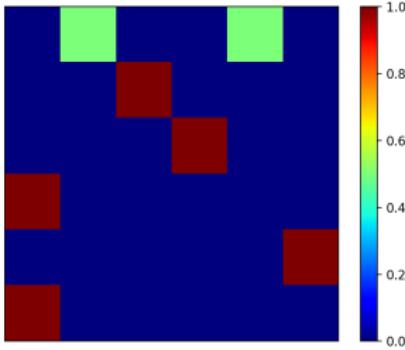
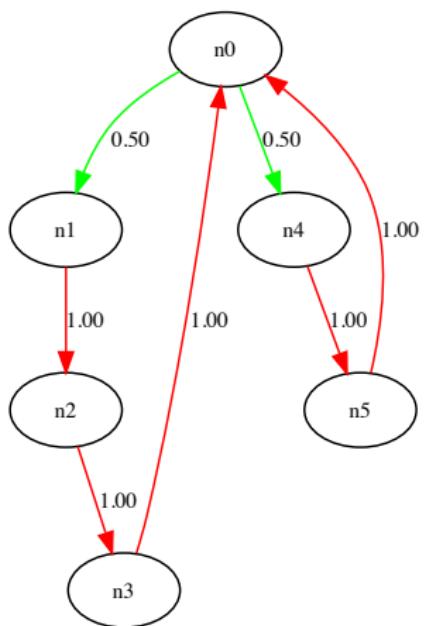
Matrice de transition :



$t = 30$

# Exemples & discussion (3)

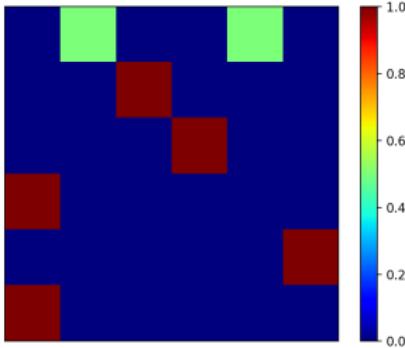
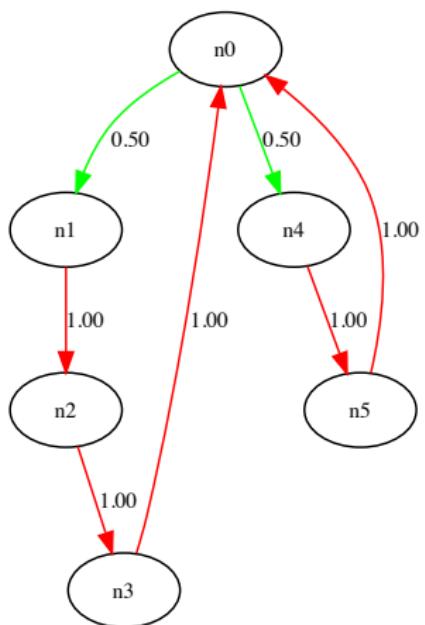
Matrice de transition :



$t = 50$

# Exemples & discussion (3)

Matrice de transition :

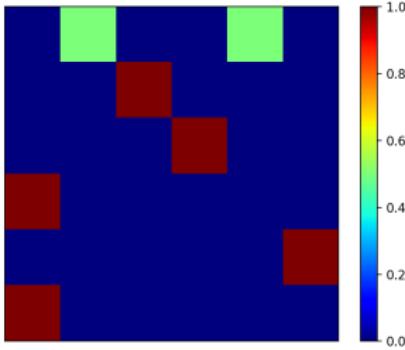
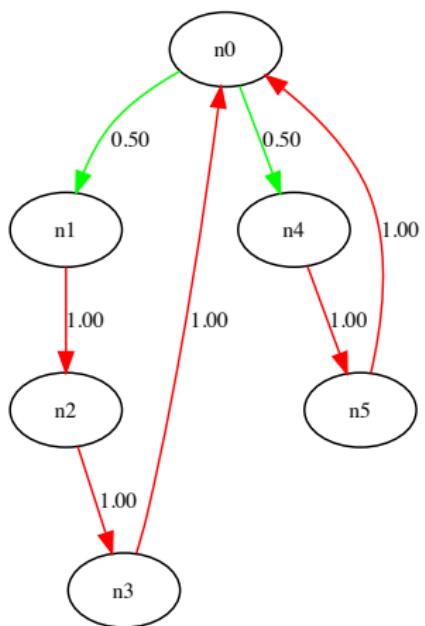


$t = 50$



# Exemples & discussion (3)

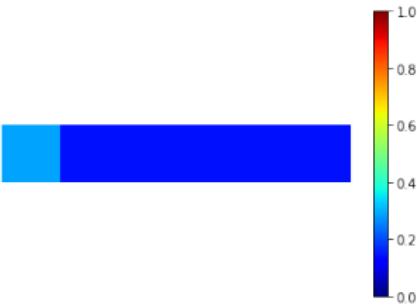
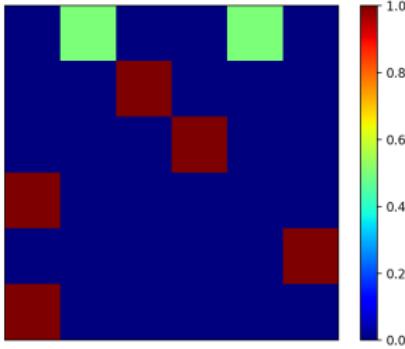
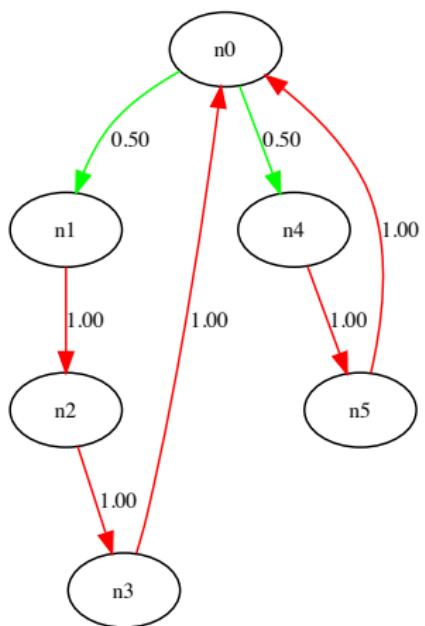
Matrice de transition :



$t = 60$

# Exemples & discussion (3)

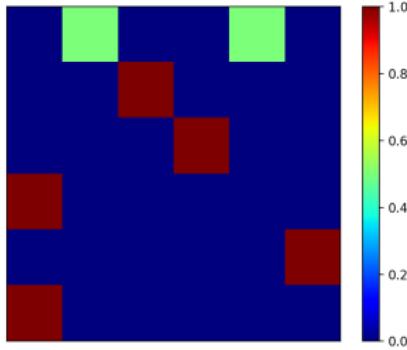
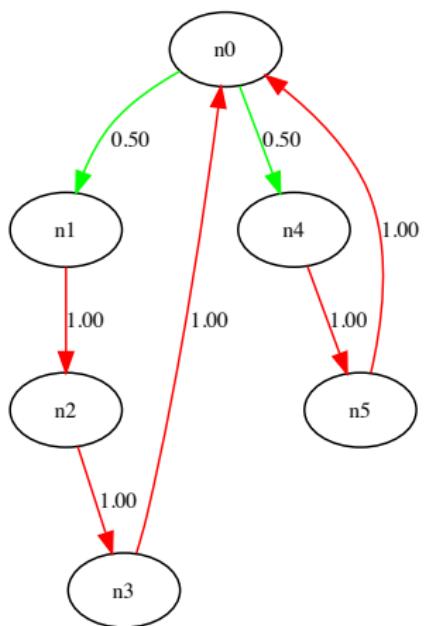
Matrice de transition :



$t = 70$

# Exemples & discussion (3)

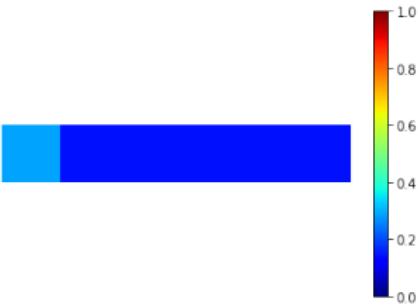
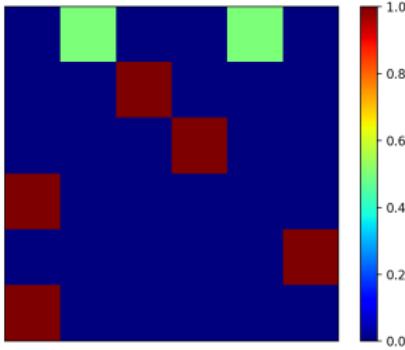
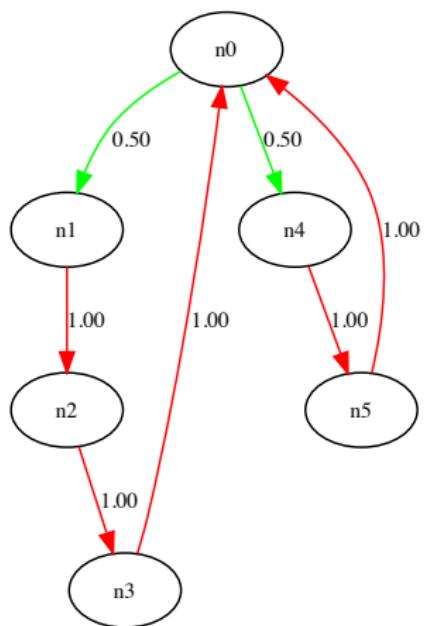
Matrice de transition :



$t = 80$

# Exemples & discussion (3)

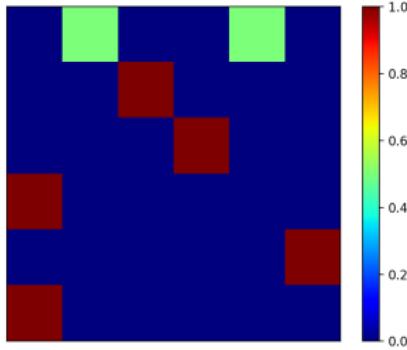
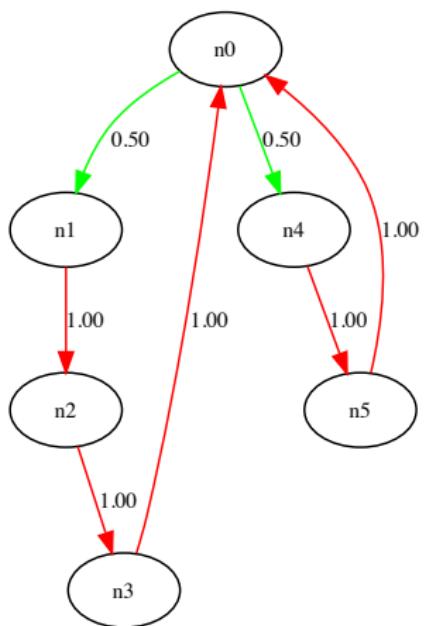
Matrice de transition :



$t = 90$

# Exemples & discussion (3)

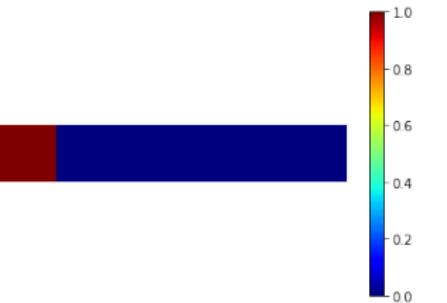
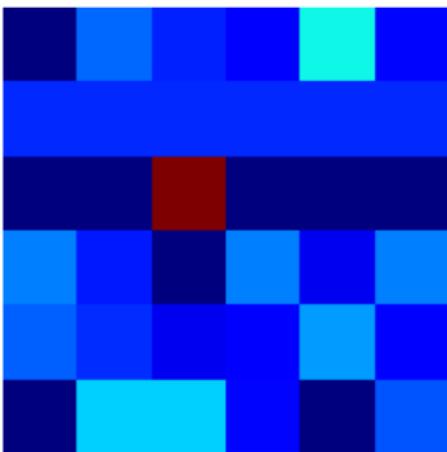
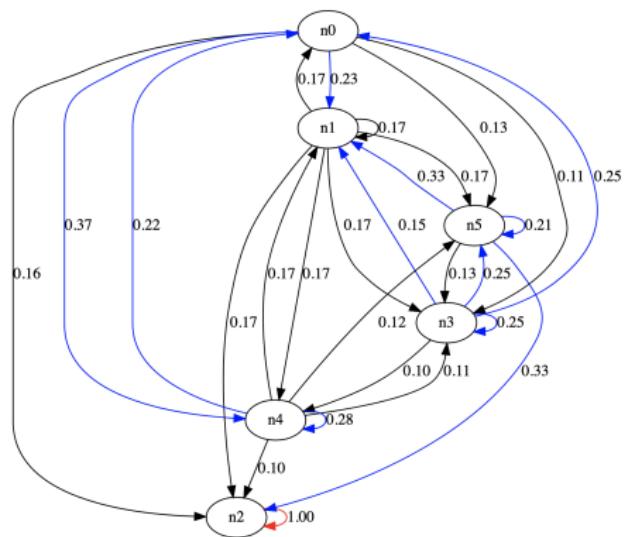
Matrice de transition :



$t = 100$

# Exemples & discussion (4)

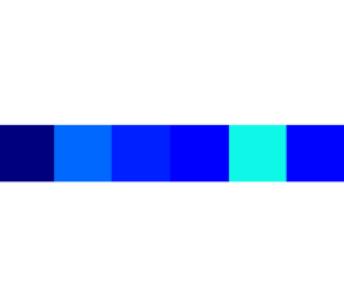
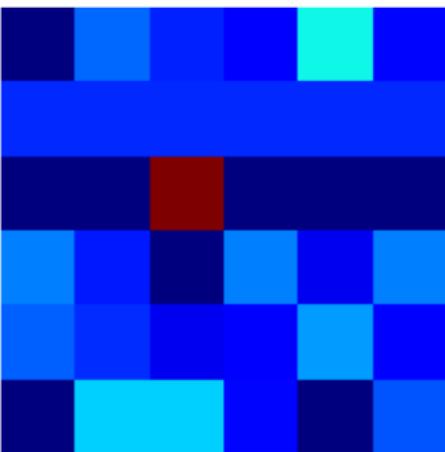
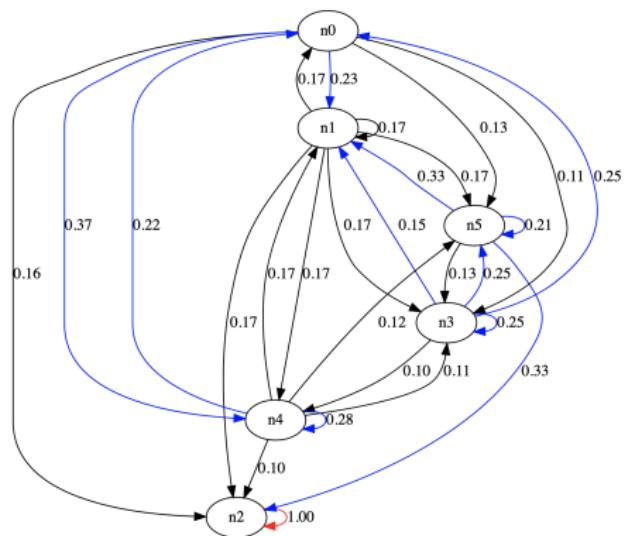
Matrice de transition :



$t = 0$

# Exemples & discussion (4)

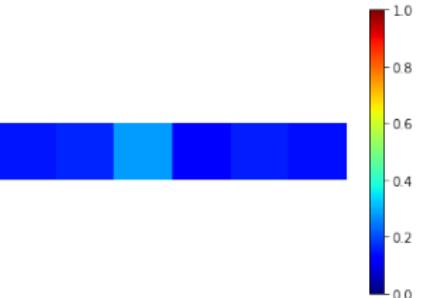
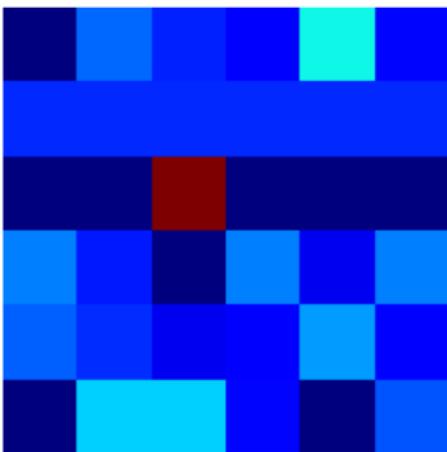
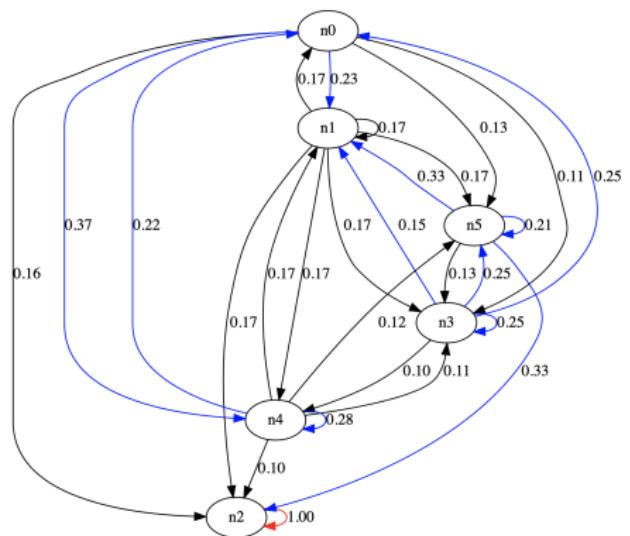
Matrice de transition :



$t = 1$

# Exemples & discussion (4)

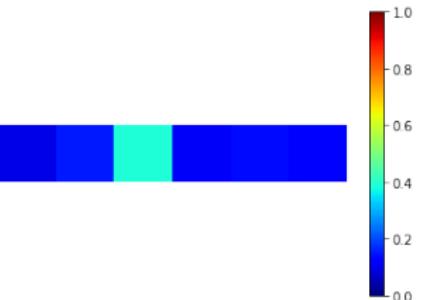
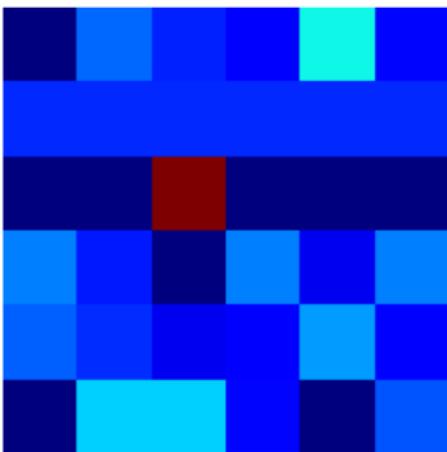
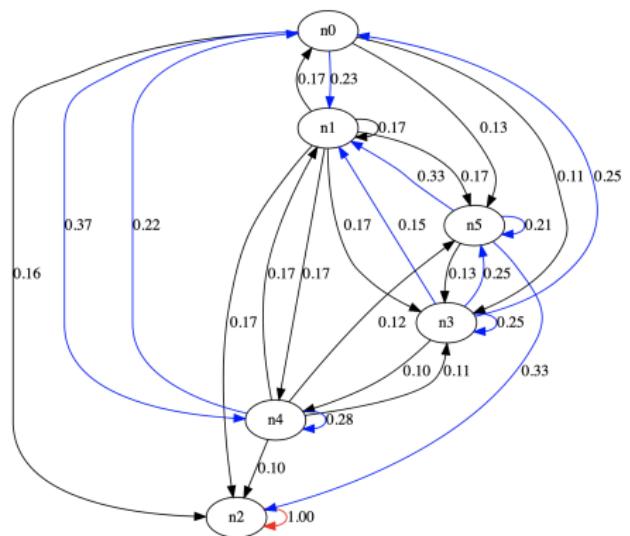
Matrice de transition :



$t = 2$

# Exemples & discussion (4)

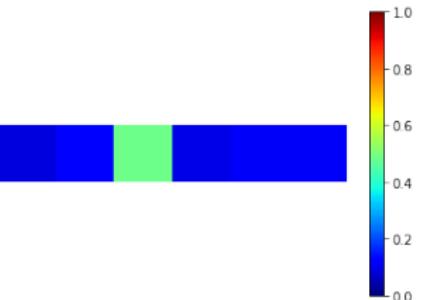
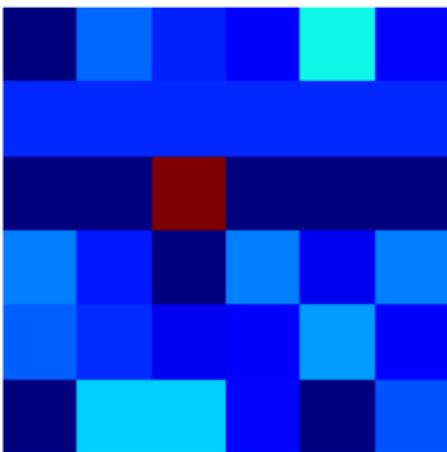
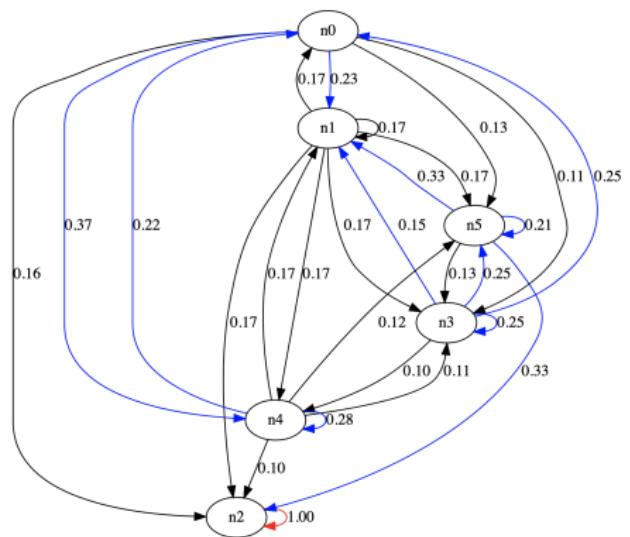
Matrice de transition :



$t = 3$

# Exemples & discussion (4)

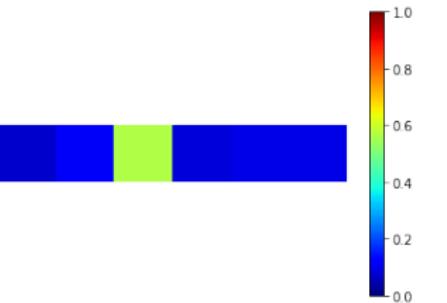
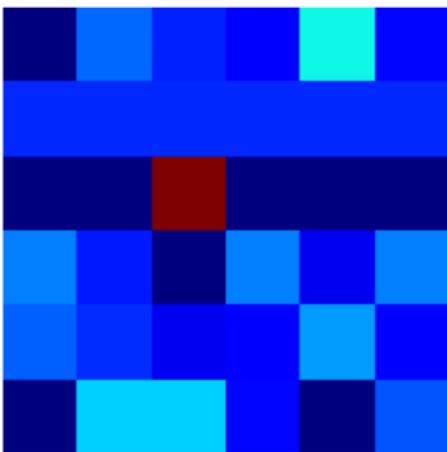
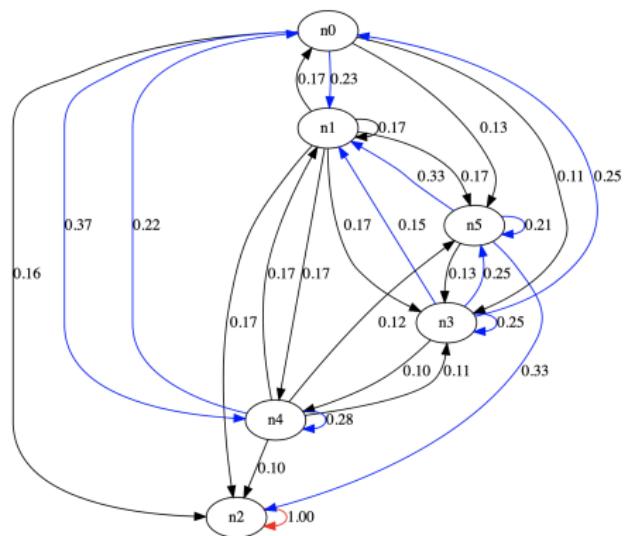
Matrice de transition :



$t = 4$

# Exemples & discussion (4)

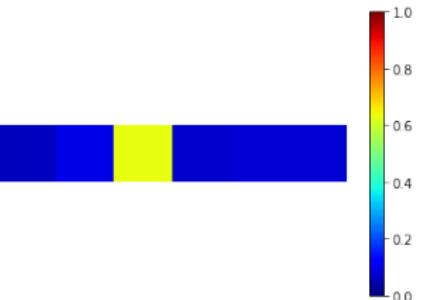
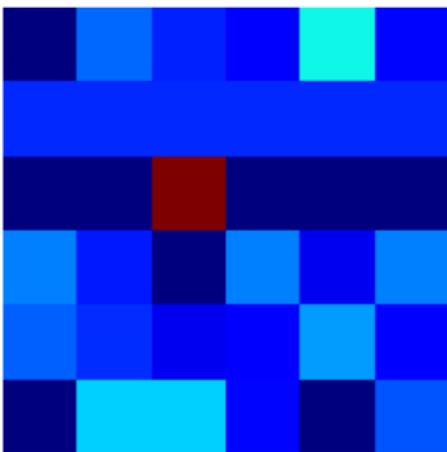
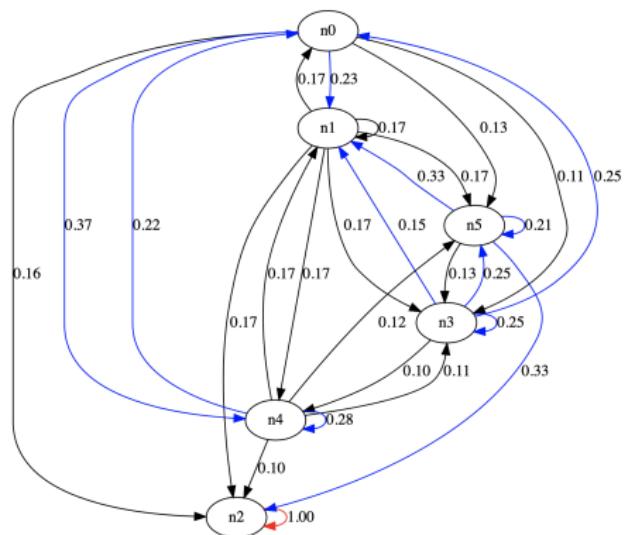
Matrice de transition :



$t = 5$

# Exemples & discussion (4)

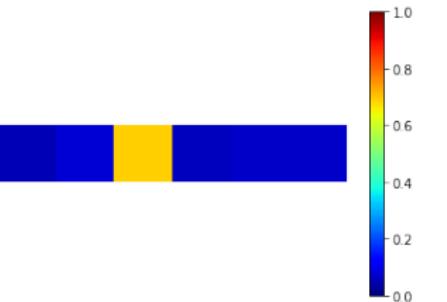
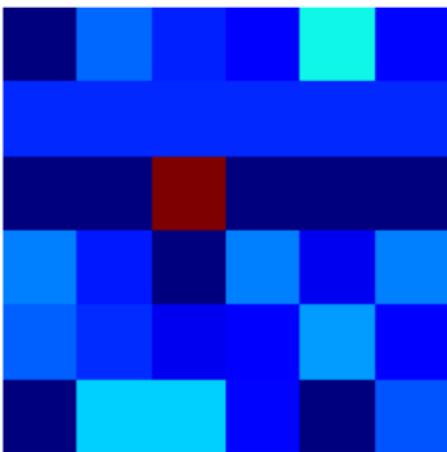
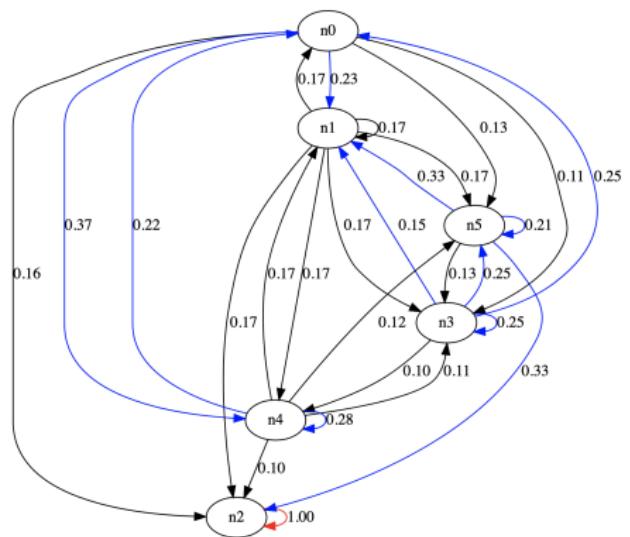
Matrice de transition :



$t = 6$

# Exemples & discussion (4)

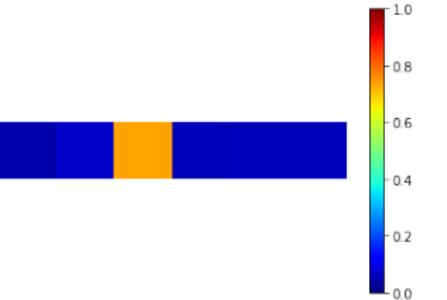
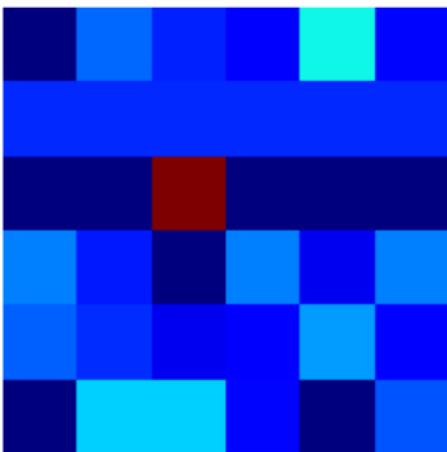
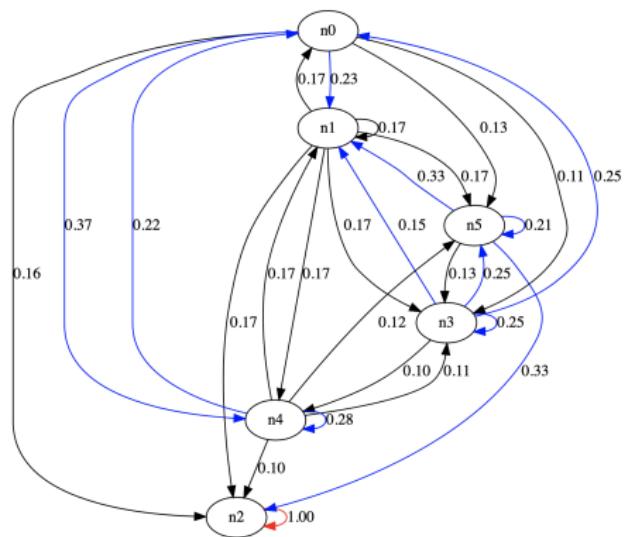
Matrice de transition :



$t = 7$

# Exemples & discussion (4)

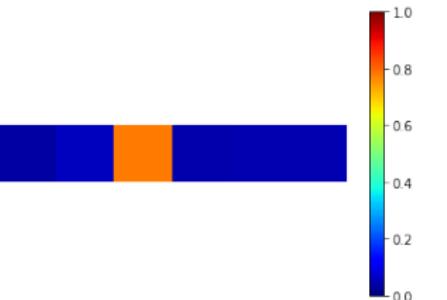
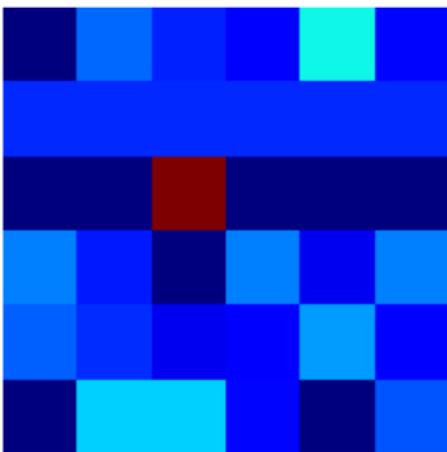
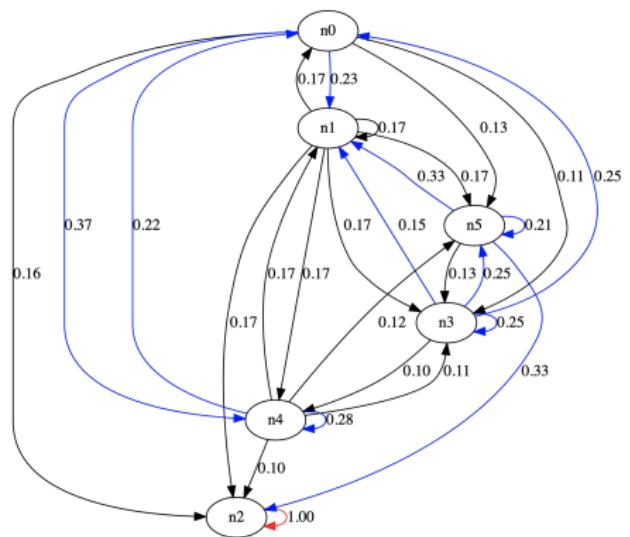
Matrice de transition :



$t = 8$

# Exemples & discussion (4)

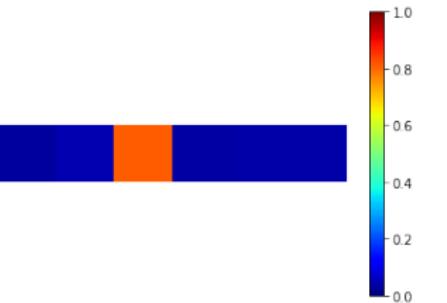
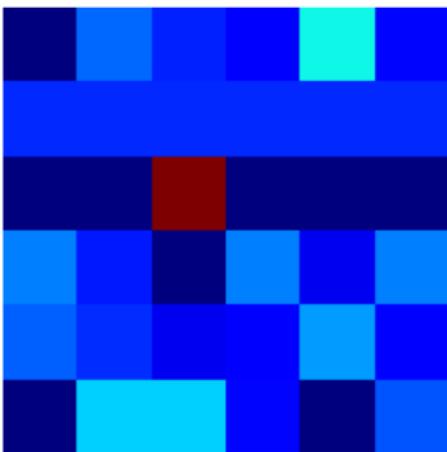
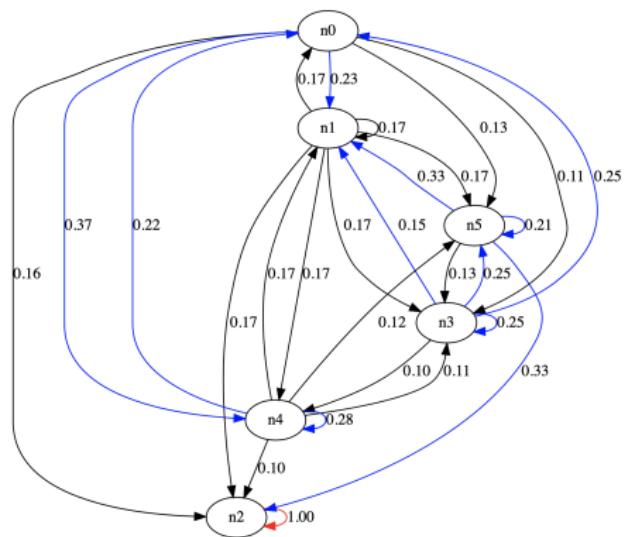
Matrice de transition :



$t = 9$

# Exemples & discussion (4)

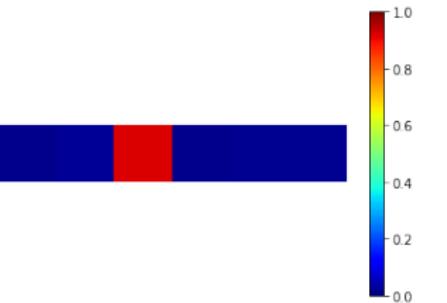
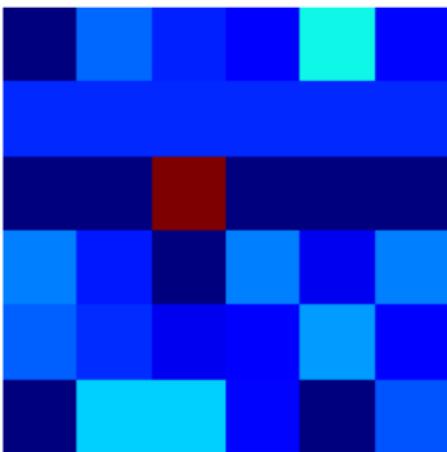
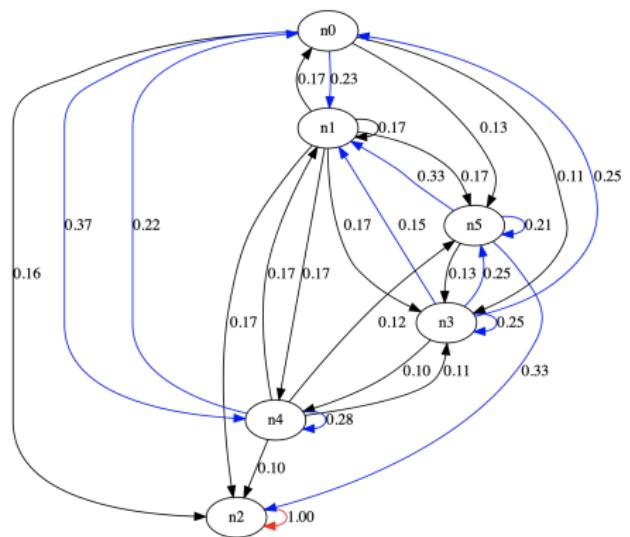
Matrice de transition :



$t = 10$

# Exemples & discussion (4)

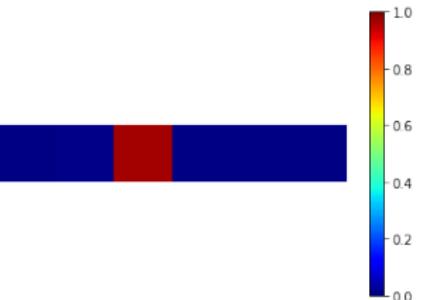
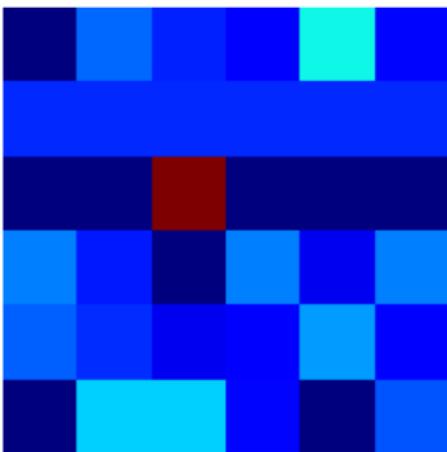
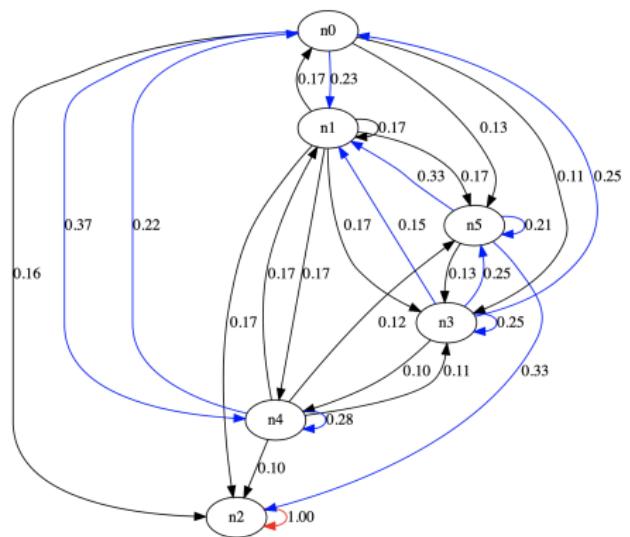
Matrice de transition :



$t = 15$

# Exemples & discussion (4)

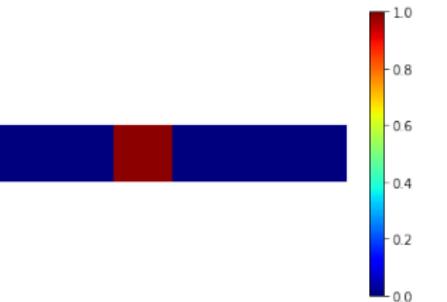
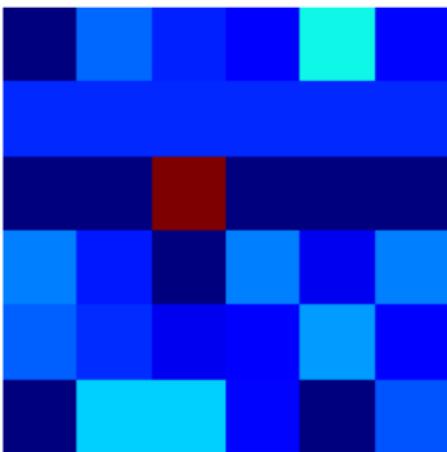
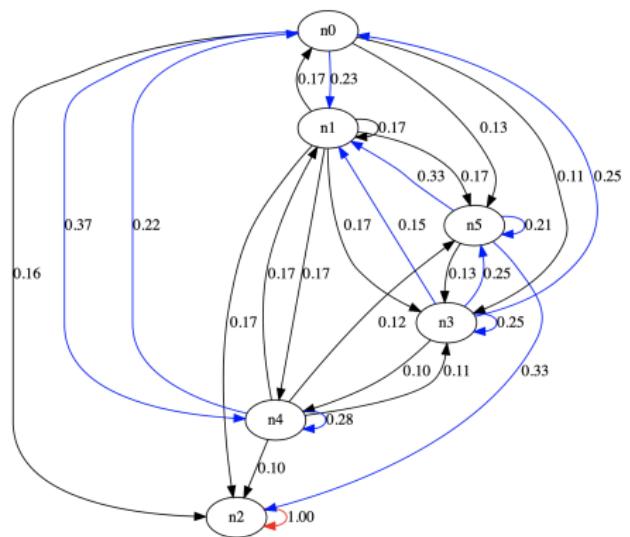
Matrice de transition :



$t = 20$

# Exemples & discussion (4)

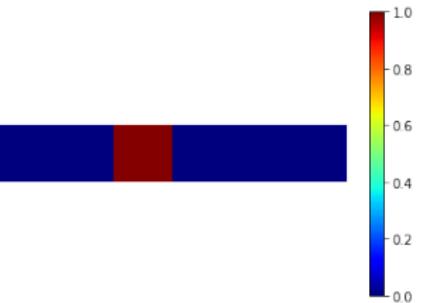
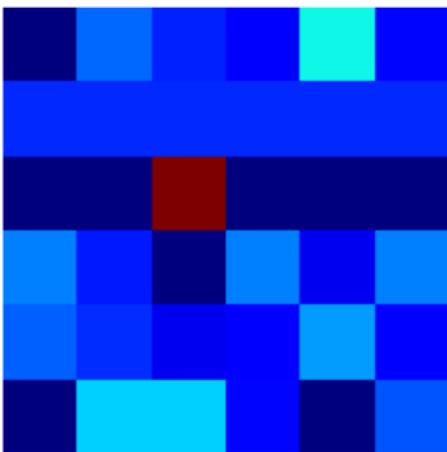
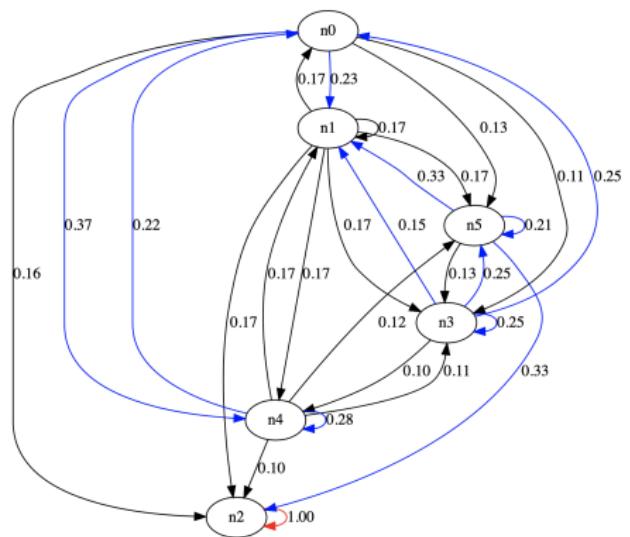
Matrice de transition :



$t = 25$

# Exemples & discussion (4)

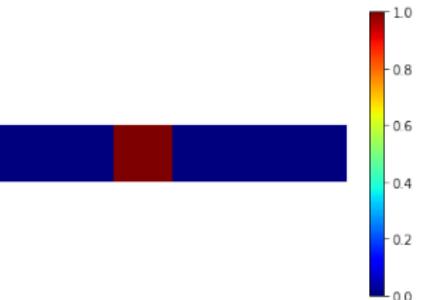
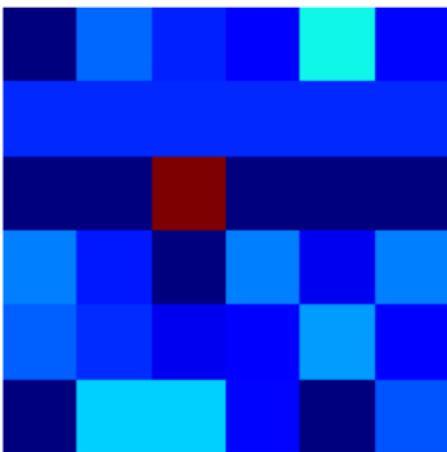
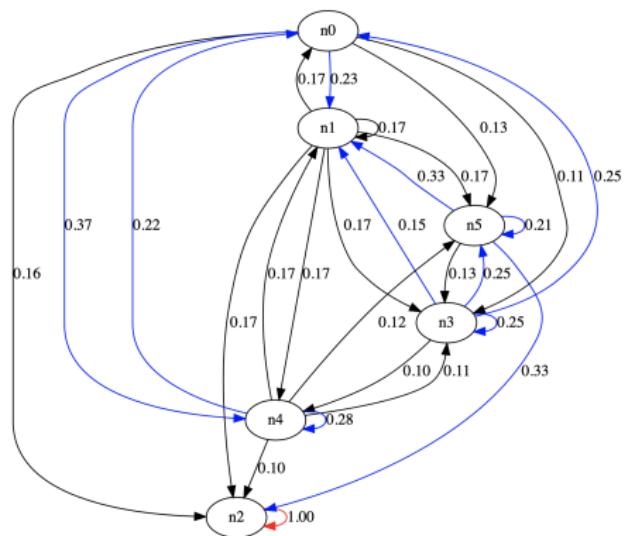
Matrice de transition :



$t = 30$

# Exemples & discussion (4)

Matrice de transition :

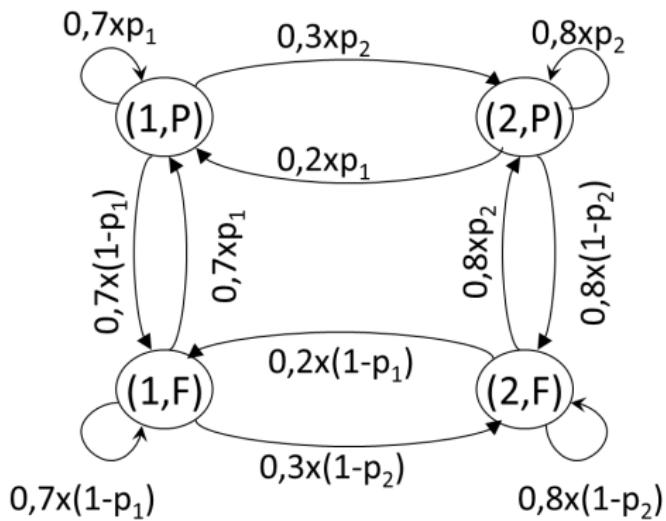
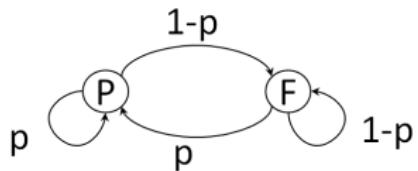


$t = 40$

- Séquence de lancers de pièce(s)...

$p_k$  : probabilité de faire *pile* avec la pièce  $k$

$p$  : probabilité de faire *pile*



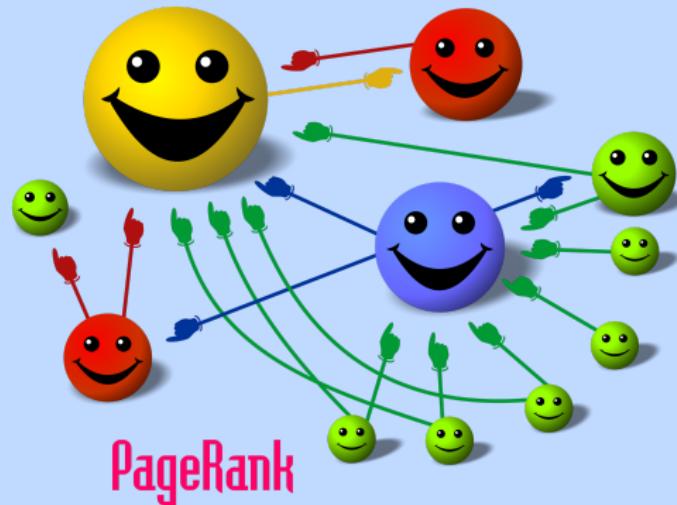
- Modélisation de parcours utilisateur sur un site web
  - Catégorisation / publicité personnalisée
  - Optimisation du site / pré-chargement de pages
- 1 trace = longueur variable...

# Modélisation (3)

- Modélisation de parcours utilisateur sur un site web
  - Catégorisation / publicité personnalisée
  - Optimisation du site / pré-chargement de pages
- 1 trace = longueur variable...

## Modèle

- **Etats :**  
page du site
- **Transitions :**  
hyperliens



# Apprentissage des chaines de Markov

Nicolas Thome  
Transparents de Vincent Guigue  
[nicolas.thome@isir.upmc.fr](mailto:nicolas.thome@isir.upmc.fr)

LIP6 / ISIR – Sorbonne Université, France

- Etant donnée une séquence d'états, calculer sa probabilité Vu précédemment :

$$\begin{aligned} p(X|\lambda) &= \prod_{t=2}^T p(x_t|x_1, \dots, x_{t-1}, \lambda) p(x_1|\lambda) = \prod_{t=2}^T p(x_t|x_{t-1}, \lambda) p(x_1|\lambda) \\ &= \pi_{x_1} \prod_{t=2}^T a_{x_{t-1}, x_t} \end{aligned}$$

- Comment apprendre une CM à partir d'exemples ?
- Comment faire de la classification de séquences avec des CM ?

- Soit une base de séquences  $B = \{X^1, \dots, X^K\}$  ( $N$  états possibles)
- Critère de vraisemblance

$$\log \mathcal{L}(B, \lambda) = \log \left( \prod_{k=1}^K p(X^k | \lambda) \right) = \sum_k \log(p(X^k | \lambda))$$

- Optimisation :

$$\lambda^* = \arg \max_{\lambda} \log \mathcal{L}(B, \lambda)$$

- Contraintes :

$$\forall i \in [1, N], \sum_{j=1}^N a_{ij} = 1$$

$$\sum_{j=1}^N \pi_j = 1$$

- Critère intégrant les **contraintes** (Lagrangien) :

$$\mathcal{C}(\lambda) = \mathcal{L}(B, \lambda) - \sum_{i=1}^N \nu_i \left( \sum_{j=1}^N a_{ij} - 1 \right) - \eta \left( \sum_{j=1}^N \pi_j - 1 \right)$$

- Si la dérivée par rapport au coefficient de contrainte est nulle, la contrainte est satisfaite :

$$\frac{\partial \mathcal{C}(\lambda)}{\partial \eta} = 0 \Leftrightarrow \sum_{j=1}^N \pi_j - 1 = 0$$

- Ecriture du Maximum de Vraisemblance :

$$L(B, \lambda) = \prod_{k=1}^K p(X^k | \lambda) = \prod_{k=1}^K \pi_{x_1^k} \prod_{t=2}^T a_{x_{t-1}^k, x_t^k}$$
$$L(B, \lambda) = \prod_{i=1}^N \pi_i^{l_i} \prod_{j=1}^N a_{ij}^{n_{ij}}$$

(1)

- $n_{ij}$  # transitions  $(ij)$  observées,  $l_i$  # état initiaux  $i$  observés
- $\ell = \log(L(B, \lambda)) = \sum_{i=1}^N \left[ l_i \log(p_i) + \sum_{j=1}^N n_{ij} \log(a_{ij}) \right]$
- Résolution :  $\frac{\partial \ell}{\partial a_{ij}} = 0$ ,  $\frac{\partial \ell}{\partial \pi_i} = 0$  + contraintes :

$$a_{ij} = \frac{n_{ij}}{n_{i \cdot}} \quad \pi_j = \frac{l_j}{K}, \text{ avec : } n_{i \cdot} = \sum_j n_{ij}$$

- Approche par comptage
- Calcul des fréquences des évènements = solution au sens MV
- Chaque ligne de  $A$  est une distribution (sommant à 1)
- En faisant une hypothèse de stationnarité, il est possible d'estimer les  $\pi_j$  sur toute la base de données :

$$\text{classique : } \pi_j = \frac{I_j}{K} \quad \text{alternative stationnaire : } \pi_j = \frac{n.j}{\sum_{ij} n_{ij}}$$

# Distance entre séquences

Nicolas Thome  
Transparents de Vincent Guigue  
[nicolas.thome@isir.upmc.fr](mailto:nicolas.thome@isir.upmc.fr)

LIP6 / ISIR – Sorbonne Université, France

- Similarité/distance = outil de base
  - k-plus proches voisins...
- La similarité peut concerner une partie seulement du signal
  - Reflexion sur les besoins spécifiques

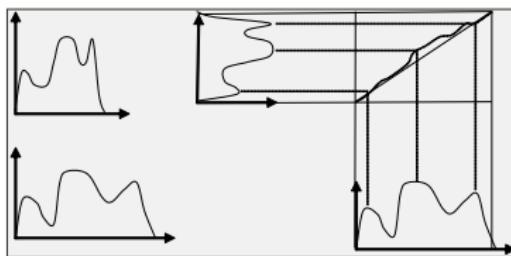
# Dynamic Time Warping (DTW)

- L'analyse du signal est locale (cf. stationnarité)
- Unités de reconnaissance plus globales (phonèmes, mots, ...)
- ⇒ Nécessité de comparer des séquences de vecteurs
- **DTW = distance entre séquences**
  - ayant des longueurs différentes
  - insensible à certaines variabilités d'élocution
  - calculable efficacement

# Distance entre séquences

## Idée :

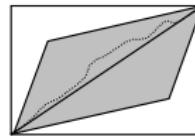
- Existence d'une distance entre séquences capable de
  - Prendre en compte les différences de rythme dans les séquences
  - Comparer des séquences de longueur différente
- Dynamic Time Warping (DTW)



La distance entre les séquences est la somme des distances entre éléments mis en correspondance par l'alignement

### Appariement avec contraintes

- Continuité locale
- Alignement quasi linéaire
- Début et fin synchrones



## Distance entre séquences (2)

Notion de chemin d'alignement :

- Chemin :  $c = \{(i_k, j_k)\}_{k=1, \dots, K}$  tel que :

$$\forall k, (i_k, j_k) = \begin{cases} (i_{k-1} - 1, j_{k-1}) & i_1 = j_1 = 1 \\ (i_{k-1} - 1, j_{k-1} - 1) & j_K = T_2 \\ (i_{k-1}, j_{k-1} - 1) & i_K = T_1 \end{cases}$$

- Distance suivant un alignement :

$$D_c(S_1, S_2) = \sum_{k=1}^K d_{c(k)}(S_1[i(k)], S_2[j(k)])$$

- Distance entre 2 séquences

$$D(S_1, S_2) = \min_c D_c(S_1, S_2)$$

# Distance entre séquences (3)

Phase avant :

- calcul des  $\forall i, j, d(S_1[i], S_2[j])$
- sommes cumulées

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

2							
1							

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

9	7	10	8	10	9	11	
6	5	6	6	7	11	13	
4	3	4	6	7	9	13	
2	2	4	5	6	9	11	
1	2	5	8	11	14	18	

# Distance entre séquences (4)

Phase retour :

- Chemin correspondant au cout minimum

9	7	10	8	10	9	11	
6	5	6	6	7	11	13	
4	3	4	6	7	9	13	
2	2	4	5	6	9	11	
1	2	5	8	11	14	18	

The diagram shows a grid of numbers with arrows indicating a path from the bottom-left cell (1) to the top-right cell (11). The path follows a sequence of moves: right, up-right, up-right.

# Distance entre séquences (5)

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

3	3	4
2	2	4
1	2	5

2	2
1	2

Phase avant

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Phase arrière

9	7	10	8	10	9	11
6	5	6	6	7	11	13
4	3	4	6	7	9	13
2	2	4	5	6	9	11
1	2	5	8	11	14	18

Alignement final et distance

$x_5$	3	2	5	2	4	2	2
$x_4$	2	2	3	2	1	4	4
$x_3$	2	1	2	2	2	3	4
$x_2$	1	1	2	1	1	3	2
$x_1$	1	1	3	3	3	3	4
	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$