Modularity Based Mapper Clustering Algorithm for Insurance Customer Data

Juhyun Kim, Kyoung-Kuk Kim
(Department of Industrial and Systems Engineering, KAIST)

CONTENTS

Modularity Based Mapper Clustering Algorithm for Insurance Customer Data

Ch 1. Introduction
- Motivation, problem, key point of our study.

Ch 2. Preliminary
- Clustering algorithm : k-means, hierarchical, SOM.
- Evaluation method : ARM, RFM model, cluster validity index, feature distribution.

Ch 3. Method
- Key component : mapper, modularity.
- Details about modularity based mapper clustering.

Ch 4. Experiment
- Apply our algorithm to insurance customer data.
- Compare the result with baseline.

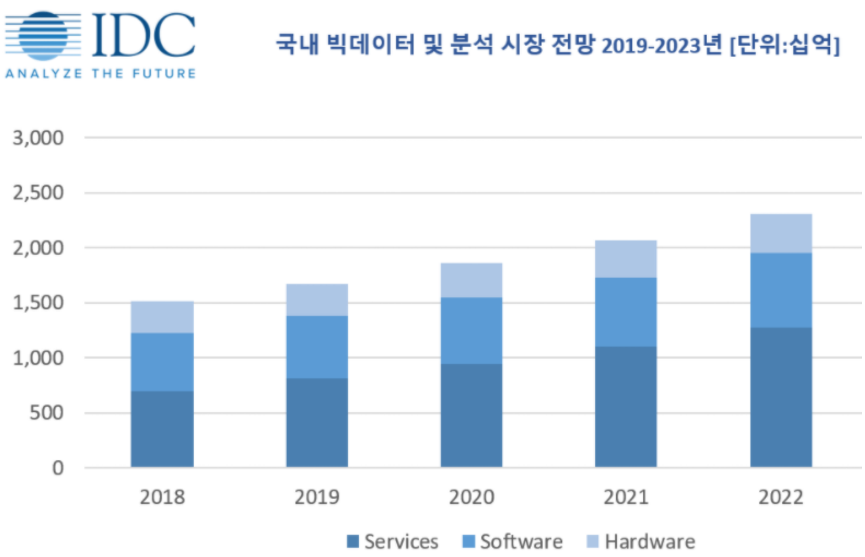
Ch 5. Conclusion
- Conclusion and future wotk.

Motivation

- Market for big data has continued to grow → financial and insurance industries are among the most keen players in using big data.

- Relevant datasets contain private information → de-identification necessary.

- Customer profiling and clustering is very important in customer-centric businesses → any dominant method is yet to be seen.



Source: IDC, 2019

Problem : Why Dosen't Clustering Work?

Binary Form of Data

- L : (1,0,0), M : (0,1,0), S : (0,0,1)
  d(L,M) = d(L,S)

Sparsity of Data

- Memory problem
- Computation time

Large Dimention of data

- Many customers
- Much information

Meaningless Distance

- Curese of dimension
- Properties of data

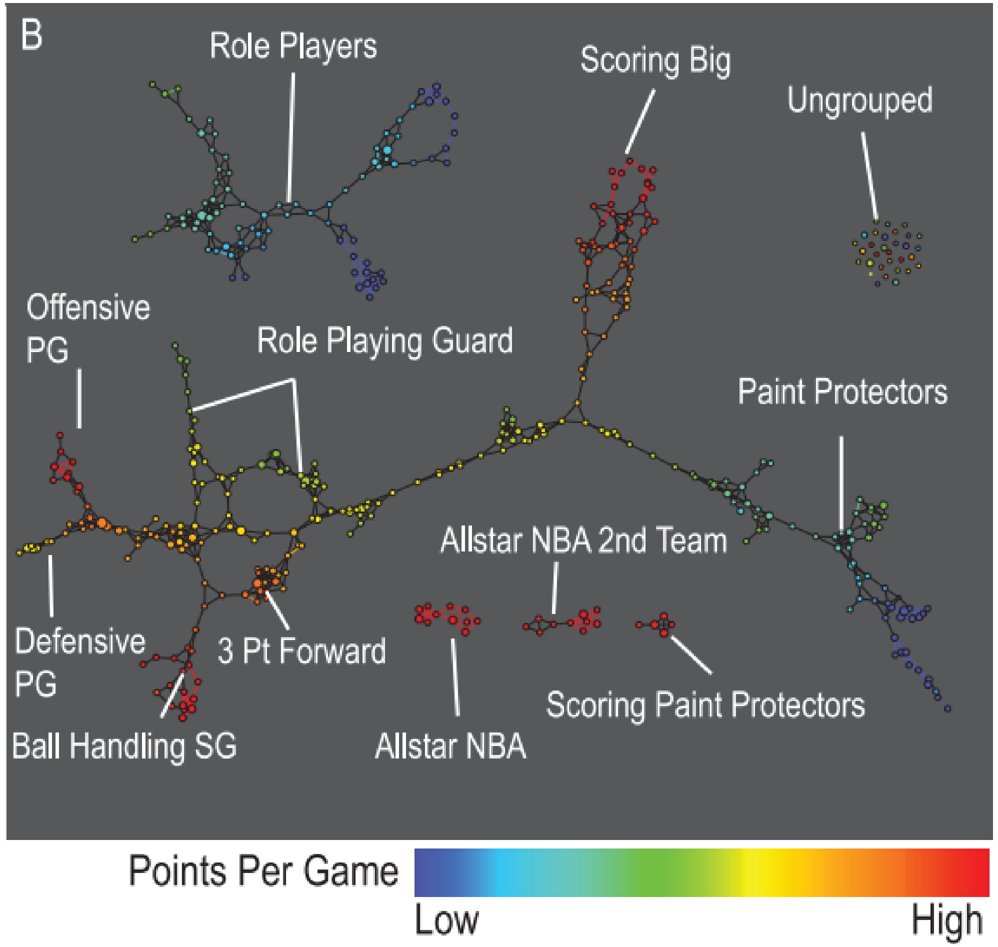
Solve problems by generating the hidden structure of data

Our Approach

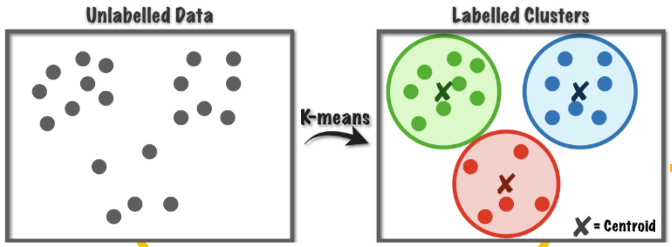
- We focus on the hidden structure of data, not a distance between data point → **our approach: generate the structure of data using mapper.**

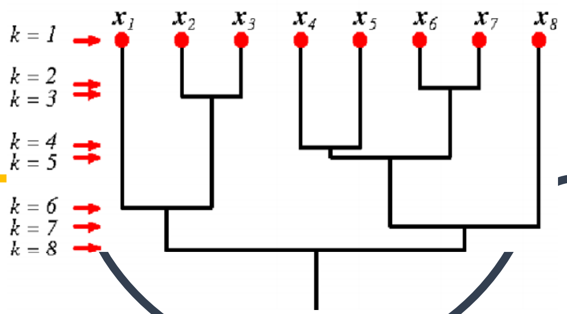
- The structure of mapper heavily depends on the parameters; yet very few research outputs → **our approach: find the persistent structure in terms of modularity.**

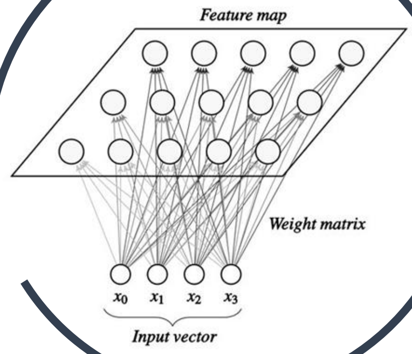
- Evaluate the clustering results from different viewpoints → **association rule, customer-centric measure, cluster validity, feature distribution.**



Clustering Algorithms







K-means Clustering

- Determine randomly initial center and each initial cluster consists of data points which share the closest initial center.
- Find better centers and clusters iteratively.

Hierarchical Clustering

- Start by letting each data point be a distinct cluster.
- Proceed by merging two closest clusters into larger ones iteratively.

Self-organizing Map

- Map the input data from a high dimensional space to a lower dimensional plot while maintaining original topological relations.

Evaluation Methods

Association Rule Mining

- Count the number of gene-rated association rules.
- Use the result for marketing.

RFM Analysis

- Find characteristics of custo-mers by using three factors : Recency, Frequency, Monetary.

Cluster Validity Index

- Measure the sparation bet-ween custers and compact-ness in a cluster.
- Use FS index, XB index, and BH index.

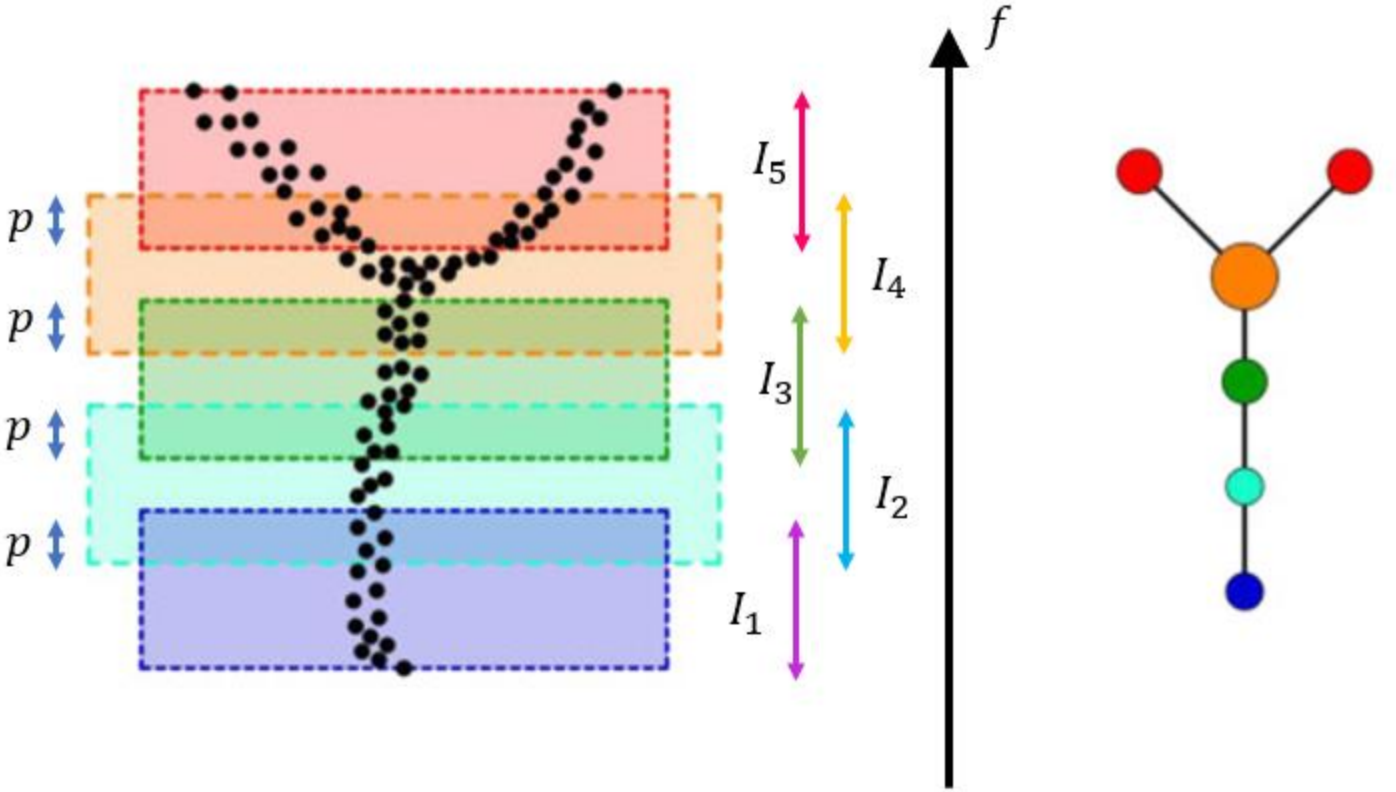
Feature Distribution

- Show important feature dis-tribution per clusters.
- Find the pattern of customers in a cluster.

Mapper

Mapper

- Useful tool which converts a data set into a graph structure.
- Mapper construction
  - Put data into overlapping bins.
  - Cluster each bin & create network.
    Vertex = a cluster of a bin.
    Edge = nonempty intersection between clusters.
- Parameters
  - Lens(f) : scalar function for input data.
  - Resolution(S) : the number of bins.
  - Gain : overlapping percentage.



Example of a mapper structure

- f = y-coordinate, S = 5, Gain = p.

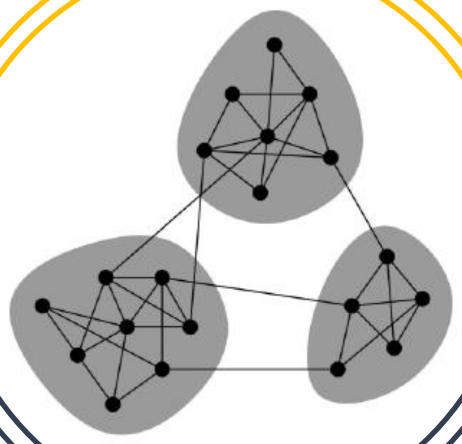
Modularity

Modularity

- Capture how good given communities are compared to a randomly wired network.

$$\mathcal{Q} = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w).$$

- Beyond about 0.3 is a good indicator of significant community in the network.



Modularity Maximizaion

- Modularity is variable in a single network affected by the shape of communities.
- Use greed approach to find an optimal modularity.
  - Calculate moduality change when communites are combined.
  - Select the largest value, join the corresponding communites.
  - Repeat above steps.

Modularity Based Mapper Clustering

GOAL : Clustering Complex & big data

Structure Finding Phase

Parameter Selection Phase

Clustering Phase

Evaluation Phase

Find persistent hidden structure

- Persistent structure in terms of the number of communities and maximum modularity value.

Select the best parameter set

- Exception for outlier cases.
- Parmeter set having the simplest mapper structure.

Make clusters using communities

- Clusters comprise data points in communities.

Evaluate the outcomes

- Association rule ming, RFM analysis, custer validity index, feature distribution.

EXPERIMENT

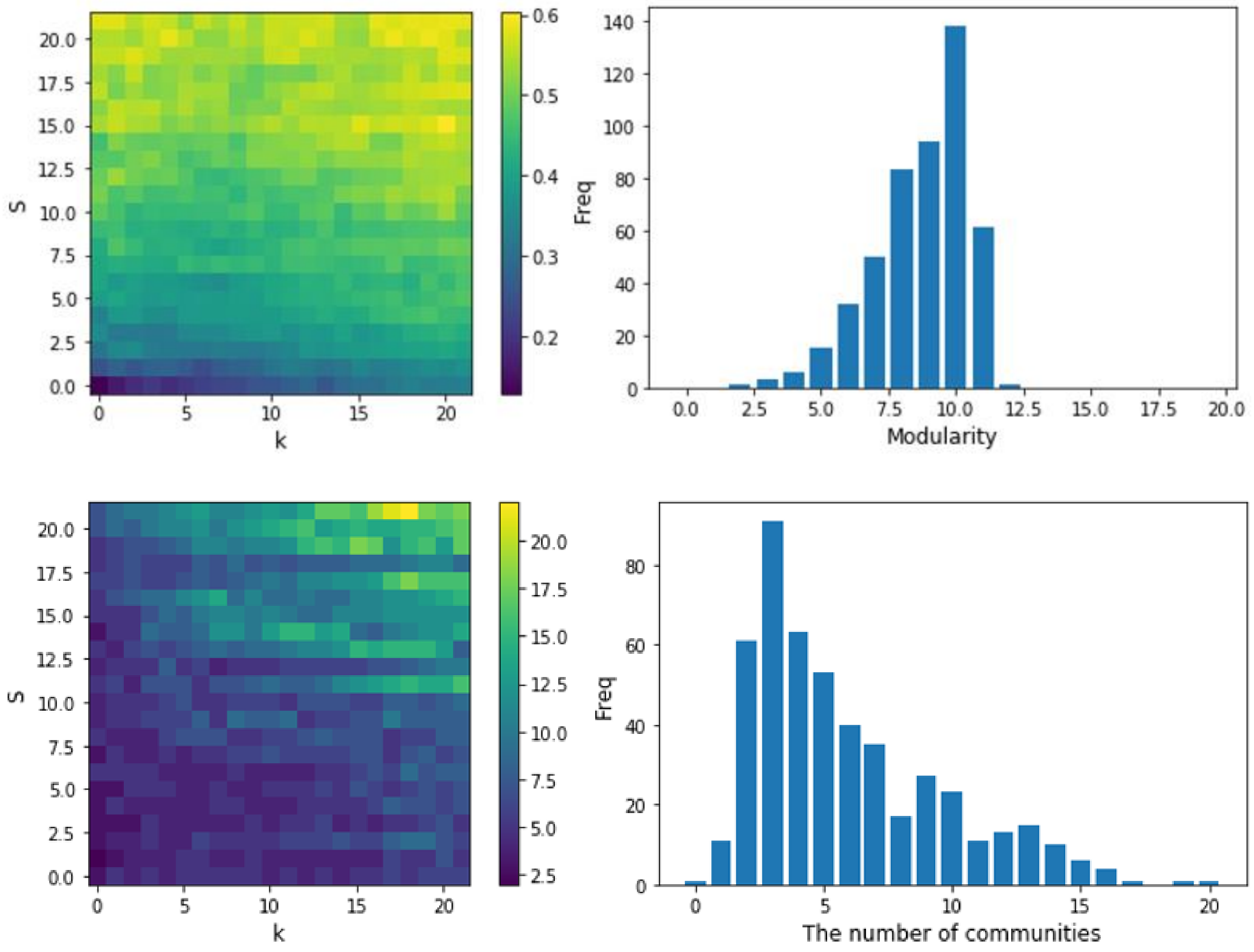
Date Set

- The data set consists of customers in insurance, card, and bank companies.
- The number of customer is about 80,000 (data matrix has over 80,000 rows).
- All information is converted into binary vector (data matrix has over 590 columns).
  - Categorical : binary vector per each category.
  - Numerical : binary vector per some unit.
- The data set includes demographic, financial, and personal information (after de-identification).

Variable name	Data type	Description
Matching	Binary	Product purchase record; Buy: 1, Not buy: 0; 307 columns(307 products: Insurance + Riders)
Gender	Binary	Male: 0, Female: 1; 1 column
Valid	Numerical	The number of valid insurance policy; 16 columns
Invalid	Numerical	The insurance is classified invalid because of non-payment; 5 columns
Cancel	Numerical	The number of insurance has been canceled; 5 columns
Insurance	Numerical	Sum of Valid and Invalid ; 7 columns
Ratio	Numerical	Riders over Insurance ; 12 columns
Premium	Numerical	total premium of a person; 17 columns
Avg Pre	Numerical	Premium over Insurance; 14 columns
CMIP	Numerical	total CMIP of a person; 13 columns
Avg CMIP	Numerical	CMIP over Insurance; 13 columns
Duration min	Numerical	The time from the last policy contract(month); 16 columns
Duration max	Numerical	The time from the first policy contract(month); 16 columns
Age	Numerical	The age of a person(years); 13 columns
Job risk	Numerical	The risk of a person determined by the firm; 5 columns(5 classes)
Job	Categorical	The job of a person; 28 columns
Address	Categorical	The address of a person; 18 columns
Hobby	Categorical	The hobby of a person; 16 columns
Card	Categorical	The grade of card firm membership; 4 columns(4 classes)
Group	Categorical	The grade of group firm membership; 5 columns(5 classes)
Home	Categorical	The shape of home; 5 columns
Life	Categorical	The shape of living; 6 columns
Credit	Binary	Hold credit card: 1, Not: 0; 1 column
Credit pay	Numerical	The monthly pay of credit card; 16 columns
Debit	Binary	Hold debit card: 1, Not: 0; 1 column
Debit pay	Numerical	The monthly pay of debit card; 13 columns
Salary	Binary	Have a pay: 1, Not: 0; 1 column
Salary pay	Numerical	The amount of a salary; 14 columns
Annuity	Numerical	Receive: 1, Not: 0 (only 3 people receive); 1 columns
Annuity pay	Numerical	The amount of an annuity; 1 columns(all people similar)

Result : Structure Finding Phase

- Let f = isoforest function, gain = 0.15 and use hierarchical clustering with k.
- The range of S and k is 4 to 25.
- Measure the maximum modularity and count the frequency of communities by increasing the interval size by 0.05.
  - The most frequent interval is 0.50~0.55
- Count the number of communities.
  - The most frequent number is 5

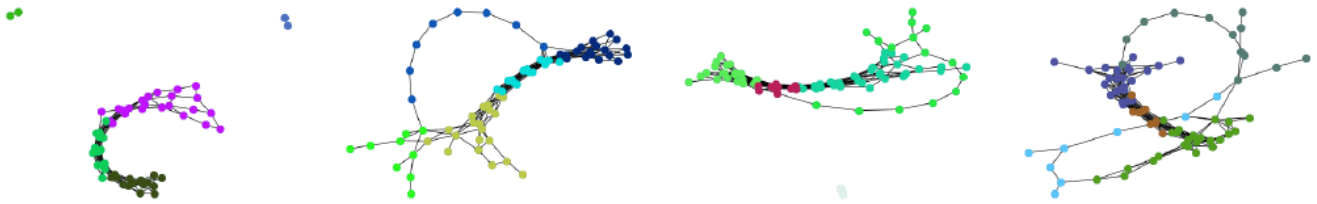


Result : Parameter Selection Phase

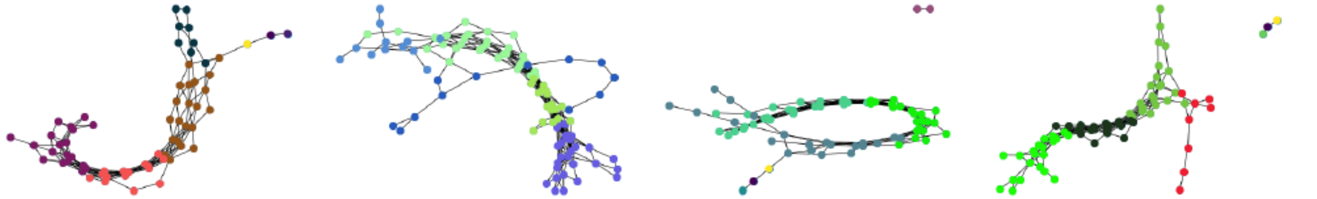
- There are 8 candidates having the persistent structure.

- Except the cases where a single node (or two nodes) consists of community: (a), (c), (e), (g), and (h).

- Select the parameter set with the simplest structure.
  - In a mapper, number of nodes = k×S → choose small value of k×S.
  - Select (b) with S = 16, k =5 , and Modularity = 0.53.



(a) $S = 15, k = 4, \mathcal{Q} = 0.50$ (b) $S = 16, k = 5, \mathcal{Q} = 0.53$ (c) $S = 16, k = 6, \mathcal{Q} = 0.50$ (d) $S = 18, k = 5, \mathcal{Q} = 0.51$



(e) $S = 19, k = 4, \mathcal{Q} = 0.54$ (f) $S = 19, k = 6, \mathcal{Q} = 0.54$ (g) $S = 20, k = 4, \mathcal{Q} = 0.52$ (h) $S = 22, k = 4, \mathcal{Q} = 0.54$

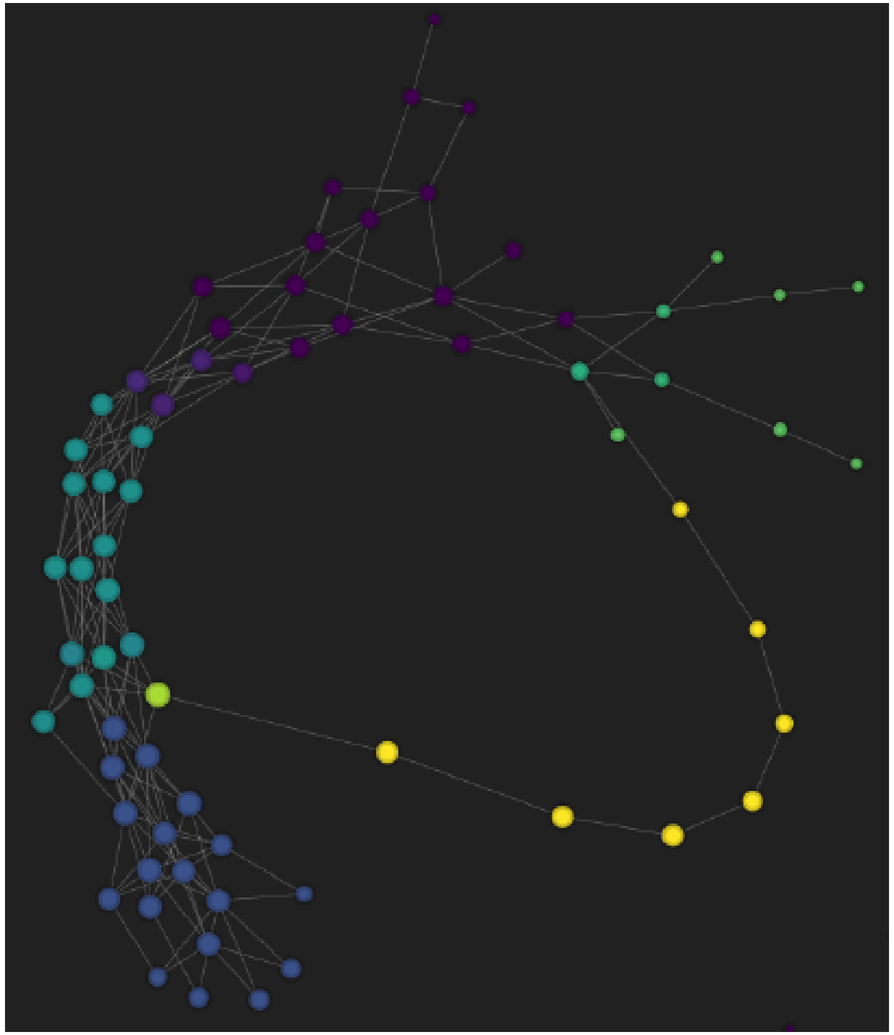
	(b)	(d)	(f)
$k \times S$	80*	90	114

Result : Clustering Phase

- Selected mapper has the most persistent and simplest structure.

- Consider each community as a cluster.
  - The data points correspond to nodes in a community consist of a cluster.

- There are 5 clusters and clusters denote different color.
  - Since this clustering has overlapping, some nodes have mixed color.



Result : Evaluation Phase

ARM Result

- Compare the number of association rules.
- Use the result for insurance recommendation.

Clustering Algorithm	The number of rules
k - means Clustering	8916
Agglomerative Clustering	2278
SOM	4552
Mapper Clustering	14989*

RFM Analysis Result

- Compare the RFM score of clusters and range of it.
- Use the result for customer valuation.

	Very High	High	Middle	Low	Very Low	Range of Score
k - means Clustering	11.96(2)	11.17(5)	9.42(3)	7.97(1)	6.95(4)	5.01
Agglomerative Clustering	11.50(2)	9.65(3)	9.08(5)	8.31(4)	6.97(1)	4.53
SOM	11.49(1)	9.39(5)	8.61(2)	7.48(4)	6.62*(3)	4.87
Mapper Clustering	13.66*(4)	13.10(1)	9.91(3)	8.94(5)	7.00(2)	6.66*

Cluster Validity Index Result

- Compare corresponding indices.
- The result shows clusters are well separated.

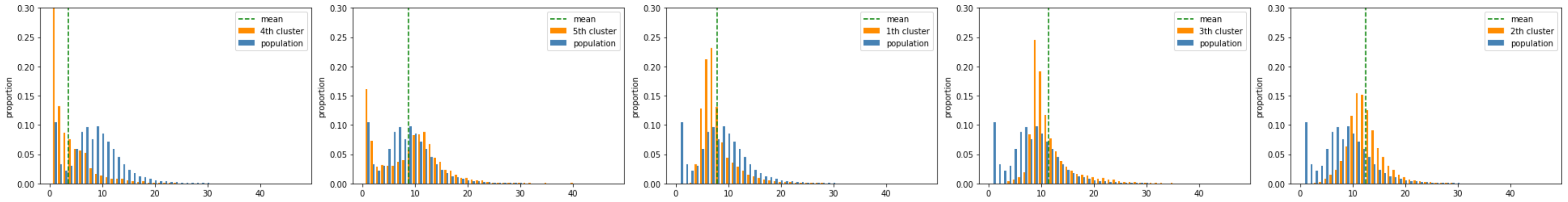
	FS index (10^6)	XB index (10^3)	BH index (10^6)
k - means Clustering	2.5769	42.3663	71.7427
Agglomerative Clustering	2.5778	170.9349	213.1737
SOM	2.5771	6.4009	19.4333
Mapper Clustering	2.5062*	5.6205*	7.2079*

EXPERIMENT

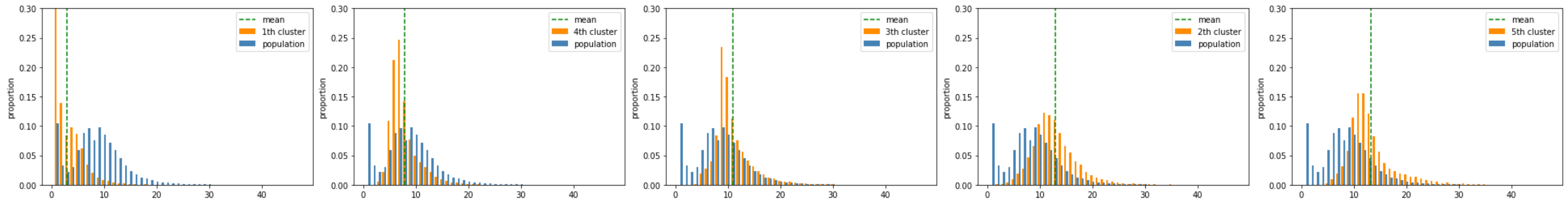
Result : Evaluation Phase

Feature Distribution Result

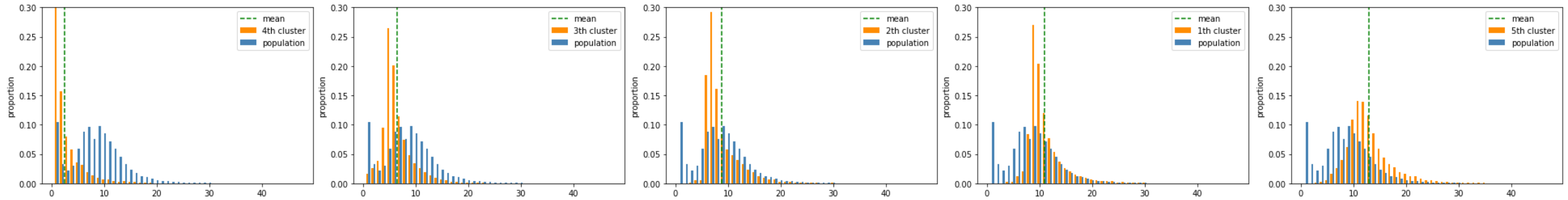
- Choose important feature → the number of buying product.
- Compare distribution of the feature for each cluster → feature for each cluster → find the pattern or properties of each cluster.
- For orther feature, we apply similar approach.
- Use the result for customer profiling → See next page.



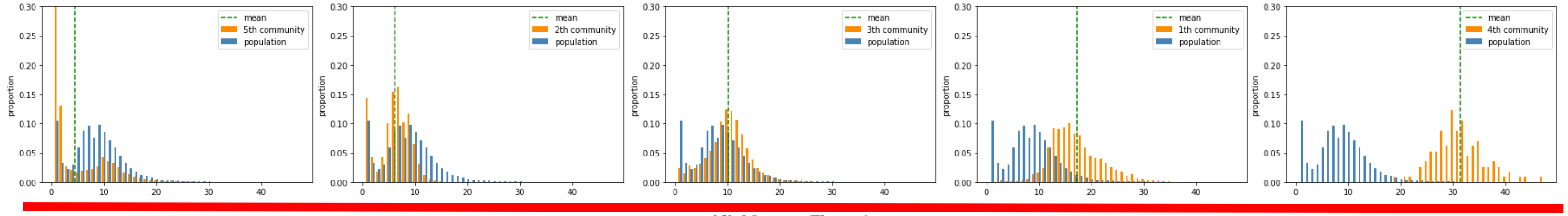
(a) Hierarchical Clustering



(b) k-means Clustering



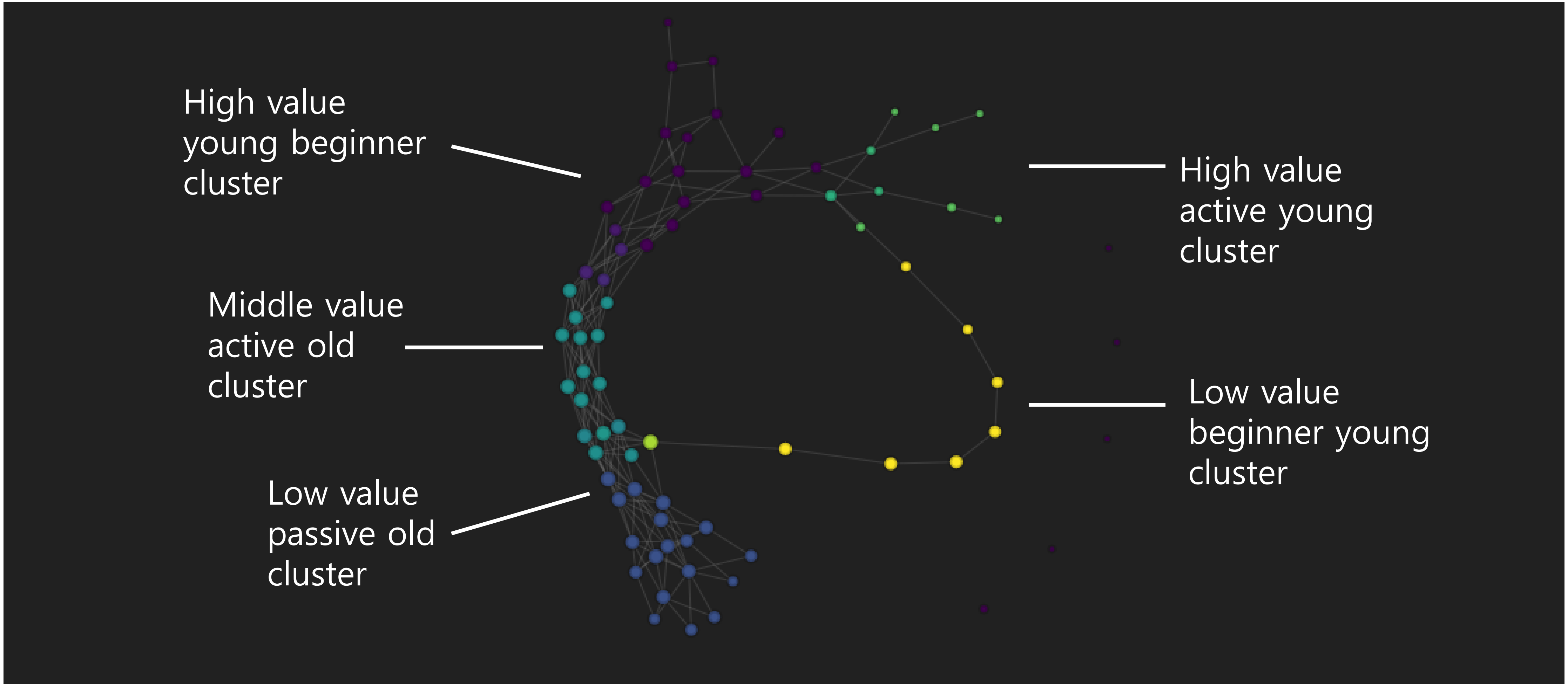
(c) SOM



(d) Mapper Clustering

Additional Result : Clustering Analysis(tagging??)

Features : the number of insurance, age, and duration(min&max).



Conclusion & Future Work

- For clustering financial and insurance data, the major challenges are the big size and complex form of data. It may make any distance measure between data points meaningless.

- We propose the modularity based clustering algorithm to find the hidden structure of data and generate clusters. To our knowledge, our model is a new approach combining a mapper algorithm and network analysis.

- We apply our algorithm to a real insurance customer dataset and find it outperform some other well known methods in terms of: recommendation, customer value, validity of clusters, and customer pattern.

- It remains to be seen the impact of TDA and network analysis together.

Thank You!!!