Lecture 1: Introduction & Roadmap

Foundations of Data Science: Algorithms and Mathematical Foundations

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk

CDT in Mathematics of Random System
University of Oxford

19 September, 2023

https://www.stats.ox.ac.uk/~cucuring/MathCDT.htm

Course overview

- combines both theoretical and practical approaches
- Goal 1: understand the mathematical, statistical and algorithmic foundations behind some of the state-of-the-art algorithms for tasks in machine learning/data mining
  - organization and visualization of data clouds
  - measures of correlation and dependence
  - dimensionality reduction •
  - clustering (point clouds and graphs/networks) •
  - network analysis •
  - ranking •
  - regression
- Goal 2: exposure to practical examples drawn from a wide range of topics including social network analysis•, finance•, image processing, biology, engineering, etc

Where appropriate, the latest research developments and trends in the respective areas will be very briefly presented.

Books & References

- Slides will be made available on my webpage at:
  `http://www.stats.ox.ac.uk/~cucuring/MathCDT.htm`

- *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani, freely available at
  `http://faculty.marshall.usc.edu/gareth-james/ISL/`
- *Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science*, by Afonso S. Bandeira
  `https://people.math.ethz.ch/~abandeira/TenLecturesFortyTwoProblems.pdf`
- *Mathematics of Data Science*, by A. Bandeira, A. Singer, T. Strohmer
  `https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf`
- *Lecture Notes for Mathematics of Machine Learning* , by Afonso S. Bandeira & Nikita Zhivotovskiy ETH Zurich
  `https://people.math.ethz.ch/~abandeira/Math_of_ML_Lecture_Notes2021.pdf`

- Popular references in the ML/data mining:
  - *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman. `http://www-stat.stanford.edu/ElemStatLearn`
  - *Pattern Recognition and Machine Learning Book*, by C. Bishop.

Prerequisites

Probability:

- event, random variable, indicator variable
- probability mass function, probability density function, cumulative distribution function
- joint distribution, marginal distribution
- conditional probability, Bayes's rule
- independence
- expectation, variance
- uniform, exponential, binomial, Poisson, Gaussian distributions

Statistics:

- sampling from a population; mean, variance, standard deviation, median, covariance, correlation, and their sample versions; histogram, scatter-plots, box-plots;
- linear regression, response and predictor variables, coefficients, residuals.

Prerequisites

Linear algebra:

- vector and matrix arithmetic; quadratic forms
- eigenvalues and eigenvectors of a matrix
- generalized eigenvalue problems

Programming:

- arithmetic (scalar, vector, and matrix operations)
- writing functions
- reading in data sets from csv files
- using and manipulating data structures (subset a data frame)
- installing, loading, and using packages
- plotting

Wilson et al., *Good enough practices in scientific computing*, PLoS computational biology, 2017 Jun 22;13(6)

A list of tentative topics

1. Introduction & syllabus
2. Statistical learning
3. Measures of correlation and dependency - (i)
4. Measures of correlation and dependency - (ii)
5. Singular Value Decomposition (SVD), rank-k approximation, Principal Component Analysis (PCA)
6. PCA in high dimensions and random matrix theory (Marcenko-Pastur); applications to finance

Nonlinear dimensionality reduction methods:

7. Multidimensional scaling, ISOMAP, Locally Linear Embedding (LLE), Laplacian Eigenmaps
8. Diffusion Maps and Vector Diffusion Maps

9. Basics of spectral graph theory
10. Networks (i) intro, summary statistics, motifs, networks models
11. Networks (ii) network centrality measures, modularity, core-periphery, further topics on networks

Topics

12. Random Graphs Properties

Clustering:

13. k-means, Spectral clustering and Cheeger's inequality
14. Stochastic Block Models: spectral & semidefinite relaxations
15. Constrained clustering, clustering of signed graphs & correlation clustering, clustering directed graphs (digraphs)

Estimation from pairwise measurements:

16. The Page-Rank algorithm
17. Ranking from pairwise incomplete noisy information
18. Angular/Group synchronization (spectral & semidefinite programming relaxations)

Regression:

19. OLS & practical considerations
20. Modern regression: Ridge, LASSO, Elastic Net

Misc:

21. Low-rank matrix completion, Procrustess analysis, Multiplicative Weights Update Algorithm

Additional Topics (if we had more time)

1. Classification: logistic regression, support vector machines, linear discriminant analysis.
2. Tree-based methods for classification and regression
3. Bagging, Boosting, Random Forests
4. The Multiplicative Weights Update Method: a Meta Algorithm and Applications
   - Arora, Sanjeev, Elad Hazan, Satyen Kale; Theory of Computing 2012
5. Johnson-Lindenstrauss Lemma; approximate nearest neighbors

Additional Topics (if we had more time)

Other potential topics

6. electrical networks and random walks
7. graph sparsification and Laplacian linear system solvers ($Lx = b$)
8. graph embeddings from noisy distances
9. non-negative matrix factorization (NMF)
10. matrix algorithms using sampling; sketch of a large matrix

What is data science/data mining?

Given (usually very large) data sets, how do you discover structural properties and make predictions?

Two very broad categories of problems:

- Unsupervised learning: discover structure. E.g., given measurements $X_1, \ldots, X_n$, learn some underlying group structure based on the pattern of similarity between pairs of points ("working in the blind")
- Supervised learning: make predictions. E.g., given measurements $(X_1, Y_1), \ldots (X_n, Y_n)$, learn a model to predict Y_i from X_i

---

- Semi-supervised learning: only for m (with $m << n$) observations we have both predictor + response measurements.
- *"the more you fit, the more you overfit"*

What is data science/data mining?

Given (usually very large) data sets, how do you discover structural properties and make predictions?

Two very broad categories of problems:

- MOST OF THIS COURSE $---->>>$ Unsupervised learning: discover structure. E.g., given measurements $X_1, \ldots, X_n$, learn some underlying group structure based on the pattern of similarity between pairs of points ("working blind")
- Supervised learning: make predictions. E.g., given measurements $(X_1, Y_1), \ldots (X_n, Y_n)$, learn a model to predict Y_i from X_i

---

- Semi-supervised learning: only for m (with $m << n$) observations we have both predictor + response measurements.
- *"the more you fit, the more you overfit"*

This course

- Combines both applied and theoretical perspectives, though for some of the topics the emphasis will be on the algorithms & methodology

Aim to understand what is it that we are trying to do

- Often, it's not enough to load your data in Python/R/Matlab, use any available packages and expect to get an answer that makes sense/is reasonable

- Understand your data! **Data is often very messy, noisy and incomplete**

- Be able to identify what the end goal is, and based on that (and the given data) identify what tools are available

Trade-offs: exact versus approximation

- Many problems are computationally hard to solve exactly
- In order to come up with tractable algorithms (that run in polynomial time) aim for an approximate solution
- **Approximation algorithms** can often perform well (sometimes they even find the exact solution (provably)!) and scale well computationally when applied to very large problems
- Polynomial time algorithms are often not enough, some applications demand close to linear-time complexity (sometimes even sublinear!)

Bias-variance tradeoff

In supervised learning, when moving beyond the training set:

- If the model is too simple, the solution is biased and does not fit the data
- If the model is too complex, the solution is very sensitive to small changes in the data

Problem of simultaneously minimizing two sources of error

- **bias**: difference btwn truth and what you expect to learn
  - high bias leads to missing out on the relevant relations between features and target outputs (*underfitting*).
  - decreases with more complex models
- **variance**: difference between what you learn from a particular data set and what you expect to learn. Arises from sensitivity to small fluctuations in the training set.
  - high variance leads to *overfitting*: modeling the random noise in the training data, rather than the intended outputs.
  - variance decreases with simpler models

Interpretability versus forecasting power

- trade-off between a model that is interpretable and one that predicts well under general circumstances

- essay on the distinction between explanatory and predictive modeling

  To Explain or to Predict?

  by Galit Shmueli, Statistical Science 2010, Vol. 25, No. 3, 289–310

  `https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf`

Occam's razor

- a problem-solving principle known as the 'law of parsimony'
- when faced with different competing hypotheses that predict equally well, choose the one with the fewest assumptions
- usually, more complex models may provide better predictions, but in the absence of differences in predictive power, the fewer assumptions the better

Statistics vs. Machine Learning

- Brian D. Ripley: *"machine learning is statistics minus any checking of models and assumptions"*

"Statistical Modeling: The Two Cultures", Leo Breiman, Statistical Science, 16 (3), 2001; argued that

- statisticians rely too heavily on data modeling and assumptions
- machine learning techniques are making progress by instead relying on the predictive accuracy of models

Statistics vs. Machine Learning

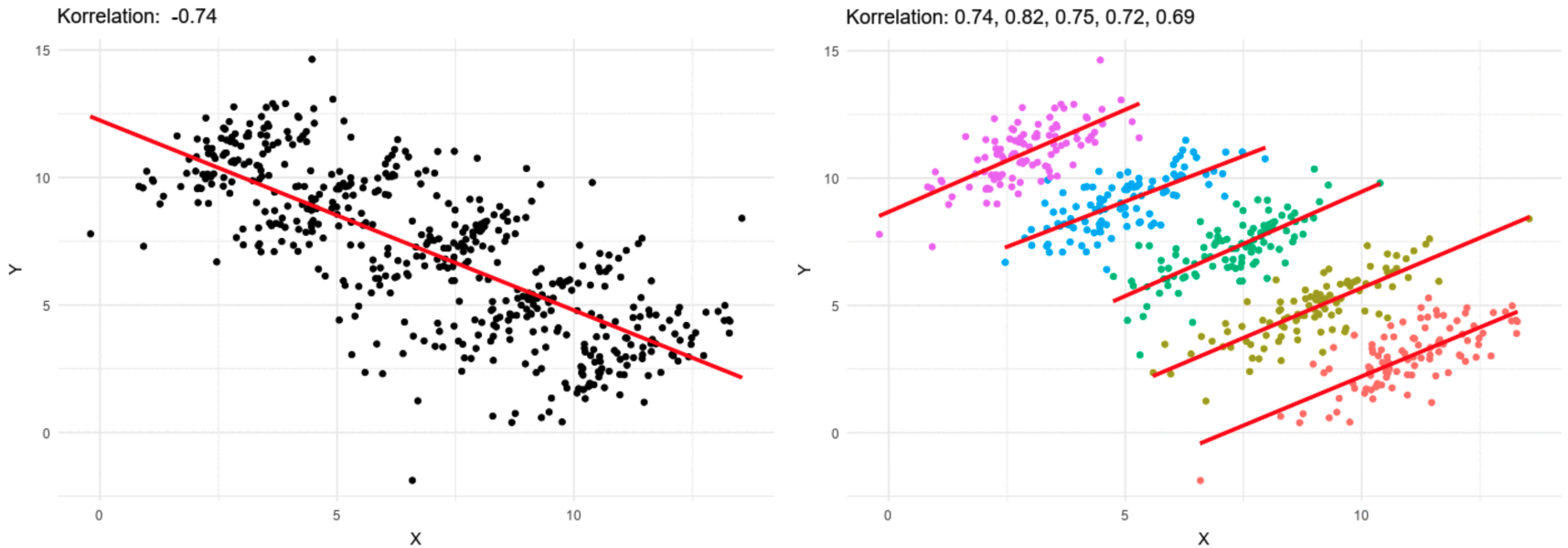
More recently, statisticians focused more on finite-sample properties, and algorithms for massive data sets (*big data*).

Some, still ongoing, differences between the two communities:

- Statistics papers are more formal and often comes with proofs, while Machine Learning papers are more open to new methodologies even if the theory is lacking (for now)
- The Machine Learning community primarily publishes in conferences and proceedings, while statisticians use journal papers (much slower process)
- Some statisticians still focus on areas which are well outside the scope of ML (survey design, sampling, industrial statistics, survival analysis, etc)

Simpson's paradox - beware!

Phenomenon in statistics when certain trends that appear when a dataset is separated into groups are reversed when the data are aggregated.



- can be resolved when confounding variables and causal relations are appropriately addressed in the statistical modeling
- misleading results that the misuse of statistics can generate

Final thoughts: No universal data mining recipe book

- hard to say which methods will work best in what situations
- sometimes, it's crystal clear what method one should follow
- most often, we have little intuition a-priori on what approach or set of tools should we use. Need to
  - understand the data first and the task at hand
  - understand the methods and their assumptions
  - make an educated guess on how to proceed
- often, customized tools are required to handle problem particularities (eg., response variable is highly imbalanced, as in default rate prediction)
- sometimes (if enough resources are available) one often tries many different methods, and chooses the one which gives best (out-of-sample) results
  - most competitions are won by **ensemble methods** - techniques that create multiple models and then combine them to produce improved results
  - **stacking**: considers heterogeneous weak learners, learns them in parallel and combines by training a meta-model

Classification of handwritten digits

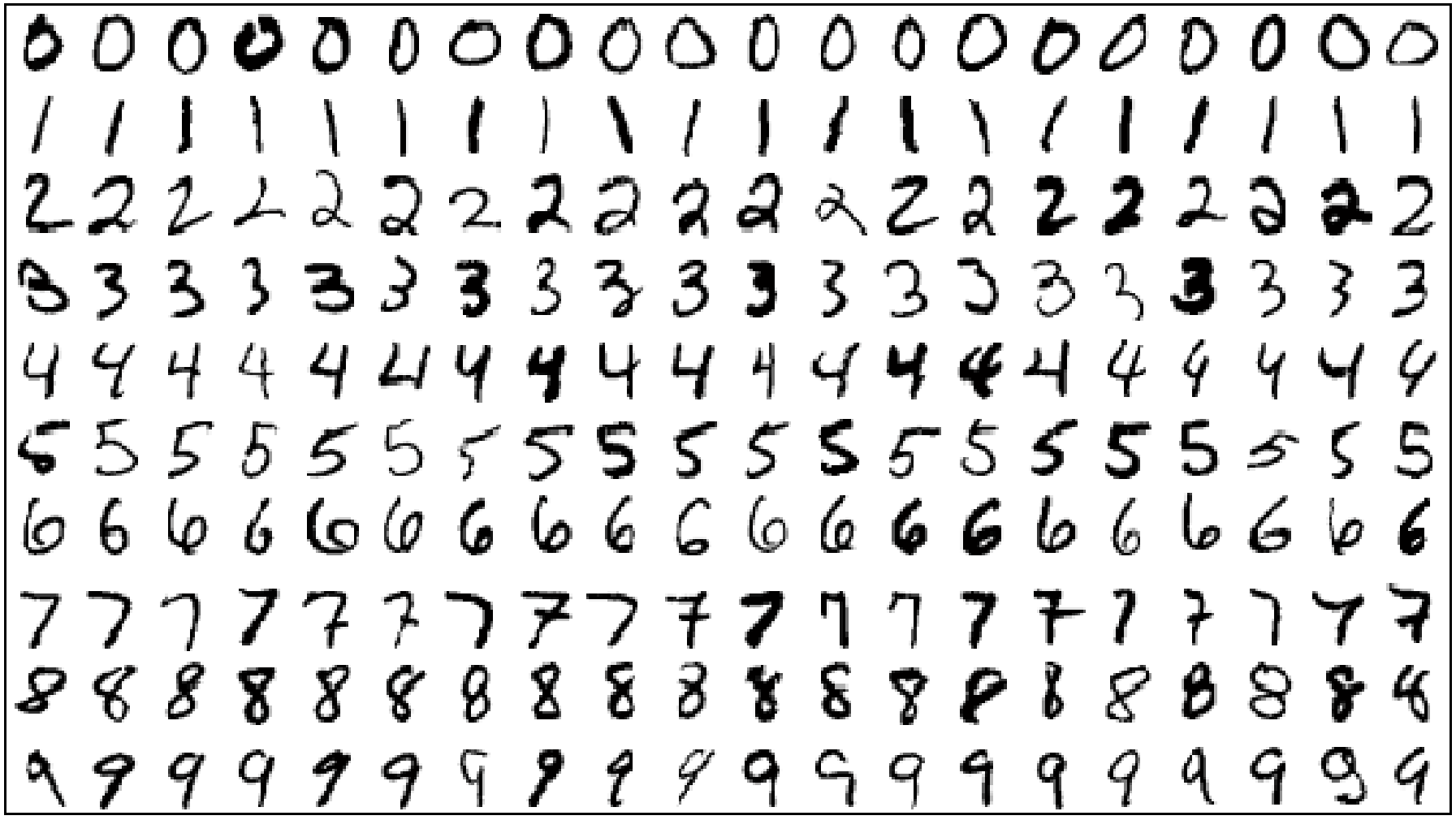


Figure: Automatic detection of handwritten postal codes.

Facebook: friend suggestions & social network analysis

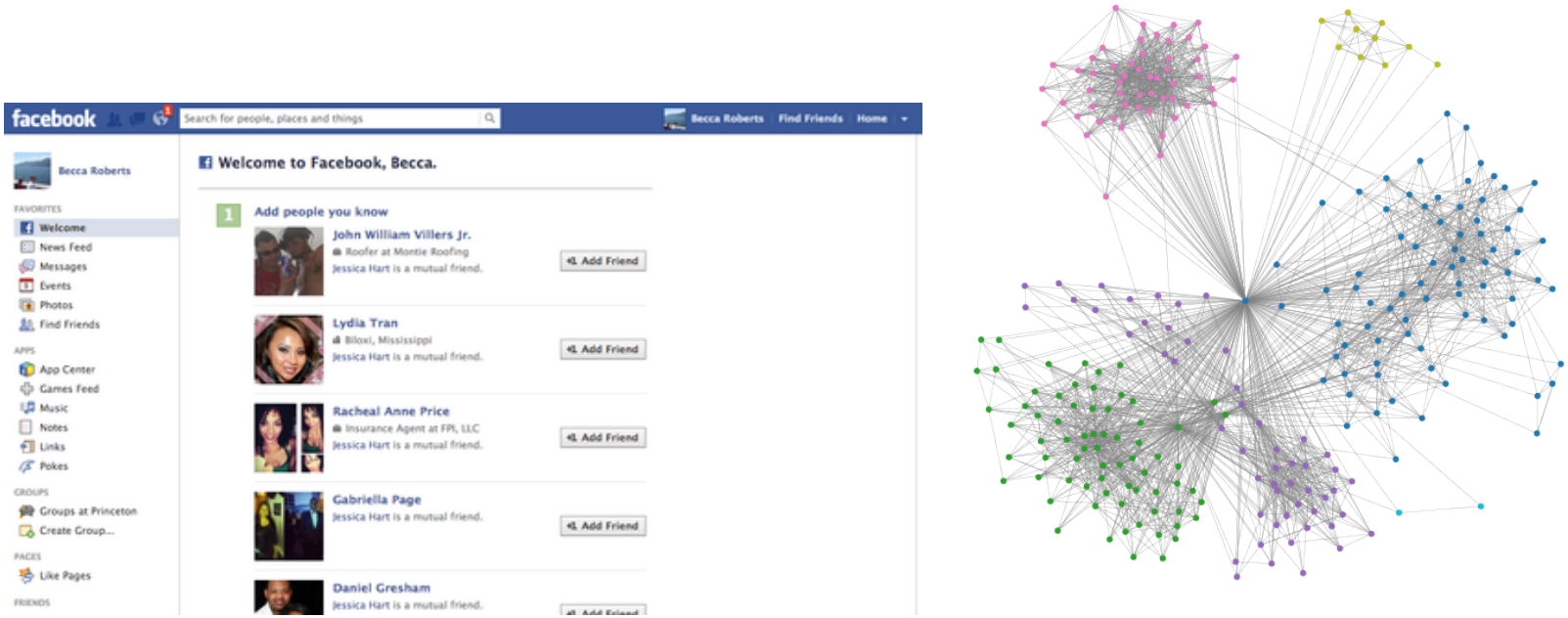


Figure: Left: People you may know. Right: Community detection in a Facebook ego network.

- based on a number of features including *"mutual friends, work and education information, networks you are part of, contacts you have imported and other factors"*.

https://www.databentobox.com/2019/07/28/facebook-friend-graph/

Forecasting the stock market

Figure: Price of Google stock.

Source: `http://businessforecastblog.com/forecasting-googles-stock-price-goog-on-20-trading-day-horizons/`

Netflix: the $ 1,000,000 Prize

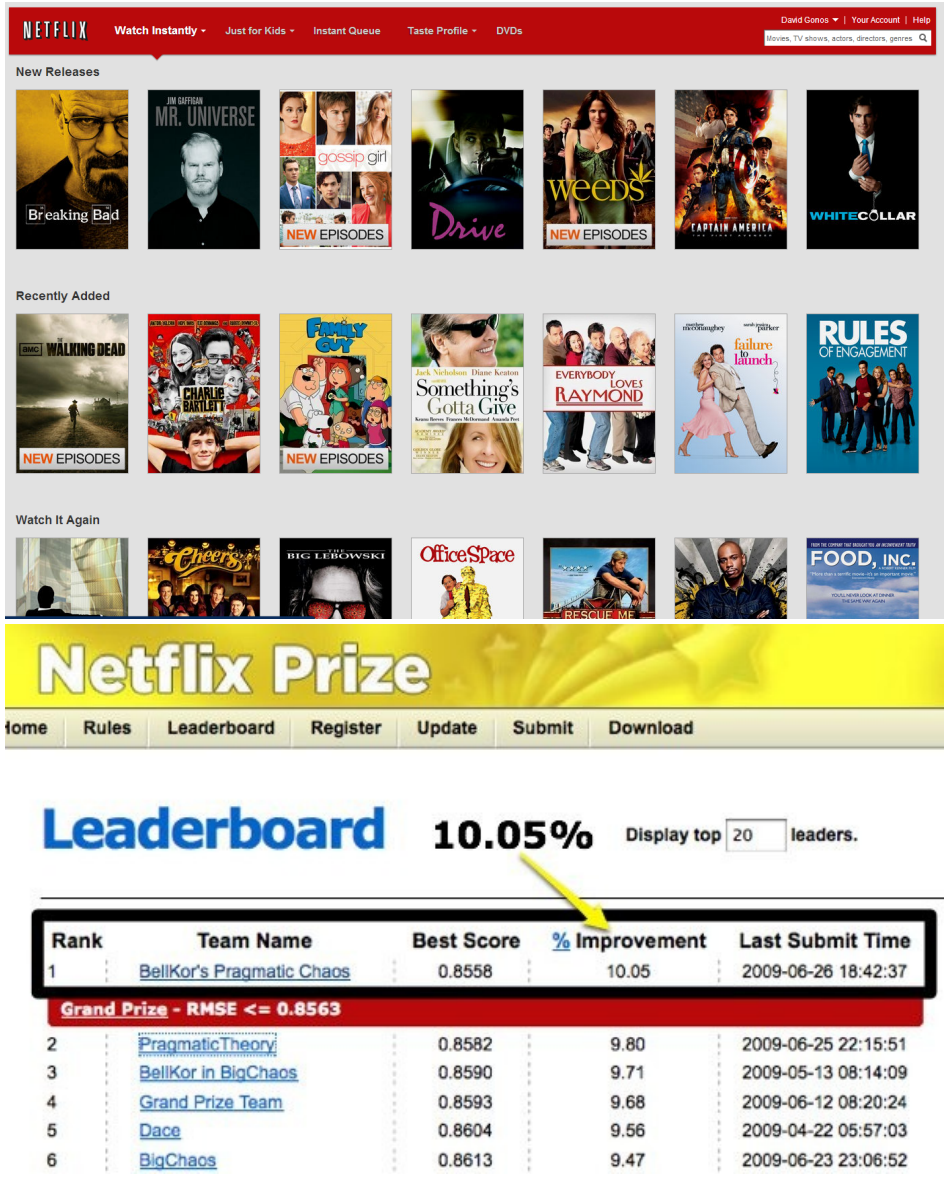


Figure: Movies you might enjoy.

Identifying patterns in migration networks

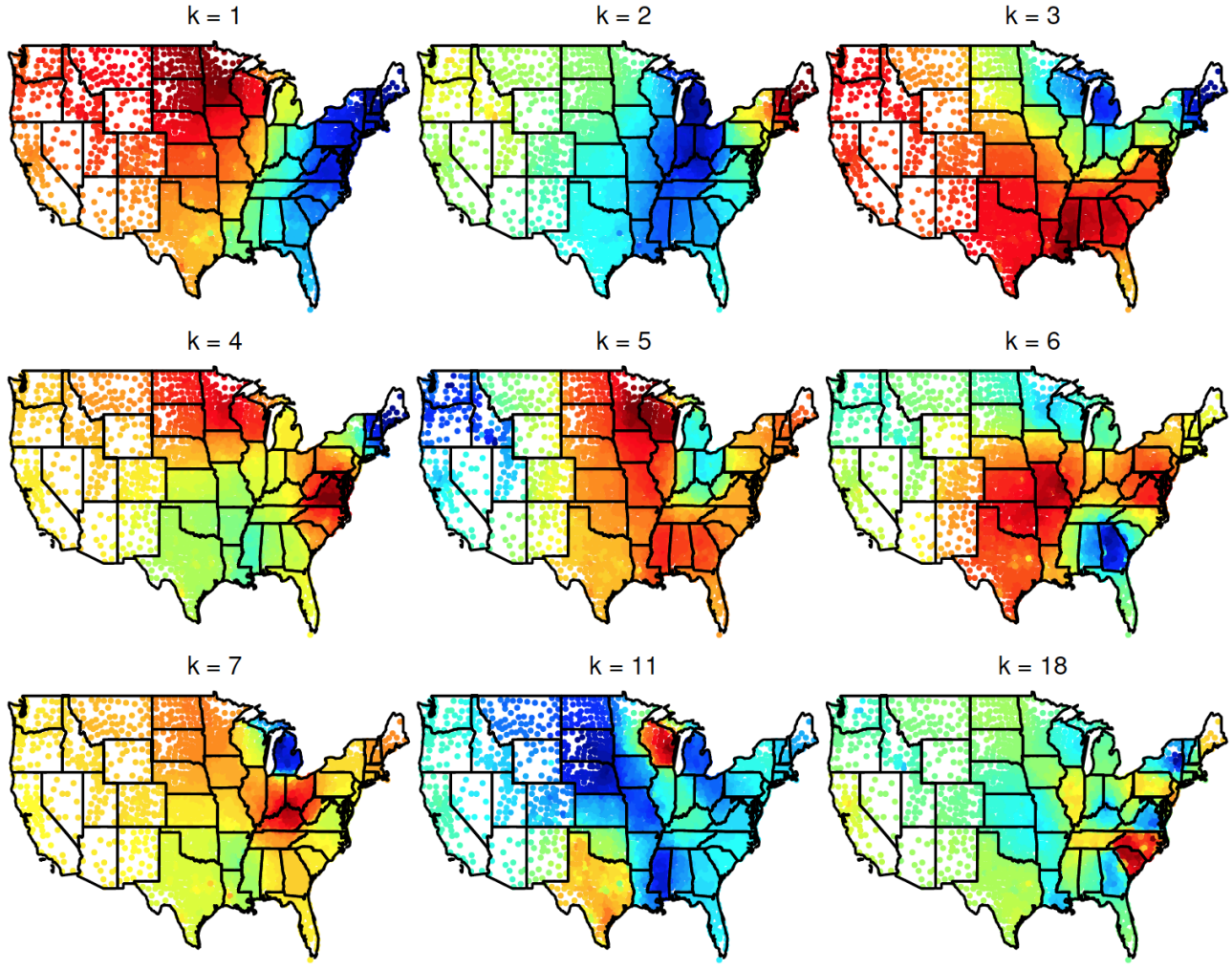


Figure: Eigenvector colourings for the similarity matrix $W_{ij} = \frac{M_{ij}^2}{P_i P_j}$, where M_{ij} denotes the number of people who migrated from county i to county j (during 1995-2000; US Census data), and P_i denotes the population of county i.

Ranking Courses in the *(UCLA) Math Curriculum*

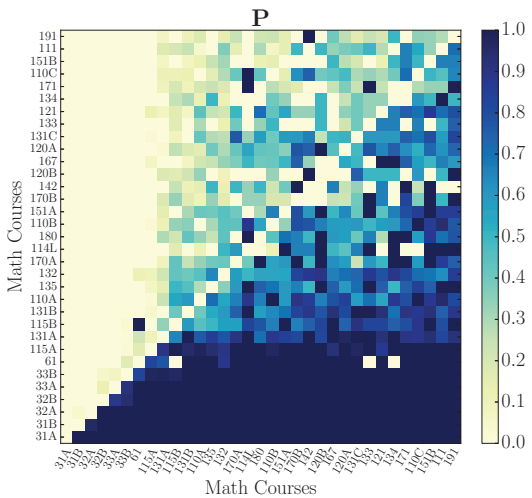


Fig. 1 The **P** matrix for A students with a Pure Mathematics focus (courses ordered by PageRank).

Table 1 Comparing the A and C students in 3 majors using SyncRank.

Applied Mathematics		**Pure Mathematics**	
A ($n_s = 140$)	**C** ($n_s = 198$)	**A** ($n_s = 86$)	**C** ($n_s = 95$)
Lin. Algebra I	Lin. Algebra I	Discr. Struct.	Lin. Algebra I
Discr. Struct.	Discr. Struct.	Lin. Algebra I	Hist. of Math
Real Analysis I	Probability I	Real Analysis I	Real Analysis I
Probability I	Real Analysis I	Lin. Algebra II	Discr. Struct.
Complex Analysis	Algebra I	Algebra I	Algebra I
Nonlin. Syst.	Num. Analysis I	Real Analysis II	Ord. Diff. Eqn.'s
Num. Analysis I	Graph Theory	Ord. Diff. Eqn.'s	Complex Analysis
Math Modeling	Real Analysis II	Complex Analysis	Game Theory
Real Analysis II	Act. Math	Probability I	Probability I
Algebra I	Nonlin. Syst.	Algebra II	Graph Theory
Graph Theory	Math Modeling	Graph Theory	Num. Analysis I
Ord. Diff. Eqn.'s	Hist. of Math	Real Analysis III	Optimization
Game Theory	Complex Analysis	Num. Analysis I	Number Theory
Research Seminar	Probability II	Logic	Algebra II

Network of financial assets

- Mel MacMahon and Diego Garlaschelli. Phys.Rev.X5, 2015. *Community Detection for Correlation Matrices*.

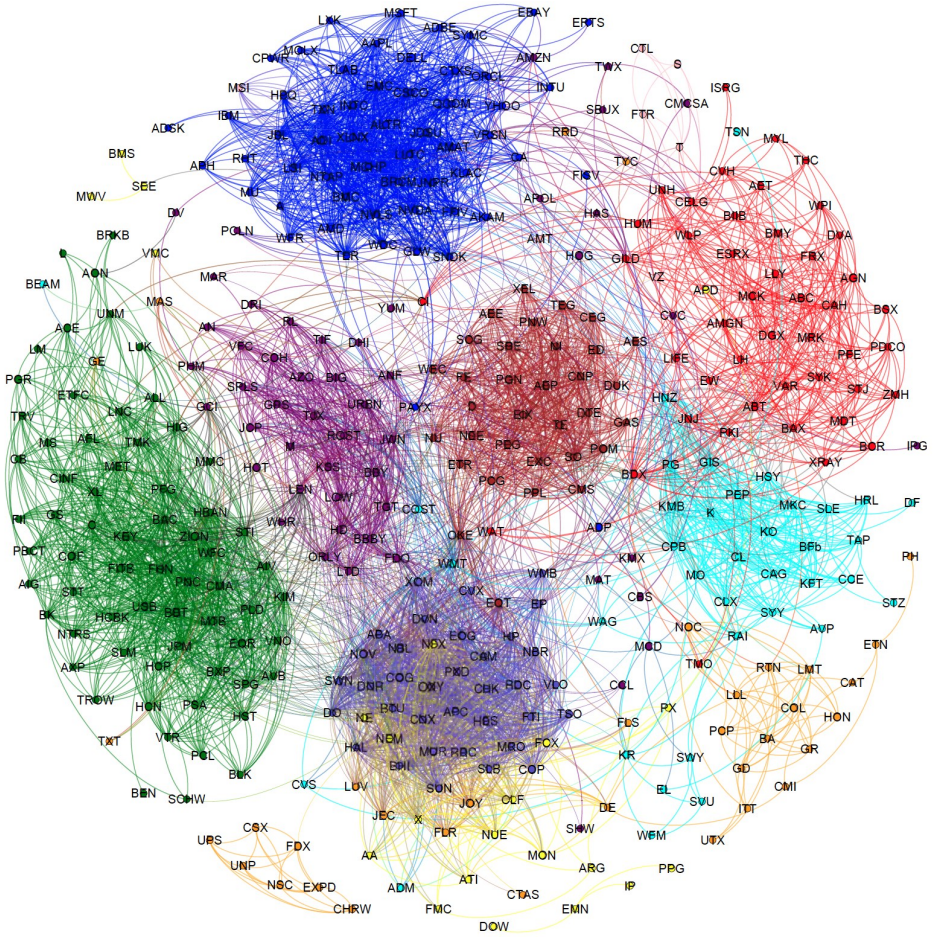


Figure: Asset correlation matrix after thresholding. The color of each node represents the industry sector to which that stock belongs. The force-based layout clearly indicates the existence of strong connections between stocks of the same industry sector.

Financial time series clustering

- Clustering of the empirical correlation matrix of 1500 time series (stocks contained in the S&P 1500 index) • Compute the bottom $k = 10$ eigenvectors of $\overline{L}$, and run a standard machine learning clustering algorithm (k-means++) to recover k clusters.

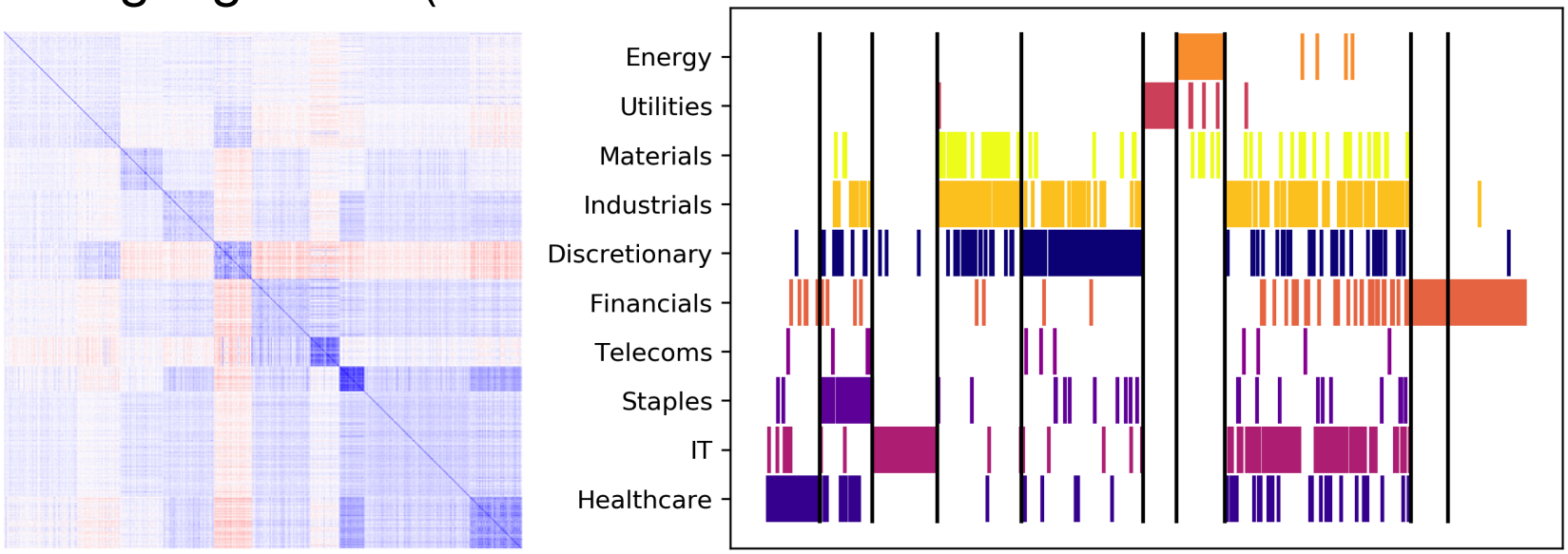


Figure: Left: the adjacency matrix A with rows/columns sorted in accordance to cluster membership. Right: Sector decomposition of the recovered clusters (based on a standard classification of the US economy into sectors). See link for details: GICS link.

M. Cucuringu, P. Davies, A. Glielmo, H. Tyagi, *SPONGE: A generalized eigenproblem for clustering signed networks*, AISTATS 2019 (python code available)

High-dimensional Covariance Regularization

To impose structure, the regularization pipeline involves truncating off-block (outside of GICS sectors) entries of the residual matrix to 0s.

- 9 sector factors proxied by their ETFs: Energy (XLE), Materials (XLB), Industrials (XLI), Consumer Discretionary (XLY), Consumer Staples (XLP), Health Care (XLV), Financial (XLF), Information Technology (XLK), Utilities (XLU).
- GICS groups can also be replaced by **data driven clusterings**.

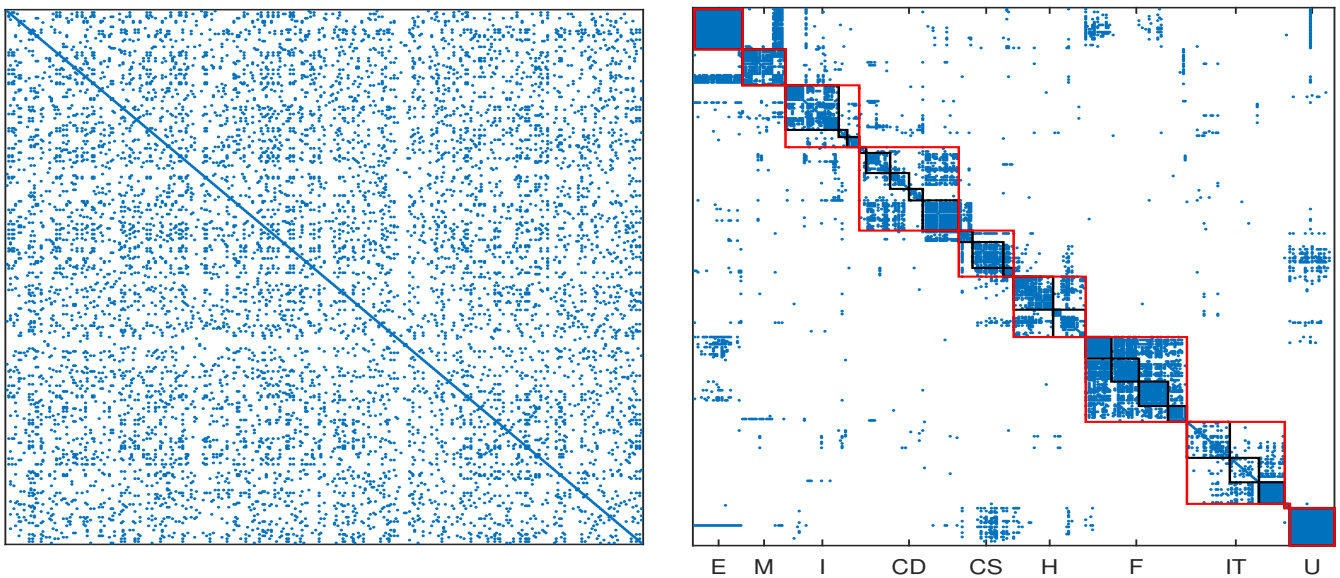
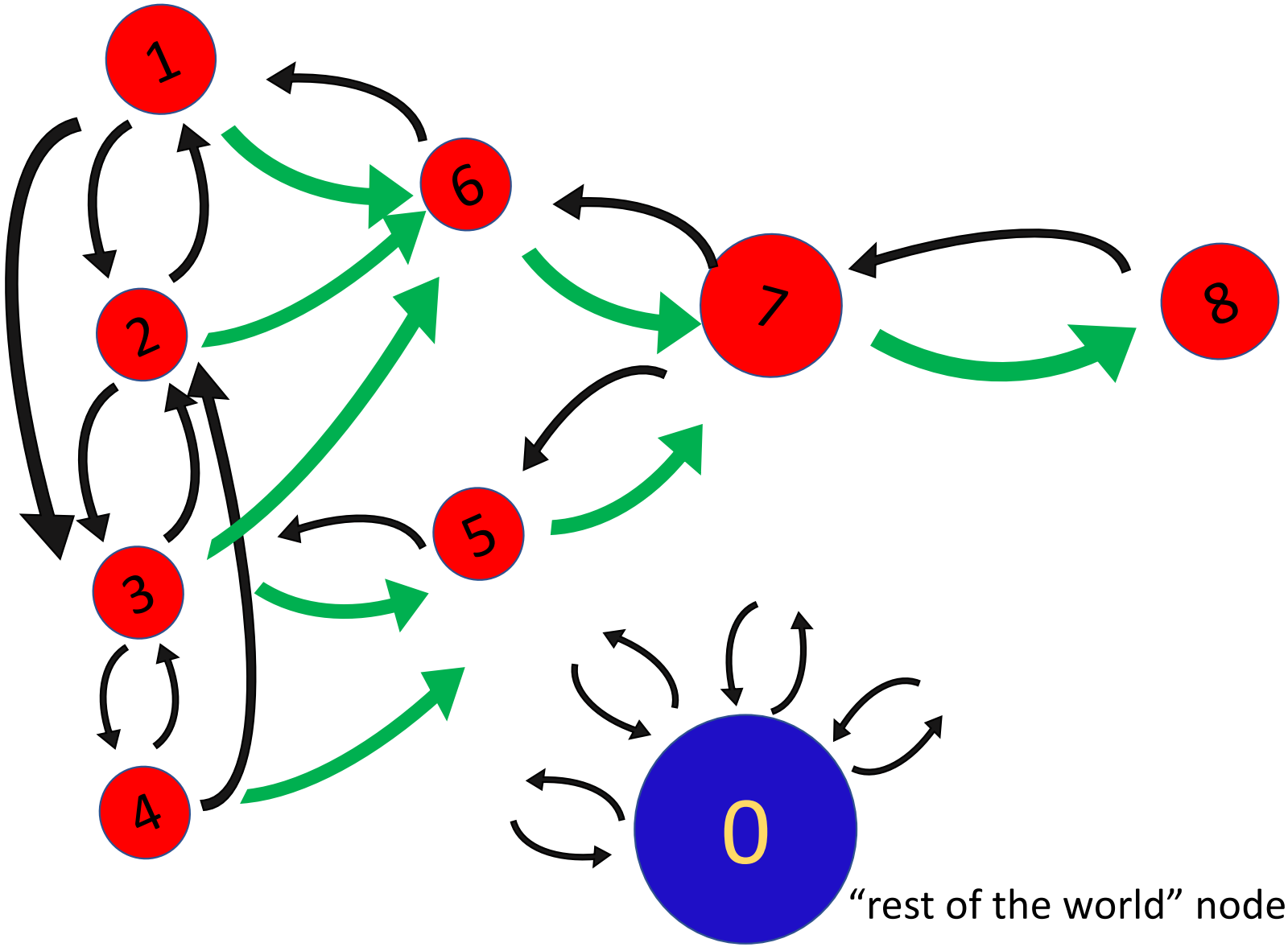


Figure: Non-zero Entries of the Residual Correlation Matrix (2007 - 2009) after taking out Fama-French factors. Based on 15 minutes data (n = 572).

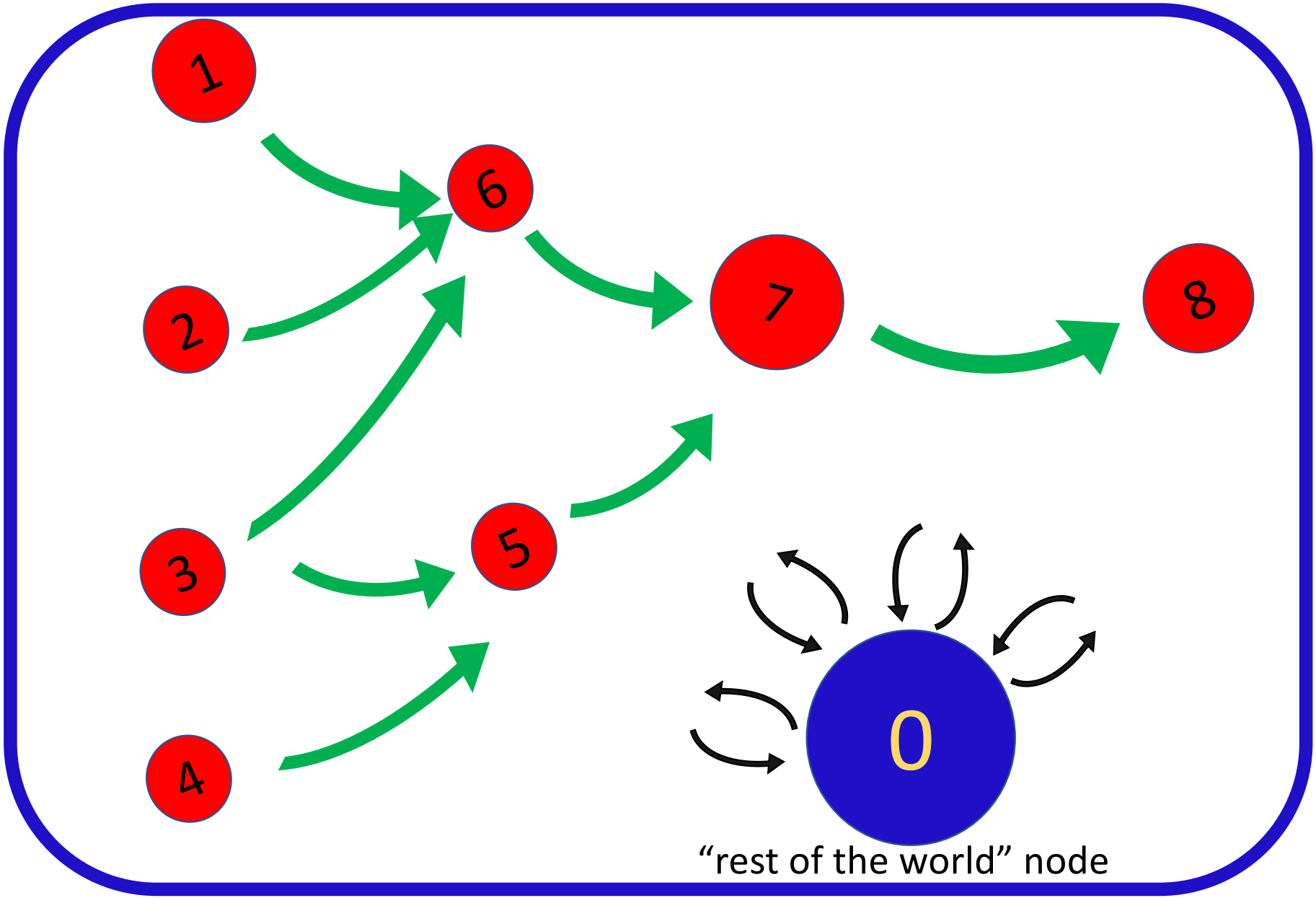
Source: *Incorporating GISC and High Frequency Data into Portfolio Allocation and Risk Estimation*, Jianqing Fan, Alex Furger, Dacheng Xiu

Anomaly detection - Transactions in a financial network (i)

Tracking the flow of money (eg, for the purpose of anomaly detection)

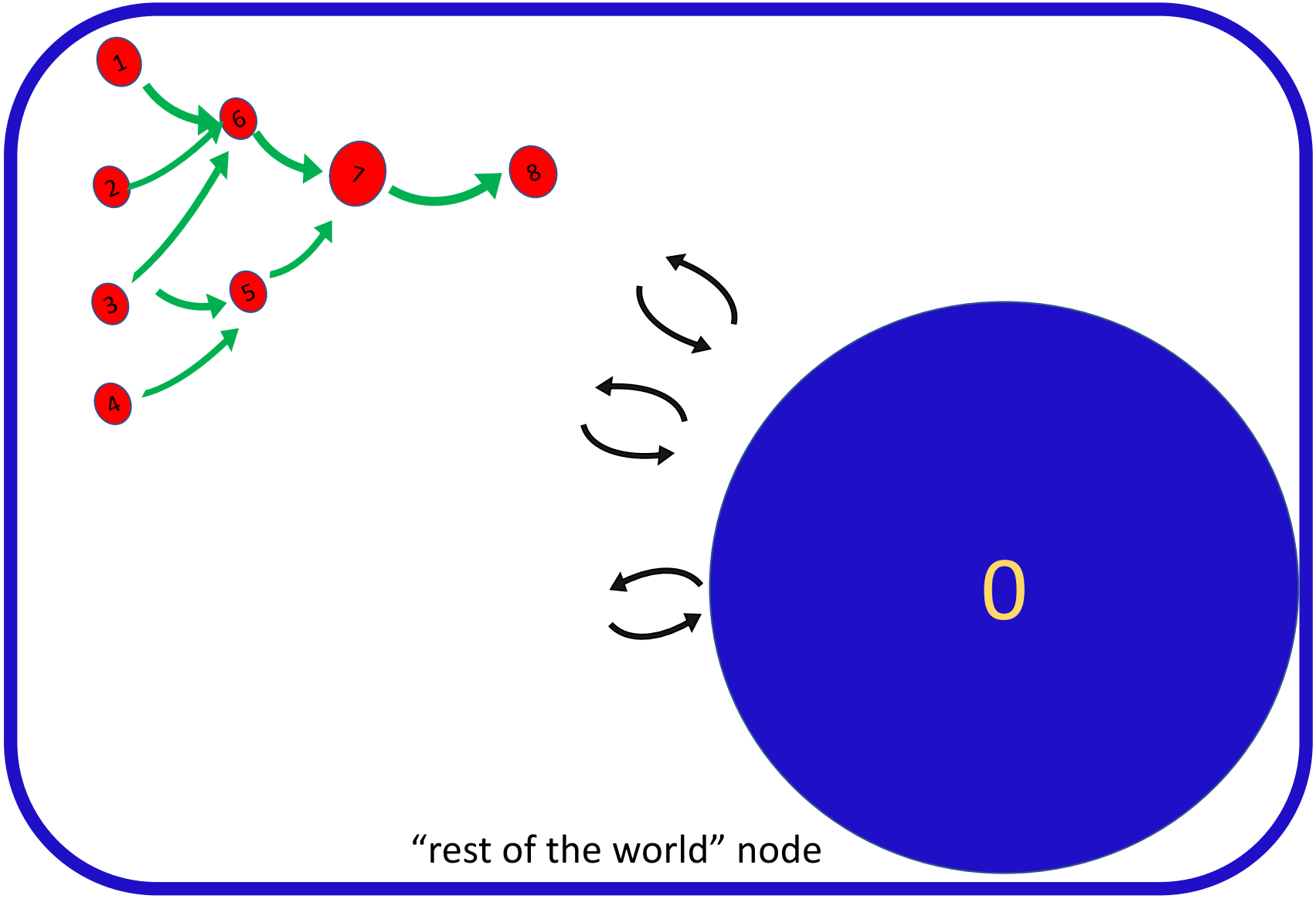


Anomaly detection - Transactions in a financial network (ii)



Anomaly detection - Transactions in a financial network (iii)

Example of a planted signal, in a (typically much larger) ambient graph.



The Group Synchronization Problem (Euc(d))

Recover group elements from a sparse noisy set of pairwise ratios.

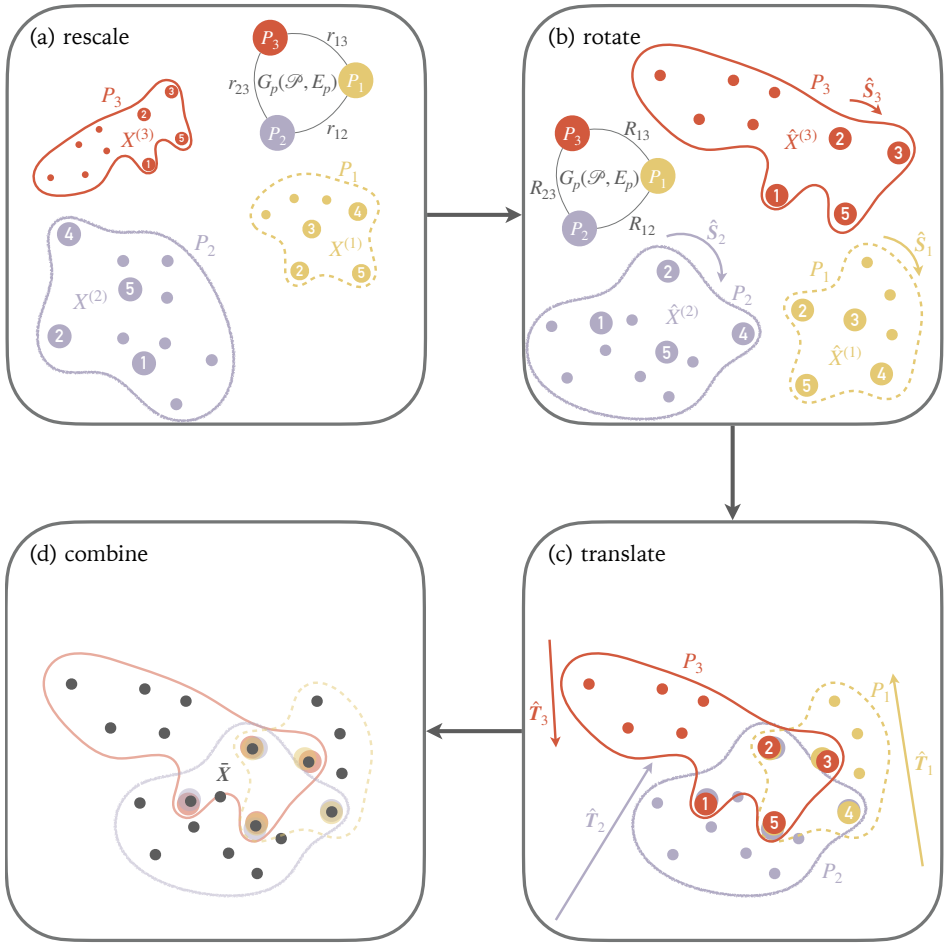


Figure: Schematic overview of the main steps of the **local2global** algorithm. (a) Synchronization over scales. (b) Synchronization over orthogonal transformations. (c) Synchronization over translations. (d) Global node embedding as centroid of aligned patch embeddings. (b-c-d) is synchronization over Euc(d) $\cong \mathbb{Z}_2 \times SO(d) \times \mathbb{R}^d$.

Large scale graph embedding (MAG240m) - divide & conquer

https://ogb.stanford.edu/kddcup2021/mag240m/

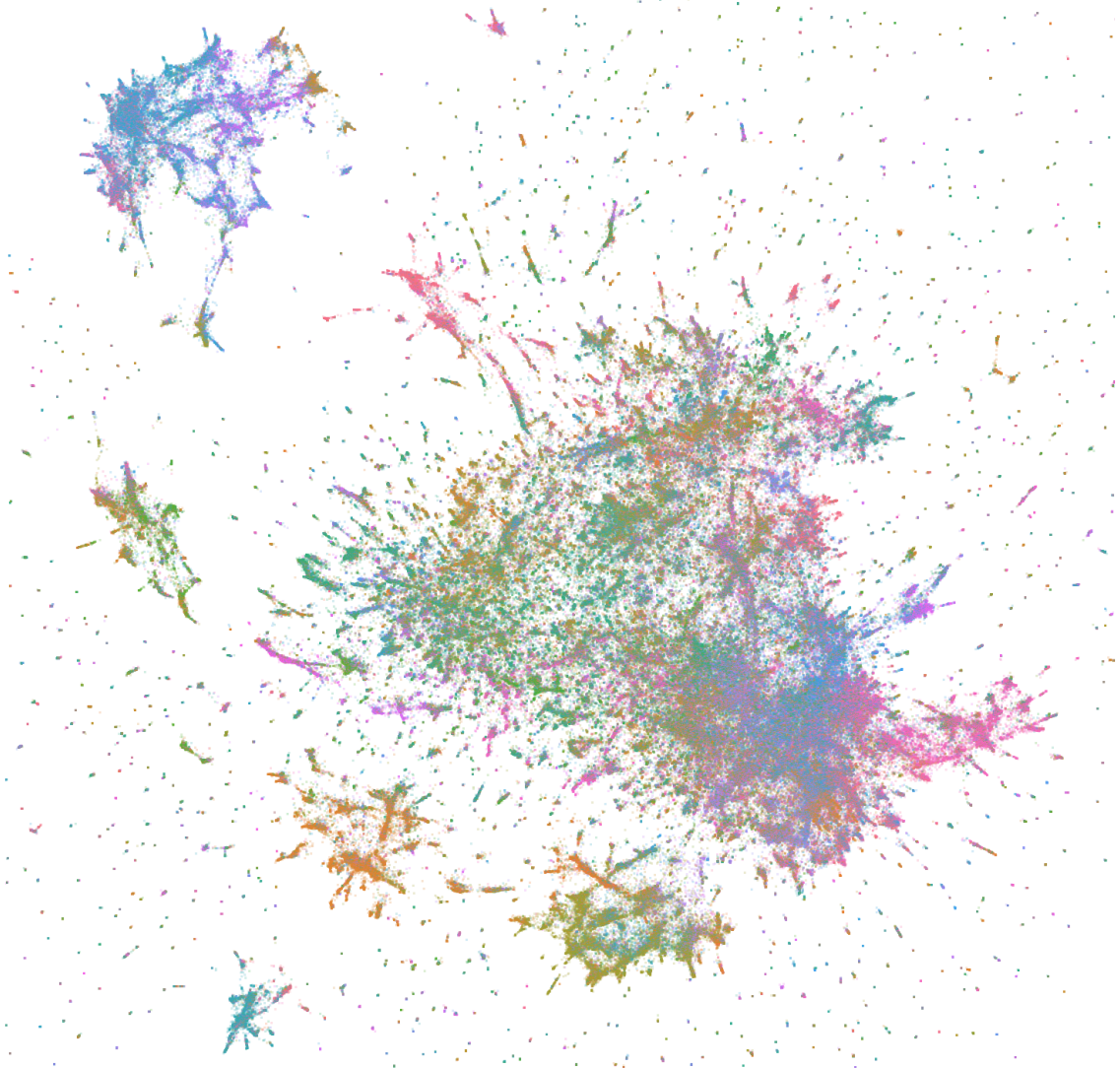


Figure: UMAP projection of 128-dim VGAE-**local2global** embedding; heterogeneous academic graph extracted from the Microsoft Academic Graph; nodes are coloured by class label/topics. Visualisation based on a sample of 500,000 labeled papers.

Lead-lag detection in multivariate time series

Given a basket of stocks, identify a subset X of stocks that lead another subset Y.

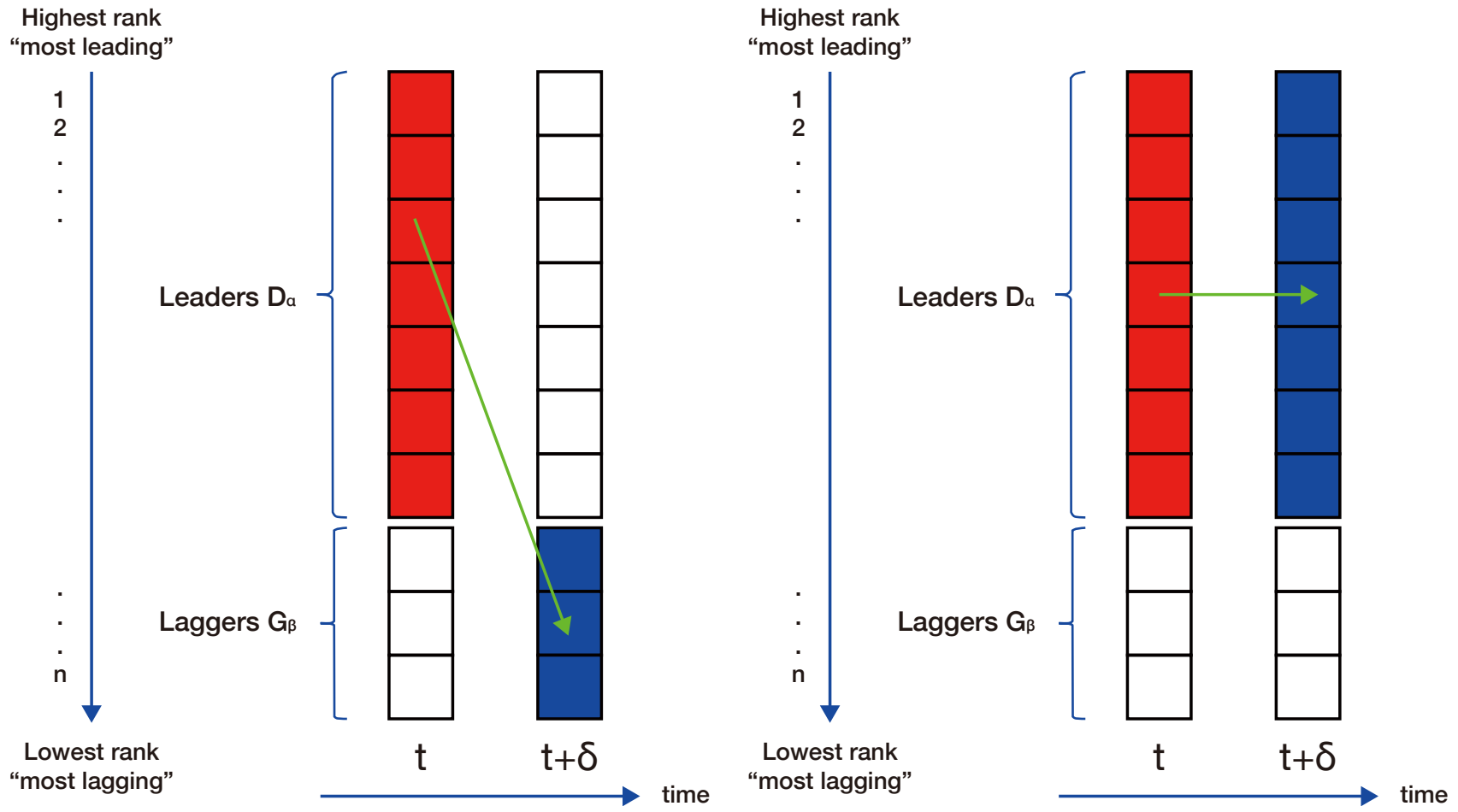


Figure: Left: Trading strategy where we expect the average price return of laggers tomorrow to move in the same direction as the average return of the leaders today. Right: strategy where we expect leaders to exhibit momentum.

Lead-lag detection in multivariate time series

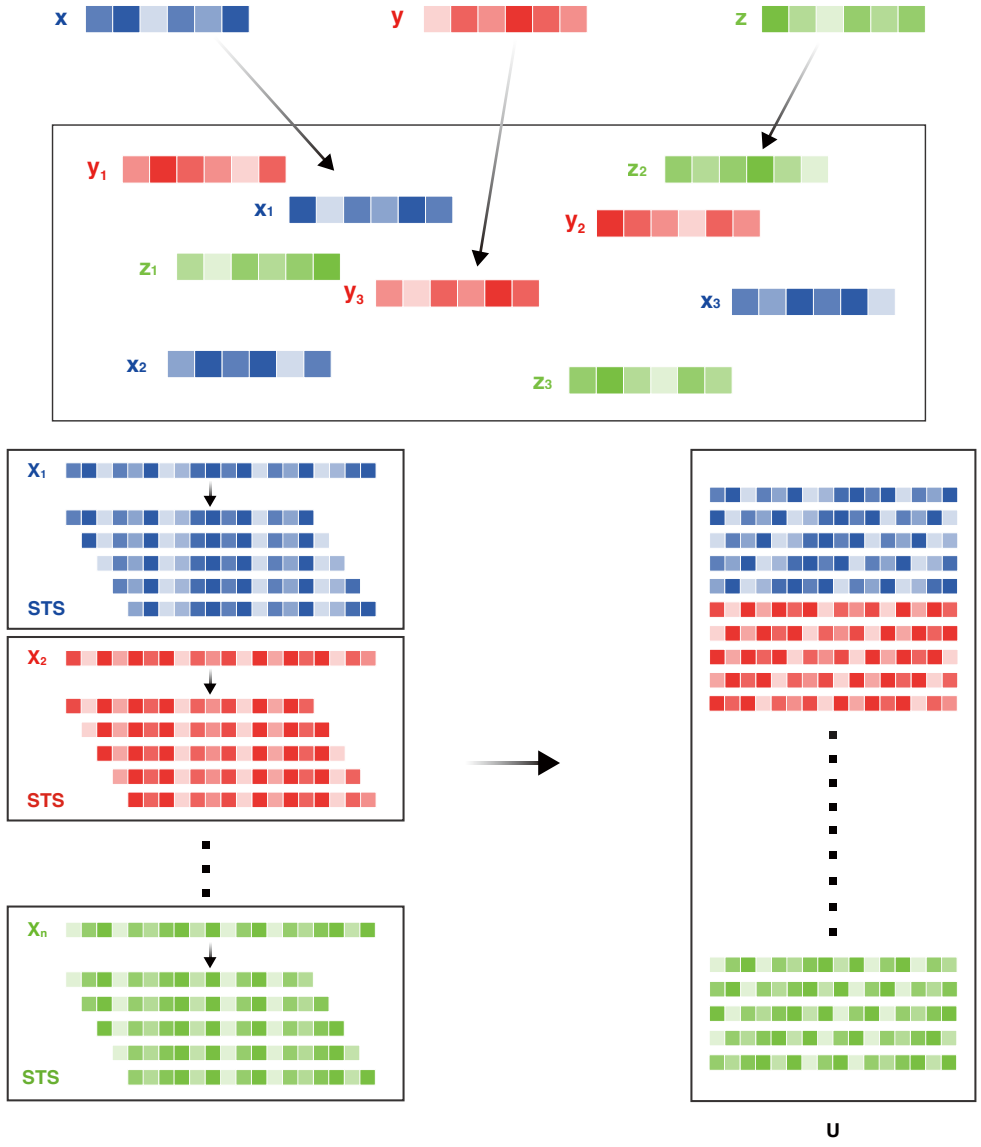


Figure: Top: each measurement is a noisy shifted version of one of the unknown three signals (x,y or z, but the mapping is not known). Bottom: extraction of sub-time series.

Trade co-occurrence analysis (fast search for nearest neighbors)

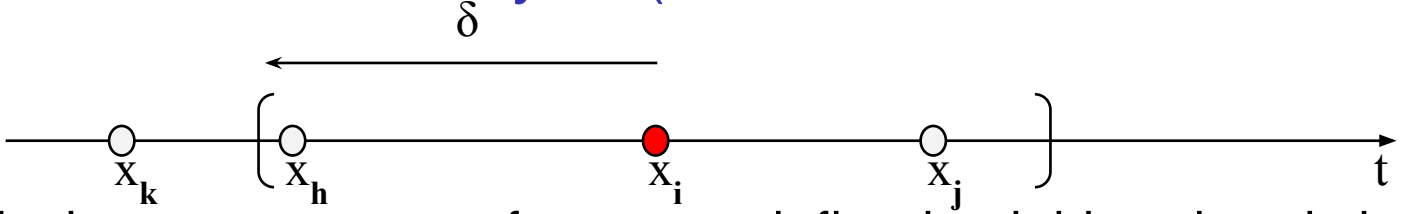


Figure: Trade co-occurrence: for a user-defined neighbourhood size δ, trade x_j arrives within the δ-neighbourhood of trade x_i, and thus they co-occur. In contrast, trade x_k locates outside x_i's neighbourhood, and thus the two trades do not co-occur. Both trades x_j and x_h co-occur with trade x_i, but they do not co-occur with each other.

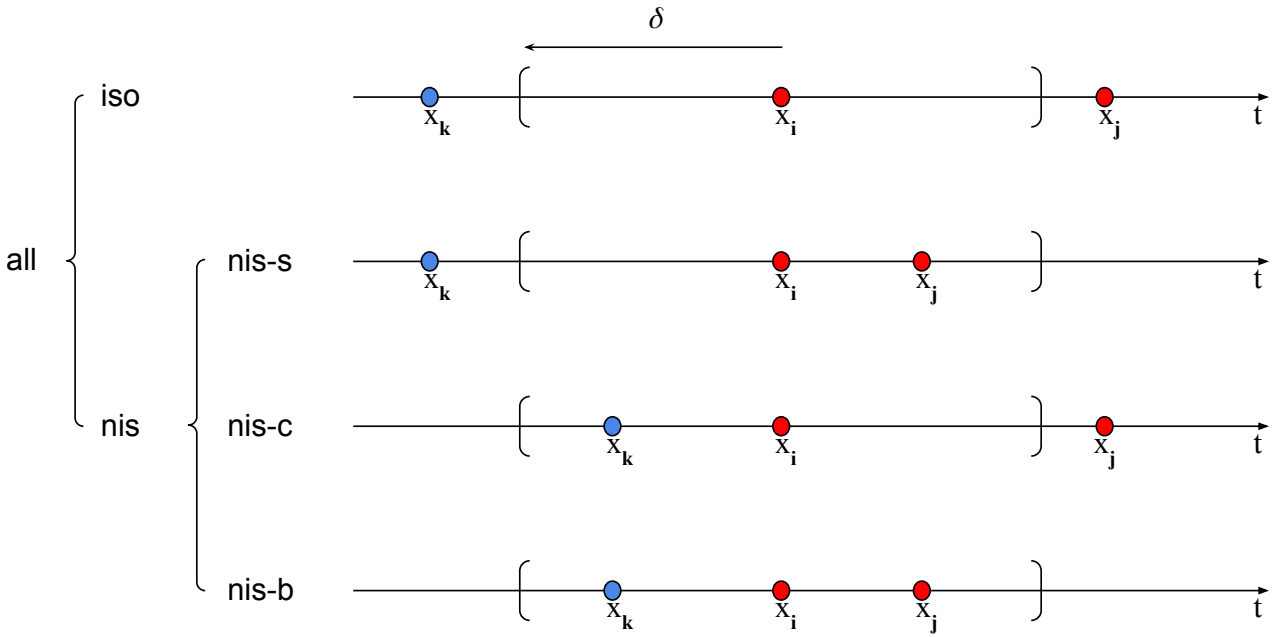
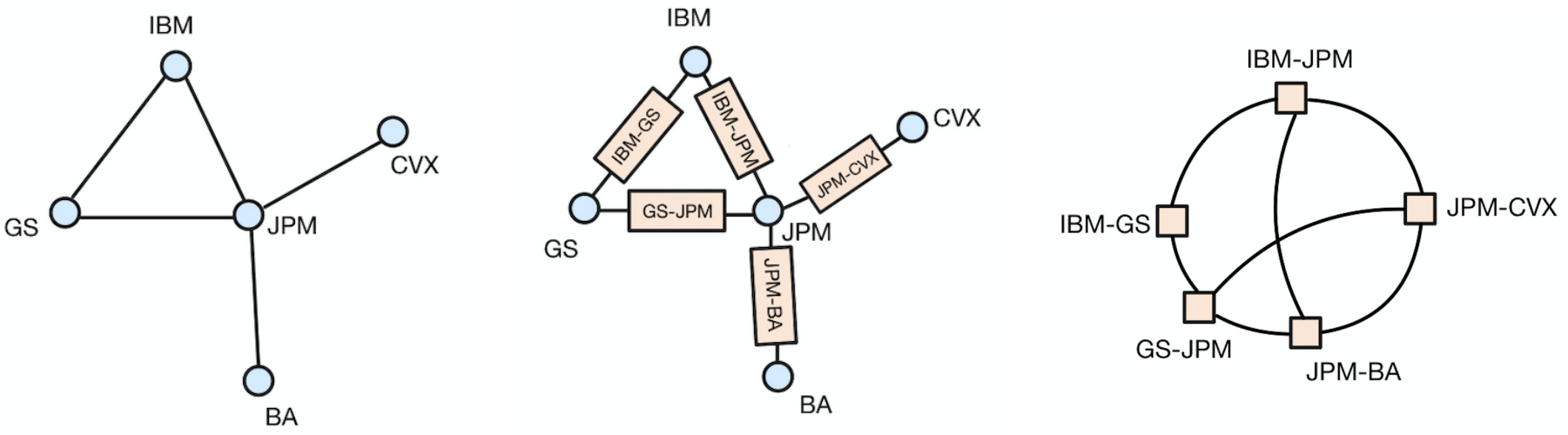


Figure: Illustration of trade types, conditioning on co-occurrenc (distinct categorical labels of trade x_i). Color indicates the stock corresponding to a trade. Thus, x_j is for the same stock as x_i, while x_k is for a different stock.

Line-graphs for forecasting realized covariance matrices.

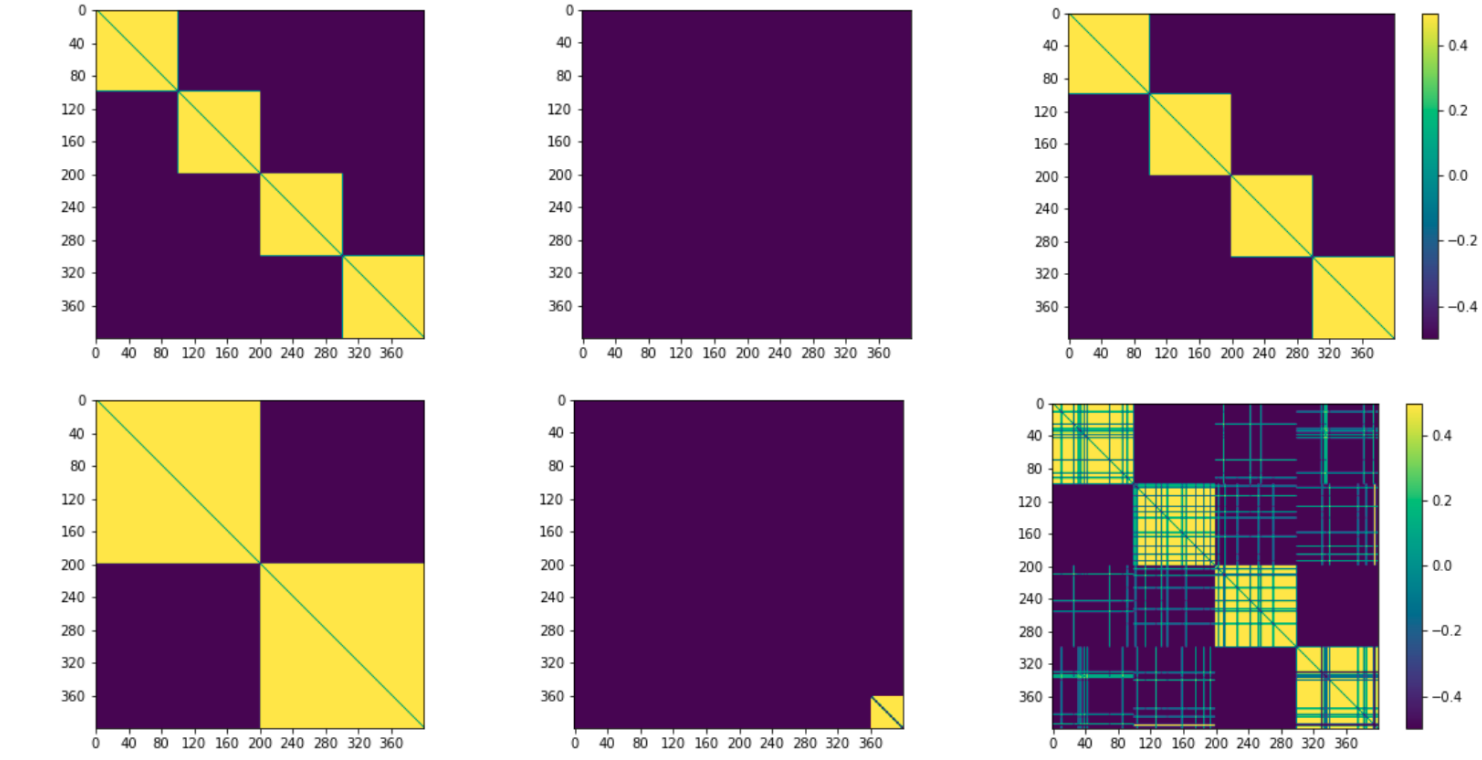


(a) Graph $\mathcal{G}$ for volatility (b) Labeled edges in $\mathcal{G}$ (c) Line graph $L(\mathcal{G})$ for correlation

Figure: Diagram of the process of building the line graph $L(\mathcal{G})$ for $N = 5$ assets.

- An edge in $\mathcal{G}$ constitutes a node in the line graph $L(\mathcal{G})$.
- The edges in the line graph capture the interdependence between two correlation pairs that have a asset in common.

Change-point detection in network time series



(a) "Merge" scenario. (b) "Birth" scenarios. (c) "Swap" scenario.

Figure: Heatmaps of the adjacency matrices of the expected graph in two stochastic block models $\mathcal{G}_1$ (first row) and $\mathcal{G}_2$ (second row), with $n = 400$ nodes, in three different scenarios of single change-point synthetic experiments:"Merge" (a), "Birth" (b) and "Swaps" (c). $\mathcal{G}_1$ and $\mathcal{G}_2$ correspond to the generative distributions of the snapshots before and after the change-point.

Summary

Many of the methods we will study are of **spectral** nature, which brings along a number of benefits:

- computational scalability
- robust to high level of noise in the data (low-SNR regime)
- theoretical signal recovery guarantees under suitably defined stochastic (block) models

Clustering, Ranking, Dimensionality Reduction:

- provide insights into the structure of various data sets
- more importantly, cluster/ranking information could be leveraged for some downstream task of interest (eg., prediction)
- unsupervised learning algorithms can be construed as a "means to an end"; in most pipelines, the ultimate task boils down to prediction or classification
- prediction & classification provide an opportunity to compare performance/utility of unsupervised learning algorithms on data sets for which no ground truth exists.

Yogi Berra: It's tough to make predictions, especially about the future