SWS3001

Summer 2018

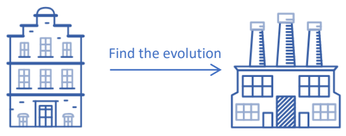
Evolution of Architectural Genre

Group 15: Deng Yiran Li Yike Tang Mingze Xing Zhuoyue

NUS National University of Singapore

Introduction

The goal of our project is to analyze the evolution of world architectural styles.



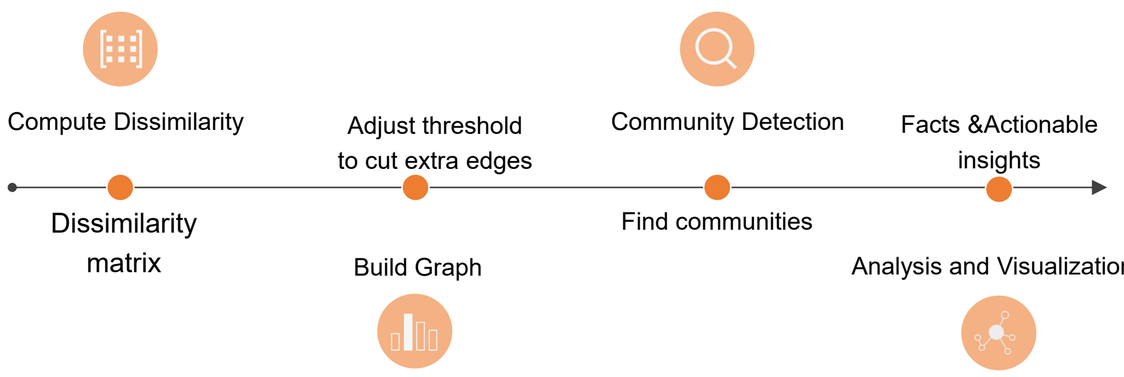
Objective

We mainly focus on representative architectural genres around the world, including 40 classes, varying from Summer Palace in Asia to Baroque styles in the west. Through neural network and clustered algorithms to observe and analyze the historical relations between different styles.

Overview

In our training dataset, we input 40 classes, each of which contains 8 different pictures of the same architecture. Each picture has been modified into a gray one with the same size of 100 times 100. Finally, we really find something interesting and astonishing.

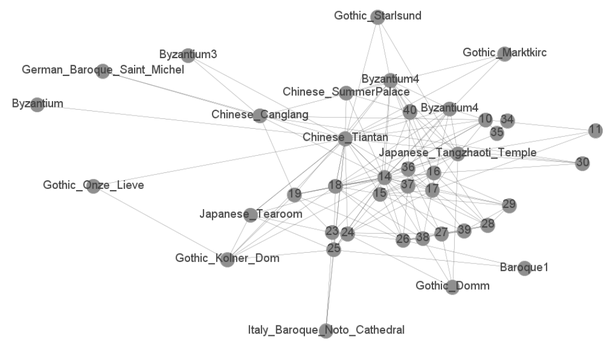
Overall Workflow



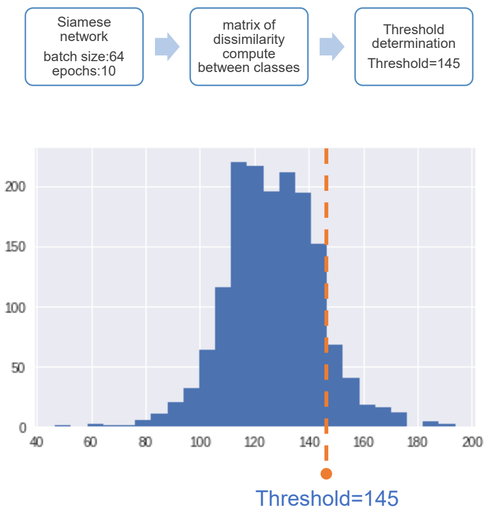
Details of Algorithms & Methods Used

Step 1

- Datasets Generation: 40 classes and 8 images of each class

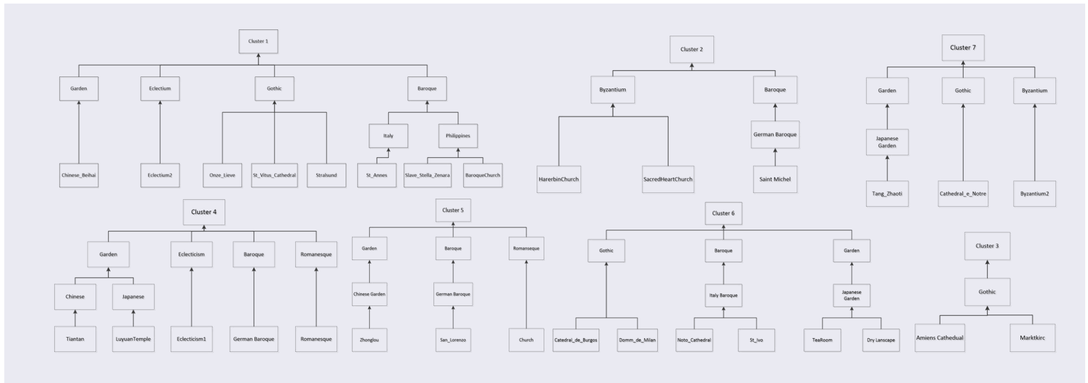


Step 2



Step 3 algorithms application

- Single-Linkage



- Markov Clustering

Inflation:

$$(\Gamma_r M)_{pq} = (M_{pq})^r / \sum_{i=1}^{k} (M_{pq})^r$$

expansion:

$$(M_{pq})' = (M_{pq})^e$$

- Girvan-Newman

Results Obtained

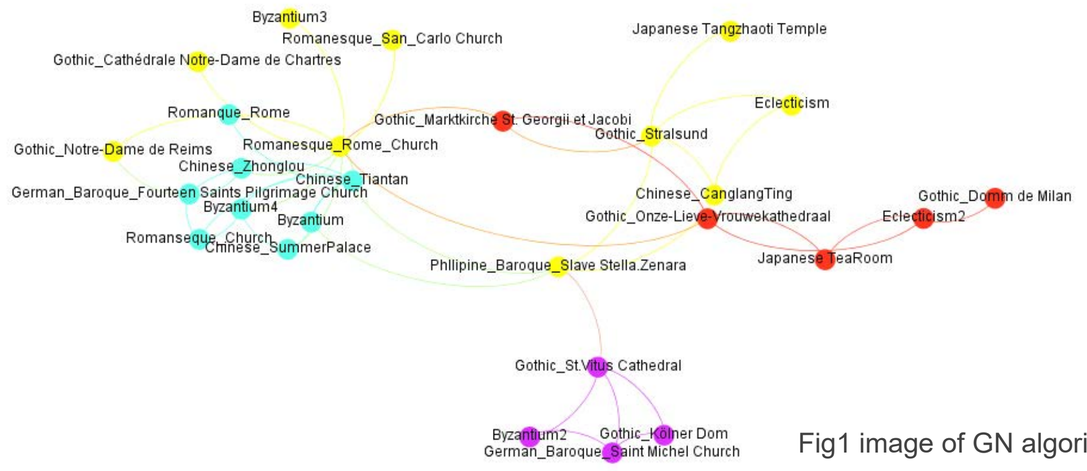


Fig1 image of GN algorithm

In the image of Girvan Newman algorithm, we could find that Paoay Church, built in Philippines, connected to both Chinese architectures and European architectures. After looking up Philippines history, we discover that church was designed by European architect but was built by Chinese workers, so it combined Chinese and western styles.

Fig2 image of single-linkage algorithm

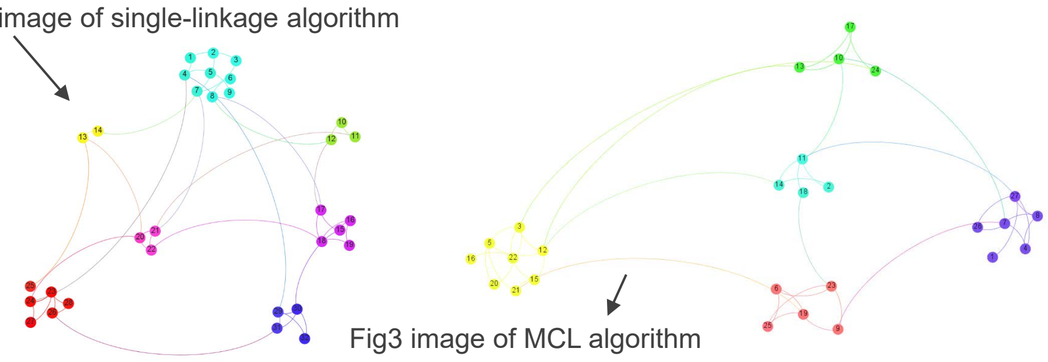


Fig3 image of MCL algorithm

Through single-linkage algorithm, we could find that the cluster where St. Vitus Cathedral in have both Gothic and Baroque styles. St. Vitus Cathedral was only a Gothic building at the beginning, but architect finished the uncompleted Gothic main tower in Baroque style, so it is a bridge of Gothic and Baroque buildings. It shows the evolution of architectures that buildings of Gothic, Baroque and the building between them are in the same cluster.

Through MCL algorithm, Schulte Cathedral, Reims Cathedral and Market Church gathered together. The classic period of Gothic architecture was formed in 12-13th century whose representative architecture is Schulte Cathedral. Reims Cathedral was built at the end of the 13th century. These two cathedrals are all in France. Most of the Market Church still maintain the architectural style of the 14th century North Germany. Thus we can see different time periods' Gothic architectures may look differently, but when they are recognized by the machine, they look the same. These architectures witness the evolution of Gothic architectural style.

Conclusions & Lessons Learnt

Mixed-style buildings can be inter-class or intra-class connections. The buildings of the same Architecture style, though in different periods, are usually clustered in the same cluster.

Different works of the same architect will appear in one cluster. Architectures in the same region usually look similar with each other.

Some later buildings, which mimic earlier ones, usually share a resemble architecture style with the earlier ones.

Technical lessons

- Datasets preprocessing
- The way to value information
- Data visualization
- Results display

Non-technical lessons

- Teamwork
- Progress of spoken English
- Carefulness
- Patience

SWS3001

Summer 2018

Analysis of Monkey Species Similarity Based on Siamese Neural Network and Community Detection Algorithm

Team G02
project ID
16

Li Chenggao Li Jiang Tan Haoyue Zeng Xiong

NUS National University of Singapore

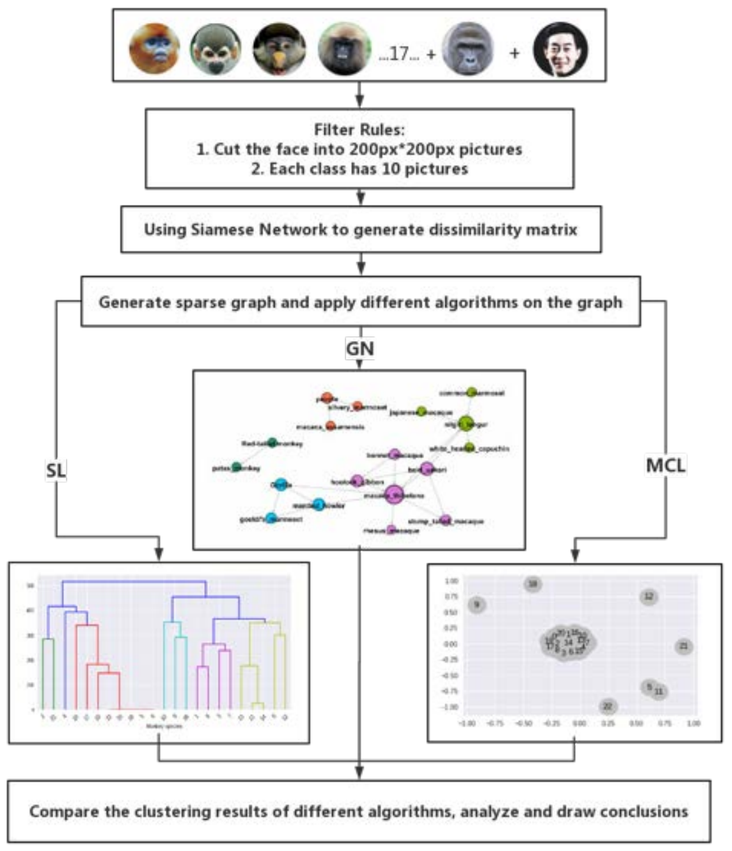
1. Introduction

- Collect facial pictures of 22 monkey species as 22 sub-classes and add human face as a control class
- Use siamese neural network to generate dissimilarity matrix
- Apply 3 different CD algorithms to cluster and evaluate clustering results of different algorithms
- Make some assumptions according to clustering
- Collect geographical and gene information to verify our hypothesis
- Questions and thinkings

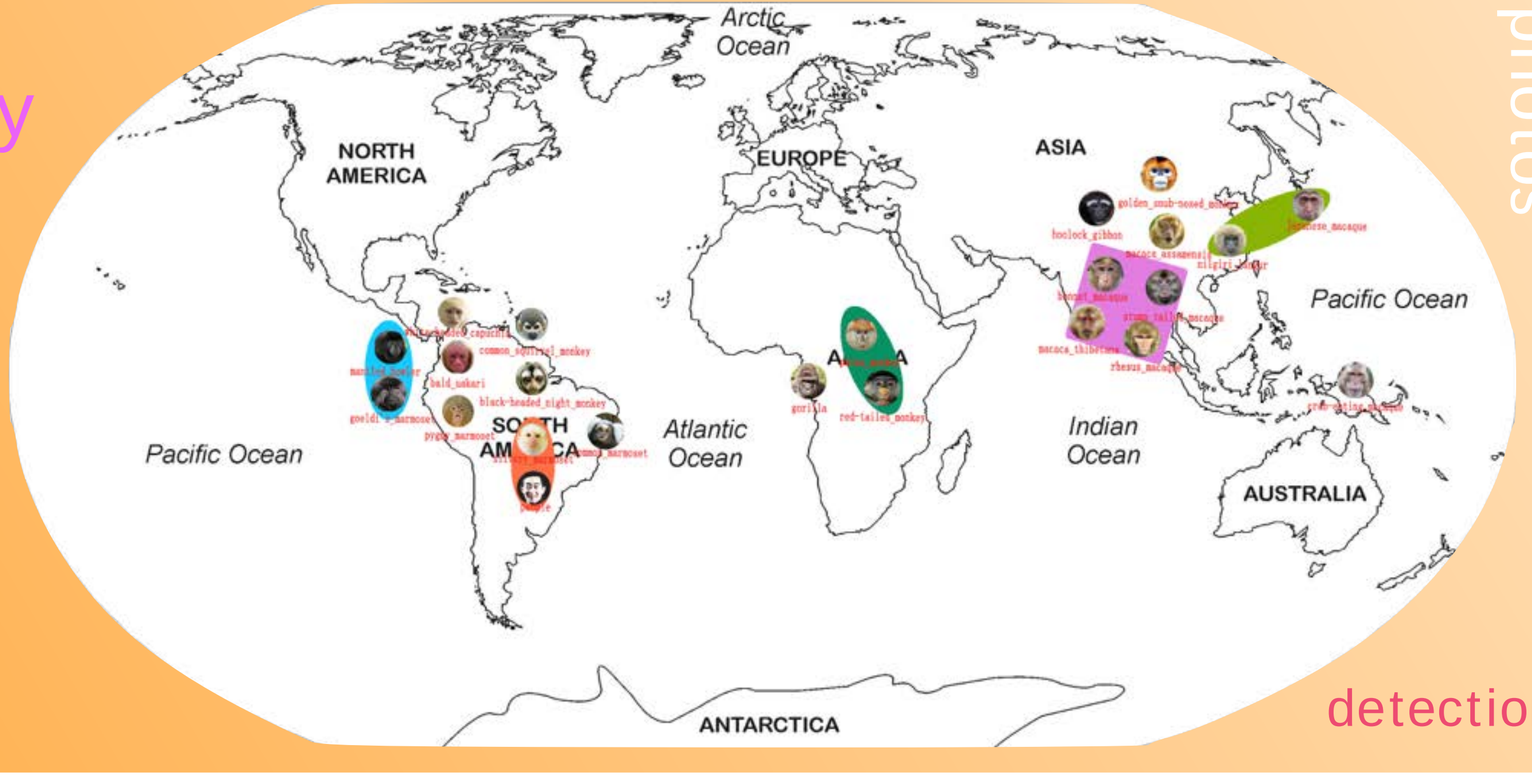
2. Details of Data & Methods Used

- Monkey image is collected from Kaggle and Internet
- Images with more than one faces are filtered and then crop the facial part of those qualified images
- 22 different species and 10 images for each
- 23*23 adjacent matrix are generated by siamese
- GN, SL and MCL algorithms are used separately for community detection and generated visualization charts

3. Overview of Our Study



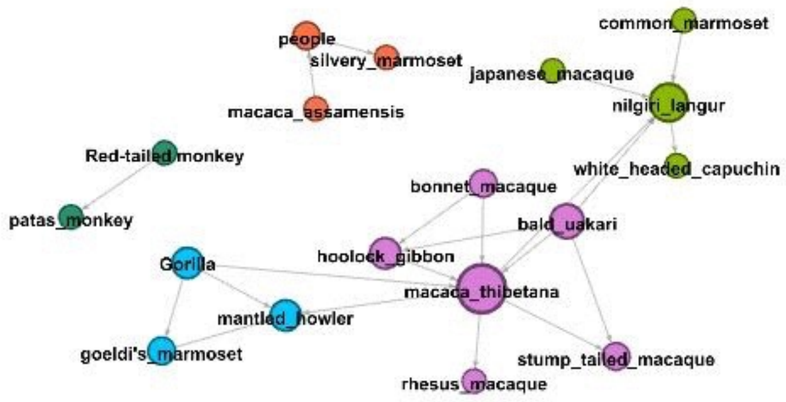
Input **facial pictures of 22 monkey species**. compare the monkey face similarity through the **siamese neural network** and then use different CD **algorithms** to find communities. We assume that the relationship between monkey species is studied by combining **geographical distribution** and **gene similarity**.



4. Results Obtained

Hypothesis:

- Monkeys with similar facial features are more likely to be in the same community
- Monkeys in the same community in a large possibility live closer



Analysis:

- Girvan–Newman algorithm gets the most interpretative results
- Cluster in purple has most species in a geographically similar location, they also belong to the same subspecies

5. Conclusions

The conclusions:

- Girvan–Newman algorithm makes the best result
- There is a certain relationship between the clustering of monkey groups and geographical distribution.
- Learning rate equals to 0.005 suits our image well

What we know we don't know:

- Except those 3 CD algorithms, is there any other cluster methods that may get better result ?
- For monkeys in the same community but not in the near habitat, is there migration or gene mutation ?
- Why the clustering is so bad when we define each image as one class by using the siamese network ?

Gossip Arts

Team 17: Liu Yihao, Shi Shangzhou, Wang Zhe, Zhang Yilun

NUS National University of Singapore
SWS3001
Summer 2018

Introduction

- Select various paintings on Wikiart as data.
- According to dissimilar matrix, the clustering is carried out by CD algorithms.
- Based on the painting's style, color, etc.
- MCL, GN, D-Cluster and other community detection algorithms were selected for image classification.
- It is expected to explore the relationship behind works through the connection of paintings.

Objective

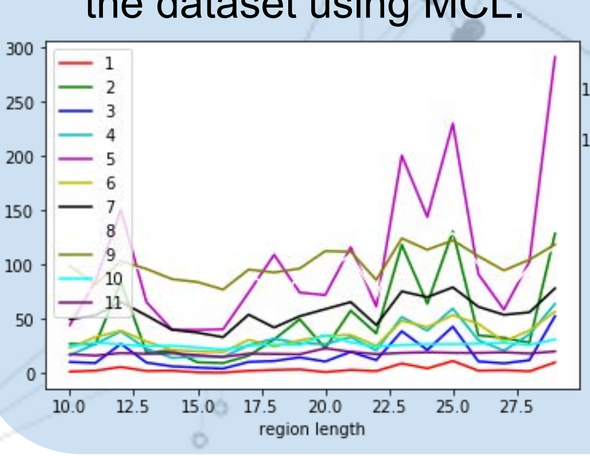
Our research takes different styles of painting as the data. According to the style and the color using of paintings, we can figure out dissimilar matrix. Then we use the MCL, GN and D-Cluster CD algorithms and find out the connection behind different works.

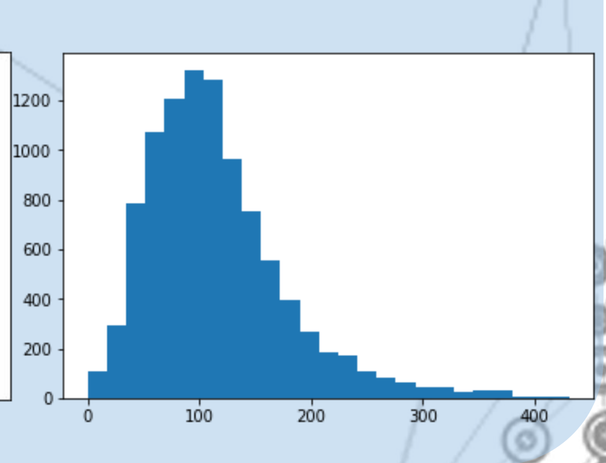
Overview of Our Study

- We choose paintings of different painters and genres from Wikiart as our data set.
- Our tags include the title, painter, genre and time of the painting. Totally 30 painter's works are included and each of them has approximately 14 paintings.
- The resolution of each piece is 100*100. The painters are from Europe, China and Japan.
- We hope that we can uncover the relations between different painting styles, the connections between different paintings of the same author and the possible historical relations among the genres.

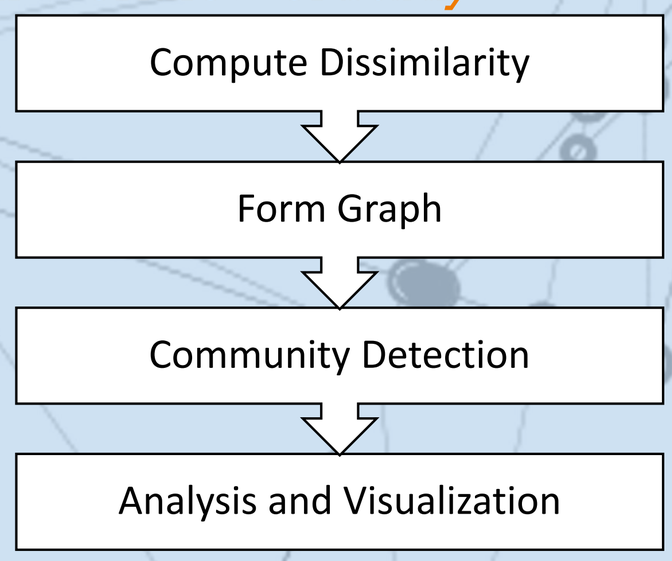
Details of Data & Methods Used

- The dataset is formed by 406 paintings in 30 clusters, each cluster represents a painter. The size of each painting is 100 by 100.
- VGG19 network and transfer learning are used to extract more accurate information of styles. After training, the images are transferred into the network and the style loss is calculated based on result of the hidden layer, which is converted into the grim matrix. What's more, the content of two images is also considered. After it, cluster analysis is performed on the dataset using MCL.





Overall Workflow of Study



Results Obtained

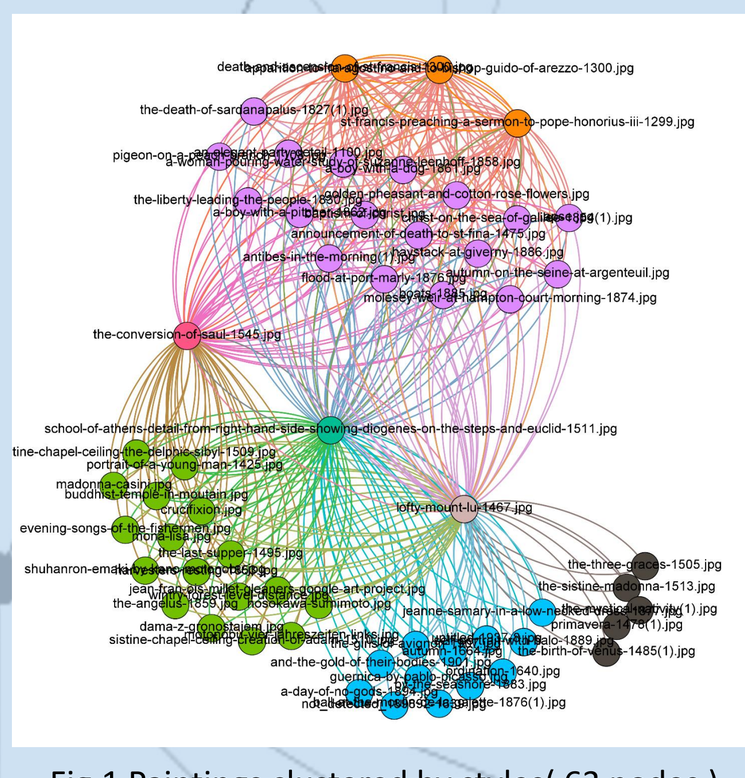


Fig.1 Paintings clustered by styles(63 nodes)

The test set includes 21 artists each with 3 paintings. By using Markov Clustering, the threshold is set to 2000. The communities showed on the Gephi directly illustrate the relationship between Eastern and Western artists and the distinct genre.(See Fig.1)

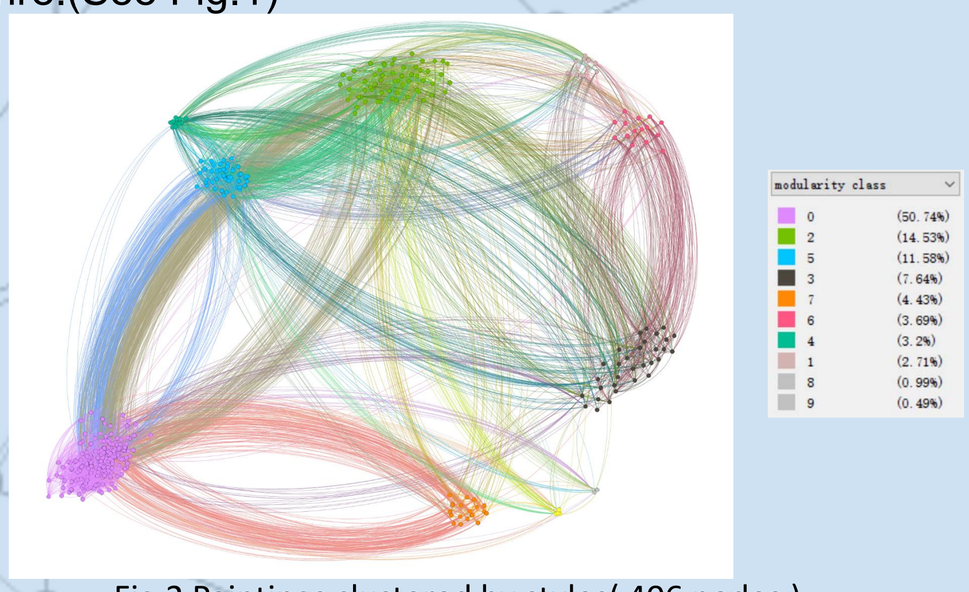


Fig.2 Paintings clustered by styles(406 nodes)

The full set with 30 artists with 406 paintings is applied by 3 different algorithms. The result shows more precise information, which intuitively describes the difference between impressionism and abstract genre and the migration of the style of Japanese arts, etc.

Conclusions, Lessons Learnt

- Arts do speak! Not only the genre is divided, but the details are showed vividly such as the similarity between co-workers and friends, but also the change of arts style as time goes by. Of course, there still be more amazing relationship left behind.
- We learnt the most beneficial thing during the class, the bravery in Skeptical learning, the global view throughout the project and the participation in the team. The project banded the relationship with my teammates to whom I am since rely grateful.

SWS3001

Summer 2018

Community Detection Of Chinese Movie Actors

Project ID: 18
Group 04

Name of Students (in alphabetical order)
Cai Derun, Lian Yibing, Zhang Mingcheng, Zhu Bingzhang

National University of Singapore

Introduction

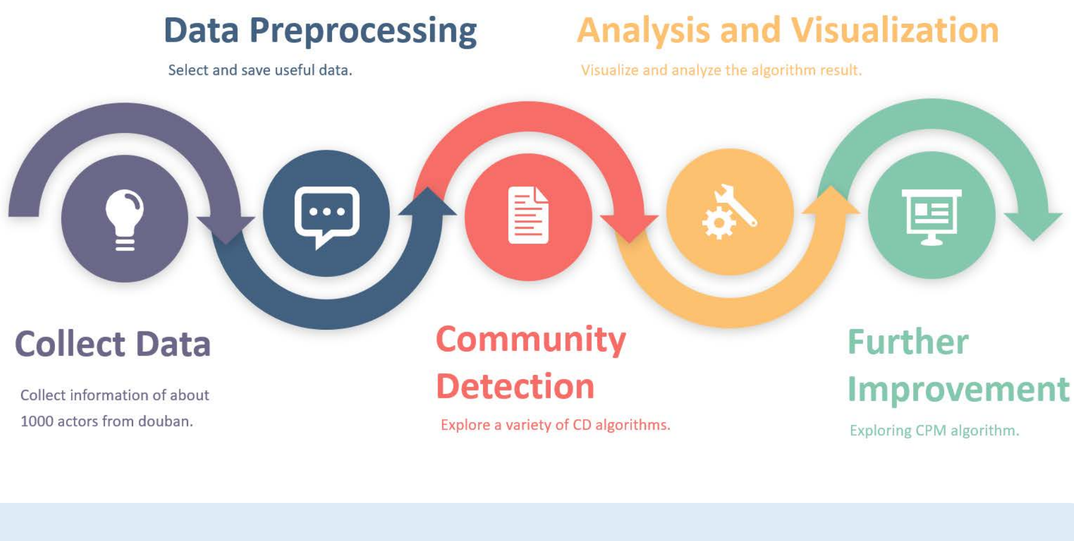
- In a network, communities are sets of nodes with dense connections internally and sparser connections between groups.
- Based on the cooperative relationship between Chinese actors, we build a graph and try several algorithms to do the community detection on this issue, in the hope that we may find some *actionable insights* behind the data.

Objective

- Find a proper method to detect communities among Chinese actors.
- Validate whether the **performance**, **birthplace** and **movie preference** of actors are related to community formation.

Overview of Your Study

We developed web crawlers to collect the basic information of actors and their cooperation times. After we found out 12 communities by Louvain Algorithm, we validated the result and analyzed the relationship between three factors and community formation.



Details of Data

- Almost 80000 pieces of data about nearly 1000 actors from the most famous Chinese movies list on *movie.douban.com*
- The data contains the cooperation times between them as well as the basic information about actors.
- Use cooperation times as edge attribute to build a graph.
- Use threshold(cooperation times<6) to cut off edges.
- Delete the vertexes which are nearly isolated.
- The basic information including birthplace, age, movie rating and genre are used to validate the cause of community formation.

Modularity Optimization Method

- Modularity is designed to measure the quality of a particular division of a network into communities.
- It is defined as the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random.

The Louvain Method:

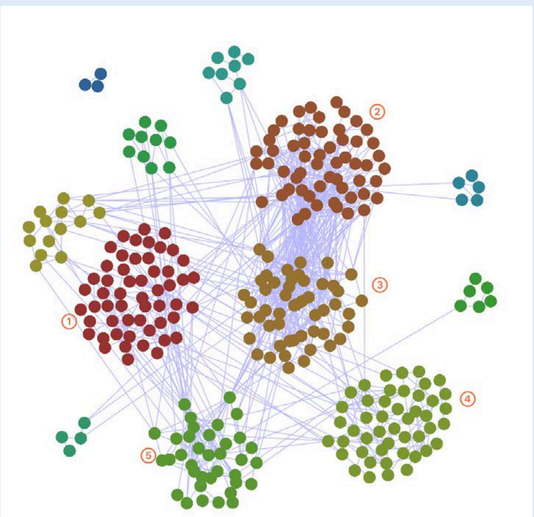
- A greedy optimization method. O(n logn) time complexity.
- Firstly, small communities are found by optimizing modularity locally on all nodes.
- Then each small community is grouped into one node and the first step is repeated.

Why Louvain Method?

- Methods based on modularity optimization focus more on the community structure than other methods (hierarchical clustering including Single Linkage and Ward's Method, Label Propagation Algorithm).
- The results of modularity optimization methods are almost the same. (Clauset, Newman, Moore's Algorithm, GN Algorithm based on modularity)
- The Louvain Method is the fastest and its result is more reasonable.

Results & Analysis

Results of Community Detection



- Twelve communities found, five large and seven small.
- The small ones are based on the personal relationships.
- So we focus on the cause of large communities.

Analysis of the Results

ID	Comedy	Romance	Action
1			
2			√
3	√		√
4	√		
5		√	

Movie preference of actors in 5 main communities

ID	Mainland	Hong Kong	Taiwan
1	48	0	0
2	18	20	1
3	6	31	1
4	45	0	0
5	30	4	0

Birthplace of actors in 5 main communities

ID	The average score of their recent movies	The average score of their best 5 movies	Their average age (ranking)
1	6.52(1)	8.71(2)	42.15(4)
2	6.06(3)	8.64(3)	56.76(1)
3	5.59(4)	8.34(4)	49.40(3)
4	6.49(2)	8.80(1)	51.16(2)
5	5.77(5)	8.27(5)	36.17(5)

Information about their average performance scores and average age

We processed the basic information of actors to measure the **performance, birthplace** and **movie preference** of actors in each community. What we found are as below:

- Communities have **unitive movie preference**, except No.1 and No.3.
- Actors of each communities are mostly **from the same region**, except No.2.
- **The average performance** scores of each communities are obviously different from each other, except No.1 and No.4.
- **The average age**, which we did not consider before, can help to differentiate the communities together with the average performance, such as No.1 and No.4.

Interesting Found

☆ Community No.1 is composed of Mainland younger method actors.
☆ Community No.2 is composed of action movie actors.
☆ Community No.3 is composed of famous Hong Kong actors.
☆ Community No.4 is composed of Mainland elder method actors.
☆ Community No.5 is composed of Mainland popular actors.

Conclusions

- Louvain Method is a way better than other algorithms to detect actor communities in our research.
- The formation of small communities is mostly based on actors' **personal relationship.**
- The formation of large communities are affect by the **performance, birthplace, movie preference** and **age** of actors actors all together rather than separately, which we did not consider before.

Further Improvement

Firstly, we assumed it as a **non-overlapping** community detection problem. Then we realized that it will make more sense by treating it as an **overlapping** community detection problem, so we tried **Cluster Percolation Method**. It can discover small communities more precisely while fail to discover big communities.

Project ID: 19

NERD'S GALLERY:
COMMUNITIY DETECTION ON PAINTINGS

Jiang Shuyu Liu Ziqian Zhang Ming Zhang Jiacan

NUS National University of Singapore

Introduction

- **Objective:** In this project, we aim to identify styles of different paintings with community detection(CD) algorithms , compare various CD algorithms' performance and take insights into the relationships of genres.
- **Dataset:** We have paintings of 12 styles , and each style contains 20 paintings.
- **Algorithms:** Girvan-Newman, Markov Clustering and Clique Percolation Method.

Details of Data & Methods Used

- **Data:** Each painting is a RGB image labeled with a certain style(abstract art, naive art, etc.).
- **Feature vector**: We use pre-trained CNN to get feature vectors. Feature vectors' size is (12, 1), and the i^{th} value indicates the probability of belonging to the i^{th} style.
- **Dissimilarity matrix**: For each pair of paintings, we compute the Euler distance of their feature vectors as their dissimilarities, set a threshold manually and remove all values that are greater than the threshold.
- **Graph**: We use NetworkX to generate graphs. We compare algorithms' performance on both weighted graphs and unweighted graphs.

Workflow

Project Proposal	Decide our project theme and gather dataset
Compute Dissimilarity	Calculate the dissimilarity matrix
Form Graph	Construct graphs for paintings
Community Detection	Detect communities using GN ,CPM and MCL

Comparison of CD Algorithms

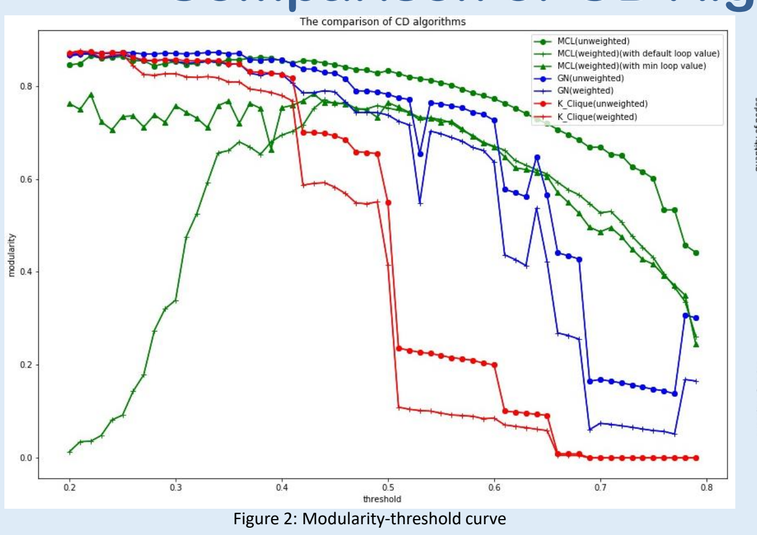


Figure 2: Modularity-threshold curve

- All three algorithms perform better on unweighted graphs than weighted graphs.(see figure 2)
- Markov Clustering is more stable than Girvan-Newman and K-Clique, and there are more fluctuations on GN's curve.(see figure 2)
- Modularity is not a good measurement when the threshold is very low, because we cut off so many edges that only few core communities can survive.
- MCL tends to divide communities strictly and generate many single-node communities, especially on weighted graphs(see figure 3.[1,2,3]). That's why the modularity is low on the MCL(weighted) curve even though the threshold is very low(see figure 3.4).
- We find that if we set the self-loop value to be the minimum in dissimilarity matrix, then MCL performs similarly on weighted graphs like other algorithms.

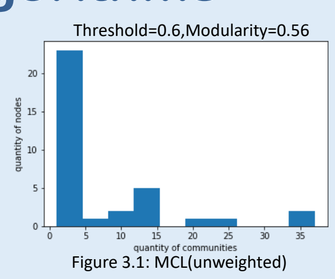


Figure 3.1: MCL(unweighted)

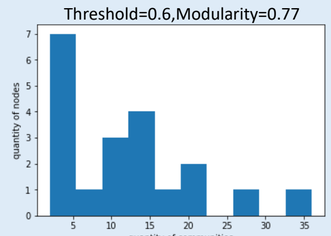


Figure 3.2: MCL(weighted)

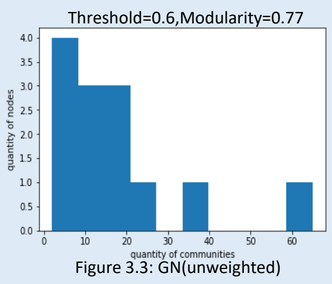


Figure 3.3: GN(unweighted)

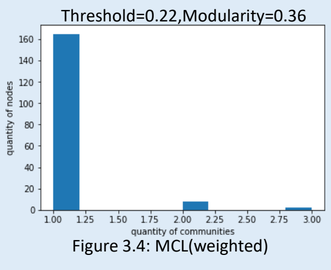


Figure 3.4: MCL(weighted)

Results & Analysis

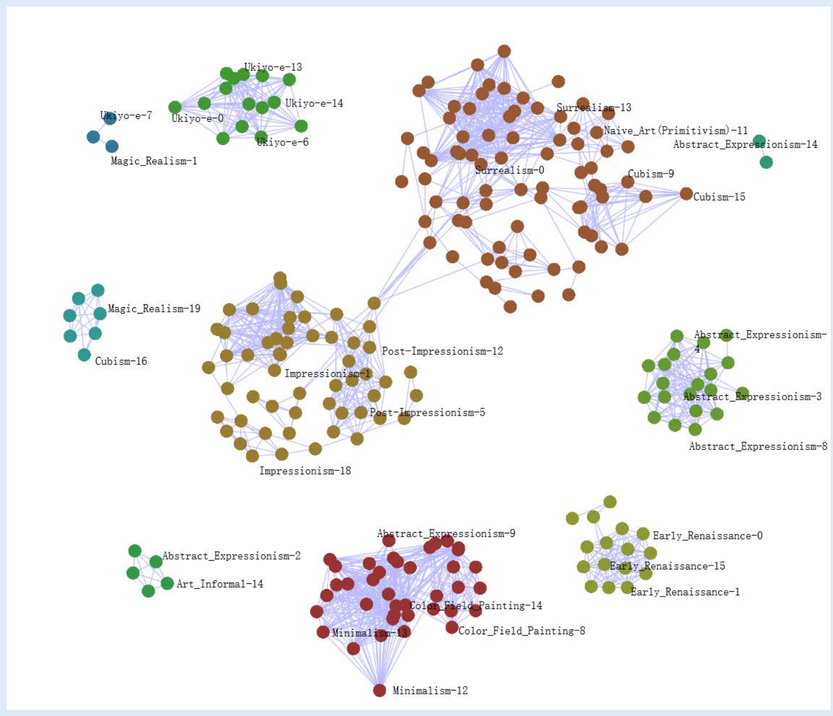


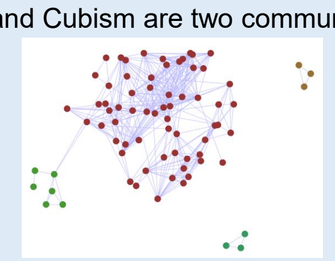
Figure1: Community Detection using Girvan-Newman Algorithm

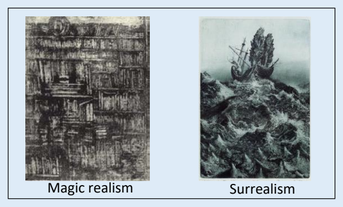
■ Results

- Some communities are closed to others, and some are far from others. It shows that some painting styles are seperated while some are intertwined.
- We have detected the unique or intertwined ones from figure 1.
  Unique: *Early Renaissance, Ukiyo-e*
  Intertwined: *<Cubism, Magic Realism, Naive Art, Surrealism>, and <Abstract Expressionism, Color Field Painting, Minimalism> and <Abstract Expressionism, Art Informal> and <Impressionism, Post-Impressionism*

■ Analysis of intertwined styles

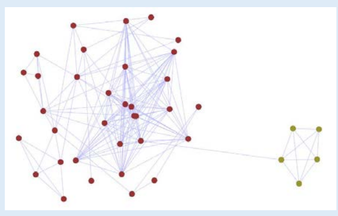
- CNN recognition accuracy is **81.25%** for the 4 painting styles: Cubism, Magic Realism, Naive Art, and Surrealism.
- Naive Art and Cubism are two communities, however, they have many connections.

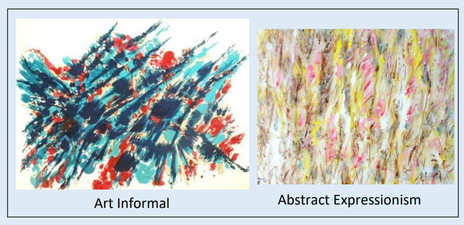




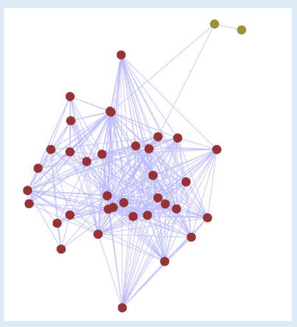
- Color Field painting and Minimalism are different styles while they have something in common with Abstract Expressionism.

- Abstract Expressionism and Art Informal have strong connections. The CNN recognition accuracy is **82.5%**.





- Impressionism and Post-Impressionism also have strong connections.



■ Findings

- Color Field painting and Minimalism are developed from Abstract Expressionism which can be validated on Wikipedia and can support our results.
- Abstract Expressionism and Art Informal are connected tightly in the communities. Actually they are called equally by artists.(*ideelart.com*)

Conclusions & Lessons

Conclusions

- Most of the relationships we find by the communities can be validated by the painting's style development history and real genres. It shows the results of our analysis are good and reasonable.
- The deep learning tool is helpful to do further work, whereas the CNN's recognition accuracy of styles should be improved.

Lessons

- Our project can be used in painting art research.
- Teamwork is very important, and everyone's work is an indispensable part.

SWS3001

Summer 2018

Fashion Miner

Team 6
Project ID:20

Jiang Zihang, Li Wei, Liu Xiao, Wu Jiaying

NUS National University of Singapore

Introduction

- A study based on 58 categories & 2000+ fashion runway photos
- Machine learning: Siamese network and Variational Autoencoder (VAE)
- Comparative analysis: Markov Clustering (MCL) and Girvan-Newman(GN) clustering algorithms
- Records of the influence by different threshold values
- Predict fashion trend

Objective

- Comparative study on two **machine learning models**: Siamese Network and VAE
- Observe the **communities** formed by MCL and GN
- Further analyze their different **clustering habits**.
- **Our plan**: To predict fashion trends & to give suggestions on how enterprises can guide and fascinate fashion lovers more efficiently.

Research Overview

Input data: 180px*270px fashion run-way photos, similar backdrops.
Labels: sub-classes of different brands
We aim to discover:

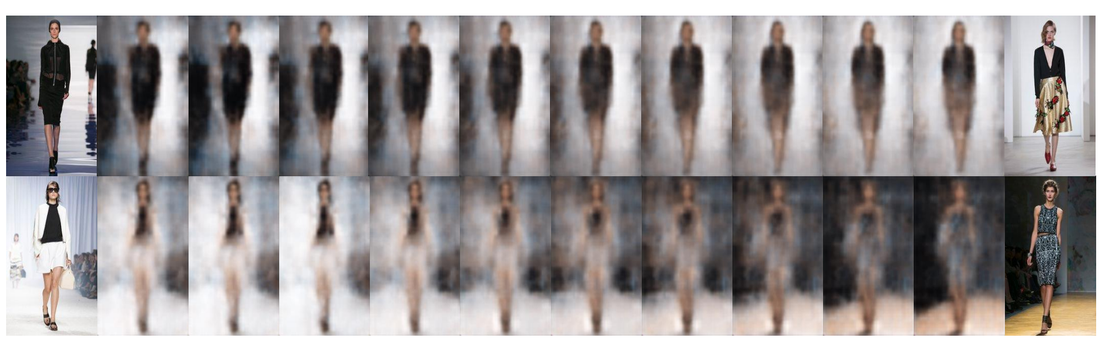
- Influence of the background
- Concise differences between two neural network models and clustering algorithms
- Fashion trends derived from communities

Overall Workflow of Study

Collect Data → Compute Dissimilarity → Form Graphs → Detect Communities → Graphic Visualization → Analysis

Details of Data & Methods Used

- Fashion dataset: 58 categories, 2,000 + fashion runway photos(all are 180px * 270px)
- Networks: Siamese Network, Variational Autoencoder
- Clustering: Girvan-Newman Clustering Algorithm
  Markov Clustering Algorithm
- Backgrounds are removed to boost accuracy



Result Obtained

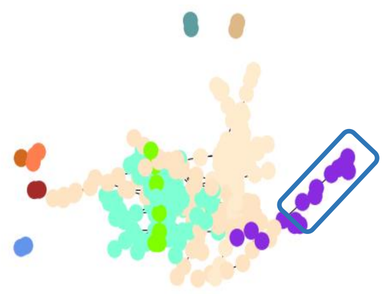


Figure 1 Siamese Network + GN
(58 * 5 dataset, with background)

Figure 3. Siamese Network + GN
(21 * 10 dataset, w/o background)

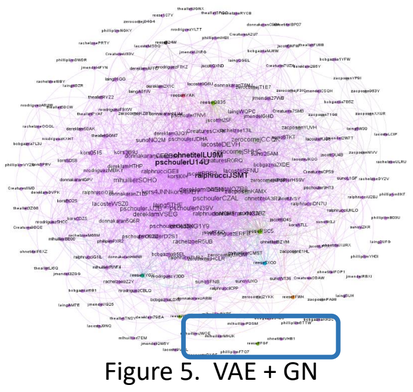


Figure 5. VAE + GN
(21 * 10 dataset, w/o background)

Figure 2.

Photos in Figure 2 are part of one cluster, they share little in common. In this case model doesn't work well.

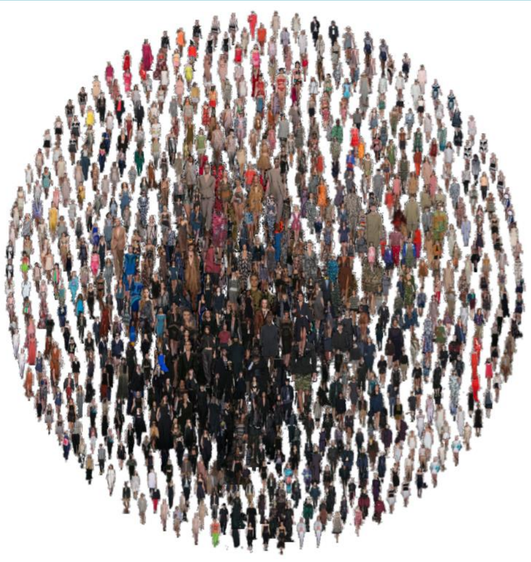
Figure 4.

As Figure 4 shows, we found a special community which is only formed by long dresses.

Images within the same cluster and near each other look similar.

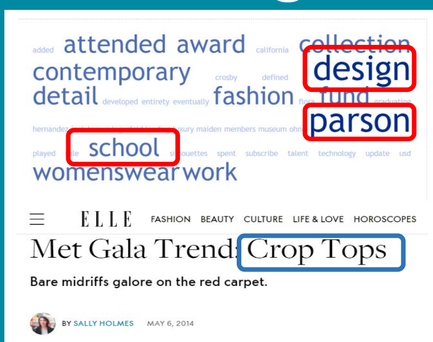
Global Trend

Black clothes tend to **cluster in the center**. So we speculate that although all colors and styles of fashion designs are entering the runway, **black is a never-fading fashion**.



Conclusive Graph (2,000+ images)

Interesting Discoveries



★ Parson School of Design goes viral!
★ Hub vertex's interesting feature: Fashion trend leader
★ Unprecedented community of long dresses (see Figure 4)

Does background affect output?

Sure it does!
To eliminate the impact, we developed a semi-supervised background deletion algorithm

Conclusions & Further Focus

- Classification ability : VAE > Siamese network
- For VAE, MCL performs better
  For Siamese Network, GN performs better
- Background does influence machine learning & classification
- Further work: More in-depth studies on the four combinations : which one is most accurate & time-efficient?

SWS3001

Summer 2018

Project ID 21

Team ID 07

Music Genre Analysis

Hu Shuang, Zhang Min, Zhi Zhuo, Zhou Chang

National University of Singapore

1 Introduction

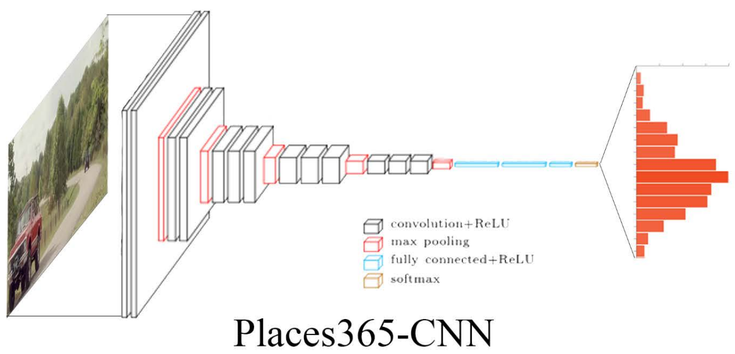
Our project intends to find the relationship between music genre and music video scenes. For example, it's very often that light music's MV is shot at comfortable places, and rock music's MV is shot at noisy and uncomfortable places.

2 Overview

We collect 173 music videos randomly, and take screenshots according to fixed time interval. In order to compare the dissimilarity between any two videos in turn, we use Places365-CNN to recognize music video's scene attributes. Then, we set up the dissimilarity matrix and apply different CD algorithms. Finally, we select the CD algorithm(GN) with the best performance and visualize the communities we find.

3 Process

Dissimilarity Calculation



Places365-CNN

top 9 attributes →

- Natural light
- Open area
- Trees
- Man-made
- Foliage
- Vegetation
- No horizon
- driving

→ $I(im_i) = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix}$

The Places365-CNN outputs a 102 dimensional attribute probability distribution vector, where we set the top 9 largest elements of the vector to 1, and let the rest of it be 0, and then get the vector $I(im_i)$.

→ $V(v_i) = I(im_1) + I(im_2) + \cdots\cdots I(im_n) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 0 \end{bmatrix}$ →

Dissimilarity matrix$= \begin{bmatrix} 0 & \cdots & 1 - Cos\,(V(v_1), V(v_n)) \\ \vdots & \ddots & \vdots \\ 1 - Cos\,(V(v_n), V(v_1)) & \cdots & 0 \end{bmatrix}$

+ means the logical or operation and n means the total amount of the images the video v_i contains.

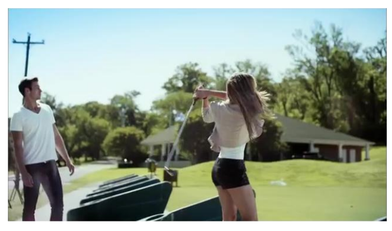
Define the value $1\text{-}Cos\,(V(v_i), V(v_j))$ as the dissimilarity of any 2 videos, and then create dissimilarity matrix to apply community algorithms.

A Showcase for CNN Predictions

man-made, natural light, grass, open area, sunny, competing, sports, no horizon, cloth

natural light, open area, boating, far-away horizon, natural, ocean, swimming, sunny, diving

no horizon, enclosed area, man-made, cloth, indoor lighting, working, stressful, vertical components, congregating



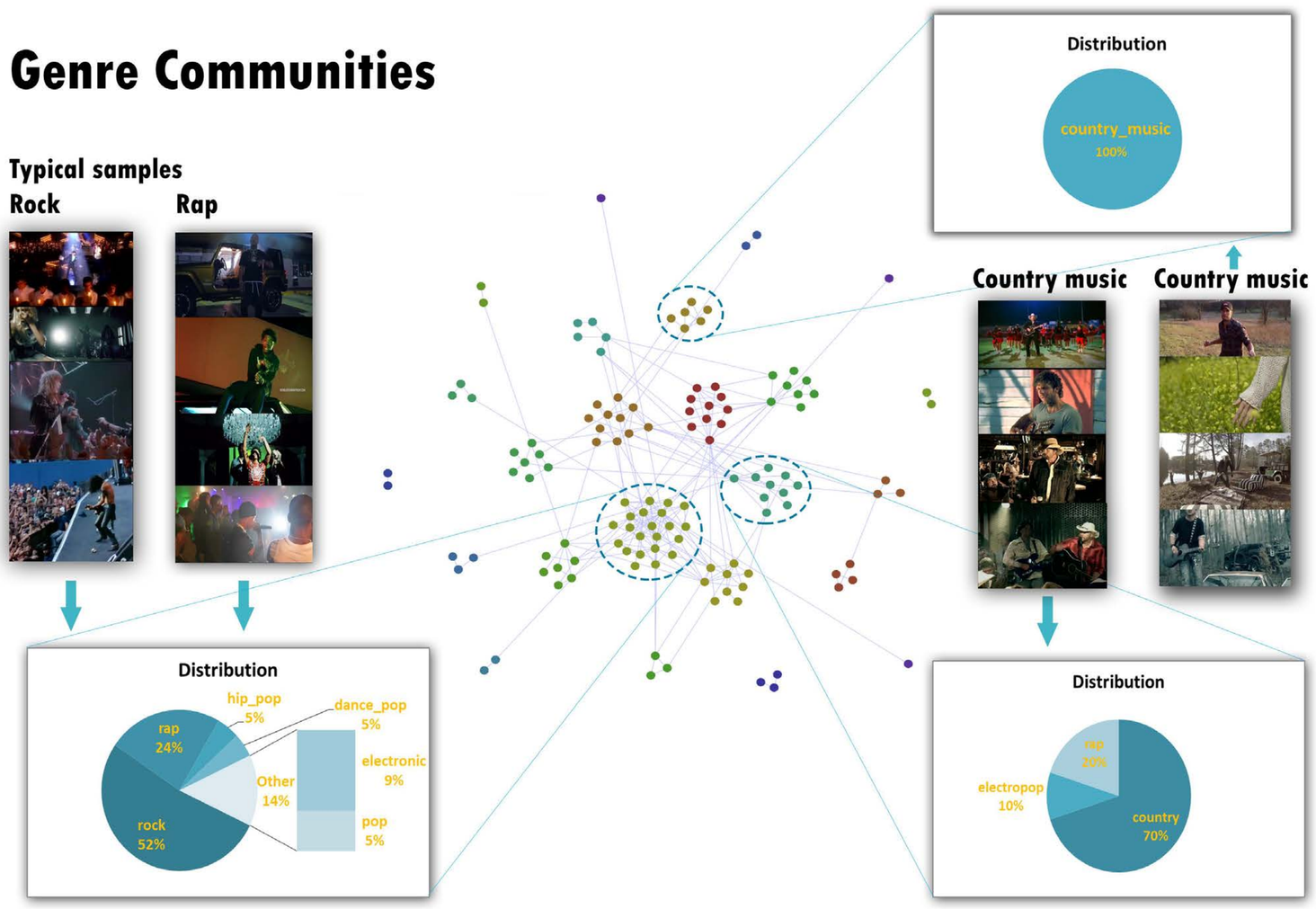
natural light, open area, man-made, sunny, trees, grass, foliage, vegetation, cloth

man-made, open area, natural light, no horizon, dry, far-away horizon, transporting, sunny, indoor lighting

no horizon, enclosed area, indoor lighting, man-made, cloth, glossy, competing, spectating, congregating

4 CD Results

Genre Communities



Interesting Findings

- Country music can be mainly divided into two types according to the filming locations. One is about typical countryside sceneries, and the other one is about indoor activity.
- As for rock and rap music, noisy, dark and stressful places are often the optimal choices.
- Pop and electronic music distribute randomly, because of their comparatively loose genre concept and the fact that it can be mixed with many other music genre.

ACTIONABLE INSIGHTS

The results may be useful in the following fields:

- Classify the most probable music genre according to the music video scene attributes.
- Recommend shooting locations and scene attributes to music video shooter.

SWS3001

Summer 2018

Team 22

The Next City

LI BAIYANG, SUN WENJING
XU ZIQIANG, YUAN XINJIE

NUS National University of Singapore

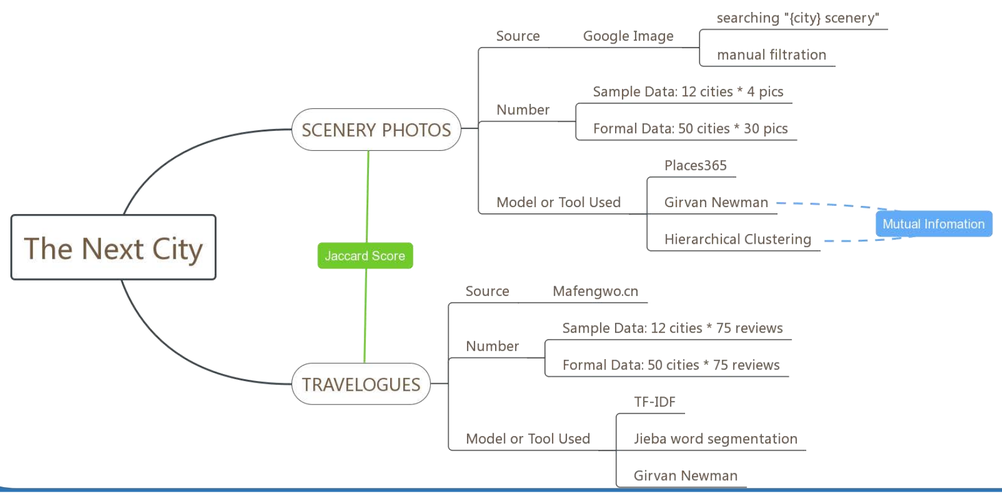
Introduction

- City Similarity
- Community Detection
- Photo vs. Travelogue
- Objectivity vs. Subjectivity
- Find the next travel destination

Objective

- To identify and analyze communities in popular cities
- Scope of project: Travel Reference

Overview of Your Study



Overall Workflow of Study

Compute Dissimilarity	Dissimilarity matrix
Form Graph	Sparse graph
Community Detection	Communities
Using different algorithms	GN&Hierarchical Clustering
Analysis and Visualization	Facts & Actionable insights

Rationale for Comparison

Mutual Information

$$\mathrm{I}(P^a, P^b) = \sum_{i=1}^{k_a}\sum_{j=1}^{k_b} \frac{n_{ij}^{ab}}{n} \log\left(\frac{\frac{n_{ij}^{ab}}{n}}{\frac{n_i^a}{n} \times \frac{n_j^b}{n}}\right)$$

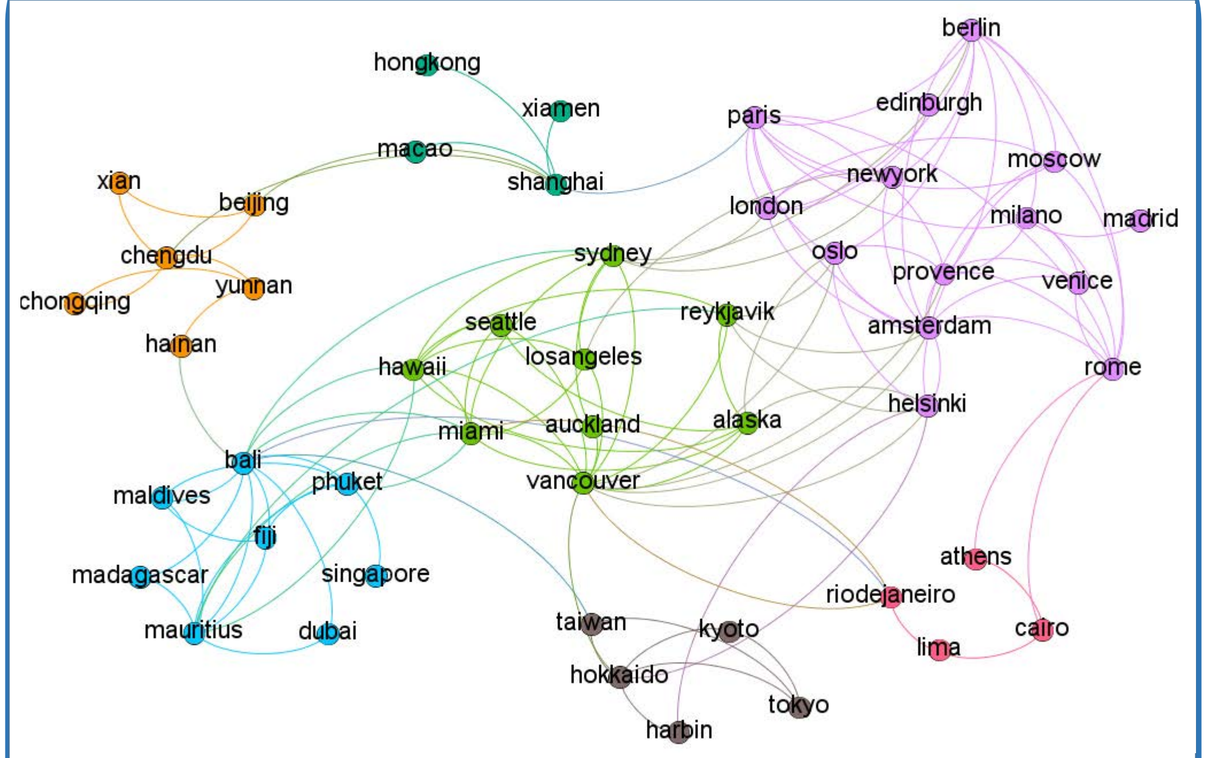
- Comparing the similarity between Girvan Newman's Algorithm and Hierarchical Clustering Algorithm
- After normalization, the higher the score is , the more similar.

Jaccard similarity coefficient

$$\mathrm{J}(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- Using Jaccard score to describe the similarity between the photo and travelogue community.
- Comparing two normalized matrix by each line(city), the higher the score is, the more the similarity is, the less the subjectivity influence the evaluation of the city.

Results Obtained



Travelogue-based communities using Girvan Newman Algorithm

Photo-based communities using Girvan Newman Algorithm

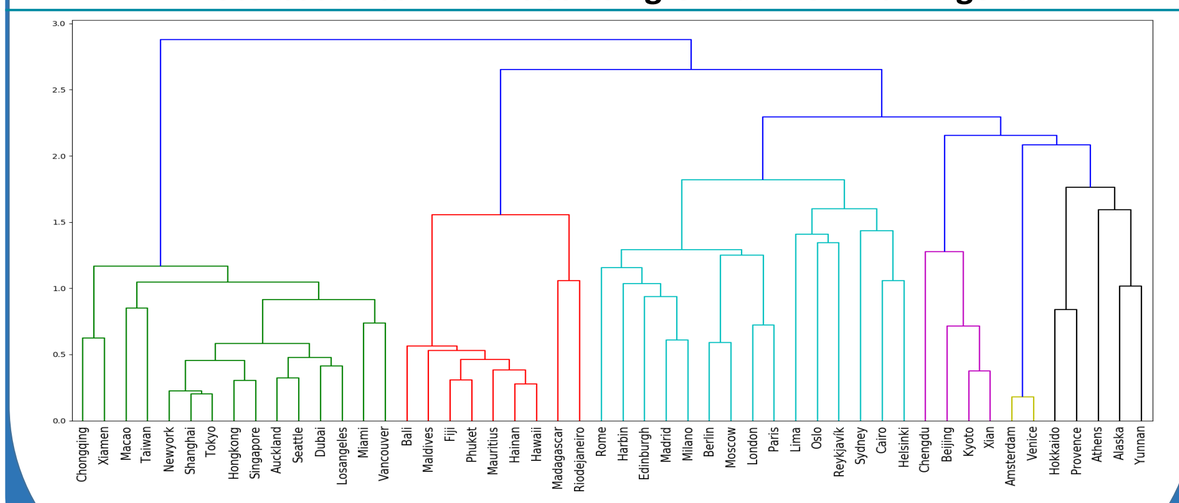


Photo-based communities using Hierarchical Clustering Algorithm

Conclusions, Lessons Learnt

Objectivity vs. Subjectivity

- Pace (slow & holiday vs. fast & capital)
- Culture(China vs. Foreign country)
- Location(Inland vs. Coastal)
- History(Old vs. Young)
- Temperature(Cold vs. Hot)

Girvan Newman vs. Hierarchical Clustering

- NMI(Normalized Mutual Information) is above 0.8
  -> quite similar

Beauty? Delicacy ? Fairyland?
For Further Amazing Findings,
Please Scan the QR Code

Browser WeChat

SWS3001

Team 23

Relationship Discovery of Wikipedia Entries

Hu Yuke, Jiang Hao, Wang Yanbo, Zhao Yuxuan

NUS National University of Singapore

Introduction

What: Relation between Wikipedia entries
Why: We wonder if the entries form "groups" ?
How: Judge the relation by hyperlink in one entry linked to another entry

Conclusions

I. Words in the same category are more likely to form communities.
II. The aggregation degree of each category varies depending on the nature of the category.
III. The result of clustering is highly dependent on the algorithm. Different algorithm results in very different communities. (modifying the structure of the graph may help)

Lessons learnt

I. Whether the edges are directed or undirected is quite important in CD. he failure at first had something to do with our approach in forming graphs.
II. It is quite important to have a clear overflow to follow, otherwise we may get lost in the process.

Entrance

Hope to discover

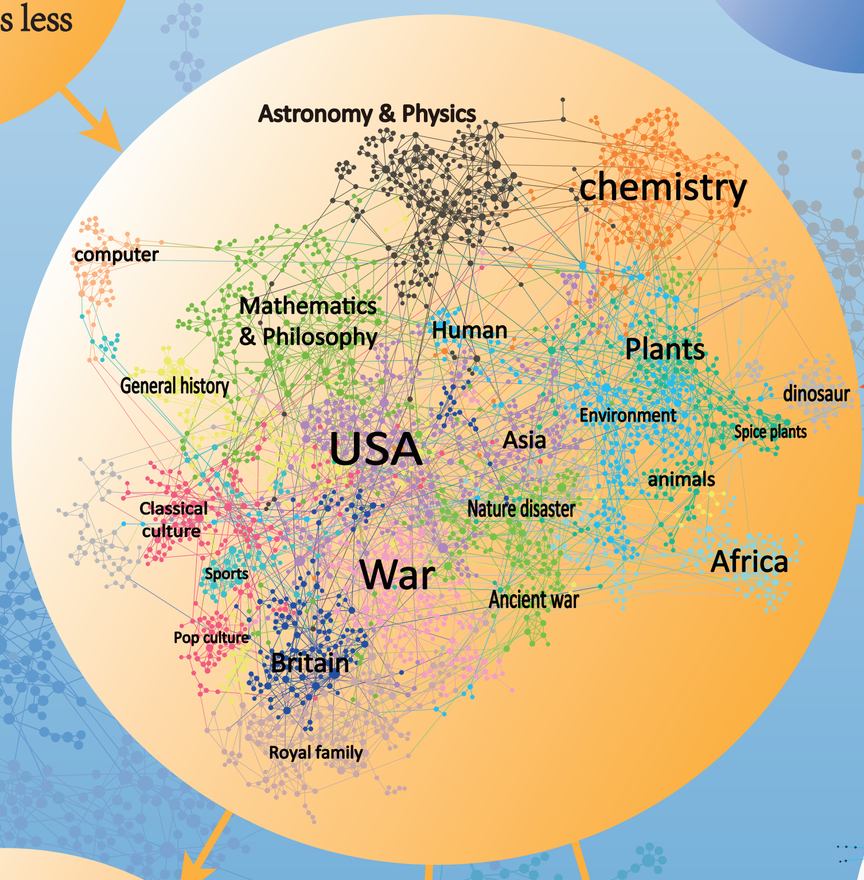
Entries in wikipedia can be clustered according to its categories

To identify and analyze communities in wikipedia entries

Filter

From the dataset which contains about 4 thousand entries, we select nodes whose degree is less than 100.

Using part of hyperlinks data downloaded from Stanford Large Network Dataset Collection



Results

The entries in the categories like physics and chemistry tend to form obvious clusters because they are closely related to each other. However, the historically related terms are often scattered due to the large space-time span.

Results

Interdisciplinary vocabulary often appears in overlapping areas of communities.

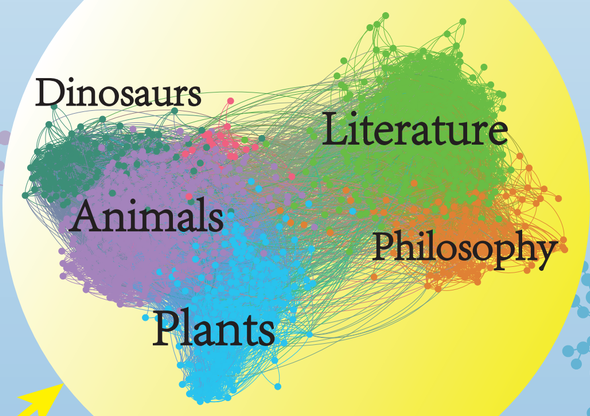
Results

Terms in the same category tend to form a community due to hyperlinks between them.

We choose two clusters (biology and humanities) from original dataset to detect communities. There are obvious boundaries between clusters.

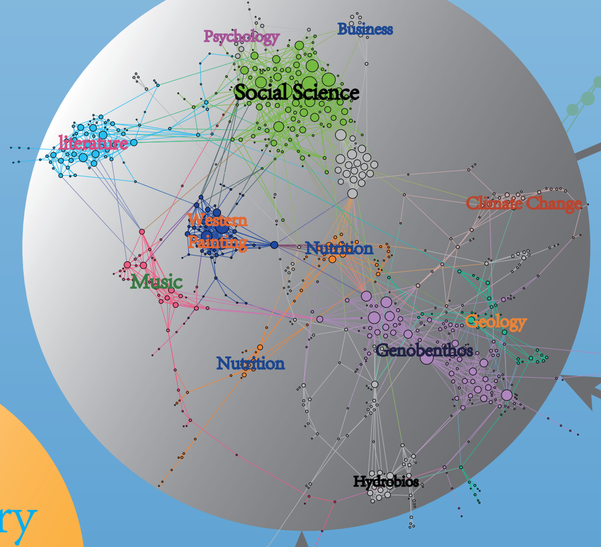
Result

Using Louvain algorithm we form 5 clusters which make sense in some extent.



Result

The results of these two method are both significant but not as good as the first one. They are also clustered according to its categories no matter from the hyperlink aspect or text aspect.



We also tried Text Similarity Analysis using TF-IDF. The similarity between entries' content formed a dissimilar matrix. Then using GN to detect the communities.

Filter

we select the edges whose similarity is great than 0.1.

Overall Workflow of Study

Fetch data

Links between entries

Form graph

Sparse graph

Community detection

Communities

Analysis and visualization

Facts & actionable insights

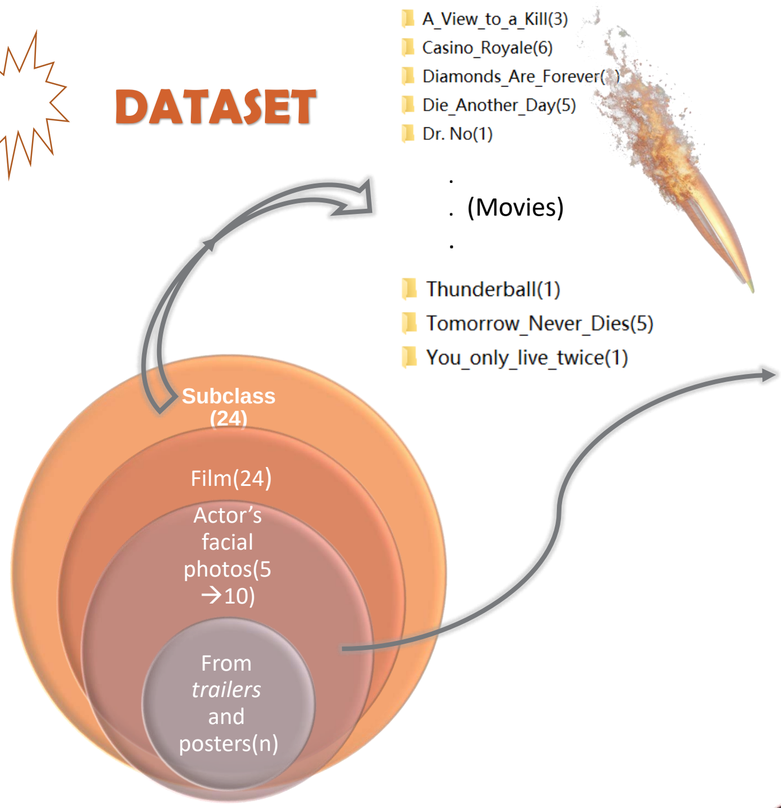
SWS3001 G10
Project-ID:24

Target: James Bond

INTRODUCTION

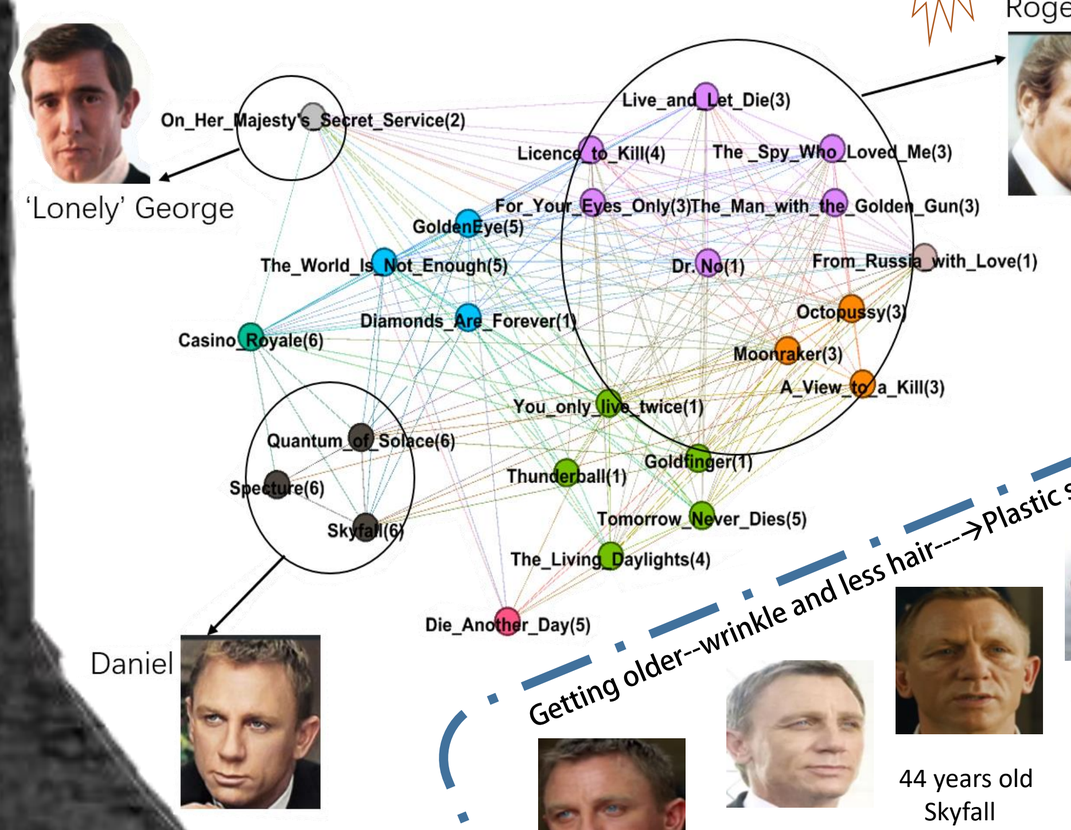
- TOPIC: '007:James Bond' series films
- SPECIFIC: 24 films including 6 actors
- WHY: The 25th film will be on show in 2019.

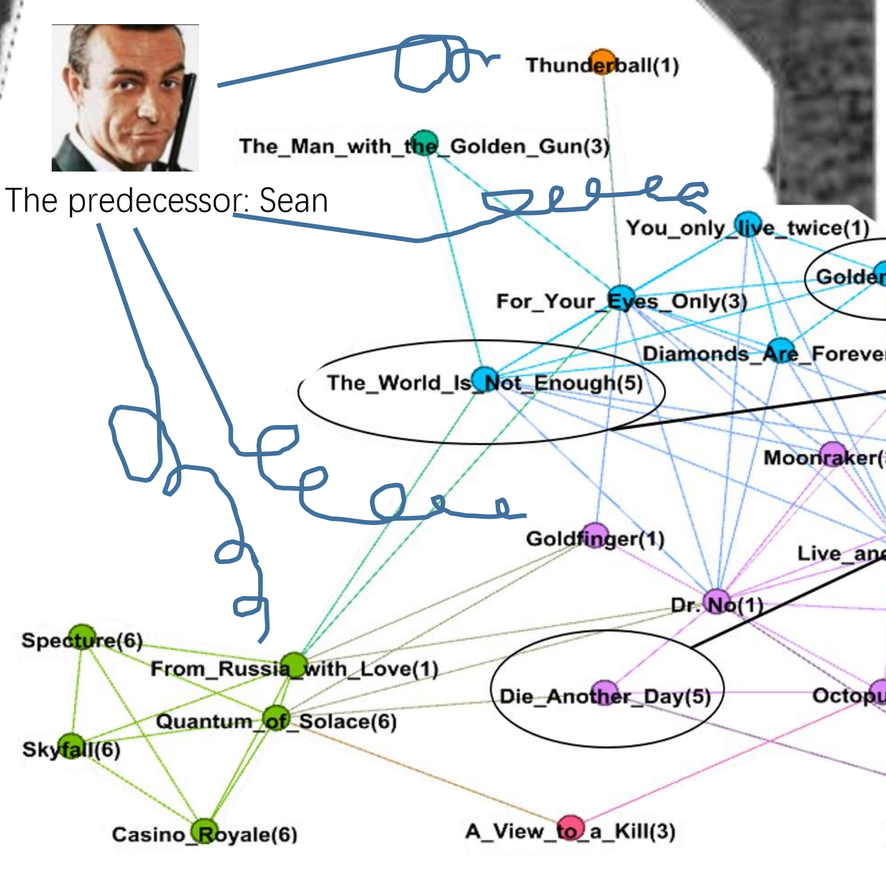
DATASET



IMMORTAL 007

Community using Complete-Linkage Algorithm





Community using MCL Algorithm

Designed by:

CHEN RUIWU
CHEN YIFAN
FENG LIKAI
YE XIANGYU

SWS3001

Summer 2018

The Influence of CNN Model on CD Algorithm

Team Group 11
Project 25
Liu Jiayi Xu Dongli Yang Xinzhao Zhang Zeyu

NUS National University of Singapore

Introduction

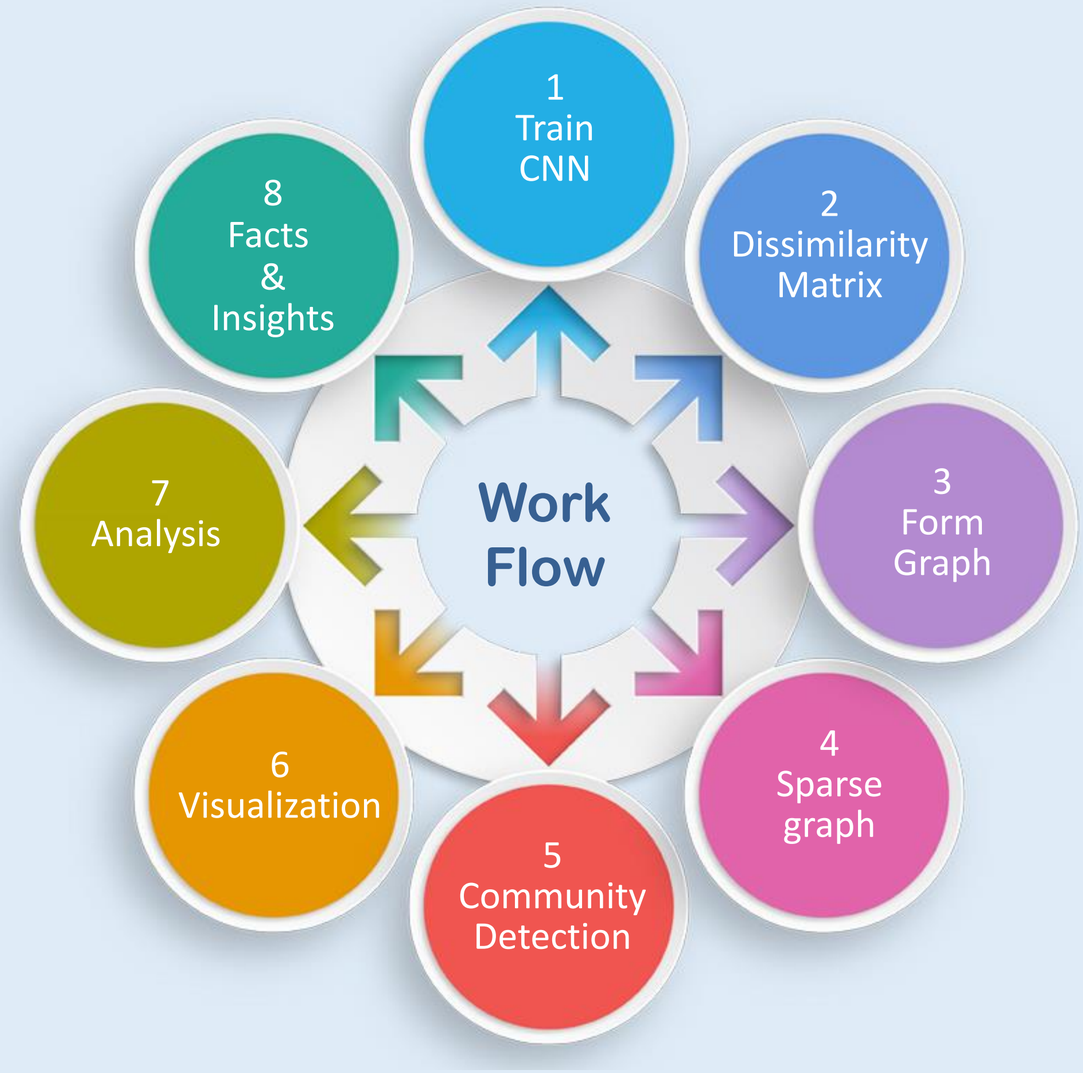
- Mining potentially unknown information on data sets.
- Training CNN → similarity matrix→ generate graph→ community discovery algorithm→analyze.
- Results obtained.
- Analyze&Conclusion.

Objective

We used two different algorithms to analyze relationships and features between the different classes in the fruit dataset. In the process of analysis, verify the pre-guess and mine unknown information from it.

Overview of Your Study

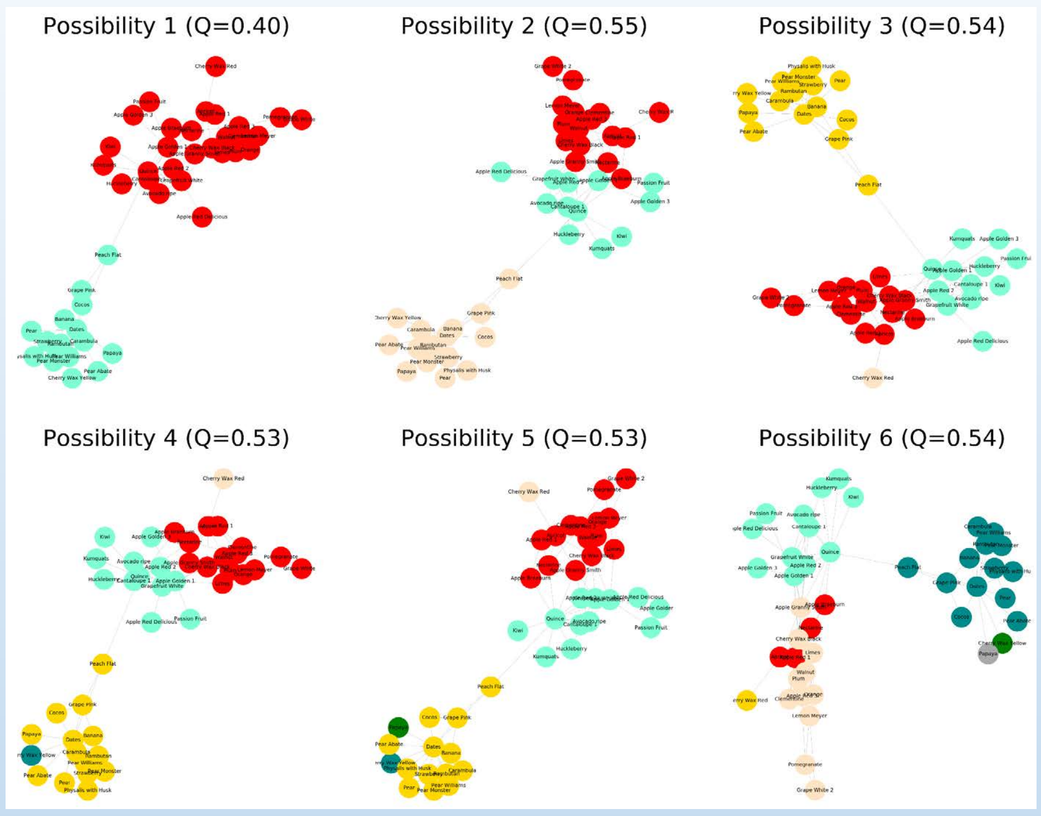
- ***Input Data:*** A dataset with 75 fruits and 450 images
- ***The labels:*** The species of fruit
- ***Outcome :*** Classified by fruit color, fruit color brightness, fruit shape, fruit size, All unknown discoveries.



Details of Data & Methods Used

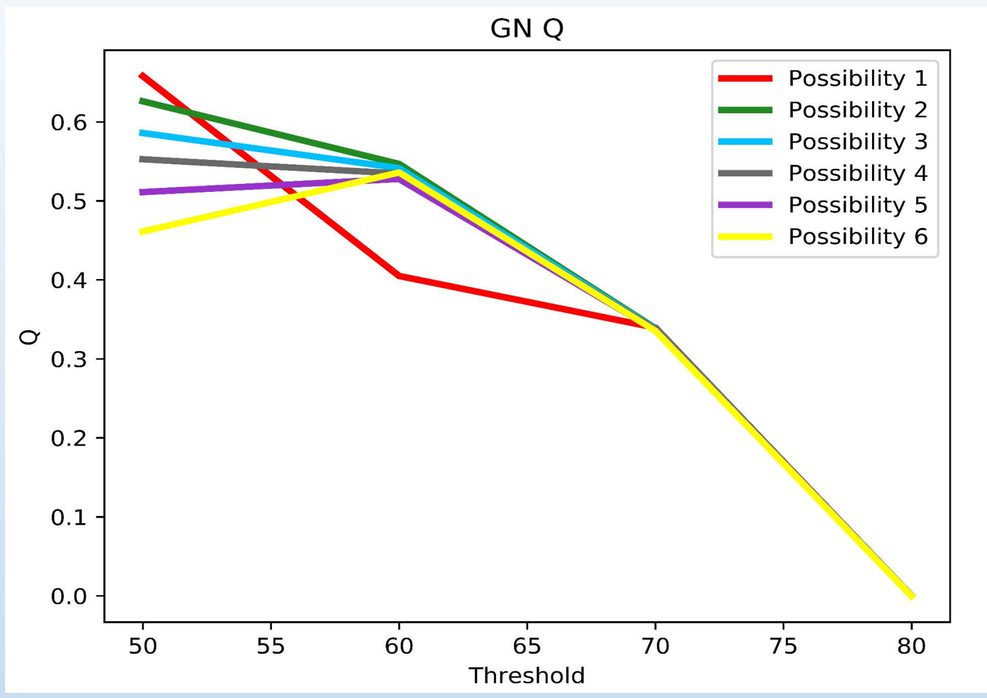
- ✓ ***Fruits dataset：*** The data comes from the kaggle dataset. The original dataset is very large. We manually screened according to the different angles of the fruit.
- ✓ ***Methods：*** Convolutional Neural Network(CNN), Siamese Network, Clique Percolation Method(CPM), Girvan Newman Algorithm(GN)

Results Obtained



Analysis： First, we use the GN method for CD analysis. We analyze distribution of dissimilarity score by histogram and select the appropriate threshold for visual analysis. We find that the clustering effect is most obvious when the threshold is equal to 70(Figure,1).

More Results



Analysis： When the value of Q is the largest, the network is ideally divided. The larger the Q value, the higher the accuracy of the community structure of the network partition. Therefore, after comparison analysis, we choose the threshold equal to 60.

Conclusions, Lessons Learnt

- We use mathematical statistics to find that as long as the colors of the fruits are not the same, no matter which method is used, they will not be assigned to the same class. We can‘t judge the quality of the similarity matrix of CNN training. We can only judge whether the results are suitable by selecting the threshold.
- So there are still many problems in this project that deserve to be studied and we will work on.
- Through this project, we have a deeper understanding of the steps of scientific research and have a more rigorous scientific attitude.