

# Extracting relevant metrics with Spectral Clustering

Evelyn Trautmann

PyData Berlin

6-8 July, 2018



# Introduction



- platform monitoring means keeping track of various metrics

# Introduction



- platform monitoring means keeping track of various metrics
- some effects are visible only in sub-dimensions like single countries, or device types

# Introduction



- platform monitoring means keeping track of various metrics
- some effects are visible only in sub-dimensions like single countries, or device types
- dimension reduction without losing important details

# Clustering



- metrics considered as points in a vector space

# Clustering



- metrics considered as points in a vector space
- similarity defines (inverse) distance measure

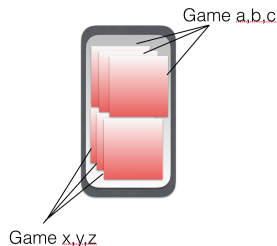
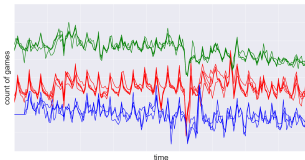
# Clustering



- metrics considered as points in a vector space
- similarity defines (inverse) distance measure
- find clusters of closely related points

# Toy Example - Gaming App

- metric 0    game a: count of games played by paying users
- metric 1    game a: count of games played by non paying users
- metric 2    game b: count of games played by paying users
- metric 3    game b: count of games played by non paying users
- ...





# Similarity Graph

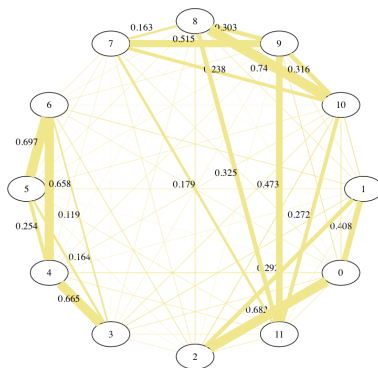


Figure: Similarity Graph

# Matrix Representation

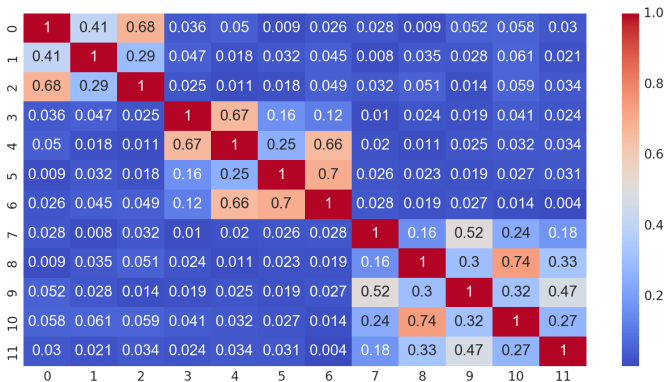


Figure: Matrix Representation

# Matrix Representation

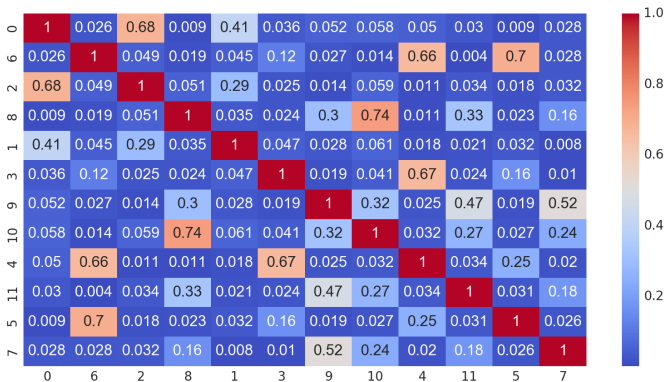


Figure: Matrix Representation

# Toy Example

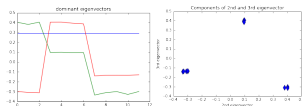


*Toy Example.ipynb*

[https://github.com/metterlein/spectral\\_clustering](https://github.com/metterlein/spectral_clustering).

# Spectral Clustering

- 1: **procedure** SPECTRALCLUSTERING( $X, k$ )
- 2:     compute similarity matrix with pairwise similarities
- 3:     Transform to Graph Laplacian  $L$
- 4:      $v_1, \dots, v_k = \text{eig}(L, k)$  ▷ first  $k$  eigenvectors
- 5:      $U = V^T$  ▷  $k$   $n$ -dimensional  $\rightarrow$   $n$   $k$ -dimensional vectors
- 6:      $\text{clusterAssignment} = \text{kmeans}(U, k)$
- 7:     **return**  $\text{clusterAssignment}$
- 8: **end procedure**[vL07]



# Similarity Function



- Similarity Function  $\phi : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, 1]$

# Similarity Function



- Similarity Function  $\phi : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, 1]$
- radial basis function [PVG<sup>+</sup>11]

$$\phi(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

# Similarity Function



- Similarity Function  $\phi : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, 1]$
- radial basis function [PVG<sup>+</sup>11]

$$\phi(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

- correlation based similarity



# Similarity Function



- Similarity Function  $\phi : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, 1]$
- radial basis function [PVG<sup>+</sup>11]

$$\phi(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

- correlation based similarity
- handle negative correlations

# Number of Clusters



- with clear block structure eigenvalue gap is obvious

# Number of Clusters



- with clear block structure eigenvalue gap is obvious
- if connectivity amongst blocks is too large, determination of cluster count is getting complex

# Number of Clusters



- with clear block structure eigenvalue gap is obvious
- if connectivity amongst blocks is too large, determination of cluster count is getting complex
- possible approach:
  - PCA explained variance

# Data Description



- metrics representing user activities, payments, etc

# Data Description



- metrics representing user activities, payments, etc
- each metric is divided in several dimensions like country, gender, etc

# Data Description



- metrics representing user activities, payments, etc
- each metric is divided in several dimensions like country, gender, etc
- combination of metrics and dimensions generates around 300 timeseries in our example.

# Data Preparation



- aggregation per day



# Data Preparation



- aggregation per day
- normalization  $X = \frac{1}{\sigma} (X - \mu)$ , with  $\mu$  mean value and  $\sigma$  standard deviation

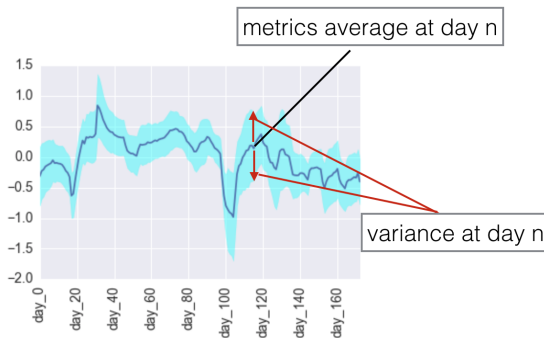
# Data Preparation



- aggregation per day
- normalization  $X = \frac{1}{\sigma} (X - \mu)$ , with  $\mu$  mean value and  $\sigma$  standard deviation
- rolling mean of 7 days to smooth weekly periodicity

# Measure Clustering Quality

- compute variance for each timepoint over all cluster members
- optimal clustering minimizes variance for each cluster



# Real Data Example



*Real Data Example.ipynb*

[https://github.com/metterlein/spectral\\_clustering](https://github.com/metterlein/spectral_clustering)

# Conclusion



- clustering metrics reduces dimensionality of observation space

# Conclusion



- clustering metrics reduces dimensionality of observation space
- small number of interpretable cluster representatives help keeping track of main platform dynamics

# Conclusion



- clustering metrics reduces dimensionality of observation space
- small number of interpretable cluster representatives help keeping track of main platform dynamics
- by cluster assignments can be discovered unexpected relations between several metrics

# Outlook



- investigate cluster assignment change over time



# Outlook



- investigate cluster assignment change over time
- hierarchical approach: clustering sub-blocks

# Outlook



- investigate cluster assignment change over time
- hierarchical approach: clustering sub-blocks
- add time shifted series to recognize Granger causalities

# References I



F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research **12** (2011), 2825–2830.



Evelyn Trautmann, *Spectral clustering demo*, [https://github.com/metterlein/spectral\\_clustering](https://github.com/metterlein/spectral_clustering), 2018.



Ulrike von Luxburg, *A tutorial on spectral clustering*, CoRR **abs/0711.0189** (2007).

# Thank You!

Questions?

evelyn.trautmann@lovoo.com

# Graph Laplacian



- Graph Laplacian with similarities as off-diagonal entries

# Graph Laplacian



- Graph Laplacian with similarities as off-diagonal entries
- negative row sums on diagonal

$$L = \begin{cases} l_{ij} \geq 0, \text{ for } i \neq j \\ l_{ii} = -\sum_{k \neq i} l_{ik}, \text{ for } i = j. \end{cases}$$

# Graph Laplacian



- Graph Laplacian with similarities as off-diagonal entries
- negative row sums on diagonal

$$L = \begin{cases} l_{ij} \geq 0, \text{ for } i \neq j \\ l_{ii} = -\sum_{k \neq i} l_{ik}, \text{ for } i = j. \end{cases}$$

- matrix exponential is a stochastic matrix

# Graph Laplacian



- Graph Laplacian with similarities as off-diagonal entries
- negative row sums on diagonal

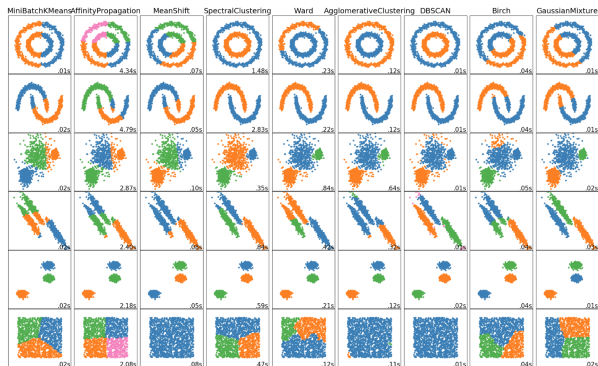
$$L = \begin{cases} l_{ij} \geq 0, \text{ for } i \neq j \\ l_{ii} = -\sum_{k \neq i} l_{ik}, \text{ for } i = j. \end{cases}$$

- matrix exponential is a stochastic matrix
- results from Markov chain theory applicable

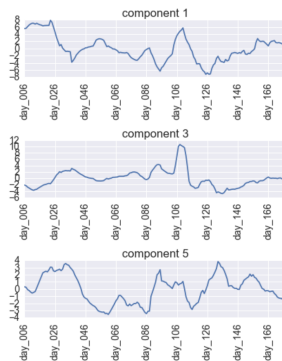
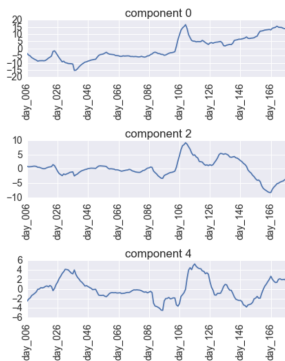


# Clustering Methods

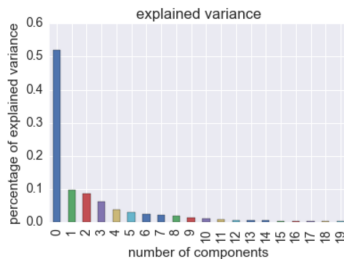
[PVG<sup>+</sup>11]



## PCA



# PCA



## Extract Platform Dynamics by means of Cluster Centers



- compute average timeseries per cluster
- illustrate average cluster timeseries with variance corridor
- visualize metrics types and subdimensions entering respective clusters

