Flavours of the World: How Russian cuisine Fits Within a Network of Recipes, Ingredients and Compounds from Around the Globe

Alden Bong (22255844)
Dylan Carpenter (21982288)
Luke Joshua Carpenter (22110274)

Semester 1, 2020

CITS4403 Computational Modeling
The University of Western Australia

Abstract
In this contribution, we attempt to recreate some key findings from the 2011 paper "Flavor network and the principles of food pairing" which uses a flavour network approach to food science. This contribution predominantly looks at Eastern European recipes. Three separate datasets were combined and mapped using 'FooDB', an ingredient and compound database. Our analysis showed that no cuisine abides by the proposed 'food pairing hypothesis'. The analysis also showed that ingredient frequency exhibits characteristics of scale free networks.

Table of Contents

Introduction

It has been observed that different cuisines tend to only utilise a small number of staple ingredients for the majority of their cooking. This cultural predisposition toward certain ingredients may produce interesting and consistent patterns in the underlying networks. This could be used to predict how to bring a nostalgic taste of a favoured cuisine to new ingredient combinations. This paper analyses attributes of single ingredients, ingredient pairs and triplets. We also analyse the food pairing principle, which states that ingredients with a high number of shared flavour compounds tend to be paired together.

Past Work

Similar to our work, the 2011 paper by Ahn, Ahnert, Bagrow and Barabasi creates a flavour network based on mappings between recipes, ingredients and compounds. This network was used to determine a wide variety of cuisines' propensity to combine ingredients with shared compounds. This network contained 281 ingredients and 1,021 compounds. The paper concluded that East Asian cuisines do not support the food pairing hypothesis, but Western cuisines do support it.

Data Gathering

In order to create a varied dataset we first used the "EasternEuropean" component of the supplementary papers dataset (Ahn, Ahnert, Bagrow, Barabasi (2011)). We also used a kaggle training data set containing Russian recipe ingredients. Lastly, a website called 'allrecipes.com' had recipes available which required some web scraping to gather. These three datasets combined to form our complete Russian recipe dataset contained 1013 unique entries.

For extra comparative data, we used the supplementary recipe data from (Ahn, Ahnert, Bagrow, Barabasi (2011)) for North America, Western European, Latin American, Southern European, East Asian cuisines.

Our ingredient flavour compound data contains 992 ingredients and 112,415 flavour compounds, 100 times the number of 1107 flavour compounds used in (Ahn, Ahnert, Bagrow, Barabasi (2011)), is sourced from (Wishart (2018)).

Data Mapping

Due to collecting recipe data from several different sources and some recipes being previously unprocessed, it was necessary to map down the large amount of ingredient names to unify references to the same ingredient (for example group "coriander" and "cilantro" under one name, and likewise "chopped onions", "sliced onions" and "brown onion"). Additionally, to access compound data from FooDB, all ingredient names had to be transformed into their naming scheme. This was accomplished in 2 stages:

- **Preprocessing:** To remove stray whitespace, convert to lowercase and perform basic vocabulary substitutions and removals (eg. swapping “broth” for “stock” and removing past particles such as “chopped” or “diced”).
- **Fuzzy Matching:** FooDB provides a list of common-names for ingredients within the database. Fuzzy string matching was used to match preprocessed ingredient names to a FooDB common name and hence the database’s ingredient name.

Since fuzzy matching is not a perfect solution to this problem, minimal manual work was required to check the resultant mappings. Of the approximately 10000 ingredients in collected russian recipes, preprocessing reduced the number of distinct ingredient names by a factor of 10, and the resulting names were mapped 83% correctly; this left less than 2% of the original dataset to be mapped manually.

Once recipes and ingredients were transformed into database compatible terms, it was possible to create a number of other maps for use in subsequent computations. This allowed for easy access to ingredient data without having to interact directly with the database.

The Flavour Network

A small example flavor network of Pelmeni and Cabbage Pie is depicted in Figure 1, showing the link between ingredients and compounds. The edges from ingredients to compounds are then condensed into edge weights depicted in Figure 2, showing the number shared compounds between ingredients.

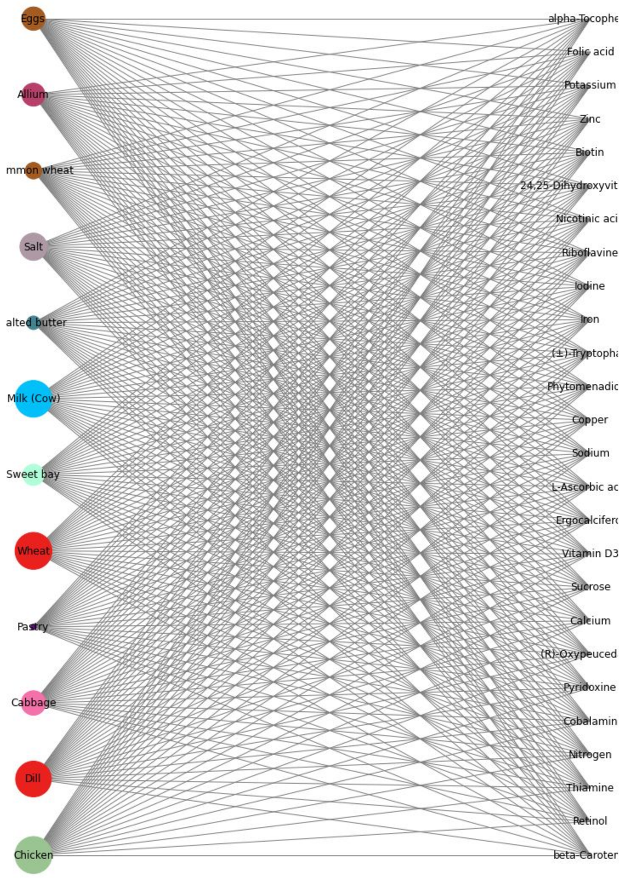


Figure 1: Partial Ingredient to Compound Bipartite Network for Pelmeni and Cabbage Pie

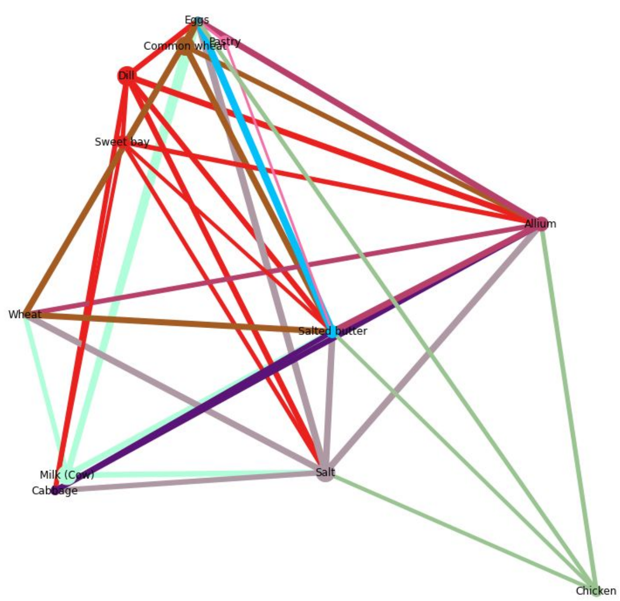


Figure 2: Ingredient network for Pelmeni and Cabbage Pie

Using our acquired recipe dataset and compound dataset, we can construct flavour networks, as depicted in figure 1 and 2, where the nodes are ingredients, the node size represents the ingredient's prevalence in recipes and the colours represent the ingredient category. The difference between figures 1 and 2 is that in figure 1 the edge size, showing size of the edge weight, represents the prevalence of that ingredient pairing in recipes while in figure 2 the edge size represents the number of shared compounds between ingredients. For viewing purposes, the edges have been trimmed using the method suggested by (Serrano, Boguná, Vespignani (2009)), the results are computed on the whole dataset.
On our second network, food pairing can be considered as a network property. Food pairing can be restated as whether the edge weight in the network depicted in figure 3 correlates with the edge weight in the network depicted in figure 4.
Figure 5 is notably denser due to the large number of flavour compounds in the flavour compound database we used.

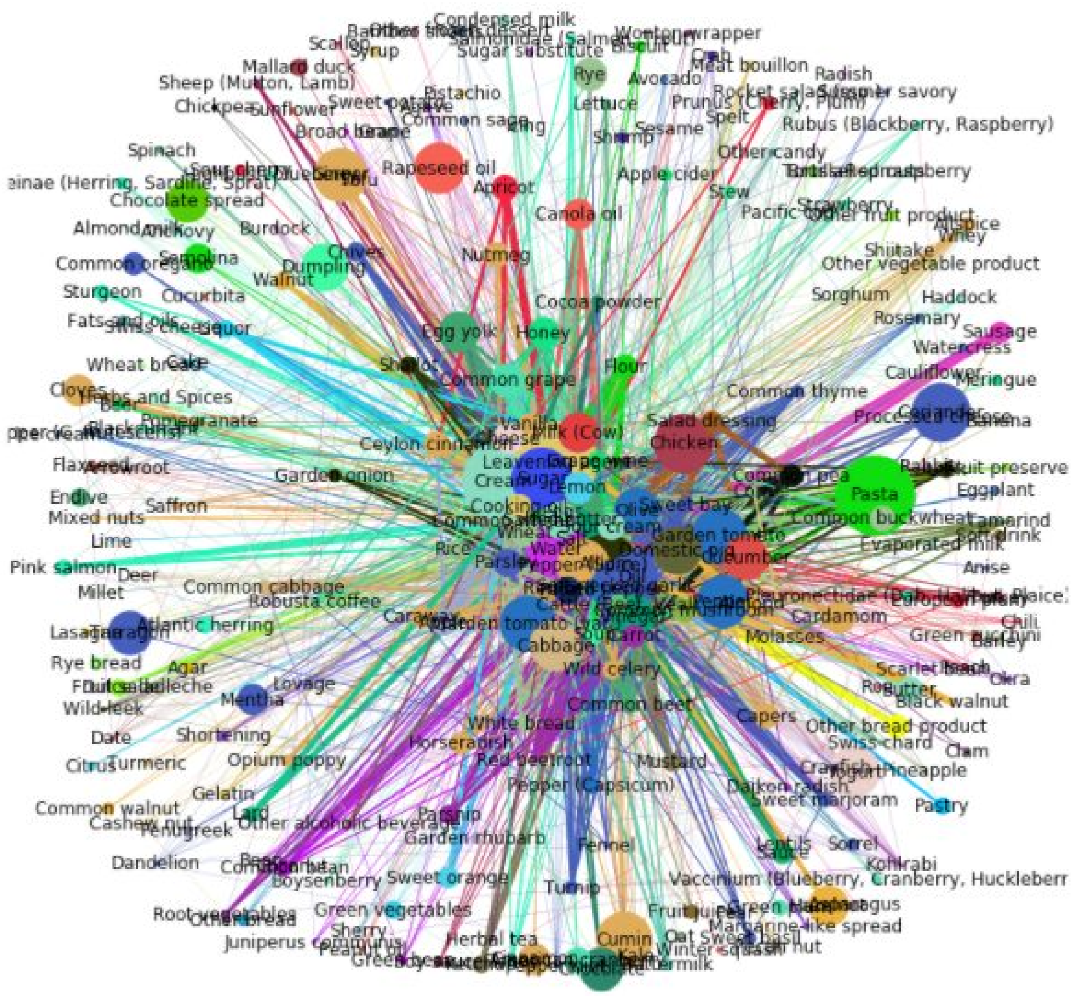


Figure 3: Backbone Network for Ingredients with Pairing Prevalence as Edge Weights

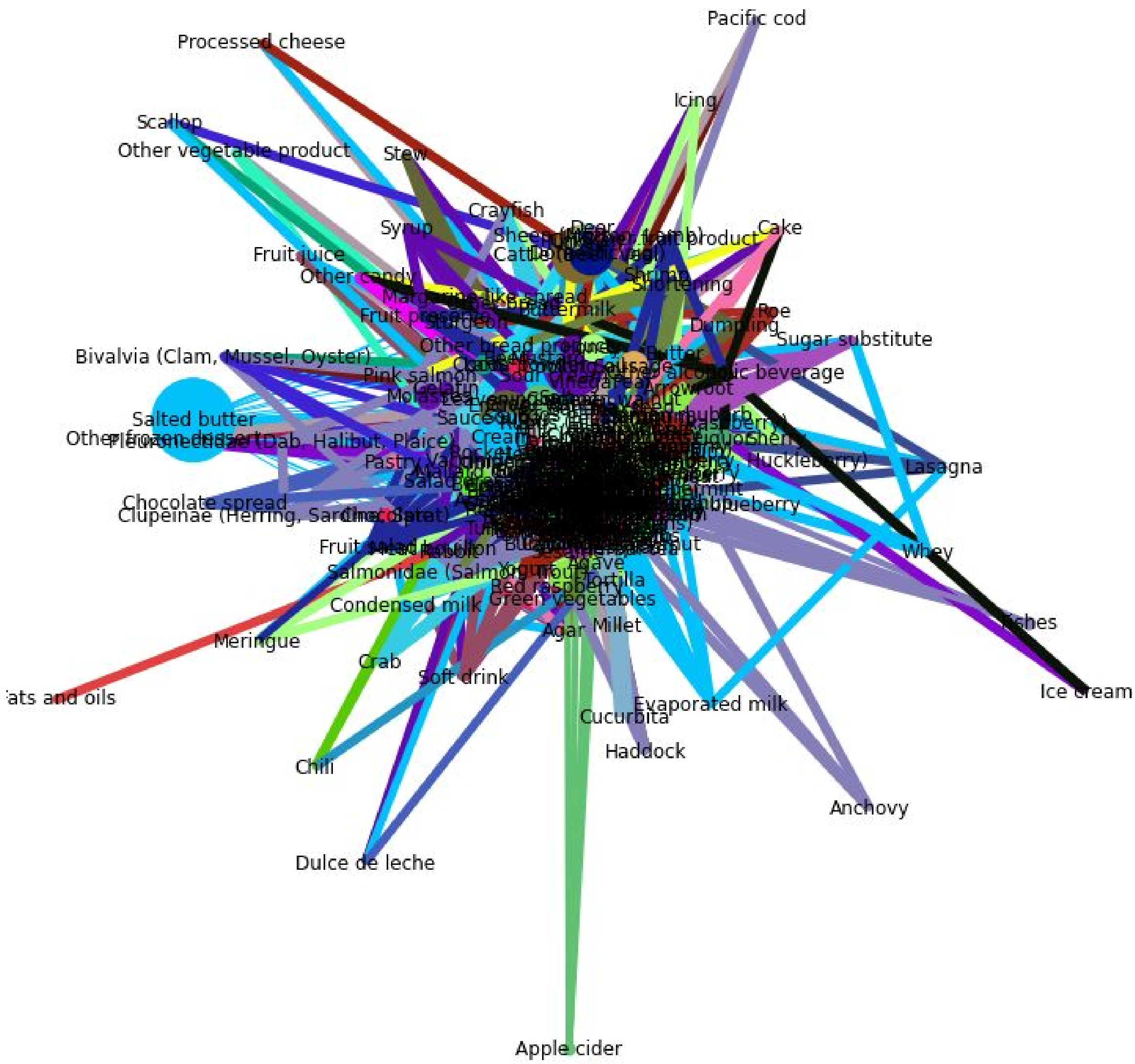


Figure 4: Backbone Network for Ingredients with Number of Shared Compounds as Edge Weights

Results

Power Laws & Scale Free Networks

As seen in figure 5, our data shows that the frequency in which people use certain ingredients follows a very robust power-law pattern. This means that there are a small number of ingredients which are responsible for a very large part of all cooking, and most ingredients are used very rarely. A more formal definition of this concept is that for *e* edges in a network, the fraction of nodes that have *P(e)* connections can be roughly estimated as:

$$P(e) = e^{-}$$

where is a value typically between 2 and 3.

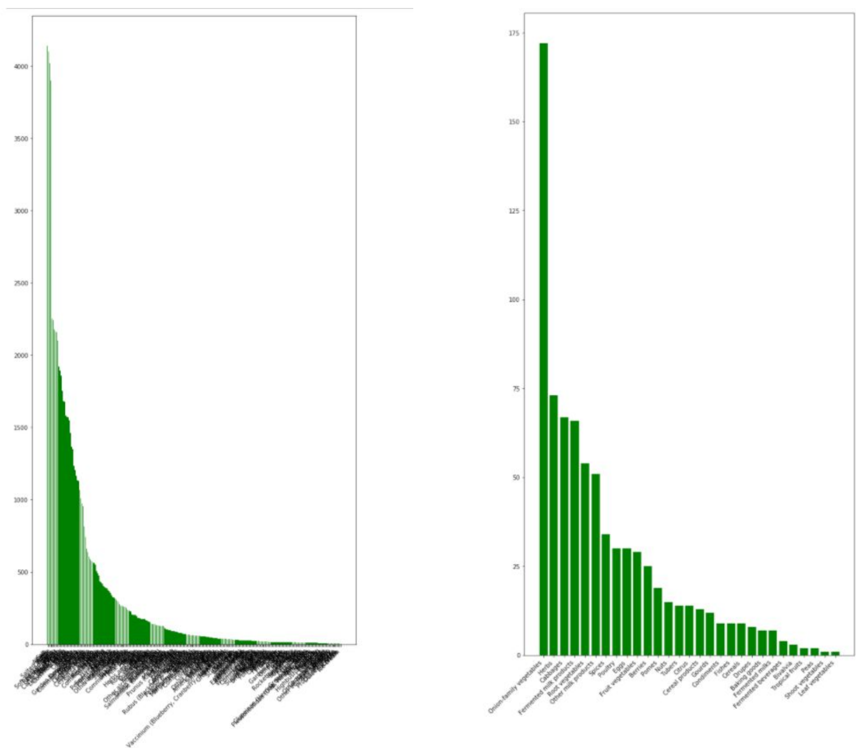


Figure 5: Frequency Histograms for ingredients (left) and categories (right)

These properties and ingredient distributions suggest that flavour and compound networks are scale-free networks. This may additionally provide evidence to suggest preferential selection of ingredients, recipes and flavour pairings.

Shared Compounds

The number of shared compounds for a recipe, $N_s(R)$ is defined as the average number of shared compounds between all ingredient pairs in that recipe, where n_R is the number of ingredients in recipe R and c_i is the compounds contained in ingredient i:

$$N_s(R) = \frac{2}{n_R(n_R - 1)} \sum_{i,j \in R, i \neq j} |C_i \cap C_j|$$

The average number of shared compounds for a cuisine, $N_s{}^c$, can be taken and compared to a pool of randomly generated recipes to give a Z-score which can be used to compare the statistical significance of the number of shared compounds in a cuisine. This metric gives insight as to whether or not a cuisine favours recipes with shared compounds between their ingredients. More details about the calculation of $N_s{}^c$ and the null models used to generate random recipes can be found in the attached python notebook (`Part 5 - Statistics.ipynb`).

(Ahn, Ahnert, Bagrow, Barabasi (2011)) found that whether ingredients were paired with ingredients with a high number of shared compounds was dependent on the cuisine.
However, our implementation of the comparison between the number of shared compounds and a random model, described in (Ahn, Ahnert, Bagrow, Barabasi (2011)), on a larger flavour compound database showed that none of our considered cuisines had a tendency towards either pairing ingredients with shared compounds or avoiding pairing ingredients with shared compounds, as seen in figures 6-9.

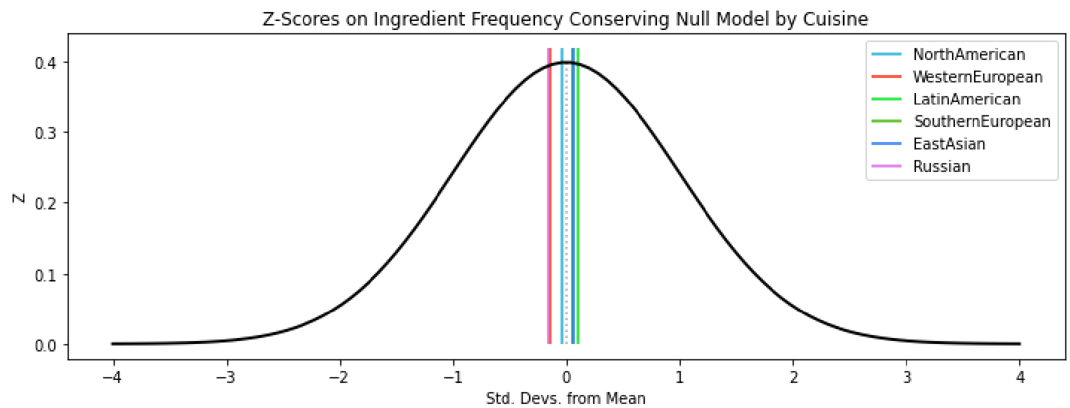


Figure 6: Z-score values for ingredient frequency

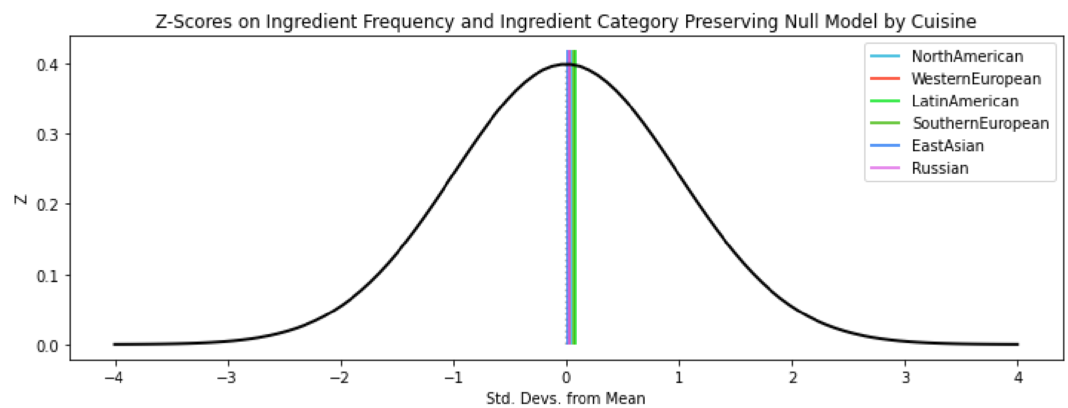


Figure 7: Z-score values for ingredient frequency and ingredient category

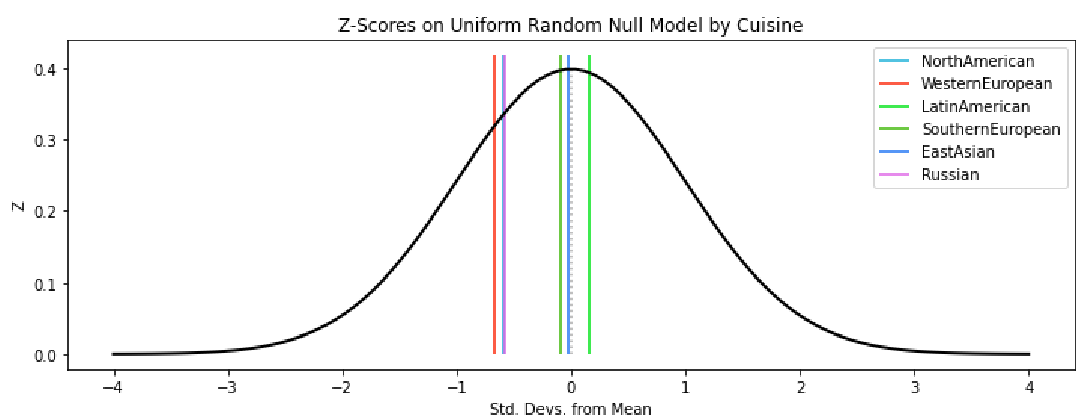


Figure 8: Z-score values for uniform random null model

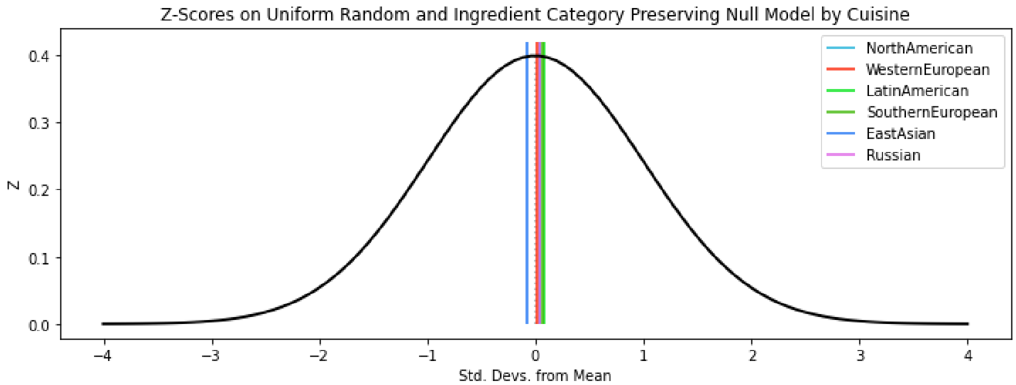


Figure 9: Z-score values for uniform random and ingredient categories

Authenticity

The authenticity of an ingredient *i* in a cuisine *c* can be determined from its prevalence P_i^c, where P_i^c is the proportion of recipes from *c* which contain *i*. A more formal definition is given by:

$$P_i^c = \frac{no.\ recipes\ in\ c\ containing\ i}{no.\ recipes\ in\ c} = \frac{n_i^c}{N_i^c}$$

The authenticity p_i^c is defined as the difference between the prevalence for c and the average prevalence for all other cuisines. A more formal definition is given by:

$$P_i^c = P_i^c - \langle P_i^{c'} \rangle_{c' != c}$$

The authenticities listed in figure 9 relate to East Asian, North American and Russian cuisines. The most authentic ingredient of East Asian cuisine is Soy Sauce. The most authentic ingredient in North American cuisine in molasses. The most authentic ingredient in Russian cuisine is salt.

Cuisine	**Single Ingredient**	**Ingredient Pair**
East Asian	Soy sauce, Sesame, Rice, Ginger, Soy bean, Pepper (Capsicum)	(Allium, Soft-necked garlic), (Allium, Soy sauce), (Soft-necked garlic, Soy sauce), (Allium, Sesame), (Soy sauce, Sesame), (Allium, Pepper (Capsicum))
North American	Molasses, Salted butter, Common wheat, Milk (Cow), Vanilla, Ceylon cinnamon	(Common wheat, Salted butter), (Common wheat, Eggs), (Common wheat, Milk (Cow)), (Common wheat, Vanilla), (Common wheat, Molasses), (Molasses, Salted butter)

Russian	Salt, Sugar, Water, Eggs, Wheat, Sour cream	(Eggs, Salt), (Allium, Salt), (Salt, Salted butter), (Eggs, Sugar), (Salt, Sugar), (Salt, Wheat)

Figure 10: Authenticities of East Asian, North American and Russian cuisines

The full list of authenticity data for 1,2,3 and 4 sets of ingredients can be found in Appendix 4.

Discussion

The Number of Shared Compounds had near-zero Z-Scores for ingredient frequency conserving, ingredient frequency conserving and category conserving, and category conserving null models.
The only model which showed larger absolute value Z-Scores was the uniform random null model, which had a higher mean shared compounds than the other 3 null models.
This is likely due to the uniform random null model pairing ingredients with a high shared compound than typically performed by a cuisine, rather than the cuisine having a propensity not to pair ingredients which share a large number compounds.
This explanation is further supported by the larger Z-Scores being negative.

Cuisine	Z-Score
North American	45.5
Western European	6
Latin American	4
Southern European	-2
East Asian	-3

Figure 11: Z-scores for each cuisine

Our results differ from (Ahn, Ahnert, Bagrow, Barabasi (2011)), which has a Z-score of 45.5, 6, 4, -2, -3 for North American cuisine, Western European cuisine, Latin American cuisine, Southern European cuisine, East Asian cuisine, respectively.
This can be explained by our use of a larger compound database, having roughly 100 times the number of compounds used in (Ahn, Ahnert, Bagrow, Barabasi (2011)), allowing the consideration of more compounds, likely detecting more shared compounds in the case of Southern European and East Asian cuisine and increasing the number of shared compounds considered by the null models, decreasing the difference between the null models and North American, Western European and Latin American cuisine.
(Ahn, Ahnert, Bagrow, Barabasi (2011)) did check for robustness of results against incompleteness of compound data, however the comparison was only a 6 times dispensary in compound data.

In addition to Number of Shared Compounds and Authenticity covered in our report, (Ahn, Ahnert, Bagrow, Barabasi (2011)) also considered the contribution of each ingredient.
We did not implement this metric as we found the cuisines had little tendency towards pairing or not pairing ingredients with shared compounds, hence none of the ingredients contributed a large amount to this effect.

Conclusion

Our results show that the food pairing principle does not hold for any cuisine. This result can likely be attributed to the fact that a vastly different compound mapping scheme to the source material. Our results additionally show that flavour compounds and recipe networks both show strong properties of being scale-free networks that follow a power law distribution and have self-similar characteristics when examined on a cuisine-by-cuisine basis or as a global network of food.

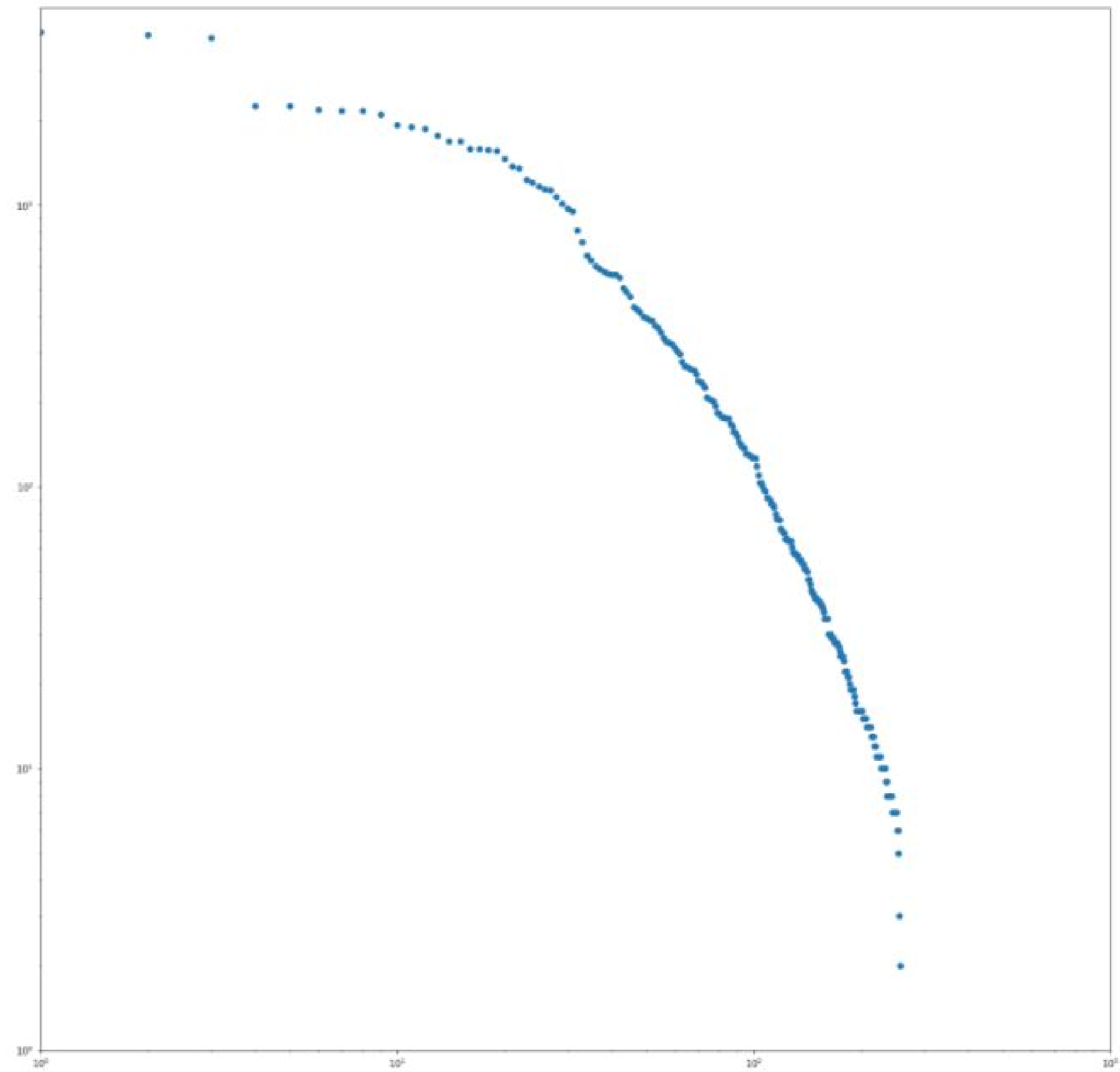
As future work, we leave measurement of the impact of compound databases, such as FlavorDB (Garg, Sethupathy, Tuwani, Nk, Dokania, Iyer, et. al (2018)) which notably combines FooDB, our compound source, with Fenaroli's Handbook of Flavor Ingredients, the compound source of (Ahn, Ahnert, Bagrow, Barabási (2011)).

References

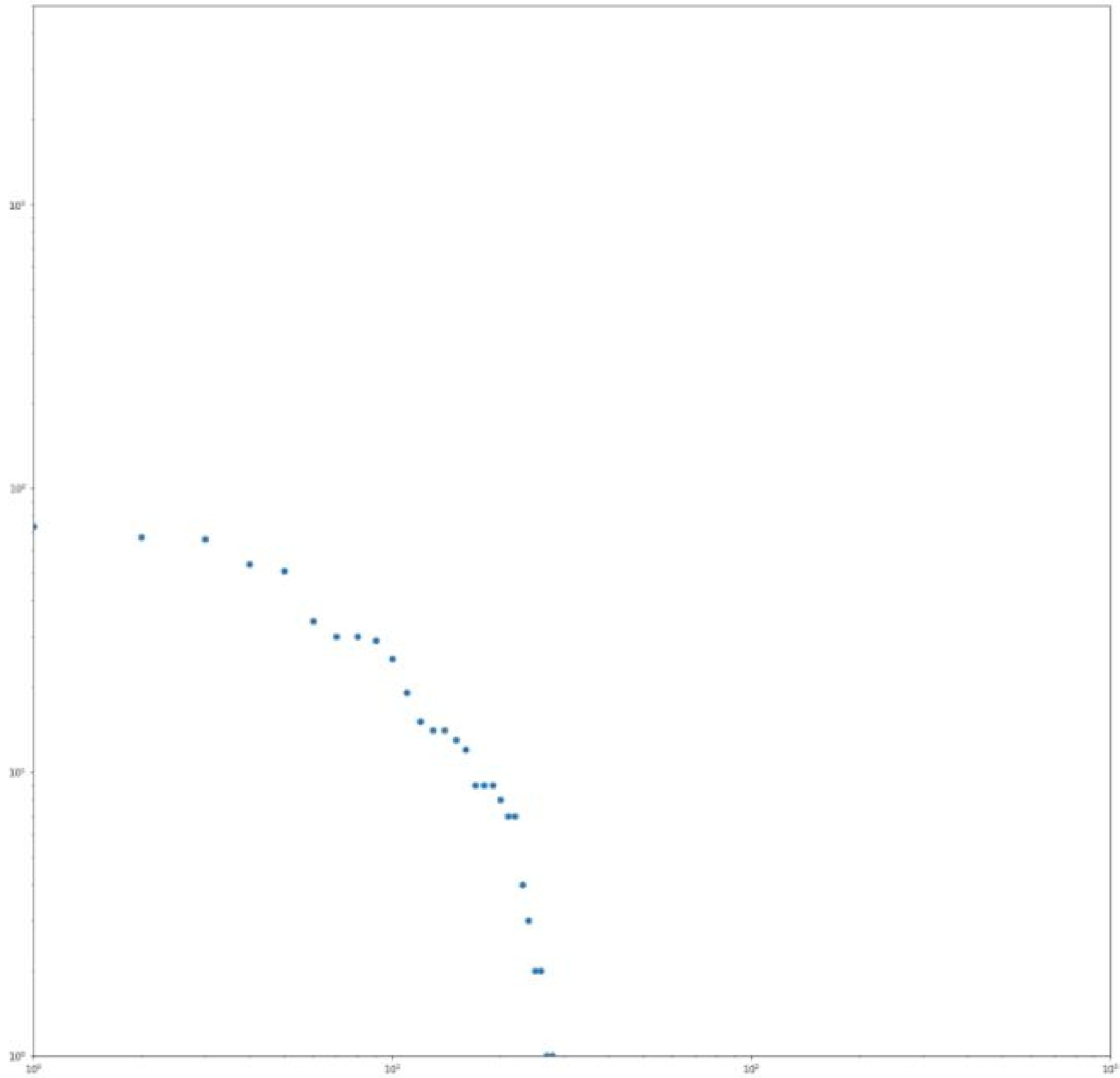
Ahn, Y. Y., Ahnert, S. E., Bagrow, J. P., & Barabási, A. L. (2011). Flavor network and the principles of food pairing. Scientific reports, 1, 196.
Wishart, D. S. (2018). FooDB: the food database.
Garg, N., Sethupathy, A., Tuwani, R., Nk, R., Dokania, S., Iyer, A., ... & Kathuria, K. (2018). FlavorDB: a database of flavor molecules. *Nucleic acids research*, *46*(D1), D1210-D1216.
Serrano, M. Á., Boguná, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, *106*(16), 6483-6488.

Appendix

Appendix 1:

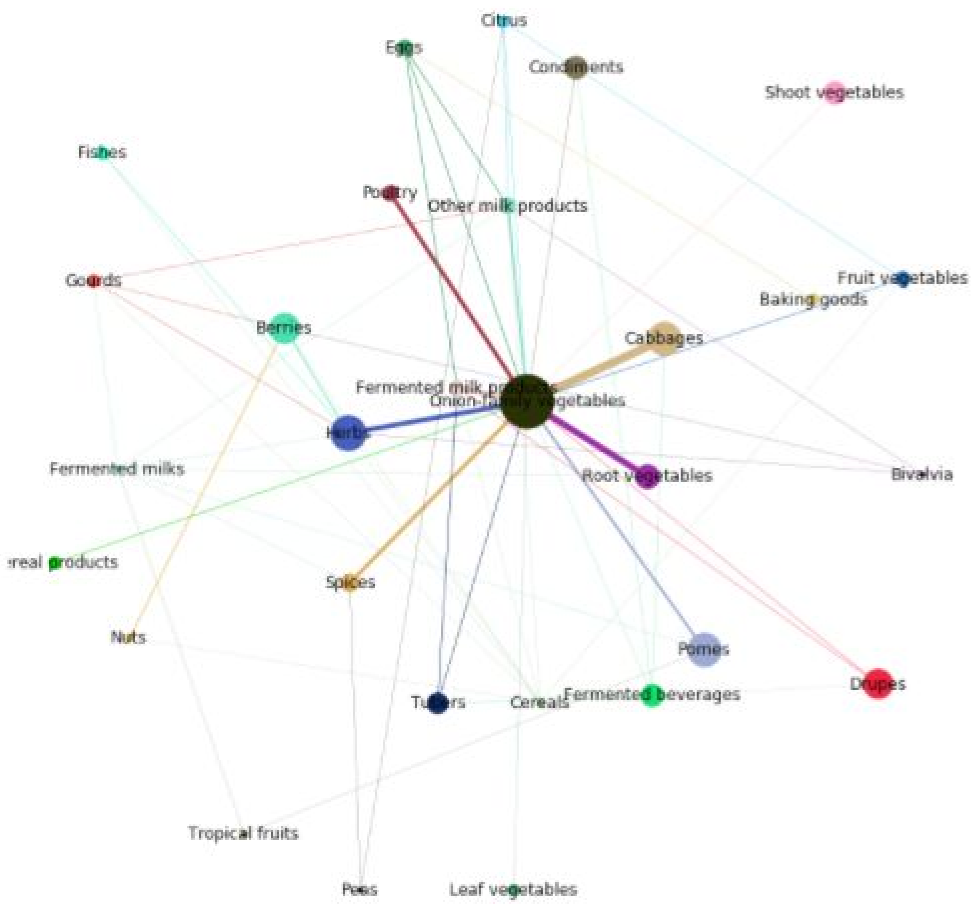


Appendix 2



Appendix 3

Backbone Network for Ingredient Categories



Appendix 4: Authenticity Data Readout

Finding Most Authentic Combination of 1 Ingredients Per Cuisine:
NorthAmerican:
[('Molasses',), ('Salted butter',), ('Common wheat',), ('Milk (Cow)',), ('Vanilla',), ('Ceylon cinnamon',)]

WesternEuropean:
[('Salted butter',), ('Common wheat',), ('Eggs',), ('Milk (Cow)',), ('Cream',), ('Common thyme',)]

LatinAmerican:
[('Pepper (Capsicum)',), ('Garden tomato (var.)',), ('Allium',), ('Cumin',), ('Soft-necked garlic',), ('Corn',)]

SouthernEuropean:
[('Olive',), ('Cheese',), ('Sweet basil',), ('Pasta',), ('Garden tomato (var.)',), ('Soft-necked garlic',)]

EastAsian:
[('Soy sauce',), ('Sesame',), ('Rice',), ('Ginger',), ('Soy bean',), ('Pepper (Capsicum)',)]

Russian:
[('Salt',), ('Sugar',), ('Water',), ('Eggs',), ('Wheat',), ('Sour cream',)]

Alden

11:00 PM
Finding Most Authentic Combination of 2 Ingredients Per Cuisine:
NorthAmerican:
[('Common wheat', 'Salted butter'), ('Common wheat', 'Eggs'), ('Common wheat', 'Milk (Cow)'), ('Common wheat', 'Vanilla'), ('Common wheat', 'Molasses'), ('Molasses', 'Salted butter')]

WesternEuropean:
[('Common wheat', 'Salted butter'), ('Common wheat', 'Eggs'), ('Eggs', 'Salted butter'), ('Common wheat', 'Milk (Cow)'), ('Cream', 'Eggs'), ('Cream', 'Salted butter')]

LatinAmerican:
[('Allium', 'Pepper (Capsicum)'), ('Garden tomato (var.)', 'Pepper (Capsicum)'), ('Pepper (Capsicum)', 'Soft-necked garlic'), ('Allium', 'Garden tomato (var.)'), ('Garden tomato (var.)', 'Soft-necked garlic'), ('Allium', 'Soft-necked garlic')]

SouthernEuropean:
[('Olive', 'Soft-necked garlic'), ('Cheese', 'Olive'), ('Garden tomato (var.)', 'Olive'), ('Cheese', 'Soft-necked garlic'), ('Allium', 'Olive'), ('Olive', 'Sweet basil')]

EastAsian:
[('Allium', 'Soy sauce'), ('Soft-necked garlic', 'Soy sauce'), ('Allium', 'Sesame'), ('Sesame', 'Soy sauce'), ('Sesame', 'Soft-necked garlic'), ('Rice', 'Soy sauce')]

Russian:
[('Eggs', 'Salt'), ('Allium', 'Salt'), ('Salt', 'Salted butter'), ('Eggs', 'Sugar'), ('Salt', 'Sugar'), ('Salt', 'Wheat')]
Finding Most Authentic Combination of 3 Ingredients Per Cuisine:
NorthAmerican:
[('Common wheat', 'Eggs', 'Vanilla'), ('Common wheat', 'Eggs', 'Salted butter'), ('Common wheat', 'Eggs', 'Milk (Cow)'), ('Common wheat', 'Milk (Cow)', 'Salted butter'), ('Common wheat', 'Salted butter', 'Vanilla'), ('Common wheat', 'Eggs', 'Molasses')]

WesternEuropean:
[('Common wheat', 'Eggs', 'Salted butter'), ('Common wheat', 'Milk (Cow)', 'Salted butter'), ('Common wheat', 'Eggs', 'Milk (Cow)'), ('Common wheat', 'Cream', 'Eggs'), ('Common wheat', 'Cream', 'Salted butter'), ('Cream', 'Eggs', 'Salted butter')]

LatinAmerican:
[('Allium', 'Garden tomato (var.)', 'Pepper (Capsicum)'), ('Allium', 'Pepper (Capsicum)', 'Soft-necked garlic'), ('Garden tomato (var.)', 'Pepper (Capsicum)', 'Soft-necked garlic'), ('Allium', 'Garden tomato (var.)', 'Soft-necked garlic'), ('Allium', 'Cheese', 'Pepper (Capsicum)'), ('Allium', 'Pepper (Capsicum)', 'Red bell pepper')]

SouthernEuropean:
[('Garden tomato (var.)', 'Olive', 'Soft-necked garlic'), ('Cheese', 'Olive', 'Soft-necked garlic'), ('Olive', 'Soft-necked garlic', 'Sweet basil'), ('Garden tomato (var.)', 'Soft-necked garlic', 'Sweet basil'), ('Garden tomato (var.)', 'Olive', 'Sweet basil'), ('Cheese', 'Garden tomato (var.)', 'Olive')]

EastAsian:
[('Allium', 'Soft-necked garlic', 'Soy sauce'), ('Allium', 'Sesame', 'Soft-necked garlic'), ('Allium', 'Sesame', 'Soy sauce'), ('Sesame', 'Soft-necked garlic', 'Soy sauce'), ('Allium', 'Ginger', 'Soft-necked garlic'), ('Allium', 'Rice', 'Soy sauce')]

Russian:
[('Eggs', 'Salt', 'Sugar'), ('Eggs', 'Salt', 'Wheat'), ('Eggs', 'Salt', 'Salted butter'), ('Eggs', 'Salted butter', 'Sugar'), ('Salt', 'Salted butter', 'Wheat'), ('Eggs', 'Sugar', 'Wheat')]
Finding Most Authentic Combination of 4 Ingredients Per Cuisine:
NorthAmerican:
[('Common wheat', 'Eggs', 'Salted butter', 'Vanilla'), ('Common wheat', 'Eggs', 'Milk (Cow)', 'Salted butter'), ('Common wheat', 'Eggs', 'Molasses', 'Salted butter'), ('Common wheat', 'Eggs', 'Milk (Cow)', 'Vanilla'), ('Common wheat', 'Eggs', 'Molasses', 'Vanilla'), ('Common wheat', 'Milk (Cow)', 'Salted butter', 'Vanilla')]

WesternEuropean:
[('Common wheat', 'Eggs', 'Milk (Cow)', 'Salted butter'), ('Common wheat', 'Cream', 'Eggs', 'Salted butter'), ('Common wheat', 'Eggs', 'Salted butter', 'Vanilla'), ('Common wheat', 'Eggs', 'Lard', 'Salted butter'), ('Common wheat', 'Cream', 'Eggs', 'Milk (Cow)'), ('Almond', 'Common wheat', 'Eggs', 'Salted butter')]

LatinAmerican:
[('Allium', 'Garden tomato (var.)', 'Pepper (Capsicum)', 'Soft-necked garlic'), ('Allium', 'Cheese', 'Pepper (Capsicum)', 'Soft-necked garlic'), ('Allium', 'Cheese', 'Garden tomato (var.)', 'Pepper (Capsicum)'), ('Allium', 'Cumin', 'Pepper (Capsicum)', 'Soft-necked garlic'), ('Allium', 'Pepper (Capsicum)', 'Red bell pepper', 'Soft-necked garlic'), ('Allium', 'Garden tomato (var.)', 'Pepper (Capsicum)', 'Red bell pepper')]

SouthernEuropean:
[('Garden tomato (var.)', 'Olive', 'Soft-necked garlic', 'Sweet basil'), ('Cheese', 'Garden tomato (var.)', 'Olive', 'Soft-necked garlic'), ('Cheese', 'Olive', 'Soft-necked garlic', 'Sweet basil'), ('Cheese', 'Garden tomato (var.)', 'Soft-necked garlic', 'Sweet basil'), ('Cheese', 'Garden tomato (var.)', 'Olive', 'Sweet basil'), ('Cheese', 'Olive', 'Pasta', 'Soft-necked garlic')]

EastAsian:
[('Allium', 'Sesame', 'Soft-necked garlic', 'Soy sauce'), ('Allium', 'Pepper (Capsicum)', 'Sesame', 'Soft-necked garlic'), ('Allium', 'Pepper (Capsicum)', 'Soft-necked garlic', 'Soy sauce'), ('Allium', 'Ginger', 'Pepper (Capsicum)', 'Soft-necked garlic'), ('Allium', 'Ginger', 'Soft-necked garlic', 'Soy sauce'), ('Allium', 'Rice', 'Soft-necked garlic', 'Soy sauce')]

Russian:

[('Eggs', 'Salt', 'Salted butter', 'Sugar'), ('Eggs', 'Salt', 'Sugar', 'Wheat'), ('Eggs', 'Salt', 'Salted butter', 'Wheat'), ('Eggs', 'Salted butter', 'Sugar', 'Wheat'), ('Eggs', 'Leavening agent', 'Salt', 'Sugar'), ('Eggs', 'Leavening agent', 'Salted butter', 'Sugar')]