De-Layering Social Networks by Shared Tastes of Friendships

Laura Dietz
dietz@cs.umass.edu
University of Massachusetts, Amherst, USA

Ben Gamari
bgamari@physics.umass.edu
University of Massachusetts, Amherst, USA

John Guiver and **Edward Snelson**
{joguiver, edsnelso}@microsoft.com
Microsoft Research, Cambridge, UK

Ralf Herbrich
ralf@fb.com
Facebook

Abstract

Traditionally, social network analyses are applied to data from a particular social domain. With the advent of online social networks such as Facebook, we observe an aggregate of various social domains resulting in a layered mix of professional contacts, family ties, and different circles. These aggregates dilute the community structure. We provide a method for de-layering social networks according to shared interests. Instead of relying on changes in the edge density, our shared taste model uses content of users to disambiguate the underlying shared interest of each friendship. We successfully de-layer real world networks from LibraryThing and Boards.ie, obtaining topics that significantly outperform LDA on unsupervised prediction of group membership.

Introduction

Many successful community detection algorithms assume that having same interests is equivalent to having high within-group density, motivated by the homophily assumption (McPherson, Lovin, and Cook 2001). But recently, social media underwent a change from small platforms centered around specific interests to general-purpose networks such as Facebook or Twitter. This leads to observed social networks that are an aggregate of several layers social interest. Unfortunately, the aggregation of link structure leads to groups with high inter-linkage but no homogeneous interests. This violation of equivalence diminishes the results of many community detection methods.

As a precursor, we suggest to decompose the observed network into layers of shared interests. First each friendship is associated with the interest shared by both users, then layers are formed by friendships with the same shared interest.

Fostering an alternative paradigm, this work disambiguates friendships into layers with the help of user-generated content. The problem is challenging, as both the observed network and observed user-content are aggregates of the social layers that we aim to decompose. In order to identify the shared interest of a friendship, we would have to know how a) the user's content expresses the shared interest, and b) with which friend this interest is shared. Estimating either of these would in turn require knowing the underlying shared interest of the friendship. Further, it is unknown which shared interests are present in the data or would give rise to useful layers.

Problem statement. Given a social network with users $\mathcal{N}$ and friendships $\mathcal{E}$. For each user $u \in \mathcal{N}$, given a set of items $\mathcal{C}(u)$ from a common item vocabulary $\mathcal{V}$ such as tags or words. The goal is to learn shared tastes $\mathcal{T}$ and associate each friendship $\{u, f\} \in \mathcal{E}$ with tastes t based on both users items and friendships with other users.

Related Work

Identifying interests of a user has been the subject of many works on recommender systems and collaborative filtering, such as from Marlin and Zemel (2009). As the goal is to recommend new items to a user, recommender systems aim to learn interests of individual nature.

Classic SNA methods are based on random walks, modularity maximization, or graph cuts, with extensions to include content (Haveliwala 2003). All these methods rely on different edge densities in the network structure.

Recently, extensions of latent Dirichlet allocation (Blei, Ng, and Jordan 2003) have been applied to various kinds of network data. Cohn and Hofmann (2000) and Erosheva, Fienberg, and Lafferty (2004) suggest models that identify groups of nodes sharing items and friends, building on the idea of hubs and authorities. Mei et al. (2008) regularize the topic model with the assumption that friends share the same topics.

A popular choice for combining topic models with observed graph structure is the mixed-membership stochastic blockmodel (Airoldi et al. 2008). Stochastic blockmodels learn a topic mixture for each user and explain the presence and absence of friendships from the compatibility of topics. Pairwise Link-LDA (Nallapati et al. 2008) and the relational topic model (Chang and Blei 2010) extend blockmodels to model node contents $\mathcal{C}$ together with the network structure $(\mathcal{N}, \mathcal{E})$.

To learn link strengths in networks, (Xiang, Neville, and Rogati 2010) provide an unsupervised method based on a Gaussian model.

Our work differs in three ways: The shared taste model does not require the graph structure to form communities, as these might get diluted when network layers are aggregated.

Algorithm 1 Generative process of the shared taste model.

1: **draw** Influence strengths for friendships $\Psi \sim$ Dirichlet(α_ψ) ranging over $\{u, f\} \in \mathcal{E}$.
2: **for all** tastes $t \in \mathcal{T}$ **do**
3: **draw** item distribution $\phi_t \sim$ Dirichlet(α_ϕ)
4: **for all** friendships $\{u, f\} \in \mathcal{E}$ **do**
5: **draw** shared taste mixture $\lambda_{\{u,f\}} \sim$ Dirichlet(α_λ).
6: **for all** users $u \in \mathcal{N}$ **do**
7: **for all** items $x_{u,i} \in \mathcal{C}(u)$ **do**
8: **draw** friendship $\{u, f_{u,i}\} \sim$ Multinomial($\Psi|_u$) restricted to friendships of user u.
9: **draw** taste $t_{u,i} \sim$ Multinomial($\lambda_{\{u,f_{u,i}\}}$) from shared tastes.
10: **draw** item $x_{u,i} \sim$ Multinomial($\phi_{t_{u,i}}$) from the tastes's item distribution.

We prioritize the search for the common interests of two friends over learning user-centric interests. The topics learned by the shared taste model are of different nature than topics of other topic models. The shared taste model also includes the notion of a link strength between friends; the strength is explained as the propensity of linked users to mutually contribute to a common interest.

Shared Taste Model

The goal is to de-layer the observed network by clustering friendships (i.e., edges in the network) and associating each cluster with one layer. All friendships in the same layer are meant to have the same underlying shared interest, such as sports, same profession, etc, but each friendship can be part of more than one layer. As both the observed network and the observed user-generated content are aggregates of layers we aim to decompose, identifying how shared interests are reflected in two users' contents is an open problem. We provide a model that helps to understand what drives a friendship and group formation without relying on graph clusters.

We suggest the shared taste model, that jointly infers:

- a set of $\mathcal{T}$ shared tastes $\{\phi_t\}_{t\in\mathcal{T}}$, also called topics or interests,
- a soft-clustering of friendships $\{u, f\} \in \mathcal{E}$ into one (or few) of $\mathcal{T}$ layers,
- for each item x in the user's content: which shared interest t it expresses and with which friend f this interest is shared with.

Generative Process

The generative process is given in Algorithm 1. The shared taste model associates each friendship with a mixture over interests λ (cf. line 5). This is in analog to a topic mixture in latent Dirichlet allocation, where first a topic is drawn from the topic mixture ($t \sim \lambda$, line 9), then the item is explained from the item distribution characteristic for the topic ($x \sim \phi_t$, line 10).

Unlike most topic models, in the shared taste model, the topic mixture and influence is coupled across the network. For each friendship $\{u, f\} \in \mathcal{E}$, both users' items will be explained by the same topic mixture ($\lambda_{\{u,f\}}$). The effect is that the topic mixture captures aspects shared between u and f.

The coupling in the strength distribution Ψ leads to the effect where the more items of u are assigned to the common topic mixture, the higher the likelihood for f to assign items as well. The model has the option to infer that the users do not share a common interest, in which case Ψ ensures that the shared topic mixture is not used at all.

All items of a user u will be explained by the topic mixture shared with one of his friends. Friends f that have same or compatible (in the sense of a topic model) items will be associated with more items, thereby leading to higher strength estimates under $\Psi(\{u, f\})$ and forming the shared interests in $\lambda_{\{u,f\}}$ simultaneously. As a consequence, the item distributions ϕ that will explain the items associated with the friendship, will capture topics of sharing rather than individual interests.

Implementation

We provide a fast multi-threaded collapsed blocked Gibbs sampling implementation in Haskell, obtainable from http://www.cs.umass.edu/~dietz/delayer/. After random initalization of variables t and f, the variable configurations are subsequencely updated conditioned on current settings of the remaining variables. To achieve better convergence properties, distribution parameters are integrated out, and variables F and T are re-estimated together as block i in the blocked sampling update equation:

$$p(f, t|i, x, u) \propto \hat{\Psi}|_u(\{u, f\}) \cdot \hat{\lambda}_{\{u,f\}}(t) \cdot \hat{\phi}_t(x) \quad (1)$$

The implementation represents collapsed representations of as count statistics on other blocks $j \neq i$, for instance $\hat{\phi}_t(x) = \frac{|\{j|T_j=t, X_j=x, i\neq j\}|+\alpha_\phi}{|\{j|T_j=t, i\neq j\}|+T\alpha_\phi}$.$\hat{\Psi}|_u$ refers to the multinomial Ψ restricted to friendships of user u with $\hat{\Psi}|_u(f) = \frac{|\{j|U_j=u, F_j=f, i\neq j\}|+|\{j|U_j=f, F_j=u, i\neq j\}|+\alpha_\psi}{|\{j|U_j=u, i\neq j\}|+|\{j|F_j=u, i\neq j\}|+\alpha_\psi \cdot F(u)}$,where $F(u)$ is the number of friends of u.

Following Wallach, Mimno, and McCallum (2009), the inference algorithm re-estimates Dirichlet hyperparameters every 400 iterations. Hyperparameters α_ψ cannot be re-estimated as there is only one multinomial distribution Ψ. Therefore we let $\alpha_\psi = 0.5$ which reflects a mildly sparse prior.

We further study an interpolated version, where $x\%$ of the observations are explained by shared tastes, the remaining by user's individual topics as in LDA. Both components draw from the same set of item distributions ϕ.

Making Predictions

After inference converged, point estimates of the multinomial parameters are obtained from count statistics of all blocks i, e.g., $\hat{\phi}_t(x) = \frac{|\{i|T_j=t, X_j=x\}|+\alpha_\phi}{|\{i|T_j=t\}|+T\alpha_\phi}$.

If required, user-specific topic mixtures, such as obtained by latent Dirichlet allocation, can be inferred by $\hat{\theta}_u(t) \propto \sum_{f:\{u,f\}\in\mathcal{E}} \hat{\Psi}(\{u, f\}) \cdot \hat{\lambda}_{\{u,f\}}(t)$.

Aside from uncertainty, a given item can only be assigned to one friendship at a time. This introduces a competition in

the friendship influence distribution Ψ. If a user has multiple friends with the same shared interest t, items that fit t are split across those friendships. This introduces a bias in the estimation of parameter Ψ, which we correct by re-estimating the friendship strength $\check{\psi}_u(f)$ from current configurations $\mathbf{t}_u$ as the geometric mean under the shared taste mixture λ.

$$\check{\psi}_u(f) = \sqrt[|\mathbf{t}_u|]{\prod_{t \in \mathbf{t}_u} \hat{\lambda}_{\{u,f\}}(t)}$$

Notice, that $\check{\psi}_u(f)$ is not a multinomial distribution, but a positive weight.

Applications

De-Layering. For the application of de-layering an observed aggregated network, friendships with the same shared taste t are assigned to the same layer. Layer of topic t contains all friendships where t has higher than uniform weight in t, that is $\hat{\lambda}(t) > \frac{1}{T}$. Further, we omit weakly related friendships from network layers, that is if both re-estimated influences $\check{\psi}_u(f)$ and $\check{\psi}_f(u)$ are below threshold τ.

Visualizing contacts. The shared taste model also gives rise to a new user-centric visualization of his contact list. $\check{\psi}_u$ indicates friends with whom the user shared a lot of common interests. Further, contacts can be sorted according to different shared interests, indicated by λ. This is similar to the functionality of Circles on Google +, but without manual annotations.

Friend-based subscription. Many online platforms provide subscriptions to a friend, e.g., a user's radio station on last.fm. We speculate that when a user subscribes to a friend he has the common shared interest in mind. If the friend has many interests other than the one shared with the subscribing user, this may lead to negative user experiences. We might improve user satisfaction by filtering the friend's content according to the shared interest. The shared taste model gives rise to such a filter by

$$p(x|u,f) = \sum_{t \in \mathcal{T}} \lambda_{\{u,f\}}(t) \cdot \phi_t(x).$$

Experimental Evaluation

We evaluate the shared taste model on real world datasets from LibraryThing and the discussion forum Boards.ie.

LibraryThing

LibraryThing is a social networking platform, where users display and organize their virtual book shelf using tags. The platform hosts user groups, providing a group specific discussion forum centered around a common subject. We select 194 socially active users, their tags and friendships, to study whether the model can de-layer the friendship graph according to ten held out discussion groups. Since each user participates in multiple groups, the observed friendship graph forms one tightly interconnected component. The groups are of different sizes, with varying propensity to have intra-group friendships.

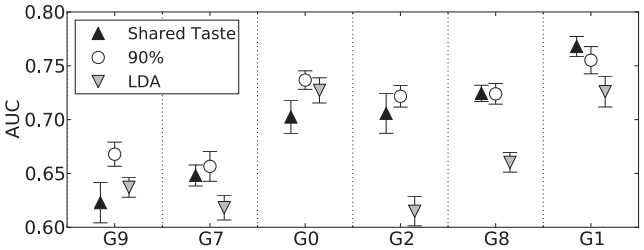


Figure 1: AUC of group membership prediction.

T12: mum's	T16: fairy tales	T13: suspense
cats(2.14e-2), crafts(1.97e-2), birds(1.79e-2), christmas(1.70e-2), scotland(1.61e-2), flowers(1.43e-2)	survival(1.71e-2), dragons(1.71e-2), vampires(1.62e-2), angels(1.54e-2), animals(1.36e-2), autism(1.36e-2)	thriller(3.10e-2), historical(2.59e-2), horror(2.52e-2), non-fiction(2.29e-2), vampires(2.22e-2), fantasy(2.07e-2)
$\rho_{G2} = 0.20$	$\rho_{G2} = 0.05$	$\rho_{G2} = -0.16$
$\rho_{G8} = 0.01$	$\rho_{G8} = 0.08$	$\rho_{G8} = 0.14$

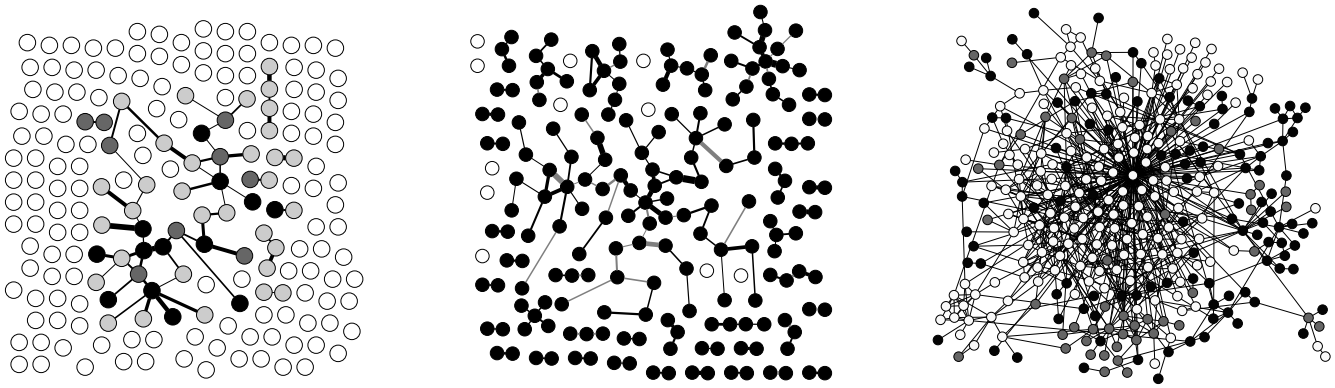
Table 1: Joint membership of friends in groups G2 and G8 with tags and Pearson correlation of topics 12, 13, an 16 with respect to $\lambda_{\{u,f\}}$.

Group membership prediction. In the lack of a natural evaluation criterion, we evaluate how well the shared tastes explain membership in user groups. We train a binary linear SVM classifier for group membership from topic features $\hat{\theta}$. The prediction quality is measured in ROC-AUC with standard error bars. Figure 1 displays results from the shared taste model in comparison to LDA, and an 90% interpolated shared taste model. Groups for which none of the model achieves an AUC of larger than 0.5 are omitted.

The shared taste model is significantly better than LDA for groups with intra-group friendships. The 90% interpolated model achieves even better performance. Interpolations with less than 90% diminish the performance gain, yielding results between LDA and the shared taste model.

Shared group membership and tags. We analyze how the joint interest of friends in groups G2 "Children's Literature" and G8 "Crime, Thriller & Mystery" correlate with identified topics in Table 1. The Pearson correlation coefficient $\rho_{g,t}$ between $\hat{\lambda}(t)$ and both users being member in group g indicates the group topic correlation for edges. Topic 12, which gather interests of mothers correlates with group 2. Topic 13 reflects interests in suspenseful literature, which is anti-correlated with G2. Topic 16 reflects mystery stories appropriate for children, correspondingly correlating with both groups.

De-layering of the social network. Figure 2 presents the predicted layers. The mutual influence is depicted in the thickness of the edges. We demonstrate the ability of identifying related user groups in Figure 2a: The layer of topic 17 contains users in groups G3 "Poetry Fool" and G5 "Tea!". Figure 2b shows that the approach prefers edges where both users share groups.



(a) LibraryThing topic 17. G3: dark grey, G5: black, other: light grey.

(b) LibraryThing topic 19. Edges with shared groups: black.

(c) Boards.ie layer "Computer Games" (white); Computer (black); TV (grey).

Figure 2: Predicted layers of the LibraryThing and Boards.ie network. Edge thickness indicates mutual influence $\check{\psi}$.

Boards.ie

We apply the shared taste model to the large dataset from the discussion forum boards.ie, provided by the organizers of ICWSM 2012. We gather the content-enriched network from the FOAF profile and thread titles. The dataset contains 1,298 users that have friends and posts. The resulting network contains 4,238 friendships. After preprocessing and stopword removal, the observed user-generated content contains 66,015 words from a vocabulary of size 9,022. We run the 90% interpolated shared taste model with 10 topics.

Scalability. Our Haskell implementation requires 110 seconds of CPU time per iteration over the data set. We ran the sampler for 1000 iterations, which took two hours on multicore computer using 24 cores. The sampler uses 10 gigabytes of memory.

Edge layers and node interests. The model identifies topics about sports, students, music bands, TV, computers in general, computer games, and an ethical discussion topic about god, abortion, and sexual orientation.

Figure 2c displays the central connected component after de-layering edges by the computer games topic. We highlight users with primary interests in the layer's topic, as well as in topic "computers", and "TV". The picture demonstrates that users with primary interest in computers mingle more tightly with gamers than those interested in TV. This kind of analysis would not have been possible without our de-layering approach.

Conclusions

We present a topic model approach for understanding shared interests in social networks, that does not rely on graph clusters to identify communities. Nevertheless, all topics obtained layers that have connected components. It is therefore suitable for de-layering aggregates of social networks, as demonstrated on data sets from LibraryThing and Boards.ie.

Acknowledgements

This work was supported in part by a Microsoft Research PhD scholarship, in part by the Center for Intelligent Information Retrieval, and in part by UPenn NSF medium IIS-0803847. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Airoldi, E. M.; Blei, D. M.; Fienberg, S. E.; and Xing, E. P. 2008. Mixed membership stochastic blockmodels. *JMLR* 9:1981–2014.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Chang, J., and Blei, D. M. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics* 4(1):124–150.

Cohn, D., and Hofmann, T. 2000. The missing link - a probabilistic model of document content and hypertext connectivity. NIPS '00.

Erosheva, E.; Fienberg, S.; and Lafferty, J. 2004. Mixed-membership models of scientific publications. *PNAS* 101(Suppl 1):5220–5227.

Haveliwala, T. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering* 15(4):784–796.

Marlin, B. M., and Zemel, R. S. 2009. Collaborative prediction and ranking with non-random missing data. RecSys '09.

McPherson, M.; Lovin, L. S.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–444.

Mei, Q.; Cai, D.; Zhang, D.; and Zhai, C. 2008. Topic modeling with network regularization. WWW '08.

Nallapati, R. M.; Ahmed, A.; Xing, E. P.; and Cohen, W. W. 2008. Joint latent topic models for text and citations. KDD '08.

Wallach, H. M.; Mimno, D.; and McCallum, A. 2009. Rethinking lda: Why priors matter. NIPS '09.

Xiang, R.; Neville, J.; and Rogati, M. 2010. Modeling relationship strength in online social networks. WWW '10.