

Introduction

We

- **Aim** to discover new knowledge from data set
- **Restrict** the study to ingredient only
- **Input** recipe data (json form) with label of country
- **Build** a graph basing on weight distribution
- **Apply** 3 different CD algorithms
- **Output** different communities of ingredients
- **Analyze** the communities and **find** new insights

SWS3001

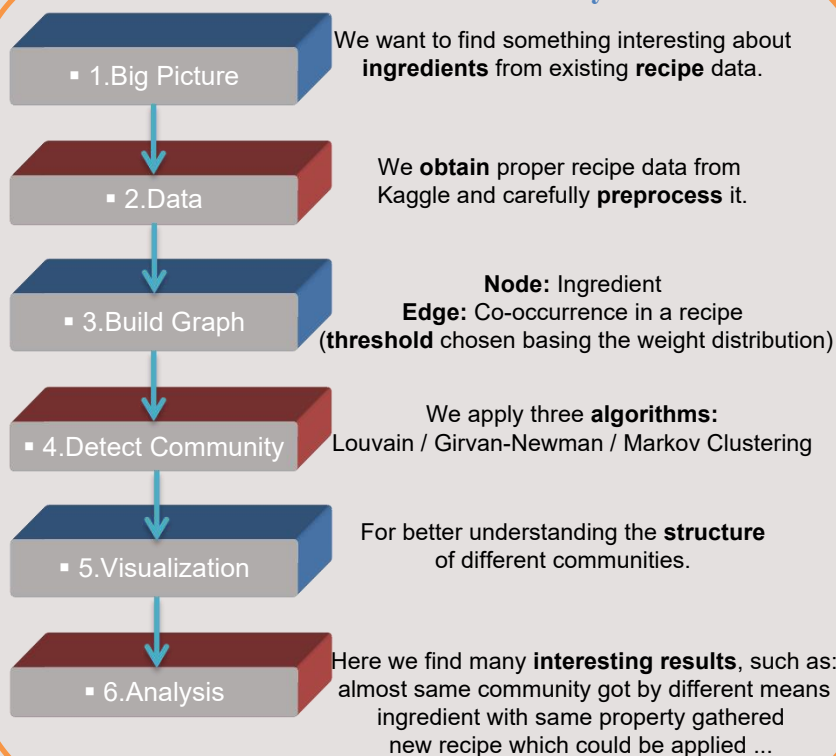
Summer 2021

Team G01

Ingredients Network

Chen Zikang, He Jiyao, Liang Junhao, Zhang Jierui

Overall Workflow of Study



What's the Future Plan?

- Apply some data mining algorithms, and compare them with CD
- Analyze the algorithms from the mechanism, test and verify result
- Enlarge the data set to see if there are more interesting things
- Is our new insight really actionable? Try the new cuisine to find out
- Build the website of our group, show our findings to the world

What We've Learnt?

- The whole process of CD
- Many useful algorithms in CD
- Dealing with little problems matters
- How to find new information
- How to make strict analysis
- How to mine potential knowledge
- Also, teamwork is very important

What's the Results?

1. The community partition of weighted graph is completely the **same** with unweighted one, using **Louvain algorithm**
2. Applying 3 algorithms, we get nearly **same partition**—7 communities, each containing similar ingredients
3. Some of the communities generated which are not in data set, however, are actually **in real world**:
tumeric, vegetable stock, spinach, Masala, ..., and sweet potatoes make an **Indian food**
broiler-fryer chicken and zesty Italian dressing make an **Italian food**
greek yogurt, lemon curd and raspberries make a **Greek food** ...
4. Some **potential** delicious cuisines are discovered, which may lead the trend:
ginger, garlic cloves, radishes, cabbage, rice vinegar, carrots, as **pickles**
5. The ingredients of some communities are actually the (almost) **same** thing:
combination of nonfat milk, evaporated milk, whole milk, low fat milk, milk
6. Some data mining algorithms don't perform well as CD algorithms

