Wagner et al. *EPJ Data Science* (2014) 3:38
DOI 10.1140/epjds/s13688-014-0036-7

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

The nature and evolution of online food preferences

Claudia Wagner[1,2*], Philipp Singer[1] and Markus Strohmaier[1,2]

*Correspondence:
claudia.wagner@gesis.org
[1]GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 5-8, Cologne, Germany
[2]University of Koblenz-Landau, Koblenz, Germany

Abstract

Food is a central element of humans? life, and food preferences are amongst others manifestations of social, cultural and economic forces that influence the way we view, prepare and consume food. Historically, data for studies of food preferences stems from consumer panels which continuously capture food consumption and preference patterns from individuals and households. In this work we look at a new source of data, i.e., server log data from a large recipe platform on the World Wide Web, and explore its usefulness for understanding online food preferences. The main findings of this work are: (i) recipe preferences are *partly driven* by ingredients, (ii) recipe preference distributions exhibit *more regional differences* than ingredient preference distributions, and (iii) weekday preferences are *clearly distinct* from weekend preferences.

Keywords: food; online preferences; server logs

1 Introduction

Italians are ?*Macaronis*", the English are ?*Roastbeef*", the French are ?*Frogs*" and the Germans are ?*Krauts*" [1]. In other words, food is often used to define and differentiate social groups. Claude Fischler [1] points out that human beings mark their membership of a culture or social group by asserting the specificity of what they eat or by defining differences with others. Eric B. Ross describes diet as an ?*evolutionary product of environmental conditions and of the basic forces, especially social institutions and social relations, that determine their use*" [2]. Work by Manuel Calvo [3] observed that in situations of migration, certain features of cuisine are sometimes retained even when the original culture and language have already been forgotten. This suggests that culture and diet are deeply connected. Understanding dietary patterns and food preferences[a] of humans is therefore central to several research communities. It is not only relevant from an anthropological and sociological view point, but also from a medical point of view since food preferences and diet obviously impact health.

Predominantly, studies of offline food preferences are based on surveys and consumer panels which continuously produce longitudinal behavioral data on the consumption behavior and preferences of individuals and households [4]. However, generating this data is a time-consuming and costly process and despite its strengths it also suffers from limitations such as high drop-out rates, high latency or *Hawthorne effects*.

Springer

© 2014 Wagner et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

Research objectives and methods. By contrast, in this work we leverage server log data from a large online recipe platform which is frequently used in the German speaking regions and present a multi-dimensional approach for exploring users? online food preferences. We infer the popularity of recipes and ingredients by counting the number of times each recipe or ingredient is visited from a certain geographic region within a certain time window.[b] These region- and time-specific popularity distributions are treated as the observable outcome of users? online food preferences and allow us to explore the nature and evolution of online food preferences using well established statistical methods. Amongst others, we apply (i) power law fitting methods by Clauset et al. [5] for explaining the intrinsic statistical properties of recipe and ingredient popularity distributions, (ii) correlation and similarity measures for explaining spatial food preferences and (iii) a stability measures [6] for exploring dynamics of temporal food preferences.

Concretely, we use these methods to explore the online food preferences of users on the following four dimensions:

- *Recipe preferences*. What are the intrinsic statistical properties of recipe popularity distributions? How general are those properties - i.e., do the recipe popularity distributions of different geographic regions reveal similar statistical properties? How do recipe popularity distributions differ from the popularity distributions of other types of online content (e.g., YouTube videos or websites in general)?
- *Ingredient preferences*. Do the ingredient popularity distributions of different regions reveal information about users? food preferences or are they just an artifact of users? recipe preferences and ingredient distribution over recipes? What are the intrinsic statistical properties of ingredient popularity distributions? How general are those properties - i.e., do the ingredient popularity distributions of different geographic regions reveal essentially the same statistical properties?
- *Spatial food preferences*. What is the relation between the geographic distance of regions and their online food preferences? Are online food preferences of geographically close regions more similar than those of distant regions?
- *Temporal food preferences*. To what extent do online food preferences change over time - i.e., change during the week or over seasons?

Contributions. The main findings of this work are: (i) Recipe and ingredient popularity distributions are heavy tail distributions and can be best approximated by truncated power law functions. The truncation is stronger for very popular recipes compared to popular ingredients. We can observe this behavior on a macro level (i.e., in the aggregation of all German-speaking regions in Europe) as well as on a meso level (i.e., in individual regions). (ii) Recipe preference distributions exhibit *more regional differences* than ingredient preference distributions. This suggests that food cultures manifest themselves more via the way food is combined and prepared, rather than what a culture ingests. (iii) Recipe preferences are *partly driven* by ingredients and (iv) weekday preferences are *clearly distinct* from weekend preferences.

Our work thereby shows that recipe visits as well as the inferred ingredient visits represent a preliminary, yet promising, *signal* for food preferences of human populations, since (a) our observations can in part be linked to real-world events, such as the asparagus season, and findings from offline studies and (b) our observations are fairly consistent on a macro and meso level which suggests that the observed online preference distributions can be reproduced at different scales.

Outline. We begin by describing our dataset in Section 2. In Section 3 we focus on investigating the nature of online food preferences on the four different dimensions described above. We discuss our main findings in Section 4, present a review of related work in Section 5 and finally conclude our work in Section 6.

2 Description of the dataset

We analyze server log data from the largest online recipe platform in Austria, ichkoche.at. The server logs describe how frequently a recipe has been visited within a certain region. A visit is defined as one or several page requests from the same IP address within the same session. We use visits rather than page hits or views to get a more accurate picture about users? recipe interests rather than their browsing practices. The 184,296 recipes have been visited by 1,695 different regions around 24 million times between August 2012 until November 2013. In addition to the log data, our dataset also contains information about the ingredients of recipes.

Even though the platform is from Austria, other German-speaking countries such as regions in Germany and Switzerland are prominently included (Bavaria, Zurich, Stuttgart, Hessen, Berlin, Worms, Bern and Saxony belong to the top 20 most active regions). Hence, we focus on data from the main German speaking regions in Europe - i.e., 50 federal states in Austria (AT), Germany (DE) and Switzerland (CH).

Figure 1 shows the number of visits for each of the 50 German speaking regions in our dataset on a log scale and the normalized entropy of the recipe and ingredient frequency vector per region. In order to be able to compare the visits of recipes and ingredients, we normalize the entropy by the logarithmic length of the vector since the number of recipes is much higher than the number of ingredients. A low entropy indicates that clear preferences have emerged (i.e., some recipes or ingredients are much more popular than others), while a high entropy indicates a more even distribution (i.e., many recipes or ingredients are equally popular). We can see that ingredient preferences are more focused than recipe preferences according to their lower entropy values. However, the list of regions ranked by their normalized recipe and ingredient entropy are strongly correlated (Kendall $\tau = 0.95$). This is not surprising since it simply shows that regions which have

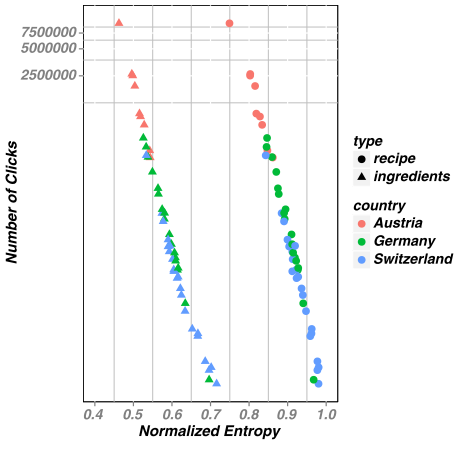

Figure 1 How active are different regions and how focused are their recipe/ingredient preferences? This figure shows the normalized entropy of the recipe and ingredient frequency vectors per region versus the number of visits per region on a log scale. Low entropy means that there is a high regional focus on just a few recipes/ingredients while a large entropy indicates a more even distribution. We see that ingredient preferences are more focused than recipe preferences according to their lower entropy values; however, they are strongly correlated (Kendall $\tau = 0.95$ for the list of regions ranked by their recipe entropy and ingredient entropy). Furthermore, we can observe that the most active regions are in Austria (due to their higher number of clicks) and that the more active a region is, the less random their recipe and ingredient preferences are (according to their lower entropy values).

a narrow recipe focus also have a narrow ingredient focus. Furthermore, we can observe almost prefect linear relationship between the activity of a region and its recipe and ingredient entropy. This indicates that the more active a region, the less random its recipe and ingredient preferences. One potential explanation is that the platform ranks popular recipes higher which will make them even more popular. Therefore, the more users from a region use the platform the more skewed the preference distribution which we observe. The most active regions are in Austria since the platform originates from this country.

3 Online food preferences

In this section we investigate online food preferences along the four introduced dimensions - i.e., recipe preferences, ingredient preferences, spatial food preferences and temporal food preferences.

3.1 Recipe preferences

Approach. To approximate the recipe preferences of one or several region(s) we count the number of times users from that region(s) have visited each recipe. We explore the intrinsic statistical properties of the popularity distributions of recipes, since the exact form of the popularity distribution often allows to infer which mechanisms might have generated the data [7, 8] and may therefore allow to gain insight into the underlying process which drives the evolution of recipe preferences. We do not only explore the properties of these distributions on a *macro level* (i.e., recipe preferences aggregated over all German-speaking regions in Europe), but also on a *meso level* (i.e., recipe preferences per region). The latter can help us to answer the question whether the shape of online recipe preferences of individual regions differs from each other and from the global, accumulated recipe preferences. That means, we explore whether different geographical regions produce popularity distributions with similar intrinsic statistical properties.

In the past, many researchers found that the *power law model* can best explain these distributions which emerge when users engage with content on the Web (cf. [6, 9–11]). Power laws are frequently appearing in social sciences, physics, biology or other sciences [12] and the probability density (mass) function of the power law distribution is defined as $f(x) = x^{-\alpha}$. Hence, we can hypothesize that our distributions at hand are also heavy tailed distributions that will most likely follow a power law model. We test this hypothesis as follows:

A simple approach to fit a power law function to data is using a least-squares linear regression. However, this method can introduce strong biases and hence, we use *maximum likelihood estimation*[c] as suggested by Clauset et al. [5] and implemented and extended by Alstott et al. [13]. Since for empirical distributions it is often difficult to find a good fit for the complete range of values, Clauset et al. [5] suggest that the power law might only hold for values that are greater than some given $x_{\min}$ value - i.e., the part of the distribution that captures popular items. Thus, we specifically focus on investigating the tail (popular recipes and ingredients) of the distribution. We use the *Kolmogorov-Smirnov statistic* as suggested by Clauset et al. [5] to find the appropriate $x_{\min}$ value that is the lowest value for which the power law model produces a good fit. Nevertheless, other candidate functions that produce heavy-tailed distributions exist. Hence, we do not only fit the power law function to our empirical data, but also other candidate functions: (a) the truncated power law function which has an exponential cut-off and is defined as $f(x) = x^{-\alpha} \exp(-\lambda x)$,

(b) the lognormal function defined as $f(x) = \frac{1}{x}\exp[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}]$ and (c) the exponential function defined as $f(x) = \exp(-\lambda x)$ which represents the lower boundary for heavy-tailed distributions. Note that for each distribution the appropriate normalization constant C is necessary such that $\sum_{x=x_{\min}}^{\infty} Cf(x) = 1$. We would like to point the interested reader to [5] and [13] for corresponding normalization definitions. In case of the lognormal distribution no discrete form for the theoretical distribution is known. Thus, we resort to the continuous counterpart for approximation by utilizing a rounding method that sums the probability mass from $x - 0.5$ to $x + 0.5$ for each data point. For comparing the candidate functions with each other - regarding their statistically significant differences - we use *likelihood ratio tests*.

One needs to note that power law fitting has some limitations (see, e.g., [6]). First of all, it is often difficult to determine which distribution has generated the data since several candidate functions might produce equally good fits. Secondly, we can only assess the goodness of fit in relative terms - i.e., we only say that a function A fits better than a function B. Thirdly, by automatically calculating $x_{\min}$ we potentially reduce the distribution to a small portion of the tail. If the tail is small enough, the power law function will always produce a good fit, but a large portion of the data will be ignored. In order to tackle this issue, we try to contrast the best fit for the whole distribution with the best fit for the tail. Finally, many different hypotheses exist that may explain why power law distributions emerge. Nevertheless, in our context some hypotheses are more plausible than others.

Macro results. Figure 2A shows a clear heavy-tailed behavior for the empirical popularity distribution of recipes since the tail of the complementary cumulative distribution function (CCDF) is heavier than one would expect by an exponential function. The figure also shows that the recipe popularity distribution does not follow a power law for the whole range of values but can best be approximated by a truncated power law function compared to other candidate functions. This is imminent as the truncated power law function is a statistically significant better fit to the empirical data compared to the pure powerlaw or lognormal function. The likelihood ratio tests between the fit of the truncated power law function and the pure powerlaw function (normalized log-likelihood ratio[d] of $R = 38.19$ with a p-value < 0.05) as well as between the truncated power law function and the lognormal function ($R = 23.60$ with a p-value < 0.05) indicate that the truncated power law function best approximates the observed distribution.

When limiting the range of values for finding the best power law fit ($\geq x_{\min}$) we see that the best $x_{\min}$ value is very high ($x_{\min} = 20{,}784$) which again indicates that power law distributions do not fit well for the whole range of values (cf. Figure 2B). Since the remaining tail of the distribution is short, we can not find statistically significant differences between the fits of the power law, the lognormal and the exponential function.

Several mechanisms such as the aging [14], information filtering [15] and content-fetching behavior [16] have been proposed to explain the sharp decay from the straight power law in the tail. In [11] the authors investigated the statistical properties of the popularity distributions of YouTube and Daum videos and argue that the so-called ?fetch-at-once" model originally introduced by Gummadi et al. [16] is most likely to explain the truncation. The model suggests that in a power-fetching scenario where users request the same content item millions of times (e.g., popular websites such as CNN) no truncation can be observed; however, if the same content is only fetched once or a limited number of times a cutoff can be observed. Cha et al. [11] conclude that it is plausible that users do

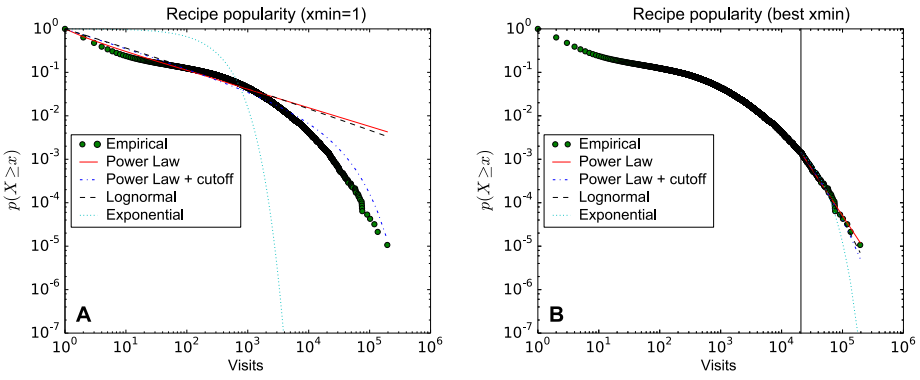

Figure 2 What are the intrinsic statistical properties of recipes? popularity distributions? The figure visualizes the CCDF of the empirical popularity distributions of recipes as well as the best fits for several candidate functions. The truncated power law function is the best fit for the whole empirical distribution (cf. panel **A**). When analyzing only the tail of the empirical distribution (cf. panel **B**) we can see that the best $x_{\min}$ is very high ($x_{\min}$ = 20,784), which indicates that power law is not a good fit for the whole range of values. Since the remaining tail of the distribution is short, we can not find statistical significant differences between the fits of the power law, lognormal and exponential function.

not watch the same video millions of times and that the limited fetching effect produces the truncation. It seems to be plausible that users of recipe platforms also fetch the same recipe a limited number of times, since the recipes do not change. If this holds true, the truncation in Figure 2 may also be explained by the *finite-size effect* [17] meaning that there is some upper limit (e.g., the number of users) that prevents the most popular recipes to be as popular as a power law distribution would suggest. A recipe can never be more popular than the number of users of the platform.

Meso results. Figure 4A shows the CCDF of the recipe popularity distribution of 50 different regions in Germany, Austria and Switzerland. We fit our candidate functions (lognormal, exponential, power law and truncated power law) to the region-specific popularity distributions. First, we use a fixed $x_{\min}$ ($x_{\min} = 1$) to cover the whole range of values for each region separately. We depict the corresponding power law fit parameters in the first row of Table 1. By comparing the power law fits of each region to the corresponding candidate functions, we can see that in almost all cases the truncated as well as the lognormal function are better fits to the data than the power law function (in 49 out of 50 cases the likelihood ratio test exhibits a p-value below the significance level of 0.05). This confirms our macro results on a regional (meso) level.

Next, we extend our analysis by finding the best $x_{\min}$ parameter for each region separately (second row of Table 1). Similar as in the macro-level analysis we again end up with extraordinary high $x_{\min}$ values. On average $x_{\min}$ is so high that the remaining tail only covers around 9% of all potential $x_{\min}$ values (i.e., bins). Again, this indicates that the power law function can only explain a very small part of the tail of the region-specific distributions.

3.2 Ingredient preferences

Approach. For analyzing ingredient preferences we infer the popularity of ingredients from the popularity of recipes - e.g., if two users visit two distinct recipes which both contain salt, than each recipe would have the popularity 1 since it received one visit, while salt would have a popularity of 2 since it received two visits. Therefore, recipe and ingredient

Table 1 Parameters of the best power law fits for the recipe and ingredient preference distributions

	α	std	$x_{\min}$	std
Recipes (all)	1.678	0.224	1.0	0.0
Recipes (best $x_{\min}$)	2.762	0.277	342.3	1428.403
Ingredients (all)	1.523	0.077	1.0	0.0
Ingredients (best $x_{\min}$)	1.883	0.712	720.94	3914.296

The fits are calculated for each region independently in two ways: (i) setting $x_{\min} = 1$ and fitting on the whole range of values and (ii) finding the best $x_{\min}$ value for the best power law fit. We report the average values over all regions and the corresponding standard deviation.

popularity distributions are obviously interdependent and it is unclear if the inferred popularity distribution of ingredients reveals information about the online food preferences of users or if it is just an artifact of the ingredient universality distribution (i.e., in how many recipes ingredients are used) and the recipe popularity distribution (i.e., how often each recipe was visited). To address this question we (i) analyze the universality distribution of ingredients and (ii) simulate the ingredient preferences of a synthetic region as follows: A synthetic region consists of a set of agents who randomly select recipes from a randomly generated recipe-popularity distribution with the same shape as our empirical recipe distribution. For each selected recipe we extract all its ingredients from our data and increase the visit count of those ingredients. Repeating this process allows generating a synthetic ingredient preference distribution which reflects how the visits would be distributed over ingredients if the recipe selection process would be random. We assume that visits are independent and use the median number of visits of all regions as the activity level of the synthetic region.

We contrast the ingredient preference distribution which is generated by the synthetic region with the empirically observed ingredient distributions. If the shapes of the two distributions do not differ significantly, we can conclude that ingredient popularity preferences are an artifact of recipe popularity and the distribution of ingredients over recipes. Otherwise, we may conclude that external forces such as users? ingredient preferences or seasonality of ingredients impact the recipe selection process. In other words, if users recipe selection process is partly driven by ingredients, we expect the empirical ingredient distributions to differ from the synthetically generated one in the sense that they should be more focused towards fewer ingredients than the synthetic one.

To investigate the shape of the popularity distribution of ingredients on a meso and macro level we adapt the same approach as for recipes which we described in the previous section.

Macro results. In the previous section we have shown that the recipe popularity distribution is best approximated by a severely truncated power law (cf. Figure 2). However, it is unclear how the ingredients are distributed over recipes (i.e., in how many recipes each ingredient is used). That means, how many ingredients are so universal that they are used in almost all recipes and how does this universality decrease? Figure 3D shows that the universality distribution of ingredients follows a truncated power law (at least for the most universal ingredients with $x_{\min} \geq 15$). This is also imminent by the results of the likelihood ratio test which indicates that the fit of the truncated power law function is a statistically significant better fit than the pure powerlaw, the lognormal and the exponential function; all p-values are below the significance level of 0.05. This observation is

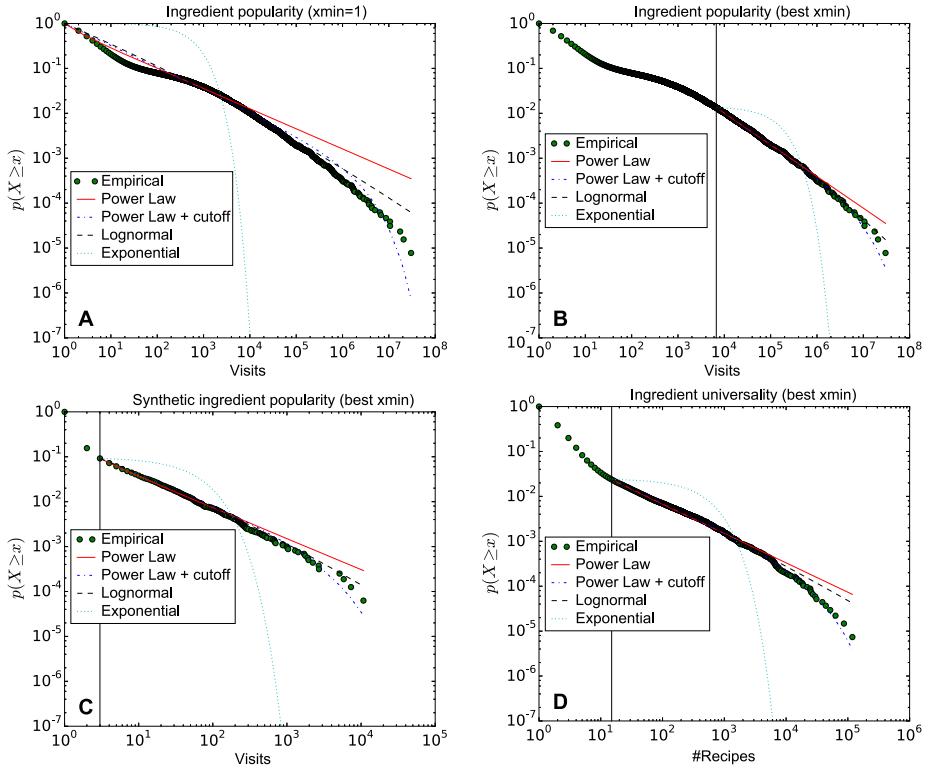

Figure 3 What are the intrinsic statistical properties of ingredients? popularity distributions? We visualize the empirical CCDFs coupled with corresponding candidate function fits. The popularity of ingredients clearly follows a truncated power law when looking at the whole range of values (cf. panel **A**). When only focusing on the most popular ingredients ($> x_{\min}$) the truncated power law function as well as the log-normal function fit best as visible in panel **B**. For the synthetic ingredient popularity distribution which we obtained by simulating a region with a random recipe selection behavior also the truncated power law and the lognormal function are the best fits (cf. panel **C**). The universality distribution of ingredients follows as well a truncated power law (cf. panel **D**).

in line with a study looking at the distribution of ingredients over recipes in cook books [18] as the truncation captures finite size effects (i.e., number of recipes is finite). We also find that the mean number of ingredients per recipe is 11 and the median is 10, which is similar to what was found in previous studies of cookbooks (mean number of ingredients per recipe was found to be between 8 and 10 [19] or between 7 and 11 [18] for different cookbooks and cuisines).

For the empirical ingredient popularity distribution we can clearly see that it follows a truncated power law for the whole range of values (cf. Figure 3A, the likelihood ratio tests between the distinct distributions all result in p-values below 0.05 indicating that the truncated power law fit is the best one out of those we tested). We also find a truncation in the ingredient popularity distribution, but the truncation is less sharp for ingredients than for recipes (cf. Figure 2). This can be derived from looking at the exponential cutoff of both truncated fits: $\lambda = 1.74e{-}05$ $(1/\lambda = 57.303)^{e}$ for recipes and $\lambda = 1.14e{-}07$ $(1/\lambda = 8{,}752.642)$ for ingredients. One potential explanation for this might be that the popularity of recipes is limited by the number of users of the platform, especially if they only fetch each recipe a limited number of time. The popularity of an ingredient depends on the number of recipes in which it is used (i.e., its universality) and the popularity of all recipes. Therefore, the

finite-size effect is less pronounced in the ingredient popularity distributions than in the recipe popularity distributions. When only focusing on the most popular ingredients with $x_{\min}$ = 6,711 (see Figure 3B), we again find that the truncated power law is a statistically significant better fit to the data than the power law function (R = 2.55 with p-value 0.002). The likelihood ratio test between the truncated power law function and the lognormal function indicates similar good fits (p-value above 0.05).

For the synthetically generated ingredient popularity distribution we find that the truncated power law function is a better fit than the power law function (p-value below 0.05). The lognormal and truncated power law function are similar good fits (p-value of 0.036). This indicates, that on the first glance our synthetic ingredient popularity distribution which is generated by a random recipe selection process does not differ from our empirical observations. However, this is only true for the most popular ingredients (such as salt, sugar, butter or oil). When taking a closer look one can see two interesting differences: the two distributions differ in (1) how unpopular ingredients are accessed by users and (2) the growth rate of popularity. Concretely, we can see in Figure 3B and C that the best $x_{\min}$ value is much smaller for the synthetic popularity distribution than for the empirical ones. This indicates, in our empirical data the shape of the distribution which also contains unpopular ingredients is different from the part which only contains more popular ingredients. We can further see that the distributions as well as corresponding fits are slightly steeper for the empirically observed ingredient preferences compared to the synthetically generated ones - the α parameter of the power law function is 1.70 for Figure 3B while it is 1.68 for the synthetic data. From these two observations we can derive that *the recipe selection process of users seems to be at least partly driven by the ingredient preferences of users*, since the ingredient preferences which are generated via the recipe selection process are more focused towards few ingredients and less focused towards others than one would expect if the process would be random. This means that users? ingredient preferences reveal stronger favor or disgust for selected ingredients than we would expect to observe if the recipe selection process would be random.

Meso results. Finally, to gain insight into the potential universality of the pattern which we observed in our macro analysis, we repeated the analysis for each region separately. We first fitted the power law and several other candidate function to the whole range of values (i.e., $x_{\min}$ = 1) of region-specific ingredient preference distributions (cf. third row in Table 1). By comparing candidate functions, we can see that for all regions the truncated as well as the lognormal function fit the data better than the power law function (p-values below 0.05). Next, we estimated the best $x_{\min}$ value for each region and fitted different candidate functions to the part of the distribution which exceeds $x_{\min}$ (cf. fourth row in Table 1). Our results show that in 46 cases the truncated power law function fits statistically significantly better our data than the power law function (positive likelihood ratio, p-values below 0.05) while in 3 regions it is worse (negative likelihood ratio, p-values below 0.05). In one case the fits are equal (p-values above 0.05). The lognormal function equals the power law function in nearly all cases (46) (p-values above 0.05). This indicates that it is indeed likely that *the ingredient distributions of different regions have been generated by the same underlying process*, since we observe the same patterns on the macro and meso level. Finally, we also observe that *the popularity distributions of different ingredients tend to be very similar, while for recipes we observe slightly more regional variability*. This observation becomes not only apparent when comparing the different regional CCDF plots

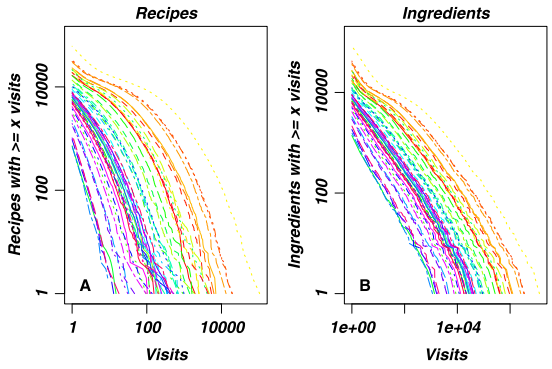

Figure 4 Are the recipe and ingredient preferences of different regions generated by the same process? The figure shows the CCDF of the ingredient and recipe popularity distributions of individual regions. The variable x on the x-axis corresponds to the number of visits, while the y-axis shows the number of recipes/ingredients that have received equal or more than x visits. One can see that the ingredient popularity distributions of different regions tend to be very similar, while for recipes we can observe slightly more regional variability. This observation is also apparent when comparing the standard deviations of the fitted α slopes (cf. Table 1) - higher for recipes (all) compared to ingredients (all).

in Figure 4, but also when comparing the standard deviations of the fitted α parameters (cf. Table 1) which is higher for recipes (all) compared to ingredients (all).

3.3 Spatial food preferences

One potential cause of shared online food preferences is geographical proximity since frequent communication and migration may explain the adoption of food preferences [20]. In the following, we test the hypothesis that geographically nearby regions are more similar regarding their online food preferences than geographically distant regions.

Approach. We compute the recipe and ingredient similarity between different regions using cosine similarity. Cosine similarity is a measure of similarity between two vectors that measures the cosine of the angle between them. Two vectors (in our case recipe or ingredient frequency vectors) with the same orientation have a cosine similarity of 1, while vectors with opposite orientation have a cosine similarity of −1.

We test the hypothesis that geographical distant regions reveal more distinct online food preferences than geographic close regions by measuring Spearman rank correlation between the geographical distance of region pairs and their recipe similarity and ingredient similarity. We use 10k bootstrap samples to estimate the confidence interval of the correlation coefficient.

We further compare the difference in the means (recipe-similarity and ingredient-similarity means) of geographical distant regions (i.e., regions which are more distant than the median distance) and geographical nearby regions (i.e., regions which are closer than the median distance) using a permutation test. For Austria, the median distance of all region pairs is 167 km, for Germany it is 283 km and for Switzerland it is 84 km. The overall median distance is 385 km. We created two groups of region pairs, distant ones D and close ones C whose sample means are $\bar{x}_D$ and $\bar{x}_C$. Let n_D and n_C be the sample size corresponding to each group. The permutation test is designed to determine whether the observed difference $T(obs)$ between the sample means is large enough to reject the null hypothesis H_0 which states that the two groups have identical probability distribution.

First, the difference in means between the two samples is calculated as $T(obs) = \bar{x}_D - \bar{x}_C$. Then, the observations of groups D and C are pooled and subsequently, the difference in sample means is calculated and recorded for every possible way of dividing these pooled values into two groups of size n_D and n_C (i.e., for every permutation of the group labels D and C). The set of the calculated differences is the exact distribution of possible differences under the null hypothesis that group labels do not matter.

Results. When analyzing all three countries together, we find a slightly negative correlation between geographic distance and the cosine similarity of the recipe and ingredient frequency vectors of different regions (Spearman?s $\rho = -0.19$, standard error $SE_\rho = 0.03$ and confidence interval CI at 95% is (–0.2535, –0.1259) for recipes and Spearman?s $\rho = -0.22$, $SE_\rho = 0.03$ and CI at 95% is (–0.2804, –0.1551) for ingredients). When only looking at Austria, we observe a much stronger correlation but with higher standard error especially for ingredients (Spearman?s$\rho = -0.74$, standard error $SE_\rho = 0.16$ and confidence interval CI at 95% is (–0.9242, –0.2340) for recipes and Spearman?s$\rho = -0.21$, $SE_\rho = 0.31$ and CI at 95% is (–0.7505, 0.4891) for ingredients).

Figures 5 and 6 show that at least in Austria and in Germany the recipe preferences of geographic close regions tend to be more similar than those of geographic distant ones. Ingredient preferences are very similar for both geographic close and distant regions (cf. Figure 6). For Switzerland we cannot observe the same pattern since geographic distances in Switzerland are very small and the diversity in the country is very large (see Figure 6). When looking at all three countries, we can still see the tendency of geographic close regions to be more similar than geographic distant ones. However, the differences are not significant and our permutation test results suggest that we cannot reject the null hypothesis - i.e., the differences within all groups can potentially be generated from the same underlying distribution. In previous work [21] the authors tested the same hypothesis for China and were able to reject the null hypothesis. However, one needs to note that our study focuses on a much smaller geographic area and therefore geographic distances may play a minor role.

3.4 Temporal food preferences

Food preferences are not static and may change over time. Therefore, we next explore the temporal evolution of human?s online food preferences and how potential temporal regularities relate to what we know about food and dietary patterns observed in offline studies.

Approach. First, we explore the normalized number of visits per day or month for selected ingredients using z-scores. The rationale behind z-score normalization is to mitigate the effect of anomalous days [22].

Next, we compute the average popularity change rate of two consecutive days or months d_i and d_{i+1} for selected ingredients as follows:

$$R(d_i, d_{i+1}) = \frac{|F(d_i) - F(d_{i+1})|}{\sum_{j=1}^{N} |F(d_j) - F(d_{j+1})|}. \quad (1)$$

N refers to the number of consecutive pairs (which is e.g., 7 in the case of a week) and $F(d)$ refers to the total access volume at day or month d.

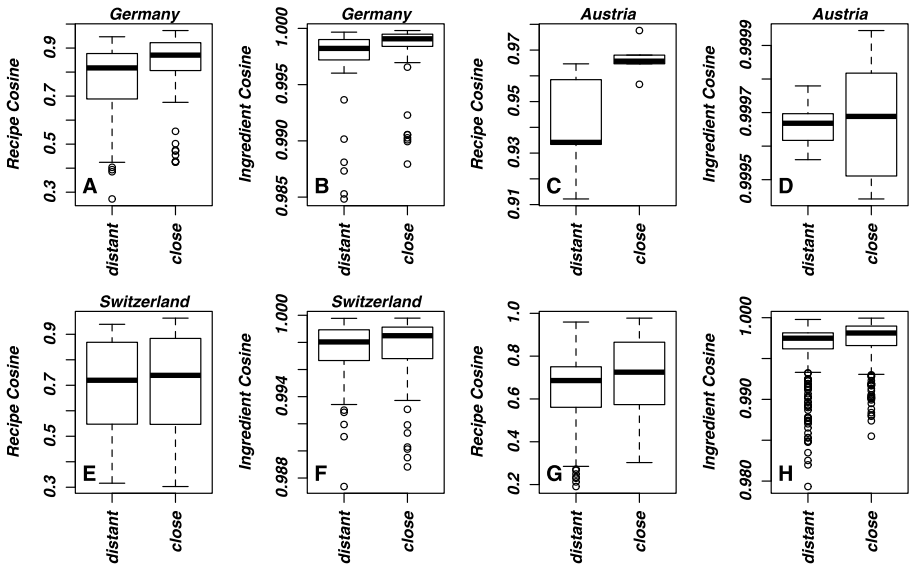

Figure 5 Do geographic nearby regions reveal more similar online food preferences than distant ones? This figure shows the cosine similarity distribution of two groups of region-pairs within each country and across countries, distant ones (which are above the median distance) and nearby ones (which are equal or below median distance). One can see that in Austria and Germany geographic close regions tend to reveal more similar recipe and ingredient preferences than distant regions; however, in Switzerland we do not observe this tendency.

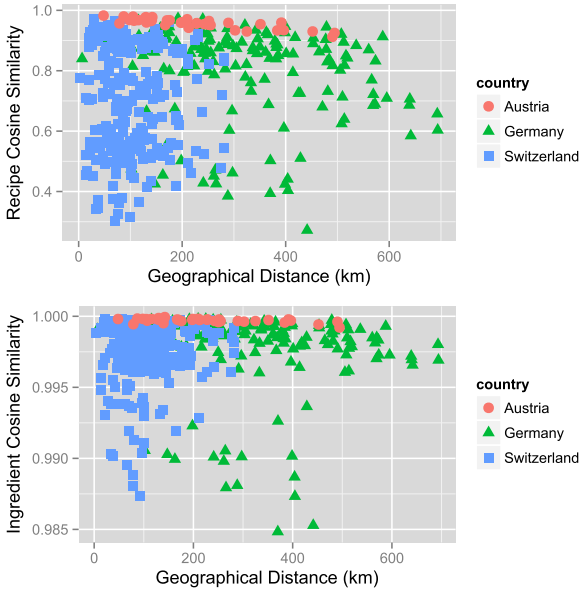

Figure 6 Do geographic nearby regions reveal more similar online food preferences than distant ones? This figure shows the cosine similarity of region pairs in relation to their physical distance. One can see that ingredient preferences tend to be very similar independent of the geographical distance in all three countries. Recipe similarity slightly decreases with increasing distance, at least in Austria and Germany.

To go beyond the exploration of the popularity of selected ingredients, we next explore the dynamics (i.e., stability and changes) of the recipe and ingredient frequency vectors between consecutive weekdays and months using the rank biased overlap (RBO) metric

[23]. RBO measures the correlation between two ranked lists of recipes/ingredients that represent the popularity of recipes/ingredients in different regions. Recipes and ingredients are ranked by the number of visits they obtained during that weekday/month within one year. RBO is a top weighted metric which means that it is more important that the ranking of the most popular recipes/ingredients does not change from one day/month to the next day/month than the ranking of recipes/ingredients in the long tail of unpopular recipes/ingredients. This makes sense, since we know that the popularity distributions of recipes and ingredients are heavy tail distribution and one might argue that recipes and ingredients which have been accessed very few times during a day or month do not reflect the online food preferences of that day or month. Therefore, we do not care if the ranking of those recipes/ingredients changes. RBO is defined as follows:

$$\mathrm{RBO}(\sigma 1, \sigma 2, p) = (1-p) \sum_{d=1}^{\infty} \frac{\sigma 1_{1:d} \cap \sigma 2_{1:d}}{d} p^{(d-1)}. \tag{2}$$

Let $\sigma 1$ and $\sigma 2$ be two not necessarily conjoint lists of ranking. Let $\sigma 1_{1:d}$ and $\sigma 2_{1:d}$ be the ranked lists at depth d. The RBO falls in the range $[0,1]$, where 0 means disjoint, and 1 means identical. The parameter p ($0 \leq p < 1$) determines how steep the decline in weights is. The smaller p, the more top-weighted the metric.

Results by week day. Figure 7 shows the normalized access volume of sample ingredients ordered by group (meat, carbohydrates, fish, vegetables and alcohol). We can see that different groups of ingredients indeed reveal similar temporal trends regarding their popularity. Meat (e.g., pork and steak) is mainly requested during the weekend with a peak on Sunday. This confirms offline observations (gained via questionnaires) which reveal that Austrians consume meat products more frequently on Sundays compared to other days in the week [24]. Carbohydrate-rich, cheap and healthy food such as pasta, vegetables and potatoes is more frequently requested at the beginning of the week and less frequently during weekends. We also observe that fish is most popular on Thursday and Friday and alcohol is more popular at weekends than during the week.

Our preliminary results raise the question whether online preferences of certain ingredients show a clear shift from weekdays to weekends. Figure 8 shows that indeed most changes happen before and after the weekend, suggesting that *online preferences for ingredients during the week are starkly different from weekend preferences*. Changes in online food preferences over the course of a week slowly accumulate, with noticeable changes starting around Thu/Fri. The end of the weekend period is clearly demarcated, evident in a high change rate across most ingredients on Sun-Mon. This means that online food preferences tend to change slowly towards weekend preferences during the week, but they change abruptly back to weekday preferences over Sun-Mon.

Finally, to complement our analysis of selected ingredients and ingredients groups, we study the dynamics of all recipes and ingredients collectively. Figure 9 shows the stability of ingredient and recipe preferences during the course of a week using the RBO metric with different top weightiness (i.e., p-values). One can see that users? ingredient preferences are very stable which can be explained by the fact that many of the most popular ingredients (such as butter, salt and pepper) are equally important on different weekdays. The shifts in

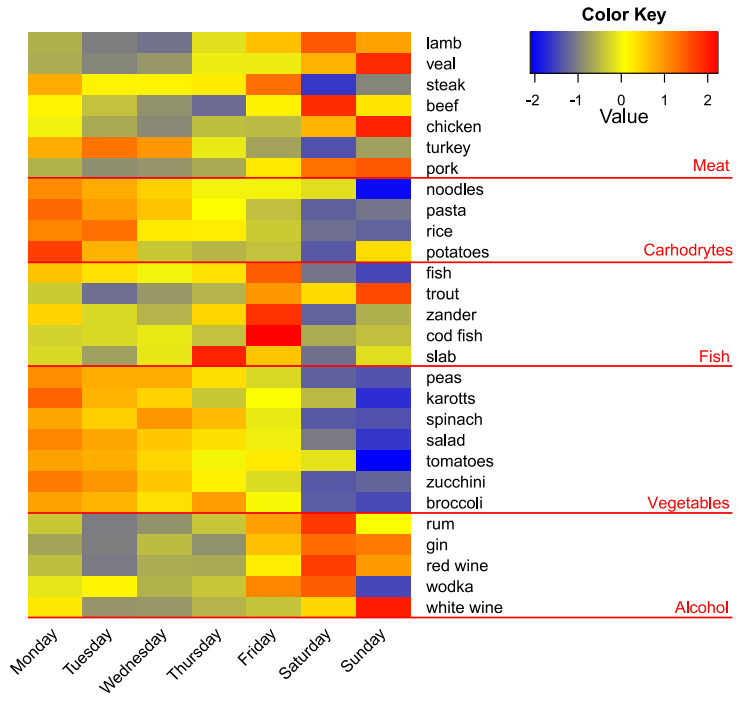

Figure 7 How popular are different ingredients over the course of a week? The figure shows *z*-scores of the request volume per weekday. One can see that fish is mainly requested Friday, meat during the weekend and carbohydrate-rich and cheap food such as pasta or potatoes during the week.

the ingredient preferences only become visible when one focuses on selected ingredients but not when analyzing the aggregation of all ingredients. For users? recipe preferences we can see that they are relatively stable during the week, but major shifts in the preferences happen during the weekends. This confirms our hypothesis that *users? online food preferences change during the weekend* which becomes visible in their recipe selection process which is in part driven by ingredient preferences as we have shown before. However, not all ingredients have the same function in the recipe selection process and it is unlikely that the most popular ingredients impact the recipe selection process since those are mainly staple food. We leave the question about which types of ingredients may drive the recipe selection process for future research.

Results by month. To characterize a typical year, we compute for a sample of ingredients their normalized access volume and the change rate between consecutive months. Figure 10 shows that the access volume of ingredients indeed allows to identify ingredients with strong seasonal prevalence such as asparagus, since recipes with asparagus are mainly requested during the asparagus season which starts at the end of April and ends in June.

Again, so far we have only explored selected ingredients. To further extend and complement our analysis we investigate the dynamics (i.e., stability and changes) of the recipe and ingredient frequency vectors between consecutive months using the rank biased overlap (RBO) metric. Figure 11 shows the RBO value between the ranked lists of recipes or ingredients of consecutive months. Recipe and ingredients are ranked by the number of

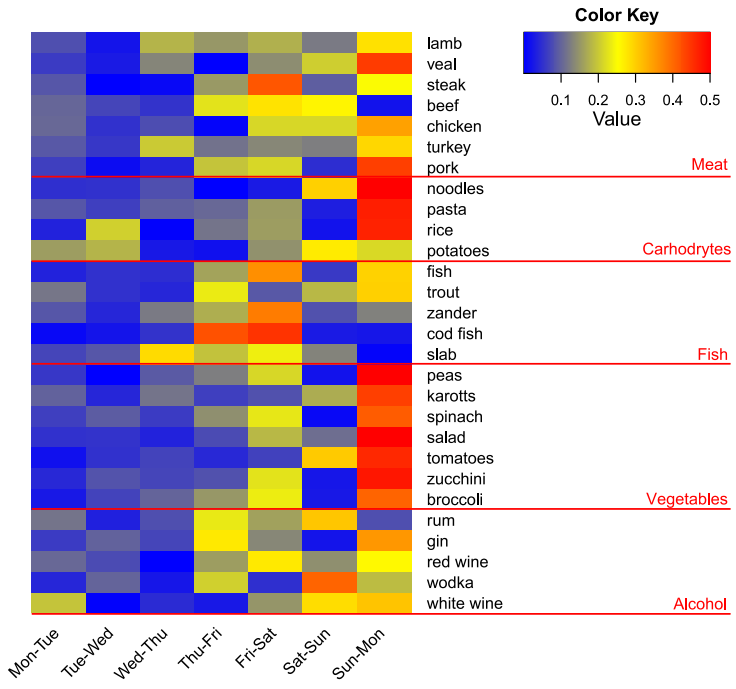

Figure 8 How do preferences change over the course of a week? The figure shows the average change rate from one day to the next one. One can see that most changes happen before and after the weekend. That means that the weekday-preferences are different from weekend-preferences.

visits they obtained during that month within one year. One can see that the recipe preferences are pretty unstable and change a lot during the course of a year. Only November and December and January, February and March seem to be exceptions since the recipe preferences remain pretty stable during these periods. However, to further dig into these patterns data collections which span over several years are required.

4 Discussion

In the following we will discuss the main findings of our work and their implications. This work is based on data stemming from one single recipe platform with unknown biases. In future work we plan to extend this study on additional log data from other recipe platforms.

Recipe and ingredient preferences. We observe that both the popularity distributions of recipes and ingredients are heavy tailed. Both can be best approximated by a truncated power law distribution, while the truncation is stronger for the recipe popularity distribution (cf. Figure 2) compared to the ingredient popularity distribution (cf. Figure 3). Our results also indicate that the recipe selection process of users is at least partly driven by their preferences towards ingredients since the empirically observed ingredient preferences are more focused towards few ingredients and less focused towards others than one would expect if the recipe selection process would be random. By investigating the online food preferences of each region separately (meso level), we obtained similar results as for the accumulated analysis (macro level). However, we see slightly more regional variability for the recipe preference distributions than for the ingredient preference distributions.

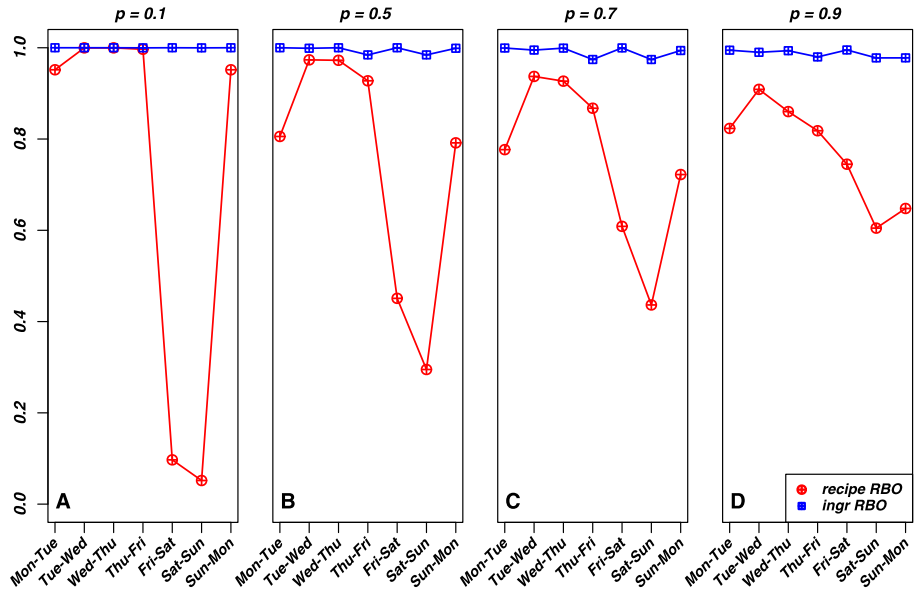

Figure 9 How stable are users? online food preferences during the course of a week? This figure shows the RBO values between the ranked lists of recipes or ingredients of two consecutive days. Recipe and ingredients are ranked by the number of visits they obtained during that weekday within one year. One can see that users online food preferences are relatively stable during the week, but major shifts in the preferences happen during weekends.

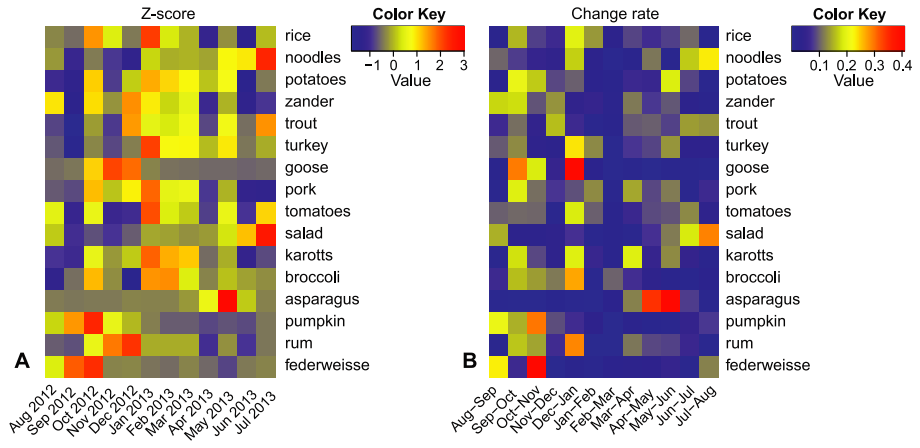

Figure 10 How seasonal are the preferences for different ingredients? This figure shows the *z*-score per month and the average change rate from one month to the next one. One can see that the popularity of certain ingredients such as asparagus can clearly be mapped to its seasonal availability in central Europe.

This indicates, that the process which generated the ingredient preferences in different regions is more similar across different regions than the process which generated the recipe preferences.

In the literature several mechanisms have been proposed and described that may produce heavy tailed and specifically power law distributions (e.g., [8, 25, 26]). The most prominent one is the so-called *Yule process* [27] (also known as preferential attachment or the rich get richer phenomenon) which can only explain certain parts of online food

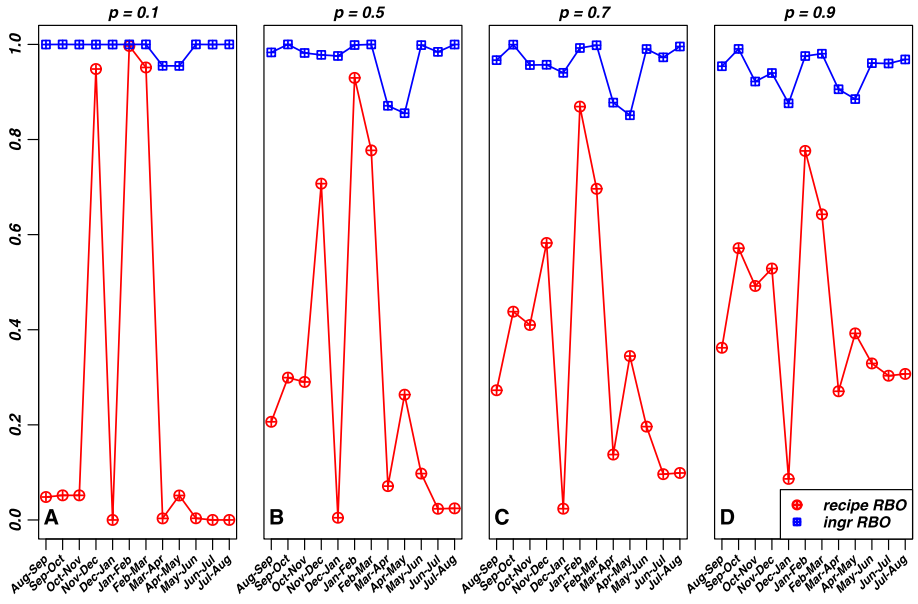

Figure 11 How stable are users? online food preferences during the course of a year? This figure shows the RBO values between the ranked lists of recipes or ingredients of consecutive months. Recipe and ingredients are ranked by the number of visits they obtained during that month within one year. One can see that the recipe preferences are pretty unstable and change a lot during the course of a year. Only November and December and January, February and March seem to be exceptions since the recipe preferences remain pretty stable during these periods.

preference distributions. As mentioned before, our results indicate that the most popular recipe and ingredient popularities are truncated (i.e., they are lower than one would expect if the power law distribution would explain the whole distribution). Several hypotheses exists that aim to explain what causes this truncation (see e.g., [11, 15, 17, 28]). In this work we presented two potential explanations of why the Yule process fails to explain the evolution of online food preferences and why we can observe a truncation for the popularity of the most popular recipes: (i) We can expect that the static nature of recipes has an influence on how they are viewed; the fact that recipes do not change and are therefore most likely only fetched a limited number of times by each user leads to the *fetch-at-once effect* which can causes the truncation of the most popular recipes. (ii) If this holds true the truncation is further impacted by the *finite-size effect* - i.e., that a physical upper limit of the maximum popularity of recipes exists which corresponds to the number of users on the platform. For ingredients, the truncation is also visible but less pronounced since universal ingredients like salt or butter are included in almost all recipes. Therefore, even if a finite number of users fetches every recipe only once or a very small number of times, universal ingredients are fetched repeatedly.

Spatial preferences. Anderson et al. [20] point out that ?*our basic nutritional needs, and some very broad preferences, are set by biology, but preferences are notoriously subject to cultural and social forces*". Therefore, it seems to be a plausible assumption, that geographic close regions have more similar food preferences than distant ones since they are more likely to be subjects of the same cultural forces. Zhu et al. [21] further point out that frequent communication and migration may explain the adoption of food preferences

and their empirical findings support this hypothesis. Our spatial preference analysis (cf. Figures 5 and 6) shows that there exists a slight tendency of geographic close regions to reveal more similar recipe and ingredient patterns; however, the differences in the German speaking part of Europe are not significant as it was observed for China [21]. One potential explanation for that is that distances in German speaking part of Europe are much smaller and also the mobility of people living in this area might be higher. Therefore, other factors like cultural similarities and transportation infrastructure might be more suitable alternatives to explain the similarities of food preferences between different regions in central Europa.

Temporal preferences. Food preferences are not static and change over time and our results clearly show that users? online food preferences change during the weekend, which becomes visible in their recipe selection process which is in part driven by ingredient preferences. Selected ingredients show a prevalence for specific seasons (cf. Figure 10) and weekdays (cf. Figure 7) which can be related with phenomena from the offline world such as the seasonal availability of some ingredients or results from reactive diet studies which showed that users tend to eat more meat during the weekend than other days in the week [4]. However, when looking at the aggregation of all ingredients we observe that the ingredient preferences are relatively stable during the course of the week, while the recipe preferences clearly change during the weekend (cf. Figure 9). This can be explained by the fact that not all ingredients have the same function in the recipe selection process and especially the most popular ingredients like salt, sugar or butter are so common that they probably do not impact the recipe selection process. Therefore, when analyzing the collection of all ingredients, their popularity appears to be very stable (cf. Figure 9), while for selected ingredients we can observe interesting changes and temporal regularities (cf. Figures 8 and 10) which can be related with the offline world (cf. Figures 7 and 10). We leave the question about which types of ingredients may drive the recipe selection process for future research.

5 Related work

Dietary trends and culinary evolution. Previous research suggests that dietary trends are affected by behavioral, socio-cultural and economic variables [29, 30], while the impact of taste factors is for adult?s food intake less apparent [31]. A question which has been of long lasting interest is: *which variables can explain the culinary variety to what extent?* Researchers, for example, found culinary regularities that are functions of the climate. The work of [32] shows that the usage of spices in a given region is highly correlated with its annual temperature. The work of West et al. [22] also suggests that climate impacts dietary patterns. Zhu et al. [21] investigate the effect of climate and geographic distance on the cuisine of different regions in China and show that geographic distance plays a more important role than climate conditions. Concretely, they showed that climate does not show any correlation with the ingredient usage similarities when controlling for geographic distance (PCC = 0.116), while geographic distance remains correlated with ingredient usage similarities also when they control for climate (PCC = −0.280).

In [18] the authors analyze the ingredient distributions of six different cookbooks. They found that the universality of ingredients varies over four orders of magnitude documenting huge differences in how frequently various ingredients are used in recipes. The authors

further find that the rank-ordered ingredient distribution (i.e., ingredients are ranked by the number of recipes in which they appear) follows a power law with an exponential cut-off to capture finite size effects. The slope of the power law parameter $\alpha = 1.72$ which cannot be explained by a general Yule process which would produce a power law with $\alpha \geq 2$. To model cuisine growth the authors propose a copy-mutate algorithm which preservers idiosyncratic ingredients in a manner akin to the founder effect in biology.

In [19] the authors empirically tested the *food pairing hypothesis* originally proposed by Benzi and Blumenthal which states that ?two ingredients which share important flavor compounds will go well together". They found that shared flavors compounds effect ingredient combinations very differently in Western and Eastern cuisines. While in Western cuisines ingredients that share flavor compounds are more frequently combined then one would expect from randomly generated recipes (with the same ingredient frequencies), in the Eastern cuisines ingredients with distinct flavor compounds are combined much more frequently than for random recipes. The authors further reveal that the food preparing differences between the Western and Eastern cuisine is due to few outliers which are frequently used in a particular cuisine, such as butter, cocoa or vanilla in North America.

While our work focuses on studying the popularity of recipes in different regions, previous work on the culinary evolution relied on a regional categorization schema of recipes to define what is typical for a region. However, it remains unclear if typical recipes of a region reflect the preferences of that region.

Online food preferences. Despite the fact that online recipe databases and community sites gain a lot of attention in the online world, little research exists today on the nature and evolution of users? online food preferences and how those preferences relate to their offline preferences (i.e., their preferences in the real world). A very inspiring piece of work was published by West et al. [22] who analyze temporal patterns and regional differences in dietary patterns online and relate them with observations from the world. Unlike in our work, West et al. use web logs recorded by a Web browser add-on provided by Bing and use the access statistics of recipes a user clicks on from a search query result page to approximate the food consumption of different regions in the US. Their user study suggests that it is a reasonable assumption that users who search for a dish and click on the recipe afterwards, are likely to cook it. Their results indicate that the online access volume of food related information may potentially allow to predict offline medical needs. They found a significant correlation between the hospital admissions of patients admitted with a diagnosis related to congestive heart failure over time and the sodium intake over time approximated via recipe visits. The curves indeed follow each other closely, however the causal relationship cannot be proved. High sodium intake may e.g. be linked to holidays which might be linked to higher traveling activities leading to a loss of compliance with medication.

Another interesting work by Teng et al. [33] shows that structural properties of nodes in ingredient networks (co-occurrence and substitution networks) can be used to improve the prediction performance of recipe ratings. The authors make the assumption that the ratings of recipes reflect the online food preferences of users, which can be inaccurate especially if only a small fraction of users uses the rating feature while most of them are lurkers.

Our work overcomes this issue by focusing on the consumption of content rather than the production of content. Unlike West et al. we focus on a rather small area in central Europe (Austria, Germany and Switzerland), while they focus on the US. Further, we use the access volume of recipes as a proxy for food preferences rather than for food consumption.

Popularity of online content. Since we analyze the popularity of recipes and ingredients over time and space, also research about the popularity of other types of online content is relevant for our work. For example, in [11] The authors found that the popularity distribution of videos on YouTube and Daum follows a power law but with a sharp decay from the straight line for the most popular videos. The curve fitting results show that the decay at the heavy tail is best fitted by adding an exponential cutoff to the power law distribution. The power law part of the distribution can be explained by the Yule process (also known as preferential attachment or the rich-get-richer phenomenon), while the sharp decay for the most popular videos can be explained by *the aging effect* [14] (i.e., high degree nodes will eventually stop receiving more links because every node ages and will stop being active at some point), *the information filtering effect* [15] (i.e., users cannot receive information about all available videos but only about a fraction of them and therefore preferential attachment is hindered), and finally *the limited fetching phenomenon* [16] (i.e., users may fetch popular videos only once or few times since they do not change, while they may fetch popular websites such as news sites million of times). In [11] the authors show via simulations that the limited-fetching phenomenon can indeed explain the sharp decay from the straight power law line for very popular videos. The higher the number of requests per users in their simulations the more visible the decay.

Server logs. In our work we use server logs as a proxy for users? online food preferences. In previous work, logs of search engine use have been successfully used to identify temporal trends (cf. [34]), geographic differences (cf. [35]) and to predict real world medical phenomena (cf. [36]). However, to our best knowledge this is the first work which analyzes server log data from recipe platforms to analyze the evolution of online food preferences.

6 Conclusions

To the best of our knowledge, our work is the first to study online food preferences of users via log data obtained from recipe websites and presents a comprehensive multidimensional approach which allows to dig into the nature and evolution of users? online food preferences. We find that recipe visits (as well as the inferred ingredient visits) may represent a plausible *signal* for food preferences of human populations, since (i) our observations can in part be linked to real-world events, such as the asparagus season, and findings from studies which e.g., showed that people eat more meat at weekends than at other days of the week and (ii) our observations are fairly consistent on a macro and meso level which suggests that the observed online preference distributions can be reproduced at different scales. We hope that this work contributes to understanding the nature and evolution of online food preferences by analyzing the observable outcome of such preferences on four different dimensions.

The main findings of this work are: (i) Recipe and ingredient popularity distributions are heavy tailed and can be approximated well by a severely truncated power law function (recipes) and a truncated power law function (ingredients). These effects can both be

found on a meso level (i.e., in individual regions) as well as on a macro level (i.e., in the aggregation of all German-speaking regions in Europe). (ii) Recipe preference distributions exhibit *more regional differences* than ingredient preference distributions. (iii) Recipe preferences are *partly driven* by ingredient preferences and (iv) weekday preferences are *clearly distinct* from weekend preferences.

Competing interests
The authors declare that they have no competing interests.

Authors? contributions
All authors designed the methodology and conceived the experiments. CW collected the data and performed the spatial and temporal analysis. PS analyzed the statistical properties of recipe and ingredient popularity distributions. All authors wrote and revised the manuscript.

Acknowledgements
We thank ichkoche.at for sharing their data with us and the anonymous reviewers for their valuable comments.

Endnotes
[a] Food preferences may not only expose what is liked but also what is disliked and avoided. Preferences assume a situation of choice but do not necessarily reflect use. One might prefer lobster over shrimps but eat more shrimps.
[b] Ingredient visits are inferred from the recipe visits in which the ingredients are used.
[c] Note that we work with discrete and not continuous data and hence, also use the exact methods necessary to cope with discrete data. For fitting the discrete power law function we use the faster analytical methods, instead of using the slow exact numerical variants.
[d] A positive value of R means that the log-likelihood of the first distribution (in this case the truncated power law function) is higher than that of the second (in this case the power law function).
[e] We also report $1/\lambda$ as it roughly tells us where the cutoff is.

Received: 20 March 2014 Accepted: 10 December 2014 Published online: 30 December 2014

References
1. Fischler C (1988) Food, self and identity. Soc Sci Inf 27(2):275-292
2. Harris M, Ross EB (1987) Food and evolution: toward a theory of human food habits. Temple University Press, Philadelphia
3. Calvo M (1982) Migration et alimentation. Soc Sci Inf 21(3):383-446
4. Prester H-G (2001) Consumer panel research at GfK. In: Social and economic analyses of consumer panel data. ZUMA-Nachrichten Spezial, vol 7
5. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661-703
6. Wagner C, Singer P, Strohmaier M, Huberman BA (2014) Semantic stability in social tagging streams. In: Proceedings of the 23rd international conference on World Wide Web, pp 735-746
7. Mitzenmacher M (2003) A brief history of generative models for power law and lognormal distributions. Internet Math 1:226-251
8. Andriani P, McKelvey B (2009) Perspective - from Gaussian to Paretian thinking: causes and implications of power laws in organizations. Organ Sci 20(6):1053-1071
9. Adamic LA, Huberman BA (2000) Power-law distribution of the World Wide Web. Science 287(5461):2115
10. Sen S, Lam SK, Rashid AM, Cosley D, Frankowski D, Osterhouse J, Harper FM, Riedl J (2006) Tagging, communities, vocabulary, evolution. In: Proceedings of the 2006 20th anniversary conference on computer supported cooperative work, pp 181-190
11. Cha M, Kwak H, Rodriguez P, Ahn Y-Y, Moon S (2009) Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Trans Netw 17(5):1357-1370.
12. Newman ME (2005) Power laws, Pareto distributions and Zipf?s law. Contemp Phys 46(5):323-351
13. Alstott J, Bullmore E, Plenz D (2014) powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1):e85777
14. Amaral LA, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci USA 97(21):11149-11152
15. Mossa S, Barthelemy M, Stanley EH, Amaral LA (2002) Truncation of power law behavior in ?scale-free" network models due to information filtering. Phys Rev Lett 88(13):138701
16. Gummadi KP, Dunn RJ, Saroiu S, Gribble SD, Levy HM, Zahorjan J (2003) Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In: Proceedings of the nineteenth ACM symposium on operating systems principles. SOSP?03, pp 314-329
17. Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. Phys Rev A 38(1):364
18. Kinouchi O, Diez-Garcia RW, Holanda AJ, Zambianchi P, Roque AC (2008) The non-equilibrium nature of culinary evolution. New J Phys 10(7):073020
19. Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L (2011) Flavor network and the principles of food pairing. Sci Rep 1:196
20. Anderson EN (2005) Everyone eats. Understanding food and culture. New York University Press, New York
21. Zhu Y-X, Huang J, Zhang Z-K, Zhang Q-M, Zhou T, Ahn Y-Y (2013) Geography and similarity of regional cuisines in China. PLoS ONE 8(11):e79161

22. West R, White RW, Horvitz E (2013) From cookies to cooks: insights on dietary patterns via analysis of web usage logs. In: Word Wide Web conference (WWW)
23. Webber W, Moffat A, Zobel J (2010) A similarity measure for indefinite rankings. ACM Trans Inf Syst 28(4):20
24. Kiefer I, Haberzettl C, Rieder C (2000) Ernährungsverhalten und Einstellung zum Essen der ÖsterreicherInnen. J Ernährmed 2(5):2-7
25. Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet Math 1(2):226-251
26. Sornette D (1998) Multiplicative processes and power laws. Phys Rev E 57(4):4811
27. Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B 213:21-87
28. Sornette D, Cont R (1997) Convergent multiplicative processes repelled from zero: power laws and truncated power laws. J Phys I 7(3):431-444
29. Logue AW (2004) The psychology of eating and drinking. Psychology Press, New York
30. Sanjur D (1982) Social and cultural perspectives in nutrition. Prentice Hall, New York
31. Drewnowski A (1997) Taste preferences and food intake. Annu Rev Nutr 17:237–253
32. Sherman PW, Billing J (1999) Darwinian gastronomy: why we use spices. BioScience J 49(6):453–463
33. Teng C-Y, Lin Y-R, Adamic LA (2012) Recipe recommendation using ingredient networks. In: Proceedings of the 3rd annual ACM web science conference. WebSci? 12, pp 298-307
34. Vlachos M, Meek C, Vagena Z, Gunopulos D (2004) Identifying similarities, periodicities and bursts for online search queries. In: Proceedings of the 2004 ACM SIGMOD international conference on management of data. SIGMOD? 04, pp 131-142
35. Bennett PN, Radlinski F, White RW, Yilmaz E (2011) Inferring and using location metadata to personalize web search. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. SIGIR? 11, pp 135-144
36. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012-1014

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com