Projet 3A - Extracting Insights From NBA Data Using Topology

Gabriel Azevedo Ferreira
Gustavo Castro
Henrique Gasparini Fiuza do Nascimento

March 16, 2017

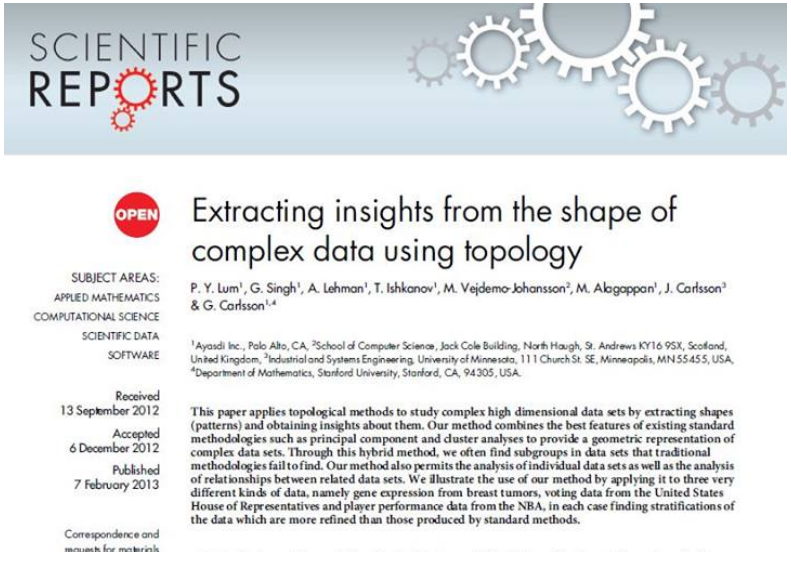
École Polytechnique

Table of contents

Introduction

Contextualizing

- Paper published by M. Alagappan in 2013



SCIENTIFIC REPORTS

OPEN

Extracting insights from the shape of complex data using topology

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

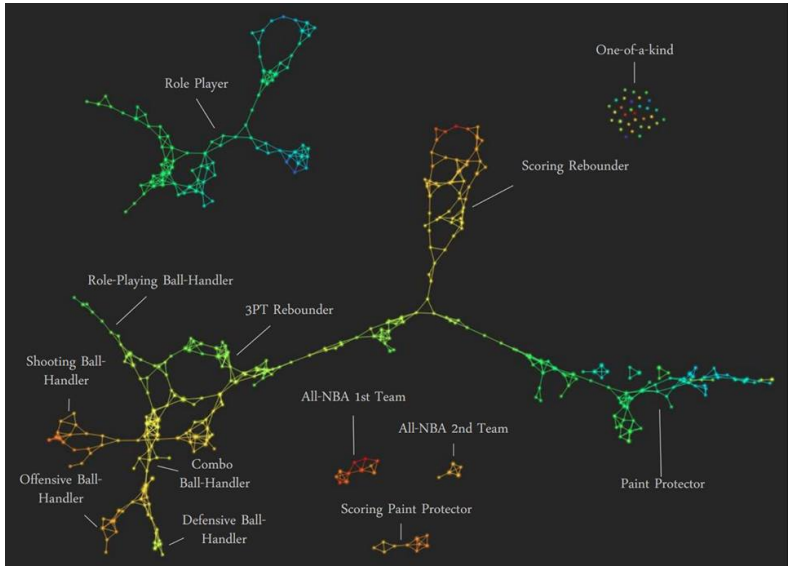
Received
13 September 2012

Accepted
6 December 2012

Published
7 February 2013

Correspondence and

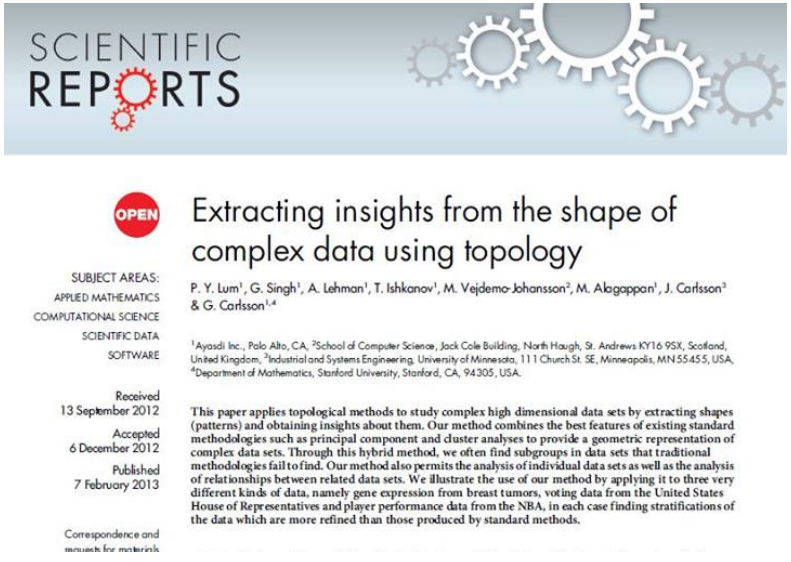
This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of complex data sets. Through this hybrid method, we often find subgroups in data sets that traditional methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



[M. Alagappan, 2012]

Contextualizing

- Paper published by M. Alagappan in 2013
  - Method Mapper applied to NBA players of 2010-2011 season



SCIENTIFIC REPORTS

OPEN

Extracting insights from the shape of complex data using topology

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

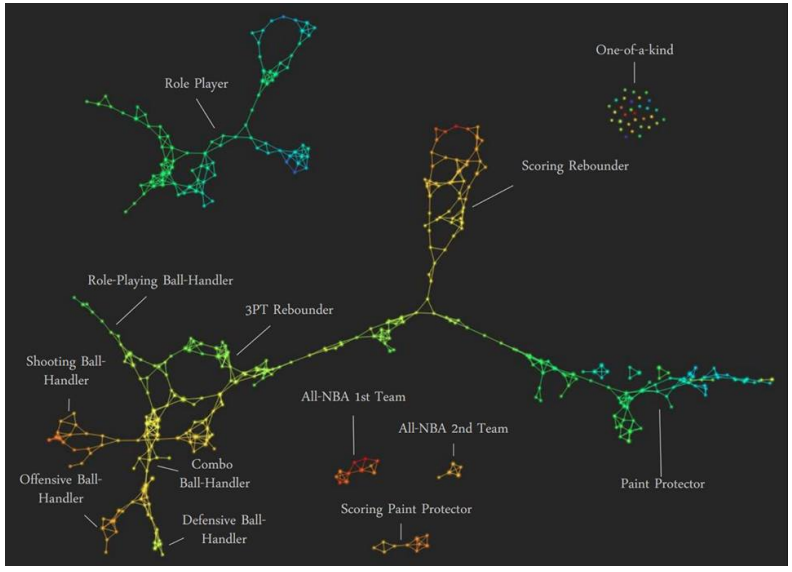
Received
13 September 2012
Accepted
6 December 2012
Published
7 February 2013

Correspondence and

P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

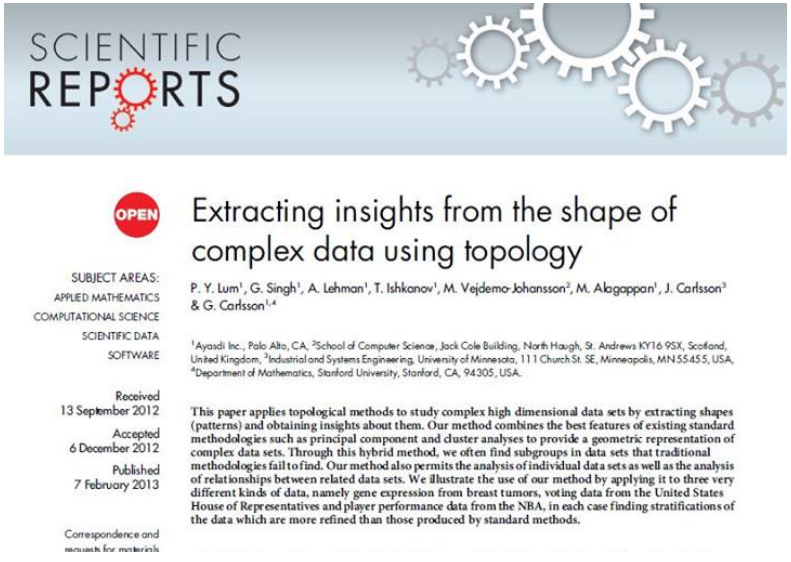
This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of complex data sets. Through this hybrid method, we often find subgroups in data sets that traditional methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



[M. Alagappan, 2012]

Contextualizing

- Paper published by M. Alagappan in 2013
  - Method Mapper applied to NBA players of 2010-2011 season
  - Lead to a different classification of playing styles



SCIENTIFIC REPORTS

OPEN

Extracting insights from the shape of complex data using topology

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

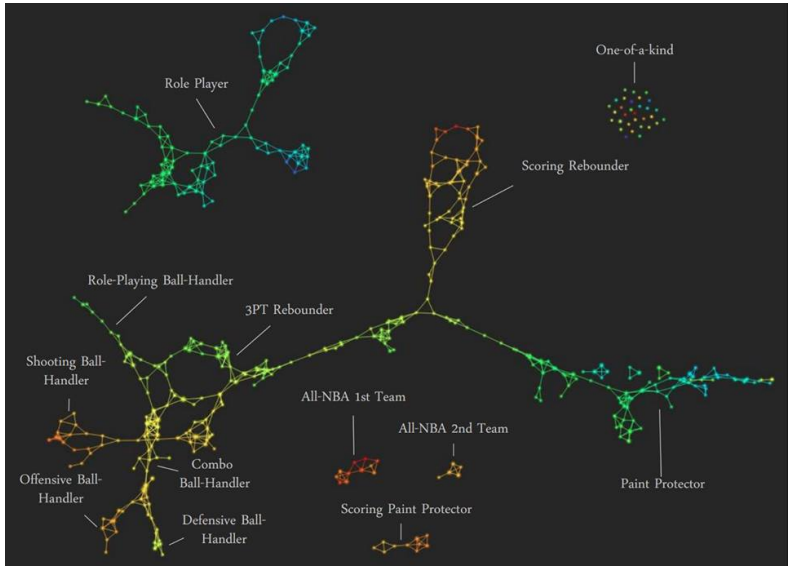
Received
13 September 2012
Accepted
6 December 2012
Published
7 February 2013

Correspondence and

P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

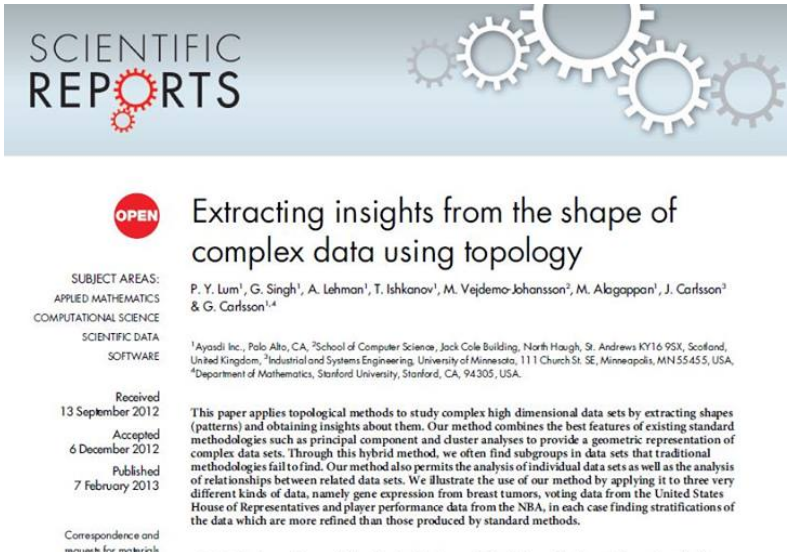
This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of complex data sets. Through this hybrid method, we often find subgroups in data sets that traditional methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



[M. Alagappan, 2012]

Our goals

- Analyse data using classical methods of clustering



SCIENTIFIC REPORTS

OPEN

Extracting insights from the shape of complex data using topology

P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

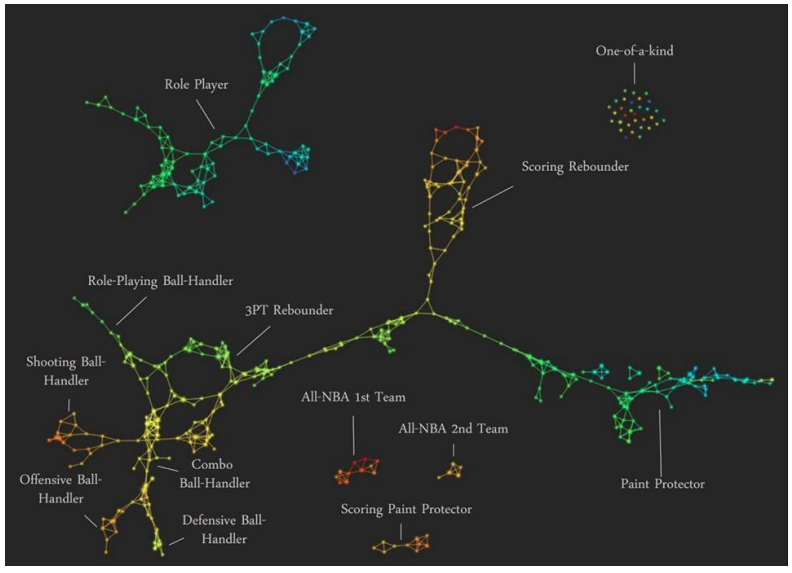
Received
13 September 2012

Accepted
6 December 2012

Published
7 February 2013

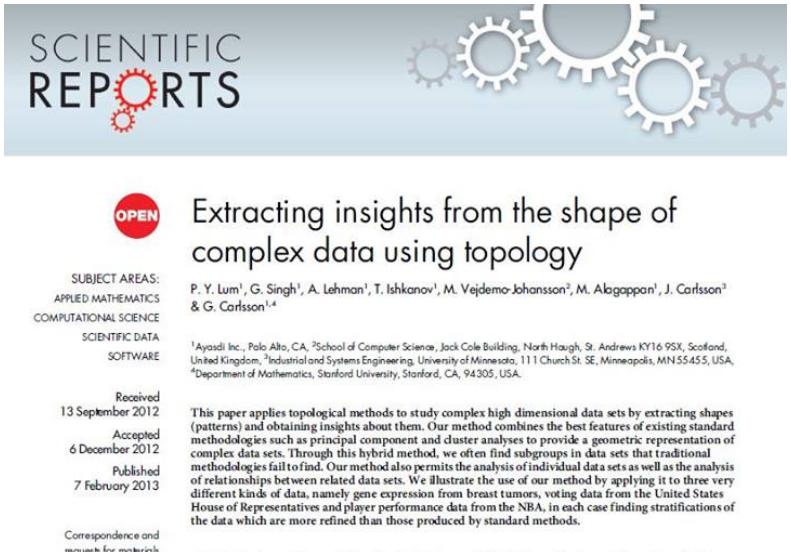
This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of complex data sets. Through this hybrid method, we often find subgroups in data sets that traditional methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.

Correspondence and



Our goals

- Analyse data using classical methods of clustering
- Use Mapper algorithm to NBA's last season



SCIENTIFIC REPORTS

OPEN

Extracting insights from the shape of complex data using topology

P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

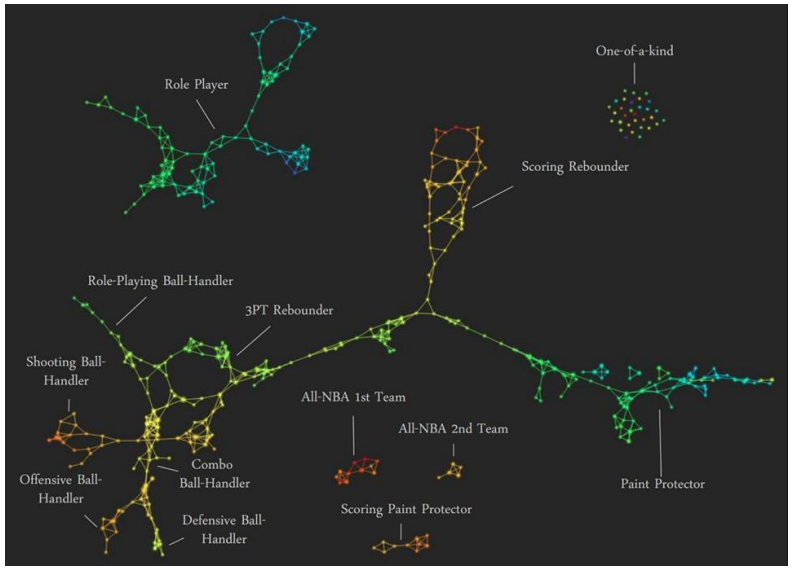
[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

Received
13 September 2012
Accepted
6 December 2012
Published
7 February 2013

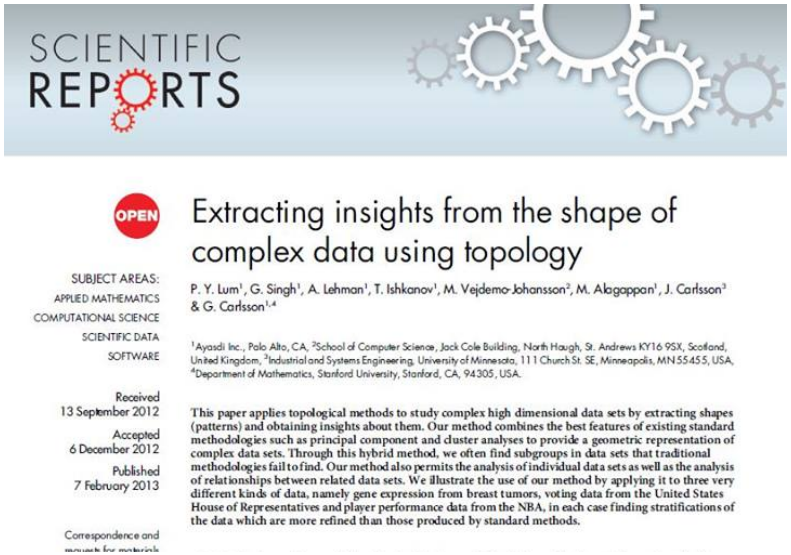
This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of complex data sets. Through this hybrid method, we often find subgroups in data sets that traditional methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.

Correspondence and



Our goals

- Analyse data using classical methods of clustering
- Use Mapper algorithm to NBA's last season
- Perform a similar analysis that Alagapan did



SCIENTIFIC REPORTS

OPEN

Extracting insights from the shape of complex data using topology

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

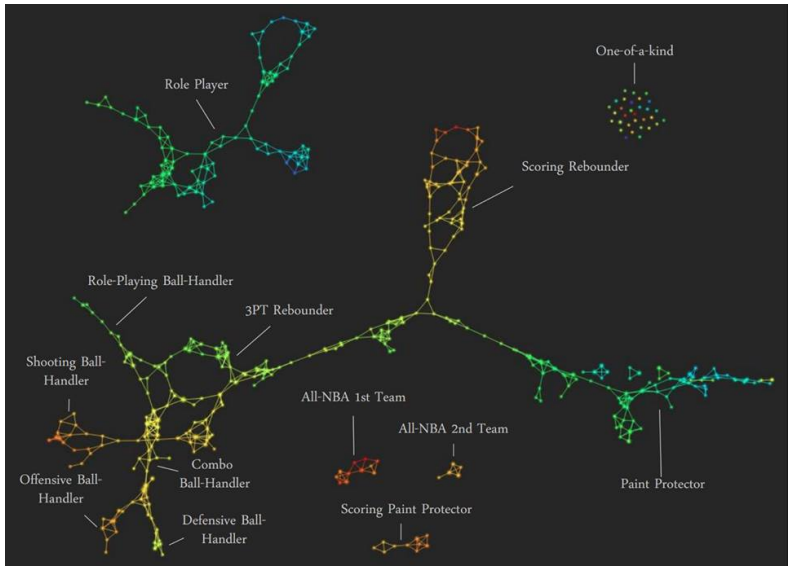
Received
13 September 2012
Accepted
6 December 2012
Published
7 February 2013

Correspondence and

P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

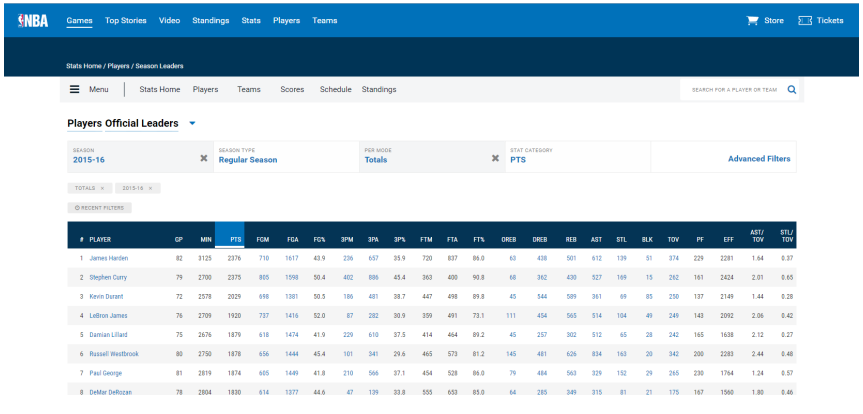
This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of complex data sets. Through this hybrid method, we often find subgroups in data sets that traditional methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



Extracting the data

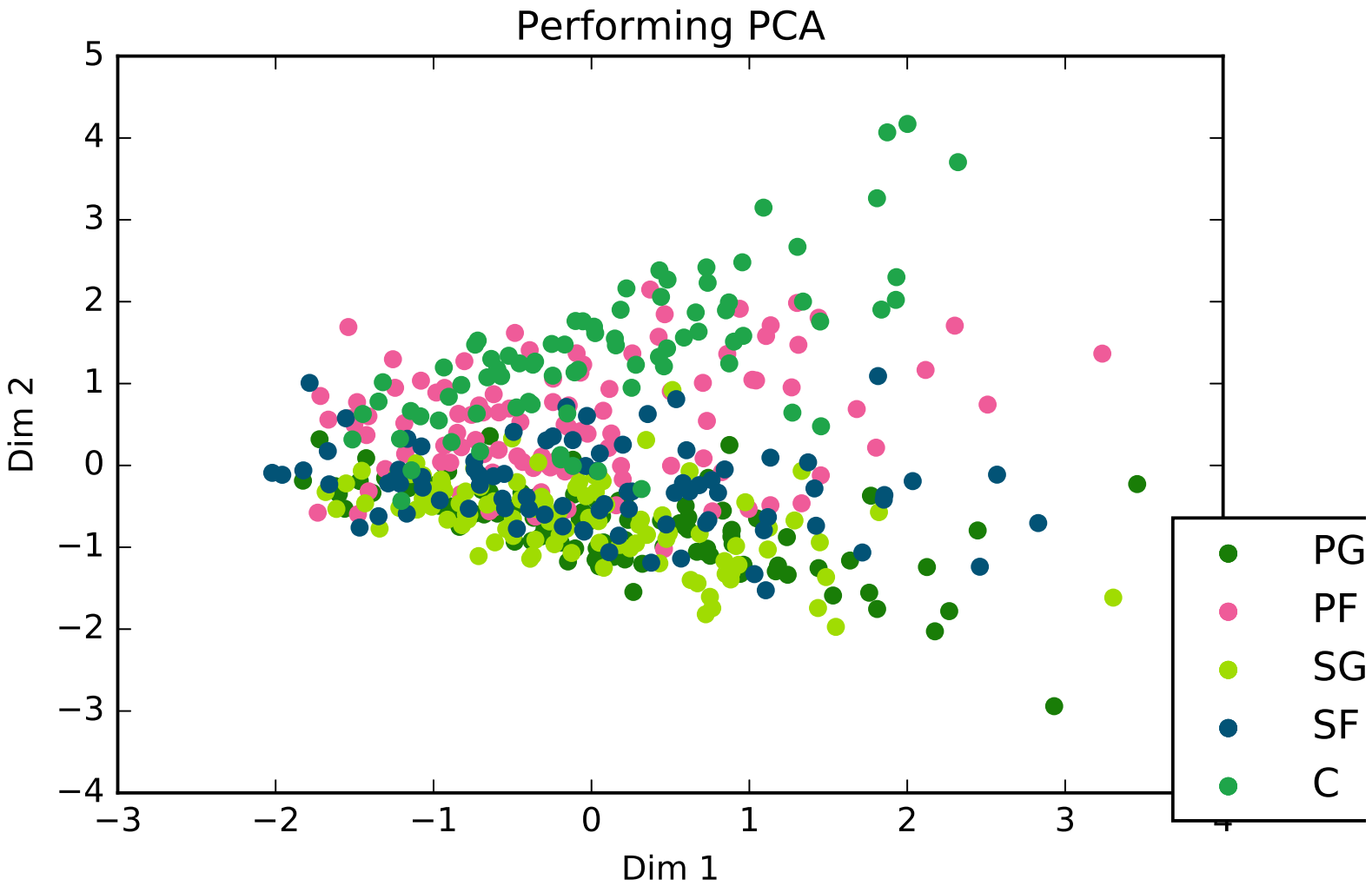
Extracting the data

- NBA official website
  - Data per Game (points per game, Blocks per game...)
  - 2015-2016 season
- ESPN website (obtain player's positions)

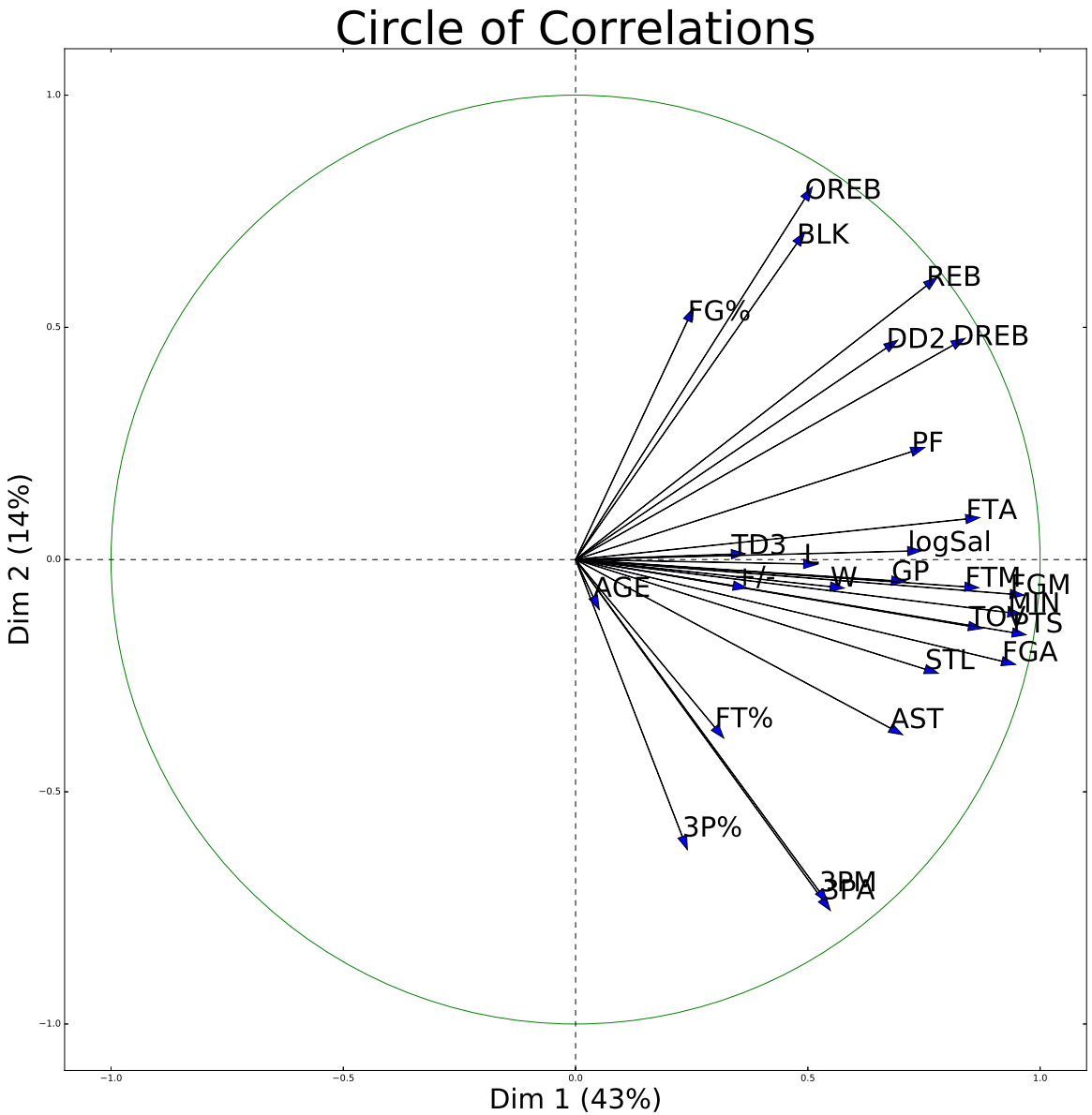


Visualizing the data via dimensionality reduction

Complete Dataset - PCA

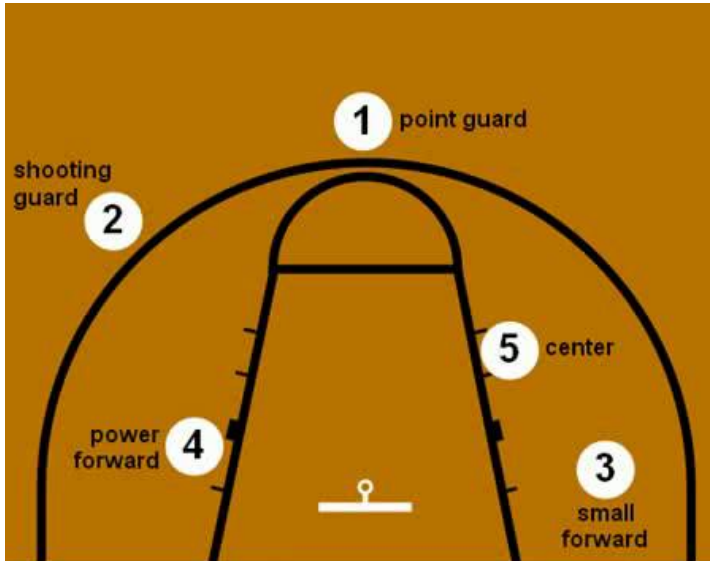


Complete Dataset - Correlation Graph



Complete Dataset

- 1^{st} Dimension: Player ability
  - Salary (log. scale)
  - Points
- 2^{nd} Dimension: Position on the Field
  - (offensive) rebounds,
  - Blocks,
  - 3 point field made/attempted



Feature selection and clustering analysis

Unsupervised methods for feature selection

- Laplacian score

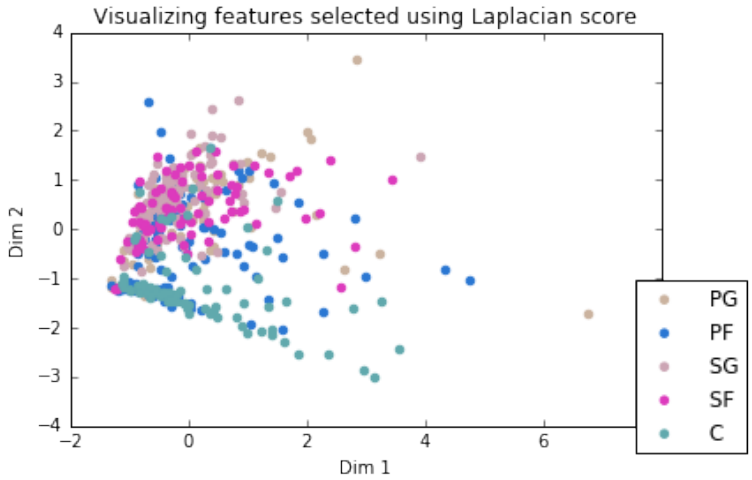
Unsupervised methods for feature selection

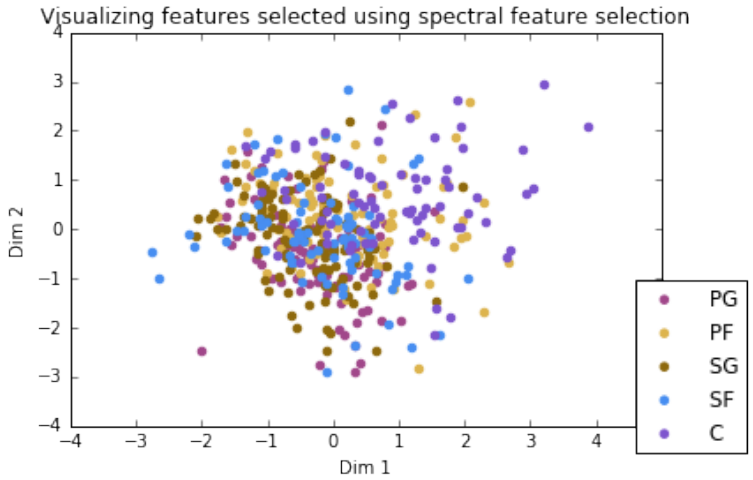
- Laplacian score
- Spectral Feature Selection

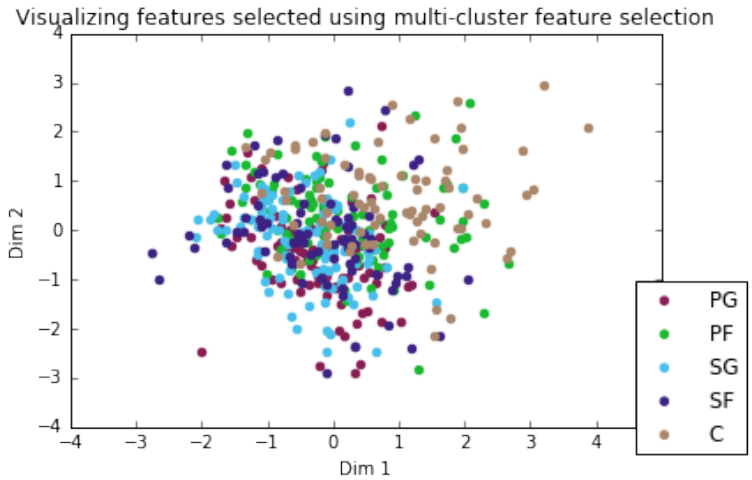
Unsupervised methods for feature selection

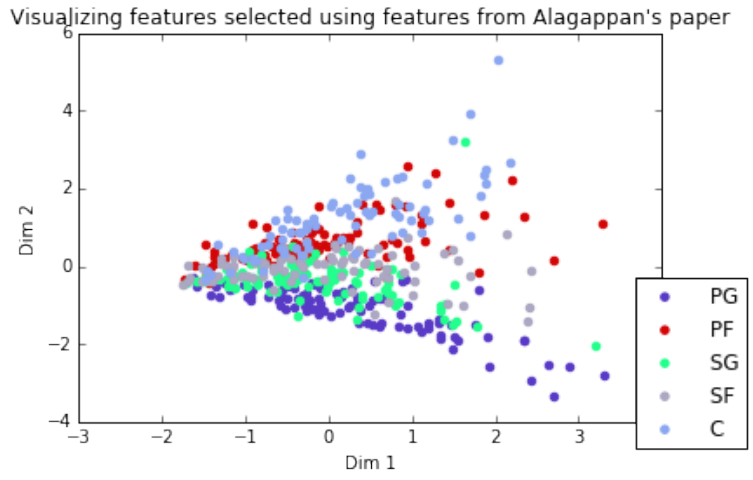
- Laplacian score
- Spectral Feature Selection
- Multi-Cluster Feature Selection

Visualizing the selected features









Evaluating a cluster

- Permutation-wise Accuracy

Evaluating a cluster

- Permutation-wise Accuracy
- Adjusted Rand's Index

Evaluating a cluster

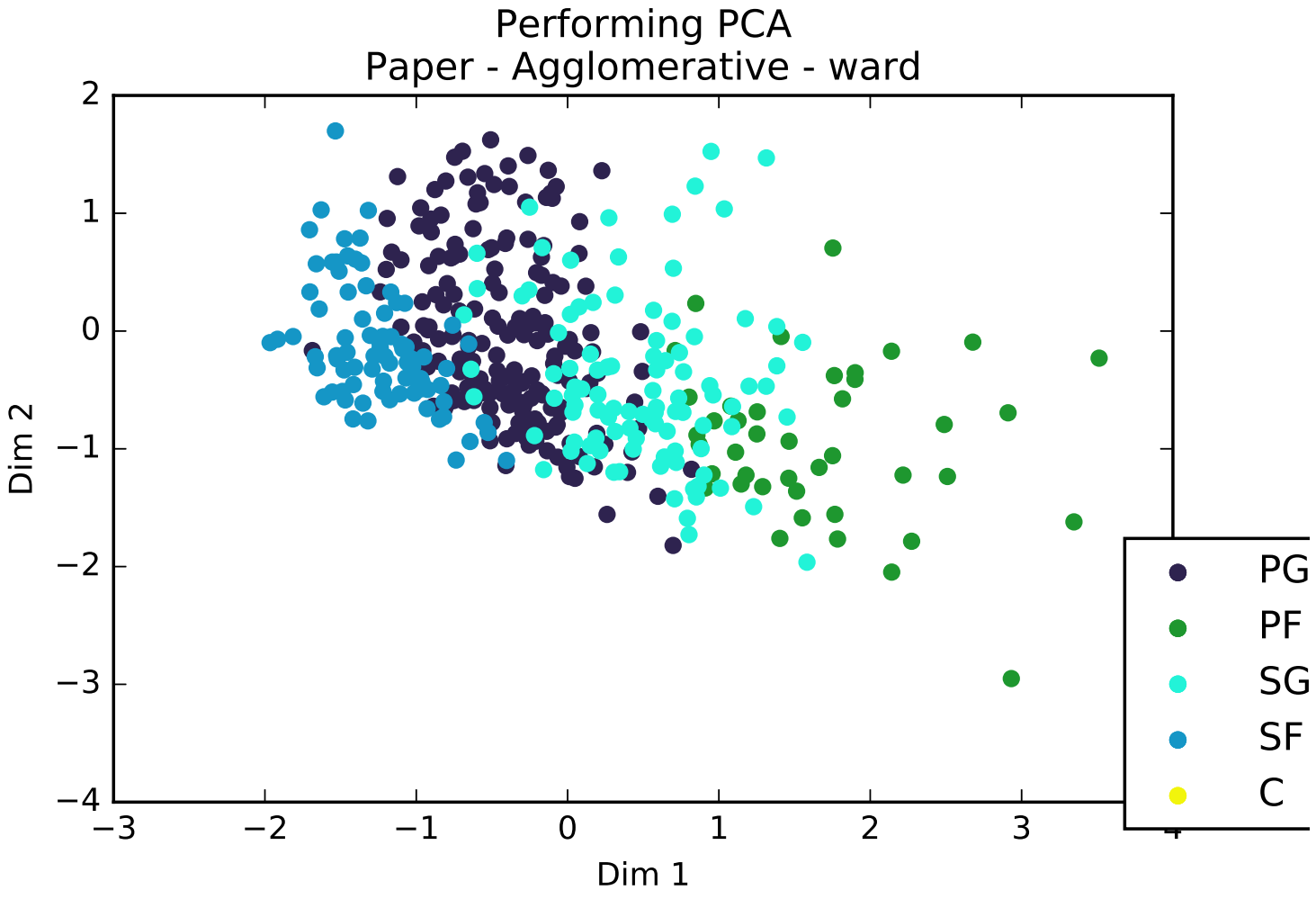
- Permutation-wise Accuracy
- Adjusted Rand's Index
- Silhouette score

Agglomerative Clustering

Table 1: Best performances for each evaluation metrics

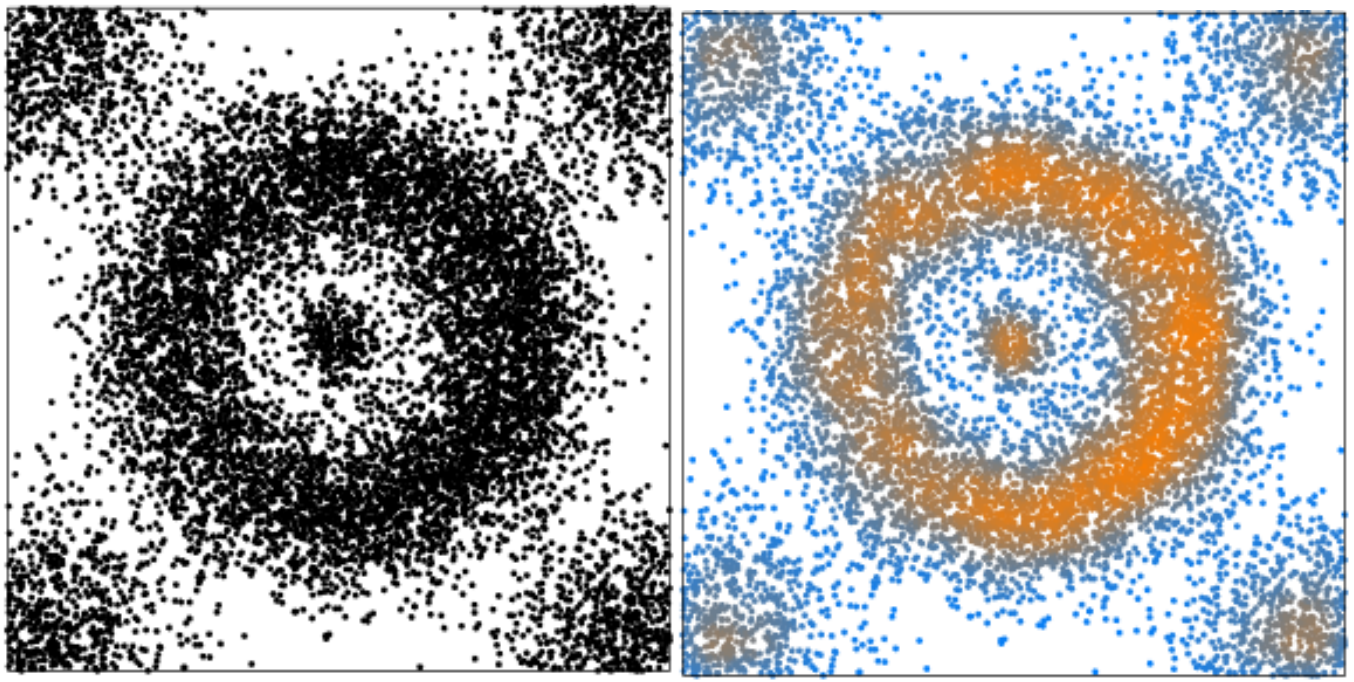
Metrics	Best score	Features	Linkage
Accuracy	37.8%	Alagappan's features	Ward
Rand's Index	0.093	Alagappan's features	Complete
Silhouette score	0.364	All the features	Average

Agglomerative Clustering



Mode-seeking

- Greater importance to peaks of density
- A Hill-Climbing scheme

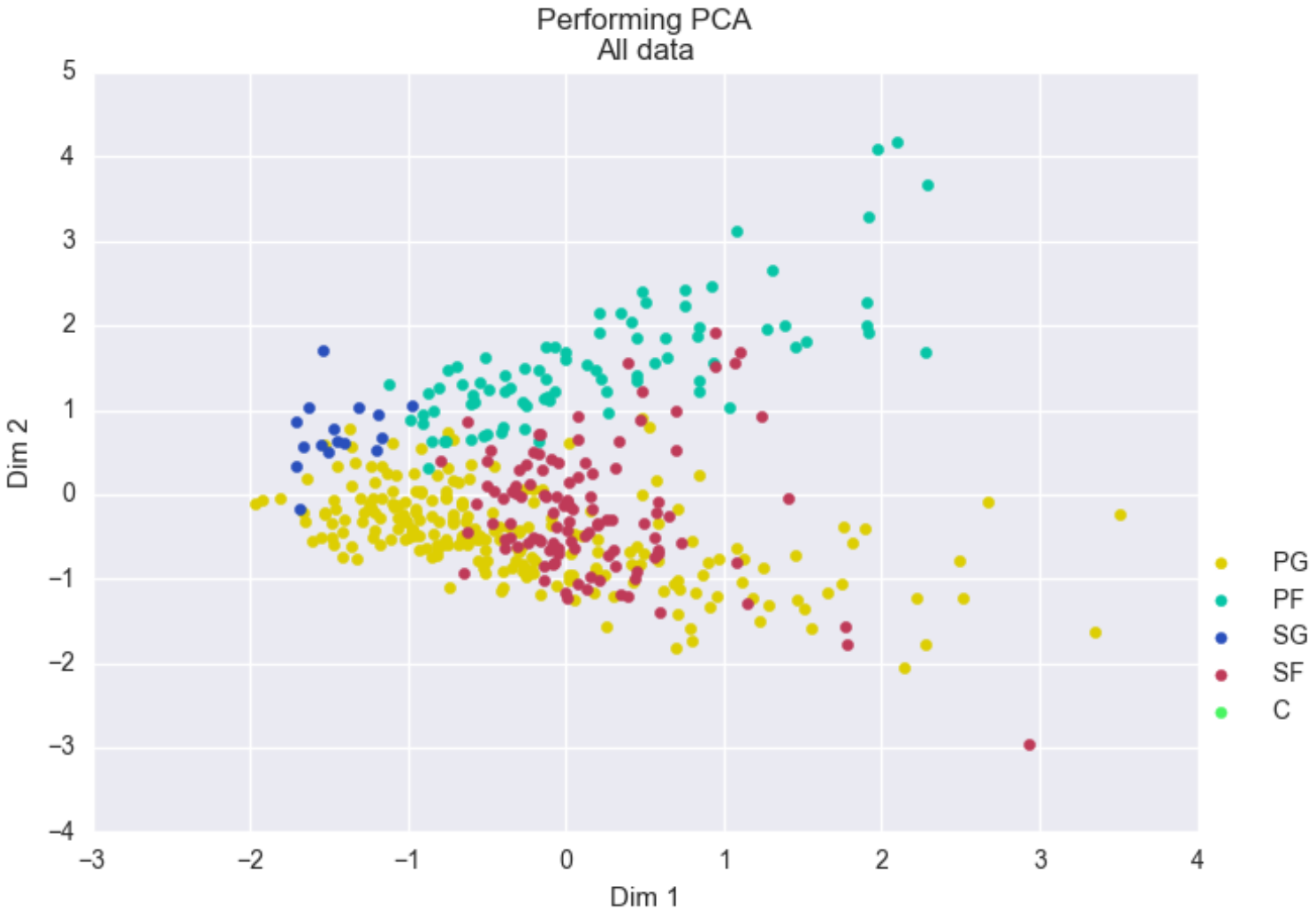


Mode-seeking

Table 2: Best performances for each evaluation metrics

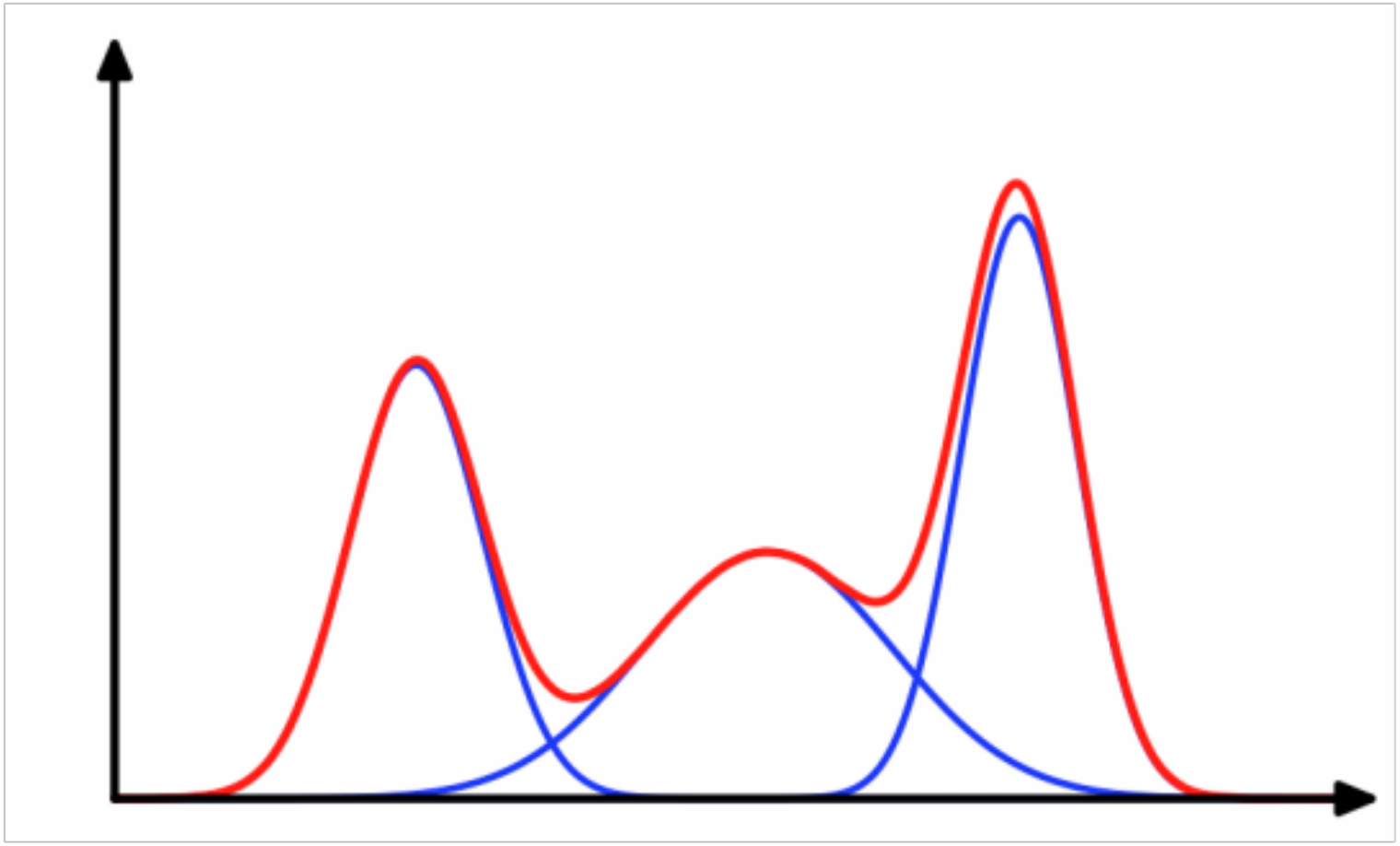
Metrics	Best score	Score range	Features	kDensity
Accuracy	34.9%	[0,100]	All the features	12
Rand's Index	0.137	[0,1]	All the features	16
Silhouette score	0.050	[-1,1]	Laplacian score	24

Mode-seeking



Gaussian Mixture Models

- Assumes data is sampled as a mixture of unknown Gaussian distributions



Gaussian Mixture Models

Table 3: Best performances for each evaluation metrics

Metrics	Best score	Features	Covariance type
Accuracy	39.67%	Alagappan's features	diagonal
Rand's Index	0.126	Alagappan's features	diagonal
Silhouette score	0.287	All the features	diagonal

Comparing with supervised algorithms

Table 4: Best performances on supervised classification

Classifier	Best accuracy	Features
Gaussian Process	55.37%	Alagappan's features
Decision Tree	51.24%	Alagappan's features
QDA	55.10%	All the features
Naive Bayes	43.80%	Alagappan's features
Linear SVM	48.98%	All the features
Neural Net	63.27%	All the features
RBF SVM	49.59%	Alagappan's features
Adaboost	54.55%	Alagappan's features
Random Forest	53.06%	All the features
Nearest Neighbors	52.07%	Alagappan's features

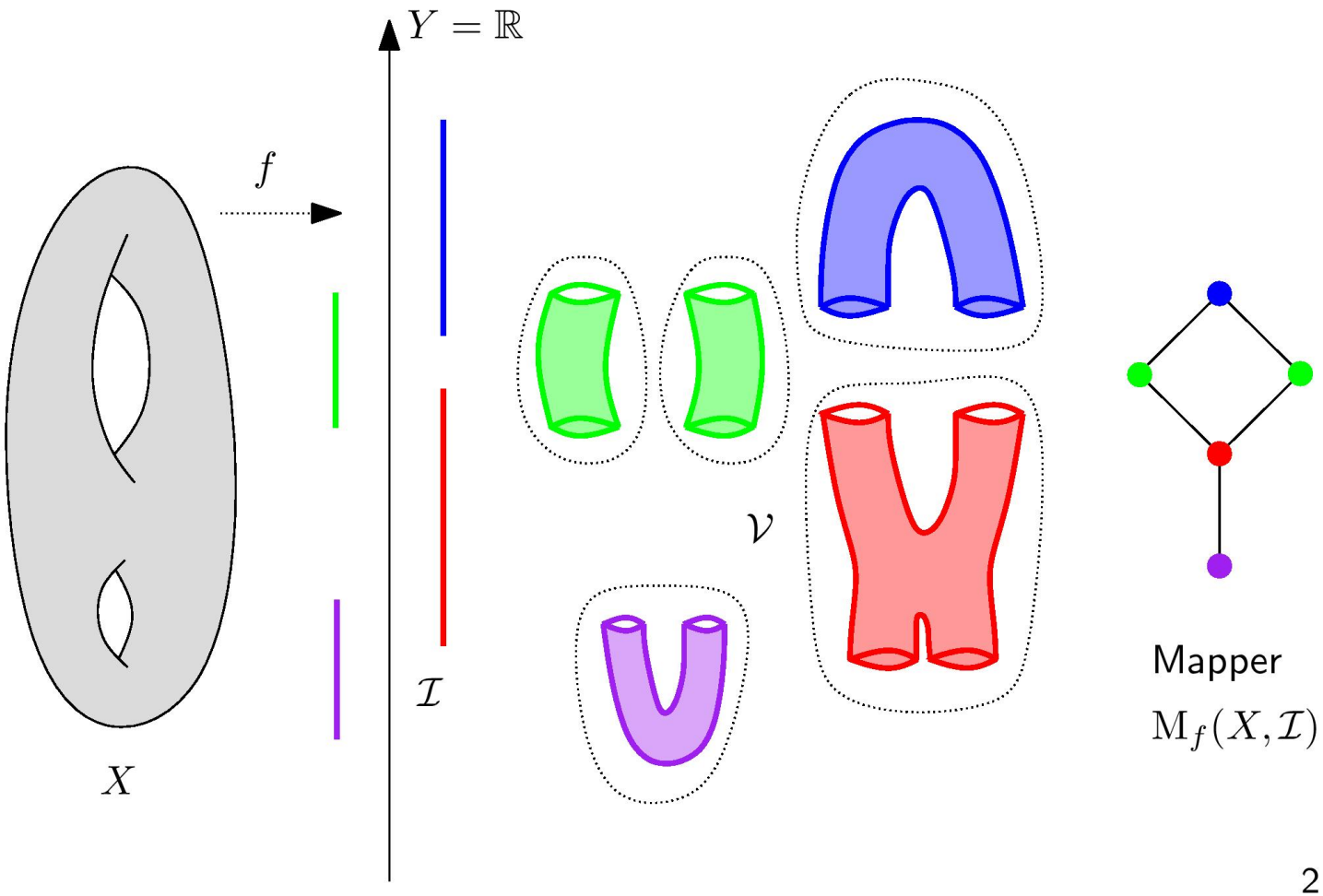
Analyzing other sets of field positions

Table 5: Best performances for each method

Method	Accuracy	Selected features	More
Common unsupervised	48.95%	Laplacian score	complete linkage
Mode-seeking	48.7%	All the features	ratio 3:1
Gaussian Mixture Models	67.5%	All the features	tied covariance
Common Supervised	79.59%	All the features	Neural net

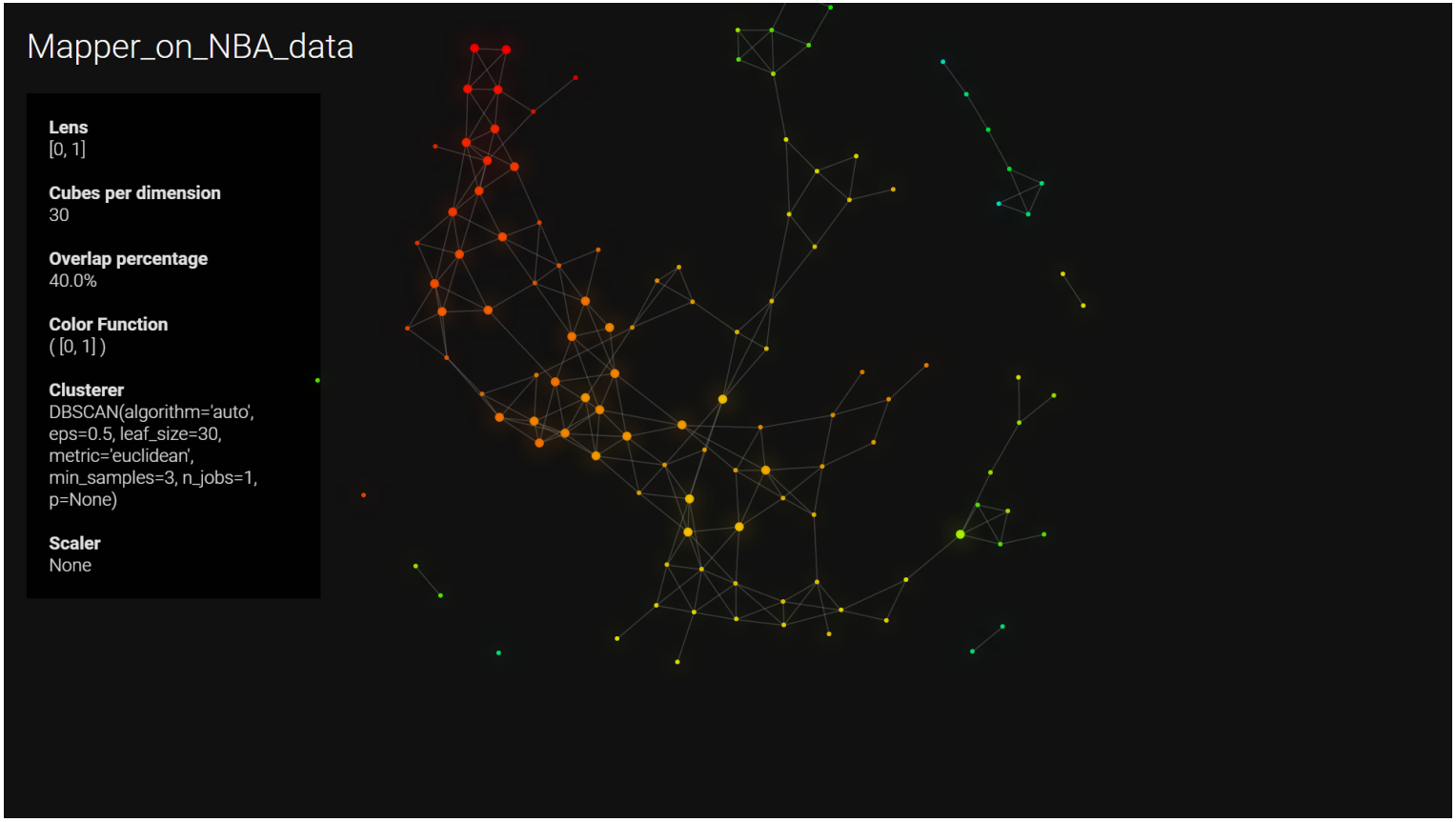
Mapper

Mapper in the continuous setting

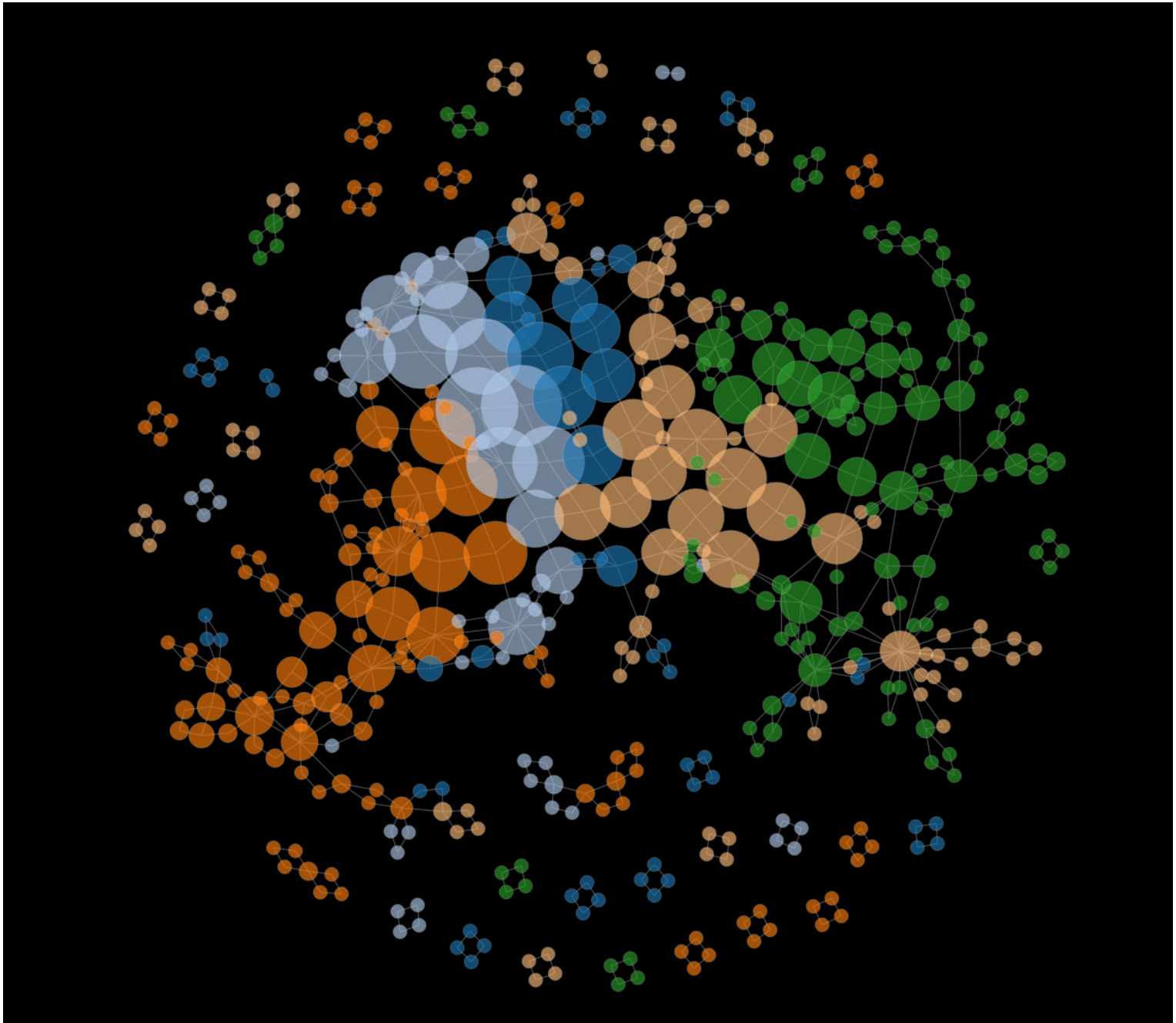


2

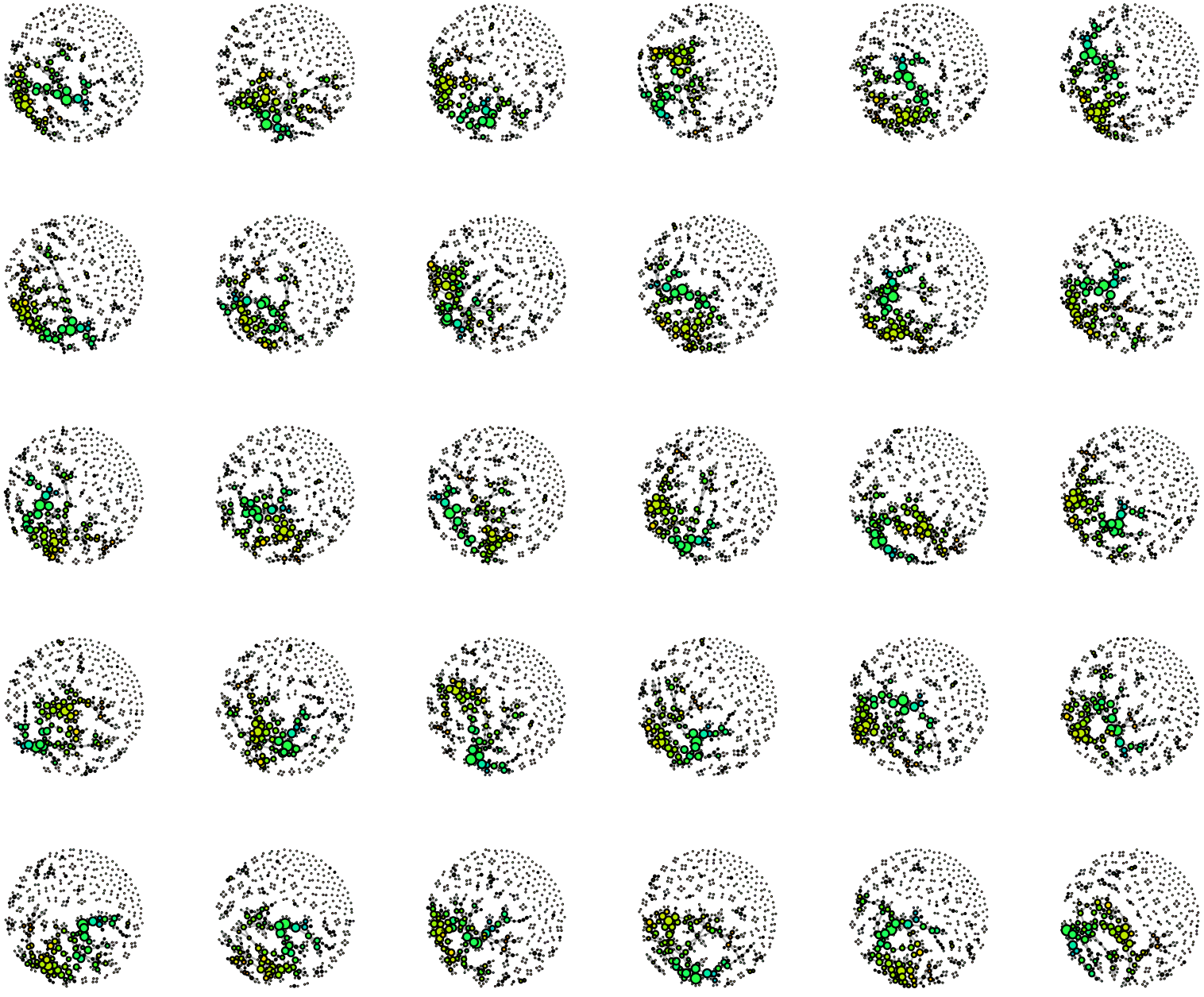
KeplerMapper



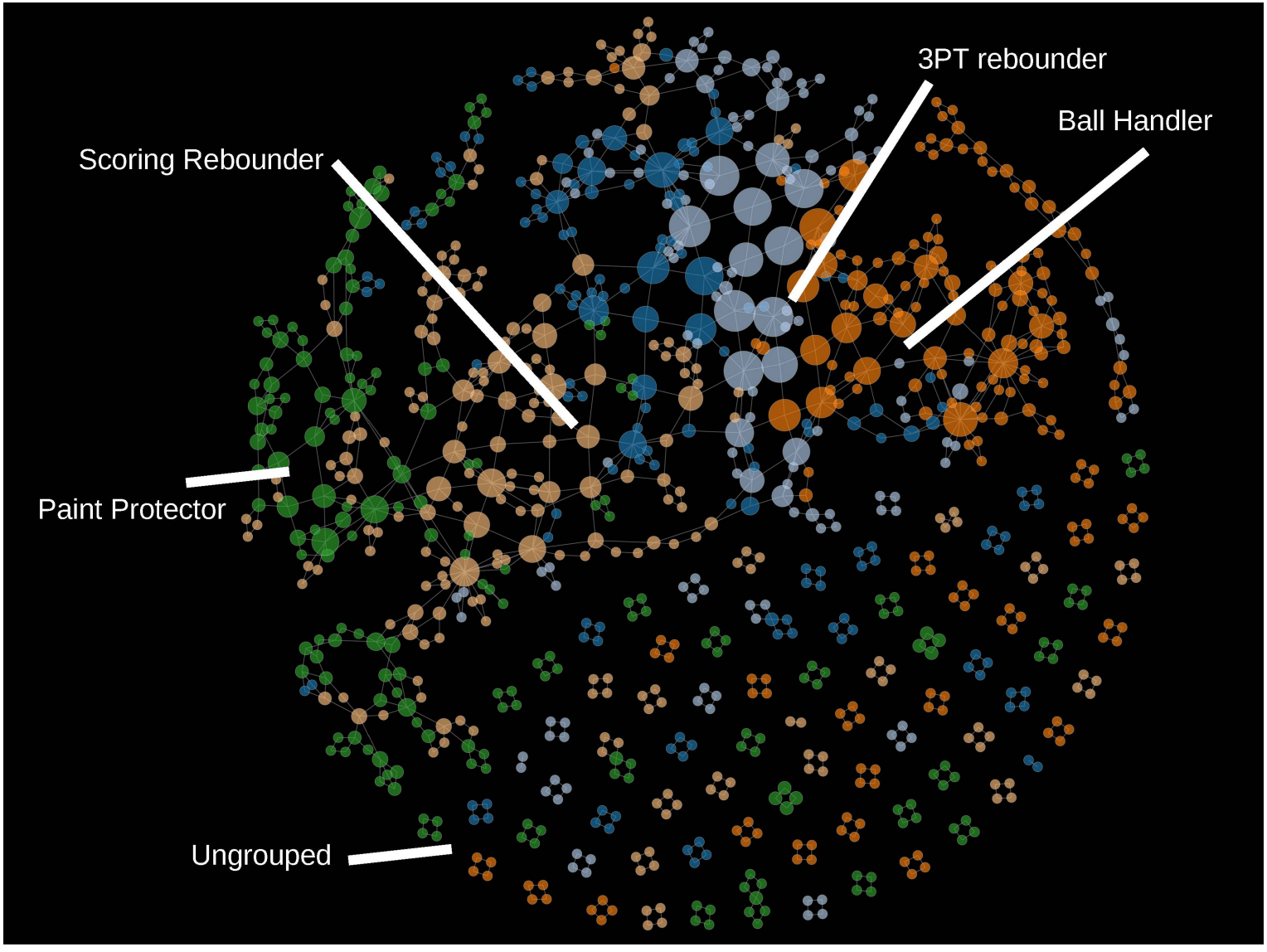
TDAMapper



Tuning parameters



Method's result



References I

M. Alagappan, P. Y. Lum, G. S. A. L. T. I. M. V.-J. J. C. . G. C. (2012).
Extracting insights from the shape of the data using topology.
SCIENTIFIC REPORTS.

Thank you

Questions?