Kissing Cuisines: Exploring Worldwide Culinary Habits on the Web

Sajadmanesh, S; Jafarzadeh, S; Ossia, SA; Rabiee, HR; Haddadi, H; Mejova, Y; Musolesi, M; Cristofaro, ED; Stringhini, G

https://arxiv.org/abs/1610.08469

Kissing Cuisines: Exploring Worldwide Culinary Habits on the Web

Sina Sajadmanesh[†], Sina Jafarzadeh[†], Seyed Ali Ossia[†], Hamid R. Rabiee[†], Hamed Haddadi[‡],
Yelena Mejova[⋆], Mirco Musolesi[♯], Emiliano De Cristofaro[♯], Gianluca Stringhini[♯]

[†]Sharif University of Technology [‡]Queen Mary University of London [⋆]Qatar Computing Research Institute [♯]UCL

ABSTRACT

As food becomes an important part of modern life, recipes shared on the web are a great indicator of civilizations and culinary attitudes in different countries. Similarly, ingredients, flavors, and nutrition information are strong signals of the taste preferences of individuals from various parts of the world. Yet, we do not have a thorough understanding of these palate varieties.

In this paper, we present a large-scale study of recipes published on the Web and their content, aiming to understand cuisines and culinary habits around the world. Using a database of more than 157K recipes from over 200 different cuisines, we analyze ingredients, flavors, and nutritional values which distinguish dishes from different regions, and use this knowledge to assess the predictability of recipes from different cuisines. We then use country health statistics to understand the relation between these factors and health indicators of different nations, such as obesity, diabetes, migration, and health expenditure. Our results confirm the strong effects of geographical and cultural similarities on recipes, health indicators, and culinary preferences between countries around the world.

1. INTRODUCTION

Nowadays, the importance of food and eating goes far beyond a means to survive. Often regarded as a social construct, food has become an essential part of modern day life. New jargon has entered our vocabulary as expressions like “foodie”, “food porn”, or “food tourism”, hint at the buzz around the entertainment arising from our culinary experiences. With the rise of social media, and the proliferation of always-on always-connected devices, this gobbling revolution is not confined to our kitchens, restaurants, and food stalls, but naturally breaks out on the social web. Sharing pictures of one’s food has become a growing passion for both tourists and locals [16], and dedicated food searching and sharing apps, along with recipe websites and the ubiquitous social presence of celebrity chefs, have all contributed to a thriving culture and passion around food worldwide.

Around the world, different cuisines are naturally intertwined with cultures, traditions, passions, and religion of individuals living in different countries and continents. Sushi, curry, kebab, pasta, tacos – these are just examples of foods conventionally associated with specific countries, as are specific cuisines and ingredients. Different dietary habits around the world are also closely related to various health statistics, including cancer incidence [3], death rates [13], cardiovascular complications [18], and obesity [14].

Although there are many common beliefs about cuisines, recipes, and their ingredients, it is still unclear what types of ingredients are unique in/about different countries, what factors make cuisines similar to each other (e.g., in terms of ingredients or flavors), and how these factors are related to individuals’ health. With this motivation in mind, in this paper, we set to investigate the way in which ingredients relate to different cuisines and recipes, as well as the geographic and health significances thereof. We use a few datasets, including 157K recipes from over 200 cuisines crawled from Yummly, BBC Food data, and country health statistics.

Overview & Contributions. First, we characterize different cuisines around the world by their ingredients and flavors. Then, we train a Support Vector Machine classifier and use deep learning models to predict a cuisine from its ingredients. This also enables us to discover the similarity across different cuisines based on their ingredients – e.g., Chinese and Japanese – while, intuitively, they might be considered different. We look at the diversity of ingredients in recipes from different countries and compare them to geographic and human migration statistics. We also measure the relationship between the nutrition value of the recipes vis-à-vis public health statistics such as obesity and diabetes.

Paper Organization. The rest of the paper is organized as follows. In Section 2, we present the datasets used in our evaluation, then Section 3 presents an analysis of the diversity of the ingredients around the world, looking at geographic diversity patterns of cuisines and notable ingredients in particular ones. In Section 4, we look at the similarity between the cuisines based on their ingredients and flavors, and use these results to train machine-learning classifiers for ingredient-based cuisine prediction models in Section 5. In Section 6, we correlate the nutrition values of recipes for different countries with their public health statistics. After reviewing related work in Section 7, the paper concludes in Section 8.

2. DATASETS

Our study relies on a number of datasets, namely, a large set of recipes collected from Yummly, a list of ingredients compiled by BBC Food, and country health statistics. In this section, we describe these datasets in detail.

2.1 Yummly data

Yummly is a website offering recipe recommendations based on the user’s taste.[1] It allows users to search for recipes, learning which dishes the user likes and providing them with recipe suggestions. It also provides a user-friendly API, which we use to collect recipes. First, we crawled Wikipedia for a list of cuisines[2], then, in Summer 2016, we queried the Yummly API for recipes belonging to each cuisine. In the end, we obtained 157,013 recipes belonging to over 200 different cuisines. Due to API restrictions, we limited the number of recipes per to 5,000.

Each recipe obtained from the Yummly API contains a number of attributes. In our study, we use the following:

[1] http://www.yummly.com

[2] https://en.wikipedia.org/wiki/List_of_cuisines

1. **Ingredients:** Each recipe contains a list of the ingredients that are required to prepare it. Since Yummly acts as a recipe aggregator from various cooking sites, the ingredients do not always appear with the same wording. In fact, it is very common to see the same ingredient written with different spellings or by using a different terminology. We overcome these issues through a standardization process described in Section 2.2.
2. **Flavors:** Recipes are identified by six flavors, specifically, saltiness, sourness, sweetness, bitterness, savoriness, and spiciness. These scores are on a range of 0 to 1.
3. **Rating:** Users are encouraged to provide a rating, from 1 to 5, for the recipes that they try. From the Yummly API, we retrieve the average review rating for each recipe.
4. **Nutrition:** Unfortunately, the Yummly search API does not directly provide nutritional information for the recipes. As a consequence, we designed a simple web crawler to fetch the corresponding web page for each recipe in our dataset, and extract information on the amount of protein, fat, saturated fat, sodium, fiber, sugar, and carbohydrate of a recipe (per serving), as well as calories.

Although some ingredients appear in other languages (e.g., German, French, etc), the recipes presented here are mostly in English; hence it is possible that some more authentic or niche local recipes might be missing from our dataset. However, considering the number of recipes and a large cut-off threshold introduced later on, we are confident this does not significantly affect our analysis. Moreover, authors of the recipes might not represent the entire population, given the fact that they are likely to be tech-savvy. This might introduce a potential bias in the dataset, but at the same time, this potential issue is compensated by its richness in terms of the variety of dishes from different countries available in it as structured information.

2.2 BBC Food Data

BBC Food[3] is a part of the BBC website providing information about recipes, ingredients, chefs, cuisines, and other information related to cooking and dishes from all BBC programs. In Summer 2016, we crawled all the ingredients from the BBC Food website, collecting about 1,000 ingredients, which we used to organize and standardize the ingredients in the Yummly dataset. The standardization process is as follows:

(i) We extracted all the 11,000 ingredients from the Yummly dataset and performed a preliminary data cleaning, i.e., removing measurement units (mass, volume, etc), numbers, punctuation marks, and other symbols.

(ii) Due to the multilingualism of the Yummly data, we used the Google Translate API to perform automatic language detection and translation of all the Yummly ingredients to English.

(iii) We used the BBC list of ingredients as a reference, and mapped all possible ingredients from the Yummly list to it.

(iv) As not all ingredients from the Yummly list were successfully mapped, we merged the similar ones into groups, and the ingredients in each group were manually mapped to the its representative ingredient.

Overall, this process yields about 3,000 standardized ingredients.

2.3 Country health statistics

As diet is directly related to the health of individuals, we also set to relate Yummly statistics to real-world health data. To this end, we will use the diabetes prevalence estimates from World Development Indicators by The World Bank[4], the health expenditure as a percentage of total GDP from The World Bank[5], and the obesity prevalence from the World Health Organization[6] in the countries to which the cuisines are mapped, using the most recent available data, which is from 2014.

3. INGREDIENTS AROUND THE WORLD

In this section, we provide a characterization of the ingredients used in dishes from all over the world. First, we investigate the diversity of ingredients in different countries. Next, we define the concept of "complexity" of a dish in terms of its ingredients and look at how complexity changes around the world. Finally, we discuss a series of case studies of most notable and significant ingredients in some eminent cuisines.

3.1 Diversity of ingredients

Aiming to investigate the diversity of ingredients in dishes of a cuisine, we set to answer the following questions:

1. How many different unique ingredients are used in total in dishes of each country? In other words, what is the number of unique ingredients the people of a country have ever used to prepare a culinary dish? The answer to this question is what we refer to as the *global diversity*.
2. How different are the dishes of an individual country relative together in terms of their ingredients combination? In other words, do different dishes usually share some ingredients or their ingredients are almost different? The answer to this question is what we call *local diversity*.

The local and global diversity of ingredients in a country depend on many parameters including the geographical location, climatic conditions, agricultural situation, or even the amount of immigration which directly influences the diversity of culinary cultures. The calculation of the global diversity is performed in two steps. Since the number of recipes per different cuisines are variable, we first set a fixed number of 100 recipes per cuisine, discarding cuisines containing fewer number of recipes, and sampling from cuisines containing more number of recipes uniformly at random, to have an equal number of recipes in all cuisines. This results in a final set of 82 different cuisines each containing 100 recipes. We then map the result obtained for each cuisine to its corresponding country. Here, some countries are mapped with more than one cuisine. For such countries, we record the average result over their associated cuisines.

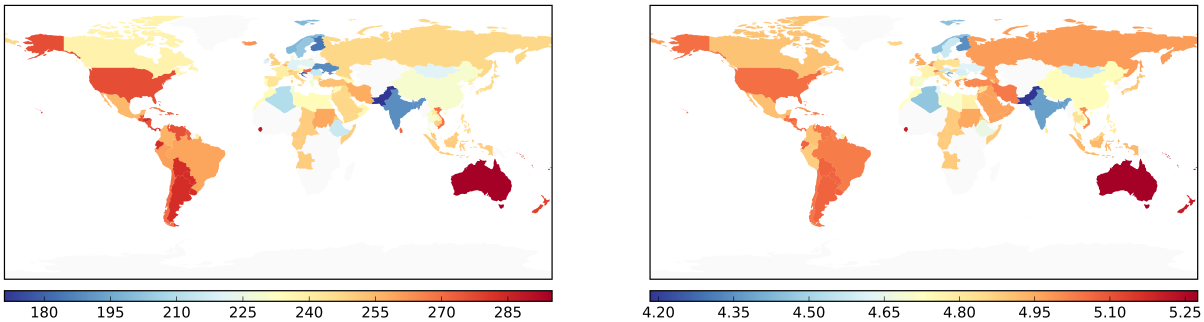
To calculate the local diversity, we look at each cuisine as a probability distribution over all standard ingredients. By counting the total number of occurrences of each ingredient in all recipes of a particular cuisine, and then normalizing the values such that they sum to one, we obtain the ingredient distribution for that cuisine. We then calculate the entropy of these distributions as the local diversity of their corresponding cuisines. The entropy of the ingredient distribution measures the unpredictability of ingredients used in the dishes. Therefore, the higher the entropy of the ingredient distribution of a particular cuisine, the more different the ingredients combination of its recipes, and thus the higher the local diversity. To preserve the smoothness of the ingredient distributions, we again keep the 82 cuisines with more than 100 recipes. After calculating the local diversity for each cuisine, we follow the

[3] http://www.bbc.co.uk/food/

[4] http://data.worldbank.org/indicator/SH.STA.DIAB.ZS

[5] http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS

[6] http://apps.who.int/gho/data/view.main.2450A



(a) Global diversity

(b) Local diversity

Figure 1: Diversity of ingredients around the world

same procedure as for the global diversity to map the cuisine-based results to countries.

Figure 1 shows the local diversity and the global diversity of ingredients for different countries around the world. According to the figure, we can see that the local and global diversities have a meaningful correlation with each other. The countries with high global diversity have also high local diversity, and countries with low global diversity have low local diversity as well. This happens because as the global diversity increases, people will have more options to choose as the ingredients for their foods, so they can prepare relatively different dishes.

Another interesting trend observable in Figure 1 is that countries like the United States and Australia which usually accept a high number of immigrants, have a relatively high global diversity. Regarding this, we hypothesized that the number of immigrants coming to a country must have an influence on the ingredient diversity of that country. This is mainly due to immigrants bringing their native culinary culture with themselves, which in turn makes the cuisines of their target country richer. To investigate this fact, we collected the net migration data from the World Bank[7] which shows the difference between the total number of immigrants and emigrants during a time period. We correlated the global diversity with the average net migration from 1960 to 2016. To this end, we fitted a polynomial curve to the data points considering the global diversity and the net migration. The result is shown in Figure 2. As we expect, the figure indicates that an increase in the net migration would result in an increase in the global diversity of ingredients.

3.2 Complexity of dishes

Another interesting concept about the culinary preferences of different countries is the complexity of dishes. The complexity of a dish is simply the number of unique ingredients required to prepare it. Accordingly, a cuisine is more complex than another one if its dishes are proportionally more complex than another one's.

Formally speaking, each cuisine is associated with the complexity distribution of its dishes. For a sample cuisine, this distribution, namely $P(X = i)$, specifies the probability of a dish from that cuisine to have exactly i unique ingredients. This way, the cumulative complexity distribution (CCD) will give us an insight about the complexity of dishes in a particular cuisine.

Figure 3 shows the cumulative complexity distribution (CCD) for Norwegian, Russian, Spanish, Tunisian, and Lao as an example. From the figure we can see that the CCD for Norwegian cuisine grows faster than the others, while for Lao, it is relatively slower. This means that Lao dishes are more complex than Norwegian ones. Thus for each cuisine, the area under its CCD is inversely related to its complexity. Hence, we use the reciprocal of

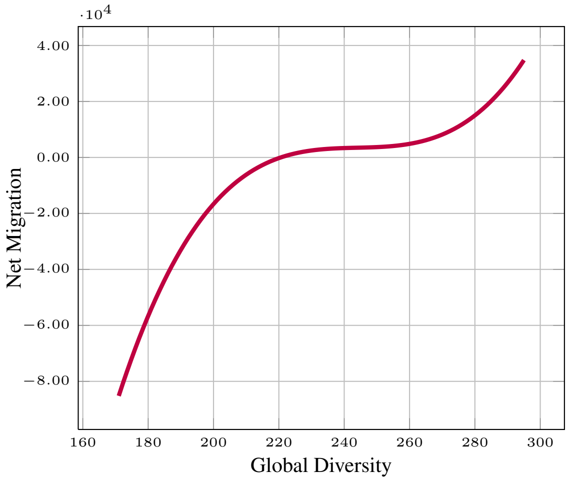


Figure 2: Relation between the global ingredient diversity and the net migration

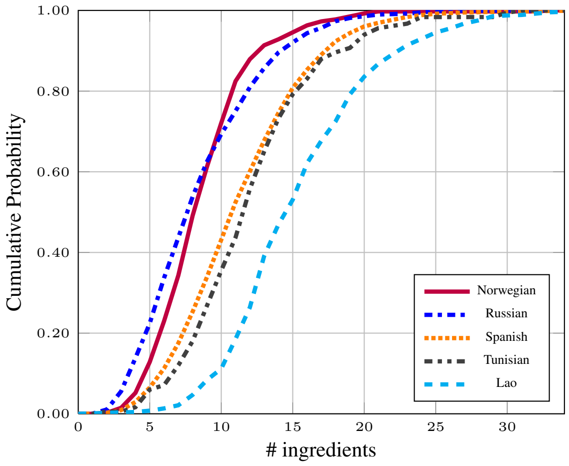


Figure 3: Cumulative complexity distribution for a number of sample cuisines.

[7] http://data.worldbank.org/indicator/SM.POP.NETM

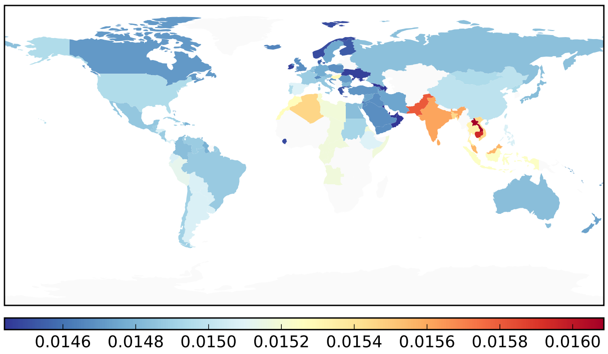


Figure 4: Complexity of dishes around the world

the area under CCD as a measure of complexity for a cuisine.

Figure 4 shows the complexity of dishes for different countries around the world. Here we have used the same approach as in Section 3.1 to map the cuisines to countries. We can see from the figure that except for some cases, the complexities are consistent with the diversities. This is due to the fact that as the number of available ingredients increases in a country (which is the result of global diversity,) people can leverage more ingredients and prepare more complex dishes. The exceptions here are China and India, two countries with the most population in the world. the complexity of dishes in these countries are relatively high, while their ingredients diversity is low. This can be the result of overpopulation or special culinary culture in these countries. Perhaps, these countries had or have good chefs that could cook more complex foods with the available ingredients!

3.3 Notable ingredients

Specific cuisines are mostly associated with different sets of ingredients due to the geographical locality of the ingredients. In this section, we study the most notable ingredient associated to some well-known cuisines using our dataset of recipes from Yummly.

We use the Term Frequency - Inverse Document Frequency (TF-IDF) calculation to find notable ingredients in each cuisine. In this approach, each ingredient is considered as an atomic word, and the collection of all the ingredients appeared within a cuisine is considered as a document. A TF-IDF calculation leads us to find the weight of different ingredients in the corpus of documents. This way, the importance of each ingredient within each cuisine is specified.

Figure 5 shows the top-50 most notable ingredients for Italian, Indian, and Mexican cuisines as a case study. We also looked at other similar cases, which we do not present here due to space constraints. According to the figure, the bigger the name of an ingredient, the more distinctive it is in its associated cuisine. We verified the soundness of results using Google Trends[8] service. For example, the term “Mozzarella” has the highest search frequency in Italy. Similarly, “Garam masala” is the most popular food additive in India according to its search volume.

4. CUISINE SIMILARITY

How do we determine the similarities between Korean and Japanese cuisine? What makes the Middle Eastern dishes seem similar to one another? In this section we answer these questions using a number of different methods and data.

4.1 Ingredient-based similarities

At first, we calculate the similarity between different cuisines based on the ingredients used in their recipes. To this end, we convert cuisines into vector space, representing each cuisine as a vector where each element indicates the frequency of an specific ingredient in that cuisine. Thereby, for each cuisine we obtain an ingredient-based feature vector which we leverage to calculate the similarity between different cuisines using the following metrics:

1. **Jensen-Shannon divergence:** If we normalize each ingredient-based feature vector such that the elements of a vector sum to one, then each vector will represent a probability distribution over standard ingredients. This way, we can use the distance measures proposed for probability distributions like Jensen-Shannon divergence. The Jensen-Shannon divergence between two probability distributions P and Q is defined as:

$$JS(P,Q) = \frac{1}{2}\left[KL(P \parallel M) + KL(Q \parallel M)\right] \quad (1)$$

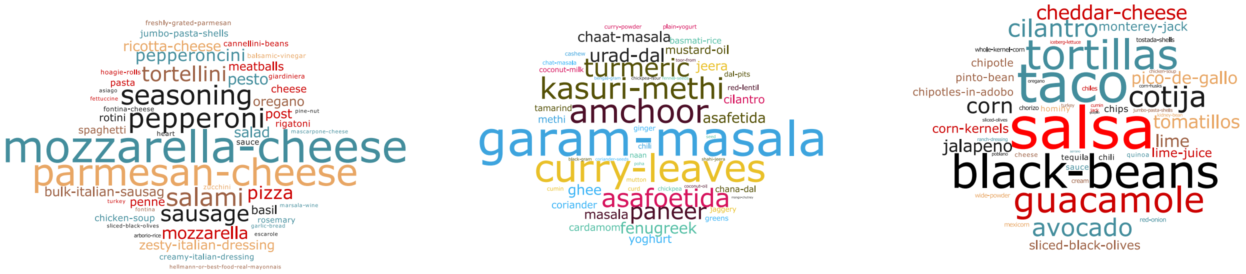
where $M = \frac{1}{2}(P+Q)$ and $KL(P \parallel M)$ is the Kullback-Leibler divergence from M to P. Since the Jensen-Shannon divergence is a distance measure between 0 and 1, we take $1 - JS(P,Q)$ as the similarity measure between two cuisines with their associated ingredient distributions P and Q. We have used Jensen-Shannon divergence instead of the simpler Kullback-Leibler divergence because $KL(P \parallel Q)$ goes to infinity when for an ingredient like i, $P(i)$ is non-zero while $Q(i)$ is. This case almost always happens in our data due to the geographical locality of ingredients. Therefore, we turned to Jensen-Shannon divergence which does not have this drawback.

2. **TF-IDF similarity:** Another measure we use to calculate the similarity between two cuisines is the well-known TF-IDF. Using this measure, we followed the approach described in section 3.3 to calculate the weight of ingredients in each cuisine. This way, each cuisine is represented as a vector where each of its elements indicates the representative power of the corresponding ingredient for that cuisine. The TF-IDF similarity between two cuisines is then simply the cosine similarity between their associated TF-IDF vectors.

Using the above similarity metrics, we calculated all similarities between each pair of cuisines. To assert the smoothness of ingredient distributions needed to compute Jensen-Shannon divergence, we limited our cuisines to those 82 ones having more than 100 recipes. Figure 6 illustrates the results for different similarity measures in a graph-based fashion. In these graphs, each node represents a cuisine and each cuisine is linked to its top-5 most similar cuisines. Link weights are proportional to the obtained similarity score between two endpoints. We colored each cuisine node according to the geographical region it resides in. The cuisines are classified into 9 regions which are North America, Latin America, Africa, Western Europe, Eastern Europe, Middle East, South Asia, East Asia, and Oceania. To visualize the graph, we have used *ForceAtlas* graph drawing algorithm implemented in Gephi tool [4]. This a force-directed algorithm which makes densely connected nodes to be grouped together [19, 24] and thus the communities become revealed.

We can see from the Figure 6 that using either of the ingredient-based similarity measures, the cuisines which reside in the same region are more similar to themselves and thus are grouped together. For example, we clearly see the clusters formed by Eastern and

[8] https://www.google.com/trends



(a) Italian **(b)** Indian **(c)** Mexican

Figure 5: Notable ingredients in three famous cuisines.

Southern Asian, Middle Eastern and African, Latin American, and Western European cuisines. Furthermore, due to the similarity of cultures in Europe and North America, and even Oceania, it can bee seen that clusters formed by the cuisines of these regions greatly overlap with each other.

4.2 Flavor-based similarities

In addition to the ingredient-based similarity, we calculate the similarity between cuisines in terms of the flavors provided in their recipes. This can help us understand how different cuisines are related to each other based on the taste of their dishes. As we mentioned in section 2.1, each recipe contains the flavor scores for six different flavors including saltiness, sourness, sweetness, bitterness, savoriness, and spiciness. To calculate the similarity between cuisines based on these flavors, like what we did for ingredient-based similarity, we consider each cuisine as a distribution over different flavors. Regarding the fact that different flavors of a recipe are correlated to each other – for instance, a dish can hardly be both sweet and spicy simultaneously – and due to the continuity of flavor scores, we hypothesize that the flavor scores are sampled from a multivariate Gaussian distribution, where each covariate corresponds a particular flavor. Considering this assumption, we fit a multivariate Gaussian distribution to each cuisine so that each cuisine become associated by a mean vector representing the average of flavor scores over all of its recipes, and a covariance matrix representing how flavors change relative to each other within that cuisine.

After fitting a multivariate Gaussian distribution to each cuisine using maximum likelihood estimation, we use Kullback-Leibler divergence to measure the distance between the distributions associated to each pair of cuisines. Since Kullback-Leibler divergence is and asymmetric distance measure, for each pair of cuisines with P and Q as their corresponding flavor distributions, we use $\left[\frac{1}{2}(KL(P \parallel Q) + KL(Q \parallel P))\right]^{-1}$ as a symmetric similarity measure between them.

Figure 7 shows the result of flavor-based similarity between different cuisines in a graph-based manner. We followed exactly the same steps as in Figure 6 to draw the graph, except that we used flavor-based similarity between cuisines. We can see from the figure that even though the flavors are not as much discriminant as ingredients, still we can observe some geographical patterns. For instance the clusters formed by Eastern Asian, Middle Eastern, Latin American, and Northern European cuisines are clear in this case as well. But what is obvious here is the fact that although there is a sense of taste similarity between the dishes from neighboring countries, the flavors are naturally shared all over the world.

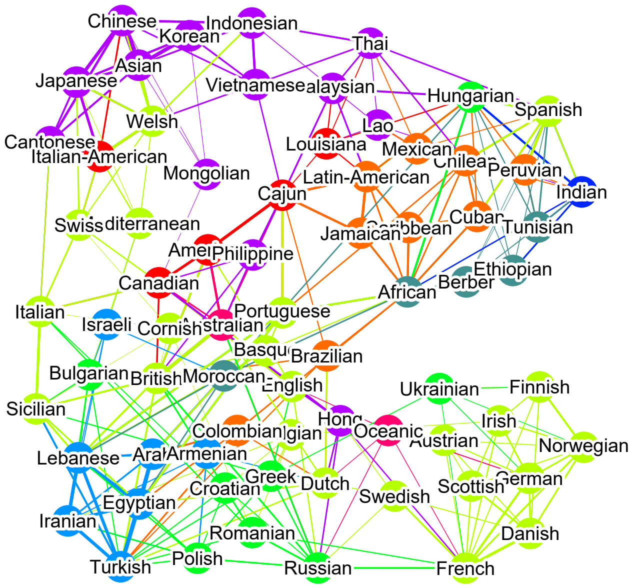
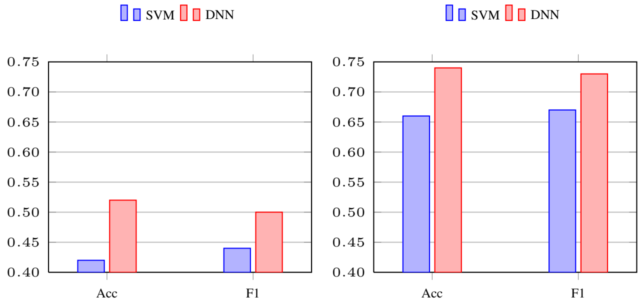


Figure 7: Graph of flavor-based similarity between different cuisines. Regional colors are the same as in Figure 6.

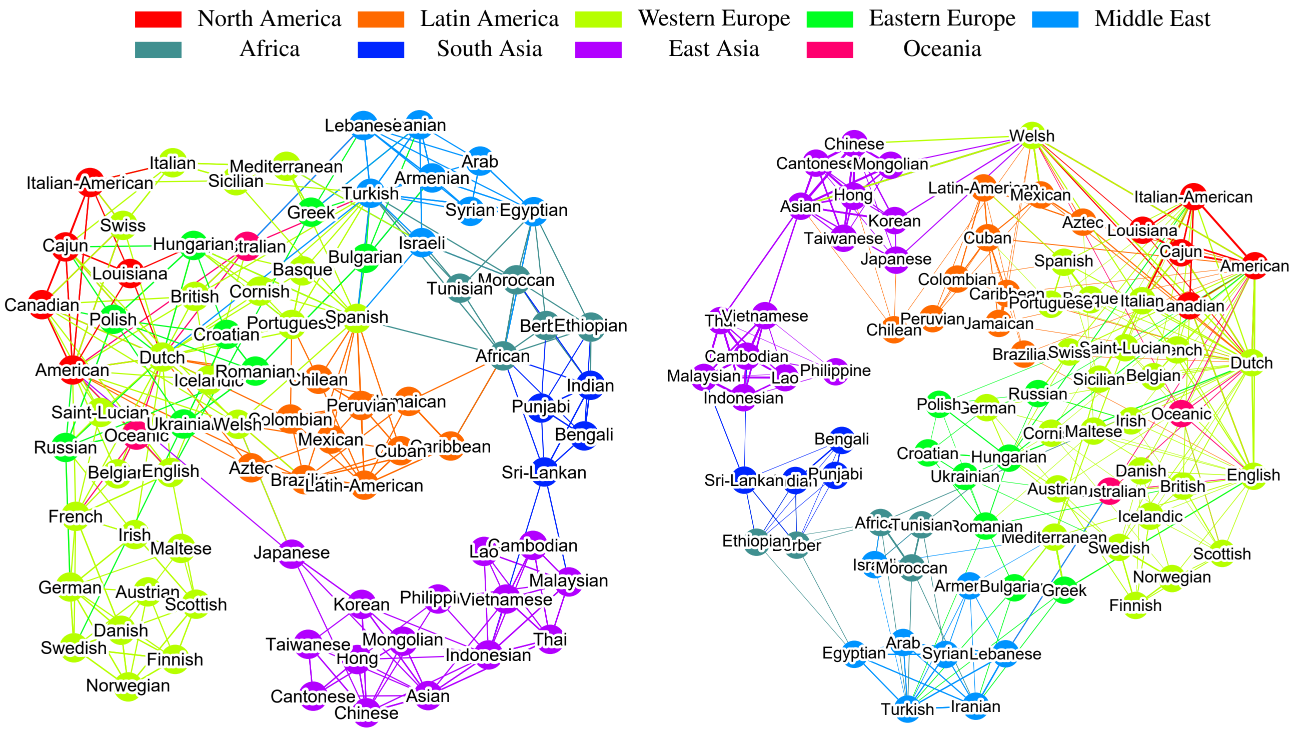


(a) Cuisine Prediction **(b)** Region Prediction

Figure 8: The prediction performance of different methods under different settings.

5. CUISINE PREDICTION

In this section we address the question of "*How good we can predict a recipe's cuisine, given its ingredients?*". To answer this question, We use two different classifiers, Support Vector Machine



(a) Jensen-Shannon divergence

(b) TF-IDF similarity

Figure 6: Graph of ingredient-based similarity between cuisines with different similarity measures.

(SVM), which is previously used in [22] for the same task, and Deep Neural Network (DNN), which is popular nowadays for classification purposes. To extract a feature vector for each recipe, we convert it into a boolean bag of words vector, considering each ingredient as an atomic word. Therefore, each recipe is represented as a vector with a length equal to the total number of ingredients, which is 3,286. The labeling of recipes are performed according to the following settings:

1. **Cuisine Prediction:** Each recipe is labeled to its cuisine. we consider 82 different cuisines having more than 100 recipes as different classes, resulting in about $100K$ recipes.
2. **Region Prediction:** Each recipe is labeled according to one of the 9 geographical regions where its cuisine belongs to. The regions are considered the same as in section 4. This results to have about $157K$ recipes.

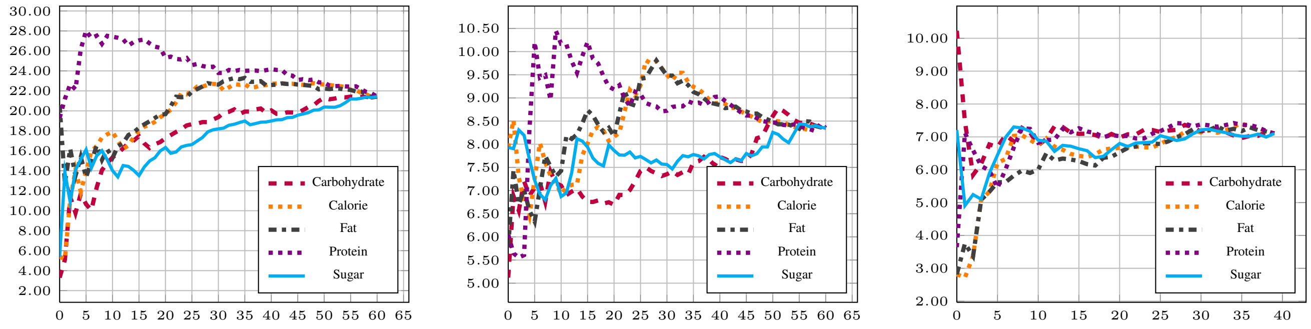
For multi-class classification with SVM, we use linear kernel with one vs. rest coding. The class imbalance problem is resolved with adjusting the weight of each cuisine inversely proportional to its frequency. The implementation is done using Scikit-learn machine learning library in python [5]. For DNN, we use Keras deep learning library [7] and create four dense hidden layers and a softmax output layer. Each of the first two hidden layers consists of 1000 neurons, and the two last ones each have 500 neurons. Dropout regularization [21] is used for all of the hidden layers. We use Adadelta [26] with default parameters as the optimizer. For both methods we take 80% of the data as training set and the remaining 20% as the test set. The prediction performance of both methods are evaluated under accuracy and F-measure.

Figure 8 shows the obtained results with both SVM and DNN methods. Figure 8a illustrates the results for cuisine prediction, while Figure 8b shows the results of region prediction task. As we can see in the figure, the DNN model performs about 24% better than SVM for cuisine prediction task under accuracy and over 13% better under F-measure. For region prediction task, since the number of classes are much fewer than cuisine prediction, both methods performed relatively better. In this case, the accuracy and F-measure achieved by the DNN model is about 12% and 9% better relative those achieved by SVM, respectively.

In order to get more intuition about the similarity of recipes in different regions, we bring the confusion matrix of the DNN model for region predictions in Table 1. Each region name is abbreviated in two letters. For example, LA means Latin American and AF means African cuisines. The number of correctly classified recipes are shown in bold and for each class, the greatest number of miss-classifications is shown in red. This table clearly demonstrates that the almost all of the miss-classifications are Western European. This is probably due to the huge ethnic composition of Western European countries which resulted in the diversity of culinary cultures of that region. The table shows that for some regions like Southern and Eastern Asian, the number of miss-classified recipes are somewhat low relative to the correctly classified ones. This result is as same as in Figure 6 in which these regions were almost disconnected from the others. On the other hand, for some cuisines like Oceanic, Eastern European, and Northern America, the number of miss-classifications are relatively high, mostly with Western European. Figure 6 shows clearly that the cultures in these regions are very similar to each other, mainly due to the common ethnics and history.

6. NUTRITION VALUES

In this section, we investigate the relation between the nutrition values of the recipes associated with countries and their hard measures of health, including obesity rate, diabetes rate and health ex-



(a) Nutritions vs obesity

(b) Nutritions vs diabetes

(c) Nutritions vs health expenditure

Figure 9: Smoothed Average Vector of nutritions and health measures.

Table 1: Confusion Matrix for DNN Region Prediction

		Prediction Outcome								
		LA	SA	OC	EA	AF	WE	ME	EE	NA
Actual Class	LA	**1888**	15	2	84	25	455	13	33	92
	SA	18	**961**	1	52	16	40	17	5	3
	OC	21	2	**177**	21	3	119	5	6	18
	EA	49	37	2	**5211**	13	342	26	24	51
	AF	31	23	2	28	**704**	136	57	7	15
	WE	453	49	21	660	85	**9430**	165	541	557
	ME	35	24	9	107	72	366	**634**	78	17
	EE	51	30	1	58	29	885	41	**1320**	94
	NA	127	7	5	128	22	1045	16	60	**1508**

penditure. Since the recipes are categorized by their cuisine but the health measures are reported by countries, we need to map each cuisine to one or more countries beforehand. To this end, we assign each cuisine to the countries that include that cuisine geographically or semantically. For example, we mapped the Kurdish cuisine to the countries: Iran, Syria, Iraq and Turkey because the Kurdistan region is divided between those countries. As another instance, we mapped the Italian-American cuisine to both Italy and USA countries because the cuisine represents both the Italian and American food culture. This way, we consider the recipes of each country as all of the recipes of the cuisines that belong to that country.

For each country, we calculate the average calorie, protein, fat, carbohydrate and sugar values over its recipes. We used the following three methods to study the relation between nutritions and health statistics:

1. **Pearson Correlation:** Pearson method calculates the linear correlation between the vectors u and v, where each element of u represents a nutrition value (e.g. sugar value) for a specific country, and each element of v represents a health measure (e.g. obesity rate) of that country.
2. **Kendall-Tau Correlation:** Kendall-Tau correlation is used to measure the ordinal correlation between u and v vectors.
3. **Smoothed Average Vector:** Given two vectors u and v, consider the vector u is sorted in ascending order as u'. Then v' would be the the reordered version of v to match the u'. Smoothed Average Vector for vectors u and v is a vector s with the same size where $s[i] = \frac{\sum_{j=1}^{i} v'[i]}{i}$. This measure captures the health trends with regards to the increases in consumption of nutritions. Moreover, It is also robust against the noisy variations of the underlying data.

Table 2 shows the correlation between different nutritions and the hard measures of health. As the results suggest, nutrition values show a significant correlation with the health related measures of countries. The dominant positive correlated nutritions are the sugar and carbohydrate. It is intuitive because those are the main elements of snack meals like cakes, creams, etc which can contribute to the health difficulties and the consequence expenditures eventually. On the other side, protein value shows strong negative correlation with the level of obesity and diabetes in societies. Noticeably, the positive impact of high-protein diets on losing weight is frequently studied in the literature [11]. Figure 9 exhibits the SAV for different nutritions and health measures. The trend of the diagrams endorses that including the countries with high average nutrition values (except protein) results in an increase in the average health measures (e.g. average obesity). Proteins show completely opposite patterns as expected. Including the countries with high protein diets decrease the rate of health difficulties (e.g. obesity or diabetes). A noticeable trait in both Table 2 and Figure 9 is that the correlations and trends are more highlighted in the obesity results rather than the diabetes and health expenditure. The reason is that the diabetes and health expenditure are more elaborate phenomena than the obesity. For example, in addition to consuming foods, there are a variety of other genetic and environmental factors that may cause the diabetes. Remarkably, the genetic susceptibility of different ethnics varies so much [9]. As another example, over intaking of proteins itself can lead to an spectrum of adverse effects [8]. Therefore the relation of protein intaking and health expenditures of the countries is not as clear as the relation between obesity and proteins.

7. RELATED WORK

Recently, public health has been increasingly analyzed through the lens of the web and social media. We refer the reader to [6] for an overview of the recent research in this area. Abbar et al. [1] relate food mentions on Twitter conversations to the obesity and diabetes rates, using caloric values, and find a high correlation (coefficient 0.77) between caloric values of tweets and obesity values in various states in the US. Low-obesity areas of USA have also been shown to be more socially active on Instagram (posting comments and likes) than those from high-obesity ones by Mejova et al. [17], who present a large-scale analysis of pictures taken at 164K restaurants in the US.

Ahn et al. [2] study culture-specific ingredient connections, creating a “flavor network” from a dataset of about 56K recipes and relating them to the geographical groupings of countries. Similar “flavor-based” food pairing studies are conducted on cuisines

Table 2: Countries Health Measures vs Their Recipes' Nutrition Values

		Correlation Values	
Health Measure	Nutrition	Pearson	Kendall-Tao
Obesity	Calorie	-0.104	-0.110
	Protein	$\mathbf{-0.483}$	$\mathbf{-0.299}$
	Fat	-0.115	-0.127
	Carbohydrate	0.300	0.201
	Sugar	**0.461**	**0.293**
Diabetes	Calorie	-0.077	-0.048
	Protein	$\mathbf{-0.162}$	-0.022
	Fat	-0.123	$\mathbf{-0.063}$
	Carbohydrate	**0.173**	**0.106**
	Sugar	0.142	0.066
Health Expend.	Calorie	0.098	0.110
	Protein	$\mathbf{-0.083}$	$\mathbf{-0.022}$
	Fat	**0.197**	**0.141**
	Carbohydrate	-0.064	-0.015
	Sugar	0.134	0.069

in distinct geographical areas such as India [12]. West et al. [25] mine logs of recipe-related queries to uncover temporal patterns in consumption. Using Fourier transforms, they show the yearly and weekly periodicity in food "density" of the searched recipes, with different trends in Southern and Northern hemispheres, suggesting a link between food selection and climate. A study of Austrian recipe sites by Wagner et al. [23] also highlights differences in the recipes of regions which are further apart. Zhu et al. [27] conduct a similar study on Chinese recipes to investigate the effect of geographical and climatic proximities on ingredients similarity of domestic cuisines.

Kular et al. [15] create a network of recipes using a dataset of 300 recipes from 15 different countries, and show the network's *small-world* and scale-free properties. As opposed to this line of work, we also exploit flavor and nutritional information, alongside health statistics countries to provide a deeper analysis about the dishes, cuisines, culinary cultures, and the impact of food on human life. Su et al. [22] investigate underlying connections between cuisines and ingredients via machine learning classification, with an application to predicting the cuisine by looking at recipes. Like our analysis, theirs is based on a large-scale data collection of recipes—specifically, 226K recipes collected from food.com. However, they only look at classifying cuisines using Support Vector Machine (SVM), while we propose a deep neural network architecture to capture the highly non-linear relation of a recipe cuisine and its associated ingredients. The results approve that the proposed deep model outperforms SVM by a significant margin in terms of prediction accuracy and F-1 measure.

There are major difference between our work and the ones discussed above, in both scale and domain. An important characteristic of our work comes from the size and the quality of the various datasets we used, which enable us to derive first-of-its-kind insight on worldwide cuisines and their relationship to health factors. In addition to the ingredients, we also exploited flavor and nutritional information, alongside health and immigration statistics, allowing us to perform a deeper analysis of the dishes, cuisines, culinary cultures, as well as the impact of food on human life.

8. CONCLUSION

This paper presented a large-scale study of user-generated recipes on the Web, their ingredients, nutrition, similarities across countries, and their relation with country health statistics. Our results have multiple implications: we found strong similarities between cuisines in neighboring countries, yet, the diversity of ingredients and flavors varies largely across the continents, mostly affected by net migration trends. *You are what you eat* might be a cliché, but we did find quantitative evidence of a strong correlation between nutrition information of the recipes (e.g., in terms of sugar intake) and obesity. Also, we demonstrated that deep learning can be used to effectively predicting cuisines from ingredients, potentially providing possibility for fine-grained analysis of food and dishes as well as improved recipe recommendations based on individuals' profile.

Our findings indicate that certain ingredients (e.g., mozzarella) uniquely represent a certain cuisine (e.g., Italian) and there are strong clusters of ingredients across neighboring countries. This feature eases the prediction of regions (e.g., continents) from the combination of ingredients in a cuisine. Moreover, the correlation between the ingredients and health conditions, such as diabetes, is naturally of great importance for public health experts, where behavior nudges or recommendation of similar dishes in flavor and ingredient complexity can be utilized to improve dietary intake [10, 20].

In future work, we plan to explore the possibility of recipe recommendation based on regional and personal tastes and user ratings. This is important as a local Chinese dish or a distinct flavor combination may be "alien" to, e.g., a Western person, but of interest to a Japanese individual. We also wish to asses the ability to model flavors with ingredients, and discover ingredients to match a specific flavor palette. Finding answers to these questions would provide a better understanding of the composition of flavors and ingredients in popular dishes and provide a better recommendation system for a healthier, tastier, and more diverse experience.

9. REFERENCES

[1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through Twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206, 2015.

[2] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási. Flavor network and the principles of food pairing. *Nature Scientific reports*, 2011.

[3] B. Armstrong and R. Doll. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International journal of cancer*, 15(4):617–631, 1975.

[4] M. Bastian, S. Heymann, M. Jacomy, et al. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *International AAAI Conference on Web and Social Media*, pages 361–362, 2009.

[5] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[6] D. Capurro, K. Cole, M. I. Echavarría, J. Joe, T. Neogi, and A. M. Turner. The use of social networking sites for public health practice and research: a systematic review. *Journal of medical Internet research*, 16(3):e79, 2014.

[7] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[8] I. Delimaris. Adverse effects associated with protein intake above the recommended dietary allowance for adults. *ISRN nutrition*, 2013, 2013.

[9] S. C. Elbein. Genetics factors contributing to type 2 diabetes across ethnicities. *Journal of diabetes science and technology*, 3(4):685–689, 2009.

[10] G. D. Foster, A. P. Makris, and B. A. Bailer. Behavioral treatment of obesity. *The American journal of clinical nutrition*, 82(1):230S–235S, 2005.

[11] T. L. Halton and F. B. Hu. The effects of high protein diets on thermogenesis, satiety and weight loss: a critical review. *Journal of the American College of Nutrition*, 23(5):373–385, 2004.

[12] A. Jain, N. Rakhi, and G. Bagler. Analysis of food pairing in regional cuisines of india. *PloS one*, 10(10):e0139539, 2015.

[13] A. Keys, A. Mienotti, M. J. Karvonen, C. Aravanis, H. Blackburn, R. Buzina, B. Djordjevic, A. Dontas, F. Fidanza, M. H. Keys, et al. The diet and 15-year death rate in the seven countries study. *American journal of epidemiology*, 124(6):903–915, 1986.

[14] M. Kratz, T. Baars, and S. Guyenet. The relationship between high-fat dairy consumption and obesity, cardiovascular, and metabolic disease. *European journal of nutrition*, 52(1):1–24, 2013.

[15] D. K. Kular, R. Menezes, and E. Ribeiro. Using network analysis to understand the relation between cuisine and culture. In *Network Science Workshop (NSW), 2011 IEEE*, pages 38–45, 2011.

[16] Y. Mejova, S. Abbar, and H. Haddadi. Fetishizing food in digital age:# foodporn around the world. In *International AAAI Conference on Web and Social Media (ICWSM 2016)*, 2016.

[17] Y. Mejova, H. Haddadi, A. Noulas, and I. Weber. #FoodPorn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 51–58, 2015.

[18] M. Michel de Lorgeril, P. Salen, J.-L. Martin, I. Monjaud, J. Delaye, and N. Mamelle. Mediterranean diet, traditional risk factors, and the rate of cardiovascular complications after myocardial infarction. *Heart failure*, 11:6, 1999.

[19] A. Noack. Modularity clustering is force-directed layout. *Phys. Rev. E*, 79:026102, Feb 2009.

[20] N. Regulating. Judging nudging: can nudging improve population health? *Bmj*, 342:263, 2011.

[21] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[22] H. Su, T.-W. Lin, C.-T. Li, M.-K. Shan, and J. Chang. Automatic recipe cuisine classification by ingredients. In *Proceedings of the 2014 ACM Joint Conference on Pervasive and Ubiquitous Computing*, pages 565–570, 2014.

[23] C. Wagner, P. Singer, and M. Strohmaier. Spatial and Temporal Patterns of Online Food Preferences. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 553–554. International World Wide Web Conferences Steering Committee, 2014.

[24] L. Waltman, N. J. van Eck, and E. C. Noyons. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4):629–635, 2010.

[25] R. West, R. W. White, and E. Horvitz. From cookies to cooks: insights on dietary patterns via analysis of web usage logs. In *WWW*, 2013.

[26] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[27] Y.-X. Zhu, J. Huang, Z.-K. Zhang, Q.-M. Zhang, T. Zhou, and Y.-Y. Ahn. Geography and similarity of regional cuisines in china. *PloS one*, 8(11):e79161, 2013.