Link Predictions in an Online Health Community for Smoking Cessation

Sulyun Lee
Interdisciplinary Graduate Program in Informatics
University of Iowa
Iowa City, IA, USA
sulyun-lee@uiowa.edu

Hankyu Jang
Department of Computer Science
University of Iowa
Iowa City, IA, USA
hankyu-jang@uiowa.edu

Kang Zhao
Department of Business Analytics
University of Iowa
Iowa City, IA, USA
kang-zhao@uiowa.edu

Michael S. Amato
Innovations Center, Truth Initiative
Washington, DC, USA
Department of Medicine, Mayo Clinic
College of Medicine and Science
Rochester, MN, USA
mamato@truthinitiative.org

Amanda L. Graham
Innovations Center, Truth Initiative
Washington, DC, USA
Department of Medicine, Mayo Clinic
College of Medicine and Science
Rochester, MN, USA
agraham@truthinitiative.org

Abstract

Effective link predictions in online social networks can help to improve user experience and engagement, which are often associated with better health outcomes for users of online health communities (OHCs). However, limited attention has been paid to predicting social network links in OHCs. This paper explores link predictions in an OHC for smoking cessation by considering it as a multi-relational social network that incorporates multiple types of social relationships. We demonstrate that leveraging information from multiple networks built based on different types of relationships is superior to using only information from a single network or the aggregated network. In addition, adding community structures and nodal similarities based on network embedding can help link predictions in different ways. Our work has implications for the design and management of a successful online health community.

CCS Concepts: • **Information systems** → **Data analytics**; **Social networking sites**.

Keywords: Social Network, Network Embedding, Multi-Relational Network, Supervised Learning, Smoking Cessation

MLG '20, August 24, 2020, San Diego, CA

https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Sulyun Lee, Hankyu Jang, Kang Zhao, Michael S. Amato, and Amanda L. Graham. 2018. Link Predictions in an Online Health Community for Smoking Cessation. In *MLG '20: 16th International Workshop on Mining and Learnign with Graphs, August 24, 2020, San Diego, CA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

Online health communities (OHCs) allow people with similar health concerns to seek and receive social support [4, 10]. Such communities have seen increased popularity during recent years [9] and have led to powerful health benefits, including improved emotional and psychological well-being [30, 34, 42] and better health outcomes [19, 21, 47]. Specifically for smoking cessation, Graham et al.[19] showed that greater participation in an OHC for smoking cessation was associated with higher abstinence rates.

One of the keys to the success and longevity of OHCs is the active and sustained engagement of members over time [46]. Being able to predict which members may benefit from connections or "links" with other members can enable OHCs to facilitate such connections to nurture and grow the community. The goal of link predictions in social networks is to infer where new ties may form in the future based on information extracted from the current network [24]. Accurate predictions of links for an OHC can help users connect with each other, and hence get them more engaged in the OHC [40]. In an OHC for smoking cessation, current smokers who are actively trying to quit may benefit from the advice and encouragement from former smokers, and they may also benefit from the validation and shared experience of others. Greater levels of engagement may lead to increases in self-efficacy for quitting[7, 12]. Evidence of a dose-response

relationship for engagement in OHCs with successful outcomes has been observed [15, 18, 20]. To date, relatively little empirical work has focused on link predictions in OHCs.

Most OHCs offer various types of communication channels to facilitate user interactions. This means that links may be established across multiple channels (e.g., public vs. private) and types of communications (e.g., one-to-one message vs. one-to-many message) which may be appropriate for different purposes [4]. For example, a messaging channel allows individuals to send direct and private messages in a one-to-one fashion, whereas in a public discussion channel, users post content and reply to others' posts as a group.

Interactions through these channels naturally form a multi-relational network [47] where individuals are connected by different types of social relationships. Previous research that has examined content in OHCs has typically focused on single communication channels [43] and has not considered the potential information available from a multi-relational network. In addition to identifying the individuals a user has interacted with, the large amount of user-generated content from OHCs may also be an important source of information for link prediction.

This paper proposes a novel approach for link prediction in OHCs. *First*, we demonstrate that for OHCs with multiple channels of communications, a multi-relational perspective is valuable in improving link prediction performance. *Second*, we show that community structures, as well as nodal similarities based on network embeddings can contribute to link predictions in OHCs, while textual similarities do not contribute much predictive value. We review related research in Section 2, describe the unique dataset involved in this work in Section 3, and present our approach for multi-relational link prediction in Section 4. Section 5 details the results of our experiments and we conclude with discussions of the practical implications of this work and future research directions.

2 Related Work

As OHCs have become more popular and important for people to access health-related information and support, many studies have investigated OHCs from different perspectives. Some analyzed or predicted users' participation in OHCs [40, 42]. Many researchers focused on the content of users' online discussions to identify common topics [29, 33], investigate the quality of information exchanged in OHCs [41], and discover information about drug use [44]. Another stream of research attempted to discover major factors that affect health outcomes of OHC users [26, 30].

There are also various studies on OHCs specifically for smoking cessation. For example, Cobb et al.[12] analyzed how users' smoking status is influenced over time by their interactions with others. A network analysis also found that adopting the multi-relational perspective can reveal interesting patterns between users' abstinence and their network centralities in different communication channels [47]. Other research attempted to identify individuals' offline smoking status from user-generated content [3, 39] or studied social support patterns [32].

In social network analysis, the main idea behind link prediction is that similar nodes are likely to be connected in the future, as reflected in the concept of homophily (a.k.a., "birds of a feather") [27]. Therefore, determining how to measure "similarity" between nodes is an important task that proceeds link predictions. Existing measures are mainly based on structural characteristics (e.g., the number of common neighbors between two nodes) or nodal attributes (e.g., keywords used by authors in their publications).

Based on this idea, there are two major ways to predict links: unsupervised learning and supervised learning [23]. Unsupervised method computes similarity scores for all pairs of nodes in a network. Then top scoring node pairs are predicted to form links in the future [24]. Such methods are often intuitive and can easily generalize to networks in different contexts.

By contrast, supervised methods uses machine learning to find the difference between node pairs that form links in the future and pair that do not [25]. The typical setup of supervised link predictions is a binary classification problem, whose feature sets can include several similarity measures computed during the training period for existing ties in a network. Then different classification algorithms can be deployed for link predictions between pairs of disconnected nodes. Because they are trained based on ground-truth data of a specific network, supervised link prediction methods usually outperform unsupervised methods for the same network [25, 37], although the model trained on one network may not work well for another network.

However, most link prediction methods treated all social relationships homogeneously by aggregating different social relationships into one network. In a multi-relational network, potential interconnections among different types of relationships may have made it challenging to predict links. At the same time, such complexity also offers opportunities to leverage more fine-grained information from different types of social relationships for link predictions, because users' interactions via one type of relationship may affect the formations of ties based on another type.

Thus, some studies have started to incorporate the multi-relational nature of the networks into link prediction [35]. For example, Davis et al.[16] proposed a link prediction method based on enumerating all possible patterns in triads, but at the cost of high computational complexity.Wang and Sukthankar[38] attempted to find different types of edges from a single-relational network with edge clustering [36], and added the type(s) of edges attached to a node to improve link predictions. However, the edge between two nodes can

only belong to one type, while in multi-relational networks, two nodes can be connected by multiple edges, each representing a different type of social relationship (e.g., two individuals can be neighbors and colleagues at the same time).

3 Data and Setup

3.1 Data Source

This study used data from BecomeAnEX.org [1], a Web-based smoking cessation program developed and managed by Truth Initiative in collaboration with Mayo Clinic. The dataset used in this study spanned the period from January 1, 2010, to May 31, 2015, and included records of both posting and reading behaviors of users who accessed content of the community on BecomeAnEX by clicking and reading a post (e.g., a blog, a message board post, or a group discussion thread) or a private message. The community was migrated from a different platform before this period, which resulted in a slightly different user experience. Our analyses focus on this time frame given the stability of the social network feature set.

During this time, there were four types of communication channels in the BecomeAnEX community: 1) blogs & comments (BC), 2) group discussions (GD), 3) message boards (MB), 4) private messages (PM). For each of these channels, we constructed one sub-network based on users' interactions through that channel. Thus, the four sub-networks, one for each channel, constitute a multi-relational network, where the same set of nodes are connected by edges that represent different types of relationships [45, 47, 48].

In the sub-network for BC, we connected the author of a blog post with those who posted comments to the blog post. Similarly, for GD, which is characterized by threaded discussions, the GD sub-network connected the original poster of a thread with others who replied to the thread. Similar to "walls" in online social networks such as Facebook, MB allows one user to post a message on another's message board. Thus the MB sub-network connected a message board owner with those who left a message on the board. PM represents one-to-one communication and ties in the PM sub-network connected the senders and recipients of private messages. Ties in all sub-networks were undirected because when it comes to social support, both seeker and providers can benefit from such activities [1].

3.2 Setup

We adopted a sliding-window approach and set up the link prediction on a weekly basis–predicting if two currently disconnected nodes will form a new tie during the next week based on what is observed during the current week. For our experiments, we selected 32 consecutive weeks of data, from week 50 to week 81, because this period represents one of the most active periods in terms of user posing activities. Four sub-networks, G^{BC}, G^{GD}, G^{MB}, and G^{PM} were constructed based on users' activities made during anytime between week 50 to week 80. In addition to these sub-networks, we also constructed an aggregated network G^{AGG} that aggregates all of the user interactions across the four channels–as long as two nodes are connected in one of the four networks, they are connected in the aggregated network. Table 1 summarizes statistics of the five networks mentioned above; $|V|$ and $|E|$ denotes number of nodes and edges, $degree_{mean}$, $degree_{max}$, $degree_{std}$ denotes mean, max, and s.t.dev of degrees, C and r denotes clustering coefficient and assortativity, and n_{comp} denotes the number of connected components.

Because real-world networks are often sparse, meaning that most node pairs are not connected, one problem in link prediction is an unbalanced dataset that has way more negative instances (i.e., node pairs with no ties) than positive instances (i.e., node pairs connected by ties). To make the dataset more balanced, we adopted a common approach–only included node pairs (i.e., instances) that are two hops away in the aggregated network G_t^{AGG} during training week t. The label of a training instance was set to 1 if the two corresponding nodes form a new tie during week $t+1$ in G_{t+1}^{AGG}, and 0 otherwise. Features for the training set were extracted based on networks based on user activities during week t.

A similar approach was used to generate the testing set for the week-t prediction. Testing instances include two-hop node pairs in the aggregated network G_{t+1}^{AGG} constructed for week $t+1$, excluding those that are already connected in the aggregated network G_t^{AGG} for week t. Features for instances in the testing set were extracted from networks based on user activities during week $t+1$. Testing instance labels were based on tie formation during week $t+2$ in G_{t+2}^{AGG}. Figure 1 illustrates the experiment setup for both training and testing.

Throughout the 30 weeks of predictions, there is a total of 21,416 newly formed links in the aggregated network that we tried to predict. Among those links, around 80% of the

Table 1. Network statistics

	G^{BC}	G^{GD}	G^{MB}	G^{PM}	G^{AGG}
$\|V\|$	1516	899	2953	369	3694
$\|E\|$	22706	1418	8873	666	27837
$degree_{mean}$	29.955	3.155	6.009	3.610	15.071
$degree_{max}$	1076	111	756	83	1303
$degree_{std}$	65.590	5.440	27.723	7.739	52.710
C	0.575	0.016	0.185	0.133	0.312
n_{comp}	2	45	20	35	33
r	-0.281	-0.096	-0.342	-0.210	-0.283

[a]The networks are based on user interactions that occurred anytime during the weeks from 50 to 81.

[1]We also considered all ties as unweighted, as adding such weights did not improve prediction results in subsequent experiments.

Figure 1. Experiment setup

links were actually formed in BC, 1% were formed in GD, 33% were formed in MB, and 2% were formed in PM – the percentages do not add up to 100% because one link might be connected in one or more channels . As explained in Table 1, most of the links that we predicted for are from the BC channel, which is also the most active channel in the OHC.

4 Method

4.1 Baseline Features for Multi-Relational Link Prediction

The three baseline features that we used in this experiment all attempt to capture similarity between nodes and have been widely adopted in the link prediction literature.

- Preferential attachment (PA) [5, 8] assumes that nodes with higher degrees tend to be connected, and uses the degree multiplication of two nodes as the similarity value of two nodes.
- Jaccard coefficient (JC) [17] is based on the idea that two nodes that share more common neighbors are more likely to connect. It is the number of common neighbors of two nodes divided by the number of total neighbors of the two nodes.
- Adamic-Adar (AA) [2] extends JC by assigning more weights to common neighbors with a lower degree.

Our multi-relational link prediction (MRLP) approach considers nodal proximity across the four sub-networks and extracts the three baseline features (PA, JC and AA) for each of the communication channels. This yields feature set F_{ALL} that consists of F_{BC}, F_{GD}, F_{MB}, and F_{PM} generated from G^{BC}, G^{GD}, G^{MB}, and G^{PM}, respectively. This design makes it possible for algorithms to learn characteristics of nodal similarity from each channel and leverage information from different types of edges. As a comparison, the baseline model for link prediction does not consider a social network as a multi-relational one and thus extracts the three baseline features for the aggregated network only, yielding feature set F_{AGG}.

4.2 Additional Features

In addition to the three baseline features, we also introduced additional features, namely community-based features, embedding-similarity features, and text-similarity features. We then evaluated if they improve the link prediction performance for the OHC.

4.2.1 Community-Based Features. Community-based features capture if two nodes belong to the same network community. A network community is a subset of nodes that are more densely connected with nodes in the same subset than with nodes outside the subset. The use of network community structure for link prediction is based on the assumption that nodes in the same network community have a higher chance of interacting and then forming ties with each other. Among several community detection algorithms, we selected two computationally efficient ones that automatically pick the number of communities.

- Modularity maximization (*CM*) was proposed by Clauset et al.[11], with a complexity of $O(MD \log N)$, where N represents the number of nodes and M represents the number of edges and D denotes the depth of the dendrogram. We used $C_i \in [1, .., k_{CM}]$, to denote k_{CM} communities detected by this algorithm. If nodes x and y are in the same community C_i, the similarity between x and y is set to 1, and 0 otherwise.

$$s_{xy} = \begin{cases} 1, & \text{if } x, y \in C_i, \ (1 \le i \le k_{CM}) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

- Label propagation (*CLP*) [13] is even more efficient with a complexity of $O(k \cdot M)$, where M represents the number of edges and k represents the number of iterations needed for the algorithm to converge. In this algorithm, each node is initialized with a unique community label. In each iteration, each node gets the majority label of its neighboring nodes, with ties randomly broken. Iterations will stop after a stable set of community labels emerge, which is observed to be after five iterations. The number of communities, k_{CLP}, is also determined by the algorithm. The notion of nodal similarity is the same as the *CM* mentioned above: two nodes in the same community C_i would have a similarity of 1, and 0 otherwise.

In sum, each algorithm added a binary feature for each pair of nodes to indicate if the two nodes are in the same network community. We denoted the feature set that includes

community-based binary features generated on each sub-network as F_{COM}.

4.2.2 Embedding-Similarity Features. Network embedding learns vector representations of nodes in networks [14]. The learned representations can reflect the structural and neighborhood properties of each node. With such vector representations, similarity can be calculated between each pair of nodes.

Among several ways of generating network embedding, we applied DeepWalk [31]. DeepWalk extends word embedding techniques [28] in text mining to networks by considering each node as a word in a corpus. Word embedding using the Skip-Gram model maximizes the probability of having the next word in a corpus, given the sequence of previous words. Likewise, DeepWalk first performs a series of random walks from a source node to produce a set of node sequences. Then, it maximizes the probability of predicting the next node, given previous nodes generated by random walks. Therefore, the learned vector representations of nodes can reflect their neighborhood characteristics as nodes close to each other would have similar vectors.

We used the same set of parameters as in the work of Perozzi et al. [31] across all sub-networks and aggregated networks to ensure a fair comparison. The length of random walks is 40, and 128 is the dimension of embedding vectors. We repeated walks 80 times starting at each node. Once we obtained vector representations for all nodes using DeepWalk, we then computed cosine similarities between two nodes' vectors as their nodal similarity. Specifically, given a node pair x and y, we learned the vector representations V_x and V_y for them and computed their cosine similarity $s_{xy} = cos(V_x, V_y)$. The feature set based on the embedding-similarity features for each channel is denoted as F_{EMB}.

4.2.3 Text-Similarity Features. So far, all features we extracted were based on network structures. As users' interactions are based on the content they generate, we also looked into such user-generated content. We assumed that users who care about similar topics in an OHC may have a higher chance of interacting with each other. Thus we calculated textual similarity among users' posts as a measure of nodal similarity. We considered the posts during the 30 weeks across BC, GD, and MB channels except for PM since the contents of private messages were excluded from this study for privacy concerns.

Texts were pre-processed by removing the stop words, lemmatizing, and stemming. Then, we applied the latent Dirichlet allocation (LDA) [6] model to the posts in three channels for each week to generate the topic distribution of each post based on the weekly posting trend. Note that our focus is topical similarity, which is usually robust against the choice of the number of topics.

For post p published by user x in channel j, during week t, its topic distribution is denoted as T_p^{xjt}. Since our prediction is on a weekly basis, to represent the topic distribution of each user in a specific channel, we averaged the topic distributions of all the user's posts during the corresponding week via the channel. Then, we computed the cosine similarity between the averaged topic distributions for a pair of the user to get text-based nodal similarity. In this way, if the averaged topic distributions of the posts are similar in the same week and the same channel, we regarded that the two users are interested in a similar topic, yielding a high similarity. The text-similarity for nodes x and y in channel j in week t can be computed as follows:

$$s_{xy} = cos(A_x, A_y), \text{ where } A_x = \frac{\sum_{p=1}^{n} T_p^{xjt}}{n} \quad (2)$$

We used feature set F_{TEX} to represent text-similarity features based on each channel.

5 Results

We first compare the prediction performance of different models with different feature sets with four different classifiers. Specifically, we used default parameter settings from scikit-learn package for random forest (RF, 100 trees), logistic regression (LR, L2 regularization), and AdaBoost (AB, 50 trees) in classification. The multi layer perceptron (MLP) was trained with one hidden layer with 32 neurons on the dataset using mini-batch gradient descent with a batch size of 20 for ten epochs by setting aside 20 percent of the data per epoch.

Classification results were then evaluated with precision (PREC) and precision@K (PREC@K) as our goal is to recommend top future links with high accuracy instead of recovering all future links. We also included normalized discounted cumulative gain (nDCG) [22] as an evaluation measure. The metric assigns weights to prediction, so that links that actually formed rank higher than those that did not form.

Link prediction results on baseline model and MRLP, both with three baseline features, are summarized in Table 2. Besides the MRLP approach and the approach base solely on the aggregated network (using feature set F_{AGG}, we also included four model approaches that only used features from each sub-network (using feature sets F_{BC}, F_{GD}, F_{MB}, and F_{PM} respectively). Each value in the table is the average of the prediction results across 30 weeks of the dataset, where the values in bold denote the best performer for each evaluation metric.

When comparing MRLP with F_{AGG}, it is clear that MRLP performs better: its PREC, PREC@10, PREC@20, nDCG@10, and nDCG@20 are 8%, 4%, 1%, 3%, and 1% better than the best performing baseline model, respectively. In other words, considering the effects of each sub-network from a multi-relational perspective works better than considering a single network or aggregating these networks into one. At the same time, using information from only one sub-network is not

as good as using the aggregated network, which combined 4 sub-networks into one. Among the four sub-networks, BC provides the best performance, followed by MB. This is likely because BC and MB included more user activities than the other two channels, which means more information can be learned from the past and more future links will be formed via these two channels as well.

Table 3 illustrates the performance of our multi-relational link prediction approach after additional features were incorporated into the model. The values in the table are the averaged performance across all 30 weeks, as in the previous table. When evaluated with PREC, PREC@10, and nDCG@10, MRLP + F_{EMB} performs the best among the additional features, performing 6%, 2%, and 1% better than MRLP only. Besides, MRLP + F_{COM} performs the best when PREC@20 and nDCG@20 were used for the evaluation, with 2% better performance than MRLP. Because both embedding features and community features capture a node's structural positions beyond its immediate neighbors (as in baseline features), the results suggest that nodal similarities measured at a higher level of network structures are valuable in improving link predictions. Nevertheless, when adding text-based nodal similarities, the performance of link prediction does not see consistent improvement. As adding embedding-similarity features to baseline feature provides the most improvement, we then experimented how embeddings features generated from each channel contribute to link prediction.

Table 2. Results for Baseline vs. MRLP

Metric	CLF	Baseline Approach					MRLP
		F_{AGG}	F_{BC}	F_{GD}	F_{MB}	F_{PM}	F_{ALL}
PREC	RF	0.249	0.229	0.000	0.114	0.070	0.282
	LR	0.466	0.418	0.000	0.315	0.084	0.445
	AB	0.439	0.426	0.000	0.207	0.115	0.388
	MLP	0.511	0.491	0.000	0.325	0.052	**0.551**
PREC @10	RF	0.400	0.300	0.008	0.157	0.038	0.370
	LR	0.617	0.583	0.024	0.380	0.121	**0.640**
	AB	0.507	0.460	0.016	0.267	0.041	0.440
	MLP	0.607	0.567	0.010	0.393	0.077	**0.640**
nDCG @10	RF	0.431	0.300	0.006	0.154	0.057	0.366
	LR	0.634	0.612	0.025	0.392	0.135	**0.655**
	AB	0.500	0.445	0.019	0.259	0.052	0.460
	MLP	0.622	0.594	0.009	0.386	0.089	0.648
PREC @20	RF	0.377	0.300	0.008	0.113	0.031	0.347
	LR	0.603	0.535	0.016	0.328	0.102	0.600
	AB	0.513	0.432	0.008	0.247	0.028	0.463
	MLP	0.597	0.533	0.008	0.318	0.057	**0.610**
nDCG @20	RF	0.405	0.301	0.007	0.124	0.045	0.351
	LR	0.619	0.567	0.019	0.351	0.117	0.622
	AB	0.507	0.432	0.012	0.248	0.039	0.470
	MLP	0.610	0.561	0.008	0.334	0.070	**0.624**

[a]Values that are in bold denote the largest value for each evaluation metric.

From the best-performing MRLP + F_{EMB} model for each performance measure, we excluded one embedding-similarity feature from the corresponding sub-network, and compared the prediction performance with that of MRLP + F_{EMB}. In other words, from the F_{ALL} + F_{EMB} model, the embedding-similarity feature based on BC, GD, MB, and PM channel is excluded once, leading to F_{ALL} + F_{EMB-BC}, F_{ALL} + F_{EMB-GD}, F_{ALL} + F_{EMB-MB}, and F_{ALL} + F_{EMB-PM} respectively. Table 4 summarises the comparison. The *italic* values denote the performances with the highest decrease compared to MRLP + F_{EMB}, whose best-performing classifiers' performance is denoted in bold. These imply that not considering the embedding in the corresponding channel reduces the prediction

Table 3. Performance of additional features on MRLP

Metric	CLF	MRLP	MRLP+More Feature Sets				
		F_{ALL}	F_{ALL} +F_{COM}	F_{ALL} +F_{EMB}	F_{ALL} +F_{TEX}	F_{ALL} +F_{COM} +F_{EMB}	F_{ALL} +F_{COM} +F_{EMB} +F_{TEX}
PREC	RF	0.282	0.290	0.318	0.311	0.323	0.338
	LR	0.445	0.463	0.446	0.448	0.462	0.458
	AB	0.388	0.387	0.389	0.388	0.385	0.376
	MLP	0.551	0.558	**0.584**	0.561	0.462	0.541
PREC @10	RF	0.370	0.373	0.367	0.393	0.360	0.433
	LR	0.640	0.637	**0.650**	0.617	0.640	0.633
	AB	0.440	0.480	0.410	0.487	0.463	0.477
	MLP	0.640	0.597	0.637	0.640	0.607	0.633
nDCG @10	RF	0.366	0.380	0.385	0.401	0.349	0.449
	LR	0.655	0.650	0.656	0.636	0.662	0.641
	AB	0.460	0.494	0.433	0.504	0.480	0.488
	MLP	0.648	0.620	**0.663**	0.646	0.623	0.649
PREC @20	RF	0.347	0.358	0.367	0.365	0.378	0.373
	LR	0.600	**0.622**	0.607	0.598	0.613	0.602
	AB	0.463	0.453	0.463	0.470	0.450	0.475
	MLP	0.610	0.573	0.607	0.600	0.572	0.570
nDCG @20	RF	0.351	0.368	0.379	0.379	0.366	0.400
	LR	0.622	**0.635**	0.623	0.616	**0.635**	0.616
	AB	0.470	0.471	0.463	0.487	0.465	0.484
	MLP	0.624	0.595	0.633	0.615	0.593	0.600

[a]Values that are in bold denote the largest value for each evaluation metric.

Table 4. Contributions of embedding-similarity features from in each sub-network.

Metric	CLF	MRLP+Emb	MRLP+Emb for only 3 channels			
		F_{ALL} +F_{EMB}	F_{ALL} +F_{EMB-BC}	F_{ALL} F_{EMB-GD}	F_{ALL} F_{EMB-MB}	F_{ALL} F_{EMB-PM}
PREC	MLP	**0.584**	*0.528*	0.543	0.551	0.538
PREC@10	LR	**0.650**	0.657	*0.643*	0.647	0.650
nDCG@10	MLP	**0.663**	0.636	*0.618*	0.626	0.659
PREC@20	LR	**0.622**	0.607	*0.603*	0.607	0.615
nDCG@20	LR	**0.635**	0.604	*0.595*	0.599	0.625

[a]Values that are in bold denote the best performing value for F_{ALL} + F_{EMB}.
[b]Values in *Italic* denote the lowest value when the embedding feature for a channel is dropped.

power the most. The comparison suggests that (1) embeddings from each sub-network contributes to the overall link prediction, as removing embedding from a sub-network generally hurts the performance; (2) embeddings from different sub-networks contribute differently, with the removal of embedding-based features from the GD sub-network leading to the greatest deterioration in performance.

6 Conclusions and Future Work

In this paper, we proposed an approach for link predictions in a multi-relational social network in an OHC for smoking cessation. We demonstrated that considering different types of social relationships in a multi-relational social network can improve the performance of link predictions in this context. Sub-networks for BC and MB, the more active channels in the OHC, provide more signals for link prediction in the multi-relational network than the other two channels.

In addition, we showed that looking beyond nodes' immediate neighborhood, and including community structures, as well as nodal similarities based on embedding, could further enhance the performance of our prediction. Among the four sub-networks we investigated, network embeddings generated from the GD channel are the most important contributor to the prediction, which is not captured by the baseline features.

The results have important implications for the design and management of OHCs. Recommender systems that suggest specific pieces of content or community threads may expose users to other members and communication channels they might not otherwise discover. The more that users get connected within the a OHC, the more opportunities there are for relevant, useful exchanges of social support. Specifically, in OHCs for smoking cessation, a recommender system could be developed to recommend content written by other smokers who are in similar situations. For instance, someone experiencing withdrawal symptoms shortly after quitting smoking may derive benefit from connecting with others at the same stage of quitting, who can empathize and share similar experiences. A recommender system based on multi-relational link prediction could recommend a group of users who are going through the same hardships and participating actively in the group discussion. By actively recommending other users with whom the user is predisposed to connect, such a system could increase the social support and information exchanged by reducing the amount of time the user would otherwise spend searching and filtering out potential connections who are not a good match. This is an empirical question that is worthy of future exploration.

There are also interesting future research directions. For example, all of the experiments in this study are based on undirected networks. In the future, we may consider the direction of ties between users Additional analyses on why incorporating text similarity into the model was not useful would be an informative as well. Also, how to better predict stronger ties via PM is worth further investigations. We are also interested in using the deep learning architectures to learn multi-relational network embeddings for better link predictions.

References

[1] 2020 Truth Initiative. [n.d.]. Communication data of users of BecomeAnEx online health community. https://www.becomeanex.org/.

[2] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social networks* 25, 3 (2003), 211–230.

[3] Michael S Amato, George D Papandonatos, Sarah Cha, Xi Wang, Kang Zhao, Amy M Cohn, Jennifer L Pearson, and Amanda L Graham. 2019. Inferring smoking status from user generated content in an online cessation community. *Nicotine and Tobacco Research* 21, 2 (2019), 205–211.

[4] A Bambina. 2007. *Online Social Support: The Interplay of Social Networks and Computer-mediated Communication.* Cambria Press, Youngstown, N.Y.

[5] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[7] Mafalda Burri, Vincent Baujard, and Jean-François Etter. 2006. A qualitative analysis of an internet discussion forum for recent ex-smokers. *Nicotine & Tobacco Research* 8, Suppl_1 (2006), S13–S19.

[8] Hsinchun Chen, Xin Li, and Zan Huang. 2005. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*. IEEE, 141–142.

[9] W. Y. Chou, Y. M. Hunt, E. B. Beckjord, R. P. Moser, and B. W. Hesse. 2009. Social media use in the United States: Implications for health communication. *J Med Internet Res* 11, 4 (2009), e48. https://doi.org/10.2196/jmir.1249 4.

[10] Katherine Y. Chuang and Christopher C. Yang. 2014. Informational support exchanges using different computer-mediated communication formats in a social media alcoholism community. *Journal of the Association for Information Science and Technology* 65, 1 (Jan. 2014), 37–52. https://doi.org/10.1002/asi.22960

[11] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

[12] Nathan K Cobb, Amanda L Graham, and David B Abrams. 2010. Social network structure of a large online community for smoking cessation. *American journal of public health* 100, 7 (2010), 1282–1289.

[13] Gennaro Cordasco and Luisa Gargano. 2010. Community detection via semi-synchronous label propagation algorithms. In *2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA)*. IEEE, 1–8.

[14] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. 2018. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering* (2018).

[15] Brian G Danaher, Keith Smolkowski, John R Seeley, and Herbert H Severson. 2008. Mediators of a successful web-based smokeless tobacco cessation program. *Addiction* 103, 10 (2008), 1706–1712.

[16] Darcy Davis, Ryan Lichtenwalter, and Nitesh V Chawla. 2011. Multi-relational link prediction in heterogeneous information networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 281–288.

[17] Société Vaudoise des Sciences Naturelles. 1864. *Bulletin de la Société vaudoise des sciences naturelles*. Vol. 7. F. Rouge.

[18] Amanda L Graham, George D Papandonatos, Sarah Cha, Bahar Erar, and Michael S Amato. 2018. Improving adherence to smoking cessation

treatment: smoking outcomes in a web-based randomized trial. *Annals of Behavioral Medicine* 52, 4 (2018), 331–341.

[19] Amanda L. Graham, Kang Zhao, George D. Papandonatos, Bahar Erar, Xi Wang, Michael S. Amato, Sarah Cha, Amy M. Cohn, and Jennifer L. Pearson. 2017. A prospective examination of online social network dynamics and smoking cessation. *PLOS ONE* 12, 8 (Aug. 2017), e0183655. https://doi.org/10.1371/journal.pone.0183655

[20] Amanda L Graham, Kang Zhao, George D Papandonatos, Bahar Erar, Xi Wang, Michael S Amato, Sarah Cha, Amy M Cohn, and Jennifer L Pearson. 2017. A prospective examination of online social network dynamics and smoking cessation. *PloS one* 12, 8 (2017).

[21] Julianne Holt-Lunstad, Timothy B. Smith, and J. Bradley Layton. 2010. Social Relationships and Mortality Risk: A Meta-analytic Review. *PLOS Medicine* 7, 7 (July 2010), e1000316. https://doi.org/10.1371/journal.pmed.1000316

[22] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/345508.345545

[23] Zhepeng Lionel Li, Xiao Fang, and Olivia R Liu Sheng. 2018. A survey of link recommendation for social networks: methods, theoretical foundations, and future research directions. *ACM Transactions on Management Information Systems (TMIS)* 9, 1 (2018), 1.

[24] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology* 58, 7 (2007), 1019–1031.

[25] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. 2010. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 243–252.

[26] Xiaoxiao Ma, Guanling Chen, and Juntao Xiao. 2010. Analysis of an online health social network. In *Proceedings of the 1st ACM international health informatics symposium*. ACM, 297–306.

[27] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[29] Sahiti Myneni, Kayo Fujimoto, Nathan Cobb, and Trevor Cohen. 2015. Content-driven analysis of an online community for smoking cessation: integration of qualitative techniques, automated text analysis, and affiliation networks. *American journal of public health* 105, 6 (2015), 1206–1212.

[30] Priya Nambisan. 2011. Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association* 18, 3 (2011), 298–304.

[31] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[32] Bernd Ploderer, Wally Smith, Steve Howard, Jon Pearce, and Ron Borland. 2013. Patterns of support in an online community for smoking cessation. In *Proceedings of the 6th International Conference on Communities and Technologies*. 26–35.

[33] Kenneth Portier, Greta E. Greer, Lior Rokach, Nir Ofek, Yafei Wang, Prakhar Biyani, Mo Yu, Siddhartha Banerjee, Kang Zhao, Prasenjit Mitra, and John Yen. 2013. Understanding Topics and Sentiment in an Online Cancer Survivor Community. *JNCI Monographs* 2013, 47 (2013), 195–198. http://jncimono.oxfordjournals.org/content/2013/47/195.abstract

[34] Baojun Qiu, Kang Zhao, Prasenjit Mitra, Dinghao Wu, Cornelia Caragea, John Yen, Greta E. Greer, and Kenneth Portier. 2011. Get online support, feel better–Sentiment analysis and dynamics in an online cancer survivor community. In *Proceedings of the Third IEEE International Conference on Social Computing (SocialCom'11)*. Boston, MA, 274–281.

[35] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2016. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2016), 17–37.

[36] Lei Tang and Huan Liu. 2009. Scalable learning of collective behavior based on sparse social dimensions. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 1107–1116.

[37] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. Acm, 1100–1108.

[38] Xi Wang and Gita Sukthankar. 2013. Link prediction in multi-relational collaboration networks. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*. ACM, 1445–1447.

[39] Xi Wang, Kang Zhao, Sarah Cha, Michael S. Amato, Amy M. Cohn, Jennifer L. Pearson, George D. Papandonatos, and Amanda L. Graham. 2019. Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach. *Decision Support Systems* 116 (Jan. 2019), 26–34. https://doi.org/10.1016/j.dss.2018.10.005

[40] Xi Wang, Kang Zhao, and Nick Street. 2017. Analyzing and Predicting User Participations in Online Health Communities: A Social Support Perspective. *Journal of Medical Internet Research* 19, 4 (2017), e130. https://doi.org/10.2196/jmir.6834

[41] Elissa R. Weitzman, Emily Cole, Liljana Kaci, and Kenneth D. Mandl. 2011. Social but safe? Quality and safety of diabetes-related online social networks. *Journal of the American Medical Informatics Association* 18, 3 (May 2011), 292–297. https://doi.org/10.1136/jamia.2010.009712

[42] Ronghua Xu and Qingpeng Zhang. 2016. Understanding Online Health Groups for Depression: Social Network and Linguistic Perspectives. *Journal of Medical Internet Research* 18, 3 (March 2016). https://doi.org/10.2196/jmir.5042

[43] Christopher C. Yang and Ling Jiang. 2018. Enriching User Experience in Online Health Communities Through Thread Recommendations and Heterogeneous Information Network Mining. *IEEE Transactions on Computational Social Systems* 5, 4 (Dec. 2018), 1049–1060. https://doi.org/10.1109/TCSS.2018.2879044

[44] Christopher C. Yang and Haodong Yang. 2018. Mining heterogeneous networks with topological features constructed from patient-contributed content for pharmacovigilance. *Artificial Intelligence in Medicine* 90 (Aug. 2018), 42–52. https://doi.org/10.1016/j.artmed.2018.07.002

[45] Yang Yang, Nitesh Chawla, Yizhou Sun, and Jiawei Hani. 2012. Predicting links in multi-relational and heterogeneous networks. In *2012 IEEE 12th international conference on data mining*. IEEE, 755–764.

[46] Colleen Young. 2013. Community management that works: how to build and sustain a thriving online health community. *Journal of medical Internet research* 15, 6 (2013), e119.

[47] Kang Zhao, Xi Wang, Sarah Cha, Amy M Cohn, George D Papandonatos, Michael S Amato, Jennifer L Pearson, and Amanda L Graham. 2016. A Multirelational Social Network Analysis of an Online Health Community for Smoking Cessation. *Journal of Medical Internet Research* 18, 8 (Aug. 2016), e233. https://doi.org/10.2196/jmir.5985

[48] Kang Zhao, John Yen, Louis-Marie Ngamassi, Carleen Maitland, and Andrea H Tapia. 2012. Simulating inter-organizational collaboration network: a multi-relational and event-based approach. *Simulation* 88, 5 (2012), 617–633.