


Multi-Relational Link Prediction for an Online Health Community

Sulyun Lee, Hankyu Jang, Kang Zhao,
Michael S. Amato, Amanda L. Graham



truth initiative[®]
INSPIRING TOBACCO-FREE LIVES

 **THE UNIVERSITY OF IOWA**[®]
Interdisciplinary Graduate Program in Informatics
Department of Computer Science
Department of Business Analytics

Motivation



Motivation

Online Health Community (OHC)

- OHC enables social networking
 - Similar health concerns
- Various communication channels
 - Blog posts & comment
 - Group discussion
 - Message board
 - Private message

Q: How can users of OHC achieve better cessation outcome?



Motivation

Multi-Relational Link Prediction

- **Networking** is a key to a better cessation outcome
- Can we utilize users' communication data for better networking?
 - Generate a network from each channel (**multi-relational network**)
 - Recommend friends to users that share similar interests utilizing information from each network (**link prediction**)



Better Networking

Better Engagement

Better Cessation
Outcome

Data & Setup

Data

- Source: **BecomeAnEx** - OHC for smoking cessation
- 6 years of users' interaction (2010 - 2015) in 4 channels
 - Blog posts & comment (**BC**)
 - Group discussion (**GD**)
 - Message board (**MB**)
 - Private message (**PM**)

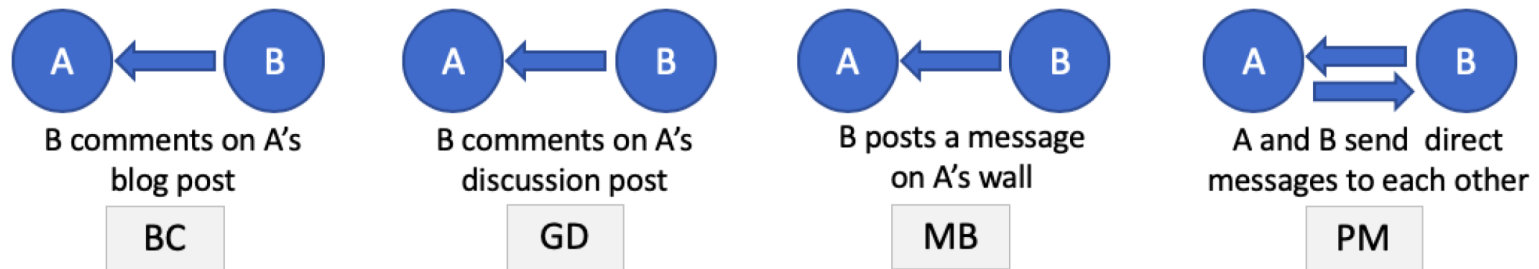


Figure 1: Four communication channels in BecomeAnEx

Data & Setup

Communication Network

- Four subnetworks: one undirected network for each channel
 - G_{BC} , G_{GD} , G_{MB} , G_{PM}
- One aggregated network (G_{AGG})
- Both support seekers and providers can benefit from participations in an OHC
- 32 consecutive weeks were considered

Table 1: Network statistics

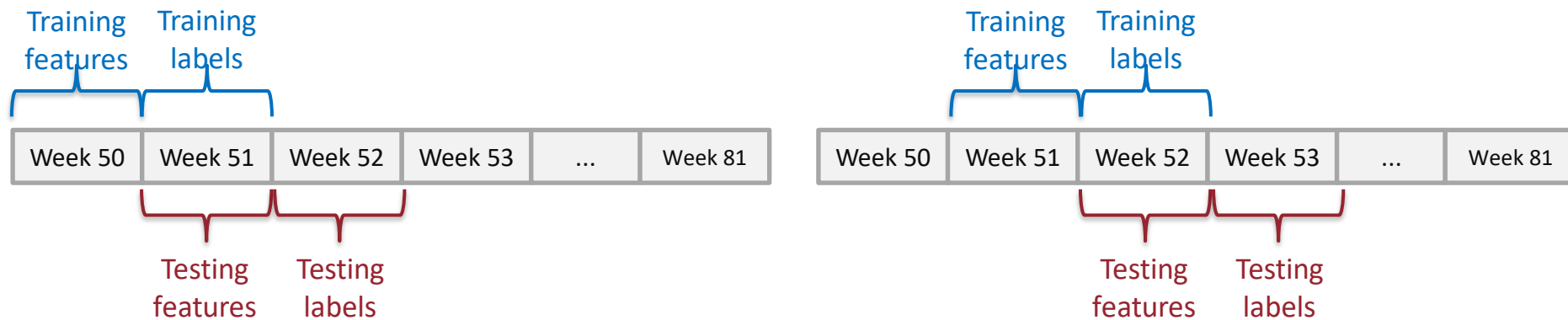
	G_{BC}	G_{GD}	G_{MB}	G_{PM}	G_{AGG}
Number of nodes	1516	899	2953	369	3694
Number of edges	22706	1418	8873	666	27837
Average degree	29.955	3.155	6.009	3.610	15.071
Maximum degree	1076	111	756	83	1303
Degree standard deviation	65.590	5.440	27.723	7.739	52.710
Clustering coefficient	0.575	0.016	0.185	0.133	0.312
Number of connected components	2	45	20	35	33
Assortativity	-0.281	-0.096	-0.342	-0.210	-0.283

^aThe graphs are constructed with the user interactions that occurred anytime during the weeks from 50 to 81.

Data & Setup

Setup

- Task:
 - Predict the next week's newly formed network ties, no matter which channel
 - Leverage the current week's network structures
- A sliding window approach:



[Supervised Link Prediction]

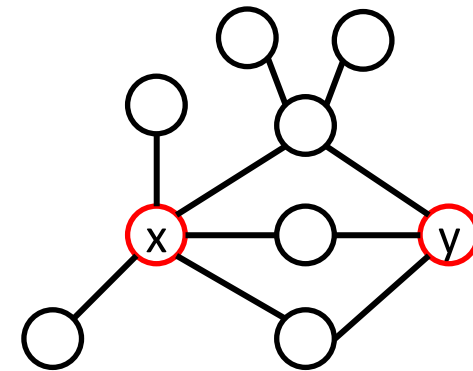
- **Instances:** Pairs of users in the G_{Agg} for week t
- **Features:** Similarity scores computed for the pair of user for week t
- **Labels:** Binary **1**: if the pair is connected in the G_{Agg} for week $t+1$
0: otherwise

Features (Similarity Measures)

Neighbor-based

- Jaccard's coefficient $\frac{|\tau(x) \cap \tau(y)|}{|\tau(x) \cup \tau(y)|}$
 - Idea: penalizing nonshared neighbor

Symbol	Definition
$\tau(x)$	Neighbor of x



$$\text{score}_{\text{JC}}(x, y) = \frac{3}{5}$$

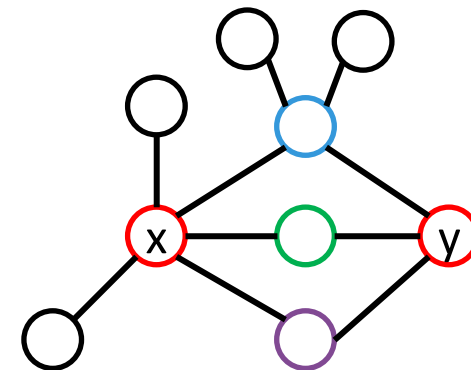


Features (Similarity Measures)

Neighbor-based

- Jaccard's coefficient $\frac{|\tau(x) \cap \tau(y)|}{|\tau(x) \cup \tau(y)|}$
 - Idea: penalizing nonshared neighbor
- Adamic Adar $\sum_{z \in \tau(x) \cap \tau(y)} \frac{1}{\log k(z)}$
 - Idea: penalizing 'shared neighbor' that has many neighbors

Symbol	Definition
$\tau(x)$	Neighbor of x
z	Common neighbor of x and y
$k(z)$	Degree of z



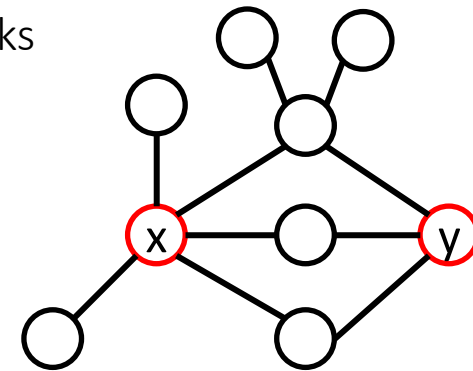
$$\text{score}_{AA}(x, y) = \frac{1}{\log 4} + \frac{1}{\log 2} + \frac{1}{\log 2}$$

Features (Similarity Measures)

Neighbor-based

- Jaccard's coefficient $\frac{|\tau(x) \cap \tau(y)|}{|\tau(x) \cup \tau(y)|}$
 - Idea: penalizing nonshared neighbor
- Adamic Adar $\sum_{z \in \tau(x) \cap \tau(y)} \frac{1}{\log k(z)}$
 - Idea: penalizing 'shared neighbor' that has many neighbors
- Preferential attachment $k(x) \times k(y)$
 - Idea of "the rich get richer" in scale-free networks

Symbol	Definition
$\tau(x)$	Neighbor of x
z	Common neighbor of x and y
$k(z)$	Degree of z



$$\text{score}_{\text{PA}}(x, y) = 5 \times 3$$

P. Jaccard. "Etude comparative de la distribution florale dans une portion des alpes et des jura", Bulletin de la Societe Vaudoise des Sciences Naturelles 1901

L. A. Adamic, E. Adar, "Friends and neighbors on the Web," Social Networks 03

A. L. Barabasi, H Jeong, Z Neda et al, "Evolution of the social network of scientific collaborations", Physica A 02

M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions", Internet Mathematics 2004

Methods

Baseline & MRLP (Proposed Method)

- Baseline features extract 3 features (JC, AA, PA) for each of the 4 subnetworks and the aggregated network
 - F_{BC} : 3 baseline features from G_{BC}
 - F_{GD} : 3 baseline features from G_{GD}
 - F_{MB} : 3 baseline features from G_{MB}
 - F_{PM} : 3 baseline features from G_{PM}
 - F_{Agg} : 3 baseline features from G_{Agg}
- MRLP considers a social network as multi-relational by stacking F_{BC} , F_{GD} , F_{MB} , F_{PM} (12 features in total)
 - F_{ALL} : $F_{BC} + F_{GD} + F_{MB} + F_{PM}$

Methods

Performance Measures

- Classifiers
 - Random Forest, Logistic Regression, AdaBoost, Neural Network
- Evaluation Measure
 - Precision, Precision@k
 - Goal is to recommend top future links with high accuracies
 - Normalized discounted cumulative gain (nDCG@k)
 - It's important to rank links that occurred higher than those that did not occur

Results

Baseline vs. MRLP

Table 2: Results for Baseline vs. MRLP

Metric	Classifier	Baseline Approach					MRLP
		F_{AGG}	F_{BC}	F_{GD}	F_{MB}	F_{PM}	F_{ALL}
Precision	Random Forest	0.249	0.229	0.000	0.114	0.070	0.282
	Logistic Regression	0.466	0.418	0.000	0.315	0.084	0.445
	AdaBoost	0.439	0.426	0.000	0.207	0.115	0.388
	Neural Network	0.511	0.491	0.000	0.325	0.052	0.551
	Random Forest	0.400	0.300	0.008	0.157	0.038	0.370
	Logistic Regression	0.617	0.589	0.024	0.280	0.121	0.640

MRLP

Considering features
from each network

>

Baseline

Considering a single
network or an
aggregated network

PREC@20	Random Forest	0.511	0.500	0.000	0.119	0.091	0.511
	Logistic Regression	0.603	0.535	0.016	0.328	0.102	0.600
	AdaBoost	0.513	0.432	0.008	0.247	0.028	0.463
	Neural Network	0.597	0.533	0.008	0.318	0.057	0.610
nDCG@20	Random Forest	0.405	0.301	0.007	0.124	0.045	0.351
	Logistic Regression	0.619	0.567	0.019	0.351	0.117	0.622
	AdaBoost	0.507	0.432	0.012	0.248	0.039	0.470
	Neural Network	0.610	0.561	0.008	0.334	0.070	0.624

^aValues that are in bold denote the largest value for each evaluation metric.



Can we do better?

More Features

- F_{COM} : Community-based features
 - Modularity maximization (4 features, each from G_{BC} , G_{GD} , G_{MB} , G_{PM})
 - Label propagation (4 features, each from G_{BC} , G_{GD} , G_{MB} , G_{PM})
- F_{EMB} : Embedding-similarity
 - DeepWalk (4 features, each from G_{BC} , G_{GD} , G_{MB} , G_{PM})
- F_{TEX} : Text-similarity
 - Latent Dirichlet allocation (LDA, 3 features, each from G_{BC} , G_{GD} , G_{MB})

More Features

Community-Based Feature (F_{COM})

- Idea: Two nodes are similar if they belong to the same community
 - Modularity maximization

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

Nodes that are in same community

Number of communities detected using modularity maximization

- Label propagation

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

Number of communities detected using label propagation

More Features

Embedding-Similarity Feature (F_{EMB})

- Idea: Two nodes with similar representations are similar
- How to learn representation of nodes in the graph?
 - DeepWalk

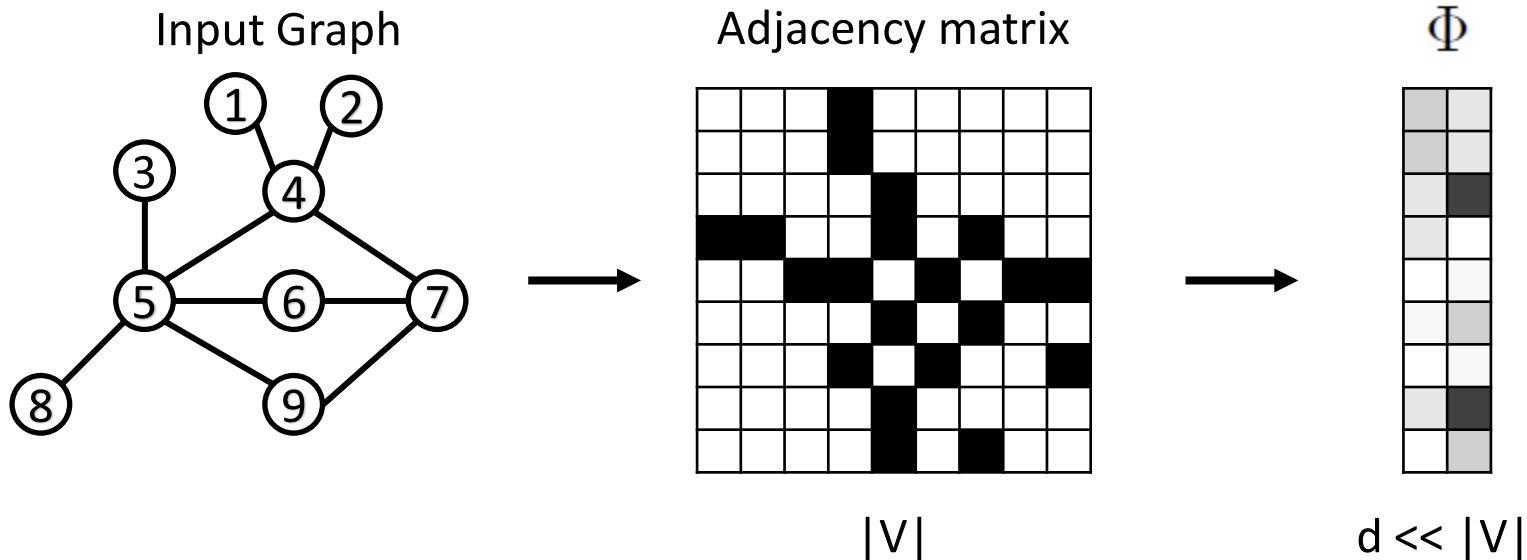


More Features

Embedding-Similarity Feature (F_{EMB})

- Idea: Skip-gram (word embedding) - Learn a vector representation of word such that nearby words would have similar representation
- Input: $G = (V, E)$
- Output: $\Phi : v \in V \rightarrow \mathbb{R}^{|V| \times d}$

DeepWalk



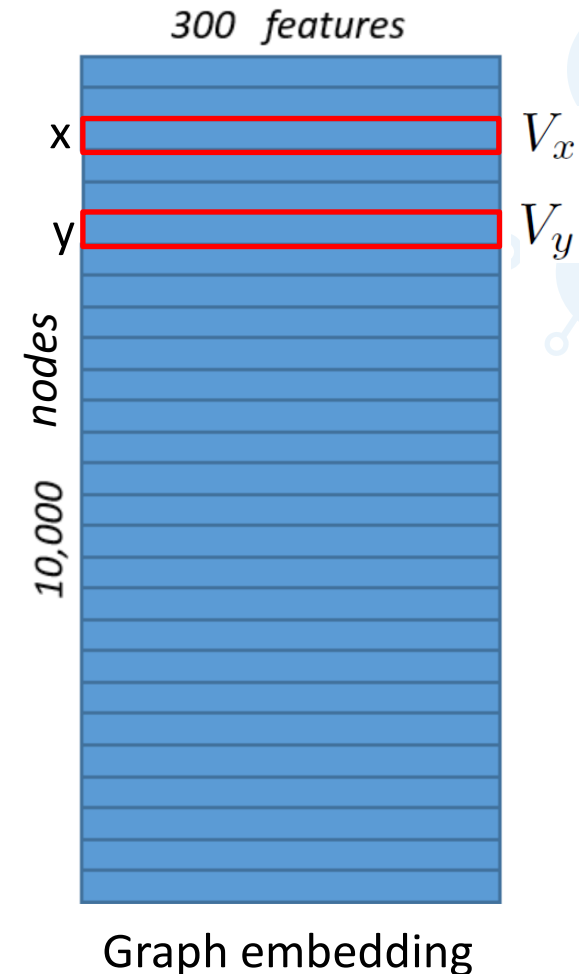
More Features

Embedding-Similarity Feature (F_{EMB})

- After the embedding is learned, compute cosine similarity of the two vectors

$$s_{xy} = \text{COS}(V_x, V_y)$$

- Nodes that have similar embedding have higher score



More Features

Text-Similarity Feature (F_{TEX})

- Idea: Users who care about similar topics in an OHC may have a higher chance of interacting with each other
- Compute text similarity among users' posts as a measure of similarity

More Features

Text-Similarity Feature (F_{TEX})

- Combined posts across 30 weeks
- Applied latent Dirichlet allocation (LDA)
 - Each post has a topic distribution (30 dimensional vector)
 - If a user posted n posts in a channel for a week, then the topic distribution is averaged over n topic distributions
- Two users have similar topic distributions if they expressed interest in similar topics with each other

$$s_{xy} = \cos(A_x, A_y), \text{ where } A_i = \frac{\sum_{p=1}^n T_{ijtp}}{n}$$

Topic distribution for

- user i
- channel j
- week t
- post p
- number of posts n

Results

MRLP vs. MRLP + more features

$$F_{ALL} : F_{BC} + F_{GD} + F_{MB} + F_{PM}$$

F_{COM} : Community-based feature

F_{EMB} : Embedding-similarity feature

F_{TEX} : Text-similarity feature

Table 3: Performance of additional features on MRLP

Metric	Classifier	MRLP	MRLP+More Feature Sets				
		F_{ALL}	$F_{ALL} + F_{COM}$	$F_{ALL} + F_{EMB}$	$F_{ALL} + F_{TEX}$	$F_{ALL} + F_{COM} + F_{EMB}$	$F_{ALL} + F_{COM} + F_{EMB} + F_{TEX}$
Precision	Random Forest	0.282	0.290	0.318	0.308	0.323	0.338
	Logistic Regression	0.445	0.463	0.446	0.448	0.462	0.458
	AdaBoost	0.388	0.387	0.389	0.386	0.385	0.376
	Neural Network	0.551	0.558	0.584	0.516	0.462	0.541
PREC@10	Random Forest	0.370	0.373	0.367	0.437	0.360	0.433
	Logistic Regression	0.640	0.637	0.650	0.623	0.640	0.633
	AdaBoost	0.440	0.480	0.410	0.487	0.463	0.477
	Neural Network	0.640	0.597	0.637	0.627	0.607	0.633
nDCG@10	Random Forest	0.366	0.380	0.385	0.467	0.349	0.449
	Logistic Regression	0.655	0.650	0.656	0.642	0.662	0.641
	AdaBoost	0.460	0.494	0.433	0.500	0.480	0.488
	Neural Network	0.648	0.620	0.663	0.629	0.623	0.649
PREC@20	Random Forest	0.347	0.358	0.367	0.382	0.378	0.373
	Logistic Regression	0.600	0.622	0.607	0.580	0.613	0.602
	AdaBoost	0.463	0.453	0.463	0.497	0.450	0.475
	Neural Network	0.610	0.573	0.607	0.577	0.572	0.570
nDCG@20	Random Forest	0.351	0.368	0.379	0.417	0.366	0.400
	Logistic Regression	0.622	0.635	0.623	0.604	0.635	0.616
	AdaBoost	0.470	0.471	0.463	0.503	0.465	0.484
	Neural Network	0.624	0.595	0.633	0.592	0.593	0.600

^aValues that are in bold denote the largest value for each evaluation metric.

Conclusion

Takeaways

- Our MRLP method of utilizing multi-relational information outperformed baseline methods
- Embedding-similarity feature further improved the performance as well as community-based features
- Adding text-similarity features did not improve the performance

Discussions

- Important implications for the design and management of an OHC
 - Recommend friends to users
 - Recommend users to read blog posts
 - Recommend users to participate in group discussions
- OHC related studies show that better engagement of users in OHC helps them to achieve their goals



X. Ma, G. Chen, J. Xiao, "Analysis of an online health social network," IHI 10

S. Myneni, K. Fujimoto, N. Cobb, T. Cohen, "Content-driven analysis of an online community for smoking cessation: integration of qualitative techniques, automated text analysis, and affiliation networks," AJP 15

P. Nambisan, "Information seeking and social support in online health communities: impact on patients' perceived empathy," JAMIA 11

Questions?

Sulyun Lee: sulyun-lee@uiowa.edu

Hankyu Jang: hankyu-jang@uiowa.edu

Kang Zhao: kang-zhao@uiowa.edu



truth initiative[®]
INSPIRING TOBACCO-FREE LIVES

 **THE UNIVERSITY OF IOWA**[®]
Interdisciplinary Graduate Program in Informatics
Department of Computer Science
Department of Business Analytics