

Heat diffusion distance processes: a statistical method to analyze graph data

October, 26th 2021
Colloque Jeunes Probabilistes et Statisticiens

Etienne Lasalle

Introduction



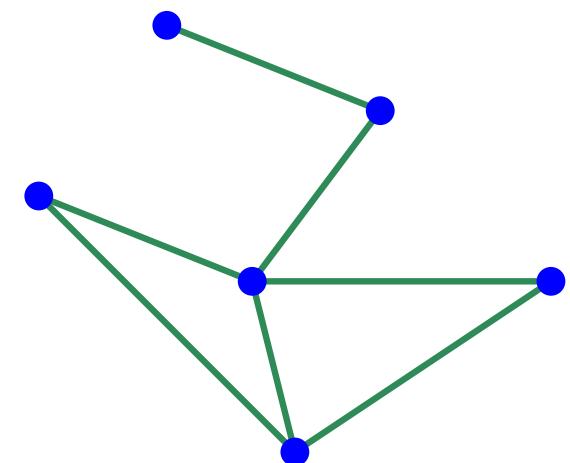
Goals :

- analysis of graph samples
 (G_1, \dots, G_N)
- theoretical results (asymptotic in N)
- useable in practice

$G = (V, E)$

Vertices

Edges



Requirements :

- take into account topological information
- graphs can be weighted
- graph sizes (same/different)
- node correspondance (known/unknown)

Outline



1. Tools

- Heat Kernel Diffusion Processes
- Heat Persistence Diffusion Processes

2. Theoretical Results

- Functional Central Limit Theorem
- Gaussian Approximation Rates

3. Simulations

- Confidence Bands
- Two-sample Tests

Comparing graphs

Assumption :

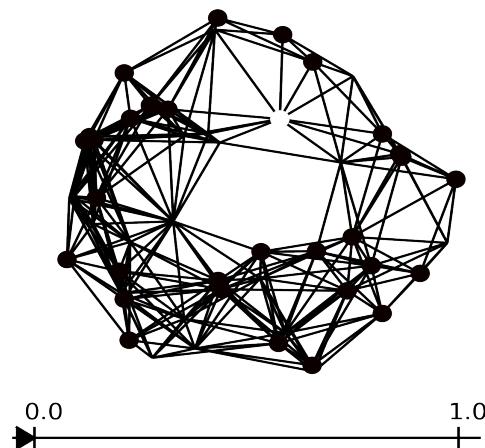
same sizes n & known node correspondance.



W , weight matrix, $W_{i,j}$: weight of $\{i, j\}$

$L = D - W$, laplacian.

D , degree matrix, $D_{i,i} = \sum_{j=1}^n W_{i,j}$



Heat Diffusion :

For $t \geq 0$, $u_t \in \mathbb{R}^n$: heat distribution,

$$\frac{d}{dt} u_t = -L u_t, \quad t \geq 0$$

$$u_t = e^{-tL} u_0,$$

e^{-tL} , heat kernel at time t .

Comparing graphs

Assumption :

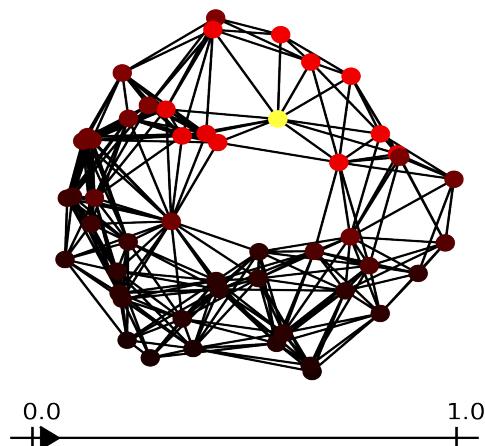
same sizes n & known node correspondance.



W , weight matrix, $W_{i,j}$: weight of $\{i, j\}$

$L = D - W$, laplacian.

D , degree matrix, $D_{i,i} = \sum_{j=1}^n W_{i,j}$



Heat Diffusion :

For $t \geq 0$, $u_t \in \mathbb{R}^n$: heat distribution,

$$\frac{d}{dt} u_t = -L u_t, \quad t \geq 0$$

$$u_t = e^{-tL} u_0,$$

e^{-tL} , heat kernel at time t .

Comparing graphs

Assumption :

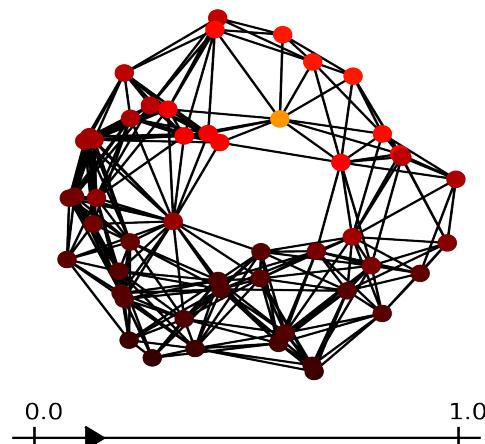
same sizes n & known node correspondance.



W , weight matrix, $W_{i,j}$: weight of $\{i, j\}$

$L = D - W$, laplacian.

D , degree matrix, $D_{i,i} = \sum_{j=1}^n W_{i,j}$



Heat Diffusion :

For $t \geq 0$, $u_t \in \mathbb{R}^n$: heat distribution,

$$\frac{d}{dt} u_t = -L u_t, \quad t \geq 0$$

$$u_t = e^{-tL} u_0,$$

e^{-tL} , heat kernel at time t .

Comparing graphs

Assumption :

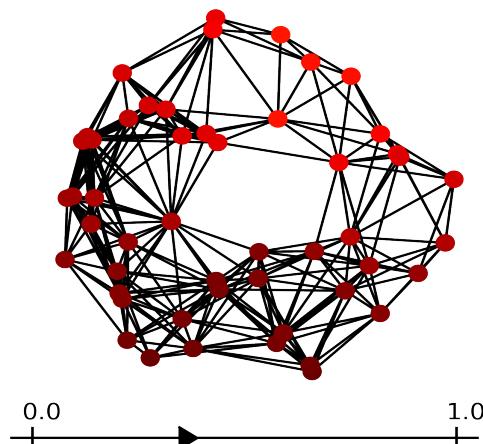
same sizes n & known node correspondance.



W , weight matrix, $W_{i,j}$: weight of $\{i, j\}$

$L = D - W$, laplacian.

D , degree matrix, $D_{i,i} = \sum_{j=1}^n W_{i,j}$



Heat Diffusion :

For $t \geq 0$, $u_t \in \mathbb{R}^n$: heat distribution,

$$\frac{d}{dt} u_t = -L u_t, \quad t \geq 0$$

$$u_t = e^{-tL} u_0,$$

e^{-tL} , heat kernel at time t .

Comparing graphs

Assumption :

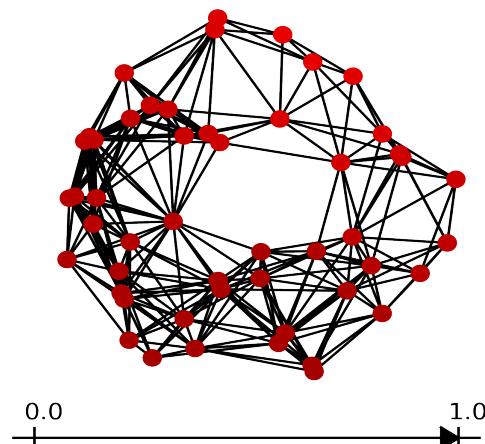
same sizes n & known node correspondance.



W , weight matrix, $W_{i,j}$: weight of $\{i, j\}$

$L = D - W$, laplacian.

D , degree matrix, $D_{i,i} = \sum_{j=1}^n W_{i,j}$



Heat Diffusion :

For $t \geq 0$, $u_t \in \mathbb{R}^n$: heat distribution,

$$\frac{d}{dt} u_t = -L u_t, \quad t \geq 0$$

$$u_t = e^{-tL} u_0,$$

e^{-tL} , heat kernel at time t .

Comparing graphs

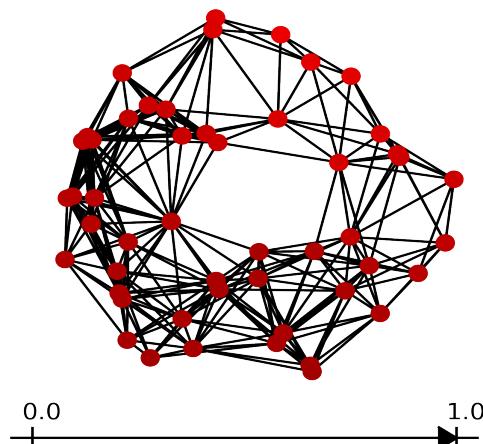
Assumption :

same sizes n & known node correspondance.

W , weight matrix, $W_{i,j}$: weight of $\{i, j\}$

$L = D - W$, laplacian.

D , degree matrix, $D_{i,i} = \sum_{j=1}^n W_{i,j}$



Heat Diffusion :

For $t \geq 0$, $u_t \in \mathbb{R}^n$: heat distribution,

$$\frac{d}{dt} u_t = -L u_t, \quad t \geq 0$$

$$u_t = e^{-tL} u_0,$$

e^{-tL} , heat kernel at time t .

Heat Kernel Distance :

$$D_t(G, G') = \|e^{-tL} - e^{-tL'}\|_F$$

[HGJ13]

- respectful of the topology ✓
- t : scale parameter ✓

Comparing graphs

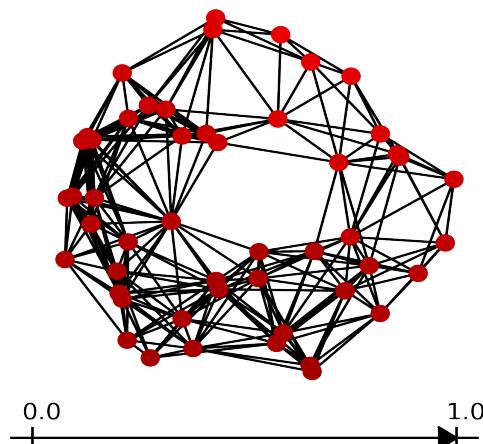
Assumption :

same sizes n & known node correspondance.

W , weight matrix, $W_{i,j}$: weight of $\{i, j\}$

$L = D - W$, laplacian.

D , degree matrix, $D_{i,i} = \sum_{j=1}^n W_{i,j}$



Heat Diffusion :

For $t \geq 0$, $u_t \in \mathbb{R}^n$: heat distribution,

$$\frac{d}{dt} u_t = -L u_t, \quad t \geq 0$$

$$u_t = e^{-tL} u_0,$$

e^{-tL} , heat kernel at time t .

Heat Kernel Distance :

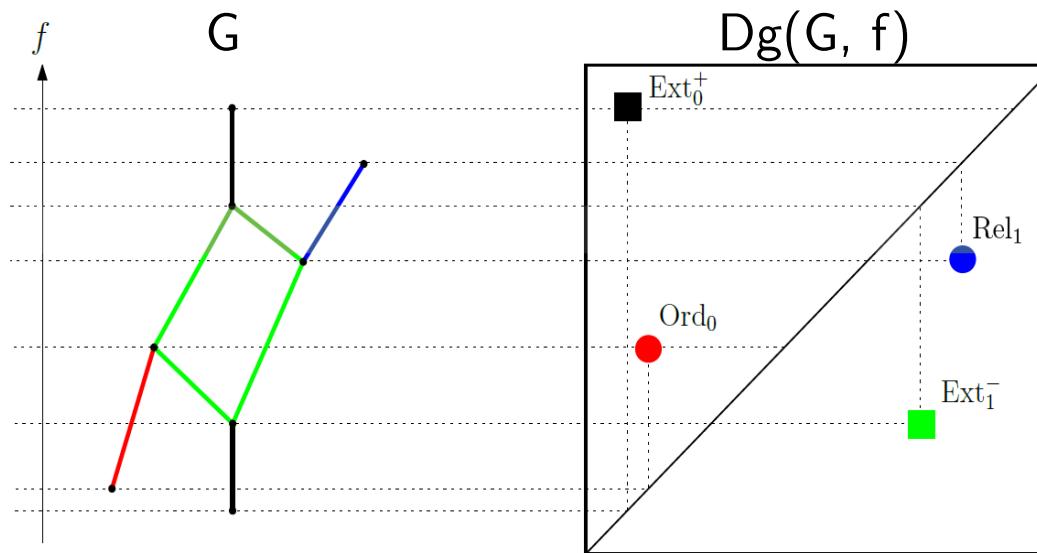
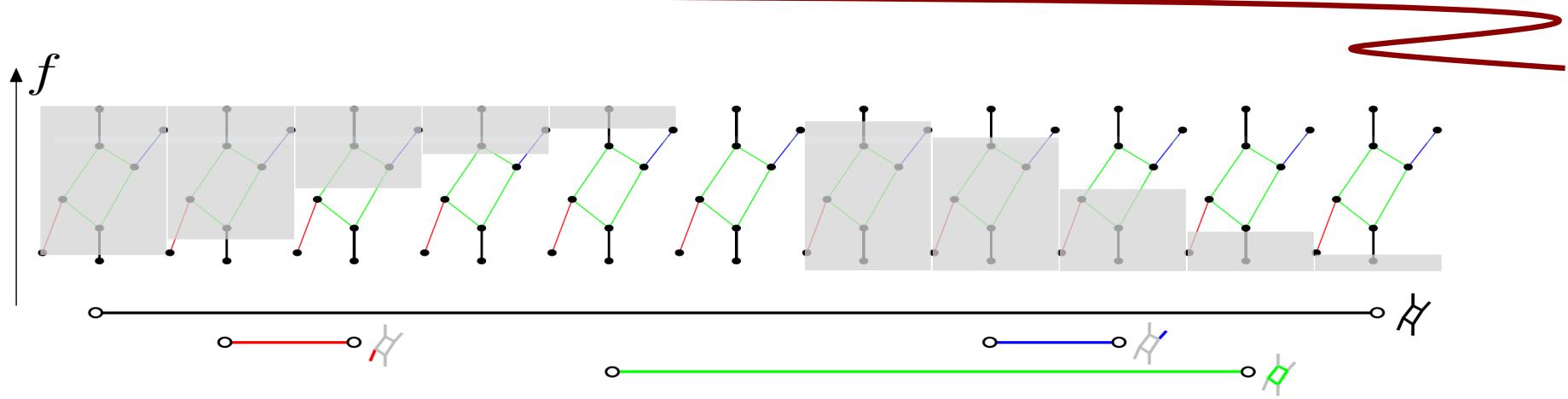
$$D_t(G, G') = \|e^{-tL} - e^{-tL'}\|_F \quad [\text{HGJ13}]$$

- respectful of the topology
- t : scale parameter



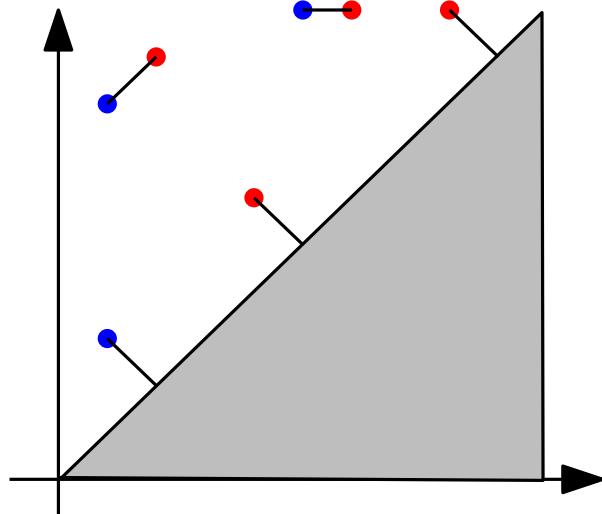
??

Using Topological Data Analysis



Figures from [CCIL+19]

Comparing persistence diagrams



μ, ν : finite multisets of points in \mathbb{R}^2 .

$\Delta = \{(a, a), \forall a \in \mathbb{R}\}$: diagonal

π : a matching from $\mu \cup \Delta$ to $\nu \cup \Delta$

$\Pi(\mu, \nu)$: set of all matchings

Bottleneck Distance :

$$d_B(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \sup_{x \in \mu \cup \Delta} \|x - \pi(x)\|_\infty$$

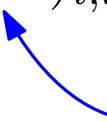
[EH10]

Choice of f



Heat Kernel Signature (HKS) : [SOG09] [HRG14]

$$h_t(G) : i \rightarrow (e^{-tL})_{i,i}$$

 "Remaining heat at node i "

Heat Persistence Distance (HPD) :

$$H_t(G, G') = \max_{D_g} d_B(Dg(G, h_t(G)), Dg(G', h_t(G')))$$

[SOG09]: A concise and provably informative multiscale signature based on heat diffusion, Sun, Ovsjanikov, Guibas, 2009

7 - 1 [HRG14] : Stable and informative spectral signatures for graph matching, Hu, Rustamov, Guibas, 2014

Choice of f



Heat Kernel Signature (HKS) : [SOG09] [HRG14]

$$h_t(G) : i \rightarrow (e^{-tL})_{i,i}$$

"Remaining heat at node i "

Heat Persistence Distance (HPD) :

$$H_t(G, G') = \max_{D_g} d_B(Dg(G, h_t(G)), Dg(G', h_t(G')))$$

Heat Kernel Distance (HKD):

$$D_t(G, G') = \|e^{-tL} - e^{-tL'}\|_F$$

[SOG09]: A concise and provably informative multiscale signature based on heat diffusion, Sun, Ovsjanikov, Guibas, 2009

7 - 2 [HRG14] : Stable and informative spectral signatures for graph matching, Hu, Rustamov, Guibas, 2014

How to choose t ?

Don't Choose !



Functional Point of View

$$D.(G, G') : \begin{array}{ccc} [0, T] & \mapsto & \mathbb{R} \\ t & \mapsto & D_t(G, G') \end{array} \quad \text{or} \quad H.(G, G') : \begin{array}{ccc} [0, T] & \mapsto & \mathbb{R} \\ t & \mapsto & H_t(G, G') \end{array}$$

Empirical Process Point of View

$$\{D_t((G, G')), \quad t \in [0, T]\} \quad \text{or} \quad \{H_t((G, G')), \quad t \in [0, T]\}$$

$$\mathcal{F}_{HKD} = \{D_t(\cdot), \quad t \in [0, T]\} \quad \text{or} \quad \mathcal{F}_{HPD} = \{H_t(\cdot), \quad t \in [0, T]\}$$

How to choose t ?

Don't Choose !

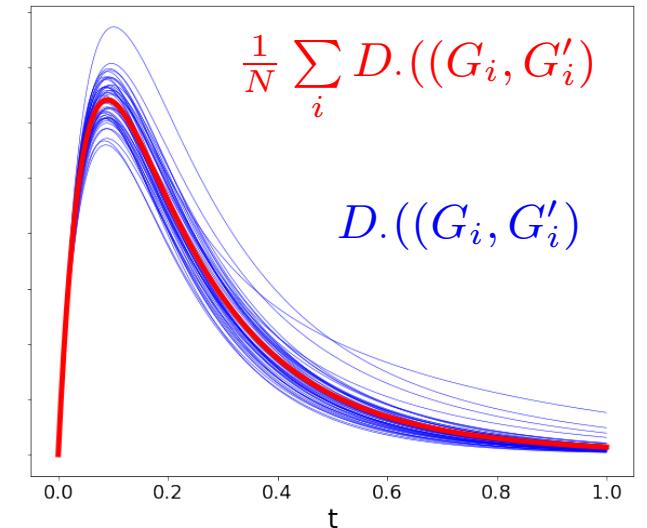
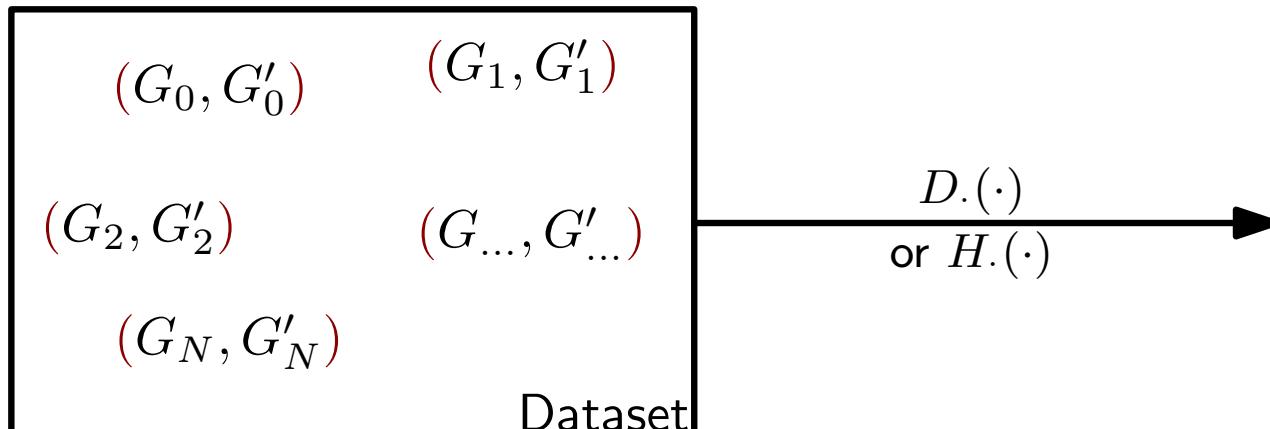


Functional Point of View

$$D.(G, G') : \begin{array}{ccc} [0, T] & \mapsto & \mathbb{R} \\ t & \mapsto & D_t(G, G') \end{array} \quad \text{or} \quad H.(G, G') : \begin{array}{ccc} [0, T] & \mapsto & \mathbb{R} \\ t & \mapsto & H_t(G, G') \end{array}$$

Empirical Process Point of View

$$\begin{array}{lll} \{D_t((G, G')), \quad t \in [0, T]\} & \text{or} & \{H_t((G, G')), \quad t \in [0, T]\} \\ \mathcal{F}_{HKD} = \{D_t(\cdot), \quad t \in [0, T]\} & \text{or} & \mathcal{F}_{HPD} = \{H_t(\cdot), \quad t \in [0, T]\} \end{array}$$



Lipschitz continuity



Proposition. Fix n and $w_{\max} > 0$.

For all (G, G') of size n , with weights in $[0, w_{\max}]$,

$$\begin{array}{ccc} [0, T] & \rightarrow & \mathbb{R}^+ \\ t & \rightarrow & D_t(G, G') \end{array} \text{ is } (n^{3/2}w_{\max})\text{-lipschitz continuous.}$$

Proposition. Fix n and $w_{\max} > 0$.

For all (G, G') of size **at most** n , with weights in $[0, w_{\max}]$,

$$\begin{array}{ccc} [0, T] & \rightarrow & \mathbb{R}^+ \\ t & \rightarrow & H_t(G, G') \end{array} \text{ is } (2nw_{\max})\text{-lipschitz continuous.}$$

Functional central limit theorem

- $(G_1, G'_1), \dots, (G_N, G'_N) \sim P$ (i.i.d sample)
- P_N : empirical measure



$$P_N D_t = \frac{1}{N} \sum_{i=1}^N D_t((G_i, G'_i))$$

$$P D_t = \mathbb{E}_P [D_t((G, G'))]$$

Theorem. Fix n and $w_{\max} > 0$. For all distribution P over pairs of graphs of size n , with weights in $[0, w_{\max}]$,
the family $\mathcal{F}_{HKD} = \{D_t(\cdot), t \in [0, T]\}$ is P -Donsker

$$\{\sqrt{N}(P_N - P)D_t, t \in [0, T]\} \xrightarrow{\text{weak}} \text{Gaussian Process } \mathbb{G}$$

$\forall h : \mathcal{C}([0, T]) \rightarrow \mathbb{R}$, continuous and bounded,

$$\lim_{N \rightarrow \infty} \mathbb{E} \left[h \left(\sqrt{N}(P_N - P)D_t \right) \right] = \mathbb{E} [h(\mathbb{G})]$$

Functional central limit theorem

- $(G_1, G'_1), \dots, (G_N, G'_N) \sim P$ (i.i.d sample)
- P_N : empirical measure



$$P_N D_t = \frac{1}{N} \sum_{i=1}^N D_t((G_i, G'_i))$$

$$PD_t = \mathbb{E}_P [D_t((G, G'))]$$

Theorem. Fix n and $w_{\max} > 0$. For all distribution P over pairs of graphs of size n , with weights in $[0, w_{\max}]$,
the family $\mathcal{F}_{HKD} = \{D_t(\cdot), t \in [0, T]\}$ is P -Donsker

$$\{\sqrt{N}(P_N - P)D_t, t \in [0, T]\} \xrightarrow{\text{weak}} \text{Gaussian Process } \mathbb{G}$$

$\forall h : \mathcal{C}([0, T]) \rightarrow \mathbb{R}$, continuous and bounded,

$$\lim_{N \rightarrow \infty} \mathbb{E} \left[h \left(\sqrt{N}(P_N - P)D_t \right) \right] = \mathbb{E} [h(\mathbb{G})]$$

Theorem. Fix n and $w_{\max} > 0$. For all distribution P over pairs of graphs of size **at most** n , with weights in $[0, w_{\max}]$,
the family $\mathcal{F}_{HPD} = \{H_t(\cdot), t \in [0, T]\}$ is P -Donsker

Consequences: consistent confidence bands and two-sample tests

Results



Gaussian Approximation with rate r_N :

$\forall \lambda > 1, \exists C$ s.t. $\forall N \geq 1,$

one can construct on the same probability space both X_N and a version of the Gaussian process $\mathbb{G}^{(N)}$, s.t.

$$\mathbb{P} \left(\|X_N - \mathbb{G}^{(N)}\|_\infty > C.r_N \right) \leq N^{-\lambda}.$$

$\left\{ \sqrt{N}(P_N - P)D_t, \quad t \in [0, T] \right\}$ and $\left\{ \sqrt{N}(P_N - P)H_t, \quad t \in [0, T] \right\}$
admit Gaussian Approximations with rate :

$$r_N = N^{-1/7} \log(N)^{9/14}.$$

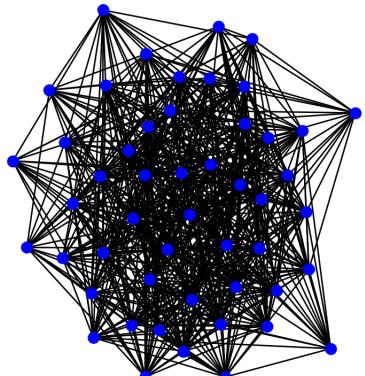
Remark: r_N is independent of n (graph size)

Simulations : Stochastic Models

Erdös-Renyi (ER)

$n = 50$

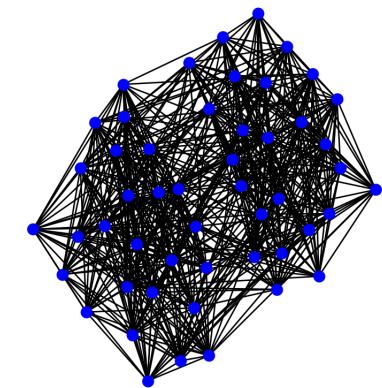
$p = 0.5$



Stochastic Block Model (SBM)

$n_1 = n_2 = 25$

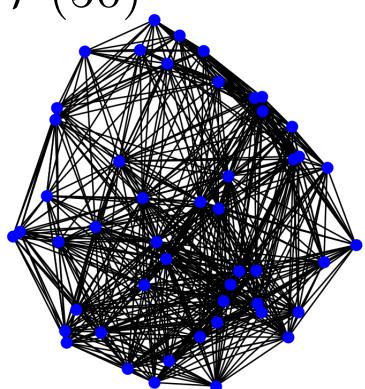
$$p = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$$



Geometric (Disk)

$n = 50$ or $\mathcal{P}(50)$

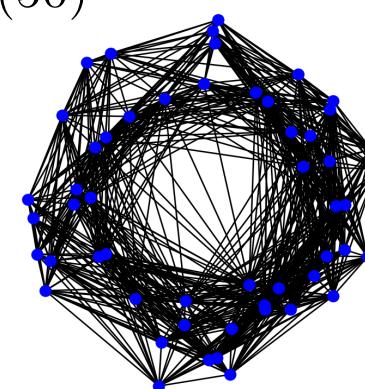
$p = 0.5$



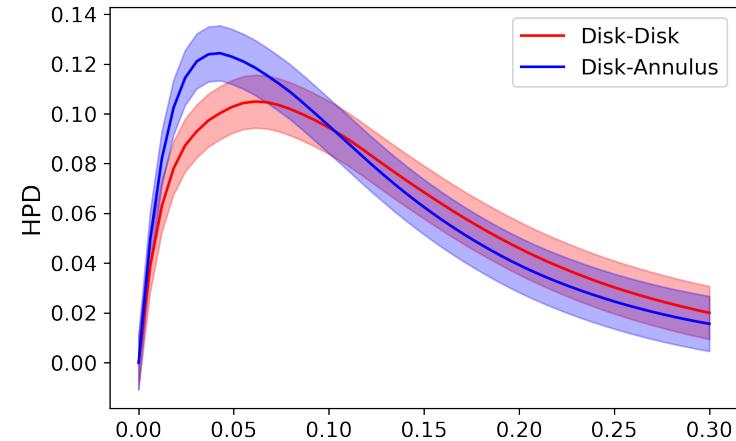
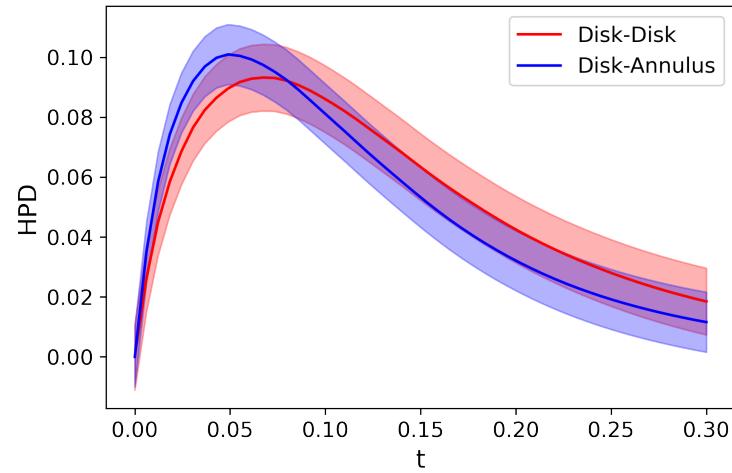
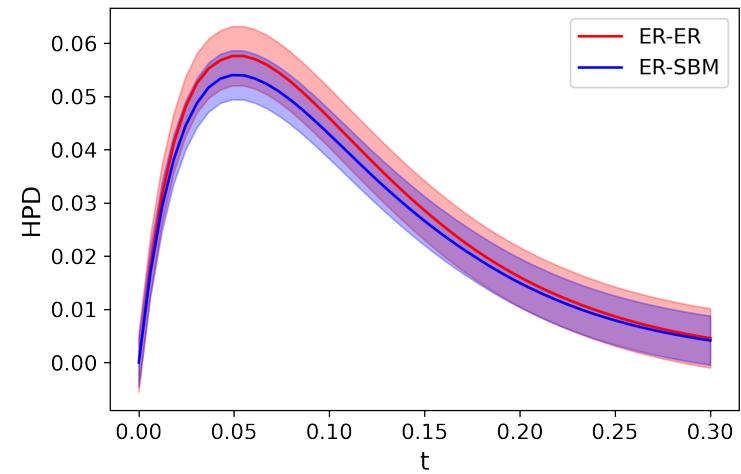
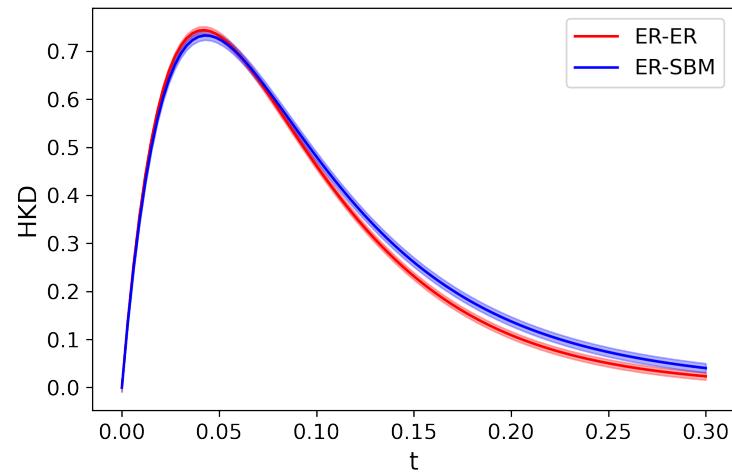
Geometric (Annulus)

$n = 50$ or $\mathcal{P}(50)$

$p = 0.5$



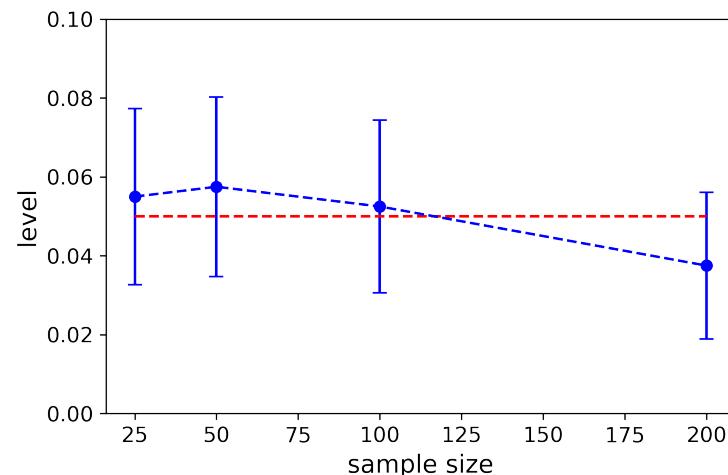
Simulations : Confidence Bands



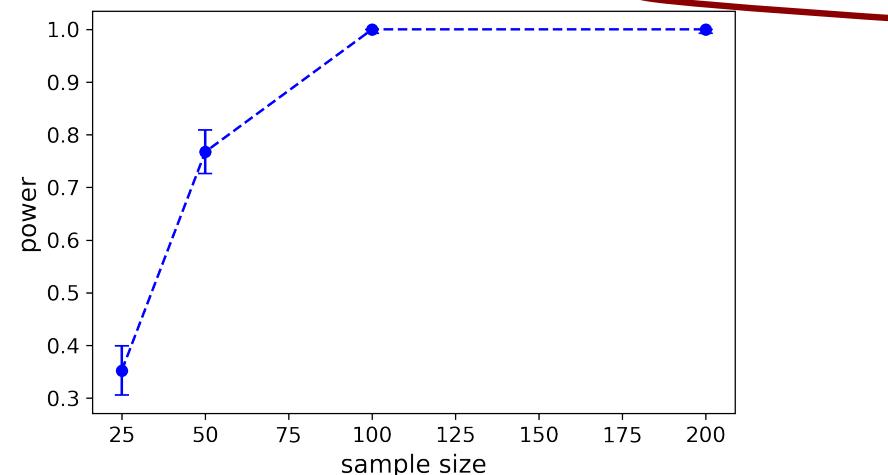
Graphs with random size.

Simulations : Two-sample Tests

HKD



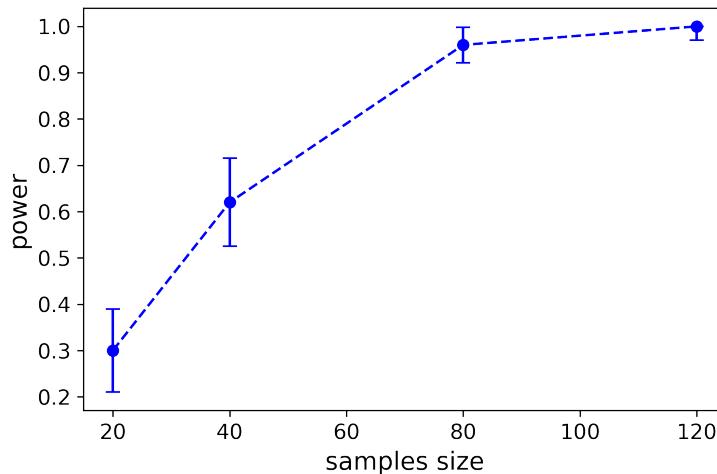
ER-ER



ER-ER vs ER-SBM

Level 95%, bootstrap sample size : 1000, number of tests : 400

HPD



Disk-Disk vs Disk-Annulus

Level 95%, bootstrap sample size : 1000, number of tests : 100

Simulations : Two-sample Tests

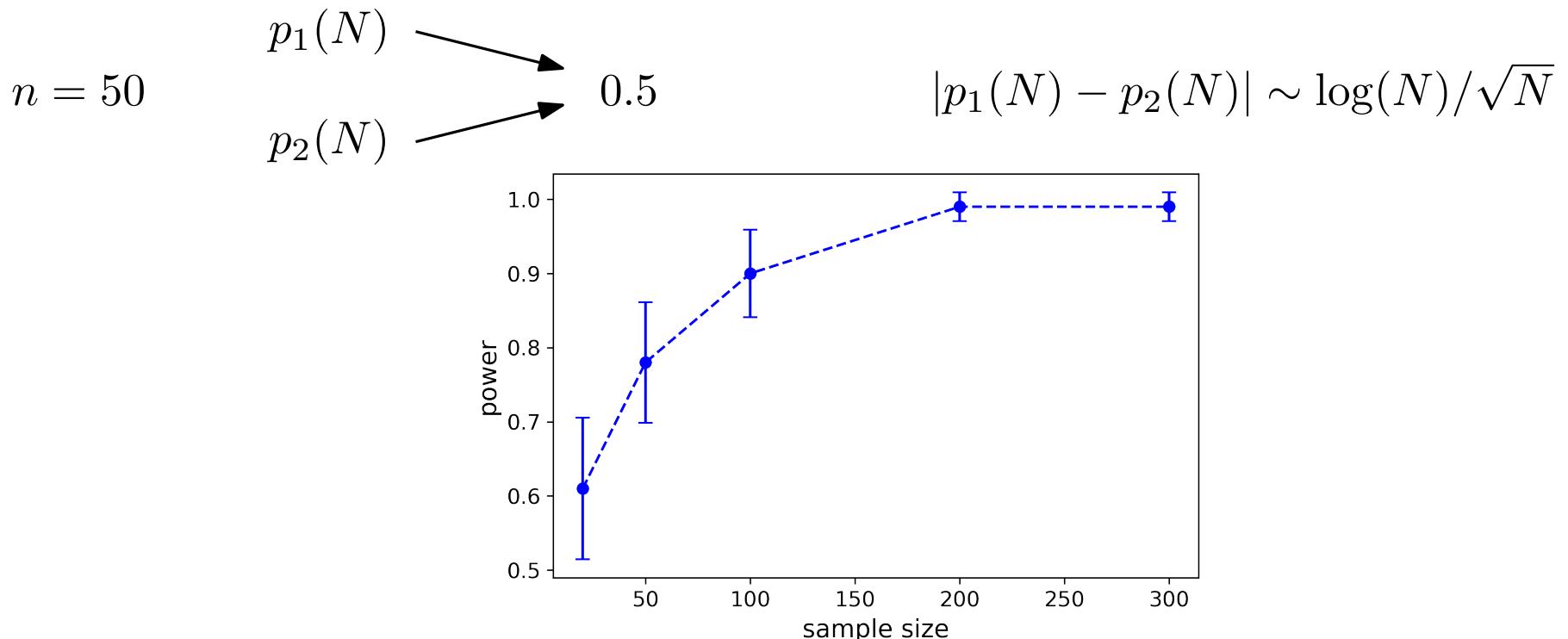


Neyman-Pearson regime :

sample of size N

Neyman-Pearson test : $ER(p_1(N))$ vs $ER(p_2(N))$

$$|p_1(N) - p_2(N)| \gg 1/\sqrt{N}$$



Conclusion



L. (2021) Heat diffusion distance processes: a statistically founded method to analyze graph data sets. arXiv preprint arXiv:2109.13213.

Future work:

- Applications to real datasets (activation graphs from NN)
- Learning tasks : classification, change point detection, ...
- Relationship between n and N

Conclusion



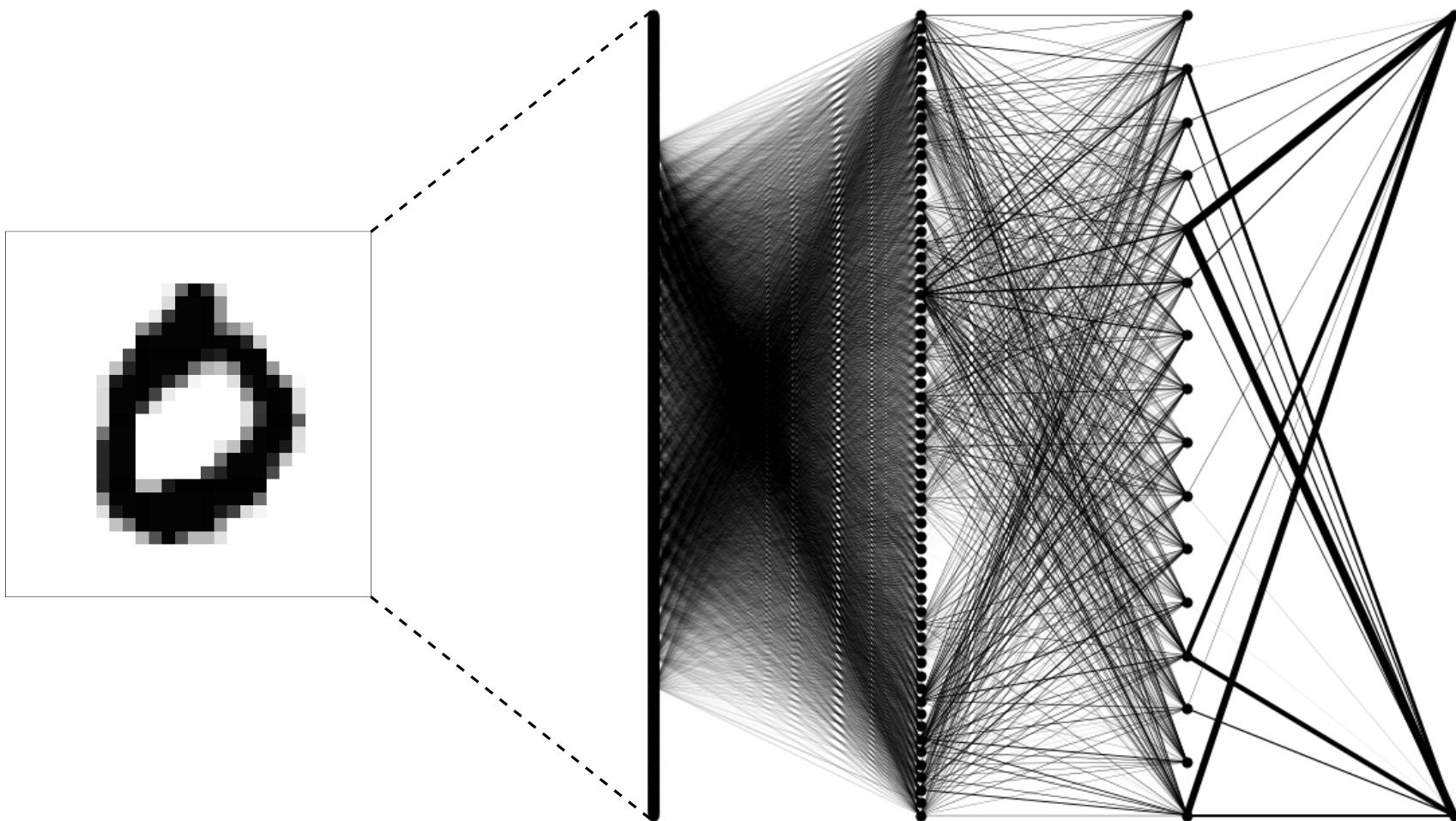
L. (2021) Heat diffusion distance processes: a statistically founded method to analyze graph data sets. arXiv preprint arXiv:2109.13213.

Future work:

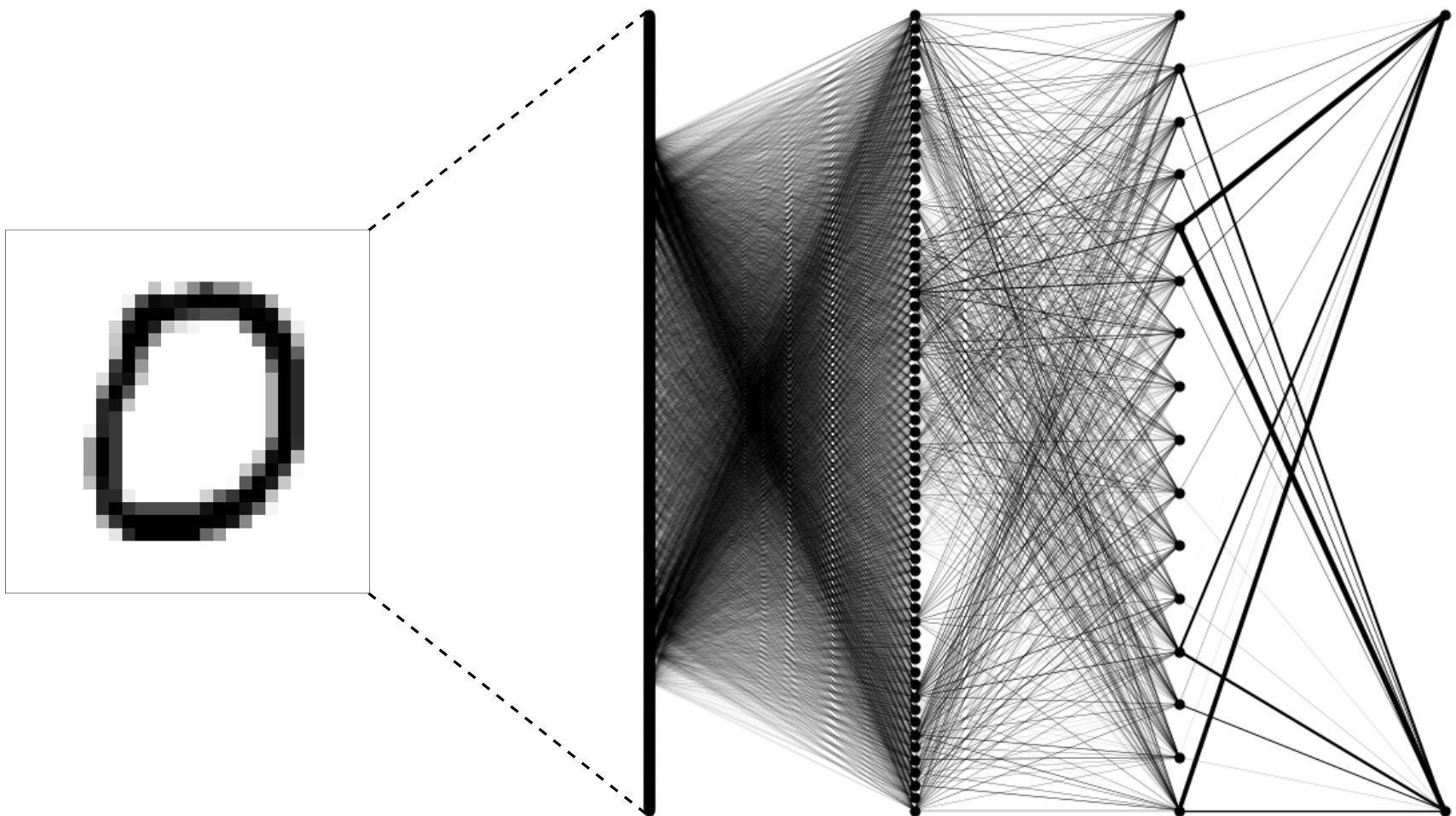
- Applications to real datasets (activation graphs from NN)
- Learning tasks : classification, change point detection, ...
- Relationship between n and N

Thank you for your attention!

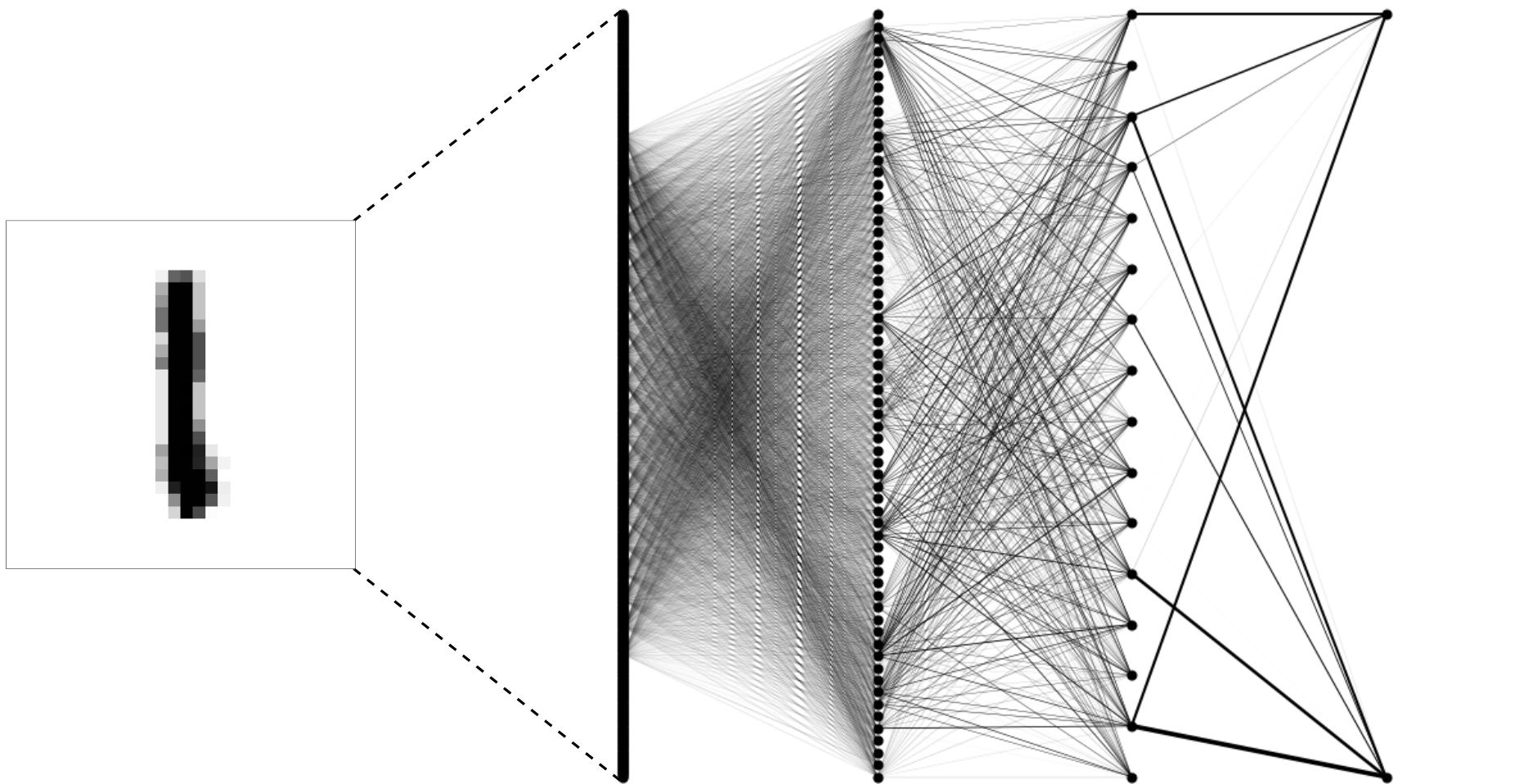
Neural Network (on MNIST)



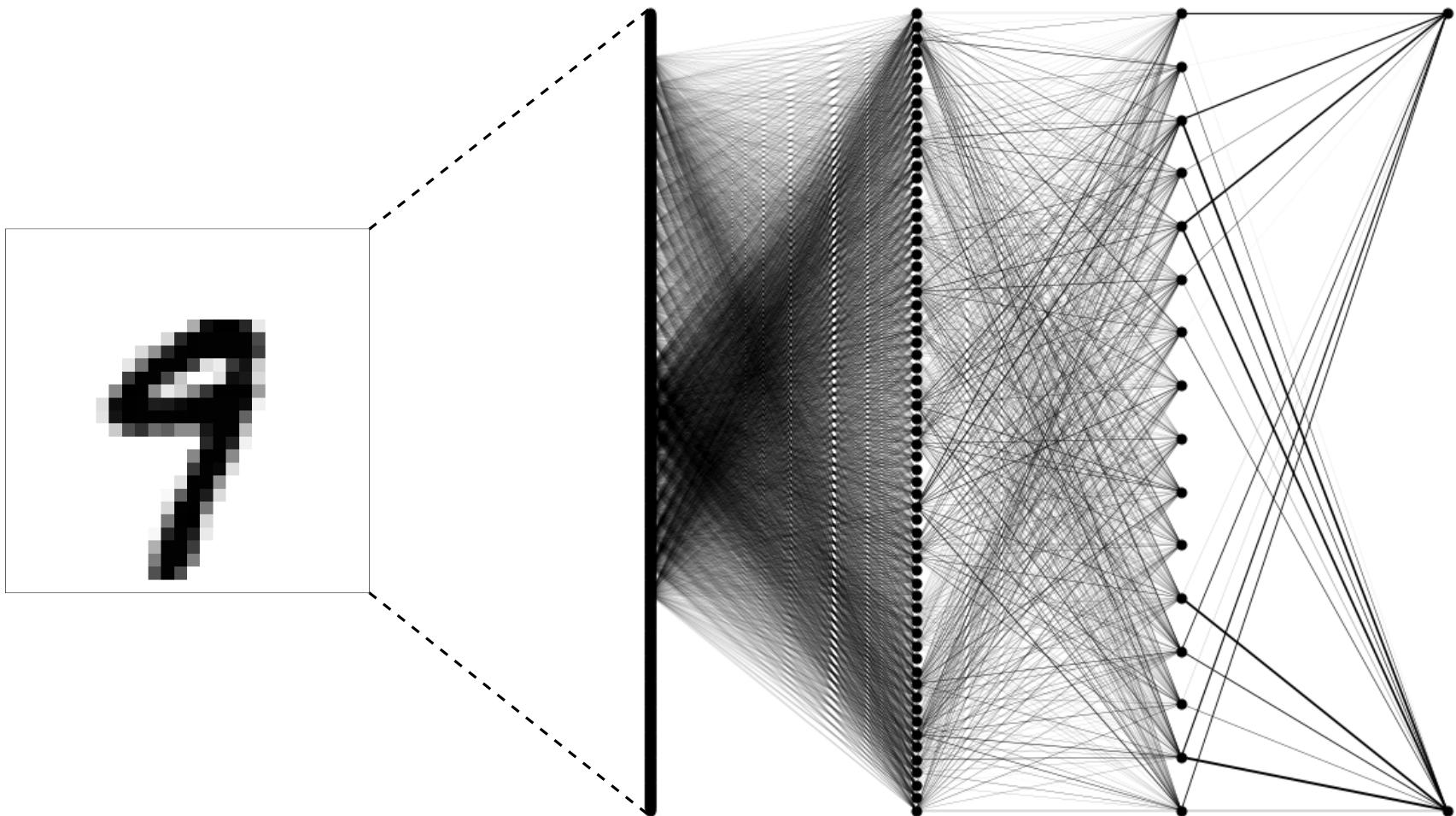
Neural Network (on MNIST)



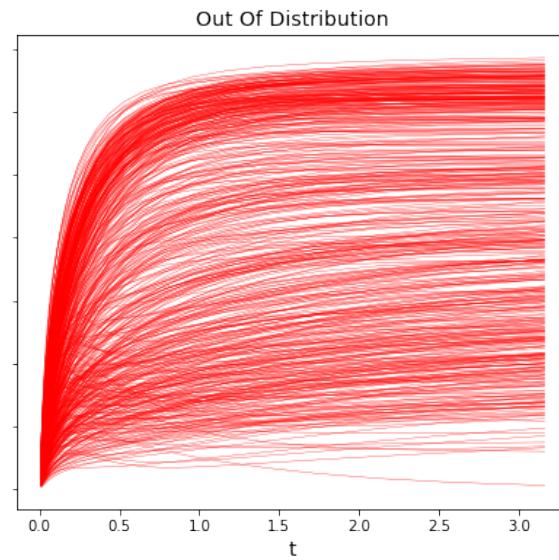
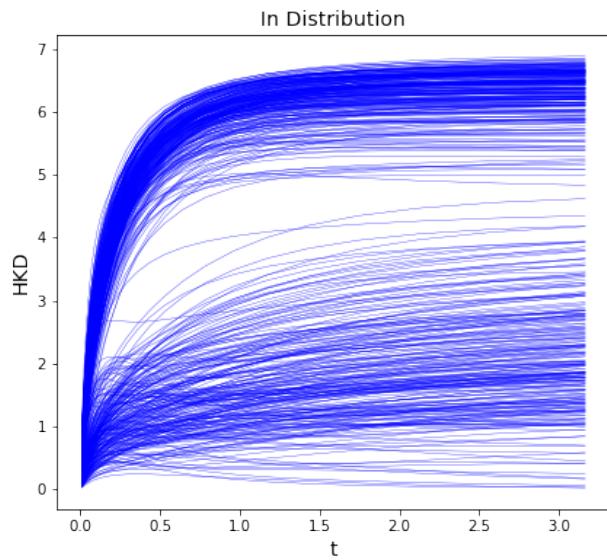
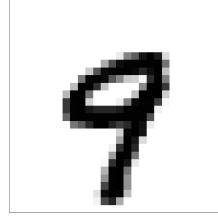
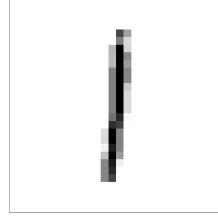
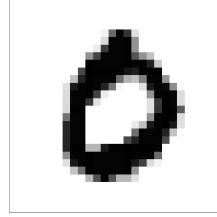
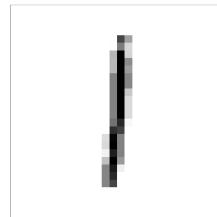
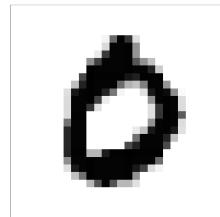
Neural Network (on MNIST)



Neural Network (on MNIST)



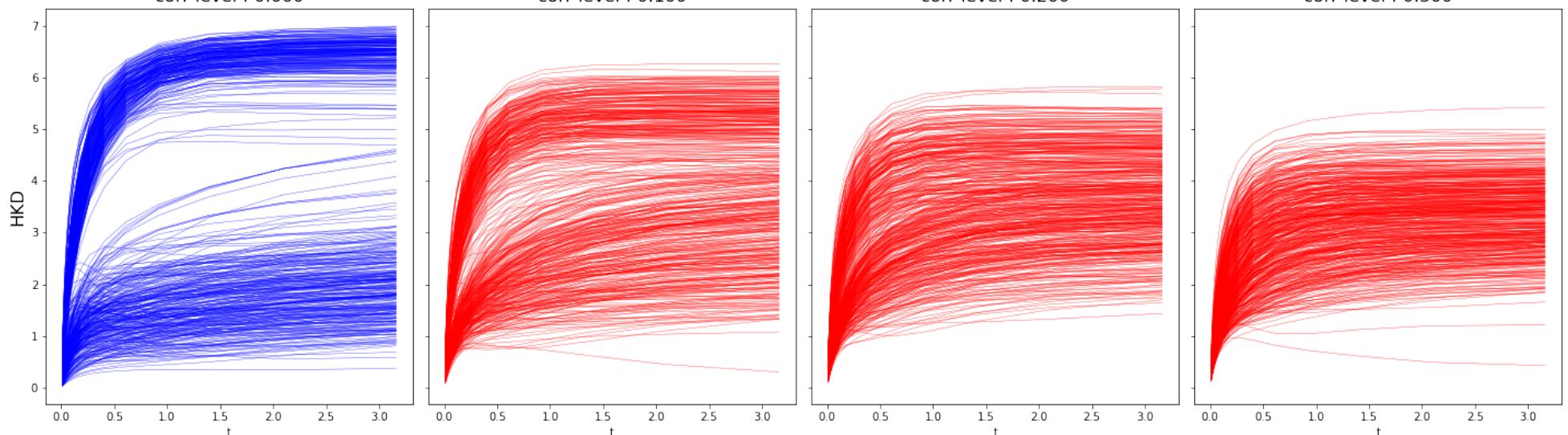
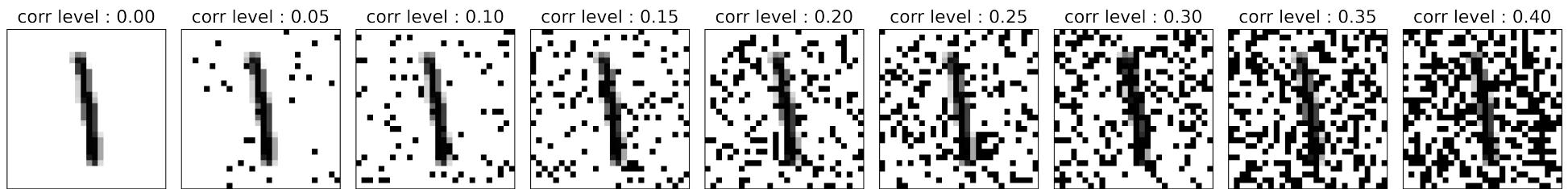
Out of distribution detection



two-sample test : ✓

Corrupted distribution (random pixels)

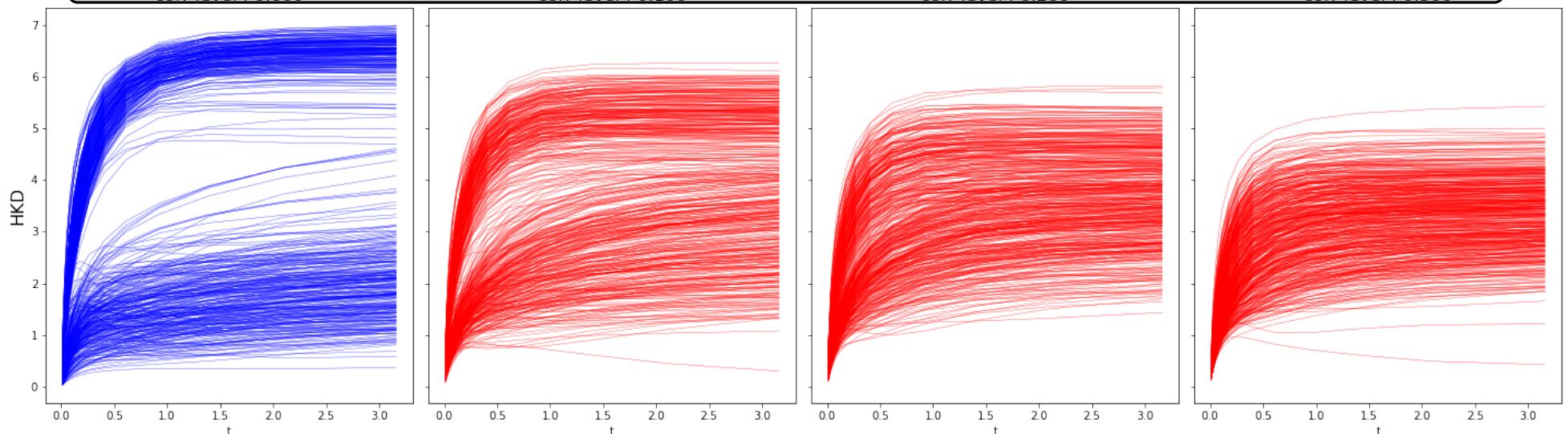
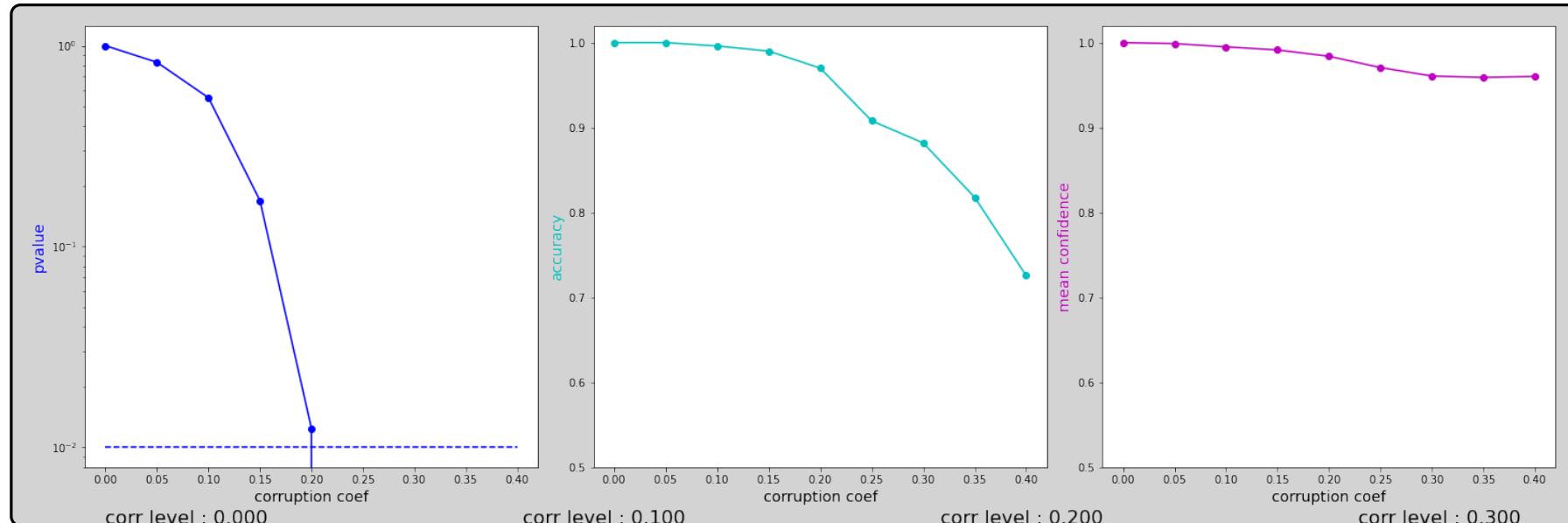
Training with digits : 0, 1.
Testing with corrupted digits.



Corrupted distribution (random pixels)

Training with digits : 0, 1.

Testing with corrupted digits.



Confidence bands

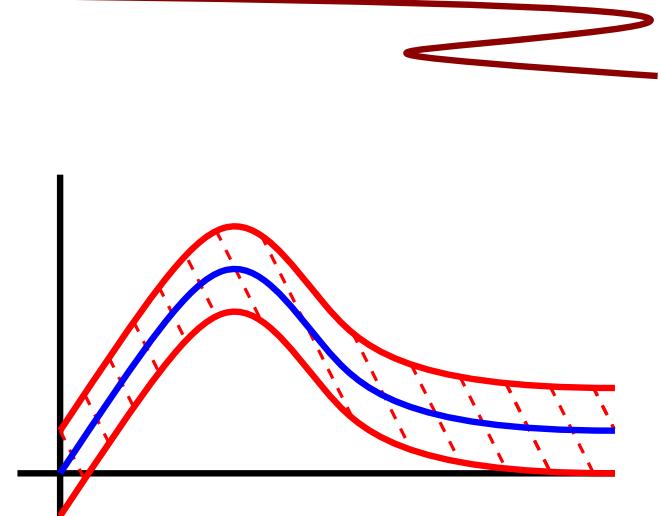
$$\mathcal{D} = (G_1, G'_1), \dots, (G_N, G'_N) \sim P$$

$$P_N = N^{-1} \sum_i \delta_{(G_i, G'_i)}$$

$$\alpha \in]0, 1[$$

$$\mathbb{P} (\|P_N D_{\cdot} - P D_{\cdot}\|_{\infty} \geq T_{\alpha, P}) \leq \alpha$$

unknown



From the Donsker property:

$$\sqrt{N}(P_N D_{\cdot} - P D_{\cdot}) \xrightarrow{\text{weak}} \mathbb{G} \xleftarrow{\text{weak}} \sqrt{N}(\hat{P}_N D_{\cdot} - P_N D_{\cdot}) \mid \mathcal{D}$$

\tilde{c}_{α} Monte Carlo estimator of c_{α} , s.t

$$\mathbb{P} \left(\|\hat{P}_N D_{\cdot} - P_N D_{\cdot}\|_{\infty} \geq c_{\alpha}/\sqrt{N} \mid \mathcal{D} \right) \leq \alpha.$$

$$\lim_{N \rightarrow \infty} \mathbb{P} \left(\|P_N D_{\cdot} - P D_{\cdot}\|_{\infty} \geq \tilde{c}_{\alpha}/\sqrt{N} \right) \leq \alpha$$

Two-sample Tests



$$X_1, \dots, X_N \sim P \text{ a sample}$$

$$P_N = N^{-1} \sum_i \delta_{X_i}$$

$$Y_1, \dots, Y_M \sim Q \text{ a sample}$$

$$Q_M = M^{-1} \sum_i \delta_{Y_i}$$

$$\mathcal{H}_0 : P = Q \quad \text{or} \quad \mathcal{H}_1 : P \neq Q$$

Idea : compute $T_{N,M} = \|P_N D_{\cdot} - Q_M D_{\cdot}\|_{\infty}$.

- reject \mathcal{H}_0 , if $T_{N,M} > \textcolor{red}{T}$
- accept \mathcal{H}_0 , otherwise

$$\mathbb{P}_{\mathcal{H}_0} (T_{N,M} > \textcolor{red}{T}) \leq \alpha$$

\tilde{c} : Monte-Carlo estimator of c , s.t.

$$\mathbb{P} \left(\|\hat{P}_N D_{\cdot} - \hat{Q}_M D_{\cdot}\|_{\infty} \geq \textcolor{red}{c} \frac{\sqrt{N+M}}{\sqrt{NM}} \mid \mathcal{D} \right) \leq \alpha.$$

resampled from

$$Z = (X_1, \dots, X_N, Y_1, \dots, Y_M)$$

$$\lim_{N,M \rightarrow \infty} \mathbb{P}_{\mathcal{H}_0} \left(T_{N,M} \geq \tilde{c} \frac{\sqrt{N+M}}{\sqrt{NM}} \right) \leq \alpha$$

if $P D_{\cdot} \neq Q D_{\cdot}$,

$$\lim_{N,M \rightarrow \infty} \mathbb{P}_{\mathcal{H}_1} \left(T_{N,M} \geq \tilde{c} \frac{\sqrt{N+M}}{\sqrt{NM}} \right) = 1$$