Supporting Online Material

Flavor network and the principles of food pairing
by Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, Albert-László Barabási

Table of Contents

List of Figures

List of Tables

S1 Materials and methods

S1.1 Flavor network

S1.1.1 Ingredient-compounds bipartite network

The starting point of our research is Fenaroli's handbook of flavor ingredients (fifth edition [1]), which offers a systematic list of flavor compounds and their natural occurrences (food ingredients). Two post-processing steps were necessary to make the dataset appropriate for our research: (A) In many cases, the book lists the essential oil or extract instead of the ingredient itself. Since these are physically extracted from the original ingredient, we associated the flavor compounds in the oils and extracts with the original ingredient. (B) Another post-processing step is including the flavor compounds of a more general ingredient into a more specific ingredient. For instance, the flavor compounds in 'meat' can be safely assumed to also be in 'beef' or 'pork'. 'Roasted beef' contains all flavor compounds of 'beef' and 'meat'.

The ingredient-compound association extracted from [1] forms a bipartite network. As the name suggests, a bipartite network consists of two types of nodes, with connections only between nodes of different types. Well known examples of bipartite networks include collaboration networks of scientists [2] (with scientists and publications as nodes) and actors [3] (with actors and films as nodes), or the human disease network [4] which connects health disorders and disease genes. In the particular bipartite network we study here, the two types of nodes are food ingredients and flavor compounds, and a connection signifies that an ingredient contains a compound.

The full network contains 1,107 chemical compounds and 1,531 ingredients, but only 381 ingredients appear in recipes, together containing 1,021 compounds (see Fig. S1). We project this network into a weighted network between ingredients only [5, 6, 7, 8]. The weight of each edge w_{ij} is the number of compounds shared between the two nodes (ingredients) i and j, so that the relationship between the $M \times M$ weighted adjacency matrix w_{ij} and the $N \times M$ bipartite adjacency

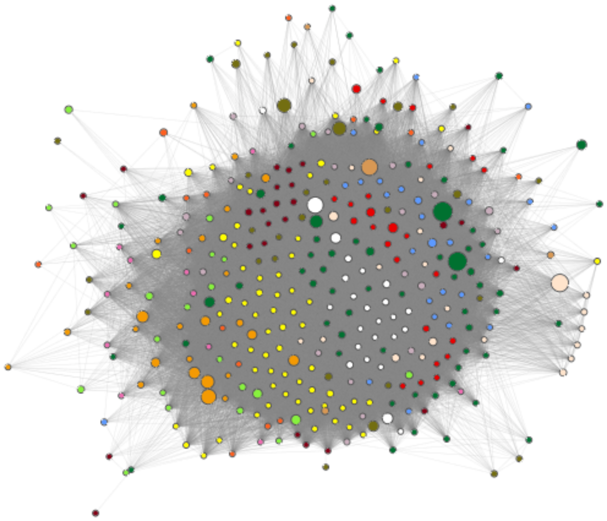

Figure S1: The full flavor network. The size of a node indicates average prevalence, and the thickness of a link represents the number of shared compounds. All edges are drawn. It is impossible to observe individual connections or any modular structure.

	3rd eds.	5th eds.
# of ingredients	916	1507
# of compounds	629	1107
# of edges in I-C network	6672	36781

Table S1: The basic statistics on two different datasets. The 5th Edition of Fenaroli's handbook contains much more information than the third edition.

matrix a_{ik} (for ingredient i and compound k) is given by:

$$w_{ij} = \sum_{k=1}^{N} a_{ik} a_{jk} \tag{S1}$$

The degree distributions of ingredients and compounds are shown in Fig. S2.

S1.1.2 Incompleteness of data and the third edition

The situation encountered here is similar to the one encountered in systems biology: we do not have a complete database of all protein, regulatory and metabolic interactions that are present in the cell. In fact, the existing protein interaction data covers *less than 10% of all protein interactions* estimated to be present in the human cell [9].

To test the robustness of our results against the incompleteness of data, we have performed

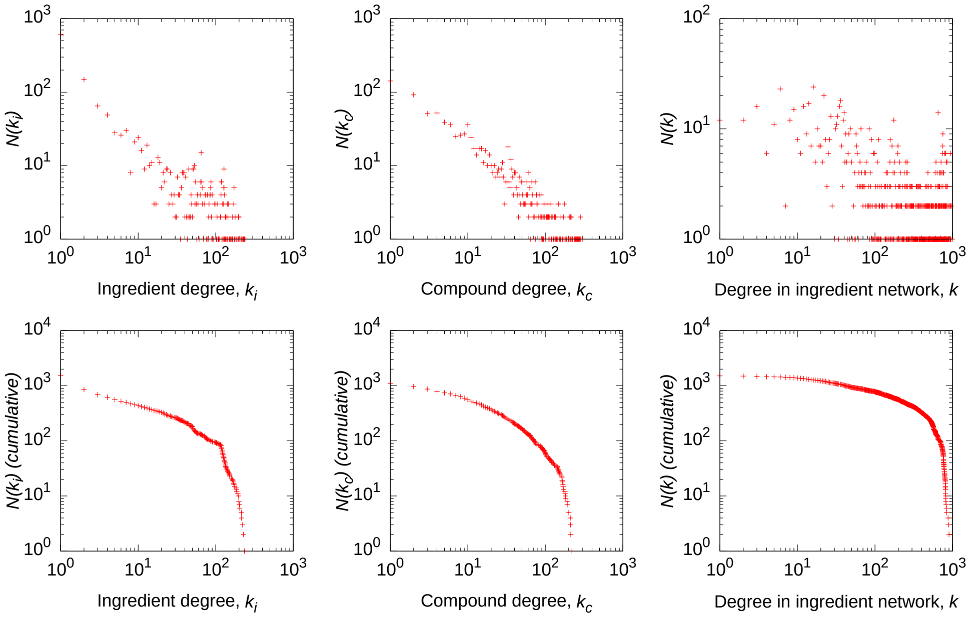

Figure S2: Degree distributions of the flavor network. Degree distribution of ingredients in the ingredient-compound network, degree distribution of flavor compounds in the ingredient-compound network, and degree distribution of the (projected) ingredient network, from left to right. Top: degree distribution. Bottom: complementary cumulative distribution. The line and the exponent value in the leftmost figure at the bottom is purely for visual guide.

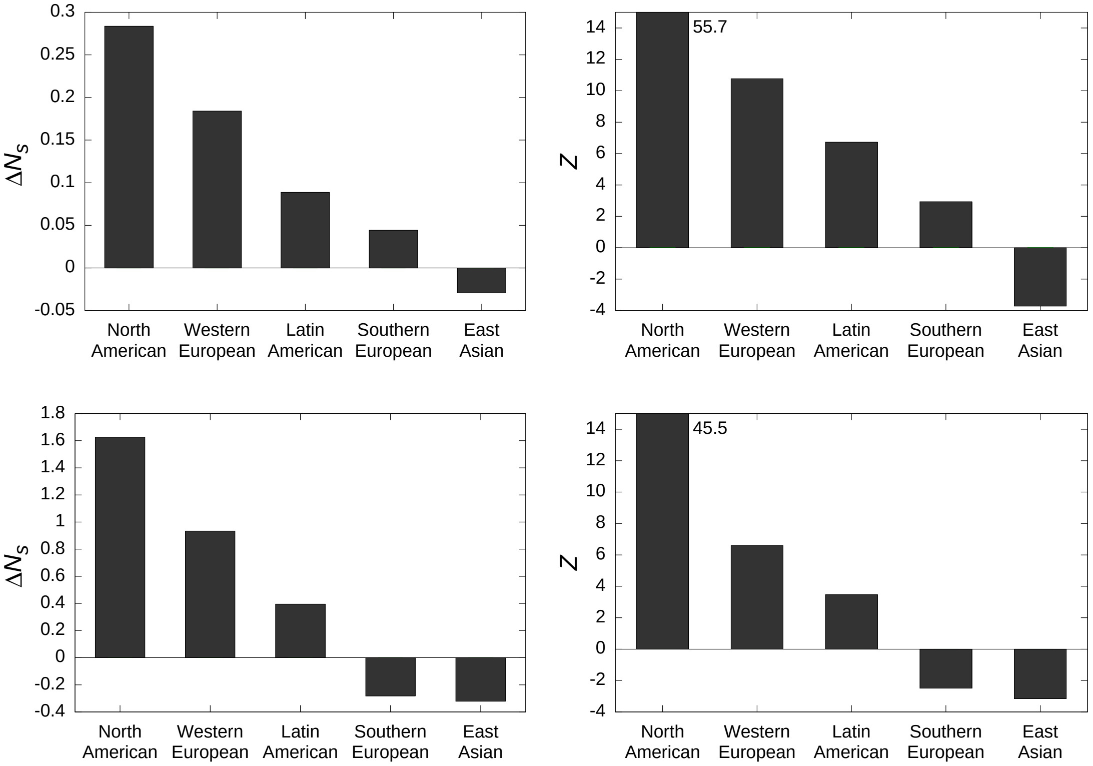

Figure S3: Comparing the third and fifth edition of Fenaroli's to see if incomplete data impacts our conclusions. The much sparser data of the 3rd edition (**Top**) shows a very similar trend to that of the 5th edition (**Bottom**, repeated from main text Fig. 3). Given the huge difference between the two editions (Table S1), this further supports that the observed patterns are robust.

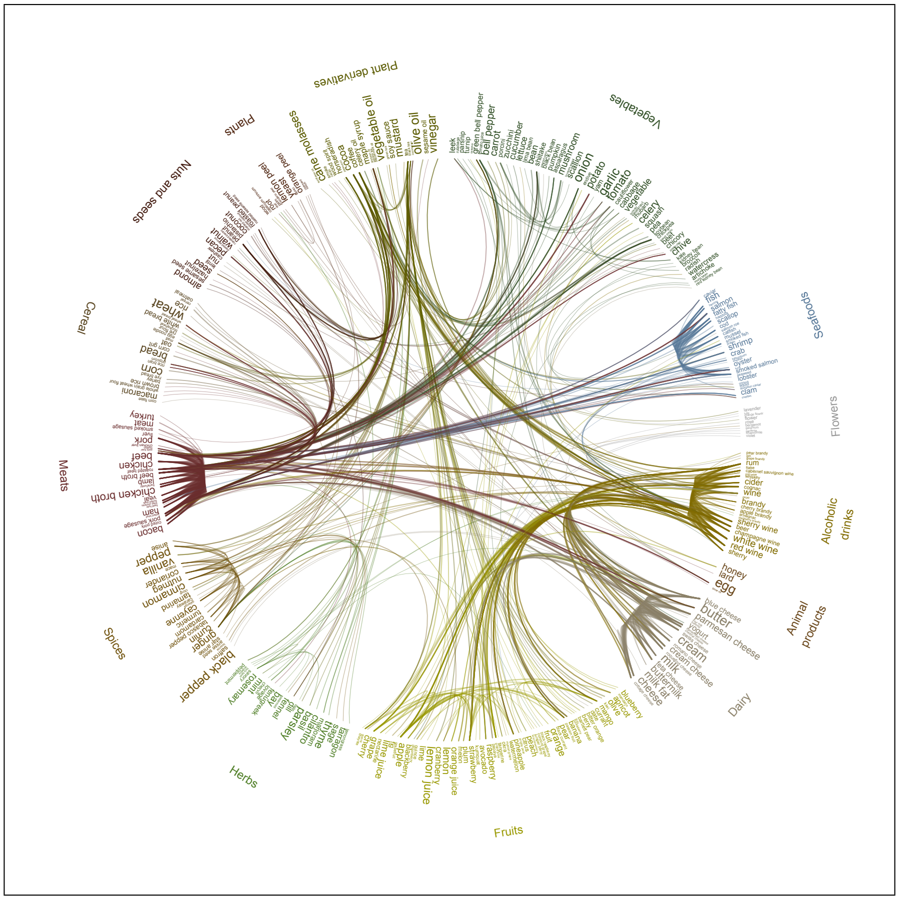

Figure S4: The backbone of the ingredient network extracted according to [10] with a significance threshold $p = 0.04$. Color indicates food category, font size reflects ingredient prevalence in the dataset, and link thickness represents the number of shared compounds between two ingredients.

the same calculations for the 3rd edition of Fenaroli's handbook as well. The 5th edition contains approximately six times more information on the chemical contents of ingredients (Table S1). Yet, our main result is robust (Fig. S3), further supporting that data incompleteness is not the main factor behind our findings.

S1.1.3 Extracting the backbone

The network's average degree is about 214 (while the number of nodes is 381). It is very dense and thus hard to visualize (see Fig. S1). To circumvent this high density, we use a method that extracts the backbone of a weighted network [10], along with the method suggested in [11]. For each node, we keep those edges whose weight is statistically significant given the strength (sum of weight) of the node. If there is none, we keep the edge with the largest weight. A different visualization of this backbone is presented in Fig. S4. Ingredients are grouped into categories and the size of the name indicates the prevalence. This representation clearly shows the categories that are closely connected.

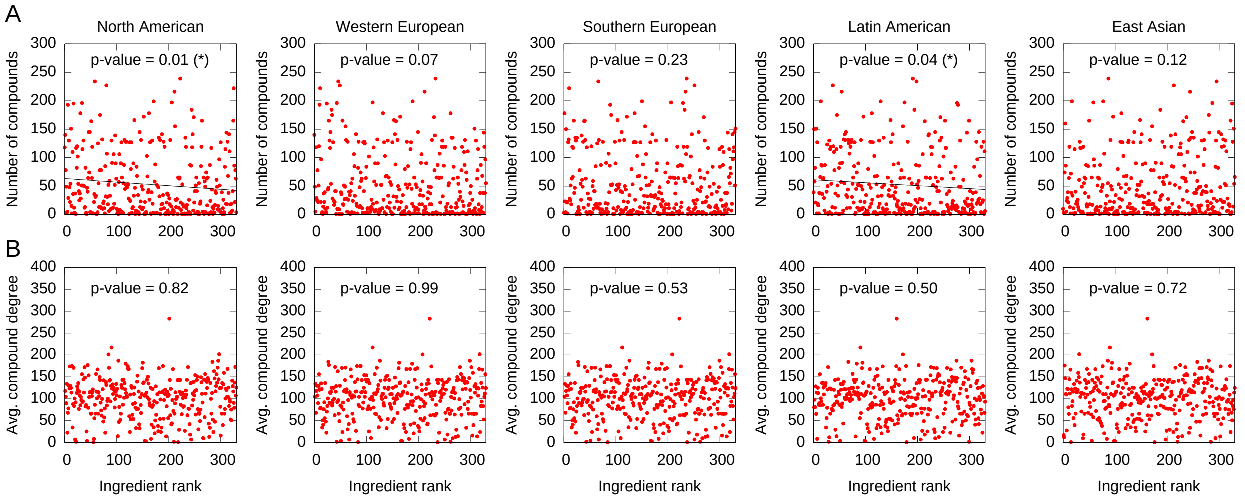

Figure S5: Are popular, much-used ingredients more studied than less frequent foods, leading to potential systematic bias? (**A**) We plot the number of flavor compounds for each ingredient as a function of the (ranked) popularity of the ingredient. The correlation is very small compared to the large fluctuations present. There is a weak tendency that the ingredients mainly used in North American or Latin American cuisine tend to have more odorants, but the correlations are weak (with coefficients of -0.13 and -0.10 respectively). A linear regression line is shown only if the corresponding p-value is smaller than 0.05. (**B**) If there is bias such that the book tends to list more familiar ingredients for more common flavor compounds, then we can observe the correlation between the familiarity (how frequently it is used in the cuisine) and the degree of the compound in the ingredient-compound network. The plots show no observable correlations for any cuisine.

S1.1.4 Sociological bias

Western scientists have been leading food chemistry, which may imply that western ingredients are more studied. To check if such a bias is present in our dataset, we first made two lists of ingredients: one is the list of ingredients appearing in North American cuisine, sorted by the relative prevalence p_i^c (i.e. the ingredients more specific to North American cuisine comes first). The other is a similar list for East Asian cuisine. Then we measured the number of flavor compounds for ingredients in each list. The result in Fig. S5A shows that any potential bias, if present, is not significant.

There is another possibility, however, if there is bias such that the dataset tends to list more familiar (Western) ingredients for more common flavor compounds, then we should observe a correlation between the familiarity (frequently used in Western cuisine) and the degree of compound (number of ingredients it appears in) in the ingredient. Figure S5B shows no observable correlation, however.

S1.2 Recipes

The number of potential ingredient combinations is enormous. For instance, one could generate $\sim 10^{15}$ distinct ingredient combinations by choosing eight ingredients (the current average per recipe) from approximately 300 ingredients in our dataset. If we use the numbers reported in Kinouchi et al. [12] (1000 ingredients and 10 ingredients per recipe), one can generate $\sim 10^{23}$ ingredient combinations. This number greatly increases if we consider the various cooking methods.

Table S2: Number of recipes and the detailed cuisines in each regional cuisine in the recipe dataset. Five groups have reasonably large size. We use all cuisine data when calculating the relative prevalence and flavor principles.

Cuisine set	Number of recipes	Cuisines included
North American	41525	American, Canada, Cajun, Creole, Southern soul food, Southwestern U.S.
Southern European	4180	Greek, Italian, Mediterranean, Spanish, Portuguese
Latin American	2917	Caribbean, Central American, South American, Mexican
Western European	2659	French, Austrian, Belgian, English, Scottish, Dutch, Swiss, German, Irish
East Asian	2512	Korean, Chinese, Japanese
Middle Eastern	645	Iranian, Jewish, Lebanese, Turkish
South Asian	621	Bangladeshian, Indian, Pakistani
Southeast Asian	457	Indonesian, Malaysian, Filipino, Thai, Vietnamese
Eastern European	381	Eastern European, Russian
African	352	Moroccan, East African, North African, South African, West African
Northern European	250	Scandinavian

Regardless, the fact that this number exceeds by many orders of magnitude the $\sim 10^6$ recipes listed in the largest recipe repositories (e.g. `http://cookpad.com`) indicates that humans are exploiting a tiny fraction of the culinary space.

We downloaded all available recipes from three websites: *allrecipes.com*, *epicurious.com*, and *menupan.com*. Recipes tagged as belonging to an ethnic cuisine are extracted and then grouped into 11 larger regional groups. We used only 5 groups that each contain more than 1,000 recipes (See Table S2). In the curation process, we made a replacement dictionary for frequently used phrases that should be discarded, synonyms for ingredients, complex ingredients that are broken into ingredients, and so forth. We used this dictionary to automatically extract the list of ingredients for each recipe. As shown in Fig. 1D, the usage of ingredients is highly heterogenous. Egg, wheat, butter, onion, garlic, milk, vegetable oil, and cream appear more than 10,000 recipes while geranium, roasted hazelnut, durian, muscat grape, roasted pecan, roasted nut, mate, jasmine tea, jamaican rum, angelica, sturgeon caviar, beech, lilac flower, strawberry jam, and emmental cheese appear in only one recipe. Table S3 shows the correlation between ingredient usage frequency in each cuisine and in each dataset. Figure. S6 shows that the three datasets qualitatively agree with each other, offering a base to combine these datasets.

S1.2.1 Size of recipes

We reports the size of the recipes for each cuisine in Table S4. Overall, the mean number of ingredients per recipe is smaller than that reported in Kinouchi et al. [12]. We believe that it is

	Epicurious vs. Allrecipes	Epicurious vs. Menupan	Allrecipes vs. Menupan
North American	0.93	N/A	N/A
East Asian	0.94	0.79	0.82
Western European	0.92	0.88	0.89
Southern European	0.93	0.83	0.83
Latin American	0.94	0.69	0.74
African	0.89	N/A	N/A
Eastern European	0.93	N/A	N/A
Middle Eastern	0.87	N/A	N/A
Northern European	0.77	N/A	N/A
South Asian	0.97	N/A	N/A
Southeast Asian	0.92	N/A	N/A

Table S3: The correlation of ingredient usage between different datasets. We see that the different datasets broadly agree on what constitutes a cuisine, at least at a gross level.

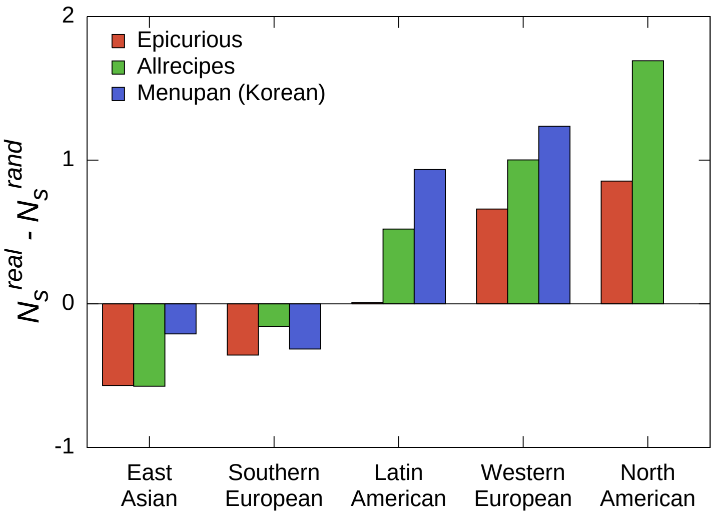

Figure S6: Comparison between different datasets. The results on different datasets qualitatively agree with each other (except Latin American cuisine). Note that *menupan.com* is a Korean website.

North American	7.96
Western European	8.03
Southern European	8.86
Latin American	9.38
East Asian	8.96
Northern European	6.82
Middle Eastern	8.39
Eastern European	8.39
South Asian	10.29
African	10.45
Southeast Asian	11.32

Table S4: Average number of ingredients per recipe for each cuisine.

mainly due to the different types of data sources. There are various types of recipes: from quick meals to ones used in sophisticated dishes of expensive restaurants; likewise, there are also various cookbooks. The number of ingredients may vary a lot between recipe datasets. If a book focuses on sophisticated, high-level dishes then it will contain richer set of ingredients per recipe; if a book focuses on simple home cooking recipes, then the book will contain fewer ingredients per recipe. We believe that the online databases are close to the latter; simpler recipes are likely to dominate the database because anyone can upload their own recipes. By contrast, we expect that the cookbooks, especially the canonical ones, contain more sophisticated and polished recipes, which thus are more likely to contain more ingredients.

Also, the pattern reported in Kinouchi et al. [12] is reversed in our dataset: Western European cuisine has 8.03 ingredients per recipe while Latin American cuisine has 9.38 ingredients per recipe. Therefore, we believe that there is no clear tendency of the number of ingredients per recipe between Western European and Latin American cuisine.

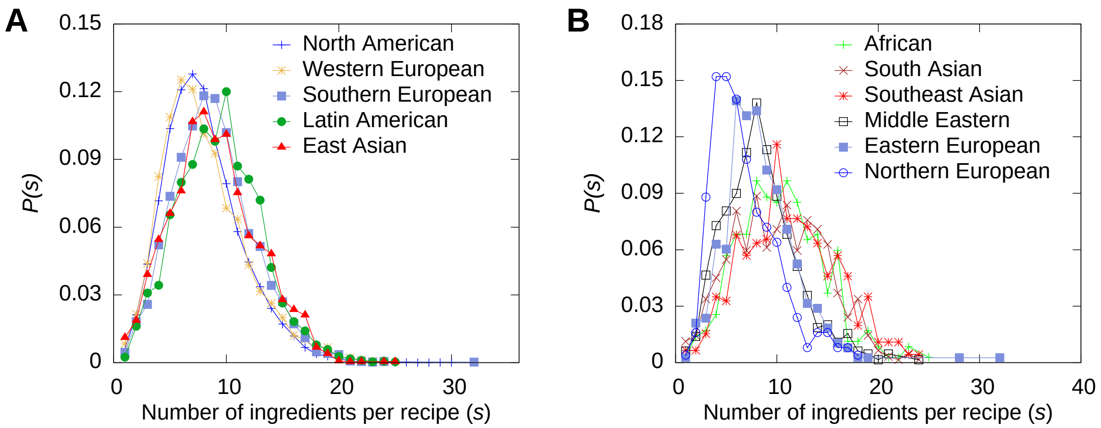

Figure S7: Number of ingredients per recipe. North American and Western European cuisine shows similar distribution while other cuisines have slightly more ingredients per recipe.

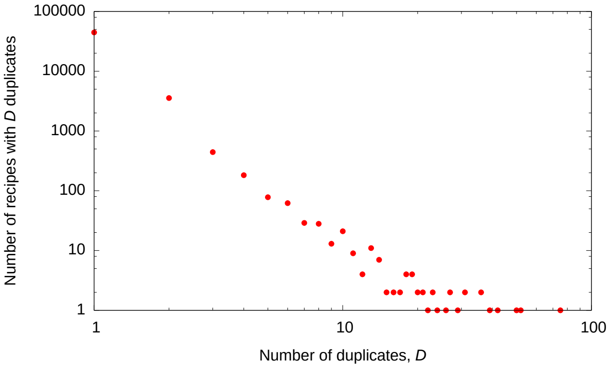

Figure S8: If a recipe is very popular, the recipe databases will have a tendency to list more variations of the recipe. This plot shows that there are many duplicated recipes that share the same set of ingredients. The number of duplicates exhibits a heavy-tailed distribution.

Yet, there seems to be an interesting trend in our dataset that hotter countries use more ingredients per recipe, probably due to the use of more herbs and spices [13, 14] or due to more diverse ecosystems. (6.82 in Northern European vs. 11.31 in Southeast Asian). Figure S7 shows the distribution of recipe size in all cuisines.

S1.2.2 Frequency of recipes

In contrast to previous work [12] that used published cookbooks, we use online databases. Although recipes online are probably less *canonical* than established cookbooks, online databases allow us to study much larger dataset more easily. Another important benefit of using online databses is that there is no *real-estate issue* in contrast to physical cookbooks that should carefully choose what to include. Adding a slight variation of a recipe costs virtually nothing to the websites and even enhances the quality of the database. Therefore, one can expect that online databases capture the frequency of recipes more accurately than cookbooks.

Certain recipes (e.g. signature recipes of a cuisine) are much more important than others; They are cooked much more frequently than others. Figure S8 shows that there are many duplicated recipes (possessing identical sets of ingredients), indicating that popularity is naturally encoded in these datasets.

S1.3 Number of shared compounds

Figure S9 explains how to measure the number of shared compounds in a hypothetical recipe with three ingredients.

S1.4 Shared compounds hypothesis

S1.4.1 Null models

In order to test the robustness of our findings, we constructed several random recipe datasets using a series of appropriate null models and compare the mean number of shared compounds N_s between the real and the randomized recipe sets. The results of these null models are summarized in Fig. S10, each confirming the trends discussed in the paper. The null models we used are:

(A, B) Frequency-conserving. Cuisine c uses a set of n_c ingredients, each with frequency f_i. For a given recipe with N_i ingredients in this cuisine, we pick N_i ingredients randomly from the set of all n_c ingredients, according to f_i. That is the more frequently an ingredient is used, the more likely the ingredient is to be picked. It preserves the prevalence of each ingredient. This is the null model presented in the main text.

(C, D) Frequency and ingredient category preserving. With this null model, we conserve the category (meats, fruits, etc) of each ingredient in the recipe, and when sample random ingredients proportional to the prevalence. For instance, a random realization of a recipe with beef and onion will contain a meat and a vegetable. The probability to pick an ingredient is proportional to the prevalence of the ingredient in the cuisine.

(E, F) Uniform random. We build a random recipe by randomly choosing an ingredient that is used at least once in the particular cuisine. Even very rare ingredients will frequently appear in random recipes.

(G, H) Uniform random, ingredient category preserving. For each recipe, we preserve the category of each ingredient, but not considering frequency of ingredients.

Although these null models greatly change the frequency and type of ingredients in the random recipes, North American and East Asian recipes show a robust pattern: North American recipes always share more flavor compounds than expected and East Asian recipes always share less flavor compounds than expected. This, together with the existence of both positive and negative N_s^{real} −

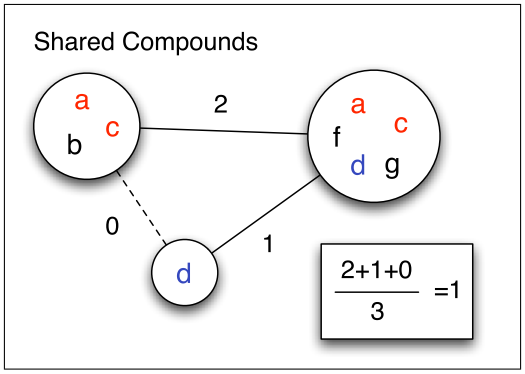

Figure S9: For a recipe with three ingredients, we count the number of shared compounds in every possible pair of ingredients, and divide it by the number of possible pair of ingredients.

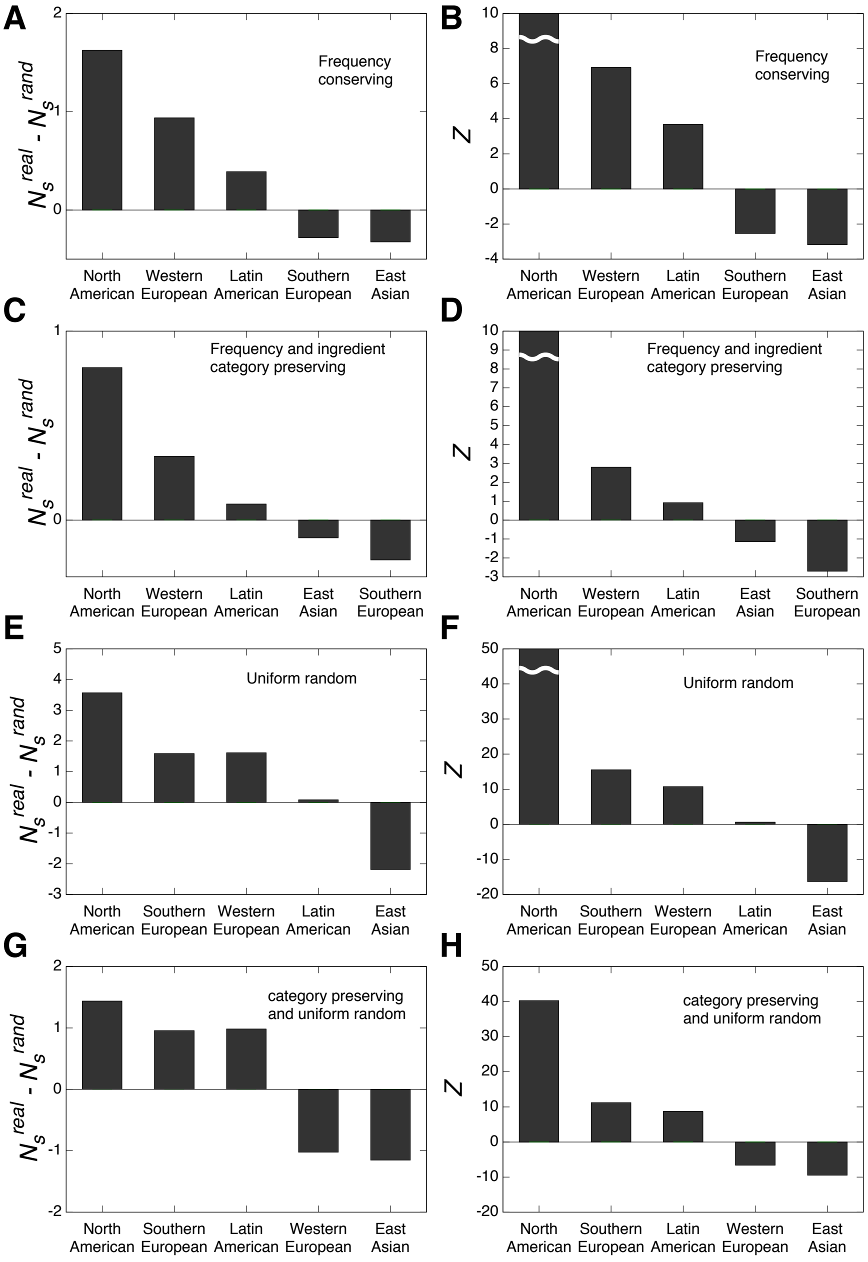

Figure S10: Four different null models. Although the size of the discrepancy between cuisines varies greatly, the overall trend is stable.

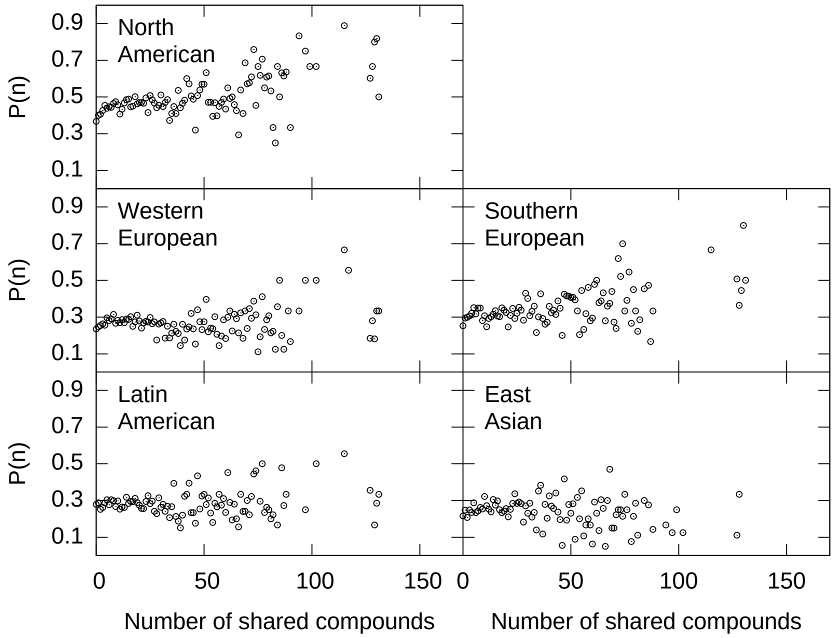

Figure S11: The probability that ingredient pairs that share a certain number of compounds also appear in the recipes. We enumerate every possible ingredient pair in each cuisine and show the fraction of pairs in recipes as a function of the number of shared compounds. To reduce noise, we only used data points calculated from more than 5 pairs.

N_s^{rand} in every null model, indicates that the patterns we find are not due to a poorly selected null models.

Finally, Fig. S11 shows the probability that a given pair with certain number of shared compounds will appear in the recipes, representing the raw data behind the generalized food-pairing hypothesis discussed in the text. To reduce noise, we only consider N_s where there are more than five ingredient pairs.

S1.4.2 Ingredient contributions

To further investigate the contrasting results on the shared compound hypothesis for different cuisines, we calculate the contribution of each ingredient and ingredient pair to ΔN_s. Since $N_s(R)$ for a recipe R is defined as

$$N_s(R) = \frac{2}{n_R(n_R-1)} \sum_{i,j \in R, i \neq j} \left| C_i \cap C_j \right| \tag{S2}$$

(where n_R is the number of ingredients in the recipe R), the contribution from an ingredient pair (i,j) can be calculated as following:

$$\chi_{ij}^c = \left(\frac{1}{N_c} \sum_{R \ni i,j} \frac{2}{n_R(n_R-1)} \left| C_i \cap C_j \right| \right) - \left(\frac{f_i f_j}{N_c^2} \frac{2}{\langle n_R \rangle (\langle n_R \rangle - 1)} \left| C_i \cap C_j \right| \right), \tag{S3}$$

where f_i indicates the ingredient i's number of occurrences. Similarly, the individual contribution can be calculated:

$$\chi_i^c = \left(\frac{1}{N_c} \sum_{R \ni i} \frac{2}{n_R(n_R-1)} \sum_{j \neq i (j, i \in R)} \left| C_i \cap C_j \right| \right) - \left(\frac{2 f_i}{N_c \langle n_R \rangle} \frac{\sum_{j \in c} f_j \left| C_i \cap C_j \right|}{\sum_{j \in c} f_j} \right). \tag{S4}$$

We list in Table. S5 the top contributors in North American and East Asian cuisines.

	North American		East Asian	
	Ingredient *i*	χ_i	Ingredient *i*	χ_i
Positive	milk	0.529	rice	0.294
	butter	0.511	red bean	0.152
	cocoa	0.377	milk	0.055
	vanilla	0.239	green tea	0.041
	cream	0.154	butter	0.041
	cream cheese	0.154	peanut	0.038
	egg	0.151	mung bean	0.036
	peanut butter	0.136	egg	0.033
	strawberry	0.106	brown rice	0.031
	cheddar cheese	0.098	nut	0.024
	orange	0.095	mushroom	0.022
	lemon	0.095	orange	0.016
	coffee	0.085	soybean	0.015
	cranberry	0.070	cinnamon	0.014
	lime	0.065	enokidake	0.013
Negative	tomato	-0.168	beef	-0.2498
	white wine	-0.0556	ginger	-0.1032
	beef	-0.0544	pork	-0.0987
	onion	-0.0524	cayenne	-0.0686
	chicken	-0.0498	chicken	-0.0662
	tamarind	-0.0427	onion	-0.0541
	vinegar	-0.0396	fish	-0.0458
	pepper	-0.0356	bell pepper	-0.0414
	pork	-0.0332	roasted sesame seed	-0.0410
	celery	-0.0329	black pepper	-0.0409
	bell pepper	-0.0306	shrimp	-0.0408
	red wine	-0.0271	shiitake	-0.0329
	black pepper	-0.0248	garlic	-0.0302
	parsley	-0.0217	carrot	-0.0261
	parmesan cheese	-0.0197	tomato	-0.0246

Table S5: Top 15 (both positive and negative) contributing ingredients to each cuisine.

References

[1] Burdock, G. A. *Fenaroli's handbook of flavor ingredients* (CRC Press, 2004), 5th edn.

[2] Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98**, 404 (2001).

[3] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Natue* **393**, 440 (1998).

[4] Goh, K.-I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685 (2007).

[5] Newman, M. E. J., Barabási, A.-L. & Watts, D. J. *The structure and dynamics of networks:* (Princeton University Press, 2006).

[6] Caldarelli, G. *Scale-free networks: complex webs in nature and technology* (Oxford University Press, USA, 2007).

[7] Dorogovtsev, S. N. & Mendes, J. F. F. *Evolution of networks: from biological nets to the internet and WWW* (Oxford University Press, USA, 2003).

[8] Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Nat. Acad. Sci.* **111**, 3747 (2004).

[9] Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nature Methods* **6**, 83–90 (2009).

[10] Serrano, M. A., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences* **106**, 6483–6488 (2009).

[11] Lee, S. H., Kim, P.-J., Ahn, Y.-Y. & Jeong, H. Googling social interactions: web search engine based social network construction. *PLoS One* **5**, e11233 (2010).

[12] Kinouchi, O., Diez-Garcia, R. W., Holanda, A. J., Zambianchi, P. & Roque, A. C. The non-equilibrium nature of culinary evolution. *New Journal of Physics* **10**, 073020 (2008).

[13] Billing, J. & Sherman, P. W. Antimicrobial functions of spices: why some like it hot. *The Quarterly Review of Biology* **73**, 3–49 (1998).

[14] Sherman, P. W. & Hash, G. A. Why vegetable recipes are not very spicy. *Evolution and Human Behavior* **22**, 147–163 (2001).