Cuisine Analysis using Monte Carlo Clustering

s1999555

University of Edinburgh

May 21, 2020

Abstract

In this project the distinctiveness of different cuisines in terms of ingredients and cultural origin is quantified in terms of network clusters. Recipes were extracted using a commercial database and a network was constructed on the basis of a sub-sample using similarity of ingredients to determine edges. Clusters were approximated with a greedy modularity algorithm. This process was conducted repeatedly for small samples and the distribution cluster properties used for inferences. Asian and, to a lesser extent, American cuisine showed consistently high distinctiveness, although the former was confounded by differences between baking and cooking. This project presents a method to make inferences on large populations using a method based on small networks and therefore low working-memory demand.

1 Introduction

What did you have for lunch today? Or is it still early and you are contemplating what you feel like today. If you are eating out, you might have quite a selection to choose from. However, the first choice you will need to make is which type of cuisine might fit your appetite. Cuisine, often associated with a geographical region, is a set of typical ingredients and preparation techniques and therefore helps to classify food. But how distinct are these different cuisines in fact? To answer this question, this project focuses on ingredients.
In this report cuisines will be assumed to complete, meaning each dish is part of a cuisine, and disjunct at the most general level, meaning that one dish may not belong to two different cuisines of the highest-level of classification, in this project it may only belong to "American" or "European" or "Asian" or "Mexican" cuisine. I will approach the question of cuisine distinctiveness based on the ingredients used in a dish. Particularly, I will construct a network of dishes based on their shared ingredients and analyze this network in terms of communities and whether those map to culinary styles. Following the assumptions of cuisine properties these communities should be disjunct and cover all recipes.

2 Experiments

The information on dishes, their ingredients and the cuisine they belong to was accessed from Spoonacular, a database of 330,000 recipes [2]. Dishes, which did not include any of the cuisine tags corresponding to the cuisines above were excluded.
Cuisine tags were organized in a hierarchical structure, with varying depth for different dishes. For example "Greek" dishes were also marked as "European" and "Eastern European" and "Mediterranean", while "American" dishes did not have any additional tags.
The recipes, which were included, were each assigned a node each. Edges would exist between two nodes if the number of ingredients two recipes shared would exceed threshold t. In order to detect communities, which are both disjunct and comprehensive, a greedy modularity maximization algorithm will be applied. The distinctiveness of a cuisine is assessed by whether it is predominant in a community, meaning a large proportion of the community is classified as that cuisine. It is possible that there are several clusters in which a cuisine covers a large proportion of dishes.

2.1 First experiment

This first experiment was conducted on a random sample of 1000 recipes at a threshold of t=6, from which a graph of 1000 nodes and 944 edges was constructed. The largest connected component of this graph, consisting of 319 nodes and 942 edges, was subjected to a cluster analysis using a greedy modularity maximization algorithm. [1] The initial clustering revealed 5 clusters of size 84, 52, 50, 48, 13 and three others with less than 10 members. In the second and third largest clusters American dishes were the most frequent. Figure 1 shows some characteristics in respect to the largest connected component of the graph as well as each of its communities. Firstly, the sample in unbalanced as American and European cuisines are over-represented and South American and African are extremely under-represented. Community 1 and 3 are clearly dominated by American cuisine. An initial inspection of the ingredients of community 1 suggests that this community consists of baking dishes. Community 2 is mostly dominated by Mexican dishes but includes a number of European and European dishes. Community 4 is mostly European, while community 5 is dominated by Asian dishes. Additionally, only few Asian dishes occur in other cluster, suggesting, that the Asian culinary category is matching particularly closely with this cluster. Conversely communities with a majority of American dishes also include a number of European dishes and vice versa. Most Mexican dishes are clustered in one community (2), but this group includes roughly a third of non-Mexican dishes as well, making it less distinct.

2.2 Balanced Experiment

An initial analysis of the data showed a bias towards American and European dishes, with two thirds of the recipes belonging to American or European cuisine. Therefore the sampling was refined in order to include a more equal distribution. A random sample of 250 each was drawn from the distribution of American, European, Asian and Mexican dishes respectively. This combined to a balanced sample of 1000 recipes.
A network was constructed in the same manner as mentioned in the first experiment and subjected to

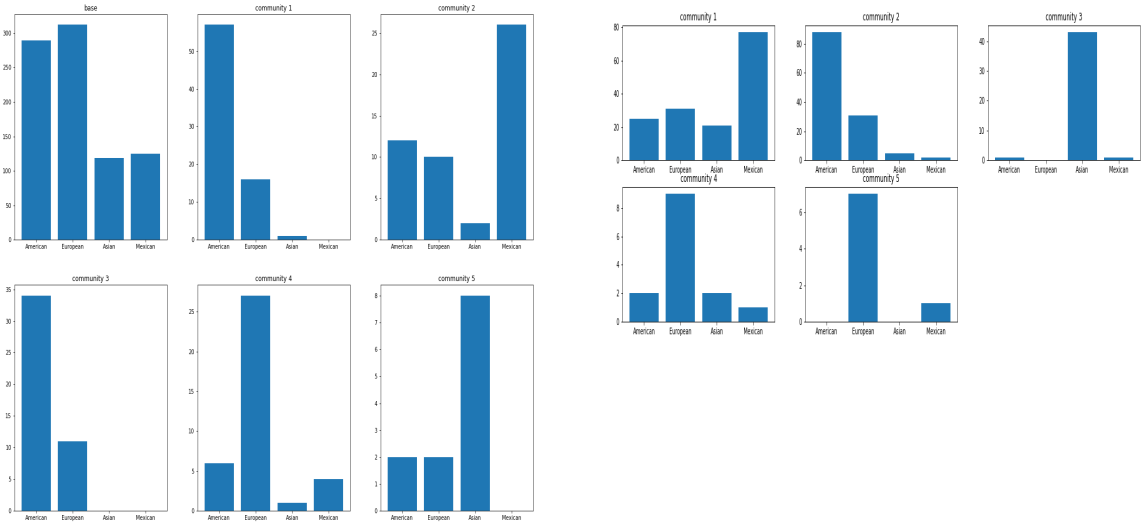

Figure 1: Right: Distribution of Cuisines in the largest connected components and its largest communities. Right: Distribution of Cuisines for the largest clusters based on a balanced sample

greedy modularity maximization clustering. The clustering revealed 7 communities of sizes 154, 126, 45, 14 and three clusters of less than 10 nodes. In these communities (See Figure 1, right plot) there was each a majority for Mexican (proportion of 50%; community 1), American (proportion of 69.8%; community 2), Asian (proportion of 95.6%; community 3) and two European communities (highest proportion of 64.3% in community 4). Typical baking ingredients, such as eggs, flour sugar and butter, were found in community 1 and 3.

2.3 Small Network Sampling

A crucially limiting factor within this project for analyzing a larger network was working memory. However basing the network construction and clustering steps on a sub-network, as I have done in both experiments discussed above, produces different results depending on the sample, severely limiting the ability to infer properties of the population. This is why I will present a distribution of cluster properties. In order to generate this data, the balanced experiment will be run 100 times using different random samples from the original data-set. The highest proportion of each cuisine in any community (which consists of at least 10 nodes) was recorded for each experiment and will thereby indicate the distinctiveness of a cuisine within a network. This method increases the computational complexity of the experiment linearly by the number of iterations, but adds only a minute memory demand, as only network statistics rather than the entire network are saved.
100 random balanced samples of size 1000 of recipes were drawn from the data and subjected to the same method as described in 2.2..

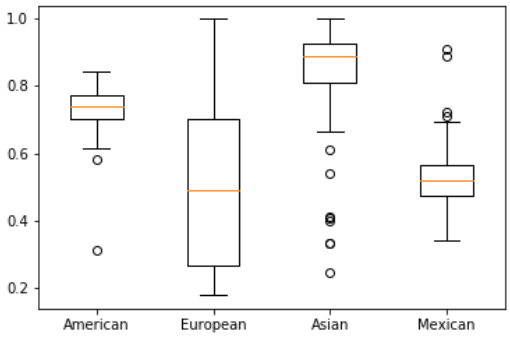

Figure 2: Boxplot of the distribution of cuisine distinctiveness based on networks built on 100 randomly drawn distributions. Cuisine distinctiveness was measured as the highest proportion of each cuisine in any community of a network

The distribution shows that American cuisines and particularly Asian cuisines had some community in which that cuisine consistently occupied a large majority in for most networks. While the distinctiveness of European was mediocre and highly inconsistent across networks constructed from different samples. Mexican cuisine made up 50% of dishes in the community it had the largest majority in, but shows more consistency in this finding across different small networks than European cuisine.

3 Discussion

The initial experiment revealed community, which clearly have substantially different distributions of dishes in respect to cuisines. It also revealed a bias in the number of dishes of each cuisine. There were a large number of European and American dishes and fewer Asian or Mexican dishes. The follow-up experiment with balanced data showed similar trends as the first experiment, but with a clearer distinctiveness of Asian dishes, as these were almost exclusively clustered in one community. Repeated iterations of the same procedure allowed for a more robust estimation of the extent to which clustering encapsulates the concept of cuisines. Reinforcing the initial observations that Asian dishes and to a lesser extent American dishes are highly distinct from others. These finding indicate that cuisine is a valid concept in categorizing food, but showcases that some cuisines, as the concept of Asian cuisine coincides with a more distinct set ingredients than others.
However, this projects' results remain subject to several limitations. Even though the bias in the number of recipes from each cuisine was controlled for, this lack of balance in numbers may indicate a more pervasive bias. Concretely, if the source of the data is biased in favour of American and European cuisine, it might also mis-represent ingredients in other cuisines by favoring ingredients which are more accessible and more common to Western cooking, even if they are not typical for the respective dish. This experiment is therefore likely to underestimate cuisine distinctness. Furthermore, preparation style (e.g. baking vs cooking) was reflected in the ingredients and affected clustering to some extent, potentially inflating the distinctiveness of American cuisine. Moreover, the selection of the cuisine categories does not represent a complete selection by any means.
The statistical checks may help to make more confident statements about the results of a smaller network, it is not clear how this relates to larger networks, with otherwise similar design. Also, this statistical method does not model the dependency between community sizes and their respective cuisine profile. Still, as the method has only low linear memory requirements in respect to the number of samples drawn, a more careful modelling procedure following a similar multiple-sample design could enable more powerful analysis in memory-constrained systems.

References

[1] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, dec 2004.

[2] Spoonacular. spoonacular recipe and food API. https://spoonacular.com/food-api. Online; accessed 2019-11-11.