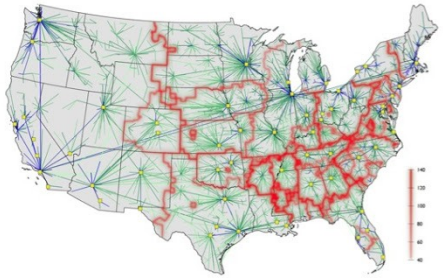
Entropic Optimal Transport in Random graphs

Nicolas Keriven

CNRS, GIPSA-lab

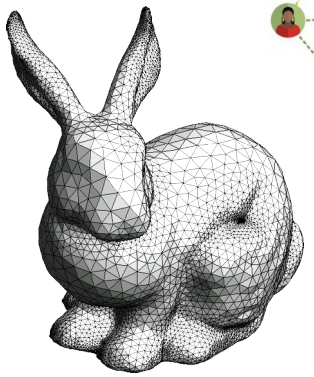
(Optimal) Transport in Graphs

- How to transport stuff on a road/computer/etc network...?



- How different two groups of people are in a social network? (w.r.t. unobserved preferences)

- How "far" apart are different parts of a manifold? (w.r.t. geodesic distance)



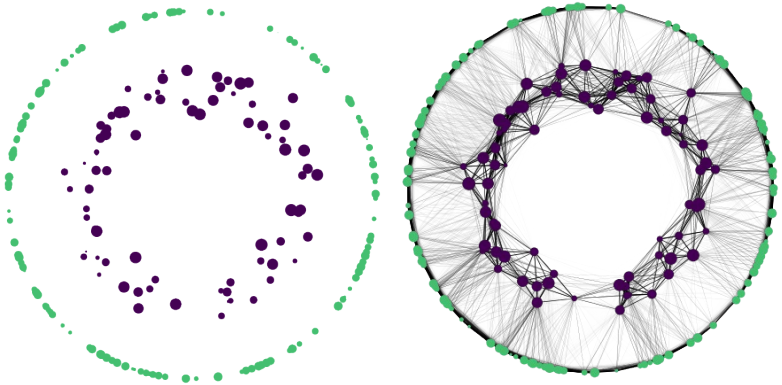
Entropic OT... in random graphs

Distributions	**Cost Matrix**
$\alpha \in \Delta_n$	$C \in \mathbb{R}_+^{n \times m}$
$\beta \in \Delta_m$	

Entropic OT... in random graphs

Distributions	**Cost Matrix**
$\alpha \in \Delta_n$	$C \in \mathbb{R}_+^{n \times m}$
$\beta \in \Delta_m$	

$$\{x_1, \dots, x_n\}$$
$$\{x_{n+1}, \dots, x_{n+m}\}$$
$$C = [c(x_i, x_{n+j})]_{ij}$$



Entropic OT... in random graphs

Distributions

$\alpha \in \Delta_n$

$\beta \in \Delta_m$

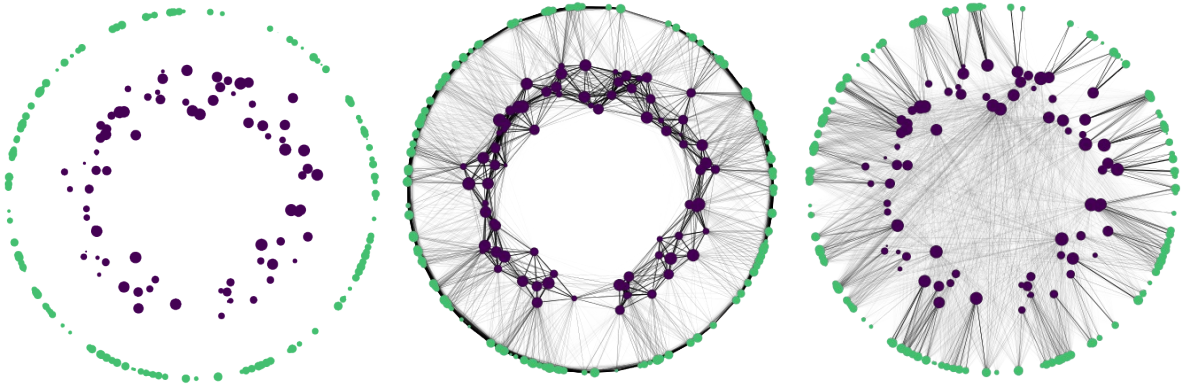
Cost Matrix

$C \in \mathbb{R}_+^{n \times m}$

$\{x_1, \ldots, x_n\}$

$\{x_{n+1}, \ldots, x_{n+m}\}$

$C = [c(x_i, x_{n+j})]_{ij}$



$$\mathcal{W}_\epsilon^C(\alpha, \beta) = \min_{P \in \Pi(\alpha,\beta)} \langle C, P \rangle + \epsilon KL(P|\alpha \otimes \beta)$$

NB: Sinkhorn only uses $K = e^{-C/\epsilon}$

Entropic OT... in random graphs

Distributions

$\alpha \in \Delta_n$

$\beta \in \Delta_m$

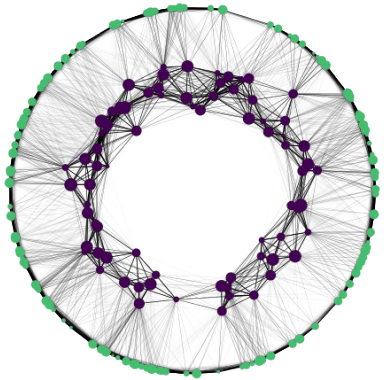
Cost Matrix

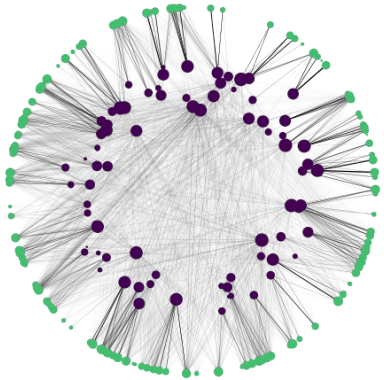
$C \in \mathbb{R}_+^{n \times m}$

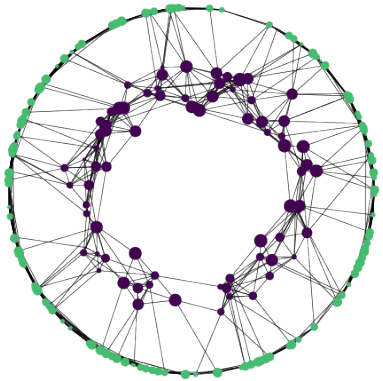
$\{x_1, \ldots, x_n\}$

$\{x_{n+1}, \ldots, x_{n+m}\}$

$C = [c(x_i, x_{n+j})]_{ij}$







Random graph/edges

$a_{ij} \sim \mathrm{Ber}(w_n(x_i, x_j))$

$$\mathcal{W}_\epsilon^C(\alpha, \beta) = \min_{P \in \Pi(\alpha,\beta)} \langle C, P \rangle + \epsilon KL(P|\alpha \otimes \beta)$$

NB: Sinkhorn only uses $K = e^{-C/\epsilon}$

Entropic OT... in random graphs

Distributions

$\alpha \in \Delta_n$

$\beta \in \Delta_m$

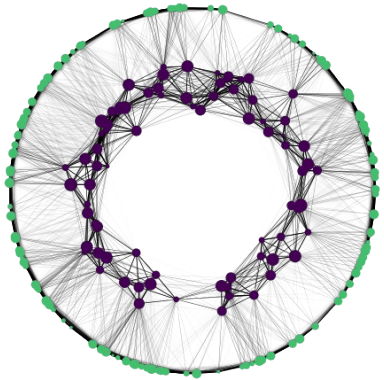
Cost Matrix

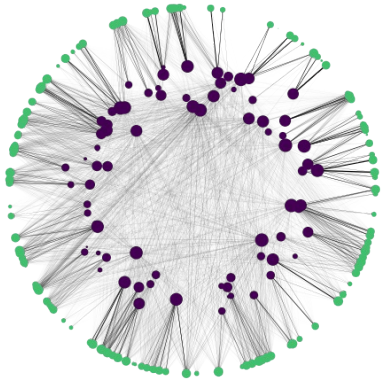
$C \in \mathbb{R}_+^{n \times m}$

$\{x_1, \dots, x_n\}$

$\{x_{n+1}, \dots, x_{n+m}\}$

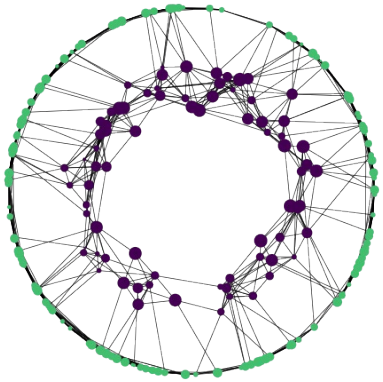
$C = [c(x_i, x_{n+j})]_{ij}$





$$\mathcal{W}_\epsilon^C(\alpha, \beta) = \min_{P \in \Pi(\alpha,\beta)} \langle C, P \rangle + \epsilon KL(P|\alpha \otimes \beta)$$

NB: Sinkhorn only uses $K = e^{-C/\epsilon}$



Random graph/edges

$$a_{ij} \sim \mathrm{Ber}(w_n(x_i, x_j))$$

- Estimate $\hat{C}$

Entropic OT... in random graphs

Distributions

$\alpha \in \Delta_n$

$\beta \in \Delta_m$

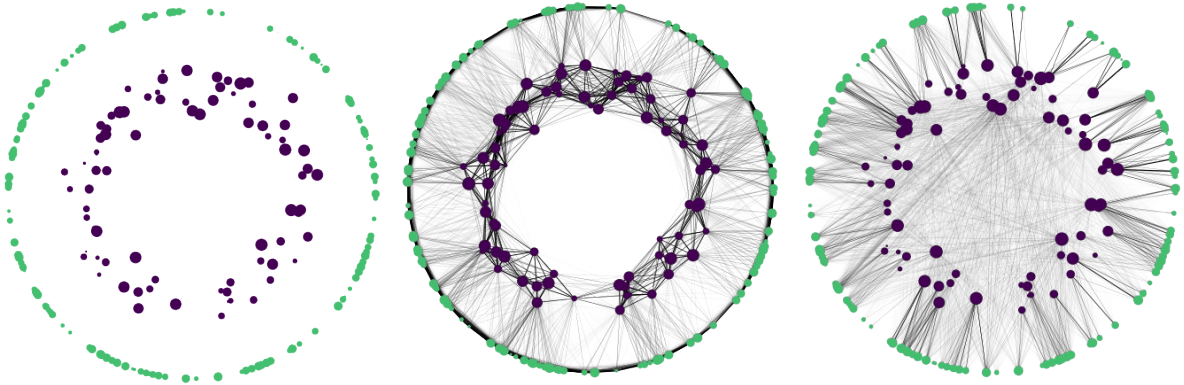
Cost Matrix

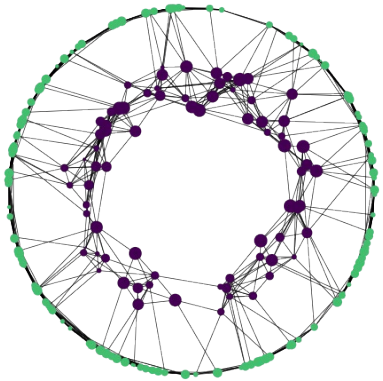
$C \in \mathbb{R}_+^{n \times m}$

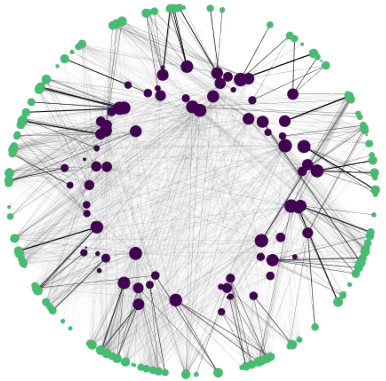
$\{x_1, \ldots, x_n\}$

$\{x_{n+1}, \ldots, x_{n+m}\}$

$C = [c(x_i, x_{n+j})]_{ij}$







Random graph/edges

$$a_{ij} \sim \mathrm{Ber}(w_n(x_i, x_j))$$

$$\mathcal{W}_\epsilon^C(\alpha, \beta) = \min_{P \in \Pi(\alpha,\beta)} \langle C, P \rangle + \epsilon KL(P|\alpha \otimes \beta)$$

NB: Sinkhorn only uses $K = e^{-C/\epsilon}$

- Estimate $\hat{C}$
- **How close is** $\mathcal{W}_\epsilon^{\hat{C}}(\alpha, \beta)$ **?**

Stability to inexact cost

Stability to inexact cost matrix?

Immediate: $\forall \epsilon \geq 0 \quad |\mathcal{W}_\epsilon^C(\alpha, \beta) - \mathcal{W}_\epsilon^{\hat{C}}(\alpha, \beta)| \leq \sup_P |\langle P, C - \hat{C} \rangle| \leq \|C - \hat{C}\|_\infty$

Stability to inexact cost

Stability to inexact cost matrix?

Immediate: $\forall \epsilon \geq 0 \quad |\mathcal{W}_\epsilon^C(\alpha, \beta) - \mathcal{W}_\epsilon^{\hat{C}}(\alpha, \beta)| \leq \sup_P |\langle P, C - \hat{C} \rangle| \leq \|C - \hat{C}\|_\infty$

May not be sufficient! E.g., obviously $\|A - \mathbb{E}A\|_\infty$ does not converge...

Stability to inexact cost

Stability to inexact cost matrix?

Immediate: $\forall \epsilon \geq 0 \quad |\mathcal{W}_\epsilon^C(\alpha,\beta) - \mathcal{W}_\epsilon^{\hat{C}}(\alpha,\beta)| \leq \sup_P |\langle P, C - \hat{C}\rangle| \leq \|C - \hat{C}\|_\infty$

May not be sufficient! E.g., obviously $\|A - \mathbb{E}A\|_\infty$ does not converge...

Theorem (K.): If $\ell \leq C_{ij}, \hat{C}_{ij} \leq L$ and $\alpha_i \lesssim \frac{1}{n}, \beta_j \lesssim \frac{1}{m}$

$\forall \epsilon > 0$

$$|\mathcal{W}_\epsilon^C(\alpha,\beta) - \mathcal{W}_\epsilon^{\hat{C}}(\alpha,\beta)| \lesssim \epsilon e^{(2L-\ell)/\epsilon} \frac{\|e^{-C/\epsilon} - e^{-\hat{C}/\epsilon}\|}{\sqrt{nm}}$$

$$\lesssim e^{2(L-\ell)/\epsilon} \frac{\|C - \hat{C}\|_F}{\sqrt{nm}}$$

Stability to inexact cost

Stability to inexact cost matrix?

Immediate: $\forall \epsilon \geq 0 \quad |\mathcal{W}_\epsilon^C(\alpha,\beta) - \mathcal{W}_\epsilon^{\hat{C}}(\alpha,\beta)| \leq \sup_P |\langle P, C - \hat{C}\rangle| \leq \|C - \hat{C}\|_\infty$

May not be sufficient! E.g., obviously $\|A - \mathbb{E}A\|_\infty$ does not converge...

Theorem (K.): If $\ell \leq C_{ij}, \hat{C}_{ij} \leq L$ and $\alpha_i \lesssim \frac{1}{n}, \beta_j \lesssim \frac{1}{m}$

$\forall \epsilon > 0$

$$|\mathcal{W}_\epsilon^C(\alpha,\beta) - \mathcal{W}_\epsilon^{\hat{C}}(\alpha,\beta)| \lesssim \epsilon e^{(2L-\ell)/\epsilon} \frac{\|e^{-C/\epsilon} - e^{-\hat{C}/\epsilon}\|}{\sqrt{nm}}$$

$$\lesssim e^{2(L-\ell)/\epsilon} \frac{\|C - \hat{C}\|_F}{\sqrt{nm}}$$

- Invariant to translating $C, \hat{C}$
- Exponential in ϵ
- First bound stronger, second bound more "usable"
- Proof: classical, bound the dual potentials
  (re-inventing the wheel?)

Application: geodesics on manifolds

RGs with "local kernels": close nodes are connected, radius decreases when #nodes increases

*Known: **weighted** **shortest paths** converge to geodesic distance*

[Bernstein et al. 2000]

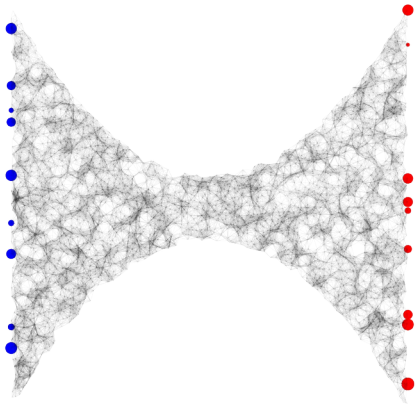
Application: geodesics on manifolds

RGs with "local kernels": close nodes are connected, radius decreases when #nodes increases

- k-manifold $\mathcal{M}_k \subset \mathbb{R}^d$ with geo. dist. $d_{\mathcal{M}}(x, y)$
- Fixed $\{x_1, \ldots, x_{n+m}\} \subset \mathcal{M}_k$
- Nodes $\{x_{n+m+1}, \ldots, x_N\} \overset{iid}{\sim} \nu$ with $N \to \infty$
- Kernel $w_N(x, y) = 1_{\|x-y\| \leq h_N}$ with $\frac{\log(1/h_N)}{N h_N^k} \to 0$

*Known: **weighted** **shortest paths** converge to geodesic distance*

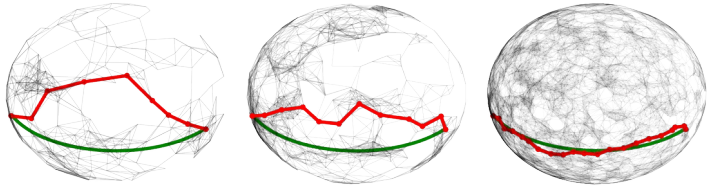
[Bernstein et al. 2000]



Application: geodesics on manifolds

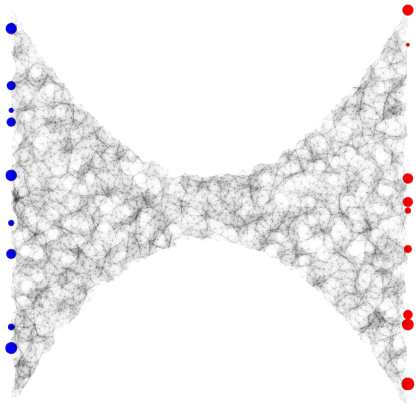
RGs with "local kernels": close nodes are connected, radius decreases when #nodes increases

- k-manifold $\mathcal{M}_k \subset \mathbb{R}^d$ with geo. dist. $d_{\mathcal{M}}(x, y)$
- Fixed $\{x_1, \dots, x_{n+m}\} \subset \mathcal{M}_k$
- Nodes $\{x_{n+m+1}, \dots, x_N\} \overset{iid}{\sim} \nu$ with $N \to \infty$
- Kernel $w_N(x, y) = 1_{\|x-y\| \leq h_N}$ with $\frac{\log(1/h_N)}{N h_N^k} \to 0$



*Known: **weighted** **shortest paths** converge to geodesic distance*

[Bernstein et al. 2000]



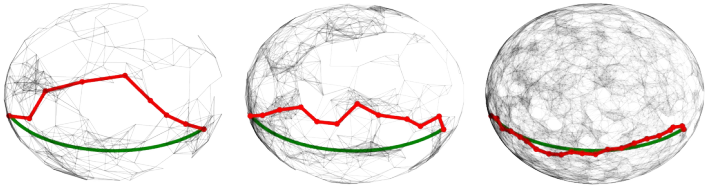
Theorem (K.): if ν has a lower-bounded density, whp

$$h_N \mathrm{SP}(v_i, v_{n+j}) = d_{\mathcal{M}}(x_i, x_{n+j}) + \mathcal{O}\left(\left(\frac{\log 1/h_N}{N h_N^k}\right)^{\frac{1}{k}}\right)$$

Application: geodesics on manifolds

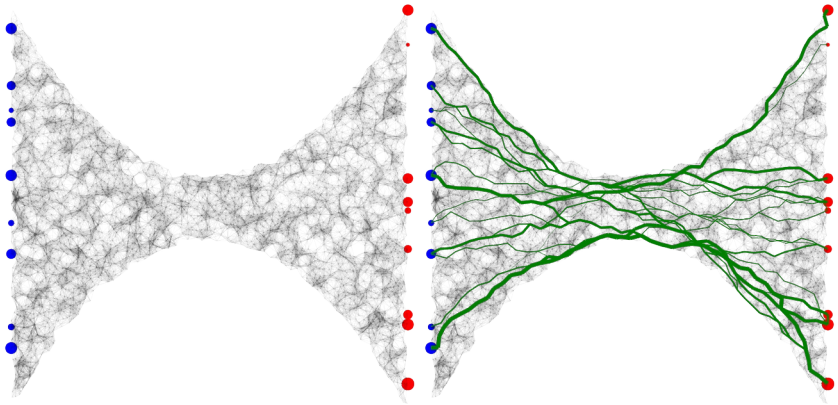
RGs with "local kernels": close nodes are connected, radius decreases when #nodes increases

- k-manifold $\mathcal{M}_k \subset \mathbb{R}^d$ with geo. dist. $d_{\mathcal{M}}(x, y)$
- Fixed $\{x_1, \dots, x_{n+m}\} \subset \mathcal{M}_k$
- Nodes $\{x_{n+m+1}, \dots, x_N\} \overset{iid}{\sim} \nu$ with $N \to \infty$
- Kernel $w_N(x, y) = 1_{\|x-y\| \leq h_N}$ with $\frac{\log(1/h_N)}{N h_N^k} \to 0$



*Known: **weighted shortest paths** converge to geodesic distance*

[Bernstein et al. 2000]



Corollary

$C_{ij} = f(d_{\mathcal{M}}(x_i, x_{n+j}))$

leads to

$\|\hat{C}_{\mathrm{SP}} - C\|_\infty \to 0$

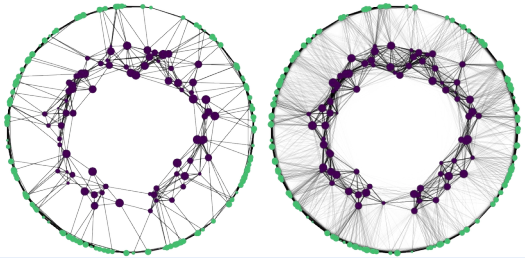
Theorem (K.): if ν has a lower-bounded density, whp

$$h_N \mathrm{SP}(v_i, v_{n+j}) = d_{\mathcal{M}}(x_i, x_{n+j}) + \mathcal{O}\left(\left(\frac{\log 1/h_N}{N h_N^k}\right)^{\frac{1}{k}}\right)$$

Application: USVT estimator

RGs with "nonlocal kernels": fixed kernel, multiplying factor decreases when #nodes increases

- Nodes $\{x_1, \dots, x_{n+m}\}$
  with $n \sim m \to \infty$
- Kernel $w_n(x, y) = \rho_n w(x, y)$
  with $\rho_n \gtrsim (\log n)/n$
  and psd kernel
- Cost $c(x, y) = f(w(x, y))$
  with Lipschitz f



Application: USVT estimator

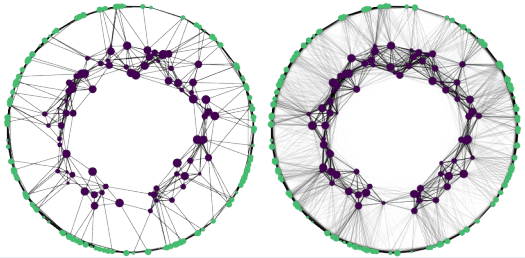
RGs with "nonlocal kernels": fixed kernel, multiplying factor decreases when #nodes increases

- Nodes $\{x_1, \dots, x_{n+m}\}$
  with $n \sim m \to \infty$
- Kernel $w_n(x,y) = \rho_n w(x,y)$
  with $\rho_n \gtrsim (\log n)/n$
  and psd kernel
- Cost $c(x,y) = f(w(x,y))$
  with Lipschitz f

[Lei&Rinaldo 2015]

Pbm: $\frac{1}{n}\|A/\rho_n - W\| \lesssim (n\rho_n)^{-\frac{1}{2}}$

but $\frac{1}{n}\|A/\rho_n - W\|_F \not\to 0$



Application: USVT estimator

RGs with "nonlocal kernels": fixed kernel, multiplying factor decreases when #nodes increases

- Nodes $\{x_1, \dots, x_{n+m}\}$ with $n \sim m \to \infty$
- Kernel $w_n(x,y) = \rho_n w(x,y)$ with $\rho_n \gtrsim (\log n)/n$ and psd kernel
- Cost $c(x,y) = f(w(x,y))$ with Lipschitz f

Universal Singular Value Thresholding (USVT) [Chatterjee 2015]

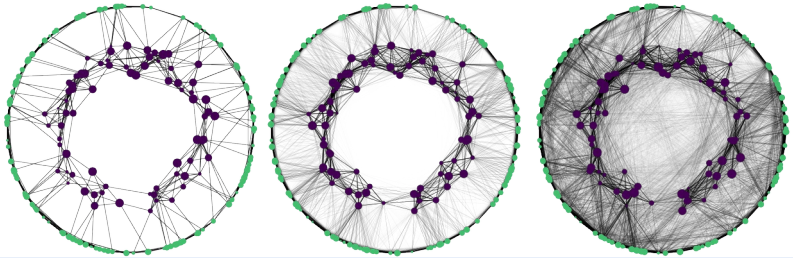
- Diagonalize $A = \sum_i \sigma_i a_i a_i^\top$

$$\hat{W}_\gamma = \mathrm{cut}_{[w_{\min}, w_{\max}]}(\rho_n^{-1} \sum_{\sigma_i \geq \gamma\sqrt{\rho_n n}} \sigma_i a_i a_i^\top)$$

[Lei&Rinaldo 2015]

Pbm: $\frac{1}{n}\|A/\rho_n - W\| \lesssim (n\rho_n)^{-\frac{1}{2}}$

but $\frac{1}{n}\|A/\rho_n - W\|_F \not\to 0$



Application: USVT estimator

RGs with "nonlocal kernels": fixed kernel, multiplying factor decreases when #nodes increases

- Nodes $\{x_1, \dots, x_{n+m}\}$ with $n \sim m \to \infty$
- Kernel $w_n(x,y) = \rho_n w(x,y)$ with $\rho_n \gtrsim (\log n)/n$ and psd kernel
- Cost $c(x,y) = f(w(x,y))$ with Lipschitz f

[Lei&Rinaldo 2015]

Pbm: $\frac{1}{n}\|A/\rho_n - W\| \lesssim (n\rho_n)^{-\frac{1}{2}}$

but $\frac{1}{n}\|A/\rho_n - W\|_F \not\to 0$

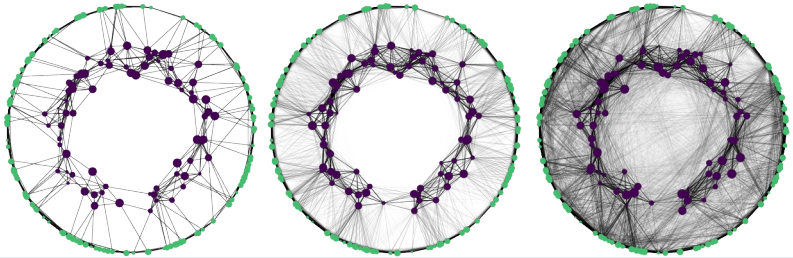
Universal Singular Value Thresholding (USVT)

[Chatterjee 2015]

- Diagonalize $A = \sum_i \sigma_i a_i a_i^\top$

$$\hat{W}_\gamma = \mathrm{cut}_{[w_{\min}, w_{\max}]}(\rho_n^{-1} \sum_{\sigma_i \geq \gamma\sqrt{\rho_n n}} \sigma_i a_i a_i^\top)$$

Theorem (K.): for all $r > 0$, there is γ_r such that, with proba $1 - n^{-r}$, $\frac{1}{n}\|\hat{W}_{\gamma_r} - W\|_F \lesssim (n\rho_n)^{-1/4}$



Application: USVT estimator

RGs with "nonlocal kernels": fixed kernel, multiplying factor decreases when #nodes increases

- Nodes $\{x_1, \dots, x_{n+m}\}$ with $n \sim m \to \infty$
- Kernel $w_n(x,y) = \rho_n w(x,y)$ with $\rho_n \gtrsim (\log n)/n$ and psd kernel
- Cost $c(x,y) = f(w(x,y))$ with Lipschitz f

[Lei&Rinaldo 2015]

Pbm: $\frac{1}{n}\|A/\rho_n - W\| \lesssim (n\rho_n)^{-\frac{1}{2}}$

but $\frac{1}{n}\|A/\rho_n - W\|_F \not\to 0$

Universal Singular Value Thresholding (USVT)

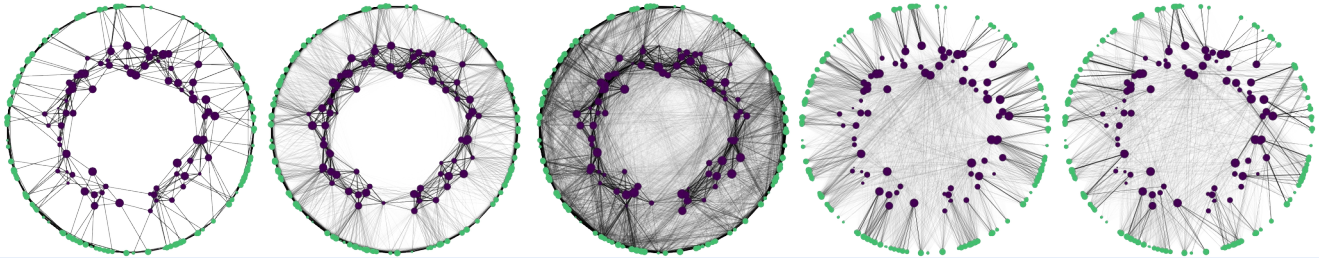
[Chatterjee 2015]

- Diagonalize $A = \sum_i \sigma_i a_i a_i^\top$

$$\hat{W}_\gamma = \mathrm{cut}_{[w_{\min}, w_{\max}]}(\rho_n^{-1} \sum_{\sigma_i \geq \gamma\sqrt{\rho_n n}} \sigma_i a_i a_i^\top)$$

Theorem (K.): for all $r > 0$, there is γ_r such that, with proba $1 - n^{-r}$, $\frac{1}{n}\|\hat{W}_{\gamma_r} - W\|_F \lesssim (n\rho_n)^{-1/4}$

Corollary: $|\mathcal{W}_\epsilon^{\hat{C}_{\gamma_r}}(\alpha,\beta) - \mathcal{W}_\epsilon^{C}(\alpha,\beta)| \lesssim e^{2(L-\ell)}(n\rho_n)^{-1/4}$



Application: “fast” rate

When $w(x,y) = e^{-\frac{\|x-y\|^p}{\sigma}}$, the matrix W is **directly** the “Sinkhorn” matrix $K = e^{-C/\sigma}$

when $\epsilon = \sigma$ **and** $c(x,y) = \|x-y\|^p$

Application: "fast" rate

When $w(x,y) = e^{-\frac{\|x-y\|^p}{\sigma}}$, the matrix W is **directly** the "Sinkhorn" matrix $K = e^{-C/\sigma}$

when $\epsilon = \sigma$ **and** $c(x,y) = \|x-y\|^p$

Theorem (K.):

Defining $\mathcal{L}_\epsilon^K(\alpha,\beta) = \max_{f,g} f^\top \alpha + g^\top \beta - \epsilon(e^{\frac{f}{\epsilon}} \odot \alpha)^\top K (e^{\frac{g}{\epsilon}} \odot \beta) + \epsilon$

the dual OT cost with matrix K, whp

(plus some bounding conditions on the potentials)

$$|\mathcal{L}_\sigma^{A/\rho_n}(\alpha,\beta) - \mathcal{W}_\sigma^C(\alpha,\beta)| \lesssim (n\rho_n)^{-1/2}$$

Application: "fast" rate

When $w(x,y) = e^{-\frac{\|x-y\|^p}{\sigma}}$, the matrix W is **directly** the "Sinkhorn" matrix $K = e^{-C/\sigma}$

when $\epsilon = \sigma$ **and** $c(x,y) = \|x-y\|^p$

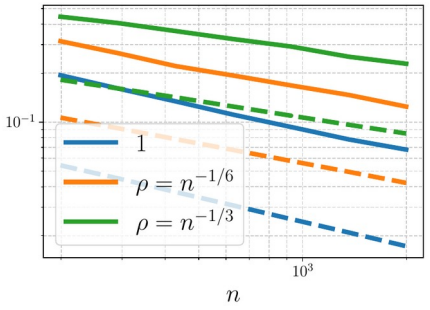
Theorem (K.):

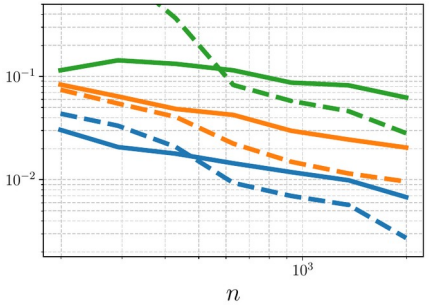
Defining $\mathcal{L}_\epsilon^K(\alpha,\beta) = \max_{f,g} f^\top \alpha + g^\top \beta - \epsilon(e^{\frac{f}{\epsilon}} \odot \alpha)^\top K (e^{\frac{g}{\epsilon}} \odot \beta) + \epsilon$

the dual OT cost with matrix K, whp

(plus some bounding conditions on the potentials)

$$|\mathcal{L}_\sigma^{A/\rho_n}(\alpha,\beta) - \mathcal{W}_\sigma^C(\alpha,\beta)| \lesssim (n\rho_n)^{-1/2}$$





Conclusion

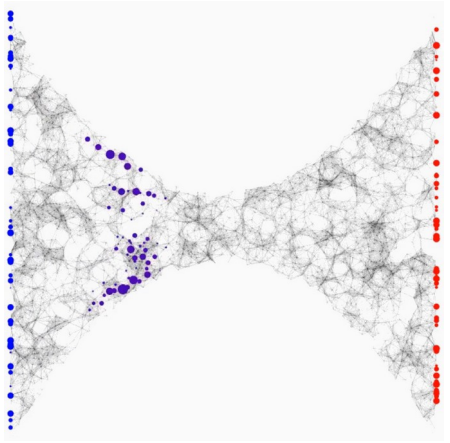
- OT can be done when the cost matrix is not known exactly
- *Maybe "reinventing the wheel" a bit, but* interesting results in the context of random graphs

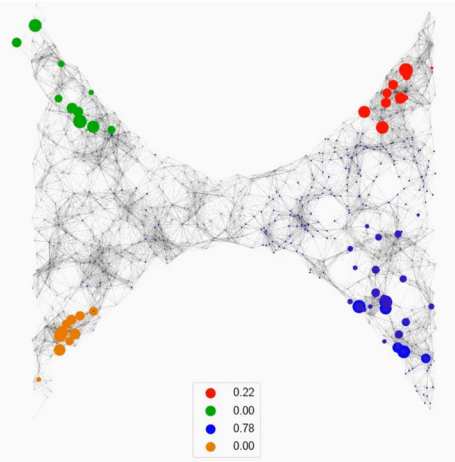
Conclusion

- OT can be done when the cost matrix is not known exactly
- *Maybe "reinventing the wheel" a bit, but* interesting results in the context of random graphs
- First steps, many **outlooks**:
  - More integrated, data-driven way of estimating the cost?

Conclusion

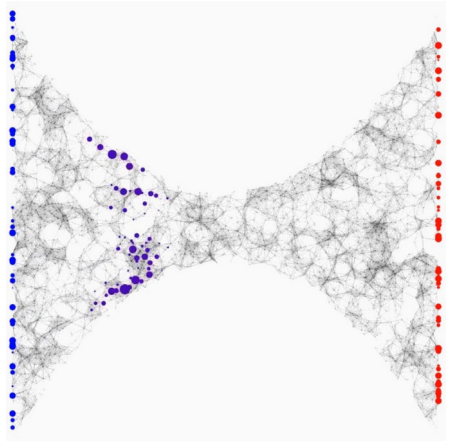
- OT can be done when the cost matrix is not known exactly
- *Maybe "reinventing the wheel" a bit, but* interesting results in the context of random graphs
- First steps, many **outlooks**:
  - More integrated, data-driven way of estimating the cost?
  - The **"by-products" of OT** are often more interesting than the OT distance!

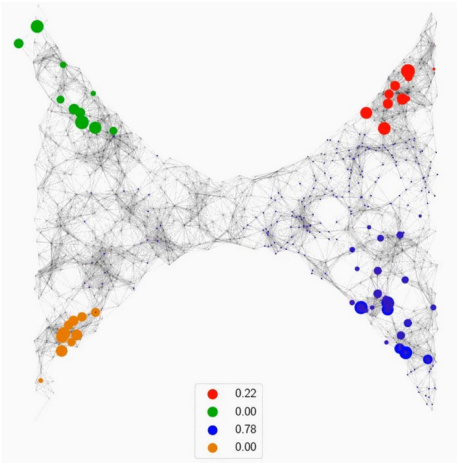




Conclusion

- OT can be done when the cost matrix is not known exactly
- *Maybe "reinventing the wheel" a bit, but* interesting results in the context of random graphs
- First steps, many **outlooks**:
  - More integrated, data-driven way of estimating the cost?
  - The **"by-products" of OT** are often more interesting than the OT distance!





Preprint coming soon

nkeriven.github.io

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR — ANR

GRandMa