Lecture 2: From Structured Data to Graphs and Spectral Analysis

Radu Balan

February 9, 2017

Data Sets

Today we discuss type of data sets and graphs. The overarching problem is the following:

Main Problem

Given a graph, discover if it can be explained as a structured data graph, or just as a random graph.

Data Sets

Today we discuss type of data sets and graphs. The overarching problem is the following:

Main Problem

Given a graph, discover if it can be explained as a structured data graph, or just as a random graph.

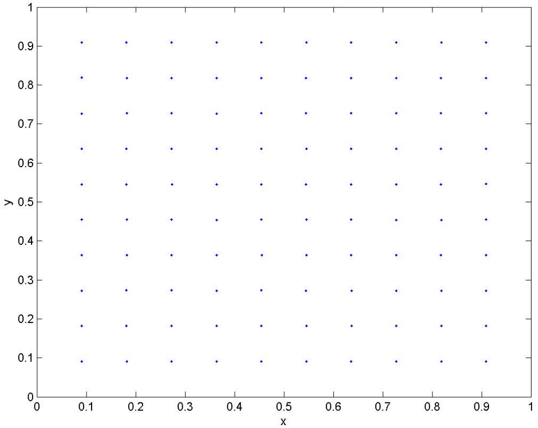
We shall discuss first how to construct a sequence of nested graphs from a data set.

Two types of data:

1. Percolation model
2. Weighted graphs

Data Sets
Percolation Models

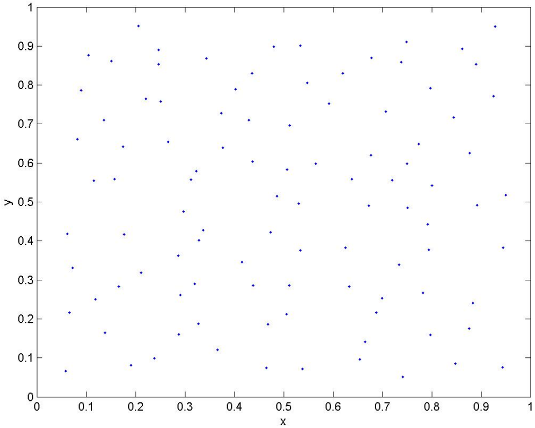
Fix a set of points in $\mathbb{R}^d$. Example, for $d = 2$:



$n = 10^2 = 100$
Uniform (regular) lattice.

Data Sets
Percolation Models

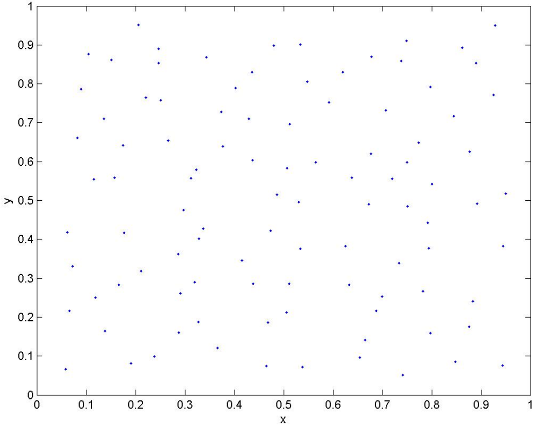
Fix a set of points in $\mathbb{R}^d$. Example, for $d = 2$:



$n = 10^2 = 100$
Nonuniform (irregular) lattice.

Data Sets
Percolation Models

Fix a set of points in $\mathbb{R}^d$. Example, for $d = 2$:



$n = 10^2 = 100$
Nonuniform (irregular) lattice.
Created by random perturbation of the regular lattice.

Data Sets
Percolation Models

Construct the matrix of pairwise distances:

$$V = \left(\|r^k - r^j\| \right)_{1 \leq k,j \leq n} \quad , \quad r^k = (x_k, y_k).$$

Data Sets
Percolation Models

Construct the matrix of pairwise distances:

$$V = \left(\| r^k - r^j \| \right)_{1 \leq k,j \leq n} \quad , \quad r^k = (x_k, y_k).$$

Then sort the set of distances in an ascending order. This way we define an order on the set of pairs of points. Implicitly this defines an ascending order on the set of edges. We obtain a sequence of nested graphs

$$(G_t)_{t \geq 0} \quad 0 \leq t \leq m = n(n-1)/2$$

where t indicates the number of edges in the graph G_t.
Thus G_t has n nodes and t edges.

Data Sets
Percolation Models

Play Examples: $n = 100$, regular/irregular, different types of norms:

$$\|r^k - r^j\|_2 = \sqrt{(x_k - x_j)^2 + (y_k - y_j)^2}$$

$$\|r^k - r^j\|_\infty = max(|x_k - x_j|, |y_k - y_j|)$$

Data Sets
Weighted Graphs

A different class of graphs: weighted graphs, $(\mathcal{V}, \mathcal{E}, W)$. Examples:

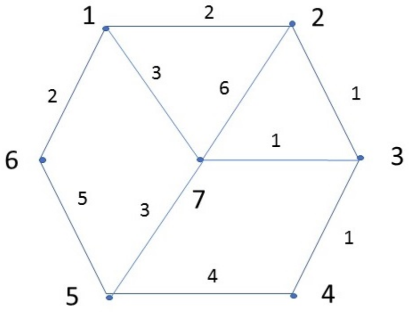
1. Joint co-authorship papers: $\mathcal{V}$ is the set of all authors; $\mathcal{E}$ is the list of joint papers; $w(e_{i,j})$ is the number of papers where both i and j are co-authors.
2. Protein-protein interaction or simultaneous expression.
3. Social networks: Facebook, LinkedIn: $\mathcal{V}$ is the set of users; $\mathcal{E}$ is the list of friendship links, or connections; $w(e_{i,j})$ is a measure of activity between i and j, e.g. number of endorsements, or 'like', or comments between i and j.
4. Communication networks ...
5. Email datasets (Enron)

Data Sets
Weighted Graphs

The sequence of nested graphs is obtained by sort the edges according to their weights: start with the largest weight first, and then pick the next largest weight, and so on.

Data Sets
Weighted Graphs

The sequence of nested graphs is obtained by sort the edges according to their weights: start with the largest weight first, and then pick the next largest weight, and so on.



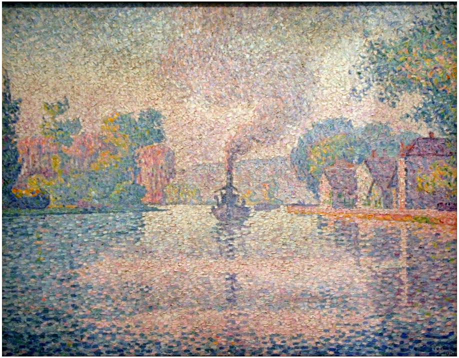
Data Sets

Data Size

Size matters:

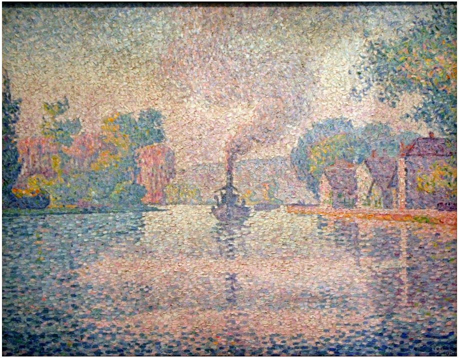
Data Sets
Data Size

Size matters:



Data Sets
Data Size

Size matters:



By Paul Signac - Ophelia2, Public Domain,
https://commons.wikimedia.org/w/index.php?curid=12570159
(L'Hirondelle Steamer on the Seine)

Data Sets
Public Datasets

On Canvas you can find links to several public databases:

1. Duke: https://dnac.ssri.duke.edu/datasets.php
2. Stanford: https://snap.stanford.edu/data/
3. Uni. Koblenz: http://konect.uni-koblenz.de/
4. M. Newman (U. Michigan): http://www-personal.umich.edu/ mejn/netdata/
5. A.L. Barabasi (U. Notre Dame): http://www3.nd.edu/ networks/resources.htm
6. UCI: https://networkdata.ics.uci.edu/resources.php

Spectral Analysis
Basic Properties

Last time we learned how to construct: the Adjacency matrix A, the Degree matrix D, the (unnormalized symmetric) graph Laplacian matrix $\Delta = D - A$, the normalized Laplacian matrix $\tilde{\Delta} = D^{-1/2}\Delta D^{-1/2}$, and the normalized asymmetric Laplacian matrix $L = D^{-1}\Delta$.

Spectral Analysis
Basic Properties

Last time we learned how to construct: the Adjacency matrix A, the Degree matrix D, the (unnormalized symmetric) graph Laplacian matrix $\Delta = D - A$, the normalized Laplacian matrix $\tilde{\Delta} = D^{-1/2}\Delta D^{-1/2}$, and the normalized asymmetric Laplacian matrix $L = D^{-1}\Delta$.

We denote: n the number of vertices (also known as the *size* of the graph), m the number of edges, $d(v)$ the degree of vertex v, $d(i,j)$ the distance between vertex i and vertex j (length of the shortest path connecting i to j), and by D the diameter of the graph (the largest distance between two vertices = "longest shortest path").

Spectral Analysis
Basic Properties

In this section we summarize spectral properties of the Laplacian matrices.

Theorem

1. $\Delta = \Delta^T \geq 0$, $\tilde{\Delta} = \tilde{\Delta}^T \geq 0$ *are positive semidefinite matrices.*
2. $eigs(\tilde{\Delta}) = eigs(L) \subset [0, 2]$.
3. 0 *is always an eigenvalue of* Δ, $\tilde{\Delta}$, L *with same multiplicity. Its multiplicity is equal to the number of connected components of the graph.*
4. $\lambda_{max}(\Delta) \leq 2 \max_v d(v)$, *i.e. the lagest eigenvalue of* Δ *is bounded by twice the largest degree of the graph.*

Spectral Analysis
Basic Properties

Theorem

Let $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1} \leq 2$ *be the eigenvalues of* $\tilde{\Delta}$ *(or* L*), that is* $eigs(\tilde{\Delta}) = \{\lambda_0, \lambda_1, \cdots, \lambda_{n-1}\} = eigs(L)$. *Then:*

1. $\sum_{i=0}^{n-1} \lambda_i \leq n$.
2. $\sum_{i=0}^{n-1} \lambda_i = n$ *if and only if the graph is connected (i.e. no isolated vertices).*
3. $\lambda_1 \leq \frac{n}{n-1}$.
4. $\lambda_1 = \frac{n}{n-1}$ *if and only if the graph is complete (i.e. any two vertices are connected by an edge).*
5. *If the graph is not complete then* $\lambda_1 \leq 1$.
6. *If the graph is connected then* $\lambda_1 > 0$. *If* $\lambda_i = 0$ *and* $\lambda_{i+1} \neq 0$ *then the graph has exactly* $i+1$ *connected components.*
7. *If the graph is connected (no isolated vertices) then* $\lambda_{n-1} \geq \frac{n}{n-1}$.

Spectral Analysis

Smallest nonnegative eigenvalue

Theorem

Assume the graph is connected. Thus $\lambda_1 > 0$. Denote by D its diameter and by $d_{max}, \bar{d}, d_H$ the maximum, average, and harmonic avergae of the degrees $(d_1, \cdots, d_n)$:

$$d_{max} = \max_j d_j \quad , \quad \bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j \quad , \quad \frac{1}{d_H} = \frac{1}{n}\sum_{j=1}^{n} \frac{d_j}{.}$$

Then

1. $\lambda_1 \geq \frac{1}{nD}$.
2. *Let* $\mu = \max_{1 \leq j \leq n-1} |1 - \lambda_j|$. *Then*

$$1 + (n-1)\mu^2 \geq \frac{n}{d_H}(1 - (1+\mu)(\frac{\bar{d}}{d_H} - 1)).$$

Spectral Analysis

Smallest nonnegative eigenvalue

Theorem

[continued]

3. *Assume $D \geq 4$. Then*

$$\lambda_1 \leq 1 - 2\frac{\sqrt{d_{max} - 1}}{d_{max}}(1 - \frac{2}{D}) + \frac{2}{D}.$$

Spectral Analysis

Comments on the proof

"Ingredients" and key relations:

1. Let $f = (f_1, f_2, \cdots, f_n) \in \mathbb{R}^n$ be a n-vector. Then:

$$\langle \Delta f, f \rangle = \sum_{x \sim y} (f_x - f_y)^2$$

where $x \sim y$ if there is an edge between vertex x and vertex y (i.e. $A_{x,y} = 1$).

This proves positivity of all operators.

2. Last time we showed $eigs(\tilde{\Delta}) = eigs(L)$ because $\tilde{\Delta}$ and L are similar matrices.

3. 0 is an eigenvalue for Δ with eigenvector $1 = (1, 1, \cdots, 1)$. If multiple connected components, define such a 1 vector for each component (and 0 on rest).

4. $\lambda_{max}(\tilde{\Delta}) = 1 + \lambda_{max}(D^{-1/2} A D^{-1/2})$.

Spectral Analysis
Comments on the proof - 2

$$\lambda_{max}(D^{-1/2}AD^{-1/2}) = \max_{\|f\|=1} \langle D^{-1/2}AD^{-1/2}f, f\rangle = \max_{\|f\|=1} \sum_{i,j} A_{i,j} \frac{f_i}{\sqrt{d_i}} \frac{f_j}{\sqrt{d_j}}$$

Next use Cauchy-Schwartz to get

$$\left| \sum_{i,j} A_{i,j} \frac{f_i}{\sqrt{d_i}} \frac{f_j}{\sqrt{d_j}} \right| \leq \sum_i \frac{f_i^2}{d_i} \sum_j A_{i,j} = \sum_i f_i^2 = \|f\|^2 = 1.$$

Thus $\lambda_{max}(\tilde{\Delta}) \leq 2$. Similarly $\lambda_{max}(\Delta) \leq 2(\max_i d_i)$.

Spectral Analysis
Comments on the proof - 2

$$\lambda_{max}(D^{-1/2}AD^{-1/2}) = \max_{\|f\|=1} \langle D^{-1/2}AD^{-1/2}f, f\rangle = \max_{\|f\|=1} \sum_{i,j} A_{i,j}\frac{f_i}{\sqrt{d_i}}\frac{f_j}{\sqrt{d_j}}$$

Next use Cauchy-Schwartz to get

$$\left|\sum_{i,j} A_{i,j}\frac{f_i}{\sqrt{d_i}}\frac{f_j}{\sqrt{d_j}}\right| \leq \sum_i \frac{f_i^2}{d_i}\sum_j A_{i,j} = \sum_i f_i^2 = \|f\|^2 = 1.$$

Thus $\lambda_{max}(\tilde{\Delta}) \leq 2$. Similarly $\lambda_{max}(\Delta) \leq 2(\max_i d_i)$.

5. If the graph is connected, $trace(\tilde{\Delta}) = n = \sum_{i=0}^{n-1} \lambda_i$. Since $\lambda_0 = 0$ we get all statements of Theorem 2.

Spectral Analysis
Comments on the proof - 2

$$\lambda_{max}(D^{-1/2}AD^{-1/2}) = \max_{\|f\|=1} \langle D^{-1/2}AD^{-1/2}f, f\rangle = \max_{\|f\|=1} \sum_{i,j} A_{i,j}\frac{f_i}{\sqrt{d_i}}\frac{f_j}{\sqrt{d_j}}$$

Next use Cauchy-Schwartz to get

$$\left|\sum_{i,j} A_{i,j}\frac{f_i}{\sqrt{d_i}}\frac{f_j}{\sqrt{d_j}}\right| \leq \sum_i \frac{f_i^2}{d_i}\sum_j A_{i,j} = \sum_i f_i^2 = \|f\|^2 = 1.$$

Thus $\lambda_{max}(\tilde{\Delta}) \leq 2$. Similarly $\lambda_{max}(\Delta) \leq 2(\max_i d_i)$.

5. If the graph is connected, $trace(\tilde{\Delta}) = n = \sum_{i=0}^{n-1}\lambda_i$. Since $\lambda_0 = 0$ we get all statements of Theorem 2.
6. Theorem 3 is slightly more complicated (see [2]).

Spectral Analysis
Special graphs: Cycles and Complete graphs

Cycle graphs: like the hexagon in HW2.
Remark: Adjacency matrices are circulant, and so are Δ, $\tilde{\Delta} = L$.

Spectral Analysis

Special graphs: Cycles and Complete graphs

Cycle graphs: like the hexagon in HW2.
Remark: Adjacency matrices are circulant, and so are Δ, $\tilde{\Delta} = L$.

Then argue the FFT forms a ONB of eigenvectors. Compute the eigenvalues as FFT of the generating sequence.

Spectral Analysis

Special graphs: Cycles and Complete graphs

Cycle graphs: like the hexagon in HW2.
Remark: Adjacency matrices are circulant, and so are Δ, $\tilde{\Delta} = L$.

Then argue the FFT forms a ONB of eigenvectors. Compute the eigenvalues as FFT of the generating sequence.

Consequence: The normalized Laplacian has the following eigenvalues:

1. For cycle on n vertices: $\lambda_k = 1 - cos\frac{2\pi k}{n}$, $0 \leq k \leq n-1$.
2. For the complete graph on n vertices:

$$\lambda_0 = 0 \;, \;\; \lambda_1 = \cdots = \lambda_{n-1} = \frac{n}{n-1}.$$

References

- B. Bollobás, **Graph Theory. An Introductory Course**, Springer-Verlag 1979. **99**(25), 15879–15882 (2002).
- F. Chung, **Spectral Graph Theory**, AMS 1997.
- F. Chung, L. Lu, The average distances in random graphs with given expected degrees, Proc. Nat.Acad.Sci. 2002.
- R. Diestel, **Graph Theory**, 3rd Edition, Springer-Verlag 2005.
- P. Erdös, A. Rényi, On The Evolution of Random Graphs
- G. Grimmett, **Probability on Graphs. Random Processes on Graphs and Lattices**, Cambridge Press 2010.
- J. Leskovec, J. Kleinberg, C. Faloutsos, Graph Evolution: Densification and Shrinking Diameters, ACM Trans. on Knowl.Disc.Data, **1**(1) 2007.