Graph Mining
SD212
6. Diffusion in Graphs

Thomas Bonald
Institut Polytechnique de Paris

2019 – 2020

These lecture notes introduce techniques for ranking or classifying the nodes of the graph based on heat diffusion. The interest compared to other techniques like PageRank is to enable *contrast* (e.g., to identify nodes that are both close to some nodes and far from some other nodes).

1 Heat diffusion

Consider a graph $G = (V, E)$ of n nodes and m edges. We assume that the graph is undirected, with adjacency matrix A. Let $d = A1$ be the vector of degrees, which we assume positive, and $D = \text{diag}(d)$.

The idea of heat diffusion is to view the graph as a thermodynamic system. Let $T(t)$ be the vector of temperatures of the n nodes at time t. Heat exchanges occur through each edge of the graph proportionally to the temperature difference between the corresponding nodes, so that:

$$\forall i \in V, \quad \frac{\mathrm{d}T_i}{\mathrm{d}t} = \sum_{j \in V} A_{ij}(T_j - T_i). \tag{1}$$

This equation can be written in vector form:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = -LT,$$

where $L = D - A$ is called the *Laplacian* matrix of the graph and plays the role of the Laplace operator in the usual heat equation.

Conservation. Let $\bar{T} = \frac{1}{n}\sum_{i \in V} T_i$ be the average temperature of the nodes. It follows from the heat equation that

$$\frac{\mathrm{d}\bar{T}}{\mathrm{d}t} = \frac{1^T}{n}\frac{\mathrm{d}T}{\mathrm{d}t} = -\frac{1^T}{n}LT = 0.$$

The average temperature is preserved over time.

Equilibrium. In steady state, the vector of temperatures satisfies Laplace's equation:

$$LT = 0. \tag{2}$$

The vector T is said to be *harmonic*. Laplace's equation can be written equivalently $DT = AT$, namely

$$\forall i \in V, \quad T_i = \frac{1}{d_i}\sum_{j \in V} A_{ij}T_j.$$

This shows that the temperature of each node at equilibrium is the average temperature of its neighbors.

If the graph is connected, the solution is a constant vector: all nodes have the same temperature at equilibrium. By the conservation property, the temperature of each node at equilibrium is the average temperature in the initial state. If the graph is not connected, the solution is a constant vector per connected component, with temperature in each connected component equal to the average temperature in this connected component in the initial state.

Convergence. The solution to the differential equation (1) is given by:

$$T(t) = e^{-Lt}T(0).$$

The matrix $H(t) = e^{-Lt}$ is known to as the *heat kernel*. It gives the solution at any time t starting from any initial state $T(0)$ as $T(t) = H(t)T(0)$. It can be expressed through the spectral decomposition[1] of the Laplacian matrix L, $L = U\Lambda U^T$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ the diagonal matrix of eigenvalues $\lambda_1 = 0 \leq \lambda_2 \leq \ldots \leq \lambda_n$ and $U = (u_1, \ldots, u_n)$ an orthogonal matrix of corresponding eigenvectors:

$$H(t) = \sum_{k=1}^{n} e^{-\lambda_k t} u_k u_k^T.$$

Observe that the first eigenvector is a solution to Laplace's equation (2). If the graph is connected, then $u_1 = 1/\sqrt{n}$ and $\lambda_2 > 0$ (because there is a unique solution to the Laplace's equation, up to some multiplicative constant), so that

$$H(t) \to u_1 u_1^T \quad \text{when } t \to +\infty,$$

and the vector of temperatures converges to the average temperature in the initial state:

$$T(t) \to \bar{T}(0)1 \quad \text{when } t \to +\infty.$$

The convergence is exponential at rate λ_2, the spectral gap of the Laplacian matrix.

2 Heat diffusion in discrete time

Let $P = D^{-1}A$ be the transition matrix of the random walk in the graph. Consider the heat diffusion in discrete time defined by:

$$T(s+1) = PT(s). \tag{3}$$

The temperature of each node at time step $s+1$ is the average temperature of its neighbors at time step s. To compare with the heat diffusion in continuous time, it is worth writing this equation as follows:

$$T(s+1) - T(s) = -(I - P)T(s). \tag{4}$$

Observe that $I - P = D^{-1}L$, so that the heat diffusion in discrete time (3) is similar to that in continuous time (1). The major difference lies in the inertia of the nodes: nodes of high degree are less sensitive to temperature differences than nodes of low degree (due to the term D^{-1}).

Conservation. Let $\hat{T} = \frac{1}{2m}\sum_{i \in V} d_i T_i$ be the *weighted* average temperature of the nodes. It follows from (3) that

$$\hat{T}(s+1) = \frac{d^T}{2m}T(s+1) = \frac{d^T}{2m}PT(s) = \frac{d^T}{2m}T(s) = \hat{T}(s).$$

The weighted average temperature is preserved over time.

[1] The Laplacian matrix is symmetric and positive semi-definite.

Equilibrium. In steady state, the vector of temperatures satisfies:

$$T = PT,$$

which is equivalent to Laplace's equation (2).

If the graph is connected, the solution is a constant vector. By the conservation property, the temperature of each node at equilibrium is the *weighted* average temperature in the initial state. If the graph is not connected, the solution is a constant vector per connected component, with temperature in each connected component equal to the *weighted* average temperature in this connected component in the initial state.

Convergence. The solution to the differential equation (3) is given by:

$$T(s) = P^s T(0).$$

The matrix P is not symmetric, unlike the normalized adjacency matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. Consider the spectral decomposition[2] of this matrix, $D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U\Lambda U^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ the diagonal matrix of eigenvalues, with $\lambda_1 = 1 \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$, and $U = (u_1, \ldots, u_n)$ an orthogonal matrix of corresponding eigenvectors. Let $V = D^{-\frac{1}{2}}U$. Then:

$$P = V\Lambda V^T D \quad \text{with} \quad V^T D V = I. \tag{5}$$

We have:

$$PV = V\Lambda,$$

showing that $V = (v_1, \ldots, v_n)$ is a matrix of right eigenvectors of P, for the respective eigenvalues $\lambda_1, \ldots, \lambda_n$. Observe that the first eigenvector is a solution to Laplace's equation.

In view of (5),

$$P^s = V\Lambda^s V^T D,$$

that is

$$P^s = \sum_{k=1}^{n} \lambda_k^s v_k v_k^T D.$$

If the graph is connected and not bipartite, then $v_1 \propto 1$ and $|\lambda_2| < 1$ (because there is a unique solution to the Laplace's equation, up to some multiplicative constant), so that

$$P^s \to v_1 v_1^T D \quad \text{when } s \to +\infty,$$

and the vector of temperatures converges to the weighted average temperature in the initial state:

$$T(s) \to \hat{T}(0)1 \quad \text{when } s \to +\infty.$$

The convergence is geometric at rate $|\lambda_2|$, the spectral gap of the transition matrix.

3 Dirichlet problem

The Dirichlet problem consists in solving Laplace's equation in the presence of boundary conditions. Let S be some strict subset of V and assume that the temperature of each node $i \in S$ is set at some fixed value T_i. We are interested in the evolution of the temperatures of the other nodes. The Dirichlet problem consists in finding the vector of temperatures at equilibrium, that is

$$\forall i \notin S, \quad (LT)_i = 0, \tag{6}$$

with the boundary conditions T_i for all $i \in S$.

Equivalently,

$$\forall i \notin S, \quad T_i = (PT)_i. \tag{7}$$

In other words, the temperature of each node $i \notin S$ at equilibrium is the average of the temperature of its neighbors.

[2] We use the same notation as for the spectral decomposition of the Laplacian matrix, but of course they are different.

Uniqueness. The solution to the Dirichlet problem is unique, provided that the graph is connected. If the graph is not connected, the solution is unique provided there is at least one node of S in each connected component.

Proposition 1 *If the graph is connected, there is at most one solution to the Dirichlet problem.*

Proof. We first prove that the maximum and the minimum of the vector T are achieved on the boundary S. Let i be any node such that T_i is maximum. If $i \notin S$, it follows from (7) that T_j is maximum for all neighbors j of i. If no such node belongs to S, we apply again this argument until we reach a node in S. Such a node exists because the graph is connected. It achieves the maximum of the vector T. The proof is similar for the minimum.

Now consider two solutions T, T' to the Dirichlet problem (9). Then $\Delta = T' - T$ is a solution of the Dirichlet problem with the boundary condition $\Delta_i = 0$ for all $i \in S$. We deduce that $\Delta_i = 0$ for all i (because both the maximum and the minimum are equal to 0), that is $T' = T$. □

Random walk. Let P^S_{ij} be the probability that a random walk in the graph first hits the set S in node j when starting from node i. Observe that P^S is a stochastic matrix, with $P^S_{ij} = \delta_{ij}$ (Kronecker delta) for all $i \in S$. By first-step analysis, we have:

$$\forall i \notin S, \quad P^S_{ij} = \sum_{k=1}^{n} P_{ik} P^S_{kj}. \tag{8}$$

The following result provides a simple interpretation of the solution to the Dirichlet problem in terms of random walk in the graph: the temperature of each node is the average of the temperatures of the nodes at the boundary, weighted by the probabilities of hitting each of these nodes first:

Proposition 2 *The solution to the Dirichlet problem is*

$$\forall i \notin S, \quad T_i = \sum_{j \in S} P^S_{ij} T_j. \tag{9}$$

Proof. The vector T defined by (9) satisfies for all $i \notin S$:

$$\sum_{j=1}^{n} P_{ij} T_j = \sum_{j=1}^{n} P_{ij} \sum_{k \in S} P^S_{jk} T_k = \sum_{k \in S} P^S_{ik} T_k = T_i,$$

where we have used (8). Thus, T satisfies (7). The proof then follows from Proposition 1. □

Exact solution. We now characterize the solution to the Dirichlet problem. Without any loss of generality, we assume that nodes with unknown temperatures (i.e., not in S) are indexed from 1 to $n - s$ so that the vector of temperatures can be written

$$T = \begin{bmatrix} X \\ Y \end{bmatrix},$$

where X is the unknown vector of temperatures at equilibrium. Writing the transition matrix in block form as

$$P = \begin{bmatrix} Q & R \\ \cdot & \cdot \end{bmatrix},$$

it follows from (7) that:

$$X = QX + RY, \tag{10}$$

so that:

$$X = (I - Q)^{-1} RY. \tag{11}$$

Note that the inverse of the matrix $I - Q$ exists whenever the graph is connected, which implies that the matrix Q is sub-stochastic with spectral radius strictly less than 1 [1].

Approximate solution. The exact solution (11) requires to invert a (potentially large) matrix. In practice, a very good approximation is provided by a few iterations of (10), the rate of convergence depending on the spectral radius of the matrix Q. The small-world property of real graphs suggests that a relatively small value can be chosen for the number of iterations.

4 Ranking

We now show how to use heat diffusion to get some personalized ranking of the nodes. The technique is similar to Personalized PageRank, except that we must specify both "hot" nodes and "cold" nodes. Specifically, let S be some strict subset of V whose temperatures are fixed (the seeds). If all temperatures are equal, the Dirichlet problem is trivial. We need some *contrast* in the temperatures on the boundary. For instance, some seeds (the hot nodes) may have temperature 1 and some other seeds (the cold nodes) may have temperature 0: the solution to the Dirichlet problem will give a temperature between 0 and 1 for all nodes and can be used to rank these nodes, as illustrated by Figure 1.

Ranking by Dirichlet

Input:
P, transition matrix of the random walk
S, set of seeds
T^S, temperature of seeds
K, number of iterations

Do:
$T \leftarrow \text{mean}(T^S)1$, $T_S \leftarrow T^S$
For $t = 1, \ldots, K$,
  $T \leftarrow PT$, $T_S \leftarrow T^S$

Output:
T, vector of temperatures

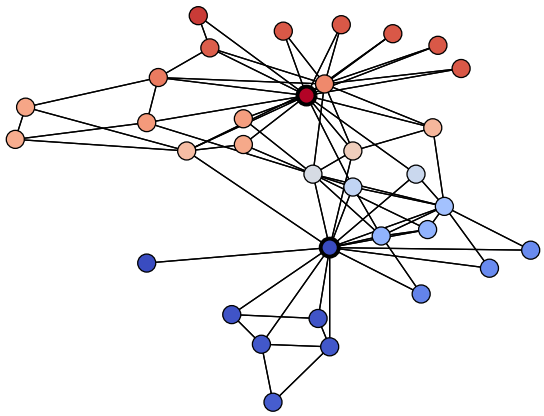


Figure 1: Ranking of nodes by the Dirichlet problem on the Karate Club graph (34 nodes, 2 seeds).

Another way to rank nodes is to use heat diffusion without constraints: the temperatures of the seeds are only set in the initial state, then evolve as those of the other nodes. The main differences with the technique based on the Dirichlet problem are the following:

- The temperatures at equilibrium are equal (at least if the graph is connected) and thus cannot be used to rank nodes. It is the transient state that is interesting, before convergence.
- The solution depends on the type of diffusion (continuous time or discrete time). The diffusion in discrete time is usually preferred as it is simpler to implement.
- The solution depends on the initial temperatures of non-seed nodes. A natural choice is to take the average temperature of seed nodes, possibly weighted by their degrees. Another option is to set positive temperatures to seed nodes (hot nodes) and null temperatures to non-seed nodes (cold nodes). The final ranking then gives nodes that are closer to seed nodes than to other nodes (see Figure 2).

Ranking by diffusion

Input:
P, transition matrix of the random walk
S, set of seeds
T^S, temperature of seeds
K, number of iterations

Do:
$T \leftarrow 0$, $T_S \leftarrow T^S$
For $t = 1, \ldots, K$,
    $T \leftarrow PT$

Output:
T, vector of temperatures

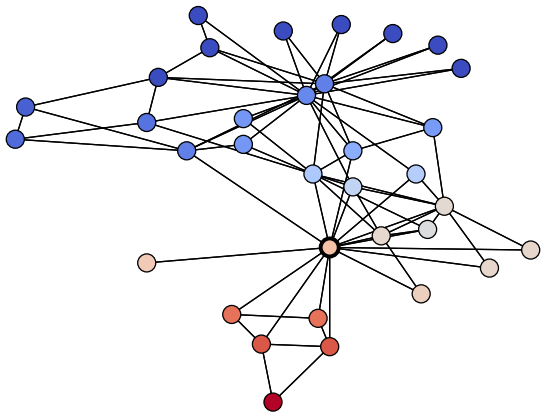
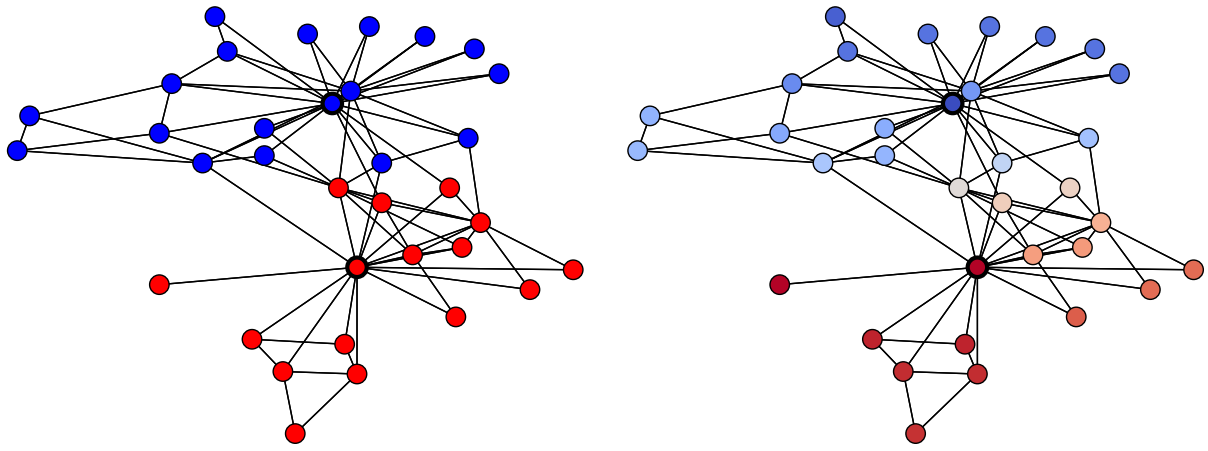


Figure 2: Ranking of nodes by heat diffusion on the Karate Club graph (34 nodes, 1 seed).

5 Classification

Heat diffusion can also be applied to node classification. The objective is to infer the labels of all nodes given the labels of a few nodes callsed the *seeds*. We here present the results based on the Dirichlet approach but the classification can rely on free diffusion as well. We denote by S the set of seeds.

Binary classification. When there are only two different labels, the classification can be done by solving one Dirichlet problem. The idea is to use the seeds with label 1 as hot sources, setting their temperature at 1, and the seeds with label 2 as cold sources, setting their temperature at 0. The solution to this Dirichlet problem gives temperatures between 0 and 1, as illustrated by Figure 3.



(a) Ground truth.

(b) Solution to the Dirichlet problem.

Figure 3: Binary classification of the Karate Club graph with 2 seeds. Red nodes have label 1, blue nodes have label 2.

A natural approach consists in assigning label 1 to all nodes with temperature above 0.5 and label 2 to other nodes. In practice, it is preferable to set the threshold to the mean temperature,

$$\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

Specifically, all nodes with temperature above $\bar{T}$ are assigned label 1, the other are assigned label 2. Equivalently, temperatures are centered before classification: after centering, nodes with positive temperature are assigned label 1, the other are assigned label 2.

Note that the temperature of each node can be used to assess the confidence in the classification: the closer the temperature to the mean, the lower the confidence. This is illustrated by Figure 3 (the lighter the color, the lower the confidence). In this case, only one node is misclassified and has indeed a temperature close to the mean.

Multi-class classification In the presence of more than 2 labels, a *one-against-all* strategy can be used: the seeds of each label alternately serve as hot sources while all the other seeds serve as cold sources. After centering the temperatures (so that the mean temperature of each diffusion is equal to 0), each node is assigned the label that maximizes its temperature. This algorithm is parameter-free.

Classification by Dirichlet

Input:
P, transition matrix of the random walk
S, set of seeds
y_S, labels of seeds (L labels)

Do:
For $\ell = 1, \dots, L$,
  For $i \in S$, $T_i \leftarrow 1$ if $y_i = \ell$, $T_i \leftarrow 0$ otherwise
  $T^{(\ell)} \leftarrow \text{Dirichlet}(P, S, T^S)$
  $\Delta^{(\ell)} \leftarrow T^{(\ell)} - \text{mean}(T^{(\ell)})$
For $i \notin S$
  $\hat{y}_i \leftarrow \arg\max_\ell(\Delta_i^{(\ell)})$

Output:
$\hat{y}$, vector of predicted labels

6 Extensions

Weighted graphs. The results readily apply to weigthed graphs, the weight of an edge corresponding to the strength of the connection. In the heat equation (1), the weight A_{ij} of edge i, j, if any, is interpreted as the thermal conductivity between nodes i and j. The Laplacian matrix $L = D - A$ and the transition matrix $P = D^{-1}A$ are defined in the same way, with $D = \text{diag}(A1)$ the diagonal matrix of node weights.

Directed graphs. The extension of the results to directed graphs requires some care. First, the graph may have sinks, from which the random walk is not defined. A natural approach consists in letting the random walk jump to any node chosen uniformly at random in V. The transition matrix becomes:

$$P_{ij} = \begin{cases} \frac{A_{ij}}{d_i^+} & \text{if } d_i^+ > 0, \\ \frac{1}{n} & \text{otherwise,} \end{cases}$$

with $d^+ = A1$ the vector of out-degrees (out-weights if the graph is weighted).

The heat diffusion in discrete time (3) is then well defined. Observe that heat exchanges between nodes are no longer symmetric (as opposed to actual physical systems). In particular, the conservation property is no longer satisfied, and the uniqueness of the solution to Laplace's equation is not guaranteed, unless the graph is strongly connected. To avoid these problems, a damping factor can be introduced as for PageRank. The transition matrix with damping factor $\alpha \in (0, 1)$ becomes:

$$P^{(\alpha)} = \alpha P + (1 - \alpha)\frac{11^T}{n},$$

with P defined above. In terms of heat diffusion (3), this means that the temperature of each node at time $t+1$ is equal to some weighted average of the average temperature of its successors at time t (with weight α) and the average temperature of all nodes at time t (with weight $1 - \alpha$). The solution to Laplace's equation (and to the Dirichlet problem) is then unique and can be used to rank or classify nodes.

References

[1] Fan R.K. Chung. *Spectral graph theory.* American Mathematical Soc., 1997.