Social Affinity Filtering: Recommendation through Fine-grained Analysis of User Interactions and Activities

Suvash Sedhain
ANU & NICTA
Canberra, Australia
suvash.sedhain@anu.edu.au

Scott Sanner
NICTA & ANU
Canberra, Australia
Scott.Sanner@nicta.com.au

Lexing Xie
ANU & NICTA
Canberra, Australia
lexing.xie@anu.edu.au

Riley Kidd
ANU
Canberra, Australia
rileyjkidd@gmail.com

Khoi-Nguyen Tran
ANU
Canberra, Australia
khoi-nguyen.tran@anu.edu.au

Peter Christen
ANU
Canberra, Australia
peter.christen@anu.edu.au

ABSTRACT

Content recommendation in social networks poses the complex problem of learning user preferences from a rich and complex set of interactions (e.g., likes, comments and tags for posts, photos and videos) and activities (e.g., favourites, group memberships, interests). While many social collaborative filtering approaches learn from aggregate statistics over this social information, we propose a different approach: we first define social affinity groups (SAGs) of a target user by analysing their fine-grained interactions (e.g., users who have been tagged in the target user's video) and activities (e.g., users who have joined the same special interest group that the target user has joined). Then we learn which SAGs are most predictive of the target user's preferences in a method we term social affinity filtering (SAF). We apply SAF to preference data from a set of Facebook users and their complete interactions with 38,000+ friends collected over a four month period. Our analysis demonstrates that SAF yields higher accuracy than a range of state-of-the-art (social) collaborative filtering approaches and that not all interactions and activities are equally predictive: among many insights, we show certain user-to-user interactions are more informative than others and we show that activity informativeness varies drastically with type, size, and exposure. In summary, this work demonstrates the previously untapped predictive power of fine-grained social interaction and activity features and the novel method of SAF to leverage them for state-of-the-art social recommender systems.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

Keywords

social networks, collaborative filtering, recommender systems

1. INTRODUCTION

Online social networks such as Facebook record a rich set of user preferences (likes of links, posts, photos, videos), user traits, interactions and activities (conversation streams, tagging, group memberships, interests, personal history and demographic data). This presents myriad new dimensions to the recommendation problem by making available a rich labeled graph structure of social interactions and content from which user preferences can be learned and new recommendations can be made.

Existing recommendation methods for social networks aggregate this rich social information into a simple measure of user-to-user interaction that can be leveraged to model user homophily via social regularization [34, 9, 22, 19, 23], a trust ensemble [20], or a low-rank factorization of the social interactions matrix [21]. But in aggregating all of these interactions and common activities into a single strength of interaction, we ask whether important preference information has been discarded? Indeed, the point of departure for this work is the hypothesis that different fine-grained interactions (e.g. commenting on a wall vs. getting tagged in a video) and activities (e.g., a university alumni group vs fans of a TV series) *do* represent different preferential *affinities* between users, and moreover that effective *filtering* of this information will lead to better results in social recommendation.

To quantitatively validate our hypotheses and evaluate the informativeness of different fine-grained features for social recommendation, we have built a Facebook App to collect detailed user interaction and activity history available through the Facebook Graph API along with user preferences solicited by the App on a daily basis. Specifically, each day our App recommends three links to each App user that are collected from the timeline of other users (both friends and non-friends) and we record users' explicit likes and dislikes of these recommended links. Given this data, we define *social affinity groups (SAGs)* of a target user by analysing their

fine-grained interactions (e.g., users who have been tagged in the target user's video) and activities (e.g., users who have joined the same special interest group that the target user has joined). Given these SAGS, we (a) learn to predict whether a user will like an item based on members of other SAGs who have also liked the item using a novel recommendation method we call *social affinity filtering (SAF)*, and (2) analyse the relative informativeness of different SAGs based on various properties such as type, size, and exposure.

In the four months that our App was active, we collected data for a set of Facebook app users and their full interactions with 38,000+ friends along with 22 distinct types of interaction and users activity for 3000+ groups, 4000+ favourites, and 10,000+ pages. In subsequent sections that outline our experimental methodology and results in detail, we make the following critical observations:

- We found that SAF significantly outperforms numerous state-of-the-art collaborative filtering and social recommender systems, by up to 6% in accuracy – in short, fine-grained interactions are very informative, bringing into question the efficacy of previous social recommendation approaches that aggregate user-to-user interactions into a single numerical value.
- We also found that groups, pages, and favourites make for more informative SAGs than those defined by user-to-user interactions – likely because the former can be applied to SAGs over the entire Facebook population rather than just a user's friends (where the data is considerably limited).
- Among the interactions, we found that those on videos are more predictive than those on other content types (photos, post, link), and that outgoing interactions (performed by the ego on the alter) are more predictive than incoming ones (performed by alters on the ego's timeline), although the level of exposure of an ego to an alter's preferences is more important than the directionality of the interaction with the alters.
- Among *groups*, *pages* and *favourites*, we found that the most predictive features have smaller membership size. We also found features corresponding to "long-tailed" content (such as music and books) tend to be more predictive than those without a large number of choices (e.g. sports or favorite teams).

As detailed in the subsequent sections, these findings not only demonstrate the power of leveraging fine-grained interaction and activity features but they also suggest which of these features are most important to collect when building SAF-based recommenders. This latter point is quite important since as we note later – the more permissions an App requests, the less likely a user is to grant permissions to the App, so choosing permissions (i.e., social features) well is crucial for achieving good recommendations with minimal intrusion into a user's privacy.

2. SOCIAL AFFINITY GROUPS

As illustrated in Fig 1, the high-level objective of this paper is to predict whether or not a user u will like a digital item i (in our test case, a link). We define user u's preference for link i as $like(u, i)$, this will be our prediction target.

$$likes(u,i) = \begin{Bmatrix} true & \text{if user } u \text{ likes item } i \\ false & \text{otherwise} \end{Bmatrix}$$

We define the social affinity between two users via their direct *interactions* on various digital items, and their shared *activities* in different communities of the social network. We call our recommendation algorithm *Social Affinity Filtering (SAF)*, as it infers $like(u, i)$ through the preference of their *Social Affinity Groups (SAG)*, i.e. a set of users with known preferences to link i, and who has at least one interaction or activity in common with u.

2.1 Action types on Facebook

On Facebook, We use the term *Interactions* and *Activities* to refer to the range of user-user and user-community actions, respectively.

Interactions describes the communication between Facebook users. There are a few dozen different interaction types that have distinct item modality, action and direction.

- **Modality:** (4 possibilities) User u can interact with another user v via *links, posts, photos* and *videos* that appear in either user's timeline.
- **Action type:** (3 possibilities) A user u can *comment* or *like* user v's item. He/she can also *tag* user v on an item, often indicating that user v is present when the content is created (for photo/video/post), or to explicitly raise user v's attention for a post – with one exception in Facebook that u cannot tag a link with users v since the link is created by third parties and merely shared on Facebook.
- **Directionality:** (2 possibilities) We look at *incoming* and *outgoing* interactions, i.e., if user u comments on, tags, or likes user v's item, then this is an outgoing interactions for u, and an incoming interactions for v. Although high correlation between *incoming* and *outgoing* interactions is recently observed [28], whether or not interaction direction affects user preferences differently is still an open question we wish to answer in this work.

There are a total of 22 interaction types. Namely the cross-product of modalities, actions and directions, minus *link-tag-*$\{$*incoming, outgoing*$\}$ since links cannot be tagged in Facebook.

Activities describes the user interactions with Facebook communities like groups, pages, favourites.

- **Groups** on Facebook [1] are analogous to community organizations in the real-world. It allows users to de-

[1]From Facebook Blog: http://www.facebook.com/blog/blog.php?

Figure 1: Overview of social affinity for link recommendation. A *social affinity group (SAG)* consists of the set of alters of a user u (ego) who have a certain interaction or share an activity membership with u. Alters defined by SAGs serve as proxies for an ego's interest with some SAGs showing stronger affinity with an ego as learned by *social affinity filtering (SAF)* and analysed subsequently.

clare membership and supports people to organize activities, to post related content, and to have recurring discussions about them. Examples of groups include *Stanford Thai* (Fig 1 bottom left), or *Harvard Debate Club*.

- **Pages** on Facebook [2] are analogous to the homepages of people, organizations and events on the world-wide-web. They are publicly visible, and users can subscribe to the updates on the page, and also engage in discussions. Example pages include *DARPA* (an organization, Fig 1 bottom middle), or *Beyonce* (a singer).

- **Favourites** are analogous to bookmarks (on physical books or on the web browser). It it is a user-created list containing various digital items such as Facebook apps, books, music, and many other types of items to indicate their interest. Example favourite items include *Big Bang Theory* (TV series), or *FC Barcelona* (soccer club). Fig 1 bottom right show a Facebook screenshot when a user adds a favourite. [3]

post=324706977130, "Groups are the place for small group communication and for people to share their common interests and express their opinion. Groups allow people to come together around a common cause, issue or activity to organize, express objectives, discuss issues, post photos and share related content".

[2]From Facebook Blog: (http://www.facebook.com/blog/blog.php?post=324706977130 "Facebook Pages enable public figures, businesses, organizations and other entities to create an authentic and public presence on Facebook. Facebook Pages are visible to everyone on the internet by default. Facebook user can connect with these Pages by becoming a fan and then receive their updates and interact with them."

[3]According to Facebook Blog, (https://www.facebook.com/help/232262810142682 "Facebook facilitates a wide variety of user selected favourites (Activities, Favorite Athletes, Books, Interests, Movies, Music, Sports, Favorite Teams, Television). These favourites allow a user to associate themselves with other people who share their same favourite tendencies.

Our evaluation includes 3000+ *group*, 4000+ *page* and 10,000+ *favourite* features as detailed in Sec 4.1.

2.2 Social Affinity Groups

Based on the definitions of *interaction* and *activities* above, we define two types of *social affinity groups* of a user with a target item, namely *Interaction Social Affinity Groups (ISAG)* and *Activity Social Affinity Groups (ASAG)*.

- **Interaction Social Affinity Groups**. Let the set of interaction affinity classes be the cross-product of Interaction modality, action and direction:

$$\begin{aligned} \textit{Interaction Affinity Classes} := &\{link, post, photo, video\} \\ &\times \{like, tag, comment\} \\ &\times \{incoming, outgoing\} \end{aligned}$$

  Then we define

  > *ISAG(u, k)* := the set of the users who have interaction k with user u.

  For example,

  > *ISAG(u, link-like-incoming)* is the set of all users who have liked link posted by user u.
  >
  > *ISAG(u, photo-comment-outgoing)* is the set of all users whose photos received at least one comment from u.

- **Activity Social Affinity Groups**: We define activity affinity groups based on group membership, page likes and user favourites.

  > *ASAG(u, k)* := the set of the users who have common preference for entity k (group, page, favourite) with user u.

2.3 Social Affinity Features

- **Interaction Social Affinity Features** : We define Interaction affinity features for target user u and item i for ISAG's classes $\langle X_1, X_2 \ldots, X_k \rangle$ as

$$X_{k,u,i} = \begin{cases} true & \text{if } \exists v \in ISAG(u,k) \wedge likes(v,i) \\ false & \text{otherwise} \end{cases}$$

  Additionally we add friend feature which encodes whether the target item i is liked by friend or not.

- **Activity Social Affinity Features** : We define activity affinity features for target user u and item i $\langle X_1, X_2 \ldots, X_k \rangle$ as

$$X_{k,u,i} = \begin{cases} true & if\ \exists\, v \in\ ASAG(u,k) \wedge likes(v,i) \\ false & otherwise \end{cases}$$

In our analysis we use only those features (groups, pages and favourites) that are joined/liked by at least one of our app users.

3. SOCIAL AFFINITY FILTERING

Social affinity filtering (SAF) is based on the idea that affinity between users expressed in social networks via interactions and activities captured by SAGs is predictive of user preferences. With SAGs now defined as in Sec 2.2, the task of SAF is simply one of classification of a user u's preference for a link i as outlined in Fig 1. While a classification approach to recommendation might evoke comparisons to standard *content-based filtering* (CBF) [18], we remark that CBF is not a social recommendation approach and unlike CBF, SAF does not require explicit user features (e.g., age, gender, location, etc.) or item descriptors (link text, link genre, etc.); in contrast, SAF requires only social interactions and learns the affinities between a user (ego) and the different set of alters as represented by SAGs that the user interacts or shares common activities with. SAF represents a simple and efficient yet nonetheless novel approach to *social recommendation* using fine-grained interaction and activity features that has not been previously proposed in the literature.

Our task in SAF is to predict for a given user u and item i whether $likes(u,i)$ is $true$ or $false$. For this purpose, we have the social affinity features $X_{k,u,i}$ defined in Sec 2.3 based on the various SAGs k defined in Sec 2.2; the $X_{k,u,i}$ specifically correspond to boolean features indicating whether any users in the kth SAG of user u also liked i. For example, k could be the SAG of u for the interaction of *link-like-incoming* or the activity of liking the *Obama Re-Election Headquarters* Facebook page. Then knowing whether anyone in each SAG k for user u likes item i provides a rich set of fine-grained features for prediction. It is up to SAF to learn how to weight each SAG k to aggregate all SAF preferences into one final prediction, which is done by training a classifier on historical data.

Formally, given a user u and item i, a SAF classifier is simply a function

$$f : \mathbf{X}(u,i) \rightarrow likes(u,i)$$

where $likes(u,i) \in \{true(like), dislike(false)\}$ and $\mathbf{X}(u,i) = \langle \cdots, X_{k,u,i}, \cdots \rangle$ for all SAG's k. To train f, one simply provides a dataset of historical observations $D = \{\mathbf{X}(u,i) \rightarrow likes(u,i)\}$ where f could be a linear classifier trained by an SVM, logistic regression, or naïve Bayes. Then for future predictions, we simply are given a new user u and item i to predict for and build the feature vector $\mathbf{X}(u,i)$ and predict $likes(u,i)$ using the trained $f(\mathbf{X}(u,i))$.

4. EVALUATION

4.1 Data Description

We built a Facebook App[4] to collect detailed user interaction and activity history available through the Facebook Graph API to build the SAG information as defined in Sec 2.2 along with user u's preferences on recommended links i solicited by the App on a daily basis to build the $likes(u,i)$ relation required in Sec 2.3. The data collection is performed with full permission from the user and in accordance with an approved Ethics Protocol[5].

Over 119 users installed the Facebook App during the evaluation period.[6] From these core App users, the App has access to their detailed Facebook profiles and their *complete* interactions with a total of 38,259 friends.

Our App tracks app user's (and their friends') details and interactions on Facebook. Interactions that occur through wall posts provide a rich variety of content and interaction data. We distinguish four Facebook items from wall posts: general posts (e.g., status updates, activity updates such as new friends), links, photos and videos. Four main interactions on these items are permitted by Facebook: posting an item to a friend's wall, commenting, liking, and tagging. Furthermore, Facebook allows user to interact with entities (groups, pages and favourites) via membership and liking. The App also keep tracks of the information of user's group membership, page likes and favourites. The App does not track deletions of these items, interactions (e.g., unlike) and memberships for performance reasons and we found very few deletions during an initial testing stage. We summarise relevant basic statistics of the data in Tables 2–4. Table 2 summarizes the number of records for each item (row) and interaction (column) combination. Table 1 shows some demographics from user profiles. Table 3 shows the group membership, page like and favourites counts for users.

Our app recommends three links to users every day, where the users may give their feedback on the links indicating whether they liked it or disliked it. Users are recommended links that are collected either from both friends and non-friends. Throughout this paper, App user rated link data is used to evaluate the recommendation algorithms. Table 4 shows the statistics of the app user rating for friend and non-friend recommendations. We chose to only display three links per day in order to avoid rank-bias with preferences;

[4]Name and link omitted for anonymity.

[5]Link omitted for anonymity.

[6]In order to collect full interactions among App users as required in our experimentation, our App requested to collect information posted to *friends'* timelines. With such expressive permissions concerning friend interactions, most potential users were hesitant to install the App; hence, after an intensive one month user drive period at our University, we were able to attract the pool of users used in the experiments. The difficulty of collecting data with such expressive App permissions suggests the importance of identifying a small subset of social features and permissions required to obtain them in order to encourage App adoption by a wide audience.

	App Users	Ego network of App Users
Users	119	38,378
Male	85	20,840
Female	33	17,032

Table 1: App user demographics

App Users	Tags	Comments	Likes
Post	7,711	22,388	15,999
Link	—	7,483	6,566
Photo	28,341	10,976	8,612
Video	2,525	1,970	843
Ego network of App Users	**Tags**	**Comments**	**Likes**
Post	1,215,382	3,122,019	1,887,497
Link	—	891,986	995,214
Photo	9,620,708	3,431,321	2,469,859
Video	904,604	486,677	332,619

Table 2: Statistics on user *Interactions*, grouped by item *Modality* and *Action type*.

	App Users	Ego Network of App Users
Groups	3,469	373,608
Page Likes	10,771	825,452
Favourites	4,284	892,820

Table 3: Statistics on user *Actions*, counted for *groups, pages* and *favorites* over the App users and their ego network.

	Friend recommendation	Non-Friend recommendation
Like	1392	1127
Dislike	895	2111

Table 4: Dataset breakdown of prediction target $like(u, i)$, by the source of the link (friend/non-friend) and value (true/false).

furthermore each link could be *independently rated* and in general user's rated all three links on a given day that they viewed their recommendations.

Crucially we note that all data used in the subsequent experiments is offline batch data that has been stored and analyzed after the four month data collection period.

4.2 SAF Analysis

In this section, we wish to compare novel SAF-based methods with a variety of (social) collaborative filtering baselines. For baselines, we examine the most likely class constant predictor (Const), and state-of-the-art collaborative filtering algorithms: Nearest Neighbor (NN) [6], Matrix Factorization (MF) [29] and Social MatchBox (SMB) [23] – an extension of matrix factorization that encourages more homophily for pairs of users with more interactions. For SAF methods, we analyse variants using interaction features (ISAF) and activity features (ASAF) for each activity class: group memberships, page likes, and favourites. We prefix the SAF method (ISAF or ASAF) according to the classifier used for training: naïve Bayes (NB-ISAF/NB-ASAF), support vector machine (SVM-ISAF,SVM-ASAF), and logistic regression (LR-ISAF,LR-ASAF). We report the average accuracy result using 10-fold cross validation and show standard error bars corresponding to 95% confidence intervals on the accuracy.

Fig 2 compares the accuracies of constant predictor and social matchbox with SAF based on naïve Bayes, logistic regression and SVM classification for a range of interaction and activity (group, page, favourite) SAGs. In all of our experiments SAF variants performed significantly better than Social Matchbox and the constant predictor except for naïve Bayes, we conjecture this is due to the correlations between SAGs that cause naïve Bayes to over- or under-estimate the true probability of likes.

In general we note in Fig 2 that activities appear to be more predictive than interactions, but we believe the reasons for this are quite simple: Interaction SAGs can only evaluate the friends of user u whereas Activity SAGs are able to look at all users, independent of u's friends. Hence, given the general sparsity of "likes" data in Facebook, Activity SAGs appear to draw on a much larger group of SAGs with much more activity.

Among activities, Fig 2 shows that activities are more predictive than interactions. Among the activities, page likes are the most predictive followed by group membership and favourites. Returning to our conjecture that data sparsity can hurt SAF, we note from Table 3 that page likes are more prevalent than groups and favourites. Moreover, these results may indicate that there is more social affinity between comembers of inherently social activities like groups and pages than between users who simply have favourites in common.

Comparing SAF to the state-of-the-art in social collaborative filtering (SCF) as represented by Social Matchbox [23], we observe that SAF consistently outperforms it. We note that the key difference of SAF vs. SCF is that SAF exploits the predictiveness of fine-grained interactions — it breaks down into groups, whereas most SCF approaches [23, 34, 9, 22, 19, 21, 20] is that instead of collapsing a diverse set of interactions into aggregate statistics such as the number of interactions between user u_1 and user u_2, regardless of whether u_1 tagged u_2 in a photo or u_1 liked a photo on u_2's wall. Clearly there is a great deal of benefit deriving from the fine-grained interactions indicating why without modeling any latent space and using a simple linear classifier, SAF can outperform SCF methods based on matrix factorization approaches that attempt to learn latent user and item features.

On two final notes, we remark that SAF yields a computational and optimization advantage over SCF in that it is

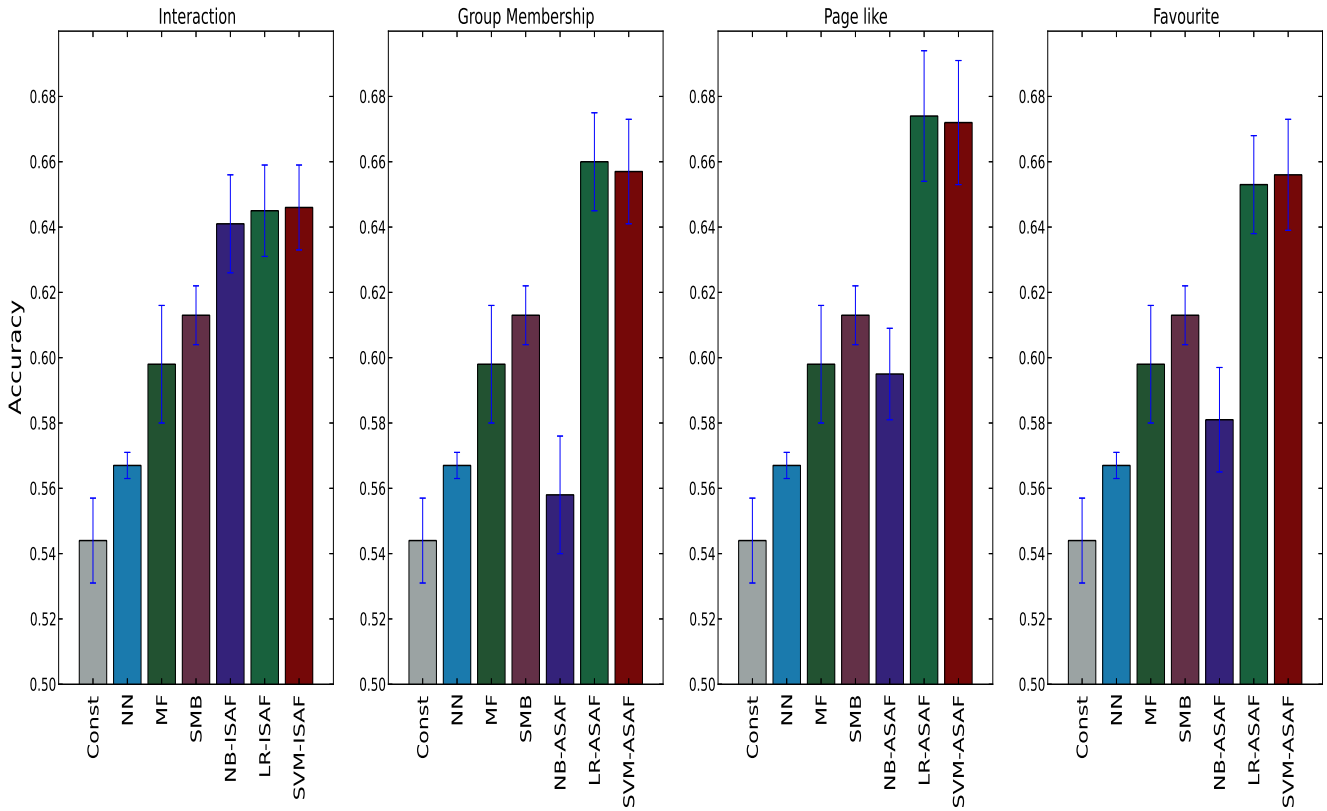


Figure 2: Comparision of the standard collaborative filtering algorithms with proposed SAF based models. SAF based models significantly outperforms standard collaborative filtering and social recommender systems, by up to 6 % in accuracy.

Table 5: Conditional entropy of various interactions (lower conditional entropies are more informative).

Modality (X)	$H(Y\|X = true)$
video	0.850
link	0.915
post	0.918
photo	0.926

Action Type (X)	$H(Y\|X = true)$
tags	0.920
comments	0.921
likes	0.924

Direction (X)	$H(Y\|X = true)$
outgoing	0.928
incoming	0.935

Modality-Direction (X)	$H(Y\|X = true)$
tags-outgoing	0.885
likes-outgoing	0.885
tags-incoming	0.900
likes-incoming	0.902
comments-outgoing	0.908
comments-incoming	0.912

Action-Direction (X)	$H(Y\|X = true)$
photo-outgoing	0.857
video-outgoing	0.863
link-outgoing	0.895
link-incoming	0.896
post-incoming	0.902
post-outgoing	0.906
video-incoming	0.915
photo-incoming	0.921

straightforward and efficient to find a globally optimal classifier with respect to certain training criteria (e.g., optimising log loss in logistic regression or hinge loss in SVMs) unlike SCF approaches that generally rely on computationally expensive matrix factorization techniques that lack optimality guarantees. Further, we also note quite surprisingly that SAF inherently scales to a large number of users and generalizes to completely new users without suffering from the cold-start problem: this is simply because nothing SAF learns is user dependent, it learns to weight SAGs independent of any particular user.

Given the clearly demonstrated benefits of SAF, we now proceed in the next two sections to analyse the two primary types of SAG features (interactions and activities) to better understand characteristics of both informative and uninformative SAGs in each context and the social phenomena that may be responsible for these characteristics.

4.3 Interaction Analysis

In this section we analyze the informativeness of Interaction SAGS, namely user interactions according to their modality, type, and direction, as described in Section 2.

A general method for measuring the amount of information that a feature X provides w.r.t. predicting another variable Y (in this case likes or dislikes) is to calculate its conditional entropy:

$$H(Y|X = True) = -\sum_{y \in (like, dislike)} p(y|X = true) \ln(p(y|X = true))$$

In general, a lower conditional entropy indicates a more informative feature. Here we use entropy conditioned on sparse features $X = True$ instead of just $H(Y \mid X)$ or mutual information $I(Y; X)$, as we found that the alternatives are highly correlated with (or dominated by) the number of occurrences of the feature (or $P(X = True)$), and give trivial results.

First we analyze various interactions individually and jointly to understand what interactions define SAGs with a high social affinity for a user u's preferences. To this end, we make a few observations from the conditional entropy analysis of Table 5:

- Interaction on *videos* seem to have a stronger preferential affinity than other modalities such as links, posts and photos. This could be due to the fact that video viewing is time-consuming and users inherently only watch the videos of those whose preferences they often share.
- Tagging has a slightly lower conditional entropy than commenting and liking, likely because the majority of tags contains person name(s) on Facebook, indicating a direct social interaction.
- A user is more likely to share preferences with someone who she initiates the interaction with (outgoing) vs. with someone who initiates the interaction with her (incoming). As an extreme instance of this, we note that while outgoing photo and video interactions are most informative, it appears that incoming photo and video interactions are least informative.

In figure 3 we plot conditional entropy of modality and action for incoming/outgoing interactions constrained to links liked by at least k friends in the SAG. Figure 3 reiterates many observations made above for various k. In addition, we note that preference affinity with a SAG increases as more people in the SAG like the item — then the more likely a user is to like an item. While incoming interactions were not as predictive as outgoing interactions for the same k, we note that higher k for an incoming interaction can be more predictive than lower k for an outgoing interaction. Note that these graphs are cumulative in k, different from the exposure curve on exactly k friends [27]. Our observations on user preference on items like by a number of Facebook friends suggest large cumulative number of friend interactions is more predictive – this can be translated into recommender system design. Further investigation is needed to pinpoint whether or not there is diminishing returns on repeated exposures [31, 27] on k, and how this could be leveraged to design recommendation algorithms.

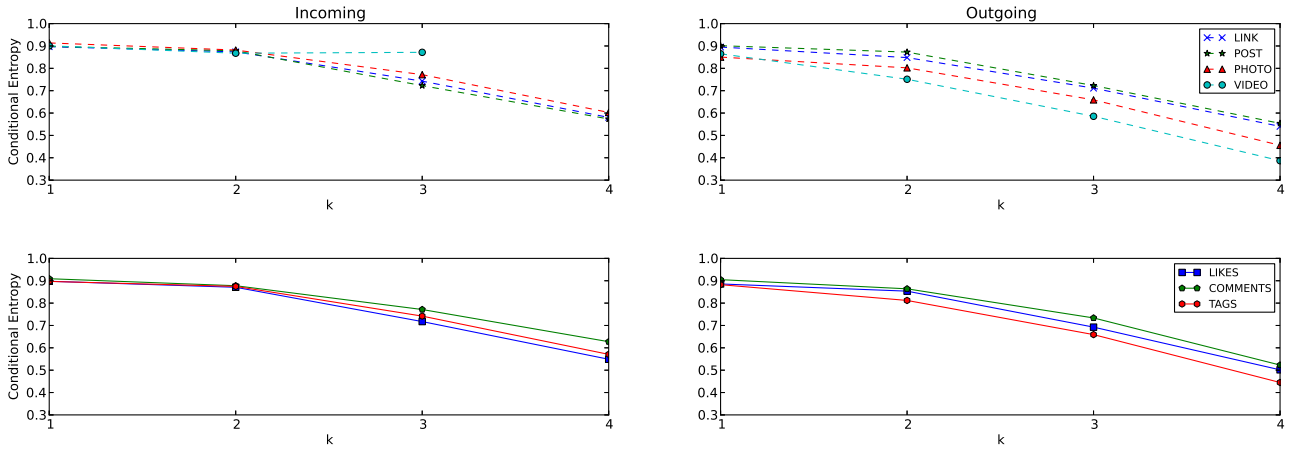


Figure 3: Conditional Entropy of modalities/activities for incoming/outgoing interactions vs item liked by at least k friends

4.4 Activity Analysis

We analyze the informativeness of Activity SAGS by looking at the correlation between the size and type of groups, pages and favourites.

Fig 4 shows the relationship between conditional entropy and logistic regression weights versus the size of activity groups. Here the size of a *group*, *page* and *favorite* is the number of total users in the activity group. For *Pages* and *Favorites* this is the total number of Facebook users, whether or not they are in the App users' ego network, while for *Groups* only the number of users in the App users' ego network is visible to our app. Both scatter plots shows that the activity groups of small size can be highly predictive (low conditional entropy or weights that deviate extremely from zero) whereas large groups are rarely predictive.

In Fig 5 we plot the average conditional entropy of the top 10% of features cumulative up to the size of the activity group given on the x-axis; this allows us to determine the marginal contribution of larger groups to the average conditional entropy as larger groups are incrementally added in. This graph distinctly shows that the small sizes of groups, pages and favourites have low average conditional entropy that transitions sharply to a higher average once a size threshold has been met. From Fig 5 we can infer that the group sizes up to 50 and page/favourite sizes up to 10^5 are most predictive.

We also analyze predictiveness of favourites by categories in Fig 6, where the favorite categories are obtained from Facebook API. We can see that contents in the "long-tail", i.e., having a large number of occurrences far from the most popular choices, tend to have be the most predictive affinities. Examples of these include music, books, movies. On the contrary, generic affinities (e.g. interests) and those with a smaller number of choices (e.g. sports or fav-teams) tend to be less predictive.

5. RELATED WORK

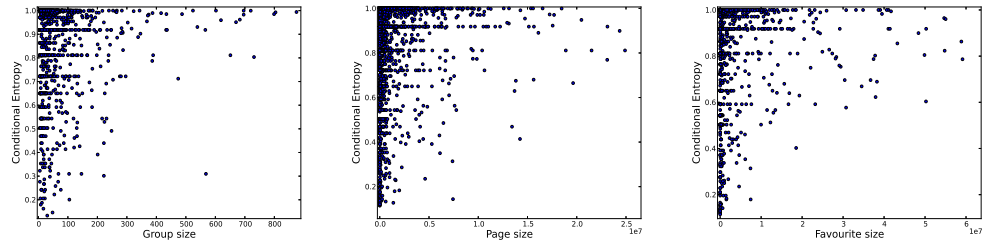
This work relates to many others in inferring user preferences on social and information networks. We structure the discussion into three parts: the first is concerned with the nature and observations on user traits, interactions and diffusion mechanisms; the second is concerned with correlating these user traits and interactions to user preferences and interests; the third is concerned with methods that uses these observations for predicting user interest or recommending content on social networks.

The first group of related work studies the nature of user profile, interactions, and diffusion. Profile information and demographics is correlated with user behavior patterns. Chang et al.[8] showed that the tendency to initiate a Facebook friendship differs quite widely across ethnical groups, while Backstorm et al.[3] have additionally showed that female and male users have opposite tendencies for dispersing attention for within-gender and across-gender communication. Two particular measurement studies on Facebook attention [33, 3] have inspired our work. Although the average number of friends for a Facebook user is close to the human psychological limit, known as the Dunbar number [14], the findings concur that a user's attention (i.e., interactions) are divided among a much smaller subset of Facebook friends. [3] studied two types of attention: communication interaction and viewing attention (e.g. looking at profiles or photos). Users' communication attention is focused on small numbers of friends, but viewing attention is dispersed across all friends. This finding supports our approach of looking at many types of user interactions across all of a user's contact network, as a user's interest is driven by where he or she focuses attention on.

The mechanisms of diffusion invites interesting mathematical and empirical investigations. The Galton-Watson epidemics model suits the basic setup of social message diffusion, and can explain real-world information cascade such as

Top Groups	Top Pages	Top Favourites
Heavy Metal - (city name)	Avascular Necrosis	Avascular Necrosis
Stephen Conroy Should Not Filter Our Internet	Assidian	Tortured
Silicone Stripper	Tortured	Elysian
Hardcore dancing is not moshing	Elysian	Anno Domini
Metal bands come to (city name) cause I'm sick of...	Darker Half	Hellbringer
(city name) Rock Gigs	Johnny Roadkill	Johnny Roadkill
Let's Mosh - (city name) metal radio show - 2XX 9	Anno Domini	Darker Half
Bring Steel Panther to Australia	Billy Madison	Bane Of Isildur
(city name) Death/Heavy Metal Appreciation	Hellbringer	Katabasis
Robert The Bruce (Band)	Metalocalypse	Aeon of Horus

Table 6: Top 10 Groups/Pages/Favourites ranked by Conditional Entropy. The city name where our institution and many Facebook users resides is anonymized.



(a) (b) (c)

(d) (e) (f)

Figure 4: Conditional entropy vs size (a-c); logistic regression feature weights vs size (d-f)

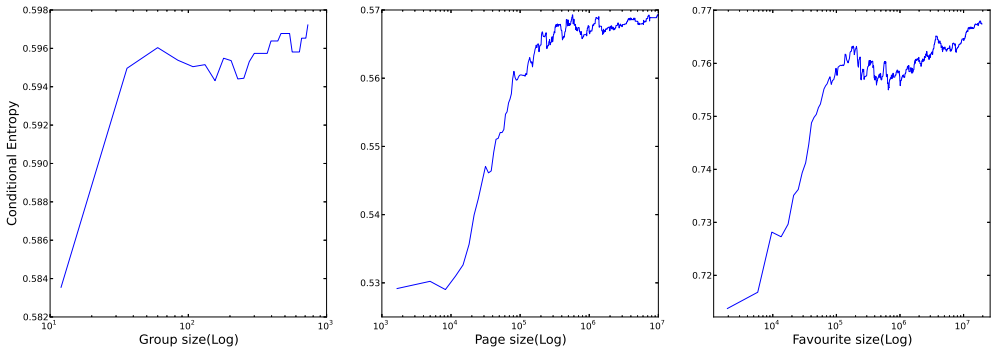


Figure 5: Average conditional entropy of top 10% groups, pages and favourites features cumulative over the size

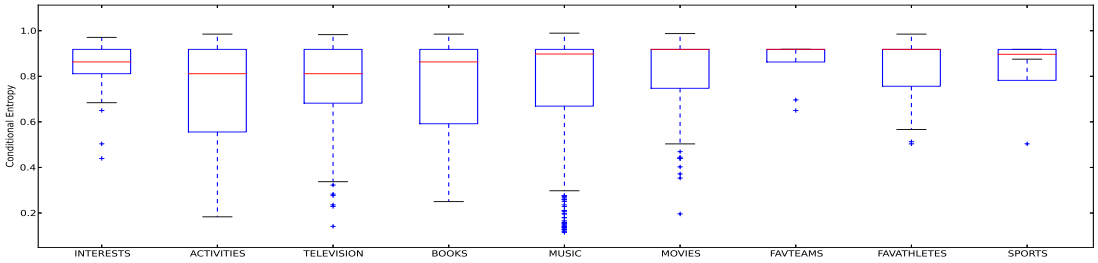


Figure 6: Conditional entropy for top 1000 favourites breakdown by categories

email chain-letters when adjusted with selection bias [12]. For social diffusions in a one-to-many setting, however, the epidemics model has been less accurate. Ver Steeg et al. [31] found that online message cascades (on Digg social reader) are often smaller than prescribed by the epidemics model, seemingly due to the diminishing returns of repeated exposure. Romero et al. [27], in an independent study, confirmed the effect of diminishing returns with Twitter hashtag cascades, and further found that cascade dynamics differ across broad topic categories such as politics, culture, or sports. Our observations on user preference on items like by a number of Facebook friends suggest large cumulative number of friend interactions is more predictive, further investigation is needed to pinpoint the effect of diminishing returns on repeated exposures.

The nature of social diffusion seem to be not only democratic [2, 4], but also broadening for users [5]. While influential users are important for cascade generation [4], large active groups of users are needed to contribute for the cascade to sustain [2]. Moreover, word-of-mouth diffusion can only be harnessed reliably by targeting large numbers of potential influencers, confirmed by observations on Twitter [4] and online ads [32]. In a study facilitated by A/B testing on Facebook links, [5] found that while people are more likely to share the information they were exposed to by their strong ties than by their weak ties, the bulk of information we consume and share comes from people with different perspectives (weak ties). Our Facebook App is intended to bridge this gap between insights from these observations and predicting user actions.

The second group of related work tries to correlate from user interactions to preferences and tie strength. Saez-Trumper et al [28] found that incoming and outgoing actives are highly correlated on broadcast platforms such as Facebook and Twitter, and such correlation does not hold in one-to-one mode of communication such as email. Multiple studies have found that online interactions tend to correlate more with interests than with user profile. Singla et al. [30] found that user who frequently interact (via MSN chat) tend to share (web search) interests. Anderson et al. [1] concluded that the level of user activities correlate with the positive ratings that they give each other, i.e., it is less about what they say (content of posts) but more about who they interacted with. Such findings echo those by Brandtzag [7] that real-world interactions (e.g., appearing in the same photo) further strengthens friendship on Facebook, while virtual interactions reveal interests. Furthermore, ratings of real-world friendship strength and trust [10] seems to be better predicted from the intimacy, intensity, and duration of interactions, than from social distance and network structure. Our work is not only inspired by these observations, we also quantifying the strength of correlations of user interest with a large variety of user affinities – namely, activities, and group preferences in different categories.

The last group of related work is concerned with using social network and behavior information for recommendation. Matrix factorization is one of the prevailing approaches for recommender systems [17, 21]. Recent advances include extending matrix factorization to user social relation in regularization [22, 19], to take into account multiple relations [26, 16], and to model social context [15]. In particular, there are different designs for using social information to regularize objective functions [34], a trust ensemble [20], a low-rank factorization of the social interactions matrix [21], or social-spectral regularization that takes into account user and item features [23]. These systems have shown very promising performance across a range of problems, but their all collapse social affinity (fine-grained interactions and group affinity) into one or a very low-dimensional representation. The point of departure of this work is to explore the rich affinity structure, we compared a recent matrix factorization approach [23] and found SAF with simple classifiers outperform state-of-the-art.

Additional work on predicting user actions join multiple social networks, and explores logical representation of user actions. Nori et al. [24] examines the predictability of user actions on Twitter from actions in both Twitter and Del.icio.us. The study uses both linear regression and an bipartite graph model that outperformed state-of-the-art models. Gomes et al. [13] derived rules for Facebook interactions using a psychology-inspired formal symbolic language. Our affinity definition is based on direct interactions within a users' ego network, this is complementary to a recent alternative [25] that uses number of paths between two users encodes the resilience

of network structure, as it was recently found [11] that the vast majority of information diffusion happens within one step from the source node. These work are most closely related to ours, yet none has examined such a diverse set of user actions in the same context: one-on-one interactions (e.g. commenting), broadcast (e.g. posting, sharing), and co-preference (e.g. likes).

In summary, our study is motivated by overall utility of diverse and fine-grained user interactions. To the best of our knowledge, this is the first work that look at 10,000+ different types of social affinity. We show that rich affinity features outperform state-of-the-art recommendation approaches, and our observations confirm the effect of demising returns on repeated exposure, we observe that contents in the *long tail* tend to be more predictive, and quantified the correlation of a large variety of affinity traits with user preferences.

6. CONCLUSIONS

We proposed Social Affinity Filtering (SAF) as a new method for social recommendation. We first defined social affinity groups (SAGs) of a target user by analysing their fine-grained interactions and activities. Then we learned which SAGs were most predictive of the target user's preferences leading to SAF. We evaluated the proposed algorithm on a dataset collected from a Facebook App we built, showing that SAF yields 6% absolute improvement in accuracy with respect to a state-of-the-art social recommendation engine. This is an important result given that SAF is built on standard supervised classification techniques unlike more complex matrix factorization approaches typically used in social collaborative filtering. Furthermore, we quantified the relative importance of interactions and activities for recommendation and we analysed what properties made some SAGs more predictive than others. Among many insights, our results show that video and photo interactions are more predictive than other modalities, and outgoing interactions are more predictive than incoming, and that smaller social groups are more predictive than larger ones. Future directions of research can investigate the nature of interactions by measuring the level of user engagement – e.g. if videos are inherently more engaging, or examine the nature of social groups via additional metrics – e.g. the social network within members of the group, or activity level of the group.

7. ADDITIONAL AUTHORS

8. REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Effects of User Similarity in Social Media. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2012.

[2] S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in social media: Persistence and decay. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[3] L. Backstrom, E. Bakshy, J. Kleinberg, T. Lento, and I. Rosenn. Center of attention: How facebook users allocate attention across friends. In *Proc. 5th International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[4] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. WSDM '11, 2011.

[5] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. *Facebook report, http://www.scribd.com/facebook*, 2012.

[6] R. M. Bell and Y. Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *ICDM-07*, 2007.

[7] P. B. Brandtzg and O. Nov. Facebook use and social capital — a longitudinal study. ICWSM'11, 2011.

[8] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow. epluribus : Ethnicity on social networks. In *ICWSM '10*, pages 18–25, 2010.

[9] P. Cui, F. Wang, S. Liu, M. Ou, and S. Yang. Who should share what? item-level social influence prediction for users and posts ranking. In *International ACM SIGIR Conference (SIGIR)*, 2011.

[10] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proc. CHI*, pages 211–220. ACM, 2009.

[11] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 623–638, New York, NY, USA, 2012. ACM.

[12] B. Golub and M. O. Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proc. Nat. Academy Sci.*, 107(24):10833–10836, 2010.

[13] A. Gomes and M. da Graca C Pimentel. Social interactions representation as users behavioral contingencies and evaluation in social networks. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 275–278. IEEE, 2011.

[14] R. Hill and R. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.

[15] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang. Social contextual recommendation. CIKM '12, pages 45–54, 2012.

[16] M. Jiang, P. Cui, F. Wang, Q. Yang, W. Zhu, and S. Yang. Social recommendation across multiple relational domains. CIKM '12, pages 1422–1431, 2012.

[17] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[18] K. Lang. NewsWeeder: Learning to filter netnews. In *12th International Conference on Machine Learning*

ICML-95, pages 331–339, 1995.

[19] W.-J. Li and D.-Y. Yeung. Relation regularized matrix factorization. In *IJCAI-09*, 2009.

[20] Ma, King, and Lyu. Learning to recommend with social trust ensemble. In *SIGIR-09*, 2009.

[21] H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: Social recommendation using probabilistic matrix factorization. In *CIKM-08*, 2008.

[22] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *WSDM-11*, 2011.

[23] J. Noel, S. Sanner, K.-N. Tran, P. Christen, L. Xie, E. V. Bonilla, E. Abbasnejad, and N. Della Penna. New objective functions for social collaborative filtering. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 859–868, New York, NY, USA, 2012. ACM.

[24] N. Nori, D. Bollegala, and M. Ishizuka. Exploiting user interest on social media for aggregating diverse data and predicting interest. ICWSM '11, 2011.

[25] R. Panigrahy, M. Najork, and Y. Xie. How user behavior is related to social affinity. WSDM '12, pages 713–722, 2012.

[26] S. Rendle, L. B. Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *KDD-09*, 2009.

[27] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. WWW '11, pages 695–704. ACM, 2011.

[28] D. Saez-Trumper, D. Nettleton, and R. Baeza-Yates. High correlation between incoming and outgoing activity: A distinctive property of online social networks? In *Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11, 2011.

[29] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS 20*, 2008.

[30] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. WWW '08, pages 655–664, New York, NY, USA, 2008. ACM.

[31] G. Ver Steeg, R. Ghosh, and K. Lerman. What stops social epidemics? ICWSM '11, 2011.

[32] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, December 2007.

[33] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao. User interactions in social networks and their implications. EuroSys'09, pages 205–218, 2009.

[34] Yang, Long, Smola, Sadagopan, Zheng, and Zha. Like like alike: Joint friendship and interest propagation in social networks. In *WWW-11*, 2011.