Random Walks in Graphs

Thomas Bonald

Stage LIESSE
2018

Schedule

- **9:30 - 12:30** Tutorial
- **12:30 - 13:30** Lunch
- **13:30 - 17:00** Lab session (python)

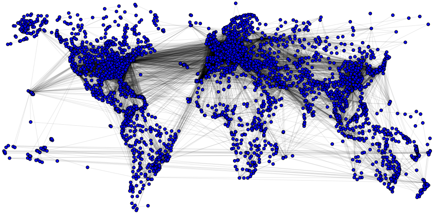
Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...

Main European highways

Graph data

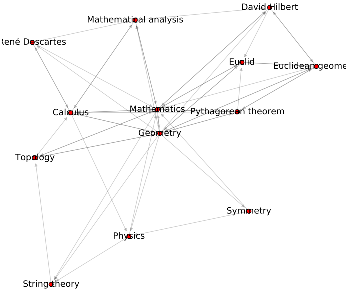
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...



International flights

Graph data

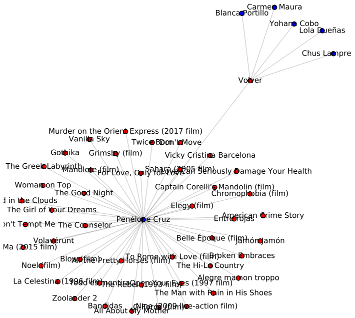
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...



Extract from Wikipedia

Graph data

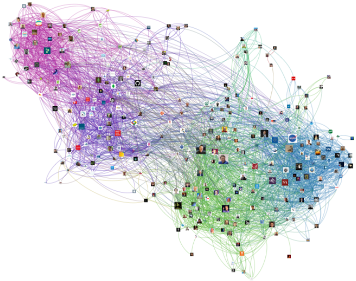
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...



Extract from the movie-actor graph

Graph data

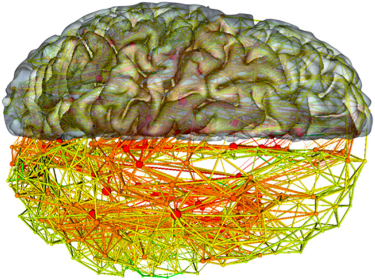
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...



Extract from Twitter
Source: AllThingsGraphed.com

Graph data

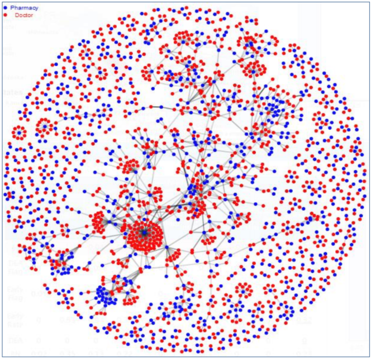
- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...
- **Biology:** brain, proteins, phylogenetics, ...



The brain network
Source: Wired

Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...
- **Biology:** brain, proteins, phylogenetics, ...
- **Health:** genetic diseases, patient-doctor-pharmacy-drugs, ...



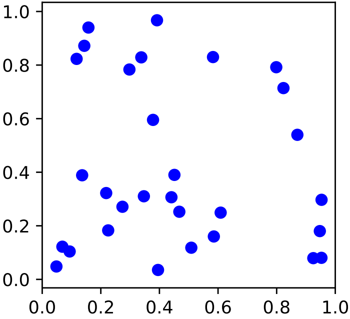
Pharmacy-doctor network
Source: IAAI 2015

Graph data

- **Infrastructure:** roads, railways, power grid, internet, ...
- **Communication:** phone, emails, flights, ...
- **Information:** Web, Wikipedia, knowledge bases, ...
- **Social networks:** Facebook, Twitter, LinkedIn, ...
- **Biology:** brain, proteins, phylogenetics, ...
- **Health:** genetic diseases, patient-doctor-pharmacy-drugs, ...
- **Marketing:** customer-product, bundling, ...

Data as graph

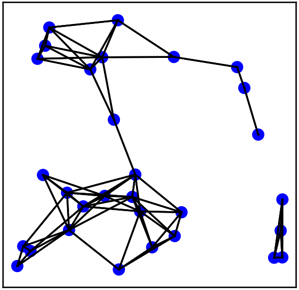
- Dataset $x_1, \ldots, x_n \in \mathcal{X}$
- Similarity measure $\sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$
- Graph of n nodes with weight $\sigma(x_i, x_j)$ between nodes i and j



Example: $\mathcal{X} = [0,1]^2$, $\sigma(x,y) = 1_{\{d(x,y)<1/4\}}$

Data as graph

- Dataset $x_1, \dots, x_n \in \mathcal{X}$
- Similarity measure $\sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$
- Graph of n nodes with weight $\sigma(x_i, x_j)$ between nodes i and j



Example: $\mathcal{X} = [0,1]^2$, $\sigma(x,y) = 1_{\{d(x,y)<1/4\}}$

Motivation

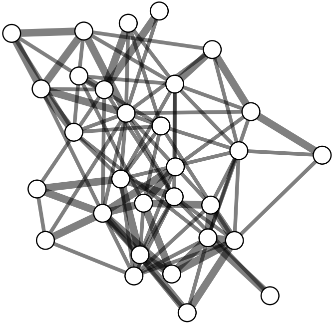
- Information retrieval
- Content recommandation
- Advertizing
- Anomaly detection
- Security

Graph analysis

- What are the most important nodes? $\rightarrow$ Ranking
- Can we predict new links? $\rightarrow$ Local ranking
- What is the graph structure? $\rightarrow$ Clustering
- Can we predict labels? $\rightarrow$ Classification

Setting

A weighted, undirected, connected graph of n nodes
No self-loops
Weighted adjacency matrix A
Vector of node weights $d = A\mathbf{1}$



Outline

1. Random walk
2. Laplacian matrix
3. Spectral analysis
4. Graph embedding
5. Applications

Outline

1. Random walk → Statistical physics
2. Laplacian matrix → Heat equation
3. Spectral analysis → Mechanics
4. Graph embedding → Electricity
5. Applications

Outline

1. **Random walk** → Statistical physics
2. Laplacian matrix → Heat equation
3. Spectral analysis → Mechanics
4. Graph embedding → Electricity
5. Applications

Random walk

Consider a **random walk** in the graph G where the probability of moving from node i to node j is A_{ij}/d_i

Random walk

Consider a **random walk** in the graph G where the probability of moving from node i to node j is A_{ij}/d_i

The sequence of nodes $X_0, X_1, X_2, \ldots$ defines a **Markov chain** on $\{1, \ldots, n\}$ with transition matrix $P = D^{-1}A$

Random walk

Consider a **random walk** in the graph G where the probability of moving from node i to node j is A_{ij}/d_i

The sequence of nodes $X_0, X_1, X_2, \ldots$ defines a **Markov chain** on $\{1, \ldots, n\}$ with transition matrix $P = D^{-1}A$

- Dynamics:

$$\mathrm{P}(X_{t+1} = i) = \sum_j \mathrm{P}(X_t = j) P_{ji}$$

Random walk

Consider a **random walk** in the graph G where the probability of moving from node i to node j is A_{ij}/d_i
The sequence of nodes $X_0, X_1, X_2, \ldots$ defines a **Markov chain** on $\{1, \ldots, n\}$ with transition matrix $P = D^{-1}A$

- Dynamics:

$$\mathrm{P}(X_{t+1} = i) = \sum_j \mathrm{P}(X_t = j) P_{ji}$$

- Stationary distribution π:

$$\mathrm{P}(X_\infty = i) = \sum_j \mathrm{P}(X_\infty = j) P_{ji} \quad \iff \quad \pi_i = \sum_j \pi_j P_{ji}$$

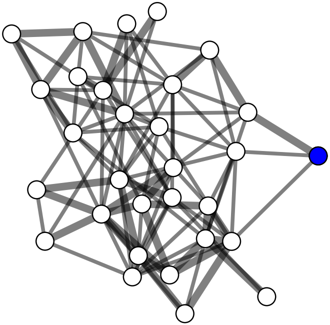
(global balance)

Return time

Since π_i is the frequency of visits of node i in stationary regime, the **mean return time** to node i is given by

$$\sigma_i = \mathrm{E}_i(\tau_i^+) = \frac{1}{\pi_i}$$

with $\tau_i^+ = \min\{t \geq 1 : X_t = i\}$



Reversibility

A Markov chain is called **reversible** if in stationary regime, the probability of any sequence of states is the same in both directions of time

Reversibility

A Markov chain is called **reversible** if in stationary regime, the probability of any sequence of states is the same in both directions of time

- Transition from state i to state j:

$$\begin{aligned} &\mathrm{P}(X_t = i, X_{t+1} = j) = \mathrm{P}(X_t = j, X_{t+1} = i) \\ \iff \quad &\pi_i P_{ij} = \pi_j P_{ji} \quad \text{(local balance)} \end{aligned}$$

Reversibility

A Markov chain is called **reversible** if in stationary regime, the probability of any sequence of states is the same in both directions of time

- Transition from state i to state j:

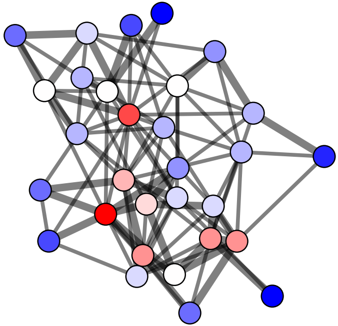
$$\mathrm{P}(X_t = i, X_{t+1} = j) = \mathrm{P}(X_t = j, X_{t+1} = i)$$
$$\iff \quad \pi_i P_{ij} = \pi_j P_{ji} \quad \text{(local balance)}$$

- Sequence of states $i_0, i_1, \dots i_\ell$:

$$\mathrm{P}(X_t = i_0, \dots, X_{t+\ell} = i_\ell) = \mathrm{P}(X_t = i_\ell, \dots, X_{t+\ell} = i_0)$$
$$\iff \quad \pi_{i_0} P_{i_0 i_1} \dots P_{i_{\ell-1} i_\ell} = \pi_{i_\ell} P_{i_\ell i_{\ell-1}} \dots P_{i_1 i_0}$$

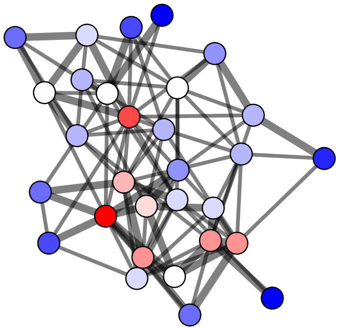
Reversibility & random walks

- The **random walk** in a graph is a reversible Markov chain, with stationary distribution $\pi \propto d$



Reversibility & random walks

- The **random walk** in a graph is a reversible Markov chain, with stationary distribution $\pi \propto d$



- Conversely, any **reversible** Markov chain is a random walk in a graph, with weights $\pi_i P_{ij} = \pi_j P_{ji}$

Reversibility in physics

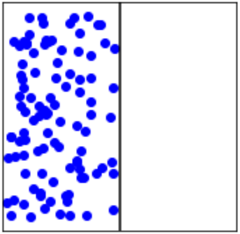
- All microscopic laws of physics are **reversible**

Reversibility in physics

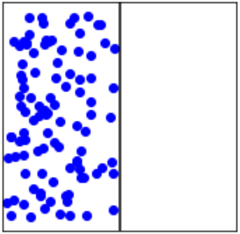
- All microscopic laws of physics are **reversible**
- The second law of thermodynamics states that the evolution of any isolated system is **irreversible**

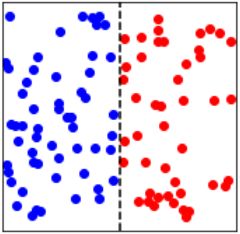
Reversibility in physics

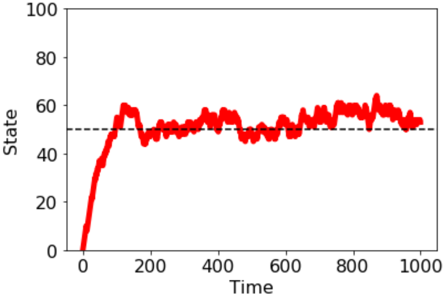
- All microscopic laws of physics are **reversible**
- The second law of thermodynamics states that the evolution of any isolated system is **irreversible**
- This apparent paradox was solved by **Tatiana & Paul Ehrenfest** in 1907



Example







Hitting time, commute time & escape probability

- Mean **hitting time** of node j from node i:

$$H_{ij} = \mathrm{E}_i(\tau_j), \quad \tau_j = \min\{t \geq 0 : X_t = j\}$$

- Mean **commute time** between nodes i and j:

$$\rho_{ij} = H_{ij} + H_{ji}$$

- **Escape probability** from node i to node j:

$$e_{ij} = \mathrm{P}_i(\tau_j < \tau_i^+)$$

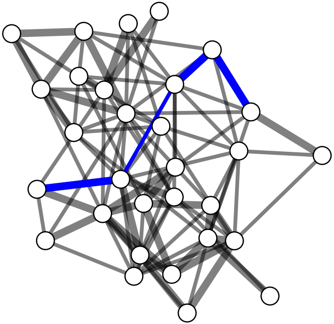
Proposition

$$\rho_{ij} = \frac{1}{\pi_i e_{ij}}$$

Proof

Frequency of no-return paths

$$\forall i \neq j \quad \pi_i e_{ij} = \pi_j e_{ji}$$



Outline

1. Random walk → Statistical physics
2. **Laplacian matrix** → Heat equation
3. Spectral analysis → Mechanics
4. Graph embedding → Electricity
5. Applications

Laplacian matrix

Let $D = \mathrm{diag}(A1)$.

Definition

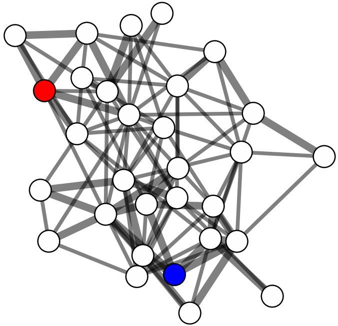
The matrix $L = D - A$ is called the **Laplacian matrix**.

Heat equation

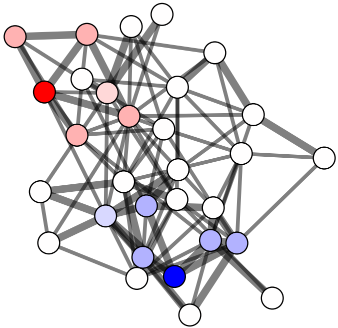
- Fix the temperature of some nodes $S \subset \{1, \ldots, n\}$
- Interpret the weight A_{ij} as the **thermal conductivity**
- Then for any node $i \notin S$,

$$\frac{dT}{dt} = \sum_j A_{ij}(T_j - T_i) = -(LT)_i$$

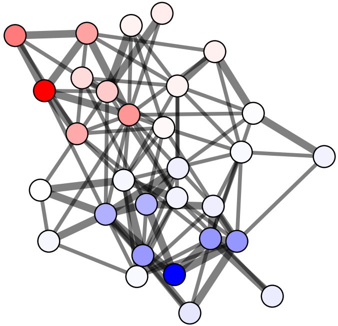
Example



Example



Example



Equilibrium

Dirichlet problem

- For any node $i \notin S$,
$$(LT)_i = 0$$
with boundary condition T_i for all $i \in S$
- The vector T is said to be **harmonic**

Uniqueness

There is **at most one** solution to the Dirichlet problem

Proof based on the **maximum principle**

The maximum principle

Back to random walks

- Consider the probability that the random walk first hits S in j when starting from i:

$$P_{ij}^S = \mathrm{P}_i(\tau_j = \tau_S)$$

  with $\tau_S = \min\{t \geq 0 : X_t \in S\}$
- This defines a **stochastic matrix** P^S

Back to random walks

- Consider the probability that the random walk first hits S in j when starting from i:

$$P_{ij}^S = \mathrm{P}_i(\tau_j = \tau_S)$$

with $\tau_S = \min\{t \geq 0 : X_t \in S\}$

- This defines a **stochastic matrix** P^S

Existence

The solution to the Dirichlet problem is

$$\forall i \notin S, \quad T_i = \sum_{j \in S} P_{ij}^S T_j$$

Solution to the Dirichlet problem

Outline

1. Random walk → Statistical physics
2. Laplacian matrix → Heat equation
3. **Spectral analysis** → Mechanics
4. Graph embedding → Electricity
5. Applications

Spectral analysis

The Laplacian matrix L is **symmetric** and **positive semi-definite**

Proposition

$$\forall v \in \mathbb{R}^n, \quad v^T L v = \sum_{i<j} A_{ij}(v_i - v_j)^2$$

Spectral analysis

The Laplacian matrix L is **symmetric** and **positive semi-definite**

Proposition

$$\forall v \in \mathbb{R}^n, \quad v^T L v = \sum_{i<j} A_{ij}(v_i - v_j)^2$$

Spectral decomposition

$$L = V \Lambda V^T$$

- $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of **eigenvalues**, with $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$
- $V = (v_1, \ldots, v_n)$ is a unitary matrix of **eigenvectors**, with $v_1 = 1/\sqrt{n}$

Mechanics

Consider a mechanical system of n particles of unit mass located on a **line** and linked by **springs** with stiffness A_{ij} (Hooke's law)

Mechanics

Consider a mechanical system of n particles of unit mass located on a **line** and linked by **springs** with stiffness A_{ij} (Hooke's law)

Denoting by $v \in \mathbb{R}^n$ the location of these particles, the **force** between i and j is:

$$A_{ij}|v_i - v_j|$$

Mechanics

Consider a mechanical system of n particles of unit mass located on a **line** and linked by **springs** with stiffness A_{ij} (Hooke's law)

Denoting by $v \in \mathbb{R}^n$ the location of these particles, the **force** between i and j is:

$$A_{ij}|v_i - v_j|$$

We deduce the **potential energy** of the system:

$$\frac{1}{2}\sum_{i<j} A_{ij}(v_i - v_j)^2 = \frac{1}{2}v^T L v$$

Energy minima

The minimum of $v^T L v$ under the constraint $v^T v = 1$ is:

- 0 (take $v = v_1$)
- λ_2 under the constraint $1^T v = 0$ (take $v = v_2$)

Theorem

For all $k = 1, \ldots, n$,

$$\lambda_k = \min_{\substack{v : v^T v = 1 \\ v_1^T v = 0, \ldots, v_{k-1}^T v = 0}} v^T L v$$

and the minimum is attained for $v = v_k$.

Proof

Physical interpretation

Assume each particle has unit mass and let the mechanical system rotate with **angular velocity** $\omega > 0$

Physical interpretation

Assume each particle has unit mass and let the mechanical system rotate with **angular velocity** $\omega > 0$
By Newton's law,

$$\forall i, \quad \sum_j A_{ij}(v_j - v_i) = -v_i\omega^2$$

$$\iff \quad Lv = \omega^2 v$$

Physical interpretation

Assume each particle has unit mass and let the mechanical system rotate with **angular velocity** $\omega > 0$
By Newton's law,

$$\forall i, \quad \sum_j A_{ij}(v_j - v_i) = -v_i\omega^2$$

$$\iff \quad Lv = \omega^2 v$$

Observations

- The only possible values of angular velocity are $\sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}$
- The corresponding equilibra are proportional to $v_2, \ldots, v_n$

Physical interpretation (energy)

At equilibrium, the **potential energy** is equal to the (rotational) **kinetic energy**:

$$\frac{1}{2} v^T L v = \frac{1}{2} v^T v \omega^2$$

where $v^T v$ is the **moment of inertia** of the system.

Physical interpretation (energy)

At equilibrium, the **potential energy** is equal to the (rotational) **kinetic energy**:

$$\frac{1}{2}v^T L v = \frac{1}{2}v^T v \omega^2$$

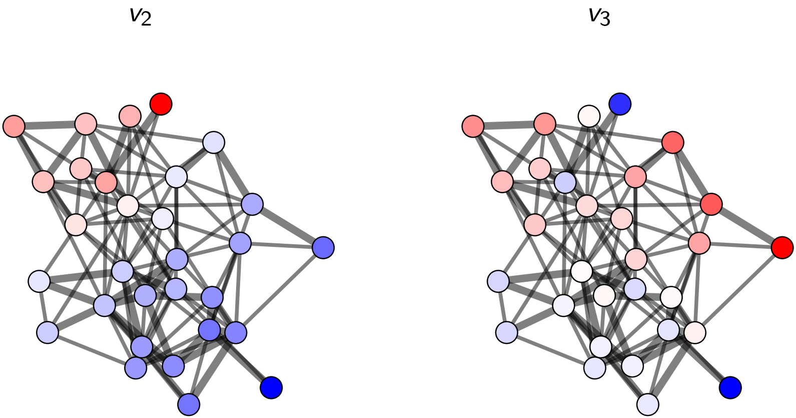
where $v^T v$ is the **moment of inertia** of the system.

Observations

For unit moments of inertia,

- The only possible values of energy are (half) $\lambda_2, \ldots, \lambda_n$
- The corresponding equilibra are $v_2, \ldots, v_n$

Example



Back to random walks

- The **normalized symmetric** Laplacian is defined by:

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}$$

- This matrix is **symmetric** and **positive semi-definite**
- By the spectral theorem,

$$\mathcal{L} = \mathcal{V} \Gamma \mathcal{V}^T$$

where $\Gamma = (\gamma_1, \ldots, \gamma_n)$, with $\gamma_1 = 0 < \gamma_2 \leq \ldots \leq \gamma_n$

Observation

The transition matrix P has eigenvalues $1 > 1 - \gamma_2 \geq \ldots \geq \gamma_n$, with corresponding matrix of eigenvectors $D^{-1/2}\mathcal{V}$

Outline

1. Random walk → Statistical physics
2. Laplacian matrix → Heat equation
3. Spectral analysis → Mechanics
4. **Graph embedding** → Electricity
5. Applications

Pseudo-inverse

Recall that

$$L = V\Lambda V^T$$

The **pseudo-inverse** of L is

$$L^+ = V\Lambda^+ V^T$$

with

$$\Lambda^+ = \mathrm{diag}\left(0, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n}\right)$$

Proposition

$$LL^+ = L^+L = I - \frac{11^T}{n}$$

Proof

First graph embedding

Consider the embedding $Z = (z_1, \ldots, z_n)$ of the nodes in $\mathbb{R}^n$, with

$$Z = \sqrt{\Lambda^+} V^T$$

First graph embedding

Consider the embedding $Z = (z_1, \ldots, z_n)$ of the nodes in $\mathbb{R}^n$, with

$$Z = \sqrt{\Lambda^+} V^T$$

Observations

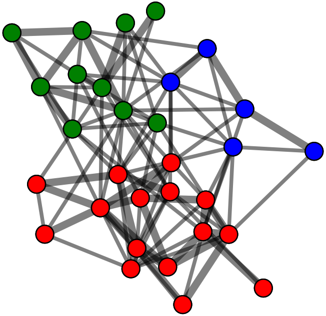
- The first coordinate is 0
- The k-th coordinate is $v_k / \sqrt{\lambda_k}$, with energy

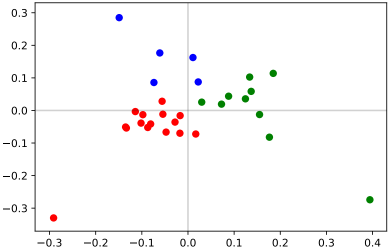
$$\frac{1}{2} \frac{v_k^T L v_k}{\lambda_k} = \frac{1}{2}$$

- Null component-wise averages, $Z1 = 0$
- The **Gram matrix** of Z is the pseudo-inverse of L

$$Z^T Z = V \Lambda^+ V^T = L^+$$

Example in $\mathbb{R}^2$





Second graph embedding

Consider the embedding $X = (x_1, \dots, x_n)$ of the nodes in $\mathbb{R}^n$, with

$$X = \sqrt{|d|} Z(I - \pi 1^T)$$

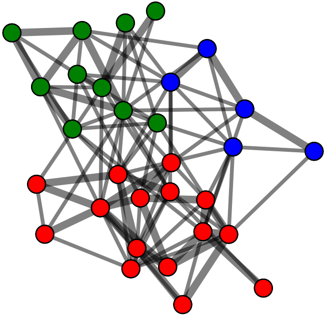
Observations

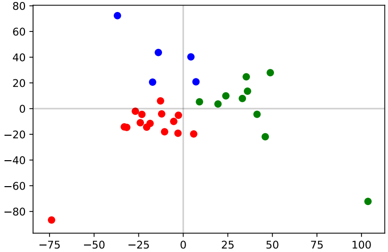
- **Shifted, normalized** version of Z
- Null component-wise **weighted** averages, $X\pi = 0$
- **Gram matrix** of X:

$$G = X^T X = |d|(I - 1\pi^T)L^+(I - \pi 1^T)$$

$$G\pi = 0$$

Example in $\mathbb{R}^2$





Back to random walks

- The mean **hitting time** of node j from node i satisfies:

$$H_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 + \sum_{k=1}^{n} P_{ik} H_{kj} & \text{otherwise} \end{cases}$$

Back to random walks

- The mean **hitting time** of node j from node i satisfies:

$$H_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 + \sum_{k=1}^{n} P_{ik} H_{kj} & \text{otherwise} \end{cases}$$

- We deduce that the matrix $(I - P)H - 11^T$ is diagonal
- Equivalently, the matrix $LH - d1^T$ is diagonal

Back to random walks

- The mean **hitting time** of node j from node i satisfies:

$$H_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 + \sum_{k=1}^{n} P_{ik} H_{kj} & \text{otherwise} \end{cases}$$

- We deduce that the matrix $(I - P)H - 11^T$ is diagonal
- Equivalently, the matrix $LH - d1^T$ is diagonal

Theorem

$$H = 11^T d(G) - G$$

where $G = X^T X$ is the Gram matrix of X

Back to random walks

- The mean **hitting time** of node j from node i satisfies:

$$H_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 + \sum_{k=1}^{n} P_{ik} H_{kj} & \text{otherwise} \end{cases}$$

- We deduce that the matrix $(I - P)H - 11^T$ is diagonal
- Equivalently, the matrix $LH - d1^T$ is diagonal

Theorem

$$H = 11^T d(G) - G$$

where $G = X^T X$ is the Gram matrix of X

Observation

$$H = 1h^T - G \quad \text{with } h^T = \pi^T H$$

Graph embedding and random walk

- Square distance to the origin:

$$||x_i||^2 = h_i \quad \text{(hitting time)}$$

- Scalar product:

$$x_j^T(x_j - x_i) = H_{ij} \quad \text{(hitting time)}$$

- Square distance between nodes i and j:

$$||x_i - x_j||^2 = \rho_{ij} \quad \text{(commute time)}$$

Proof of the Theorem

Lemma

There is at most one matrix H such that $LH - d1^T$ is diagonal and $d(H) = 0$

Proof of the Theorem

Theorem

$$H = 11^T d(G) - G$$

Mean return times

- The mean return time to node i satisfies

$$\sigma_i = 1 + \sum_j P_{ij} H_{ji}$$

- Thus the diagonal of $PH + 11^T$ gives the mean return times

Corollary

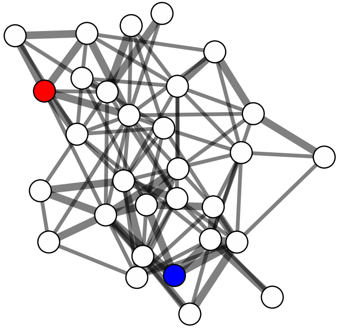
$$d(PH + 11^T) = \mathrm{diag}(\pi)^{-1}$$

Electricity

- Consider the electric network induced by the graph, with a resistor of **conductance** A_{ij} between nodes i and j

Electricity

- Consider the electric network induced by the graph, with a resistor of **conductance** A_{ij} between nodes i and j
- We look for the vector U of **electric potentials** given $U_s = 1$ (source) and $U_t = 0$ (sink)



A Dirichlet problem

- By **Ohm's law**, the current that flows from i to j is

$$A_{ij}(U_i - U_j)$$

- By **Kirchoff's law**, the net current at any node $i \neq s, t$ is null:

$$\sum_j A_{ij}(U_i - U_j) = 0$$

  that is $(LU)_i = 0$

- The vector U is the solution to the **Dirichlet problem** with boundary conditions $U_s = 1$ and $U_t = 0$

Energy dissipation

- Energy dissipation = differential of potential × current
- Total energy dissipation

$$\sum_{i<j} A_{ij}(U_j - U_i)^2$$

Thompson's principle

The potential vector U **minimizes** energy dissipation

Taking the derivative in U_i

$$\sum_{j} A_{ij}(U_j - U_i) = 0$$

that is $(LU)_i = 0$, which is the Dirichlet problem

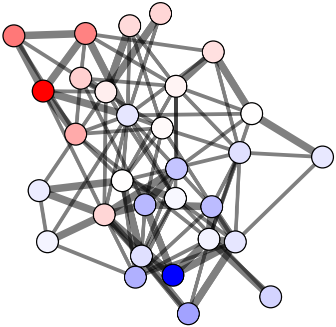
Solution to the Dirichlet problem

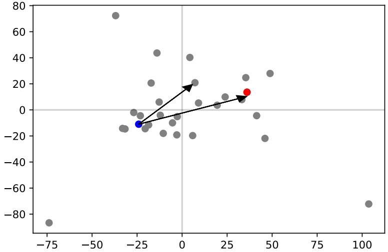
Proposition

The electric potential of node i is

$$U_i = \frac{(x_i - x_t)^T (x_s - x_t)}{||x_s - x_t||^2}$$

Example





Effective conductance, effective resistance

- The **current** that goes from node s to node t is

$$\frac{|d|}{||x_s - x_t||^2} = \frac{|d|}{\rho_{st}}$$

- This is the **effective conductance** between s and t
- The **effective resistance** between s and t is proportional to ρ_{st}, the mean commute time between nodes s and t

Electricity and random walks

The vector U of electric potential is the solution to the **Dirichlet problem** with $U_s = 1$ and $U_t = 0$

Interpretation of voltage

The voltage of any node is the **probability** that the random walk starting from this node reaches node s before node t

Electricity and random walks

The vector U of electric potential is the solution to the **Dirichlet problem** with $U_s = 1$ and $U_t = 0$

Interpretation of voltage

The voltage of any node is the **probability** that the random walk starting from this node reaches node s before node t

Interpretation of current

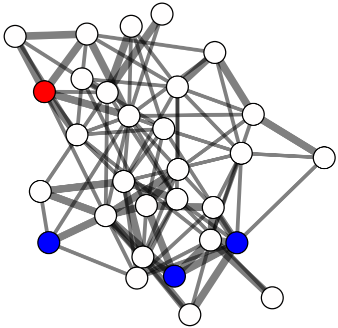
The net current from node i to node j is the **net frequency** of particles moving from node i to node j, with a flow of particles entering the network at node s at rate

$$\frac{|d|}{\rho_{st}}$$

The current as the net flow of particles

Extension

- A single source s, at electric potential 1
- Multiple sinks $t_1, \ldots, t_K$, at electric potential 0



Solution to the Dirichlet problem

Proposition

The electric potential of node i is:

$$U_i = \sum_{k=1}^{K} \alpha_k (x_i - x_{t_l})^T (x_s - x_{t_k})$$

where

- l is an arbitrary element of $\{1, \ldots, K\}$
- α is the unique solution to the equation $M\alpha = |d|1$, with M the Gram matrix of the vectors $(x_s - x_{t_1}, \ldots, x_s - x_{t_K})$

General solution to the Dirichlet problem

- For each $s \in S$, apply previous result to get $P_{is}^S \equiv U_i$
- The potential of each node $i \notin S$ is $U_i = \sum_{j \in S} P_{ij}^S U_j$

Outline

1. Random walk $\rightarrow$ Statistical physics
2. Laplacian matrix $\rightarrow$ Heat equation
3. Spectral analysis $\rightarrow$ Mechanics
4. Graph embedding $\rightarrow$ Electricity
5. **Applications**

Graph embedding

Method

1. Check that the graph is connected
2. Form the Laplacian $L = D - A$
3. Compute $v_1, \ldots, v_k$, the k eigenvectors of L associated with the lowest eigenvalues, $\lambda_1 \leq \ldots \leq \lambda_k$
4. Compute $Z = \mathrm{diag}\left(\frac{1}{\sqrt{\lambda_2}}, \ldots, \frac{1}{\sqrt{\lambda_k}}\right)(v_2, \ldots, v_k)^T$
5. Return $X = \sqrt{|d|}Z(I - \pi 1^T)$ where $\pi = d/|d|$

Observation

The dimension of the embedding must be chosen so that λ_k is large compared to λ_2

Ranking

Centrality

- **Output**: nodes in increasing order of $||x_i||^2$

Local centrality

- **Input**: node s of interest
- **Ouput**: nodes in increasing order of $x_i^T(x_i - x_s)$

Local centrality (multiple nodes)

- **Input**: nodes $s_1, \dots, s_K$ of interest (with weights)
- **Ouput**: nodes in increasing order of $x_i^T(x_i - x)$
  with x the weighted sum of $x_{s_1}, \dots, x_{s_K}$

Ranking with repulsive nodes

Directional centrality

- **Input**: node s of interest, repulsive node t
- **Ouput**: nodes in increasing order of $x_i^T(x_s - x_t)$

Directional centrality (multiple repulsive nodes)

- **Input**: node s of interest, repulsive nodes $t_1, \ldots, t_K$
- **Ouput**: nodes in increasing order of $x_i^T x$ with

$$x = \sum_{k=1}^{K} \alpha_k (x_s - x_{t_k})$$

where α is the solution to $M\alpha = 1$, with M the Gram matrix of $(x_s - x_{t_1}, \ldots, x_s - x_{t_K})$

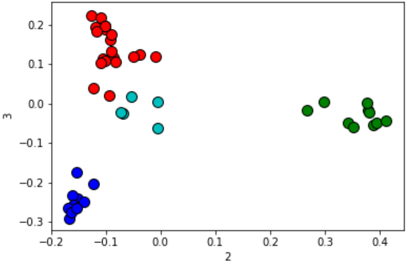
Clustering

Partition $C_1, \ldots, C_K$ of the nodes

- **Objective:** Minimizing

$$J = \sum_k \sum_{i \in C_k} ||x_i - \mu_k||^2 \quad \text{with } \mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

- A **combinatorial** problem (NP-hard)



The K-means algorithm

Algorithm

Input: K, number of clusters

Init $\mu_1, \ldots, \mu_K$ arbitrarily
Repeat until convergence:

- for each k, $C_k \leftarrow$ closest points of μ_k
- for each k, $\mu_k \leftarrow$ centroid of C_k

Output: Clusters $C_1, \ldots, C_K$

- Convergence in finite time
- Local optimum, that depends on the initial values of $\mu_1, \ldots, \mu_K$

Back to random walks

Observing that

$$J = \sum_k \frac{1}{2|C_k|} \sum_{i,j \in C_k} ||x_i - x_j||^2$$

the cost function J is, up to a factor $n/2$:

- the **mean square distance** of a random point to another random point of the same cluster
- the **mean commute time** of the random walk between a random node and another node taken uniformly at random in the same cluster

Modularity

- Given some clustering C, let

$$Q = \sum_{i,j} \pi_i (P_{ij} - \pi_j) \delta^C_{i,j}$$

where

$$\delta^C_{i,j} = \begin{cases} 1 & \text{if } i,j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

Modularity

- Given some clustering C, let

$$Q = \sum_{i,j} \pi_i (P_{ij} - \pi_j) \delta_{i,j}^C$$

where

$$\delta_{i,j}^C = \begin{cases} 1 & \text{if } i, j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

- Then Q is the difference between the probabilities that
(1) two **successive nodes** of the random walk are in the same cluster
(2) two **independent** random walks are in the same cluster
- Maximizing Q is NP-hard

The Louvain algorithm

Algorithm

Init each node in its own cluster
Repeat until convergence:

- while Q increases, change the cluster of any node to one of its neighbors
- aggregate all nodes belonging to the same cluster in a single node

Output: Clusters

- Convergence in finite time
- Local optimum, that depends on the order in which nodes are considered

Summary

- Random walks in graphs provide efficient techniques for **ranking** and **clustering** nodes
- In the **lab session**, you will learn to apply these techniques to real graphs using the Python `networkx` package