Sea Snails

Interpreting the Oceans

Sria Louis

שריה לואיס

WINDWARD

DATAHACK

Sea Snails

Interpreting the Oceans

Mor Nitzan

Maria Dyshel

Noa Weiss

Guy Keren

Who am I?

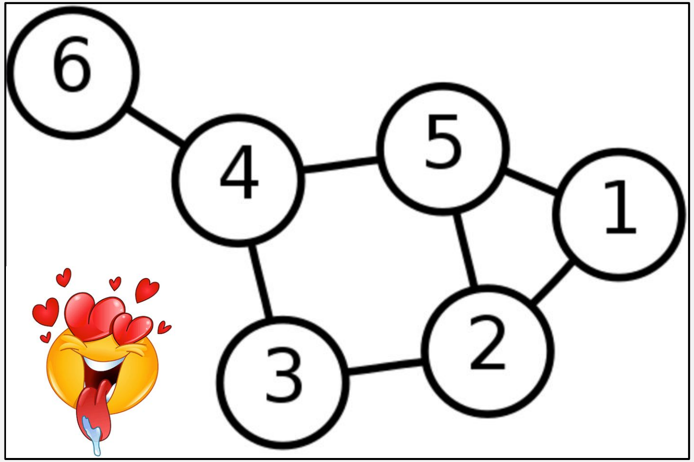
- **Sria louis**
  - BSc math & Physics
  - MSc Computer Science

Who am I?

- **Sria louis**
    - BSc math & Physics
    - MSc Computer Science
- **Not** an expert mariner.
- **Not** an expert data scientist.

Who am I?

- **Sria louis**
    - BSc math & Physics
    - MSc Computer Science
- **Not** an expert mariner.
- **Not** an expert data scientist.
- I do **LOVE** graphs!



Overview

1. The challenge and why this is interesting for you?
2. The data
3. Naïve solution
4. The trick
5. Why we failed + How to improve

The Windward Challenge

Classification problem:

Predict **ship-type**, based on a ship's **behavior**.

Underlying assumption: ships that engage in similar activities are more likely to be of the same type.

Why this is Interesting?

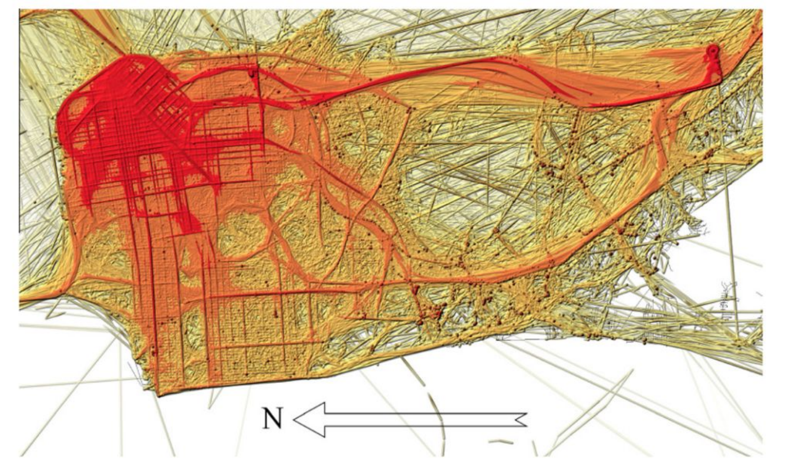
~90% of goods are shipped by sea

- Predict markets by maritime traffic
- Capturing illegal shipping and fishing

Other Spatial Data

- People, cars, animals, ideas, music, products, viruses/viruses, etc.

Trails of taxis in San Francisco

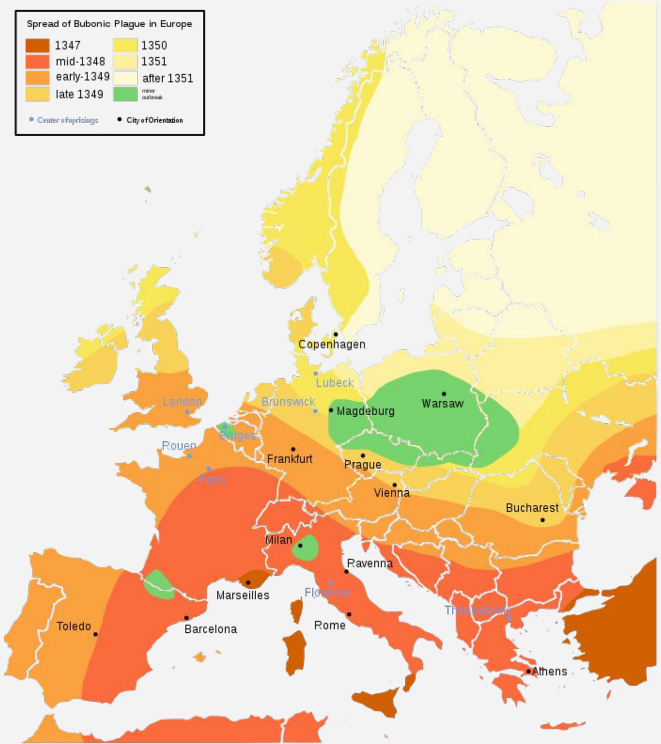


Li, B., Krushinsky, D., Reijers, H. A., & Van Woensel, T. (2014). The share-a-ride problem: People and parcels sharing taxis. European Journal of Operational Research, 238(1), 31-40.

Other Spatial Data

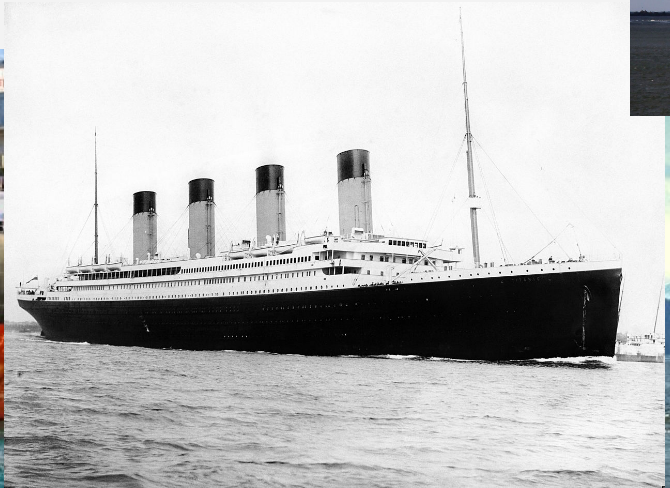
- People, cars, animals, ideas, music, products, viruses/viruses, etc.

The Black Death
Europe 1346–1353



The Data

- 7 **types** of maritime vessels


The Data

- Maritime data
  - Port visits
  - Ship-to-ship meetings

The Data

- Maritime vessel
  - **Port visits**
  - Ship-to-ship meetings

- Vessel ID
- Port ID
- Time in/out
- latitude, longitude

The Data

- Maritime events
  - Port visits
  - **Ship-to-ship meetings**

- Vessel #1
- Vessel #2
- Time
- Latitude, longitude

The Windward Challenge

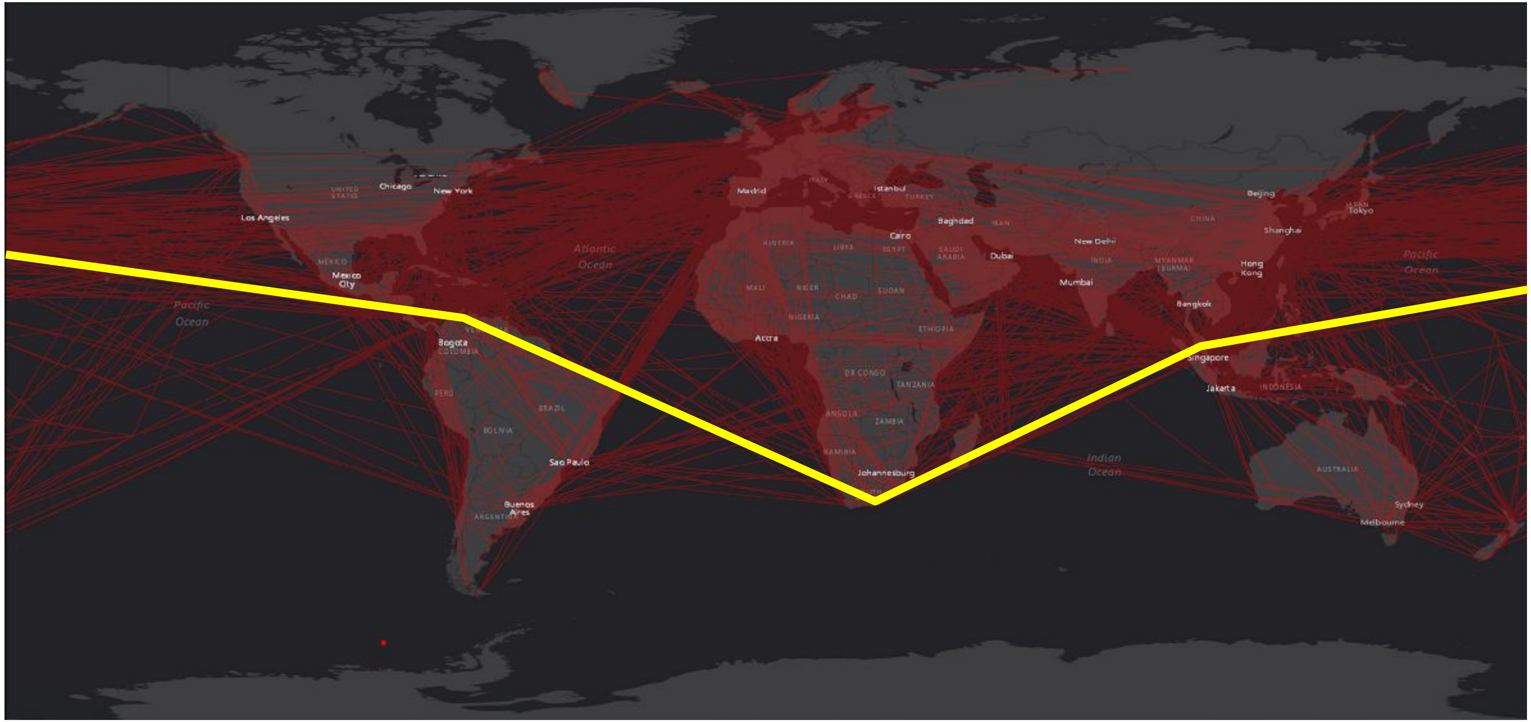
Predict ship-type, based on a ship's behavior.

Solution - Overview

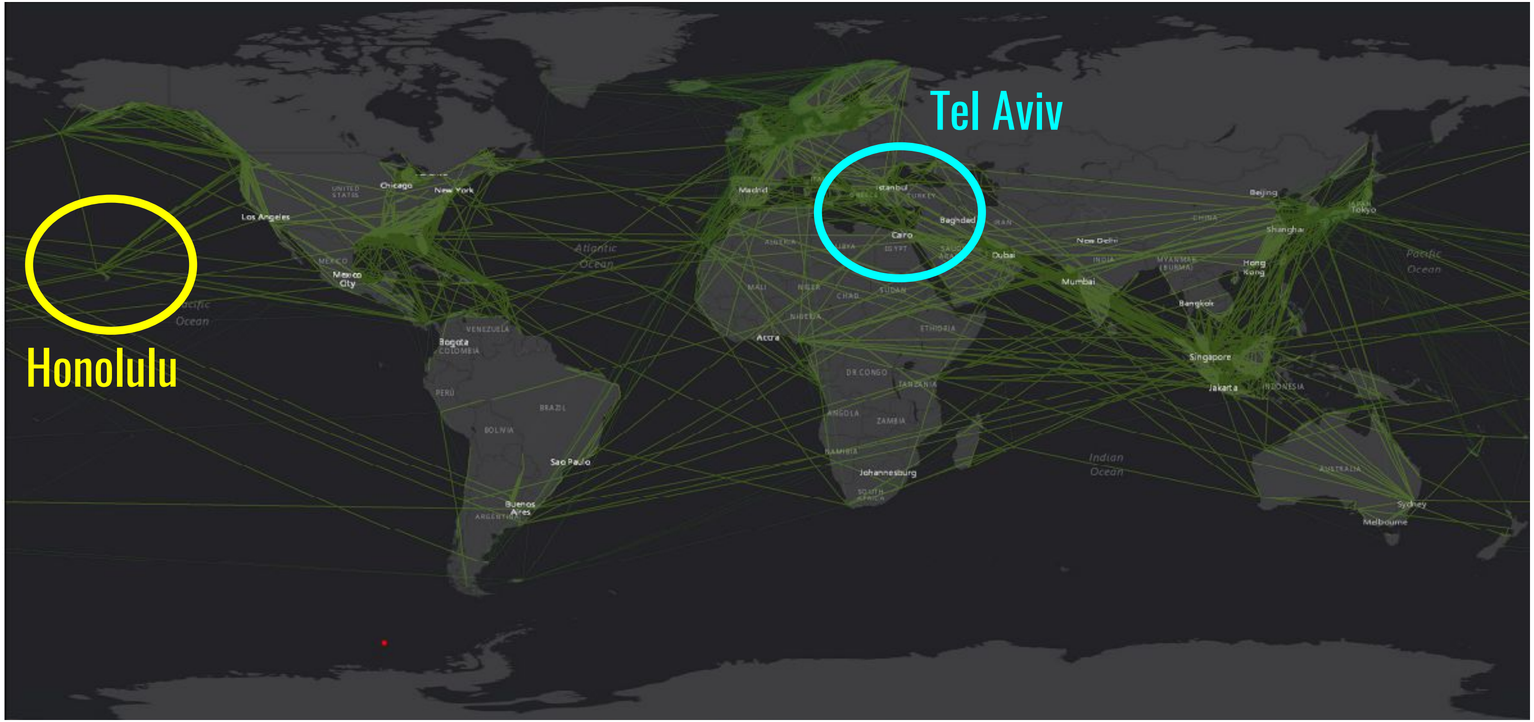
- Descriptive features:
  - Distance, velocity, port time, meeting time etc. (average/std)
  - The secret ingredient
- Different classifiers with cross validation
  - SVM, KNN, Decision Tree, etc.

Intuition: Passenger Vessels

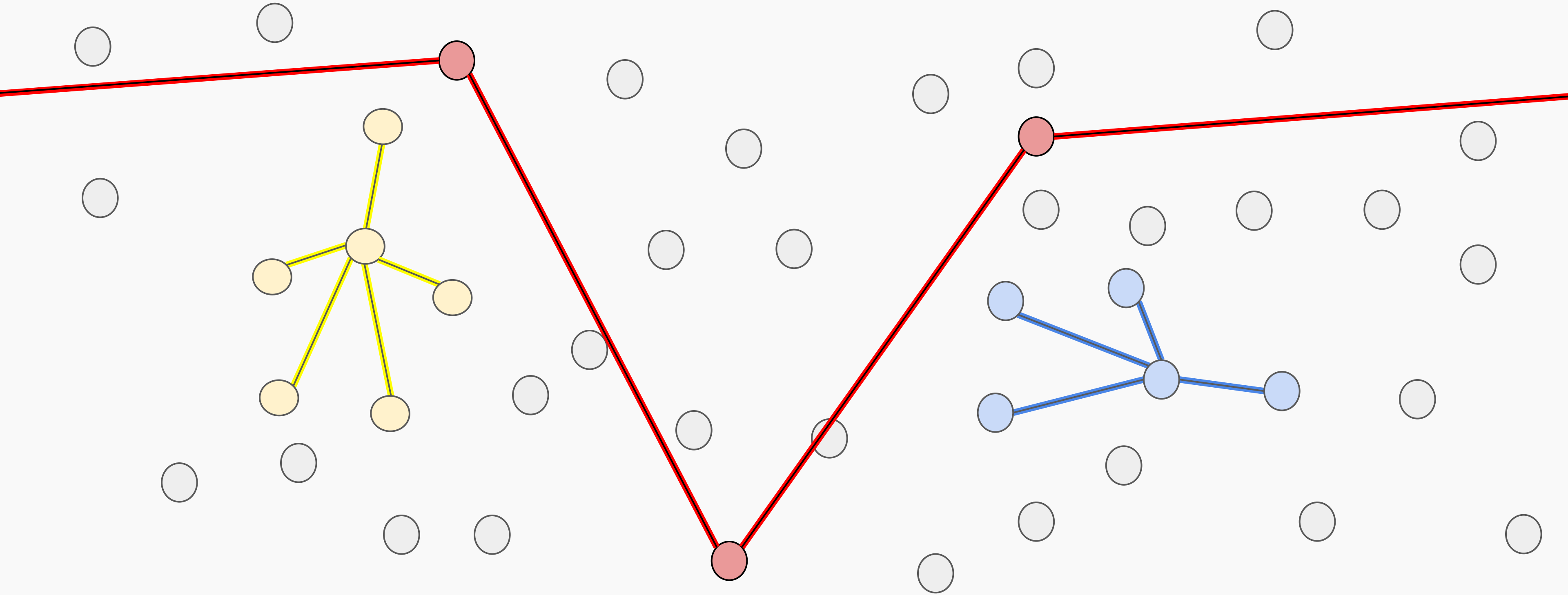
Intuition: Container Vessels



Intuition: Tug Vessels



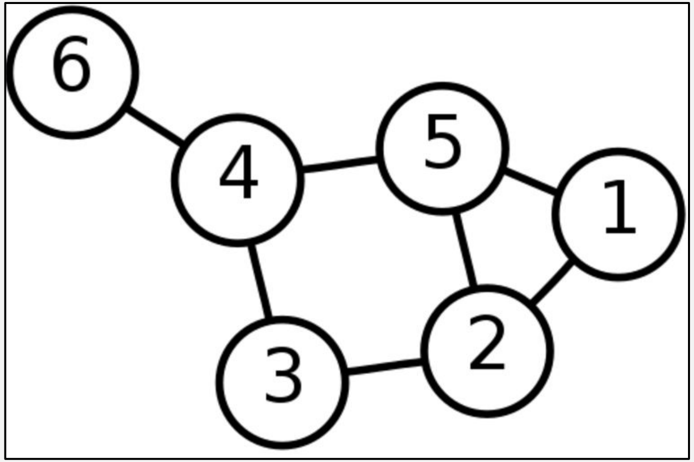
The Idea : sketch



The Idea

We are looking for features that capture the **structure** of graphs.

Recap : Graphs



- Graph G is a set of **vertices** connected with **edges**.
    - cities & roads
    - people & social links
    - **Ports & ships**
- Adjacency Matrix
    - A[i,j]=1 iff there is an edge (i,j)

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Recap : Eigenvalues

Definition: A number c is an eigenvalue of a matrix $\boldsymbol{A}$ if there is a non-zero vector $\boldsymbol{v}$ such that :

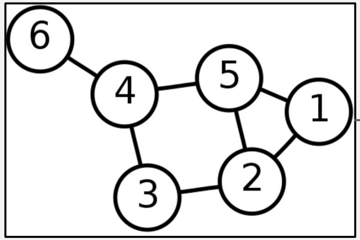
$$\boldsymbol{A}\boldsymbol{v} = c\boldsymbol{v}$$

Recap : Graphs eigenvalues

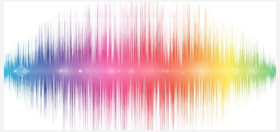
- The spectrum of a square matrix is the set of all its ***eigenvalues***, in descending order.
- The spectrum of a graph is the ***eigenvalues*** of its adjacency matrix.

Recap : Graphs eigenvalues

- The spectrum of a square matrix is the set of all its ***eigenvalues***, in descending order.
- The spectrum of a graph is the ***eigenvalues*** of its adjacency matrix.



$$\begin{bmatrix} 0&1&0&0&1&0 \\ 1&0&1&0&1&0 \\ 0&1&0&1&0&0 \\ 0&0&1&0&1&1 \\ 1&1&0&1&0&0 \\ 0&0&0&1&0&0 \end{bmatrix}$$



$[c_1,c_2,c_3,\dots,c_n]$

The Trick

We will add a new feature:

The spectrum of the ship

Ship Graph

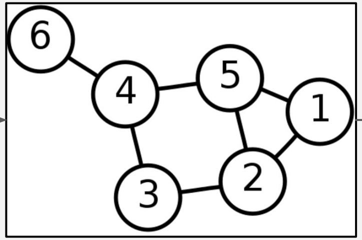
Representing ship behavior as a **graph**:

- **Vertices:** Ports & meeting points
- **Edges:** journey of ship between vertices

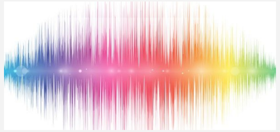
Ship Graph

Representing ship behavior as a **graph**:

- **Vertices:** Ports & meeting points
- **Edges:** journey of ship between vertices



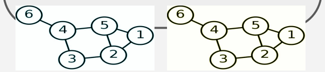
$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

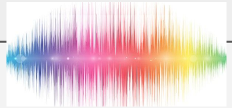


$[c_1, c_2, c_3, \ldots, c_n]$

The trick: eigenvalues!

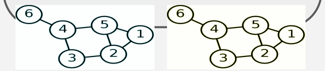
Similar **vessels** → Similar **graphs** → Similar adj. **matrices** → Similar **eigenvalues**

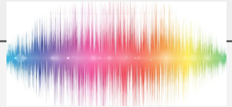




The trick: eigenvalues!

Similar **vessels** → Similar **graphs** → Similar adj. **matrices** → Similar **eigenvalues**

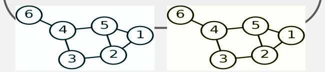


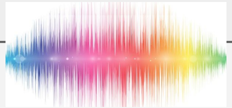


Although the other direction is **not** always true, in practice, eigenvalues are a good approximation for similarity between graphs.

The trick: eigenvalues!

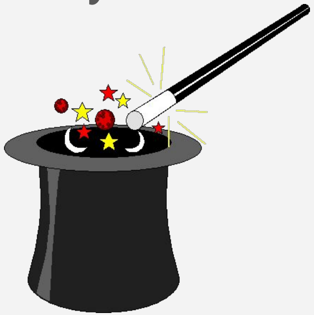
Similar **vessels** → Similar **graphs** → Similar adj. **matrices** → Similar **eigenvalues**





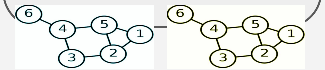
Although the other direction is **not** always true, in practice, eigenvalues are a good approximation for similarity between graphs.

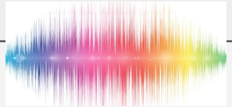
Magic?



The trick: eigenvalues!

Similar **vessels** → Similar **graphs** → Similar adj. **matrices** → Similar **eigenvalues**





<u>**Good place to start**</u>:

- Zhu, P., & Wilson, R. C. (2005, September). A study of graph spectra for comparing graphs. In BMVC.
- Spectral Graph Theory, Daniel A. Spielman - Lecture notes.

<u>**Graph matching problems**</u>:

- Umeyama, S. (1988). An eigendecomposition approach to weighted graph matching problems. IEEE transactions on pattern analysis and machine intelligence, 10(5), 695-703.
- Xu, L., & King, I. (2001). *A PCA approach for fast retrieval of structural patterns in attributed graphs*. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31(5), 812-817.

<u>**Practic usage example in social networks**</u>:

- Leskovec, J., Singh, A., & Kleinberg, J. (2006, April). *Patterns of influence in a recommendation network*. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 380-389). Springer, Berlin, Heidelberg.

To Summarize the Trick:

$$\text{Spectrum}_1 \simeq \text{spectrum}_2 \neq \text{spectrum}_3$$

תכל'ס

We add new features: the eigenvalues of the adjacency matrix of the vessel, namely, “the ship’s spectrum”. Extract using EVD.

ship_id	feature_1	feature_2	...	feature_t	ev_1	ev_2	...	ev_n
53822679								
28121984								

The usual features | The spectral features

Results

In the Windward case, the spectral features improved the results - up to 84% success.

So, what went wrong?

So, what went wrong?

Our model (implicitly) assumed ports in training set are a good representation of ports in “real world” (i.e., test set) - This was painfully far from the truth :-(

Improving Scalability

1. For each vessel, take only the vertices which were **visited**. Sometimes - it's a tiny graph!
2. Similarly to PCA, take only the *k* most prominent eigenvalues.

Sidenote: Similarity to PCA

Graph Spectrum	EVD of the Laplacian matrix (PSD - square of the incidence matrix) sklearn.decomposition.TruncatedSVD
PCA	EVD of the covariance matrix (PSD - square of the centralized data) sklearn.decomposition.PCA

Sidenote:
Similarity to PCA

Graph Spectrum	EVD of the Laplacian matrix (PSD - square of the incidence matrix) sklearn.decomposition.TruncatedSVD
PCA	EVD of the covariance matrix (PSD - square of the centralized data) sklearn.decomposition.PCA

Further Improvements

Richer graph properties can be taken as features:

1. Histogram of In-out degrees
2. Histogram of small subgraphs - “Graph k-profile”

More to research:

3. Vertex & edge weights? Multi-graphs? Digraphs?
4. Eigenvectors? Nodal domains? Laplacian? Seidel?

What next?

- Windward
- Other applications
- YOU!
- Datahack 2017 is coming!


Thanks!
Sria Louis
sria.louis@mail.huji.ac.il