A recommender-network perspective on the informational value of critics and crowds

Pantelis P. Analytis[1], Karthikeya Kaushik[2], Stefan M. Herzog[3], Bahador Bahrami[4], and Ophelia Deroy[5]

[1]Department of Business and Management, University of Southern Denmark
[2]Department of Psychology, UC Berkeley
[3]Adaptive Rationality Center, Max Planck Institute for Human Development
[4]Department of General Psychology and Education, Ludwig Maximilian University of Munich
[5]Department of Philosophy, Ludwig Maximilian University of Munich

March 29, 2024

Abstract

How do the ratings of critics and amateurs compare and how can they be combined? Previous research has produced mixed results on the first question, while the second remains unanswered. We have created a new, unique dataset, with wine ratings from critics and amateurs, and simulated a recommender system using the weighted *k*-nearest-neighbor algorithm. We then formalized the advice-seeking network spanned by that algorithm (i.e., who advises whom?) and studied people's relative influence. We find that critics are more consistent than amateurs, and thus their advice is more predictive than advice from amateurs. Getting advice from both groups can further boost performance. Our network-theoretic approach allows us to identify influential critics, talented amateurs, and the information flow between groups. Our results provide evidence about the informational function of critics, while our framework is broadly applicable and can be leveraged to devise good decision strategies and more transparent recommender systems.

Keywords: Wisdom of crowds; expert crowd; social influence; recommender network; taste homophily, social learning.

1 Introduction

Whether it is a film currently in cinemas, a restaurant that just opened, or vintage wine, people like to voice their judgments on matters of taste. Even more so, for a few select individuals—critics—expressing judgments on matters of taste has turned from a fun past time to a profession. Critics are employed in the daily and weekly press to assess restaurants, theater, movies, and wine labels, while some even run popular television shows. It has been argued that critics function as information producers, generating information about products that becomes available to the wider public [43, 71]. This key function of critics is especially pronounced on websites such as Metacritic and Rotten Tomatoes, which have based their business model entirely on devising and displaying scores that summarize critics' opinions, or when they are hired as judges in prestigious competitions. In fact, it has been shown that the judgments of influential critics can predict [29] or even alter the overall financial performance of products [3, 12]. But are the evaluations of critics more informative than those of amateur consumers, and if so, why?

Even though there is clear evidence for the sizable market impact of at least some critics, the informational value of their opinions for the broader public has often been contested. Popular lore and scholars in different traditions have argued that in matters of taste, there is often a big, unresolvable divide between critics and the general public [30]. Some would suggest that critics and other experts engage in hairsplitting over nuances that are often lost on the rest of the world, who consider expert opinions high-brow, and ivory tower, if not altogether imaginary or downright boring [34, 14]. Others emphasize that general audiences lack the ability to assess sophisticated (cultural) products, and only people who have trained their senses are capable of doing so [2]. Regardless of the point of view adopted, other people may rarely care about what critics think, whatever theirs standards of taste may be. Thus, although critics's opinions may be impactful (e.g. because they bring attention to certain categories of cultural or experience products [72]), they could also be less informative than the opinions of representative individuals in the crowd or even a randomly selected person from the street [41]. That is, the opinions of a critic or a group of critics might be less correlated with and, most importantly, less predictive of the tastes of people in the wider public, than those of other amateurs.

A number of past studies in marketing, psychology, sociology, and cultural economics have explored whether the opinions of critics and amateur consumers correlate in different taste domains and have produced mixed results [41, 42, 78, 37, 57, 13]. Holbrook, for example, finds modest correlations between the tastes of critics and the general public in the case of movies [41, 42]. Other studies looking at Broadway productions, popular music, and plays have found slightly negative [37, 57] or even strong positive [78] correlations between critics and the wider audience. Taken at face value, the previous literature remains inconclusive, and even if most studies seem to suggest that critics opinions relate to some extent to those of the wider public, it is still unclear whether critics' opinions are more informative than those of amateurs. Furthermore, although it is obvious that there is substantial variation in the extent to which different critics or amateurs can influence or inform others [19], most existing studies comparing the tastes of critics and amateurs disregard it and focus on reporting average correlations between the two groups. As a consequence, there are no methods for identifying the best critics to follow in a certain domain. Last, rather than focusing on drawing advice from specific individuals or groups, it would be valuable to know when it is beneficial combine advice from different critics and/or amateurs.

Several questions about the ways in which critics can inform and influence the broader public remain unanswered. First, are their opinions more informative than those of amateurs, and if so, why? Second, is it possible to identify informative critics and talented amateurs who have the potential to become critics? Third, how can the opinions of critics and amateurs be best combined and how would information flow between the two groups in an efficient advice network? To answer these questions we need a methodology that allows us to measure the informational value of the opinions of different groups of individuals for others as well as the individuals' potential to influence others.

We put forward an approach that combines methods from the recommender systems [15, 4], machine learning [73] and network science communities [48, 23] and from the study of expert judgments in matters of fact, where different strategies for choosing among experts or identifying the best experts have been explored in domains ranging from medicine to agriculture [69, 8, 50]. We leverage the *weighted k-nearest neighbors* algorithm (k-nn), a classic recommender systems algorithm that encodes an array of strategies for choosing among experts (and amateurs) as special cases [5] (see also 1). For each individual and item, our implementation of the algorithm draws advice from the *k* most similar other individuals in the database who have evaluated that item and weights their opinions according to a similarity sensitivity parameter to form a prediction of how much the target individual will like the item. Thus, the main difference between *k-nn* and previously proposed strategies for choosing among experts is that weighted *k-nn* relies on observed similarity between the target individual and potential advisers in past ratings rather than the observed performance in predicting some quantity of interest. By assessing the out-of-sample performance of the algorithm when drawing advice from critics, amateurs, or both groups and varying the number of neighbors *k* and the similarity sensitivity parameter ρ, we can compare the predictive performance of different strategies involving each or both of these groups for each person, and thus assess their relative informational value at the individual and aggregate level.

The *k-nn* algorithm spans an advice network among different individuals (i.e., who advises whom?) [6], and it is thus possible to visualize and study the properties of such a "taste network" in any dataset where a group of people have evaluated a set of items, even when the overlap in the ratings of different people is relatively small (i.e., sparse rater–item matrices). Individuals (advisors) whose tastes appear relevant for many similar others (advisees), be they critics or amateurs, are sought often for their advice from the *k-nn* algorithm—they have a large *recommender potential*. This potential, however, can only materialize as *recommender influence* when the advisers have experienced the items that their advisees are considering, and can thus supply their ratings when the algorithm (or an advisee) seeks it. To assess the flow of influence between two categories of individuals—critics and amateurs—we adapt the notion of homophily from network science and show how it can be applied to domains of taste. The concept of *taste homophily* is general and can be applied to any rating dataset where items have been evaluated by two or more categorical groups of raters. Thus, our methods be applied by social and behavioral scientists in other taste domains, and they can also be used by the recommender systems community to improve the transparency and interpretability of the recommendation process.

Since the advent of the internet and social media, the informational landscape in the wine world has been changing swiftly. Vivino has emerged as a reliable alternative to critics for getting access to information about the quality of different wines, and as with other opinion aggregation websites (e.g. IMDB, Yelp, Amazon) it gives wine consumers a way to consult a single information source to inform their choices, bringing forth a democracy of taste where each individual in the user base can contribute to the aggregate wine ratings [46]. Personalization algorithms can search for similarity patterns between one person and others in rating databases, and they can receive recommendations based on a number of apparently similar people from across the globe, thus circumventing critics altogether [1, 67]. To examine how the opinions of critics relate to those of amateurs we created a new, unique dataset consisting of the ratings from both renowned wine critics and regular wine consumers (i.e., amateurs, non-professionals). We obtained critics' data from Bordoverview, a website summarizing the en primeur ratings (first sampling of a production year) on Bordeaux wines, and matched these ratings with amateur data on the same wines from Vivino. The resulting dataset has the properties needed to assess how wine critics and amateurs can inform and influence others and how their judgements on matters of taste can be best combined.

2 Methods

Bordeaux wine dataset We first obtained ratings by professional critics from Bordoverview and ratings by amateurs from Vivino by scraping the two websites. We then combined the two datasets and restricted our analyses to wine labels that were included in the Bordoverview list and had at least 5 reviews in Vivino and to Vivino users with more than 50 ratings in the resulting wine label list. This resulted in a dataset comprised of 1978 wine labels (322 different wines across 15 different vintages from 2004 to 2018), 14 professional critics (or wine magazines or other outlets employing critics), and 120 Vivino amateurs. The dataset has 25,907 ratings in total, and average density (i.e., mean proportion of rated wine labels relative to all wine labels) in the dataset is 4.7% for amateurs and 53.3% for professional critics. Ratings in both datasets were normalised within each rater using *z*-scoring, that is, transformed ratings now indicate for each rater how many standard deviations a rating was above or below the mean rating of that rater. This was done to make the ratings more comparable across the different rating scales that different critics use (e.g., 10 to 20 for Jancis Robinson vs. 75 to 100 for Jeff Leve) and the 1-to-5 scale used by Vivino.

Recommendation algorithms In our analysis, we rely on the well-established *k*-nearest neighbors algorithm (*k-nn*) [66, 67, 28], that seeks the *k* most similar individuals, allowing for differential weights [15, 62].

Such a weighted nearest neighbor algorithm can be expressed as follows:

$$\widehat{u_m} = \frac{1}{\sum_{j=1}^{k} w_j} \sum_{j=1}^{k} w_j \times u_j \tag{1}$$

where $\widehat{u_m}$ is the estimate of the utility of an item m for the target individual, j is the jth nearest neighbor to that target individual, and w_j the weight put on that other individual. For $k = 1$, the algorithm seeks advice from only the most similar other individual. Setting $k = N - 1$, where N is the total number of individuals in a dataset, amounts to the weighted averaging strategy. For values of k between these two extremes, we obtain the usual *k-nn* implementation, with differential weights.

We used the Pearson correlation coefficient as a measure of similarity (w) between two individuals i and j [36], defined as follows:

$$w(i,j) = \frac{\sum_{m=1}^{M} (u_{im} - \bar{u_i})(u_{jm} - \bar{u_j})}{\sum_{m=1}^{M} \sqrt{(u_{im} - \bar{u_i})^2 (u_{jm} - \bar{u_j})^2}} \tag{2}$$

where u_{im} is the rating that the target individual i gave to item m and u_{jm} is the evaluation that the jth individual gave to the same item m. M stands for the total number of items.

We use a similarity sensitivity parameter ρ that allows us to amplify or dampen the weights $w(i,j)$ of other people [15]. We directly modify the weights obtained from Eq. 2 using the following scheme:

$$w'_j = \begin{cases} w_j^{\rho} \; if \; w(i,j) \geqslant 0 \\ 0 \; otherwise. \end{cases} \tag{3}$$

By varying k and ρ, we can produce several social learning and information aggregation strategies studied in the behavioral and management sciences (Table 1; also see [5, 6]), some of which have been used to study how to best combine expert opinions (i.e. forecasters, doctors, etc., see [8]). For instance, setting $\rho = 0$ and $k = N-1$ produces the unconditional wisdom of the crowds strategy [27]. Setting $\rho = 0$ and $k = n$ gives the original unweighted formulation of *k-nn*, which corresponds to the select crowd strategy in psychology and management [58]. Setting $\rho > 0$ weights the opinions of people more similar to the target individuals. This is common in implementations of the nearest neighbors strategy in collaborative filtering [15] and strategies aggregating opinions of more competent experts in management [18]. As ρ increases, the weight distribution becomes more unequal and the most similar individuals have a proportionally larger weight. In addition to applying *k-nn* to our entire dataset we applied the algorithm to subsets of the data, giving it either access to only the ratings of critics or amateurs (i.e. searching for the k most similar critics or amateurs).

We took two measures to make our analysis routines robust to inconsistencies that could be produced due to the sparsity of the dataset. First, we only considered correlations when the number of overlapping items between target individual i and adviser j is higher than 5 and set all the remaining correlations to the mean correlation between individual i all other individuals j with whom they had an overlap of more than 5 items. Using such thresholds (or other methods of discounting observed correlations from sparse data) is a common approach when deploying collaborative filtering algorithms. Further, when some of the k most correlated individuals had not evaluated the target item, the algorithm searched further down the list of others ranked by similarity until a committee of k people was formed or there were no further potential advisors (in which case the committee had fewer than k advisers). This is a less common implementation of the *k-nn* algorithm, but suitable for the size of our dataset and our research objectives.

Performance of *k-nn* We assessed the out-of-sample performance of the weighted *k-nn* algorithm by consistently leaving out 10 items for each individual and using the remaining items to learn the correlations between individuals. This approach is a variation of the leave-one-out approach in recommender systems [21], and ensures that that there are sufficient pair comparisons to be predicted in the test set per individual. The correlations learned in the training set were then used to find the k most similar individuals to the target individual who have evaluated the item and were then up(down)-weighted according to ρ (see Eqs. 1 & 3). Thus, for each target individual and item, k other individuals (whose opinions where weighted according to ρ) were used to predict how much the target individual would like that item. We repeated this process 1000 times and averaged the results across repetitions. As a measure of performance we used the number of correct decisions made by the model when choosing between a pair of items in the test set (45 choices in total, ties were resolved at random). We opted for this measure because it is intuitive and easily communicable (1 corresponds to perfect choices and 0.5 correspond to random choices, also see [5, 7]). It has been used before to assess the performance of decision or inference strategies in psychology and the management sciences [31] and it closely corresponds to the number of concordant pairs measure used by the recommender-systems community [47].

Social learning (taste)	**Social learning (objective)**	**Algorithm parameters**	**Cognitive strategies**
Doppelgänger [79, 5]	Follow the expert [51]	$k = 1$ and ρ = any	Take the best [32, 40]
Clique [61, 79]	Select crowd [58, 33]	$k = n$ and $\rho = 0$	–
Weighted clique	Weighted select crowd [75]	$k = n$ and $\rho > 0$	–
Weighted crowd [5]	Weighted crowd [18]	$k = N - 1$ and $\rho > 0$	Weighted additive [65, 24]
Whole crowd [79, 5]	Averaging [39]	$k = N - 1$ and $\rho = 0$	Equal weights [24, 27]

Table 1: Correspondence between the collaborative filtering algorithm parameterizations we consider (see Equations 1, 2, and 3) and the social learning and information aggregation strategies broadly studied in the social and behavioral sciences [6]

Reconstructing the *k-nn* advice network We studied the advice network spanned by *k-nn* [52, 6] by constructing advice networks—for different values of k and ρ—with nodes representing the different individuals in the dataset and directed edges representing an individual seeking advice from other individuals (or more precisely, *k-nn* seeking advice on their behalf). While all individuals had by definition the same number of k outgoing edges connecting them to other nodes, people could have a varying number of incoming edges depending on how often *k-nn* sought their advice for other individuals. We used node strength, defined as the sum of weights of the incoming edges as as a measure of social influence that naturally fits the weighted *k-nn* algorithm and weighted networks more generally [11]. For $\rho = 0$ this measure collapses to in-degree. We then measured the *recommender potential* of each individual, by calculating the node strength resulting from the *k-nn* algorithm when disregarding missing values. That is, we counted only how often people were in the first k individuals sought by the algorithm and their relative weights in the committees formed (expressed as a proportion), regardless of whether they had a rating to contribute for the item in question. We then calculated the *recommender influence* of an individual by computing who actually contributed to recommendations and how much so. That is, we calculated how often our implementation of the *k-nn* algorithm sought and used advice from an individual to predict how much another person would like an item, and the relative weight of such advice in the committees that eventually formed. Note that these two metrics of influence converge to the same metric for full (i.e., non-sparse) datasets. Following previous analyses we report results averaged across 1000 repetitions of the simulation.

Calculating taste homophily: We calculate taste homophily in the two groups by adapting the homophily index used in the work of Currarini [23] to weighted networks and then using a baseline that is appropriate for sparse data structures. We define N_i as the number of individuals of type i in the population and N the total population. Similarly, we define R_i as the number of ratings contributed from individuals of type i and R the total number of ratings. Then $p_i = N_i/N$ is the proportion of individuals of type i in the population and $r_i = R_i/R$ is the proportion of their ratings. We will use these two measures as baselines to mark whether the tastes of a group are characterized by homophily. The homophily index H_i then is defined as the proportion of weights s_i directed to members of the same group, divided by the total sum of weighted nodes (both weights directed to the same and to different groups, d_i), that is, $H_i = \frac{s_i}{s_i+d_i}$. A group is characterised by taste homophily in respect to their proportion in the population if $H_i > p_i$ and in respect to the ratings they have contributed if $H_i > r_i$. The first baseline is similar to standard definitions of homophily in the network science literature, but the second one also takes into account data density disparities in the considered groups — a feature of most real world recommender systems. The same definitions can also be used at the individual level, to evaluate whether specific individuals draw information from people belonging in the same group or from people outside their group. Note that the homophily index can be also calculated taking into account only the initial calls of the algorithm, as with recommender potential. In that case only the relation $H_i > p_i$ is relevant (we perform this analysis in the Supplementary Material).

3 Results

Critics are more consistent than amateurs In line with previous findings on the judgement consistency of experts vs. non-experts in matters of fact [9, 69], we find more agreement among professional critics than among Vivino amateurs. The average taste similarity (correlation) among critics is 0.60, whereas the average taste similarity among amateurs is 0.27 (see Figure 1 left panel). A similar result also emerges when combining the critics and amateurs datasets into a single dataset and then calculating correlations across all individuals (i.e., irrespective of group membership; see Figure 1 right panel). The average correlation between critics and everybody else is 0.39, whereas the average correlation between amateurs and everybody else is 0.29. These results are preserved even if we randomly remove ratings from the critic group to equate the average density of the two groups (see Figure 5 in the Supplementary material).

Critics tend to be representative of the amateur audience Critics also tend to be more representative of the amateur population than other amateurs. The average similarity between critics and amateurs is 0.36, which is substantially higher than the average similarity among amateurs, which is 0.27. Nonetheless, for most amateurs the highest encountered correlations in the dataset are with other amateurs. There are two reasons for that: first, there is a sizable sub-group of fifteen to twenty amateurs whose average correlations with the amateur audience is larger than the average critic-amateur correlation (also see Figure 6 left panel in the Supplementary Material). Second, the amateurs have greater dispersion in their observed taste similarities with other individuals. Thus, although the critics are quite representative of the amateur audience, it is not clear whether they are the best source of advice, and whether or how their ratings should be combined with those of amateur crowds.

Following similar critics has high prediction value We next compare the performance of recommender systems based on only amateurs or only critics in predicting the ratings of the amateur audience. When predicting amateurs, a recommender system based only on critics performs better than a recommender system based on amateurs — the difference is substantial, and larger than 3% in terms of prediction rate for the average amateur (compare the bold orange line with dash dot orange line in Figure 2 left panel; the presented results are obtained by setting $\rho = 1$) for any possible k value. Even consulting the most similar critic can improve the performance by more than 1% as compared to aggregating ratings from several similar amateurs (compare the leftmost point in the bold orange line with the rightmost point in the dash-dot orange line in Figure 2 left panel). Taking advice from additional critics can marginally further improve performance. Because there is high consistency across critics, there is not as much new information when considering the opinions of ad-

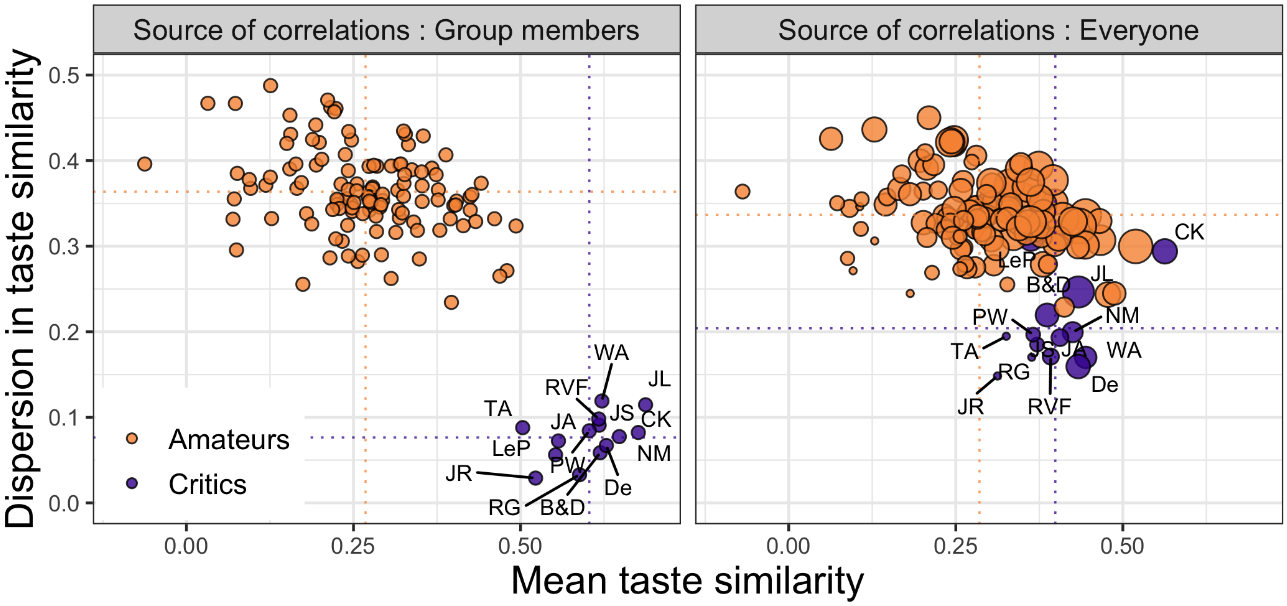


Figure 1: **Intercorrelations with members of the same group and all other individuals. Left**: The position of 14 professional critics and 120 amateurs on a 2-dimensional plane defined by mean taste similarity (i.e., mean correlation) and dispersion in taste similarity (i.e., SD of correlations) with members of the same group. **Right**: The position of the same 14 professional critics and 120 amateurs on the same plane, but this time with taste similarity calculated across all individuals (professional critics and amateurs). The color in both panels indicates whether an individual is a professional critic or an amateur and the point size in the right panel indicates recommender potential. The dotted orange and purple lines indicate the average correlations and dispersion in taste similarity for amateurs and critics. Only correlations when two individuals had an overlap of more than 5 ratings were considered in this graph. **Initials of professional critics**: WA — Lisa Perotti Brown, NM — Neal Martin, JR — Jancis Robinson, TA — Tim Atkin, B & D — Michel Bettanne and Thierry Desseauve, JS — James Suckling, JL — Jeff Leve, De — Steven Spurrier, James Lawther, Beverley Blanning and Jane Anson, RVF — Olivier Poels, Hélène Durange, and Philippe Maurange, JA — Jane Anson, LeP — Jacques Dupont, PW — Ronald DeGroot, RG — Rene Gabriel, and CK — Chris Kissack.

ditional critics. These results hold regardless of the parameter ρ used in the simulations (see Figure 3 in the Supplemental Material). In most cases, the observed performance differences translate directly to the individual level (see Figure 2 right panel). That said, there is some variability in the population. For some amateurs, for example, consulting the one or two most similar critics is the best performing strategy, whereas for other individuals, taking advice only from amateurs performs best (i.e. see leftmost individuals in Figure 10 in the Supplement and compare them to the individual at the bottom center).

Ratings from critics and amateurs can complement each other We next turn to the performance of a recommender system that uses the ratings of both critics and amateurs. For *k* values lower than five, such a recommender system would perform modestly. In fact, people would be better off discarding the data from amateurs and considering only similar critics (compare the leftmost parts of the orange bold line and the orange dashed line in Figure 2 left panel). The drop in performance for low *k* values is substantial and further stresses the informational value of critics, as their ratings generalize better to unseen items than those of apparently similar amateurs. Still, for *k* values equal to or larger than five, a recommender system using data from both

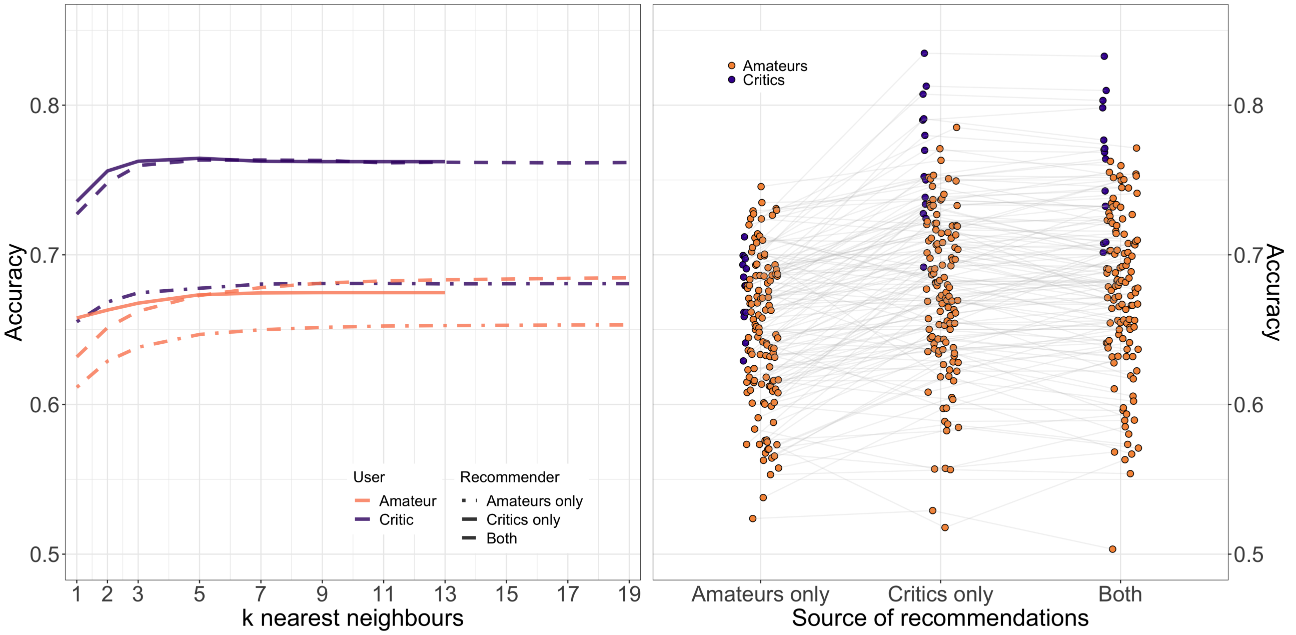


Figure 2: **Performance of the recommender system for different groups (left) and individuals (right). Left**: The average performance of the *k-nn* algorithm based only on amateurs, only on critics, or both amateurs and critics for different values of k for the amateur and critics groups. **Right**: The individual level performance of the *k-nn* algorithm based only on amateurs, only on critics, or both amateurs and critics for $k = 5$.

critics and amateurs performs best for the amateur audience, even if only slightly so in our dataset, and can further improve predictive performance (see Figure 2 left panel). This implies that the ratings generated from critics and amateurs can function as complements. Aggregating ratings from acclaimed critics and apparently similar amateurs in a single recommender system can help counteract the high statistical variance in the amateur ratings, and make the most of information from highly similar amateur individuals, leading to overall better predictive performance.

Critics (and critic-like amateurs) are more predictable Regardless of the method used, critics are much more predictable than amateurs. Predicting critics' ratings using amateur ratings (the worst performing method for critics) leads, on average, to performance similar to predicting amateurs' ratings using ratings from both critics and amateurs (the best performing method for amateurs for high k values, see left panel in Figure 2). Further, when predicting critics' ratings the less-is-more effect persists at even higher k values — using data from amateurs and critics cannot improve performance compared to using data only from critics. In addition to being more consistent, critics are, on average, much more similar to everybody else (high mean-taste similarity) and they have rated many more wines. The latter two features can capture much of performance variability in collaborative filtering recommender systems [1, 7]. The same features can also predict recommender system performance when looking only in the amateur population. In fact, the data of several amateurs in our dataset can be predicted by the recommender algorithm with an accuracy similar to that obtained for critics (see the colour of the nodes in Figure 3A). [1] The exact same features are also predictive of people's potential to

[1] A linear model using mean-taste similarity, dispersion in taste-similarity, and number of reviews as features could account for more than 65 % of the variance in the prediction rates for different individuals for k = 5 and $\rho = 1$ using data from both critics and

influence others in recommender systems [6], the topic to which we will turn next.

Some individuals have a larger potential to influence others People differ substantially in their recommender potential (see Figure 3A), that is, the total weight they could have in recommendations made for all other individuals (i.e., the sum of advice weights attributed to them by *k-nn*, not yet considering whether that individual actually rated a particular wine for which a recommendation is sought, also see Methods). Figure 3A, shows exactly how the *k-nn* algorithm, seeks advice for the target individual from other individuals for $k = 5$ and $\rho = 1$. Among the critics, Jeff Leve has the largest recommender potential. Other world renowned critics such as Jancis Robinson and Tim Atkin score relatively low in terms of this metric. Further, a subset of the amateurs have larger recommender potential than most critics (see the points in the upper left side of Figure 3), which indicates that their rating patterns appear to be similar to those of many other individuals in the amateur audience (also see Figure 6 left in the supplementary material). Some of these amateurs might have the potential to become influential critics.

The number of contributed ratings moderate people's influence To derive estimates about how much people would enjoy specific wines, our implementation of the *k-nn* algorithm calls in the adviser committee: the *k* individuals with the highest correlations to the target individual who have evaluated a specific item. That is, for the same target individual, the adviser committee may differ from item to item depending on who has evaluated specific items. This also implies that recommender potential (defined by the initial choices of *k-nn*, disregarding missing values) does not directly translate to recommender influence, as individuals with high potential may have evaluated only a few items. The lower right panel of Figure 3 shows the recommender influence of different individuals in the population for $k = 5$ and $\rho = 1$. On average, professional critics exert a much larger recommender influence than amateurs (5.54 vs .47) because they have evaluated many more items, and they are often consulted by the amateur audience in this recommender system. Even among the critics, however, there are some notable differences: Jeff Leve, the critic with the highest recommender potential, recedes in recommender influence because he has evaluated only 40% of the wines, while the journal Decanter, which has the largest number of rated items, would have the most influence in a wine recommender system built from these data. The recommender influence of different individuals can be also evaluated for specific wines, and depends on who else has rated that specific wine in the Supplementary Material (see Figure 8).

The amateurs are (mostly) seeking advice from outside their group; the critics from inside We next access the degree to which people get advice from individuals of the same or different group when the data from both groups were used in the recommender system. We calculated the homophily index (see Methods) of the amateur and critic groups for different values of *k* and ρ in the space of possible parameter configurations spanned by the *k*-nearest neighbors algorithm and we contrasted the index with the population proportion and rating proportion baselines (see Methods). When accounting for the proportion of ratings contributed from the two groups, the amateur group is characterized by slight inbreeding homophily, for low values of *k*, and heterophily for intermediate values of *k* (see Figure 4 left panel, the measure necessarily converges to the group's proportion of ratings for higher values of *k*). This result could partly explain the less-is-more effect in the recommender system performance for low values of *k* because for these values advice is mostly drawn from apparently similar amateurs, resulting in high prediction variance. Higher ρ values tend to increase homophily (decrease heterophily) for low *k* values. Note that the amateur group is strongly heterophilous considering

amateurs (adjusted $R^2 = 0.67$).

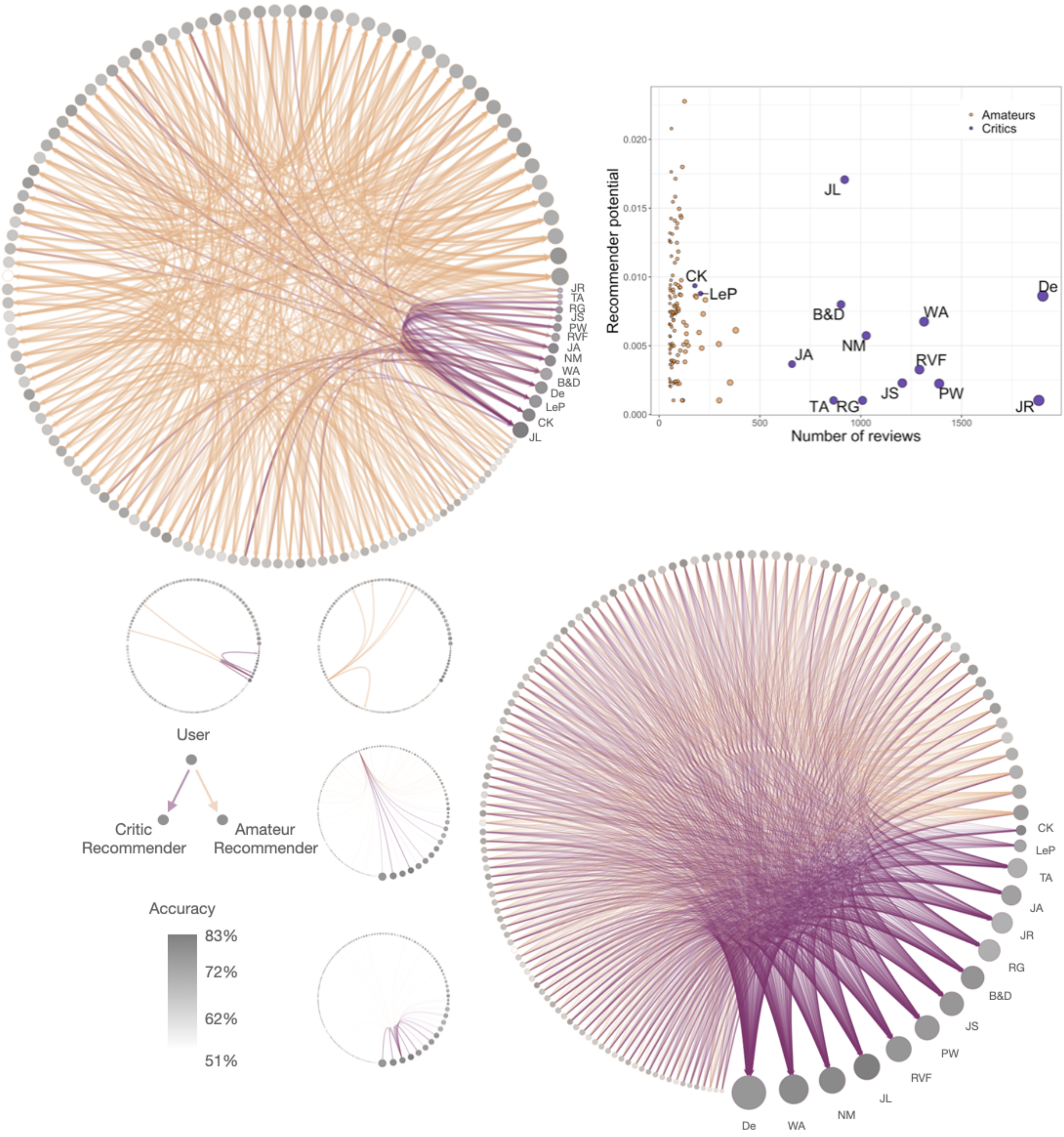


Figure 3: **The recommender potential and recommender influence of different individuals.** Nodes represent individuals, node size represents recommender potential (upper left circle) or recommender influence (bottom right circle) of different people in the recommender network spanned by the *k-nn* algorithm. **Bottom Left**: Orange edges indicate that advice is sought (or provided) from an amateur and purple edges indicate that advice is sought (or provided) from a professional critic. The colour of the nodes indicates the accuracy of the algorithm for different individuals in the dataset. Edges with weights smaller than 0.05 do not appear in the visualization to prevent overcrowding the graph. **Upper Left**: The The advice-seeking network produced by the initial call of *k-nn*, disregarding missing values (i.e., recommender potential). The edges (arrows) are pointing to the individuals from whom *k-nn* first seeks advice for the target individual. **Lower Right**: The influence graph eventually produced by *k-nn*. When an individual called by *k-nn* has not rated a wine label, the next individual in the correlation rank is consulted. This process continues until *k* advisers have been found or until the pool of potential advisers is exhausted. **Upper Right**: Amateurs and professional critics placed on a 2-dimensional plane defined by the number of items they have evaluated (x-axis) and their recommender potential (y-axis). Critics are depicted with purple color and amateurs with yellow. Node size indicates the total influence of different individuals.

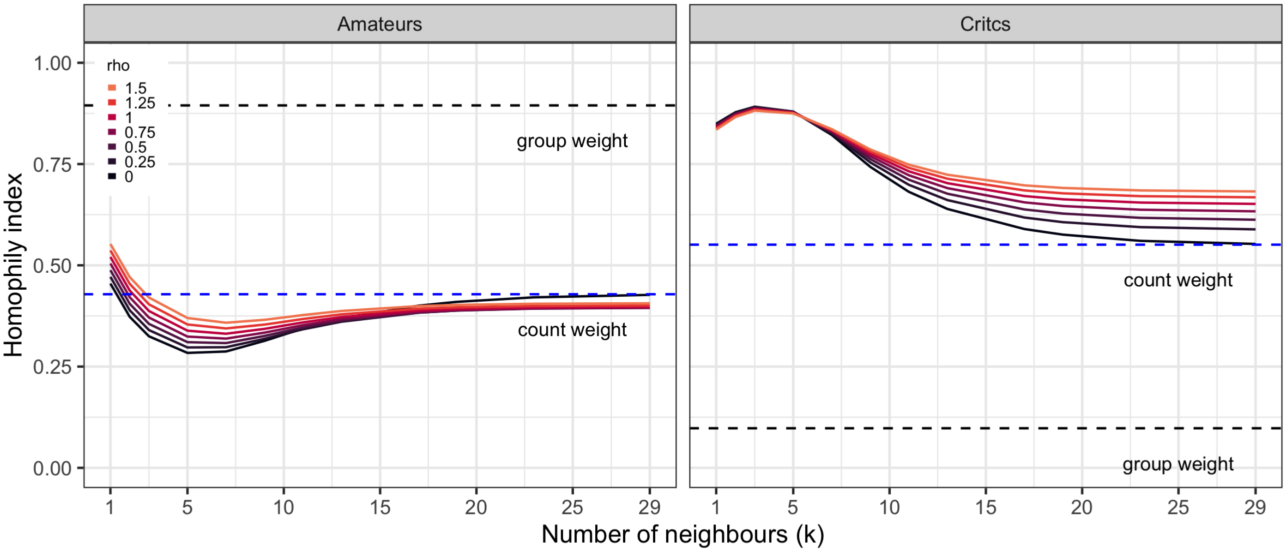


Figure 4: **Homophily index of critics and amateurs Left and Right**: The homophily index of amateurs and critics as a function of the value k in the k-nearest neighbors algorithm. Different ρ values are represented with lines of different color. The horizontal dashed lines represent homophily baselines corresponding to the proportion of group members in the population (group weight) and the proportion of the ratings contributed by the members of each group (count weight).

their proportion of the population because they draw less than 50% of advice from other amateurs although they represent more than 85% of the population. The wine critics, by contrast, are drawing most of their advice from other critics and are characterized by inbreeding homophily regardless of the baseline used. The homophily index in the critics group is particularly pronounced for a low number of neighbors k and levels off as k increases and more amateurs are (by necessity) consulted to advise the critics. For critics, higher ρ values tend to lead to increased homophily for high k values.

4 Discussion

Going back to the 18th century, the philosopher David Hume argued that critics are better equipped through training and natural predisposition to judge on matters of taste [44], and his work has sparked philosophical debates about the potentially objective nature of judgement in matters of taste [54]. Hume even preempted research on the wisdom of the crowds suggesting that two expert judges are better than one in matters of taste. But are expert judgments on matters of taste representative of those of the wider public and should they be trusted? Further, do data from select critics complement or substitute for crowd-sourced data online or advice from friends? And is it possible to identify critics whose opinions are more informative or influential or even amateurs who have the potential to become successful critics? Last, how would information flow between critics and amateurs in a recommender system that relies on the *k-nn* algorithm? We put forward a novel methodological framework that makes it possible to address these questions in any domain of taste, and applied it to the case of wine, a domain where expert judgement has been particularly revered.

4.1 Studying expertise: from matters of truth to matters of taste

Over the last few decades, disagreements among renowned wine critics such as Robert Parker and Jancis Robinson in regard to the quality of certain wine vintages have captured the attention of wine aficionados [68]. In addition, empirical studies have raised doubt about the ability of expert wine judges to discriminate across wines in blind tastings [38]. Is there a good case for expertise in wine or in other taste domains? Scholars in applied psychology and the management sciences have identified a number of conditions that characterize expert judgment or prediction. Two key properties of expert judgment are (i) consistency across judges [26, 59], and (ii) or discriminating ability [70] (i.e. or the ability of judges to evaluate similar items with similar ratings). Previous work on wine experts in blind tasting settings has noted average correlation among them ranging from 0.2 to 0.5 depending on the study [9, 22, 16]. In our data, we clearly see that renowned wine critics are highly consistent—the average mean correlation between each individual critic and other critics is higher than 0.5 for all individual critics. Thus, the average correlation rate is higher than reported in previous studies on red wines and similar to what is observed in some domains of facts in medical and business contexts [9]. The increased level of consistency across judges might reflect the higher level of expertise of the critics included in our study, who are arguably some of the most renowned tasters in the world, and might share the same evaluative schemas or the fact that en Primeur ratings are intended to be predictive of future wine quality and are not blind. Our results also imply that critics are more discriminatory than amateurs—the levels of observed agreement would not be possible if their judgments were not guided by clear and shared criteria about what makes a good wine or if their judgments were too noisy [17]. Overall, our findings indicate a high level of agreement among critics and show that that expert judgments for wine are characterised by similar statistical properties as in some domains of truth.

4.2 Amateurs and critics' ratings: substitutes or complements?

In the past, a number of studies have looked at correlations between critics and amateur raters. In their majority, these studies have found that amateur and expert opinions are only mildly correlated [41, 42, 78]. Although valuable, studies comparing the tastes of different groups using correlations leave a lot to be desired—correlations do not immediately translate to predictive power, especially when the opinions of several judges are aggregated. Going beyond previous work we evaluated the performance of different strategies that people can use to get recommendations or predict their own tastes, when seeking advice from critics, amateurs or both groups. We found that relying on the opinions of just one critic led to better predictive performance for most amateurs than seeking advice from several other amateurs. This is because critics are both more consistent and prolific in their evaluations. Having evaluated many wines helps to correctly estimate correlations between amateurs and critics and to effectively generalize in unseen data. Hume's conjecture that two critics are better than one was vindicated in our analyses, but the additional gains are modest. Overall, in the case of wine, high observed similarities with critics tend to be robust, and therefore the ratings of critics can be valuable proxies that help people predict what they like. This may explain why many renowned critics have been able to monetize the information value of their reviews by introducing subscriptions to their websites, and why the *following-the-most-similar critic* heuristic is a common decision strategy among wine afficionados [74].

There is one previous study from the recommender systems community that allows us to compare our results with those from another domain: Amatriain and colleagues compared a *k-nn* recommender system based on the ratings of select film critics to a system based on the ratings of thousands of amateurs [4]. They found that recommender system users receive slightly more predictive recommendations when relying on a large database with opinions from other amateurs than when drawing recommendations from a database of select critics (but using movie critics has a number of advantages in terms of scalability, privacy, etc.). The somewhat diverging results could reflect fundamental differences in the acquisition of expertise in the two domains: whereas watching films is almost equally accessible to everybody, and is mostly constrained by the time budget that people have available, opening and tasting wines is mostly subject to budget constraints, and can quickly become an expensive hobby, reducing the number of training samples that most amateurs have access to, but also their capacity to supply information to others. This is also reflected in the relative volume of ratings supplied by critics as compared to amateurs, which is larger in our study.

Last, going beyond considering information generated from critics and amateurs separately, as in the study by Amatriain and colleagues, we also examined what happens when information is drawn from both critics and amateurs, and found that for large values of k there is a margin to further improve recommendations. This indicates that the data points generated by critics and the potentially more voluminous data generated by amateurs in online interfaces could be complementary in a collaborative filtering recommender system. What is more, our methods make it possible to identify the informational contributions of different individuals or groups for specific items (see Figure 8 in the Supplement) and in the aggregate.

4.3 From real-world to in silico advice networks and back

Following the groundbreaking work of Katz and Lazarsleld in the 1950s [45], social scientists have used survey methodologies to elicit advice networks across domains of life (including fashion, politics, and beyond) as well as to uncover the structure of professional advice networks [53] and informal organizational networks [48]. One of the main findings emerging from this research stream is the existence of informal opinion leaders who are sought for their advice by many other individuals. Similar to the opinion leaders of real-world advice networks, some individuals are much more often sought by the *k-nn* algorithm and provide advice to many similar others [6]. Our analysis routines allowed us to uncover the position of critics and amateurs in such in wine advice networks in silico and showed how their influence is modulated by the parameters k and ρ of the *k-nn* algorithm, but also by the statistical properties of people's tastes and whether they are prolific raters. As such, the networks produced by *k-nn* can be seen as informationally efficient advice networks for matters of taste and can be compared in terms of their structure and properties to the advice networks formed by people offline or in online platforms [56, 55]. Thus, the networks produced by *k-nn* could provide new hypotheses about the formation of real-world advice networks and can help disentangle informational and other motivations when forming new ties with potential advisors.

4.4 Relative expertise and identifying talent

In domains of fact it is relatively easy to identify who is the best or the most accurate judge by looking at the prediction success of different judges in past data [70, 49]. Is there a way to assess the relative worth of judges in matters of taste, where there is no objective truth to be predicted or when long records of past data are not available? This would allow us to identify critics whose opinions are particularly valuable to others, as well as to identify talented amateurs who have the potential to become critics. One possibility is to use the average ratings for an item (i.e. wine) as a gold standard and to assess how different judges can predict it [20]. Although this could be a good assumption in some settings, it does not do full justice to the subjective nature of tastes and could break down in domains where tastes are polarized. Another approach would be to use correlations with other individuals (judges or raters) as a proxy for informational quality [46, 50, 10]. This strategy has already shown some promise in settings where there is an objective correct answer—Kurvers et al. [50], for example, have shown that using the average correlation of an individual with other judges is a reliable heuristic for judge quality in several domains, and it can lead to good predictive performance when the truth cannot be immediately verified. The approach we put forward in this paper is similar in spirit and generalizes the similarity principle in matters of taste: it relies on the node strength resulting from the advice network produced from the *k-nn* algorithm to identify influential critics and talented amateurs. Using this method we were able to identify Jeff Leve as the critic with the highest recommender potential, but also to spot several amateurs who appear to have a large capacity to inform others. Our method could be used by online platforms such as Vivino to identify and nurture talent in its user base.

4.5 Homophily and polarization in matters of taste

Many real world opinion spaces are polarized. People tend to interact, listen to and get influenced by other individuals belonging to the same groups, a property commonly referred to as homophily. The tendency to interact with similar people might be further reinforced by personalization algorithms we use in our every-day life, such as the *k-nn* algorithm that we investigated in this paper. In a similar vein, it has been argued that some personalization technologies may lead people into filter-bubbles, where most of the information they consume comes from similar individuals [63]. In the case of tastes, for instance, sociologists have long argued that people belonging to different classes, cultures, or even political affiliations also differ in their aesthetic preferences [76, 64, 14, 60]. Therefore, one would expect that at least in some domains of taste collaborative filtering algorithms might produce homophily and taste filter bubbles.

We developed a method that can be used to study whether the influence networks generated by the *k-nn* algorithm produce informational insulation for different groups by adapting a well established metric of homophily and applying it to the domain of taste. When accounting for the number of ratings contributed by the two groups (count weight), the amateur tastes would be characterized by slight homophily for low values of *k*, and by slight heterophily otherwise. By contrast, critics would get most of their information from other critics regardless of how the *k-nn* algorithm is configured. The methods that we developed in this study can be readily applied to study taste homophily in collaborative filtering systems in any other taste domain and for other types of categorical groups (e.g. men and women, Europeans and Americans, etc.). For example, Dellaposta and colleagues [25] recently pointed out that people's differences in political convictions are also reflected in their tastes. Thus, using our methods one could test whether a collaborative filtering system would generate recommendations for target users by drawing information from people with similar political affiliations, potentially further increasing the cultural divide between different groups or people.

4.6 Conclusion

In vino veritas—in wine, there is truth—says an old Latin adage. We found that critics' judgements are indeed valuable in helping amateurs identify good wines, more so than the opinion of most other amateurs. Still, there is scope for combining the opinions of both critics and amateurs, and for identifying the most influential critics and talented amateurs, whose tastes appear to be informative for many other individuals. Going beyond wine, the methods we developed are modular and generic and can be readily applied in any dataset where different groups of people have rated a number of items, even when there are discrepancies in the number of items evaluated, to tackle long-lasting research questions in the social and management sciences. Further, they can be also used as a tool for improving people's understanding of key collaborative filtering algorithms [77] by enhancing the transparency and interpretability of the recommendation process [35], and can help people hone in on the right decision strategies when seeking a good item in matters of taste.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank Orestis Kopsacheilis, Sune Lehmann, and the members of the Strategic Organization Design and the Culture, Consumption and Commerce groups at the University of Southern Denmark for helpful discussions. Pantelis P. Analytis is supported by a Sapare Aude grant by the Independent Research Fund Denmark. Bahador Bahrami is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (819040 - acronym: rid-O), and by the Templeton Religion Trust. Ophelia Deroy is supported by the Momentum grant co-Sense from Volkswagen foundation, and the BIDT Co-learn grant. We thank Rhett Nichols for editing this manuscript.

Code availability statement

Our code and datasets are available at https://osf.io/pqaw3/.

Reference

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17:734–749, 2005.

[2] Theodor W Adorno. *Aesthetic theory*. A&C Black, 1997.

[3] Héla Hadj Ali, Sébastien Lecocq, and Michael Visser. The impact of gurus: Parker grades and en primeur wine prices. *The Economic Journal*, 118(529):158–173, 2008.

[4] Xavier Amatriain, Neal Lathia, Josep M Pujol, Haewoon Kwak, and Nuria Oliver. The wisdom of the few: a collaborative filtering approach based on expert opinions from the web. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 532–539, 2009.

[5] Pantelis P Analytis, Daniel Barkoczi, and Stefan M Herzog. Social learning strategies for matters of taste. *Nature Human Behaviour*, 2(6):415–424, 2018.

[6] Pantelis P Analytis, Daniel Barkoczi, Lorenz-Spreen Philipp, and Stefan M Herzog. The structure of social influence in recommender networks. In *Proceedings of the sixth ACM conference on the Web (WWW)*, pages 415–424. ACM, 2020.

[7] Pantelis P. Analytis and Philipp Hager. Collaborative filtering algorithms are prone to mainstream-taste bias. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 750–756, 2023.

[8] Robert H Ashton. Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes*, 38(3):405–414, 1986.

[9] Robert H Ashton. Reliability and consensus of experienced wine judges: Expertise within and between? *Journal of Wine Economics*, 7(1):70–87, 2012.

[10] Pavel Atanasov and Mark Himmelstein. Talent spotting in crowd prediction. In *Judgment in Predictive Analytics*, pages 135–184. Springer, 2023.

[11] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004.

[12] Suman Basuroy, Subimal Chatterjee, and S Abraham Ravid. How critical are critical reviews? the box office effects of film critics, star power, and budgets. *Journal of Marketing*, 67(4):103–117, 2003.

[13] Myron Boor. Relationships among ratings of motion pictures by viewers and six professional movie critics. *Psychological Reports*, 70(3_suppl):1011–1021, 1992.

[14] Pierre Bourdieu. Distinction: A social critique of the judgement of taste. In *food and culture*, pages 45–53. Routledge, 2012.

[15] John S Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[16] CJ Brien, P May, and O Mayo. Analysis of judge performance in wine-quality evaluations. *Journal of Food Science*, 52(5):1273–1279, 1987.

[17] Stephen B Broomell and David V Budescu. Why are experts correlated? decomposing correlations between judges. *Psychometrika*, 74(3):531–553, 2009.

[18] David V Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 2014.

[19] Sam Cameron. On the role of critics in the culture industry. *Journal of cultural economics*, 19(4):321–331, 1995.

[20] Nicolas Carayol and Matthew O Jackson. Evaluating the underlying qualities of items and raters from a series of reviews. *Available at SSRN*, 2019.

[21] Evangelia Christakopoulou and George Karypis. Local item-item models for top-n recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 67–74, 2016.

[22] Domenic V Cicchetti. Who won the 1976 blind tasting of french bordeaux and us cabernets? parametrics to the rescue. *Journal of Wine Research*, 15(3):211–220, 2004.

[23] Sergio Currarini, Matthew O Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

[24] Robyn M Dawes and Bernard Corrigan. Linear models in decision making. *Psychological Bulletin*, 81(2):95, 1974.

[25] Daniel DellaPosta, Yongren Shi, and Michael Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

[26] Hillel J Einhorn. Expert judgment: Some necessary conditions and an example. *Journal of applied psychology*, 59(5):562, 1974.

[27] Hillel J Einhorn and Robin M Hogarth. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13(2):171–192, 1975.

[28] Michael D Ekstrand, John T Riedl, and Joseph A Konstan. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2):81–173, 2011.

[29] Jehoshua Eliashberg and Steven M Shugan. Film critics: Influencers or predictors? *Journal of Marketing*, 61(2):68–78, 1997.

[30] Herbert J Gans. *Popular culture and high culture: An analysis and evaluation of taste*. Basic books, 2008.

[31] Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996.

[32] Gerd Gigerenzer and Peter M Todd. *Simple heuristics that make us smart*. Oxford University Press, 1999.

[33] Daniel G Goldstein, Randolph Preston McAfee, and Siddharth Suri. The wisdom of smaller, smarter crowds. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pages 471–488. ACM, 2014.

[34] Paul Hekkert and Piet CW Van Wieringen. Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art. *The American Journal of Psychology*, pages 389–407, 1996.

[35] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[36] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions of Information Systems*, 22(1):5–53, 2004.

[37] Elizabeth C Hirschman and Andrew Pieros. Relationships among indicators of success in broadway plays and motion pictures. *Journal of Cultural Economics*, 9(1):35–63, 1985.

[38] Robert T Hodgson. An examination of judge reliability at a major us wine competition. *Journal of Wine Economics*, 3(2):105–113, 2008.

[39] Robin M Hogarth. A note on aggregating opinions. *Organizational Behavior and Human Performance*, 21(1):40–46, 1978.

[40] Robin M Hogarth and Natalia Karelaia. Ignoring information in binary choice with continuous variables: When is less more? *Journal of Mathematical Psychology*, 49(2):115–124, 2005.

[41] Morris B Holbrook. Popular appeal versus expert judgments of motion pictures. *Journal of Consumer Research*, 26(2):144–155, 1999.

[42] Morris B Holbrook. The role of ordinary evaluations in the market for popular culture: Do consumers have "good taste"? *Marketing Letters*, 16(2):75–86, 2005.

[43] Greta Hsu, Peter W Roberts, and Anand Swaminathan. Evaluative schemas and the mediating role of critics. *Organization Science*, 23(1):83–97, 2012.

[44] David Hume. Of the standard of taste. 1757.

[45] Elihu Katz, Paul F Lazarsfeld, and Elmo Roper. *Personal influence: The part played by people in the flow of mass communications*. Routledge, 2017.

[46] Orestis Kopsacheilis, Pantelis P. Analytis, Karthikeya Kaushik, Stefan Herzog, Bahador Bahrami, and Ophelia Deroy. Crowdsourcing the assessment of wine quality-evidence from vivino. *Preprint published on SSRN*, 2023.

[47] Yehuda Koren and Joe Sill. Ordrec: an ordinal model for predicting personalized item rating distributions. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 117–124, 2011.

[48] David Krackhardt. Cognitive social structures. *Social Networks*, 9(2):109–134, 1987.

[49] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Patricia A Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113(31):8777–8782, 2016.

[50] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Mehdi Moussaid, Giuseppe Argenziano, Iris Zalaudek, Patty A Carney, and Max Wolf. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, 5(11):eaaw9011, 2019.

[51] Kevin N Laland. Social learning strategies. *Animal Learning Behavior*, 32(1):4–14, 2004.

[52] Neal Lathia, Stephen Hailes, and Licia Capra. knn cf: a temporal social network. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 227–234. ACM, 2008.

[53] Emmanuel Lazega, Lise Mounier, Tom Snijders, and Paola Tubaro. Norms, status and the dynamics of advice networks: A case study. *Social Networks*, 34(3):323–332, 2012.

[54] Jerrold Levinson. Hume's standard of taste: The real problem. *The journal of aesthetics and art criticism*, 60(3):227–238, 2002.

[55] Kevin Lewis and Jason Kaufman. The conversion of cultural tastes into social network ties. *American Journal of Sociology*, 123(6):1684–1742, 2018.

[56] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using facebook. com. *Social Networks*, 30(4):330–342, 2008.

[57] Duane E Lundy, Grace E Allred, and Branda L Peebles. How good is this song? expert versus nonexpert aesthetic appraisal. *Psychology of Aesthetics, Creativity, and the Arts*, 13(3):293, 2019.

[58] Albert E Mannes, Jack B Soll, and Richard P Larrick. The wisdom of select crowds. *Journal of Personality and Social Psychology*, 107(2):276–299, 2014.

[59] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908, 2013.

[60] J Miller McPherson and Lynn Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, pages 370–379, 1987.

[61] Johannes Müller-Trede, Shoham Choshen-Hillel, Meir Barneron, and Ilan Yaniv. The wisdom of crowds in matters of taste. *Management Science*, 64(4):1779–1803, 2017.

[62] Robert M Nosofsky. Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43(1):25–53, 1992.

[63] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[64] Minsu Park, Jaram Park, Young Min Baek, and Michael Macy. Cultural values and cross-cultural video consumption on youtube. *PLoS one*, 12(5), 2017.

[65] John W Payne, James R Bettman, and Eric J Johnson. *The adaptive decision maker*. Cambridge university press, 1993.

[66] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–186. ACM, 1994.

[67] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[68] Mark Schatzker and Richard Bazinet. Why amateur wine scores are every bit as good as professionals'. *Vox*, 2018.

[69] James Shanteau. Competence in experts: The role of task characteristics. *Organizational behavior and human decision processes*, 53(2):252–266, 1992.

[70] James Shanteau, David J Weiss, Rickey P Thomas, and Julia C Pounds. Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, 136(2):253–263, 2002.

[71] Amanda Sharkey, Balazs Kovacs, and Greta Hsu. Expert critics, rankings, and review aggregators: The changing nature of intermediation and the rise of markets with multiple intermediaries. *Academy of Management Annals*, (ja), 2022.

[72] Wesley Shrum. Critics and publics: Cultural mediation in highbrow and popular performing arts. *American Journal of Sociology*, 97(2):347–375, 1991.

[73] Özgür Şimşek. Linear decision rule as aspiration for simple decision heuristics. In *Advances in Neural Information Processing Systems*, pages 2904–2912, 2013.

[74] George M Taber. *Judgment of Paris*. Simon and Schuster, 2006.

[75] Philip E Tetlock and Dan Gardner. *Superforecasting: The art and science of prediction*. Random House, 2016.

[76] Thorstein Veblen. *The theory of the leisure class: An economic study of institutions*. Aakar Books, 2005.

[77] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 351–362, 2013.

[78] Jules J Wanderer. In defense of popular taste: Film ratings among professionals and lay audiences. *American Journal of Sociology*, 76(2):262–272, 1970.

[79] Ilan Yaniv, Shoham Choshen-Hillel, and Maxim Milyavsky. Receiving advice on matters of taste: Similarity, majority influence, and taste discrimination. *Organizational Behavior and Human Decision Processes*, 115(1):111–120, 2011.

5 Supplementary Material

5.1 Correlation profiles in sparsity balanced data

In the results we reported in the main text the number of ratings in the two groups was unbalanced because the critics have evaluated many more of the items than the amateurs. This imbalance is an attribute of the wine market in general (i.e. critics are much more prolific in ratings than amateurs). To assess whether the correlational results we observed are influenced by this imbalance we artificially reduced the ratings of critics by removing ratings from each critic at random until their data density was approximately equal to that of the average amateur (4.7%). We then repeated the correlation analysis reported in the main text by calculating the within group correlations and the correlations with the entire population. Because the results from a single random sample of this reduced dataset can be very noisy we repeated the process 1000 times and averaged the results across repetitions. The average mean taste similarity and the dispersion in taste similarity for the critics has largely remained unchanged as compared to the results presented in Figure 1. Thus, although density has a direct impact on the recommender influence of different individuals it does not affect their expected correlations or their recommender potential substantially.

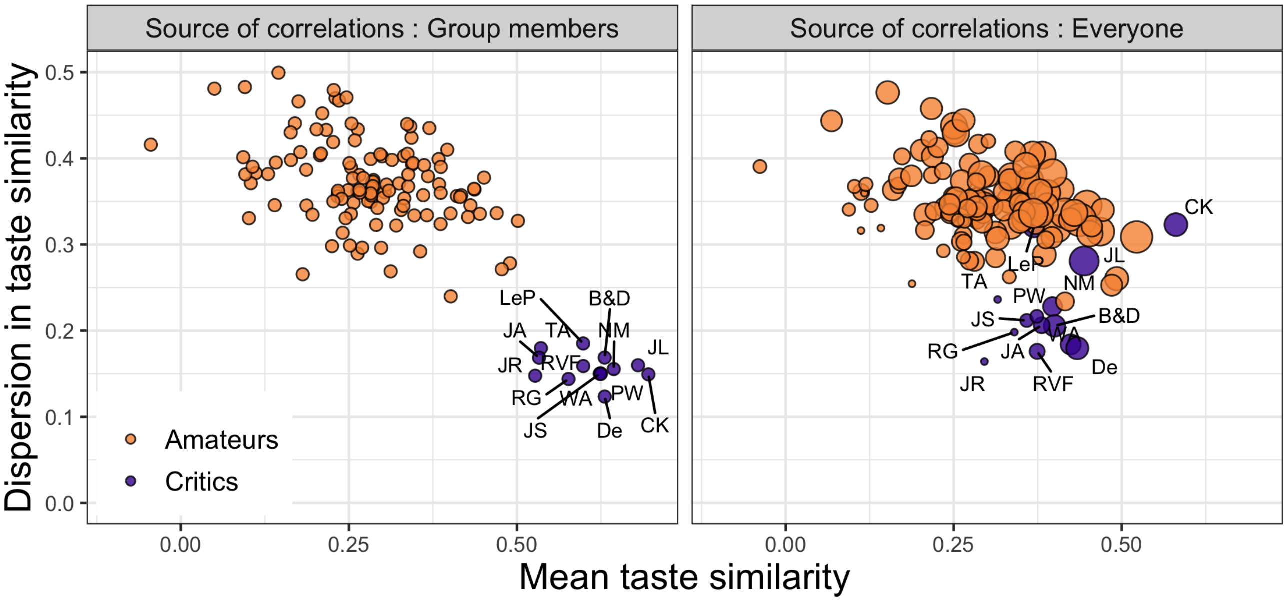


Figure 5: **Intercorrelations with individuals belonging in the same group and all other individuals when the data are balanced for sparsity. Left**: The position of 14 professional critics and 120 amateurs on a 2-dimensional plane defined by mean taste similarity (i.e., mean correlation) and dispersion in taste similarity (i.e., standard deviation of correlations) with members of the same group. **Right**: The position of the same 14 professional critics and 120 amateurs on the same plane, but this time with taste similarity calculated across all individuals (professional critics and amateurs). The color in both panels indicates whether an individual is a professional critic or an amateur and the point size in the right panel indicates the recommender potential of different individuals for $k = 5$ and $\rho = 1$.

5.2 Similarity to the amateur and critic audiences

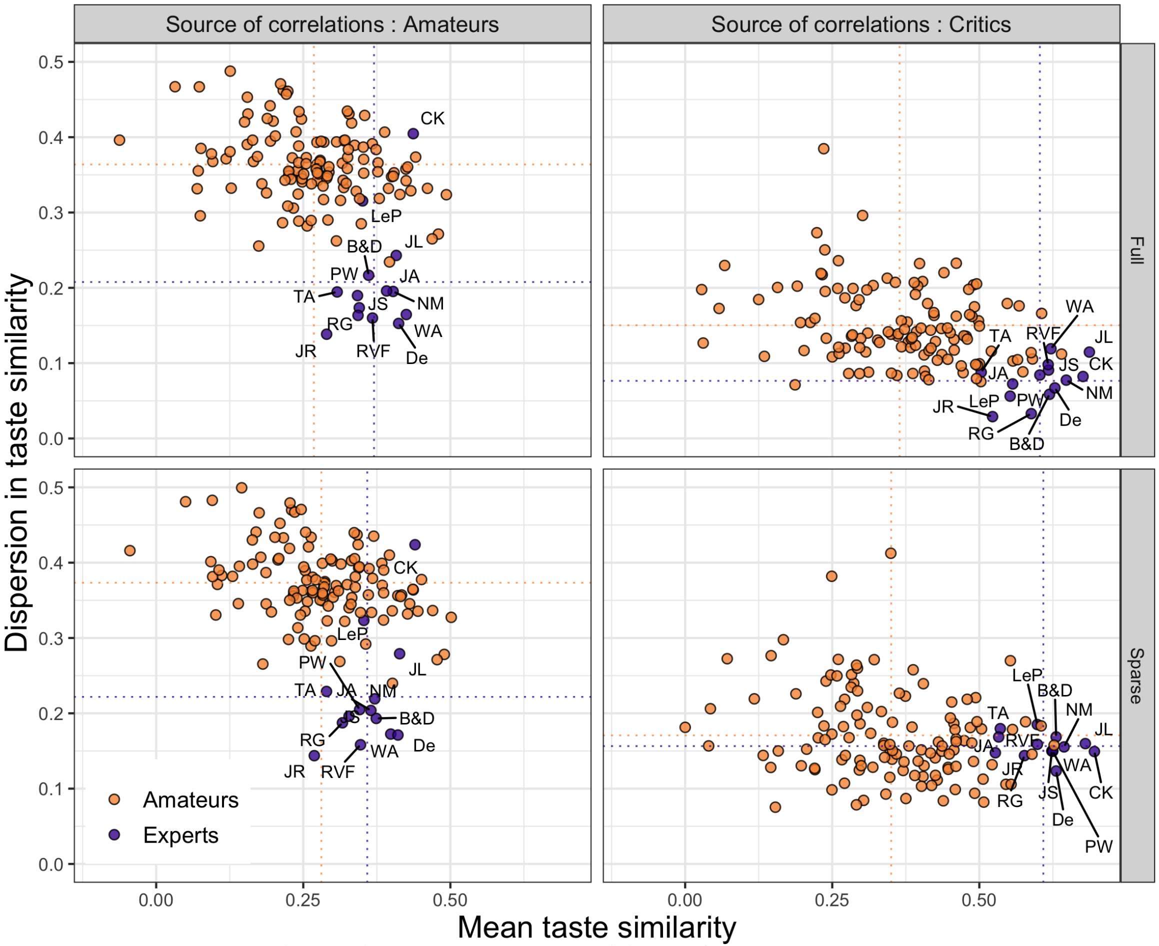


Figure 6: **Intercorrelations with amateurs and critics Left**: The position of 14 professional critics and 120 amateurs on a 2-dimensional plane defined by mean taste similarity (i.e., mean correlation) and dispersion in taste similarity (i.e., standard deviation of correlations) with amateurs. **Right**: The position of the same 14 professional critics and 120 amateurs on the same plane, but this time with taste similarity calculated across all individuals (professional critics and amateurs). The color in both panels indicates whether the an individual is a professional critic or an amateur. In the bottom row we repeat this analysis but for data balanced for sparsity. The dotted orange and purple lines indicate the average correlations and dispersion in taste similarity recorded for that group of people.

In this section we look at people's similarity exclusively with critics or amateurs, both for the full dataset and also when balancing the data for sparsity (as in the previous section). Figure 6 top left shows the correlations of the two groups with amateurs. It reveals that every single critic has correlations with amateurs higher than the average correlation among the amateurs themselves (all purple dots are on the right on the vertical dotted orange line). Figure 6 top right shows the correlations of the two groups with critics. The graph reveals that critics are more correlated among themselves than amateurs are with critics. Nonetheless, some amateurs are similar to critics in their correlation profiles, and it would have been hard to distinguish them from critics if we did not impose the critic/amateur categorization (compare with Figure 1 left, where a separation of the groups would be possible merely using their correlation profiles). These results change very little when we remove ratings from the critics until their data density becomes similar with the average amateur density (Figure 6 bottom left and right).

5.3 Strategy performance as a function of the weighting scheme

The average performance of the weighted k-nearest neighbor algorithm clearly depends of k, with larger k values leading to better performance. But how does the weighting scheme, as expressed by the parameter ρ alter the performance of the different recommendation approaches? To explore this question we show how performance varies as a function of ρ for different k values. Remember that $\rho = 0$ implies equal weights, while when $\rho = 1$ the weights directly correspond to the correlations. Overall, there appears to be little performance variation as a function of ρ (less that 2 % for our entire parameter space). For critics, larger values of ρ lead to slightly better performance, especially for high k values and irrespective of the source the data is drawn from. For amateurs, higher ρ works best when drawing advice from critics, equal weighting ($\rho = 0$) works best when drawing advice from amateurs, and intermediate ρ values work best when drawing advice from both groups (with the exception of low k values, e.g. $k = 3$, where equal weighting leads to the best results).

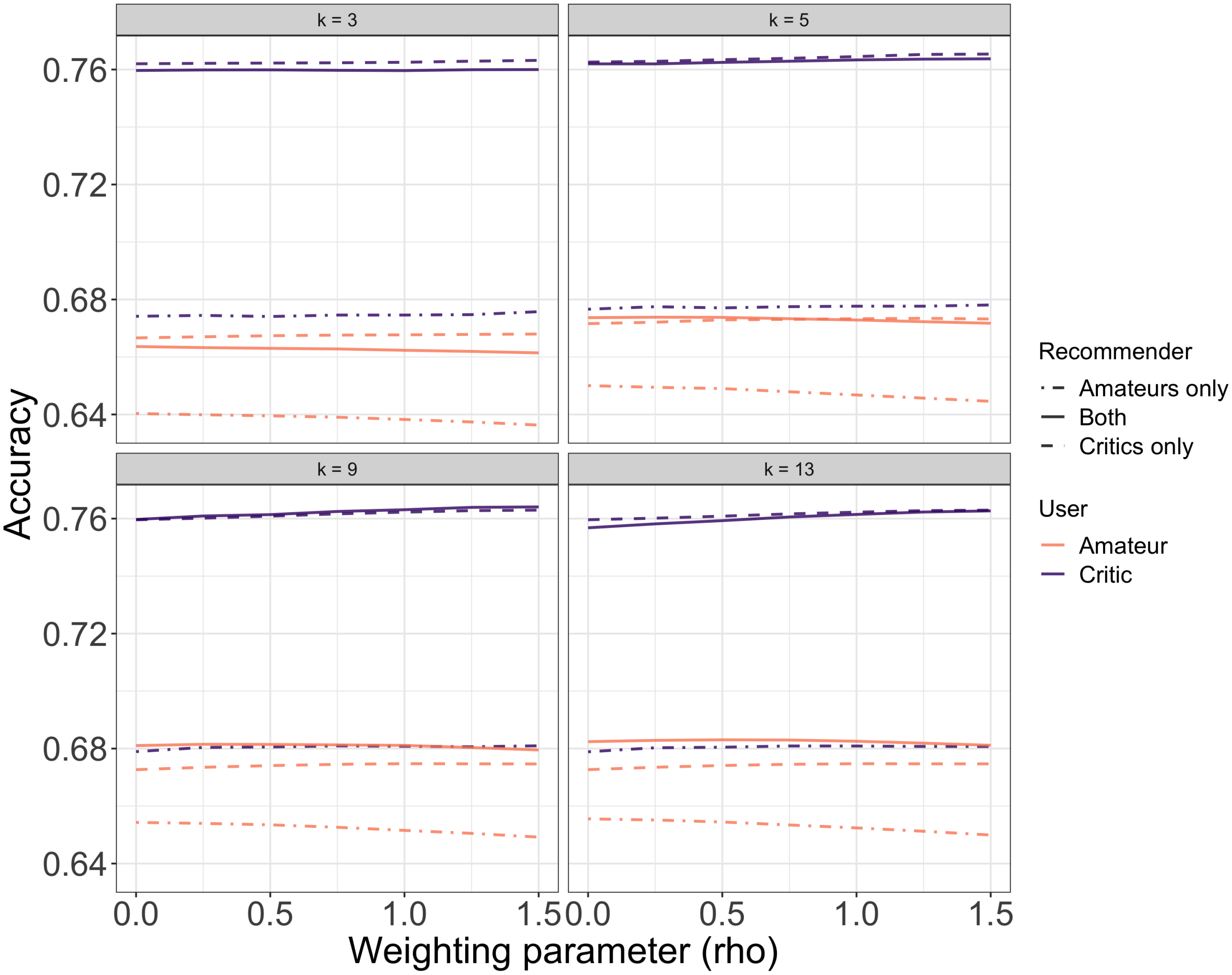
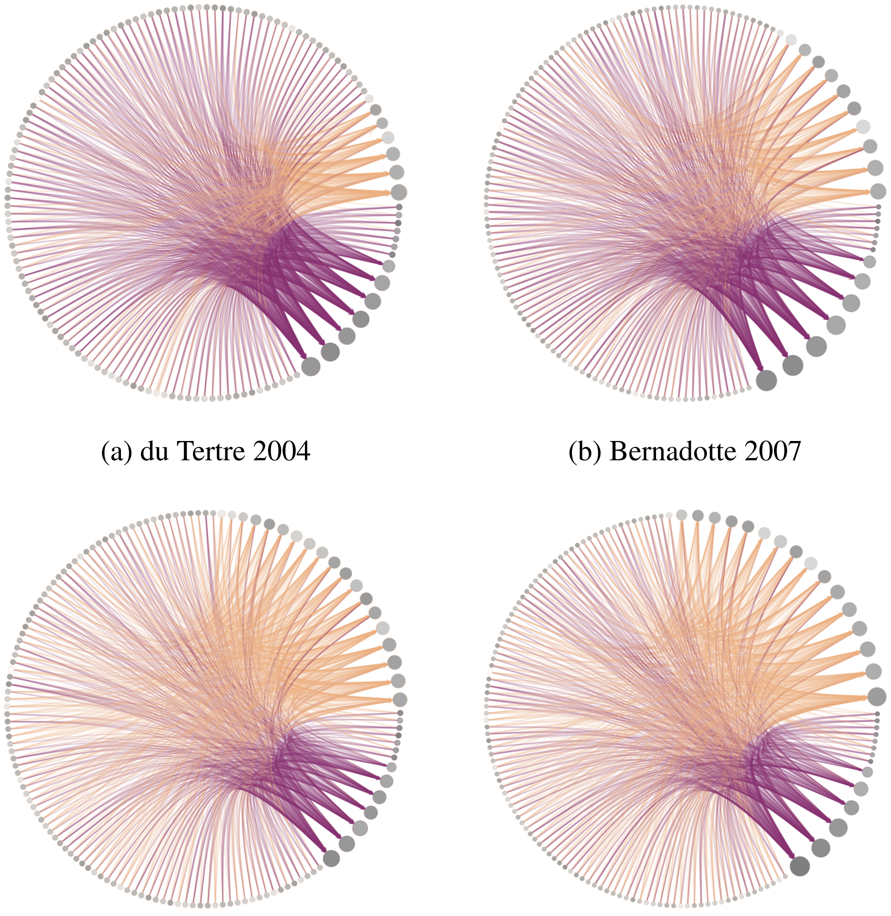


Figure 7: **Performance of the *k-nn* algorithm as a function of the parameter ρ for different *k* values Each panel**: The average performance of the *k-nn* algorithm based only on amateurs, only on critics, or both amateurs and critics for different values of k (different panels) for the amateur and critics groups, while varying the parameter ρ of the algorithm (x-axis).

5.4 Influence networks for individual wines

In this section we present examples of the influence networks for specific wines with varying popularity. For each single wine, for each repetition of the simulation, and for each target individual, the k-nearest-neighbors algorithm is searching for the *k* most similar individuals in the population who have evaluated the wine. Note that in each run of the simulation, and for each individual 10 items are withheld uniformly at random. This implies that some advisers are not available in some of the 1000 runs of the simulation and there is slight variation in the people who eventually can provide recommendations. By comparison, the influence network presented in Figure 3 in the main text is constructed by aggregating the influence networks of every single wine label included in our collection.



(a) du Tertre 2004

(b) Bernadotte 2007

(c) Pontet Canet 2008

(d) Leoville barton 2013

Figure 8: **The influence network for specific wine labels with varying popularity** Nodes represent individuals and the size of the nodes represents the recommender influence of different individuals in the recommender network spanned by the k-nn algorithm. Orange edges indicate that advice is sought (or provided) from an amateur and purple edges indicate that advice is sought (or provided) from a professional critic. The color of the nodes themselves (from light to dark grey) indicates the estimation error of the algorithm for different individuals in the entire dataset. Edges with weights smaller than 0.05 do not appear in the visualization to prevent overcrowding the graph.

5.5 Homophily index using only the initial calls

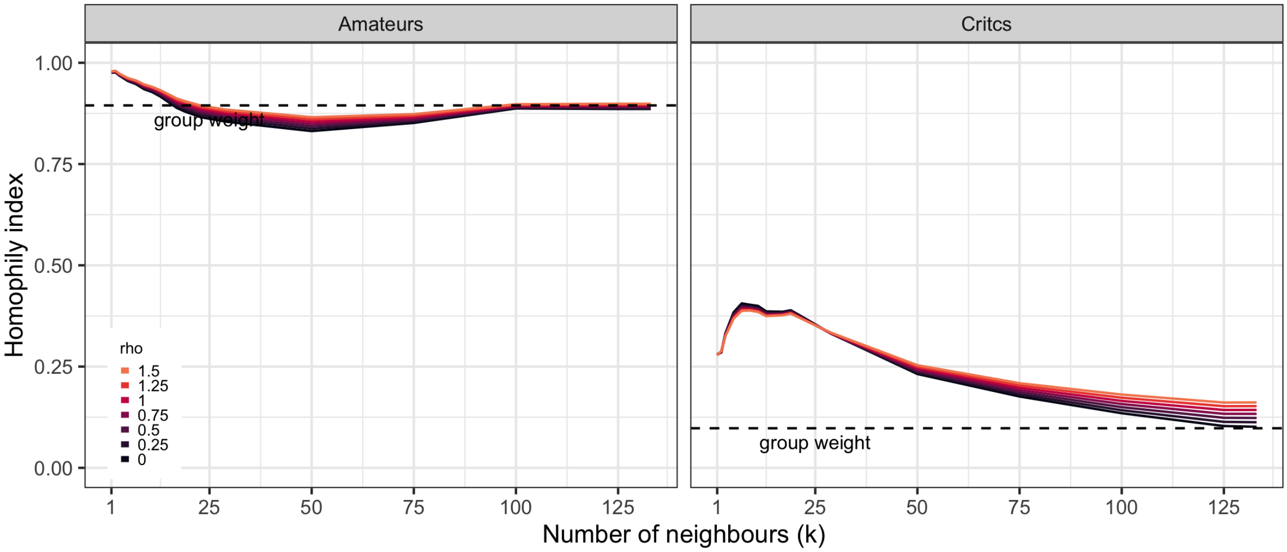


Figure 9: **Homophily index of critics and amateurs**: The homophily index of amateurs and critics as a function of k in the *k-nn* algorithm when we consider only the first calls of the algorithm. Different ρ values are represented with lines of different color. The horizontal dashed line represents a baseline corresponding to the proportion of group members in the population (group size).

The homophily index can also be seen as a measure of preference (or bias) for obtaining information from members of the same group. In the main text we presented results in terms of the ratings actually used by the *k-nn* algorithm to inform people's choices. However, when using actual ratings, the strength of the preference is not fully expressed because missing ratings from amateurs could be often substituted by the ratings of critics (who tend to be more prolific). Thus, we also visualize the homophily index when only the first k individuals called by the algorithm are included, and disregarding whether people eventually contributed ratings. This measure of homophily gives a more direct impression of the algorithm's predilection for using information from members of the same group. This measure replicates the gist of the results presented in the main text. The amateurs are characterized by inbreeding homophily for low to intermediate values of k (<17) and become slightly heterophilous for values above that. The critics, by contrast, are characterized by substantial in-breeding homophily for low values of k. As k increases, and more people are sought for advice, the homophily index converges to the group weight (for $\rho = 0$ it converges exactly to the group weight whereas for higher ρ values there might be small deviations due to the weights assigned to different individuals, see the right panel of Figure 9.)

5.6 From aggregate to individual level performance

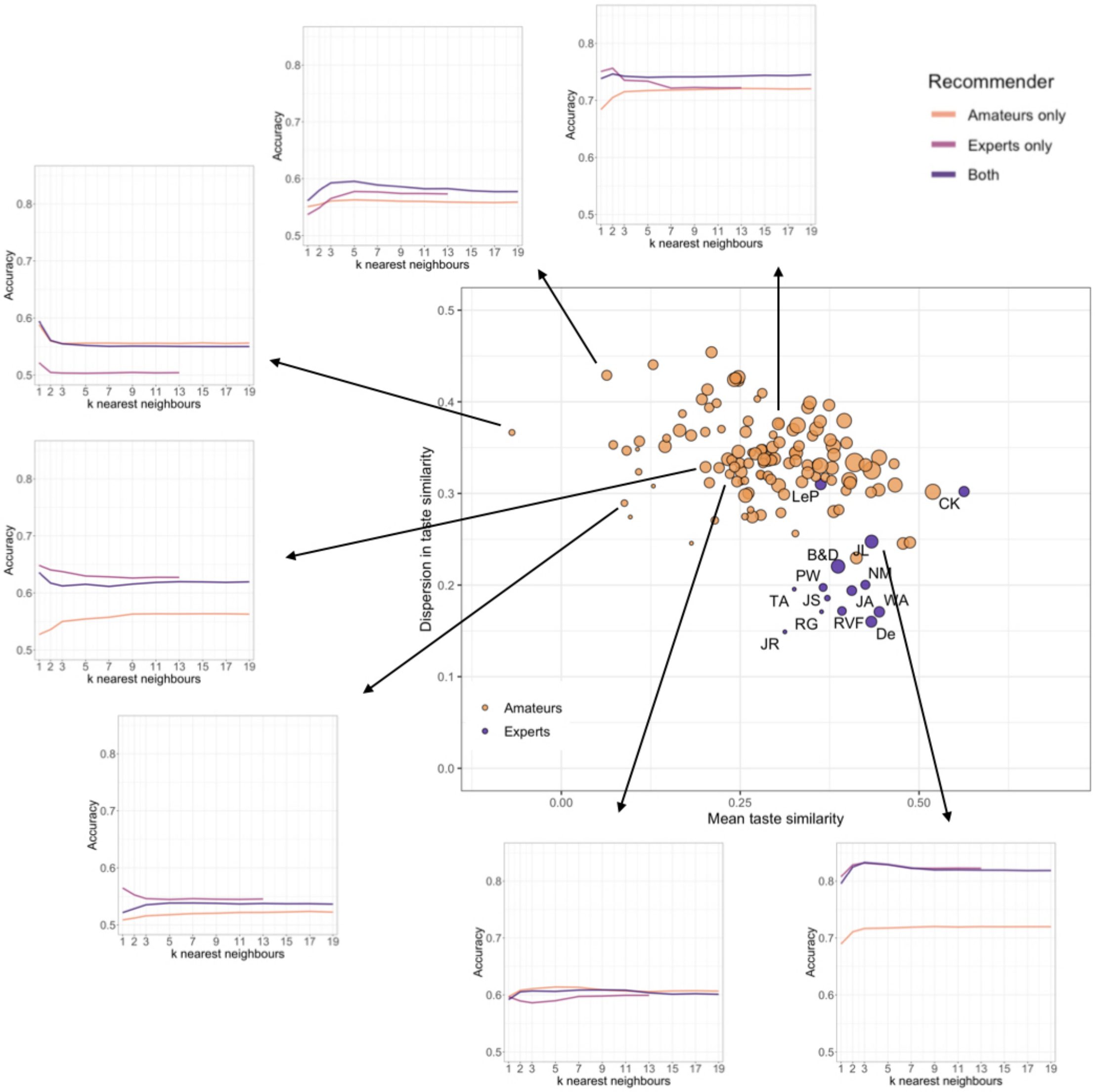


Figure 10: **The position of the 14 professional critics and 120 amateurs on the 2-dimensional plane defined by mean taste similarity and dispersion in taste similarity, but this time with taste similarity calculated. The color in both panels indicates whether an individual is a professional critic or an amateur and the point size indicates recommender potential.**

So far we have presented how the performance of different algorithms varies as a function of *k* and ρ when we average across individuals. This analysis can be also be repeated at the individual level to provide more nuanced results on how the performance of the *k-nn* algorithm (or different social learning strategies) changes for different individuals and as a function of the number of neighbors *k* for different values ρ. In Figure 10 we demonstrate the power of such individual level analysis for ρ = 1. We present three individuals for whom the heuristic *follow-the-most-similar-critic* was the best or a nearly the best strategy to follow (middle and bottom left and top right), one individual for whom following the most similar amateur performed best (top left), one individual for whom the best solution was following a clique of amateurs (bottom center), and one for whom the opinions of critics and amateurs where clearly complementary (top center). Last, we present the performance of the algorithm for Jeff Leve, the critic with the highest recommender potential (bottom right).

5.7 Correlation profiles of critics

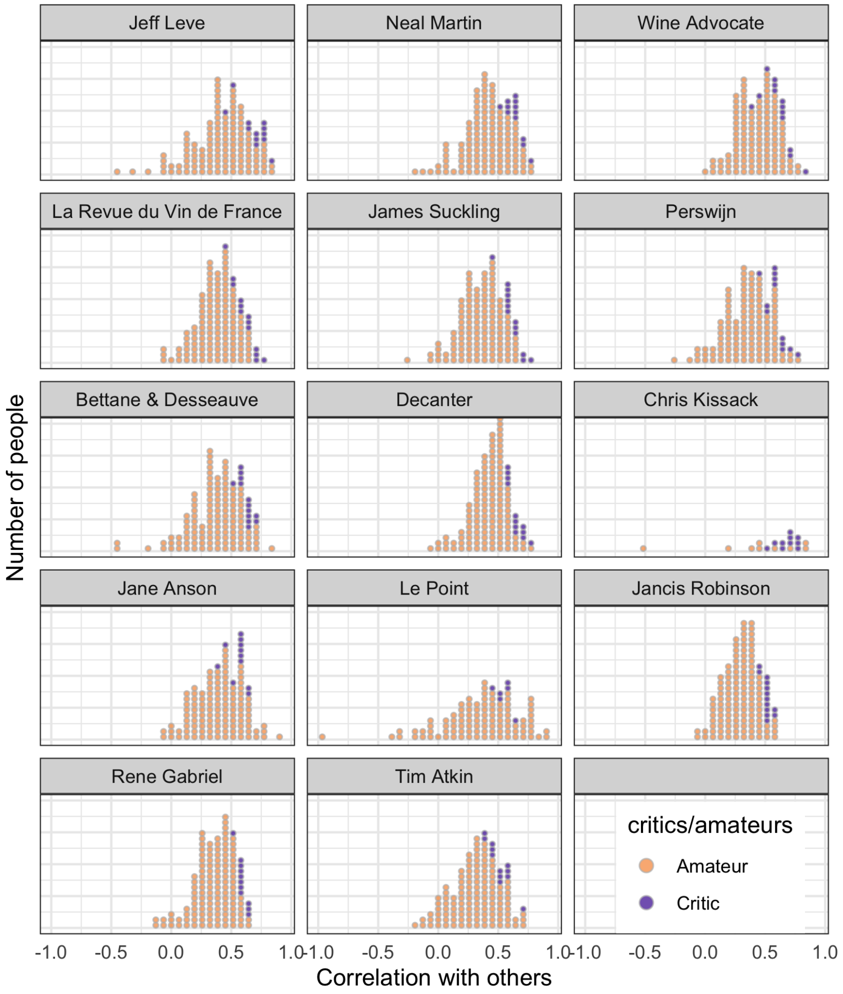


Figure 11: **The observed correlations of the 14 critics with other critics and amateurs. Orange beads correspond to correlations with amateurs and purple ones to their correlations with critics. We do not report correlations where the critic had less than 5 items in common with other individuals. The critics have been ranked in descending order, according to the prediction rate of the recommender system for them. Initials of professional critics: Wine Advocate — Lisa Perotti Brown, Decanter — Steven Spurrier, James Lawther, Beverley Blanning, and Jane Anson, Revue du Vin de France — Olivier Poels, Hélène Durange, and Philippe Maurange, Le Point — Jacques Dupont, PersWijn — Ronald DeGroot**

Our analysis revealed that there are substantial differences in the correlation profiles of critics and amateurs, but also in the influence potential of different critics. To have better insight into how these differences are produced, we looked at the correlation profiles of different critics with other critics and amateurs. It can be easily observed that the correlations of critics with other critics (purple beads) are much higher than correlations with amateur raters (orange beads). In fact, for most critics (possibly with the exception of Jacques Dupont writing for Le Point and Jane Anson) other critics are among the most correlated other individuals (the purple beads are on the far right of the correlation distribution).

5.8 Average recommender potential and influence of critics and amateurs

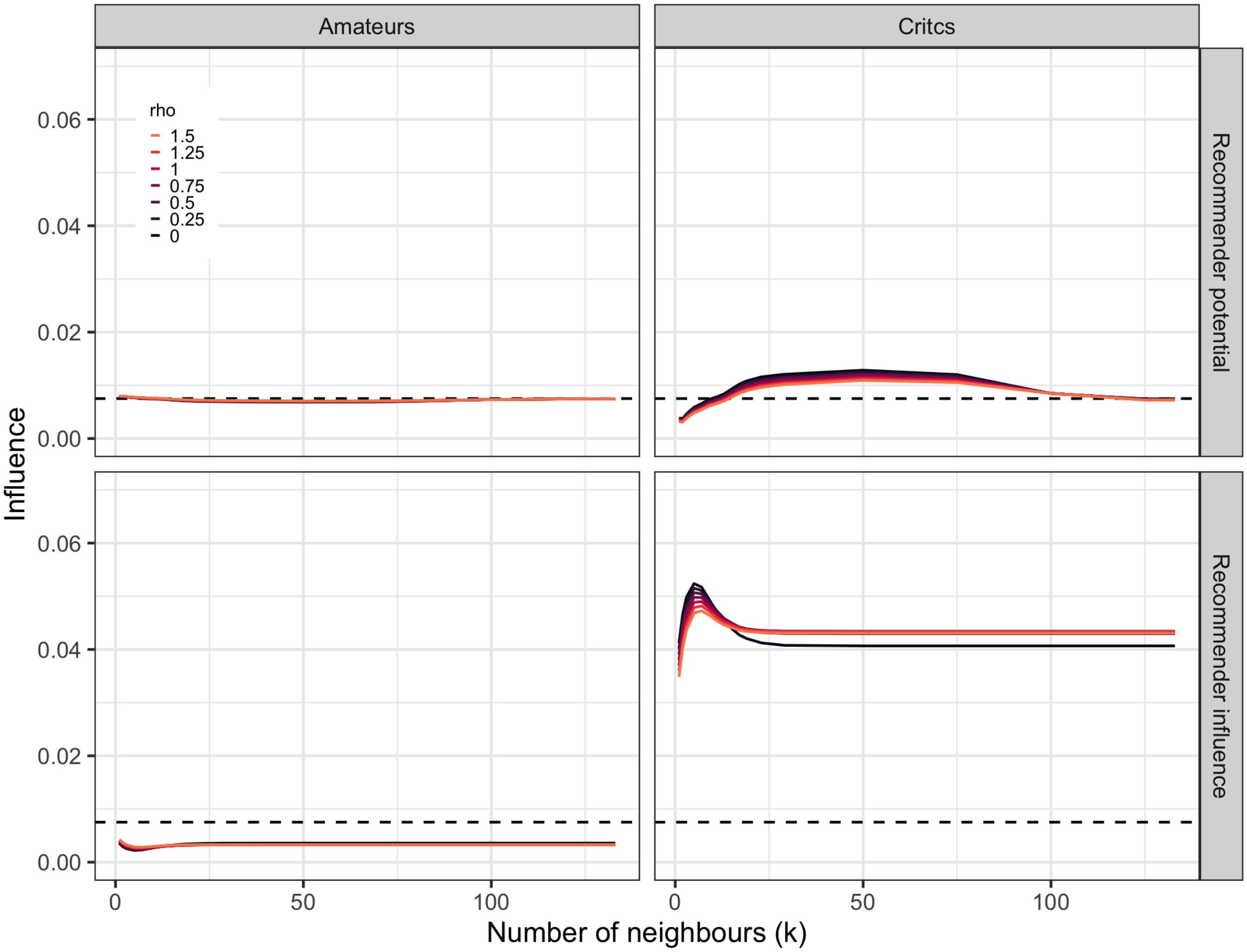


Figure 12: **The average recommender potential and recommender influence of amateurs and critics as a function of the parameters *k* and ρ. The horizontal dashed line represents the average recommender potential in the case where people consider the opinions of all other individuals using an equal weights strategy.**

For small values of *k* amateurs have a larger recommender potential than critics. This happens because the correlations among amateurs tend to be more dispersed and the people with the highest correlations with amateurs tend to be other amateurs. This picture changes as *k* increases because the opinions of critics are consistently sought by the recommender system for larger *k* values. In practice, the critics are substantially more influential than amateurs. This is because they are much more prolific raters, and they can more often supply their advice when the recommender system seeks it. Note that there are relatively small changes in the two groups as we vary the number of neighbors *k* and the similarity sensitivity parameter ρ.