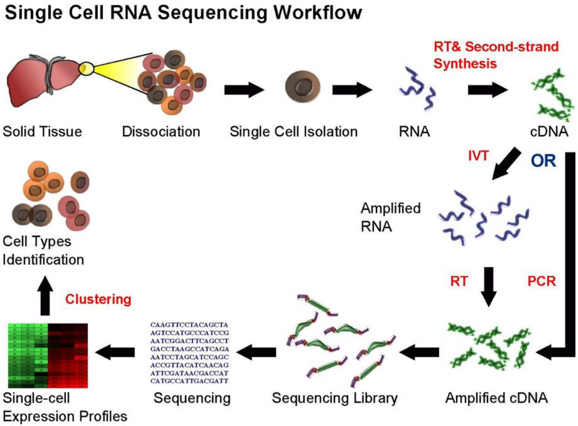
Reconstruction de trajectoires cellulaires par Transport Optimal

David ROBIN

Stage de Licence

Juin-Août 2018

Séquençage ARN sur cellule unique

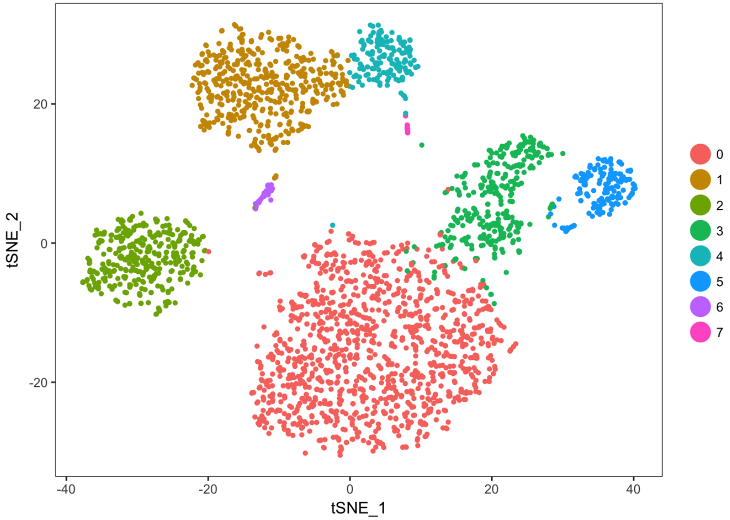


Processus scRNA-Seq destructif → pas d'évolution temporelle

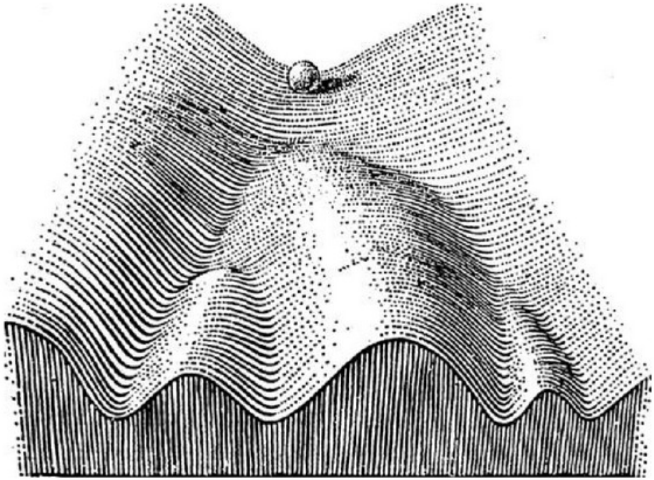
Espace d'expression génétique

	Cellule 1	Cellule 2	Cellule 3	Cellule 4	···
gene 1	32	33	124	10	···
gene 2	12	14	174	14	···
gene 3	14	12	12	12	···
gene 4	121	131	12	14	···
⋮	⋮	⋮	⋮	⋮	

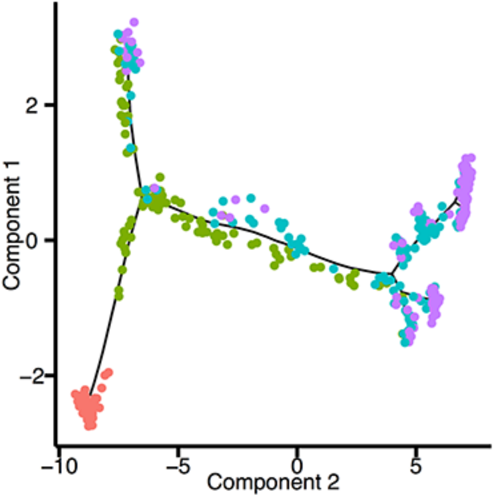
Espace d'expression génétique



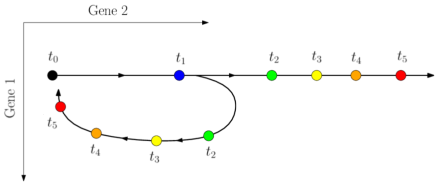
Paysages de Waddington



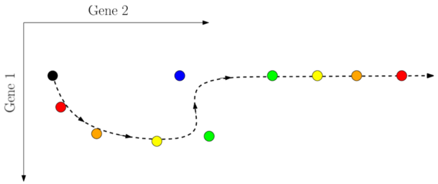
Pseudotemps



Échec du pseudotemps



Trajectoire réelle



Trajectoire inférée

Modélisation

- Ensemble des gènes: G
- Cellule: $x \in \mathbb{R}^G$
- Population: $\mathbb{P}_t = \sum_{i<n} \delta_{x_i}, \quad x_i \in \mathbb{R}^G$

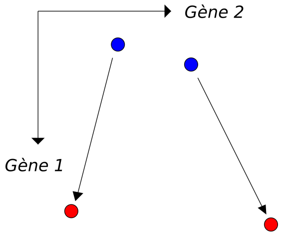
Modélisation

- Ensemble des gènes: G
- Cellule: $x \in \mathbb{R}^G$
- Population: $\mathbb{P}_t = \sum_{i<n} \delta_{x_i}, \quad x_i \in \mathbb{R}^G$
- Couplages entre $\mathbb{P}$ et $\mathbb{Q}$: $\pi \in \Gamma(\mathbb{P},\mathbb{Q})$

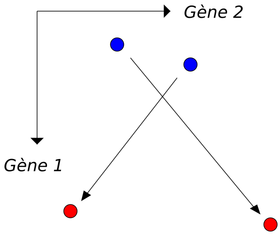
$$\forall A \subset \mathbb{R}^G, \quad \int_{x \in A} \int_{y \in \mathbb{R}^G} \pi(x,y) \cdot dxdy = \mathbb{P}(A)$$

$$\forall B \subset \mathbb{R}^G, \quad \int_{y \in B} \int_{x \in \mathbb{R}^G} \pi(x,y) \cdot dxdy = \mathbb{Q}(B)$$

L'hypothèse du Transport Optimal



faible modification du gène 2



forte modification du gène 2

Couplage et carte de transport

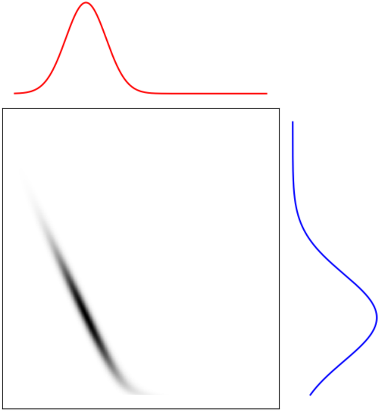


Image par une carte π:

$$\pi_\# : \mu \mapsto \int_{x\in\mathbb{R}^G} \frac{\pi(x,\cdot)}{\mathbb{P}(x)} \cdot d\mu(x)$$

$$\pi_\#\mathbb{P} = \mathbb{Q}$$

Pré-image par une carte π:

$$\pi^\# : \nu \mapsto \int_{y\in\mathbb{R}^G} \frac{\pi(\cdot,y)}{\mathbb{Q}(y)} \cdot d\nu(x)$$

$$\pi^\#\mathbb{Q} = \mathbb{P}$$

Recherche de couplage

Coût total du transport:

$$\int_x \int_y c(x,y) \cdot \pi(x,y) \cdot dxdy$$

En version discrète:

$$\langle\, c \,,\, \pi \,\rangle = \sum_x \sum_y c(x,y) \cdot \pi(x,y)$$

Problème de minimisation:

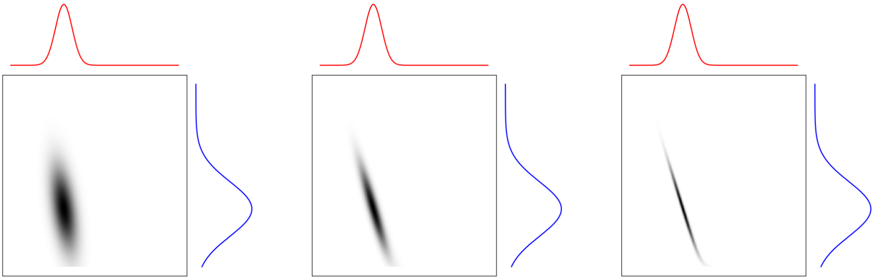
$$\min_{\pi \in \Gamma(\mathbb{P},\mathbb{Q})} \langle\, c \,,\, \pi \,\rangle$$

$$\text{où} \quad \Gamma(\mathbb{P}\,,\mathbb{Q}) = \{\, \pi \;:\; \pi \cdot 1 = \mathbb{P}\,, \pi^T \cdot 1 = \mathbb{Q} \,\}$$

Régularisation

Problème régularisé:

$$\min_{\pi \in \Gamma(\mathbb{P},\mathbb{Q})} \langle\, c\,,\pi\,\rangle - \varepsilon\mathcal{H}(\pi)$$



$\varepsilon = 10^4$ $\varepsilon = 10^3$ $\varepsilon = 10^2$

Croissance cellulaire

Hypothèse:

la position d'une cellule détermine son taux de croissance

$$r(x,y) \cdot g(x)^{\Delta t}$$

Nouvelles marginales:

$$\mathbb{P} \odot g^{\Delta t} \quad \text{et} \quad \overline{g} \cdot \mathbb{Q}$$

Transport non-équilibré

$$\min_{\pi} \langle\, c\,, \pi\,\rangle - \varepsilon\mathcal{H}(\pi) + \lambda_1 \cdot \mathrm{KL}(\pi \cdot 1 \,||\, \mathbb{P} \odot g^{\Delta t}) + \lambda_2 \cdot \mathrm{KL}(\pi^T \cdot 1 \,||\, \overline{g}\,\mathbb{Q})$$

avec KL la divergence de Kullback-Leibler:

$$\mathrm{KL}(P \,||\, Q) = \sum_x P(x) \cdot \log \frac{P(x)}{Q(x)}$$

Algorithme de Sinkhorn

$$\min_{\pi \in \Gamma(\mathbb{P},\mathbb{Q})} \langle\, c\,, \pi\,\rangle - \varepsilon \mathcal{H}(\pi) = \min_{\pi \in \Gamma(\mathbb{P},\mathbb{Q})} \varepsilon\, \mathrm{KL}(\pi \,||\, K)$$

$$\text{où} \quad K = \exp(-c/\varepsilon)$$

On projette alternativement sur $\mathcal{C}_1$ et $\mathcal{C}_2$:

$$\mathcal{C}_1 = \{\, \pi \;:\; \pi \cdot 1 = \mathbb{P} \,\} \qquad \text{et} \qquad \mathcal{C}_2 = \{\, \pi \;:\; \pi^T \cdot 1 = \mathbb{Q} \,\}$$

$$\mathcal{P}_{\mathcal{C}_1}(\pi) = \mathrm{diag}\left(\frac{\mathbb{P}}{\pi \cdot 1}\right) \cdot \pi \qquad \text{et} \qquad \mathcal{P}_{\mathcal{C}_2}(\pi) = \pi \cdot \mathrm{diag}\left(\frac{\mathbb{Q}}{\pi^T \cdot 1}\right)$$

Benamou et al., 2015,

Iterative Bregman projections for Regularized Transportation Problems

Algorithme de Sinkhorn

$$\pi^{(n)} = \mathrm{diag}(a^{(n)}) \cdot K \cdot \mathrm{diag}(b^{(n)})$$

$$a^{(n)} = \frac{\mathbb{P}}{K \cdot b^{(n)}} \qquad \text{et} \qquad b^{(n+1)} = \frac{\mathbb{Q}}{K^T \cdot a^{(n)}}$$

Extension au transport non-équilibré

$$\min_{\pi} \langle\, c \,, \pi \,\rangle \;-\; \varepsilon \mathcal{H}(\pi) \;+\; F_1(\pi \cdot 1 \,, \mathbb{P}) \;+\; F_2(\pi^T \cdot 1 \,, \mathbb{Q})$$

$$a^{(n)} = \frac{\mathrm{prox}_{F_1/\varepsilon}(K \cdot b^{(n)})}{K \cdot b^{(n)}} \qquad \text{et} \qquad b^{(n+1)} = \frac{\mathrm{prox}_{F_2/\varepsilon}(K^T \cdot a^{(n)})}{K^T \cdot a^{(n)}}$$

$$\text{où} \quad \mathrm{prox}_{F/\varepsilon}(z) = \operatorname*{argmin}_{s} \; F(s) + \varepsilon \, \mathrm{KL}(s \,||\, z)$$

Chizat et al., *Scaling algorithms for Unbalanced Transport Problems*, 2016 (ArXiv)

Extension au transport non-équilibré

$$F_1 = \lambda_1 \, \mathrm{KL}(\,\cdot\,||\,\mathbb{P}) \qquad \text{et} \qquad F_2 = \lambda_2 \, \mathrm{KL}(\,\cdot\,||\,\mathbb{Q})$$

donnent

$$a^{(n)} = \left(\frac{\mathbb{P}}{K \cdot b^{(n)}} \right)^{\frac{\lambda_1}{\lambda_1 + \varepsilon}} \qquad \text{et} \qquad b^{(n+1)} = \left(\frac{\mathbb{Q}}{K^T \cdot a^{(n)}} \right)^{\frac{\lambda_2}{\lambda_2 + \varepsilon}}$$

Réduction de dimension

- Calcul de la matrice de coût: $\mathcal{O}(n^2 G)$
- *Analyse en composante principale* pour diminuer G
- Annule les dimensions ne contenant que du bruit

Stabilisation logarithmique

Stabilisation des variables:

$$a = \tilde{a} \cdot \exp\left(u/\varepsilon\right) \qquad \text{et} \qquad b = \tilde{b} \cdot \exp\left(v/\varepsilon\right)$$

Noyau stabilisé:

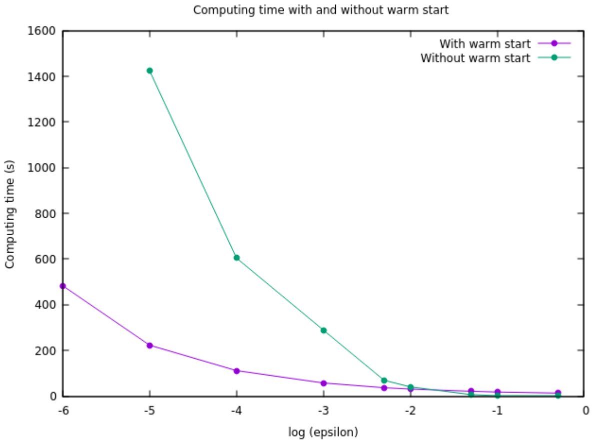
$$\tilde{K}_{i,j} = \exp(\,(\,u_i + v_j - c_{i,j}\,)\,/\,\varepsilon\,)$$

Nouvelles itérations:

$$\tilde{a}^{(n)} = \left(\frac{\mathbb{P}}{K \cdot \tilde{b}^{(n)}}\right)^{\frac{\lambda_1}{\lambda_1+\varepsilon}} \odot \exp(-\frac{u}{\lambda_1+\varepsilon})$$

$$\tilde{b}^{(n+1)} = \left(\frac{\mathbb{Q}}{K^T \cdot \tilde{a}^{(n)}}\right)^{\frac{\lambda_2}{\lambda_2+\varepsilon}} \odot \exp(-\frac{v}{\lambda_2+\varepsilon})$$

Départ à chaud



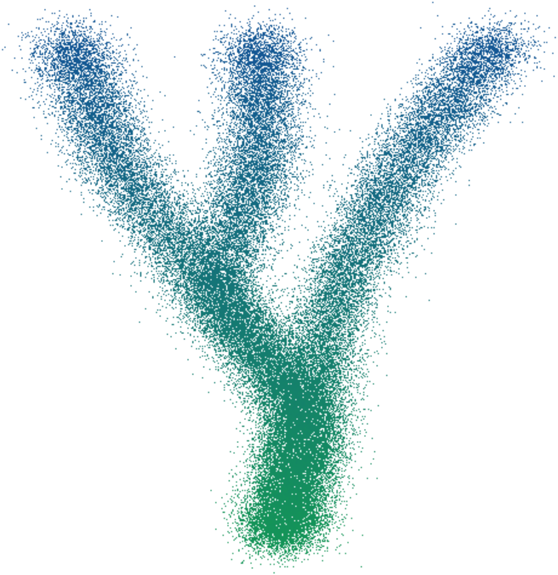
Interpolation

Population interpolée entre P et Q à la fraction α:

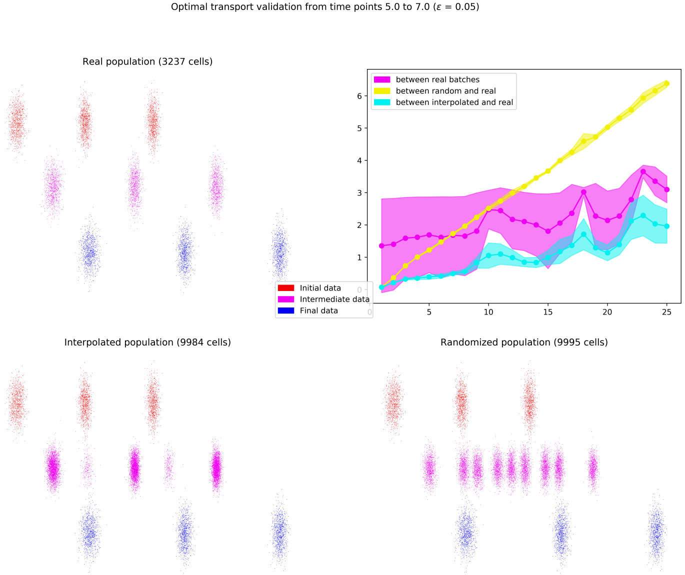
$$\mu : I_{i,j} \mapsto \tilde{\pi}_{i,j} = r_{i,j} \cdot g_i^{\alpha \cdot \Delta t}$$

$$\text{où} \quad I_{i,j} = (1 - \alpha) \cdot P_i + \alpha \cdot Q_j$$

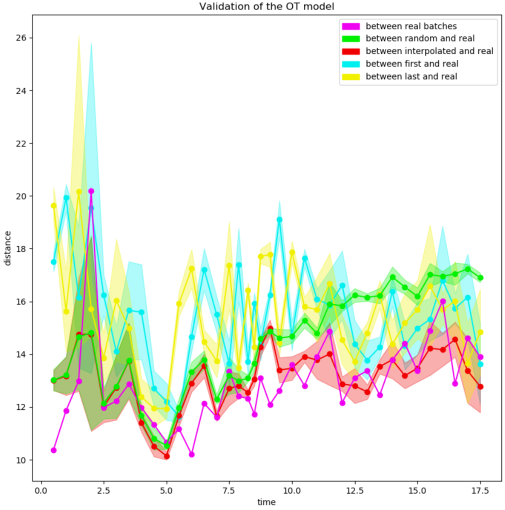
Simulation



Résultats sur simulation



Résultats sur données réelles

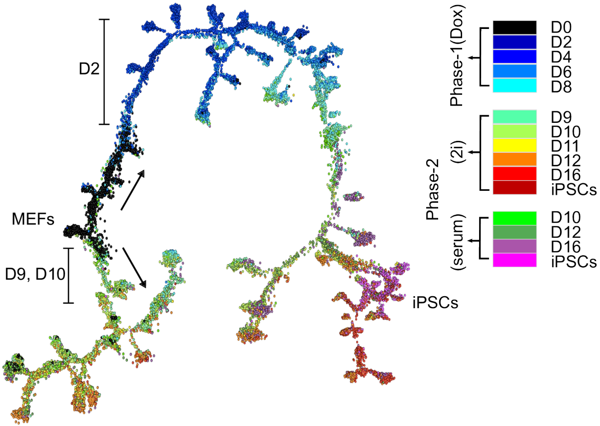


Reconstruction de trajectoires cellulaires

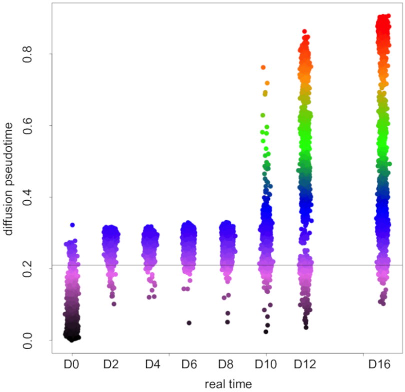
—

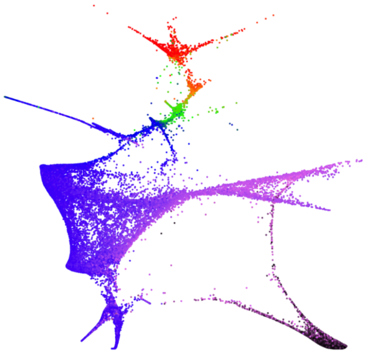
Annexes

MEF: what goes wrong without time

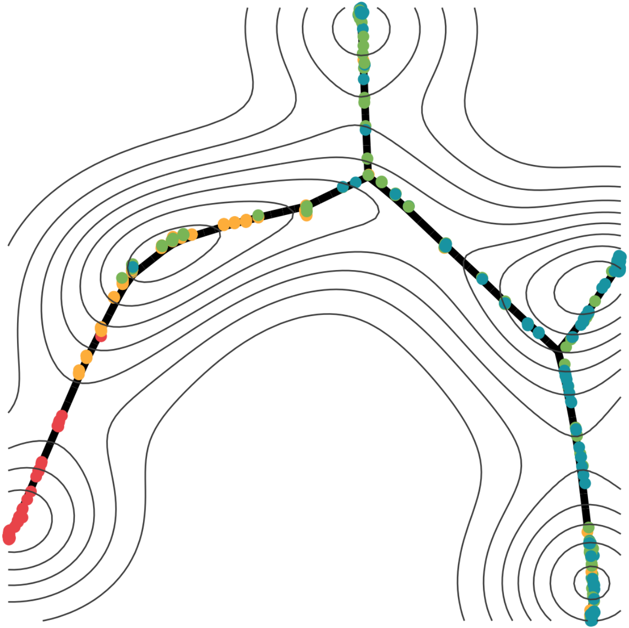


MEF: what goes wrong without time

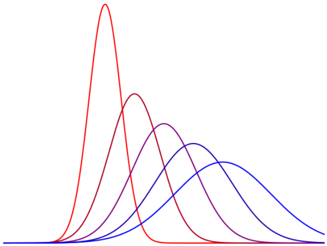




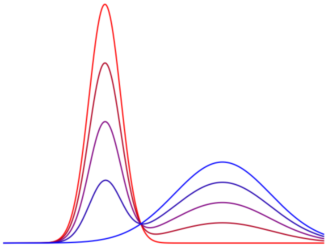
Monocle: Pseudotime



Interpolation



Interpolation par transport



Interpolation par moyenne pondérée

Transfert de couleurs

Image originale

Résultat

Couleurs à transférer

Algorithme de Sinkhorn stabilisé

function $\textsc{SinkhornSRUT}$ (c, p, q, ε, λ_1, λ_2)

$(\tilde{b}, u, v) \leftarrow (1, 0, 0)$

$\tilde{K}_{i,j} \leftarrow \exp(-c_{i,j}/\varepsilon) \quad \forall i,j$

repeat

$\tilde{a} \leftarrow \left(\frac{p}{K \cdot \tilde{b}}\right)^{\frac{\lambda_1}{\lambda_1+\varepsilon}} \odot \exp(-\frac{u}{\lambda_1+\varepsilon})$

$\tilde{b} \leftarrow \left(\frac{q}{K^T \cdot \tilde{a}}\right)^{\frac{\lambda_2}{\lambda_2+\varepsilon}} \odot \exp(-\frac{v}{\lambda_2+\varepsilon})$

if a component of $\mid \tilde{a} \mid$ or $\mid \tilde{b} \mid$ is "too big" **then**

$(u, v) \leftarrow (u + \varepsilon \log \tilde{a},\ v + \varepsilon \log \tilde{b})$

$\tilde{K}_{i,j} \leftarrow \exp((u_i + v_j - c_{i,j}) / \varepsilon) \quad \forall i,j$

$\tilde{b} \leftarrow 1$

until stopping criterion

return $(\, \tilde{a}_i \tilde{K}_{i,j} \tilde{b}_j \,)_{i,j}$

Force-directed layout embedding