Network structure in matters of taste: can social networks predict cuisine taste?

Matt Lamm
005952310

Imanol Arrieta Ibarra
05944271

Camelia Simoiu
066001410

Abstract

This paper explores the problem of inferring user interests from social network structure. Research in network-based social behavior has demonstrated that people tend to form connections with people having similar interests to their own. Known as homophily, or the "birds of a feather" phenomenon, this tendency significantly informs the structure of certain social networks. We use network analysis techniques to explore the correlation between restaurant reviewer tastes and the tastes of their friends on the Yelp! social network. Then, we present a variety of algorithms for predicting a reviewer?s tastes based on the structure of the reviewer?s ego network. These include an array of statistically learned classifiers, as well as a modified version of the deterministic autocorrelation model of Wen & Lin (2010).

Keywords — Homophily, Networks, Cliques, Community Structure

1 Introduction

People tend to form connections with people of similar interests to their own. Known as homophily, or the "birds of a feather" phenomenon, this tendency significantly informs the structure of social networks Mcpherson *et al.* (2014). For example, Singla & Richardson (2008) find that people who communicate frequently in an instant message network are more likely to share common interests. Based on these results, one anticipates that friends in a social network, i.e. members of a dyad, will be similar with significantly higher probability than a random pair of users. Within a person's ego network, there are additional structures that may influence a user's tastes. A user is more likely to be influenced by a strongly-connected community than a weakly connected one. A user is more likely to be influenced by an extremely popular user than an unpopular one. Similarly, a user is likely to be influenced by someone with many mutual friends.

We explore the relationship between user similarity and network structure in the context of a restaurant reviewer network from Yelp!. In particular, we employ the above-mentioned metrics of as features for inferring a reviewer's cuisine taste from the tastes of the reviewer's friends. We develop a modified version of the deterministic, autocorrelative model of Wen & Lin (2010), as well as several linear classification and regression techniques for inferring a user's cuisine taste and favorite cuisine type. Following Wen & Lin (2010) we consider multiple representations of a reviewer‘s taste given their reviews.

In Section 2 we discuss previous work in the domain of inferring user interests from network context. In Section 3, we present an exploratory analysis of the data. In Section 4, we present multiple algorithms for inferring a user‘s tastes from the structure and tastes of the user‘s ego network. Results are presented in Section 5.

2 Previous Work

We focus on two papers from recent work at the intersections of online user classification, social network analysis, and theories of homophily. Singla & Richardson (2008) demonstrate that individuals who speak with each other over an online IM network are more likely to be similar to each other than a random pair of users. Wen & Lin (2010) invoke insights from that paper, among others, to build a model which predicts user interests based on the interests of people they are connected with in a social network.

Singla & Richardson (2008), find that users who communicate over messenger are likely to have similar query attributes. This similarity is almost zero for the case of random pairs. Between user similarity is directly related to the amount of time users spend talking over IM. Similarity is also found to exist on an inferred network (i.e. friends of friends), though with reduced intensity. We follow Singla & Richardson (2008) and test the correlation between clique structure and homophily in the Yelp! network. With Yelp! data, we do not have information about user interaction but believe that such interactions influence the structure of the network.

Wen & Lin (2010) seek to predict the interests of users in a social network based on the interests of others in the network. They use a deterministic algorithm which leverages information from users‘ friends and the friends of their friends, to predict how much a user may be interested in a given topic. The algorithm of Wen & Lin (2010) does not incorporate any additional information regarding the structure of the network. Building upon the results of Wen & Lin (2010) and Singla & Richardson (2008), we anticipate that an autocorrelative model which incorporates more information about community structure will lead to better inferences in the context of the Yelp! social network.

3 Data and Exploratory Analysis

We will be focusing on data from the Yelp! Dataset Challenge. This is a JSON formatted data set that Yelp! made publicly available in 2013 with the purpose of promoting academic research. This set contains data from Phoenix, Las Vegas, Madison, Waterloo and Edinburgh. The analysis that follows was partially computed on the entire dataset, and partially on the Phoenix data. Comparative statistics of the entire Yelp! graph and that for Phoenix are provided in Table 1. Here edges represent an explicit friendship between two users in the network displayed publicly on Yelp! Most notably, we observe that overall, Phoenix has a slightly smaller diameter, higher clustering coefficient and a higher number of closed triangles, indicating that its users may be slightly more connected than those on the entire graph. Nevertheless, both graphs are relatively sparse with a small percentage of users being strongly connected, as can be further deduced from statistics regarding the size and number of weakly connected components (Table 2). We can see that both graphs are dominated by a singular large component, amounting to approximately over 47% and 42% of users for Yelp! and Phoenix respectively.

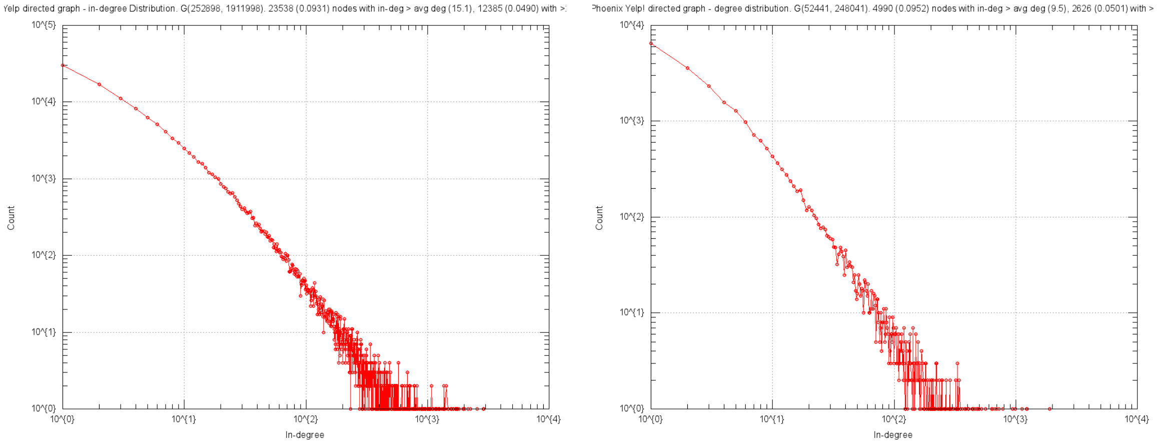
Metric	All cities	Phoenix
Number of Users	252, 898	52, 441
Undirected Edges	956, 020	124, 022
Zero Deg Nodes	129, 529	29, 389
NonZero In-Out Deg Nodes	123, 368	23, 052
Closed triangles	4, 399, 191	448, 435
Open triangles	202, 406, 349	14, 542, 578
Frac. of closed triads	0.021272	0.029914
Connected component size	0.473863	0.422418
Approx. full diameter	11	10
90% effective diameter	4.978374	4.859450
Clustering coefficient	0.063240	0.065641

Table 1: Social Network features for complete Yelp! dataset and Phoenix.

The degree distribution of the Yelp! and Phoenix networks is shown in Figure 1(a) and 1(b). While both figures show a relatively straight line at a 45 degree angle with heavy tails for high-degree nodes, we observe a slight curvature for low degree nodes, which indicates that the distribution may deviate slightly from a power law.

Size WCC	Yelp! dataset	Phoenix
$119,839$	1	0
$22,152$	0	1
9	1	0
6	3	0
5	2	1
4	21	5
3	159	45
2	$1,466$	370
1	$129,529$	$29,389$

Table 2: Size of Weakly Connected Component and Number of such components



(a)

(b)

Figure 1: Degree distribution of Yelp Users for complete dataset (all cities) 1(a) and Phoenix 1(b).

3.1 Modelling Yelp Users' Interests

We define two ground truth models of user interests for Yelp! reviewer data, focusing only on cuisine taste. *Explicit cusine taste* are treated as the reported category of reviewed restaurants, where cuisine types correspond to categories such as American, Italian, Thai and Mexican.

Reported categories in Yelp! are not always very different. For example, Yelp! has categories for Japanese restaurants and Sushi bars, despite that these are underlyingly the same cuisine type. In addition, a restaurant classified as 'American' might be reviewed most frequently for its canolis, an Italian dessert. We learn a user's *implicit cuisine taste* using a Latent Dirichlet Allocation (LDA) topic model trained on reviews. Since implicit tastes are inferred from review text, they avoid such misclassifications that may exist in reported restaurant categories.

LDA is an unsupervised algorithm which models a collection of documents as the result of a generative process Blei *et al.* (2003). Documents are represented as bags of words over fixed vocabulary V, where V is a domain-specific vocabulary set inferred from the corpus. Each document in the corpus is generated by first choosing a mixture proportion over K-topics $\theta \sim Dir(\alpha)$. Then, for each word in the document, a topic is chosen from $Z \sim Mult(\theta)$, where the topic itself is a unique distribution over all words in V. In classification of a new document an LDA model will return the maximally probable mixture of topics, a K-dimensional vector whose entries sum to one.

We've trained LDA by assuming each document corresponds to a single user, and is thus a concatenation of all of a given user's reviews. For each type of document we trained LDA on the data for $K = 5, 6, \ldots, 30$ topics. For both methods, values of K around 13 and 16 seem ideal, based on our intuitions about the relatedness of the top words in each topic.

Appendix A shows the top words from the different topics trained using LDA. It can be seen that LDA yields several coherent topics corresponding to cuisine types. It also learns multiple topics which correspond to subjects such as dining experience, service and atmosphere. These categories are still very much in line with the purpose of our research since they still correspond to consumer's interests.

4 Methodology

4.1 Exploring the Correlation of Cuisine Taste based on Homophily

Following the study of Singla & Richardson (2008) we investigate the correlation between explicit cuisine taste and friendship network structure by calculating the Jaccard index:

$$J = \frac{|A \bigcap B|}{|A \bigcup B|}$$

Where, given a pair of users i and j, A and B are their respective *explicit* cuisine tastes.
We consider the metric in three different contexts:

1. Baseline: A random sample of nodes.
2. Dyads: Nodes which are joined by one edge.
3. Triads: Sets of three nodes joined three non-self edges.

Based on the results of Singla & Richardson (2008) we hypothesized that computing the Jaccard index would be strongest for triads and lowest for the baseline, i.e. $Triads \geq Diads \geq Baseline$.

4.2 Autocorrelation Model

An autocorrelation model predicts a customer's cuisine taste as the weighted sum of the cuisine preferences of her network. As in Wen & Lin (2010) weights are defined to be inversely proportional to the distance between users and proportional to the strength of their relationship.

More specifically, with **U** as the number of users in the network, and **N** as the length of the cuisine preference vectors, we define the matrix $\mathbf{Z} \in \mathbb{R}^{UxN}$, where row $\mathbf{Z}_i^T$ represents user i's interest vector.

Thus, Wen & Lin (2010) propose the following autocorrelation model:

$$\mathbf{Z} = \rho \mathbf{W} \cdot \mathbf{Z} + \epsilon$$

or:

$$z_{ij} = \sum_{k=i}^{U} (w_{ki} \cdot z_{kj}).$$

where $z_{i,j}$ is the normalized amount user i is interested in cuisine j. $w_{k,i}$ is a weight representing the influence user k has on user i defined in this manner:

$$w_{ki} = exp(-dist(k,i)).$$

where $dist(k,i)$ is the distance between user k and user i defined as:

$$dist(i,j) = \sum_{k=1}^{K-1} \frac{1}{strength(v_k, v_{k+1})}$$

where $v_1, ..., v_k$ are nodes on the shortest path from user i to user j and $strength(v_k, v_{k+1})$ reports how strong the relationship is between nodes v_k and v_{k+1}. For Wen & Lin (2010) the strength between nodes depends on the amount of communication between those nodes. This weight calculation does not apply in our case as we do not have any measure of frequency of communication between friends in Yelp!s social network. Nevertheless we can assume that given that two users are friends in the network, the strength of their tie is maximal, in which case the distance is defined to be the shortest path between them.

4.3 Inferring User Tastes using Community Information

In the context of Yelp!, we anticipate that more information about a user's cuisine taste can be extracted from users with which she has more friends in common. In this context we can define distance in terms of the following strength metric:

$$strength(i, j) = \frac{|A \bigcap B|}{|A|}$$

Where A is the set of neighbors of node i and B is the set of neighbors of node j. This metric is more appropriate than Jaccard Index, because it gives us a way of quantifying j's influence on i (which might be differ from i's influence on j). After testing this metric on a subsample of the data, we found that it overpenalized users with too few friends in common. Thus, we ultimately define the distance between nodes using following strength metric:

$$strength(i, j) = \frac{\log(|A \bigcap B| + 1)}{\log(|A| + 1)}$$

We use cross-validation to compare the Community Information with the Auto correlation Model specified in the previous section. We divide the users into ten sets and for each group we infer their cuisine taste using the cuisine tastes from the other nine groups. Then, for each group and for each user we compute the precision and recall of our inference defined as:

$$Q_p = \frac{|INF \cap GND|}{|INF|} \qquad Q_r = \frac{|INF \cap GND|}{|GND|}$$

where INF is the set of inferred cuisine tastes with positive weights and GND is the ground truth cuisine tastes with positive weights. We then average over and across cross-validation groups to have an estimate of precision and recall. We perform equivalent tests for implicit cuisine tastes.

4.4 Logistic and Softmax Regression

We explore two statistical methods for testing our hypothesis that social ties can be used to predict the cuisine tastes of users. We implement logistic regression to predict the implicit categories in which a user is interested. This provides a comparison against the autocorrelative model, which is deterministic. We use implicit cuisine tastes, i.e. those calculated using LDA transformations, because they were found to be less sparse than explicit cuisine tastes. This increases the likelihood that users will be interested in some common implicit category. The features computed employed are as follows:

- `friends`: the average implicit tastes all of the user's neighbors
- `friends_in_triads`: the average implicit tastes of the user's neighbors that are part of a triad.
- `most_popular_friend`: the implicit taste of the neighbor with the highest degree centrality
- `most_mutual_friends`: the implicit taste of the neighbor having the most friends in common with the user

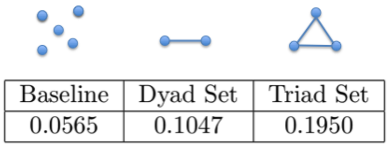
For each of these feature vectors there are fifteen components, corresponding to the number of topics used in training LDA. These features aim to capture the effect of the social influence of one's peers. For instance, a user may likely be influenced by their friends' recommendations and preferences. Similarly, users that form a triad may indicate a tight community of mutual friends. The last two features aim to detect the most influential people in a user's set of friends.

Next, we define a user's preferred category as being that for which they have submitted the highest number of reviews.Our goal is to correctly predict the user's preferred category using the same feature set as above, but for the top fifteen *explicitly reviewed* restaurant categories, we implement Softmax regression with L1 regularization to do so.

5 Results

5.1 Correlation Exploration

Contrary to Singla & Richardson (2008) we did not find a *very* strong correlation between a users' cuisine taste and the structure of their ego network. Nevertheless, our hypothesis that user connectivity is a better indicator of cuisine taste was correct. The following table shows the Jaccard index for different user relationship structures (i.e. dyads and triads). The baseline randomly sampled 200 nodes from the graph and calculated the percentage of pairs that had reviewed at least one common category. The dyad set is defined as the set of pairs of users that are connected by an edge in the network. The triads are those users belong to a fully connected set of three users. For all sets, we only consider users that have degree greater than zero. We observe almost a 100% increase in the percentage of users having reviewed a common category from the baseline to the dyads, and a 300% increase from the baseline to triads.



Baseline	Dyad Set	Triad Set
0.0565	0.1047	0.1950

Table 3: Results for Naive Bayes Model

5.2 Autocorrelation Model vs. Community Information

For the case of inferring interests using community information, we calculated the quality of inference and found that results do not differ much between our distance metric and that of Wen & Lin (2010).

Table 3 shows precision, recall and the sum of squared errors for the models inspired by Wen & Lin (2010) and the baseline random model. These results are obtained when running these algorithms on the Explicit Yelp!'s Cuisine types. Table 4 shows the same metrics when running the algorithms on the LDA categories.

Model	SS	Precision	Recall
Autocorrelation Model	0.38	0.23	0.64
Community Information	0.42	0.24	0.60
Random Assignment	2.42	0.06	0.48

Table 3: Results for Explicit Yelp!'s Cuisine Tastes.

We find that both the Autocorrelation Model and the Community Information one perform significantly better than a Random assignment. Nevertheless, the change in the strength metric does not appear to improve our results significantly. Precision does increase by a small amount (4%) for Community Information but not enough to compensate for the loss of recall.

Table 4 shows that Community Information and Autocorrelation perform marginally better when dealing with LDA categories. However, for LDA the Community Information model appears to do much better

(10.5%) than the Autocorrelation Model which agrees with our intuition that users with more friends in common are likely to be more similar.

Model	SS	Precision	Recall
AC	0.41	0.21	0.70
CI	0.20	0.28	0.69
Random Assignment	2.47	0.07	0.49

Table 4: Results for LDA Implicit Cuisine Tastes

5.3 Logistic Regression

With regards to logistic regression we found that this method performs much better than Wen & Lin (2010). It is able to capture whether a user is interested on a subject or not with higher precision and recall. This is partially due because the weights for this model are specifically trained for this task.

Another reason why this method performs so much better is that it incorporates information from all categories when making a prediction. In Wen & Lin (2010), the only way we can predict if a user likes Japanese food is if her friends like Japanese Food. In Logistic Regression we allow for the case when friends like Chinese food to be predictive of the user's interest in Japanese food.

Model	Precision	Recall
Logistic Regression	0.5	0.88

Table 5: Precision and Recall metric for Logistic Regression

5.4 Multinomial Softmax Regression

We compute results on predicting a user's favorite explicit and implicit restaurant categories. In the case of explicit cusine tastes, Softmax regression trained on the aforementioned features drastically outperfoms the baseline of simply assigning the most frequently reviewed category in the network. More than half of the time we are able to predict a user's preferred explicit cuisine type using only information from his ego network. This has enormous potential for influencing recommendation and targeted marketing systems for Yelp!. In the case of LDA-inferred implicit cuisine tastes, predictive success was very low. This is likely because LDA inferred topic proportions for a given user are often too homogeneous give a clear indication of what a user's preferred cuisine category is.

Data	Softmax	Baseline
Explicit(Yelp)	62.8%	21.8%
Implicit(LDA)	17.9%	17.5%

Table 6: Multinomial Softmax Regression

6 Conclusions and Future Work

Based on our findings, we believe there remain several paths to explore regarding strength metrics for autocorrelative models. We are now looking at ways to improve the precision of these models without affecting recall too much. Also, we are looking at new distance metrics that may allow us to further reduce the sum of squared errors.

With regards to the Multinomial Softmax Regression it may be fruitful to augment the set of features considered to improve on the accuracy we currently obtain. Still, our current performance might already be of tremendous significance for real-world application.

We sought to discover the ways in which network structure relates to restaurant reviewer similarity in the Yelp! dataset. Our hypothesis that network features can successfully be employed for inferring user interests was correct.

References

Blei, David M., Ng, Andrew Y., & Jordan, Michael I. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**(Mar.), 993–1022.

Esslimani, Ilham, Brun, Armelle, & Boyer, Anne. 2009. From social networks to behavioral networks in recommender systems. *Pages 143–148 of: Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009.*

Guy, Ido, Avraham, Uri, Carmel, David, Ur, Sigalit, Jacovi, Michal, & Ronen, Inbal. 2013. Mining expertise and interests from social media. *Proceedings of the 22nd international conference on World Wide Web*, 515–526.

Mcpherson, Miller, Smith-lovin, Lynn, & Cook, James M. 2014. BIRDS OF A FEATHER : Homophily in Social Networks. **27**(2001), 415–444.

Singla, Parag, & Richardson, Matthew. 2008. Yes, there is a correlation:-from social networks to personal behavior on the web. *. . . of the 17th international conference on World . . .*, 655–664.

Wen, Zhen, & Lin, Ching-Yung. 2010. On the quality of inferring interests from social neighbors. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, 373.

7 Appendix A: LDA Top 15 Words

environment	japanese	diner	thai	italian	chinese	service	american	environment	bar	review	italian	mexican	sportsbar
peopl	sushi	breakfast	dish	locat	chines	ask	burger	staff	wine	star	dessert	mexican	beer
way	roll	egg	thai	crust	soup	minut	hot	friendli	hour	better	bread	taco	wing
see	fish	coffe	flavor	year	rice	said	dog	famili	happi	bad	steak	salsa	game
review	fresh	bagel	spici	italian	dish	server	onion	year	select	review	pasta	chip	night
feel	happi	cream	shrimp	new	beef	manag	meat	dinner	patio	noth	dinner	burrito	hour
walk	hour	morn	rice	slice	noodl	took	bbq	atmospher	=	give	dish	bean	happi
now	tuna	friendli	fish	favorit	bowl	waitress	side	experi	enjoy	seem	wine	tortilla	sport
take	japanes	fresh	hot	fresh	egg	told	beef	enjoy	night	ok	appet	enchilada	watch
right	salmon	toast	curri	now	roll	take	grill	recommend	favorit	qualiti	italian	green	tv
first	chef	potato	sweet	top	hot	custom	potato	visit	bruschetta	mayb	entre	margarita	play
day	favorit	bacon	bean	wing	pork	anoth	bread	amaz	seat	2	chocol	rice	fun
someth	special	pancak	green	sub	buffet	busi	rib	excel	atmospher	use	special	chile	bartend
park	spici	ice	soup	friendli	veggi	call	gyro	dine	friendli	expect	potato	beef	big
star	qualiti	day	spice	famili	pho	seat	pita	friend	park	bland	serv	fresh	atmospher

Table 7: Top 15 words for LDA trained on individual reviews