Modeling Shared Tastes in Online Communities

Laura Dietz
Max-Planck Institute for Computer Science
Saarbrücken, Germany
dietz@mpi-inf.mpg.de

Abstract

Many social networking platforms track user-item interaction and friend relationships between users. The goal of this work is to learn about tastes that are shared by friends. We devise a probabilistic model in which tastes explain both friend relationships and item interactions as latent factors. After estimation, shared tastes can be used to predict common preferences of befriended users, predict further item interactions, and give an overview about the network of users that share the same taste.

1 Introduction

Many online community platforms such as Zune Social, last.fm, library thing, flicker, and CiteULike store data about users, friend relationships between users, and community artifacts associated with each. Artifacts may be songs, books, pictures, or scientific publications, respectively. It is likely that people in online communities gather in groups of shared interests, tastes, or expertises, where such interests drive friendships and friendships drive interests. Modeling this interdependency may give insights into tastes that are shared within the social network. Such shared tastes can be exploited to support users in a community platform, two examples being taste-based subscription and visualization of friend's interests, which we explain using Zune Social as a running example. Zune Social is the online community to go with Microsoft's portable music player Zune.

Zune Social allows users to subscribe to playlists of their friends. This functionality may be unsatisfactory if the friend listens to diverse kinds of music, of which only few aspects are shared with the subscribing user. Re-weighting the friend's playlist to match the taste shared by both friends will improve the user experience.

Most platforms display a user's friends as a long and unstructured list of entries; some provide a graph representation which is often too dense to provide the user with a meaningful overview. A few platforms such as Xing allow users to tag their friends manually, using those tags to improve search and visualization. This functionality can be further improved by automatically grouping friends or coloring / partitioning the friendship graph according to shared tastes, alleviating the user from manually creating and updating tags for their friends.

In the following, we refer to tastes as latent concepts that manifest in re-occurring patterns of artists in user's playlists. We assume that a user has several tastes. Tastes are called shared tastes if these patterns re-occur in playlists of befriended users. We expect shared tastes to re-occur in the network.

Problem Statement. Given playlists for each user and the underlying social network, consisting of users as nodes and friend relationships as edges. Infer shared tastes that provide good predictors for three tasks: a) cluster friendships, b) predict artists of songs a user will listen to, and c) predict for given pair of friends their shared preference for artists.

Although we use the terminology of Zune Social for understandability, these concepts can be generalized to any social network with any types of community artifacts.

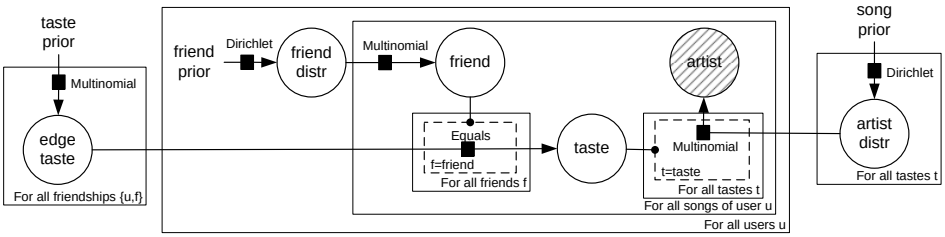

Figure 1: Shared taste model in directed factor graph notation with gates [2].

2 Our Approach

The shared taste model (cf. Figure 1) explains artists in playlists by tastes that the listener shares with his friends. In the generative process each edge in the social network is assigned one of out K tastes (called edge tastes), where each taste is associated with a distribution over artist names. For each user, a multinomial distribution over his friends is drawn from a flat Dirichlet. As users may have a different number of friends, the Dirichlet priors vary in the number of dimensions. For each song in a user's playlist, a friend is drawn from the friend distribution; the taste from the common edge (between user and friend) is associated with the song; and its artist is generated from the multinomial artist distribution of that taste. Therefore, this describes a generative model for the artists of the songs in users' playlists given the friendship graph.

The shared taste model extends Latent Dirichlet Allocation (LDA) [1] by coupling topics of friends. A user-wise topic distribution can be obtained from the shared taste model by integrating over incoming edges, that is, mixing edge taste distributions according to the friend distribution. Tastes of befriended users are coupled, encouraging an overlap in the user's topic mixtures.

Given playlists and friend relationships for all users, latent parameters of the friend and artist distributions and latent edge tastes can be estimated. We use Infer.net [3] to automatically derive a variational message passing algorithm for parameter estimation.

To predict the common preference for artists, the artist distribution associated with an edge taste acts as a predictive distribution. Predictive distributions for a user's listening behavior are obtained by integrating over friends and edge tastes, mixing the artist distributions accordingly. Inferred edge tastes provide a clustering of edges in the social network. Although the model assumes only one edge taste to be associated with an edge, its posterior distribution may have non-zero aspects for multiple tastes due to uncertainty, inducing overlapping edge clusters.

3 Extensions

Background models. We study the shared taste model with several extensions. In the shared taste model all songs have to be explained via friends. This assumption is violated in practice as users may also listen to music that they do not share with any of their friends. We study whether allowing the model to explain some songs via a background distribution will improve prediction performance. The background is either set to the empirical artist distribution to remove popular songs from tastes, or to user-specific artist distributions to remove individual aspects from the tastes.

Multiple tastes. One may argue that users with few friends have a limited choice of tastes, as the model allows for only a single taste per edge. To test whether this leads to suboptimal prediction behavior, we study an extension called the shared multi taste model. Instead of associating each edge with a single taste, the shared multi taste model associates an edge with a taste distribution, from which tastes for songs in the playlist are drawn.

This model decreases the coupling in befriended user's tastes. For instance, many songs in two users' playlists can be associated with the common edge, but be explained by disjoint tastes. The sparseness of the edge taste's prior will play a crucial role in interpolating between properties of the strongly coupled shared single taste model and an uncoupled LDA-style model.

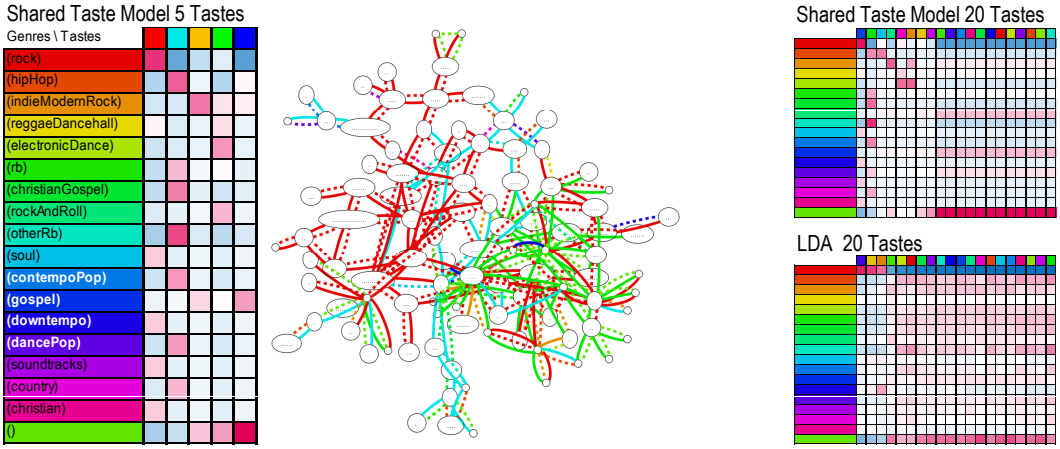

Figure 2: Center: Correlation between 5 inferred shared tastes (solid lines) and most common genre (dashed lines) per edge. Vertex labels indicate playlist lengths (one dot represents 10 songs). Pearson correlation coefficients are depicted on a temperature scale for 5 (left) and 20 tastes (right).

4 Experiments

The Zune Social platform hosts data about users, online friendships, and music listened to by the users—as required in the problem statement. We use playlists of one month as training data; the playlists of the following month as test data. The quality of the predicted ranking is evaluated in terms of area under the receiver operator curve (ROC-AUC) as well as marginal test set likelihood. Zune Social allows users to recommend songs to other users. The predicted shared preference for artists should match artists that are recommended from a user to his friend. Agreement between the predicted ranking of artists and recommendations between users is evaluated according to ROC-AUC.

Sensible predictions for edge clusterings are evaluated in terms of correlation with genres. For each friendship we use the number of co-occurring genres in both users playlists as ground truth (indicated by colors of dashed lines in Figure 2). Casting the proportions for each taste t vs. rest and each genre g vs. rest in the ground truth as binary random variables T and G for edges, we evaluate the agreement by pearson correlation coefficients $\rho(T, G)$.

Zune Social is one of few platforms that organizes its artifacts in genre categories and maintains item recommendations between users. We did not include this knowledge into the model as this would impede applicability to other platforms—we merely use it as an independent source of test data.

Sub-sampling the data. As we did not focus on scalability of the inference, we use a sub-sample of Zune's social network containing 100 connected users with recommendations and limiting vocabulary of artists to 1000. The sub-sample is selected by a Metropolis-Hastings sampler [4] to conserve properties of scale-free networks.

Reference methods. We compared the shared taste model and its extensions to results of a LDA model and a model that predicts artists according to popularity in the data set. To evaluate whether genres itself are good predictors for tastes, we study a variant of the shared taste model with known artist distributions as defined by genres which we call the shared genre model.

Results. Figure 2 indicates how well shared tastes inferred by the shared taste model correlate with genres. The shared taste model (and LDA) infer edge tastes that correlate well with genres if we use "the right" number of topics K. However, if we increase K to 20 the shared taste model still infers similar correlation patterns. The model uses an appropriate number of tastes, and infers nearly identical artist distributions for remaining tastes. In contrast, LDA overfits to subtle differences in user's playlists and is not able to give a concise clustering anymore.

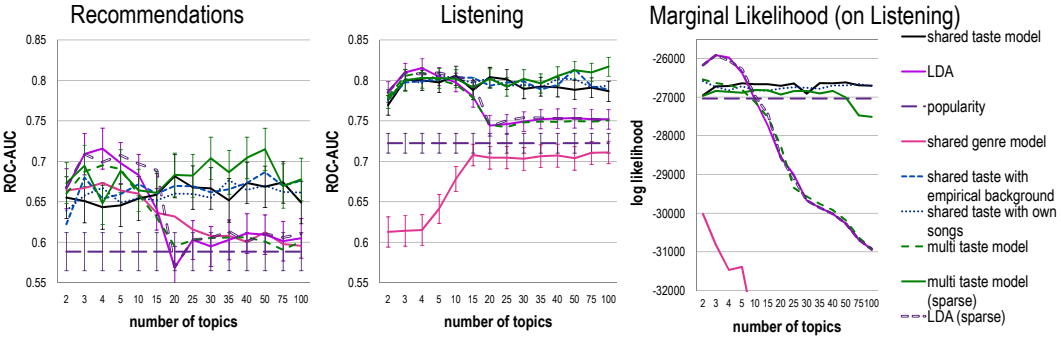

Figure 3: Prediction performance for song recommendations between users (ROC-AUC) and song listening behavior on the test set (ROC-AUC and marginal log-likelihood).

This behavior is reflected in the prediction performance as well. The shared taste model gives nearly constant prediction performance for $K \geqslant 3$ where LDA reaches the best performance for $K = 4$ and degenerates for higher K up to a point where it is merely as good as the popularity base line. The background models do not differ from the original shared taste model in terms of prediction performance. The multi shared taste model has similar behavior to either LDA for uniform Dirichlet taste prior or the shared taste model for a sparse Dirichlet(0.01) prior. It is as least as good as the best LDA predictions for 50 topics. In an experiment with 300 users and $K = 15$ (results omitted), we found that inferred tastes group genres in a meaningful way: One taste correlates with different rock and metal genres; another taste with hip hop and other r&b; a third one with r&b and reggae dancehall. LDA finds some correlations, but also indicates some counterintuitive ones. For example one taste correlates with rock and metal genres but also with electronic dance; another correlates with rock and pop as well as christian music; a third one correlates with pop but also with dancehall.

5 Conclusions

We propose a generative model that learns shared tastes from a social network structure and playlists of users. Inferred shared tastes give reasonable performances on the prediction tasks, while finding intuitive clusters that correlate well with genres. LDA gives good predictions but less intuitive clusters—especially when the number of tastes K is not chosen carefully. Performance of the shared genre model demonstrates that genres themselves do not reliably act as good predictors, and the extra effort of estimating tastes from data pays off. The shared multi taste model has the potential to give even better performances for the right sparseness of the edge taste prior. All variants of the shared taste model are insensitive to the right choice of the number of tastes K, allowing for reasonable predictions without the extra overhead of non-parametric methods.

Acknowledgments

Laura Dietz is supported by a scholarship of Microsoft Research Cambridge. Many thanks to Ed Snelson, John Guiver, and Ralf Herbrich for all the inspiring discussions and granting access to the Zune Social data set.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Tom Minka and John Winn. Gates. In *Neural Information Processing Systems*, December 2008.

[3] Tom Minka, John Winn, John Guiver, and Anitha Kannan. Infer.NET 2.3. http://research.microsoft.com/infernet, 2009.

[4] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. A walk in Facebook: Uniform sampling of users in online social networks. *ArXiv e-prints*, May 2009.