# Cross Social Networks Interests Predictions Based on Graph Features

Amit Tiroshi[†*], Shlomo Berkovsky[*], Mohamed Ali Kaafar[‡*], Terence Chen[*], Tsvi Kuflik[†]
[†]University of Haifa, Israel [*]National ICT Australia [‡]INRIA, France
[*]firstname.lastname@nicta.com.au [†]{atiroshi,tsvikak}@is.haifa.ac.il

## ABSTRACT

The tremendous popularity of Online Social Networks (OSN) has led to situations, where users have their profiles spread across multiple networks. These partial profiles reflect different user characteristics, depending mainly on the nature of the network, e.g., Facebook's social vs. LinkedIn's professional focus. Combining data gathered by multiple networks may benefit individual users, and the community as a whole, as this could facilitate the provision of more accurate services and recommendations. This paper reports on an exploratory study of the process of making such recommendations using a unique multi-network dataset containing user interests across multiple domains, e.g., music, books, and movies. We represent the data using a graph model and generate recommendations using a set of features extracted from and populated by the model. We assess the contribution of various network- and domain-related features to the accuracy of the recommendations and motivate future work into automated feature selection.

## Categories and Subject Descriptors

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

## Keywords

Social Networks; Graph Model; Interests Prediction

## 1. INTRODUCTION

The advent of OSNs has revolutionized user-to-user online and offline interactions. Users make a steadily increasing use of a plethora of OSNs, which are fast becoming the place to share and discover news, activities, and content of interest. The use cases notably vary across the networks: some facilitate microblogs, some are a place to share photos, and some serve as a professional playground. Hence, it is not unusual for a user to have accounts and profiles on multiple OSNs. Amalgamating these partial profiles can yield more accurate user models and lead to a better personalization [4].

Several prior works have focused on user modeling and personalization in a multi-OSN and cross-domain setting. In [1], a multi-OSN cross-domain dataset was collected and analysed for profile completeness across networks. The authors reported on the positive effects of the multi-OSN data in alleviating the cold start problem. A similar approach for aggregating data from multiple OSNs was presented in [10], where the dataset was enriched by semantics from external resources, like DBpedia. In [3], data from multiple OSNs was analysed in order to find what information could improve the diversity of recommendations. In [12], the authors investigated the feasibility and effectiveness of using cross-domain data for generating Facebook recommendations.

Our work relies on a rich dataset from multiple OSNs, which contains user interests belonging to multiple domains. We elaborate on prior works into cross-system and cross-domain personalization and aim to investigate what OSNs, domains, and features contribute to the accuracy of personalized interest recommendations. For this, we gather a collection of cross-OSN user profiles, and model this collection as a bipartite bidirectional graph. We engineer a set of user, interest, domain, and network related features and automatically populate these features from the graph model. Then, we apply a Random Forest classifier [6] in order to predict and recommend new topics of interest to users.

We assessed the contribution of data from six OSNs and five application domains. We identified networks and domains that improved the accuracy of the generated recommendations, while some other were neutral or even detrimental for the recommendations. Our results indicate that much attention should be given to the selection of features when aggregating partial user profiles, as the network and domain of origin play an important role and affect the value of information in the features. In summary, our work provides more evidence for the importance of feature selection in cross-system and cross-domain personalization, and calls for more research into automated feature selection.

## 2. DATASET

The dataset used in this work was collected in [7] and contains user profiles spread across six OSNs: Blogger, Facebook, Lastfm, LinkedIn, LiveJournal (LJ), and Youtube. The linkage of partial profiles across the OSNs was manually done by the users themselves. We extracted the lists of user interests and categorized these into five domains: movies, music, books, TV, and general. The categoriza-


*RecSys'13,* October 12–16, 2013, Hong Kong, China.

http://dx.doi.org/10.1145/2507157.2507206.

| OSN | total | unique | OSN | total | unique |
|---|---|---|---|---|---|
| Blogger | 27,045 | 6,587 | LJ | 30,924 | 5,198 |
| Facebook | 253,217 | 31,511 | Youtube | 2,753 | 1,561 |
| Lastfm | 63,952 | 16,483 | overall | 425,846 | 47,314 |
| Linkedin | 47,955 | 6,325 | | | |

**Table 1: Total and unique interests in each OSN.**

| domain | total | unique | domain | total | unique |
|---|---|---|---|---|---|
| general | 154,245 | 13,053 | books | 24,404 | 3,789 |
| movies | 54,382 | 5,190 | tv | 53,508 | 4,027 |
| music | 139,307 | 21,255 | all | 425,846 | 47,314 |

**Table 2: Total and unique interests in each domain.**

tion was explicitly made by the users on Blogger, Facebook, and Youtube; all Lastfm interests were categorized as music; and no categorization was available on LinkedIn and LJ, such that interests listed on these OSNs were treated as general. Users having one interest only in their aggregated profile and interests having one user only associated with them, were removed from the dataset.

The resulting dataset contains closely to 21K users and more than 47K interests, with an average of 19.46 interests listed per user. Table 1 shows the distribution of interests across the OSNs. Notably, almost 60% of the listed interests are coming from Facebook, with Lastfm and LinkedIn jointly providing more than 25% of interests on top of Facebook. Table 2 shows the distribution of interests across the domains. Domain distribution is more even than the OSN-based one, and while general interests accounts for the largest number of listed interests, music has the largest number of unique interests.

Table 3 shows the number of users who have at least one interest and Table 4 shows the average number of interests listed by a user for every domain-OSN combination. General interests and music are the dominant domains on Blogger, followed by movies and books. On Facebook, books is substantially lower than other domains, which are all comparable. The available domains are similar on Youtube, whereas TV interests are absent. As discussed earlier, music and general interests are the only domains on Lastfm and LinkedIn with LJ, respectively. Amongst the original domains (excluding general), movies are listed primarily on Facebook and Blogger, music on Lastfm and Facebook, books on Facebook and Blogger, and TV is listed only on Facebook.

# 3. METHODOLOGY

In this section we present a classifier designed to predict whether a user would like an interest. We use the Random Forest classifier [6] trained on features extracted both directly from the data (e.g., frequency) and from a graph model representation of the data.

## 3.1 Graph Model

In order to enrich the original dataset with additional features, we modeled the data using a graph. Consider a bipartite graph G = {U, I, E}, where U = $\{u_i \mid u_i$ is a user$\}$ and I = $\{i_j \mid i_j$ is an interest$\}$ are the vertices. $u_i$ and $i_j$ are connected if the $i_j$ is listed in one of the OSN profiles of $u_i$, i.e., E = $\{e_{ij} \mid u_i$ listed $i_j\}$. The edge contains labels detailing on which OSNs the interest was listed.

| | general | movies | music | books | tv |
|---|---|---|---|---|---|
| Blogger | 2,716 | 1,090 | 1,370 | 518 | – |
| Facebook | 8,391 | 8,922 | 10,453 | 6,565 | 9,619 |
| Lastfm | – | – | 7,042 | – | – |
| LinkedIn | 7,755 | – | – | – | – |
| LJ | 1,494 | – | – | – | – |
| Youtube | 448 | 484 | 650 | 552 | – |

**Table 3: Number of users with one interests listed in each domain and OSN.**

| | general | movies | music | books | tv |
|---|---|---|---|---|---|
| Blogger | 5.913 | 2.976 | 4.676 | 2.577 | – |
| Facebook | 6.999 | 5.662 | 6.509 | 3.414 | 5.562 |
| Lastfm | – | – | 9.082 | – | – |
| LinkedIn | 6.183 | – | – | – | – |
| LJ | 20.69 | – | – | – | – |
| Youtube | 1.288 | 1.278 | 1.395 | 1.177 | – |

**Table 4: Average number of interests listed by a user in each domain and OSN.**

From G, we extract a set of features, which are categorized into two groups: user features U and interest features I. Each of these groups is split into two sub-groups: basic features (IB for interests and UB for users) and graph features (IG and UG). The UB features are: number of OSNs on which the user has profile (UB1), number of interests the user has in each domain (UB2 - books, UB3 - TV, UB4 - movies, UB5 - music, UB6 - general), and total number of interests (UB7). The IB features are: number of users that liked the interest (IB1), number of OSNs on which the interest is listed (IB2), is it listed on each OSN (IB3 - Blogger, IB5 - LinkedIn, IB6 - LastFM, IB7 - LJ, IB8 - Facebook, IB9 - YouTube), and domain of the interest (IB4).

The graph features are similar for users and interests and include: degree centrality (IG1, UG2) [5], node redundancy (IG2, UG1) [9], clustering coeficient (IG3, UG4) [9], average neighbourhood degree (IG4, UG3) [2], and PageRank (IG5, UG1) [11]. Another shared feature is the shortest path length between a user and an interest, denoted by SP.

We also define $IG_{all} = \{\cup IG_i\}$, $UG_{all} = \{\cup UG_i\}$, $IB_{all} = \{\cup IB_i\}$, and $UB_{all} = \{\cup BG_i\}$. Finally, $I_{all} = \{IB_{all} \cup IG_{all}\}$ and $U_{all} = \{UB_{all} \cup UG_{all}\}$.

## 3.2 Classification

We are interested to explore the effect of various features on the accuracy of predictions for users liking interests. For the predictions, we used the Random Forest classifier [6], which was trained on the user-interest pairs that were augmented with the above mentioned graph features. The classifier was trained using 100 estimators, and 10 folds of the data were used for cross validation. A separate graph model was built for each fold, and the graph features were populated and fed into the classifier.

Random Forest is a binary classifier. As no disliked interests are included in the data, we randomly sampled per user interests not listed as liked ones, and considered these

| Feature/Group | Avg Precision | Feature/Group | Avg Precision | Feature/Group | Avg Precision |
|---|---|---|---|---|---|
| All | 0.6455 | IB4 | 0.5287 | UB5 | 0.4529 |
| IB_All | 0.5821 | IB5 | 0.5230 | UG1 | 0.4514 |
| IG1 | 0.5745 | IB6 | 0.5221 | UB6 | 0.4507 |
| IG2 | 0.5734 | IB7 | 0.5215 | UG_All | 0.4465 |
| IG3 | 0.5691 | IB8 | 0.5206 | UG1 | 0.4442 |
| IB1 | 0.5687 | IB9 | 0.5205 | UG2 | 0.4430 |
| I_All | 0.5642 | None | 0.5128 | UG3 | 0.4440 |
| IG_All | 0.5599 | SP | 0.5107 | UB7 | 0.4405 |
| IG4 | 0.5589 | UB1 | 0.4771 | UG4 | 0.4401 |
| IG5 | 0.5560 | UB2 | 0.4736 | UB_All | 0.4391 |
| IB2 | 0.5482 | UB3 | 0.4646 | U_All | 0.4376 |
| IB3 | 0.5307 | UB4 | 0.4632 | | |

**Table 5: Average precision across folds for individual features and feature groups**

as disliked[1]. The number of disliked interests was equal to the number of liked interests for each user.

We used Precision to evaluate the accuracy of the predictions, $P = \frac{TP}{TP+FP}$. Here TP is the number of correct and FP is the number of incorrect 'like' predictions.

# 4. RESULTS

## 4.1 Feature Analysis

We examine the average precision values obtained when the classifier is fed with single features, all features, or combinations of these. Table 5 summarises the results. We observe that the average precision scores vary between 0.44 and 0.58. Combining all the available features achieves the highest precision, with 65% of interests being predicted correctly. When no features except for the user-interest pair are embedded in the graph, precision falls closely to 0.5, as expected. This serves as the baseline for comparisons.

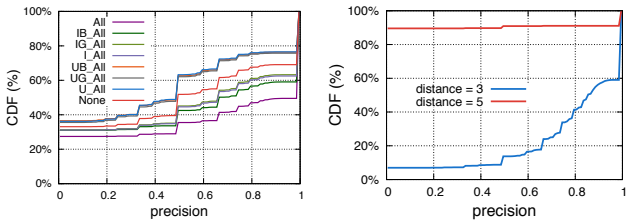
Notably, interest features, both individually and in combination, clearly outperform user features. The former score above the baseline (0.52–0.58) and the latter score below (0.44–0.48). Drilling down into specific features, we note that in the majority of cases, graph interest features IG are superior to basic interest features IB, whereas basic user features UB are superior to graph user features UG[2]. This allows us to conclude that interest features contribute valuable information and improve the accuracy of the classifier, whereas user features introduce noise into the model.

We plot in Figure 1(a) the cumulative distribution functions (CDF) of the precision of the combined user/interest features. In similar to Table 5, interest features outperform user features across the board. Consider the $I_{All}$ and $U_{All}$ curves. For $I_{All}$, precision greater than 0.5 is achieved by 56% of users and greater than 0.75 by 42% of users, while for $U_{All}$, these figures stand at 36% and 25%, respectively.

Also, we focus on the SP feature denoting the shortest path length between a user and an interest. Figure 1(b) depicts two CDF curves obtained for SP=3 and SP=5. For distant interests with SP=5, only 9% of users achieve precision to close 0.1 and the rest have 0. For interests with SP=3, 86% of users achieve precision greater than 0.5 and 66% of users – greater than 0.75. This shows that predictions for nearby nodes are easier than for distant ones.

[1]A similar approach is used in the Machine Learning community for classification of unary data [8, 13]. This setting reflects common real-life scenarios with only positive labels, e.g., purchase lists, Facebook likes, or browsing logs.

[2]Surprisingly, combining individual features results in an inverse trend. At the group of interest features, $IB_{All}$ achieves precision of 0.58 and $IG_{All}$ – 0.56, and for user features, $UG_{All}$ scores 0.45 and $IB_{All}$ – 0.44.




(a) Feature Combinations (b) Shortest Path Feature

**Figure 1: Precision CDF (a) feature combinations; (b) two values of the SP feature**

## 4.2 OSN and Domain Analysis

We use a combination of all the features and focus on the impact of OSNs and interest sources on the precision. Since the collected interests come from multiple OSNs, we evaluate the precision of the predictions for interests from a particular OSN. Figure 2 exhibits the CDF curves for precision scores for interests on the six OSNs: network=1 when an interest is listed on a network or network=0 otherwise.

As expected, Facebook (Figure 2, top right plot) as the largest source of interest in the dataset, plays a pivotal role in the overall predictions. When an interest is not listed on Facebook, only 30% of users achieve precision greater than 0.5. However, when an interest is listed on Facebook, the precision increases with 87% of users achieving a similar precision score. The gap between the two Facebook curves visually shows the importance of Facebook data; it is clearly wider than for any other OSN. Moreover, Facebook data spans across all the domains and largely overlaps with other OSNs. Hence, the shape and accuracy of the Facebook=1 curve resembles the curves of interests listed on other OSNs.

However, the explanation of this importance does not stem solely from the number of interests in each OSN. For instance, Lastfm (Figure 2, middle left plot) demonstrates a counter-intuitive result. While it provides a high number of music interests – and many of them are unique – being listed on Lastfm has a weaker effect on the predictions than being listed on Facebook. The difference between the two Lastfm curves for precision greater than 0.5 stands at 17% only. Furthermore, Lastfm is the only OSN where Lastfm=0 predictions achieve higher precision scores than Lastfm=1. We posit that these observations are explained by the low overlap between the interests listed on Lastfm[3] and on other OSNs. Hence, predictions for Lastfm-specific music interests are inherently harder than for interests on other OSNs.

We proceed with examining the importance of an interest being listed on more than one OSN. Figure 3(a) shows the precision scores averaged for interests listed on a different number of OSNs, varying from n=1 to n=5. It can be clearly seen that precision increases with n. For example, precision greater than 0.5 is achieved for 96% of interests listed on five OSNs, 70% of interests listed on three OSNs, and 57% of cases listed on one OSN. This suggests that predictions for interests listed on multiple OSNs are easy, as they are based on more reliable data, and these interests could be recommended with high confidence.

[3]Recall that Lastfm's interests belong to music only.

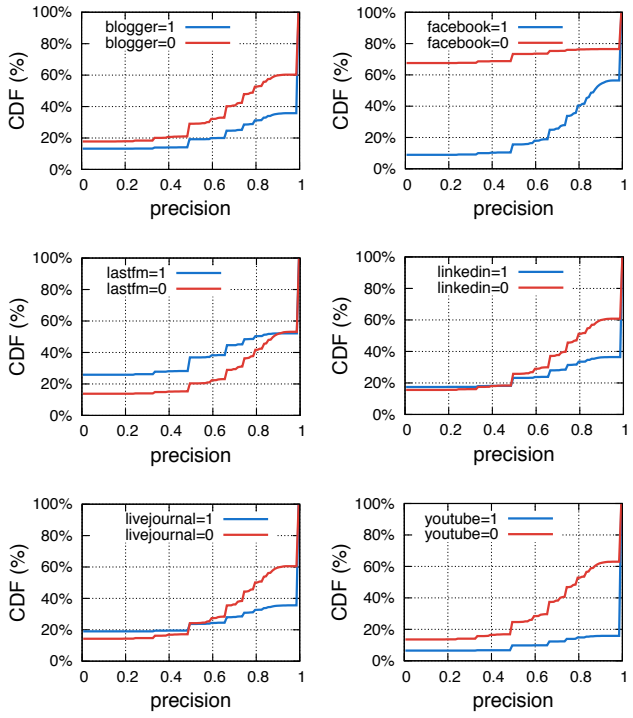



**Figure 2: Precision CDF - interests listed on the six OSNs**

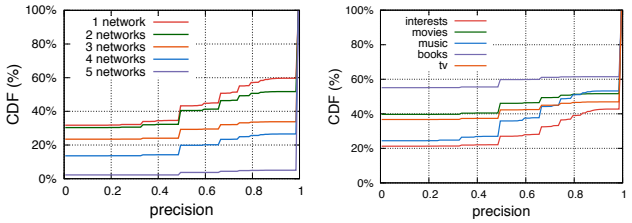
Finally, we assess the precision scores achieved for interest that belong to various domains. Figure 3(b) show the CDF curve for each domain. We observe that general interests achieve the highest precision, with 73% of interests predicted with precision greater than 0.5. This is followed by music with 64%, TV and movies coming close with 58% and 54%, and books with 40% of interests only. These scores can be explained by the volume and uniqueness of interests within each domain.

General interests include the largest number of interests, while not too many them are unique. Hence, the predictions achieve the highest precision. General interests are followed by music, but the latter have much more unique interests, such that the precision is lower. TV and movies domains have a comparable number of interests, but TV has less unique interests and its precision is slightly higher. Finally, books have the smallest number of interests, and, consequently, achieves the lower precision.

# 5. CONCLUSIONS

This work follows on previous works on multi-OSN personalization and studies how information gathered from various OSNs can lead to more accurate interest recommendations. Specifically, we investigate the potential of multi-OSN, cross-domain, and graph based features. We showed that the contribution of the gathered data to the accuracy of the predictions varies across the OSNs: it is heavily dependent on the features in use and on the source, domain, popularity, and uniqueness of the interests.

Our work represents the first step towards an aggregated view of user interests, as expressed on social networks. We plan to investigate the ways to obtain the optimal combination of features leading to higher precision rates, and also to conduct a deeper analysis of the impact of data availability on the classification process.




(a) Network Analysis (b) Domain Analysis

**Figure 3: Precision CDF (a) interests on multiple OSNs; (b) interests from various domains**

# 6. REFERENCES


[1] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system User Modeling and Personalization on the Social Web. *User Modeling and User-Adapted Interaction*, 23(2-3):169–209, 2013.

[2] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The Architecture of Complex Weighted Networks. *PNAS*, 2004.

[3] A. Bellogín, I. Cantador, and P. Castells. A Comparative Study of Heterogeneous Item Recommendations in Social Systems. *Information Sciences*, 2012.

[4] S. Berkovsky, T. Kuflik, and F. Ricci. Mediation of User Models for Enhanced Personalization in Recommender Systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286, 2008.

[5] S. P. Borgatti and D. S. Halgin. Analyzing Affiliation Networks. *The Sage handbook of social network analysis*, 2011.

[6] L. Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001.

[7] T. Chen, M. A. Kaafar, A. Friedman, and R. Boreli. Is More Always Merrier?: A Deep Dive into Online Social Footprints. In *ACM SIGCOMM WOSN*, 2012.

[8] C. Elkan and K. Noto. Learning Classifiers from Only Positive and Unlabeled Data. In *KDD*, 2008.

[9] M. Latapy, C. Magnien, and N. D. Vecchio. Basic Notions for the Analysis of Large Two-Mode Networks. *Social Networks*, 2008.

[10] F. Orlandi, J. Breslin, and A. Passant. Aggregated, Interoperable and Multi-domain User Profiles for the Social Web. In *I-SEMANTICS*. ACM, 2012.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford University, 1999.

[12] B. Shapira, L. Rokach, and S. Freilikhman. Facebook Single and Cross Domain Data for Recommendation Systems. *User Modeling and User-Adapted Interaction*, 23(2-3):211–247, 2013.

[13] J. Xie and T. Xiong. Stochastic Semi-Supervised Learning on Partially Labeled Imbalanced Data. *Active Learning Challenge Challenges in Machine Learning*, 2011.