Evaluations of Similarity Measures on VK for Link Prediction

JooYoung Lee¹ · Rustam Tukhvatov¹

Received: 31 August 2017 / Revised: 30 March 2018 / Accepted: 3 September 2018 / Published online: 14 September 2018
© The Author(s) 2018

Abstract

Recommender system is one of the most important components for many companies and social networks such as Facebook and YouTube. A recommendation system consists of algorithms which allow to predict and recommend friends or products. This paper studies to facilitate finding like-minded people with same interests in social networks. In our research, we used real data from the most popular social network in Russia, VK (Vkontakte). The study is motivated on the assumption that similarity breeds connection. We evaluate well-known similarity measures in the field on our collected VK datasets and find limited performance results. The result shows that majority of users in VK tend not to add possible users with whom they have common acquaintances. We also propose a topology-based similarity measure to predict future friends. Then, we compare our results with the results of other well-known methods and discuss differences.

Keywords Recommender system · Graph theory · Link prediction · Similarity

1 Introduction

Social networks help people to interact through Internet. Nowadays social network services allow to share interests via texts, music, video, etc. Clearly, people want to get more information and new contacts as fast as possible, and the information should be relevant to their preferences.

Link recommendation has become one of the most important features in online social networks and has been an active research area [3, 8–10, 14, 16]. There are well-known examples of link recommendation such as “People You May Want to Hire” on LinkedIn, “You May Know” on Google+ and “People You May Know” on Facebook. Given the tremendous academic and practical interests in link recommendation [6], we examined existing approaches and applied recommendation engine for VK social network.

We especially focus on the most popular similarity measures which are used in the literature, namely, cosine similarity, common neighbors, Jaccard similarity and

Adamic–Adar. We show the performance of each metric using confusion matrix plus F1 measure which is the standard baselines. We examine each metric on a set of small online social networks as well as collected VK datasets.

A social network is a graph as a data structure, the users are nodes, and the users’ friendships (relations) are edges [11]. We aim to understand factors which may affect the emergence of new edges and try to predict future connections in social networks. In Sect. 2, we introduce some of major existing methods on link prediction. In Sect. 3, we present existing methods which we test and compare our results with on VK datasets. In Sect. 4, we describe in detail about our experiment settings and datasets which we collected and preprocess. Section 5, then, presents how we can evaluate the results as well as the evaluation of the most well-known methods. In Sect. 6, we show our results and discuss implications. Finally, in Sect. 7, we conclude our research.

2 Related Work

Wu et al. [17] study understanding users’ temporal behaviors in social network platforms. The evolution of social network services is driven by the interplay between users’ preferences and social network structures. Authors argue that users’ future preference behavior is affected by the network

✉ JooYoung Lee
j.lee@innopolis.ru
Rustam Tukhvatov
r.tukhvatov@innopolis.ru

¹ Network and Information Science Laboratory, Institute of Information Systems, Innopolis University, Innopolis, Russia 420500

around them and the homophily effect. In [1], authors demonstrate the developed Social Poisson Factorization (SPF), a probabilistic model that incorporates social network information into a traditional factorization method. SPF introduces the social aspect of algorithmic recommendation. They develop a scalable algorithm for analyzing data with SPF, and demonstrate that it outperforms competing methods on six real-world datasets.

Zhang et al. [18] studied the social influence problem in a large microblogging network, Weibo.com.¹ They investigate (re)tweet behaviors of users by considering ego network of users. In our proposed method, we expand another step by considering second level degrees of users. Kooti et al. [5] demonstrate online consumers' behavior and try to explain ways to improve ad (advertising) targeting systems. Researchers use information in emails such as purchase logs and communications between users to find patterns and their interaction. For example, authors measured the effect of gender and age which showed that a female email user is more likely to be an online shopper than an average male email user. Also the spending ability goes up with the age until age of 30, then stabilizes in the early 60s and starts to drop afterward. Such findings help to predict future customers' behavior and make purchases more pleasant for consumers.

Measuring similarity between nodes is the main task in link prediction problem. In [15], authors constructed a new way to measure similarity between nodes based on game-theoretic interaction index. The basic form of the interaction index is built upon two solution concepts from game theory: the Shapley value and the Banzhaf index. It is also generalized to a wider class of solution concepts: Shapley value and Banzhaf index. Authors showed that using their approach, it is possible to improve existing results in link prediction and community detection problems.

In this paper, we consider link prediction problems, and more precisely, we discuss similarity measures for link prediction problems. One of the well-accepted hypotheses is that similar users will become future friends. Therefore, it is essential to measure similarity between users so that future links can be accurately predicted. As we will discuss throughout the paper, there exist many metrics to measure the similarity. Once the similarity is measured, there are many ways to predict future connections. One intuitive way is to have a threshold and predict users who have similarities greater than the given threshold. Other simple ways include predicting top k users and top 1% of users. More advanced techniques include learning how much to predict using machine learning algorithms.

2.1 VKontakte

In this section, we focus on the datasets which we collected over 6 months from VK network. VK (VKontakte, meaning InContact) is the largest European online social network and social media service. According to SimilarWeb,² VK is the fifth most popular website in the world. As of January 2017, daily average audience is about 87 million visitors,³ with more than 410 million registered users.⁴

Any user in VK has a profile which contains various information. First and last names are mandatory fields, and other data such as birthday, city and interests are optional. Figure 1 is an example of a profile of the social network. In addition, users can share interesting information via posting on the profile wall as well as making repost from other users or groups,⁵ and each post may contain attachments—documents or media files.

VK has open APIs for application developments. It allows developers to access the server through requests. The APIs provide the ability to access some user information with his/her consent, such as photographs, friends, profile and wall.

3 Methods

The link prediction problem is connected with network structure. In our research, we use similarity between users using information from their second level degrees.

3.1 Problem Statement

Given a snapshot of a social network $G^t = (V^t, E^t)$ where V is a set of users and E is a set of relationships, we aim to find a relationship $(v_i, v_j) \notin E^t$ and $(v_i, v_j) \in E^{t'}$ where $t < t'$.

The link prediction problem is the problem of predicting future connections in a network which may form in the network. Here, we briefly list some measures which we examine in this study.

$$\text{Cosine Similarity}(A, B) = \cos(\theta) = \frac{f(A)f(B)}{|f(A)||f(B)|} \quad (1)$$

² SimilarWeb Ltd. is a digital market intelligence company that provides web analytics, data analysis and business intelligence services for international corporations. It uses large data processing technologies to collect, measure, analyze and provide insights on behavioral patterns and statistics for the involvement of users from websites and mobile applications.

³ "Audience VKontakte" www.liveinternet.ru. Retrieved January 2017.

⁴ "List of VK users" [Vk.com](http://vk.com). Retrieved January 2017.

⁵ The VK group is a community where people can communicate with each other, exchanging ideas and proposals.

¹ The most popular Chinese microblogging service.

Fig. 1 VK profile

$$\text{Common Neighbors } (A, B) = f(A) \cap f(B) \quad (2)$$

$$\text{Jaccard Similarity } (A, B) = \frac{|f(A) \cap f(B)|}{|f(A) \cup f(B)|} \quad (3)$$

$$\text{Adamic Adar Index } (A, B) = \sum_{w \in f(A) \cap f(B)} \frac{1}{\log |f(w)|} \quad (4)$$

where $f(w)$ denotes the set of neighbors of w .

3.2 Second Common Neighbor Similarity

We propose a new similarity measure between users using second degrees [7, 12]. Second neighbors are defined as the nodes which are connected in the second hops.

$$\text{Second Common Neighbors } (A, B) = \left| \bigcup_{i \in f(A)} f(i) \cap \bigcup_{j \in f(B)} f(j) \right|$$

where $f()$ is user's friends, $f(A)$ friend of user A , $f(i)$ is the set of neighbors of i .

(5)

We propose another simple metric to compare with the proposed similarity measure which is based on the shortest distance between users.

$$\text{Shortest Path Index } (A, B) = \text{Length of Shortest Path Between } A \text{ and } B \quad (6)$$

4 Experiments

Ideally, the evaluation of link prediction algorithms should take place every timestamp to check the performance and adjust the algorithm if necessary. However, such an approach takes a long period of time and it is hard to track of incoming and outgoing users from the initial network. Therefore, it is common practice to delete a small portion of edges from the network as if they do not exist and then try to predict the deleted edges. This static approach has possible flaws since the similarity between users is not static; thus, deleted edges are easier to predict than future edges.

	id	personal	friends	wallPosts
0	2064386.0	[{'uid': 2064386, 'country': 1, 'counters': {'...}}	[2772.0, 4987.0, 8680.0, 11380.0, 11698.0, 132...	[111199386, 'Левитанский Ю.Д. Дети <b...
1	63802027.0	[{'home_town': 'Москва', 'movies': '', 'about'...	[701.0, 77777.0, 88484.0, 196957.0, 204655.0, ...	[0, 'Ура! С днем варенья, дружище! И пускай вс...
2	241284438.0	[{'home_town': '', 'movies': '', 'photo_50': '...	[36417.0, 84922.0, 125032.0, 146957.0, 211166....	[8095605, '', 4698690, '', 4247839, '', 177117...
3	7.0	[{'uid': 7, 'counters': {'followers': 5902, 'v...	[6.0, 9.0, 10.0, 14.0, 17.0, 21.0, 22.0, 23.0,....	[]
4	262153.0	[{'uid': 262153, 'country': 1, 'counters': {'u...	[651.0, 1174.0, 3775.0, 5068.0, 7720.0, 10437....	[0, '', -32179668, 'Новогоднее печенье Пече...

Fig. 2 Extracted data from VK

Table 1 VK networks with different time stamps

Date	Number of nodes	Number of edges	Average degree
23.12.2016	942469	1475811	3.1318
24.01.2017	950503	1490873	3.1370
23.02.2017	936847	1465214	3.1280
24.03.2017	939985	1469854	3.1274

In the experiment, we compare the common practice with the real link prediction based on actual future links without deleting edges. To be able to evaluate the accuracy, we collected datasets from the same set of users in different time intervals to build a learning model and then compare the predictions.

4.1 Collection of Dataset

We collected several snapshots of networks from VK to test link prediction methods. In order to make comparisons easier, we preprocess the data so that the resulting networks contain the same users since newly joined users in latter networks are not predictable. In the next step, we exclude users who are suspected to be not regular individuals such as groups, commercial accounts and celebrities. We removed users with more than 500 friends based on the mean number of friends in VK which was 240. Datasets contain users' profile information, friends and posts from their walls as shown in Fig. 2.

We use VK networks with four different timestamps as summarized in Table 1.

We further processed the data by removing nodes whose degree is less than two and newly joined nodes that did not exist in the starting graph. The processed networks are summarized in Table 2.

4.2 Implementation

We used Python programming language and developed two variants of the implementation. In the first variant, we used *Pandas* library to collect and process the dataset, and then, we wrote codes for the similarity measures, i.e., cosine similarity, common neighbors, Jaccard similarity and Adamic–Adar index.

In the second variant, we used Jaccard and Adamic–Adar similarity measures from *Networkx* library to improve the speed of our implementation.

– Python 3.5.2

Python is a high-level general-purpose programming language, focused on improving developer productivity and code readability. The syntax of the Python kernel is minimal, and at the same time, the standard library includes a large amount of useful functions.

– IPython 5.1.0

IPython is an interactive computational environment for the Python programming language, in Jupyter Note-

Table 2 VK networks after preprocessing

Date	Number of nodes	Number of edges	Average degree	Excess nodes
23.12.2016	190,550	712,014	7.4733	7412
24.01.2017	190,550	716,478	7.5201	5459
23.02.2017	190,550	712,196	7.4752	1083
24.03.2017	190,550	713,030	7.4839	340

book 4.3.1 shell. It is an open-source program and one of the popular tools for data scientists.

4.2.1 Python Libraries

– *Pandas*

Pandas library makes Python a powerful tool for data analysis. It is an open-source, easy-to-use, BSD-licensed. The package allows to build summary tables, performs a grouping, provides easy access to tabular data and has robust IO tools for loading and saving data from flat files (CSV and delimited). The library is well suited for different kinds of data: forms of statistical or observational datasets, arbitrary matrix data with row and column labels, and it might be heterogeneous or homogeneously typed, tabular data like Excel spreadsheet. *Pandas* package provides high performance because many low-level algorithms are tweaked in Cython code. Cython language is a superset of Python that allows the compiler to generate very efficient C codes from Python.

– *Matplotlib*

Matplotlib is a plotting library for the Python, which produces publication quality figures in a variety of formats. It allows to draw graphics, plots, histograms, power spectra, bar charts, error charts and scatterplots on the resulting data sets.

– *Scikit-learn*

Scikit-learn is a free, BSD-licensed machine learning library for Python. It features various classification, regression and clustering algorithms including support vector machines and random forests. The library is built on *SciPy* (fundamental library for scientific computing and other useful tools) and *NumPy* (base n -dimensional array package)

– *Networkx*

NetworkX is a free, BSD-licensed library for the creation, manipulation and study of the structure, dynamics and functions of graphs and complex networks in Python.

5 Evaluations

To evaluate the quality of predictions, we use commonly accepted measures in data mining: *precision*, *recall*, *accuracy* and *F1 score*.

- *Precision* (7) is the ratio of a number of events that are correctly predicted to a number of all predicted events.

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \quad (7)$$

- *Recall* (8) is the ratio of a number of the events that are correctly predicted to a number of all correct events.

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (8)$$

- *Accuracy* (9) is the ratio of correctly predicted events. Perhaps, it is the most intuitive performance measure, but it is only good for symmetric datasets where false positives and false negatives are roughly the same.

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{N} \quad (9)$$

where N (10) is sum of rates:

$$N = \text{fn} + \text{fp} + \text{tn} + \text{tp} \quad (10)$$

- F_1 score (11) is harmonic average of precision and recall.

$$F_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

The statistical measures of the performance of a binary classification test are also called rates:

- true positives (tp) are the categories that we expected to see and received at the output of the classifier;
- false negatives (fn) are the categories, which we expected to see, but the classifier did not define them, also known type II error;
- false positives (fp) are the categories, which should not be output, but the classifier returned them erroneously at the output, also known type I error ;
- true negatives (tn) are the categories, which should not be output, and they are also correctly absent at the output of the classifier.

5.1 Evaluation of Existing Methods

To have a base line, we started from the most commonly used measures of similarity [13] which were introduced in the previous section, cosine (1) similarity, common neighbors (2), Jaccard (3) similarity and Adamic–Adar (4).

5.1.1 Evaluation of Existing Methods on Facebook

To make evaluations, we used Facebook dataset⁶ which contains 4039 nodes and 88,234 edges. To predict connections, 30% of the edges were deleted and after deleting, the number of edges became 61,764. One of the questions that we need to answer is how many edges we should predict? The first approach we tried is that the similarity index (cosine, Jaccard and Adamic–Adar) should be greater than zero which led to very poor results. Then, we added a threshold value which decides if a future link should be predicted or not.

⁶ <https://snap.stanford.edu/data/egonets-Facebook.html>.

Fig. 3 Cosine similarity measure for Facebook

Fig. 5 Adamic-Adar similarity measure for Facebook

Fig. 4 Jaccard similarity measure for Facebook

The similarity measures have differences in their boundary values. The minimum value for all measures is 0; however, right side boundary of cosine and Jaccard is 1, whereas the right side boundary of Adamic-Adar is a float and depends on the number of intersected features. In our case, common neighbor similarity shows how many friends intersected between two users' friends lists, and therefore, the threshold is 0. The threshold value is chosen in the range from 0 to 1. The results of the threshold determination are represented in Figs. 3, 4 and 5. The influence of the threshold value is compared to F1-score.

After the previous evaluations, the thresholds for Jaccard and Adamic-Adar were not useful since the thresholds were equal to 0 and for cosine measure, best score is shown when the threshold is 0.6. The question "how many edges we should predict?" is very difficult to answer, and the performance depends on the answer. Table 3 shows the evaluation results of similarity measures using precision, recall, accuracy and F1-score.

Analyzing the results, we might say that common neighbor and Jaccard similarity measures show the same values.

Table 3 Evaluation measures for Facebook dataset

	Common neighbor	Cosine	Jaccard	Adamic-Adar
Precision	0.009341	0.006901	0.009341	0.010108
Recall	0.040130	0.346467	0.040130	0.037982
Accuracy	0.966173	0.746648	0.966173	0.969709
F1 score	0.015155	0.013532	0.015155	0.015967

Bold values indicate the best or the worst results

Adamic-Adar index shows the highest precision, accuracy and F1-score values in this example. Accuracy demonstrates high values, the reason is that TN in the is big, and accuracy contains TN in the definition. In our case, true negatives denote those connections that we did not predict and they did not really happen. Let us assume that we have a graph with 1000 nodes and 15,000 edges. The graph is not complete. Number of edges for complete graph can be found from formula:

$$\text{Number of Edges} = \frac{n(n-1)}{2} \quad (12)$$

where n is number of nodes.

So, for graph with 1000 nodes, number of possible edges is equal to 499,500. If we predict 5000 future links (new edges), true positive is equal:

$$\text{TP} = \text{APL} - \text{EL} - \text{PrL}$$

where APL is number of all possible links,

EL is number of existed links,

PrL is number of predicted links

$\text{TP} = 499,500 - 15,000 - 5000 = 479,500$, so now it is clear why TN is such big. Therefore, further accuracy does not greatly affect our analysis.

To examine popular similarity measures to predict links, we used data from Vkontakte (VK) social network. First, we processed Facebook dataset. The average number of users'

Table 4 Evaluating measures table for VK (baseline)

	Common neighbor	Cosine	Jaccard	Adamic–Adar
Precision	0.002936	0.019121	0.002936	0.014459
Recall	0.008744	0.171432	0.008744	0.005208
Accuracy	0.837793	0.959589	0.837793	0.822696
F1 score	0.004395	0.0005134	0.004395	0.007658

Bold values indicate the best or the worst results

Table 5 Datasets

Dataset	Nodes	Edges	Avg. degree
Facebook-1	324	2218	13.6914
Facebook-1_del	320	1553	9.7063
Last.fm	1892	12,717	13.4429
Last.fm_del	1801	8902	9.8856
GrQc	5242	14,496	5.5307
GrQc_del	4729	10,147	4.2914
HepTh	9877	25,998	5.2644
HepTh_del	9020	18,199	4.0353
CondMat	23133	93,497	8.0834
CondMat_del	21982	65,448	5.9547

friends is 23 in Facebook dataset and 246 in VK dataset. To efficiently process the big data, we have reduced from 10,728 nodes (2,332,068 edges) to 4000 user so that it contains 4000 nodes and 1,122,226 edges. To test the common practice, we deleted 30% edges and then try to predict the deleted ones. After deleting 30% of the edges, the number of edges was 784,499.

The threshold value is chosen to maximize the accuracy: 0.6 for cosine and 0 for others similarity measures. Table 4 illustrates the performance of the most popular similarity measures: precision, recall, accuracy and F1-score.

As shown in the table, the quality of predictions for VK network is lower than for Facebook dataset. One of the reasons is that VK network is very sparse and the relationships are concentrated in small regions. The results of the VK dataset show that Adamic–Adar similarity index is the best evaluating measure according to F1-score and cosine for others.

5.2 Evaluation of Existing Methods on Different Datasets

To further examine the similarity measures for link prediction, we used six different datasets shown in Table 5. To evaluate predictions was deleted 30% of the edges. Table 5 summarizes the original dataset and the dataset with deleted edges.

Tables 6, 7, 8, 9, 10, 11, 12, 13, 14 and 15 illustrate performance of predictions.

Table 6 Illustration of prediction rates for the small dataset from Facebook

	Cosine
(a) <i>Cosine rates</i>	
tp	102
fn	22
tn	8337
fp	60,167
Jaccard	
(b) <i>Jaccard rates</i>	
tp	89
fn	35
tn	56,306
fp	9504
Common	
(c) <i>Common neighbor rates</i>	
tp	89
fn	35
tn	56,306
fp	9504
Adamic	
(d) <i>Adamic–Adar rates</i>	
tp	59
fn	65
tn	58,243
fp	74,874

Table 7 Evaluating similarity measures for the small dataset from Facebook

	Precision	Recall	Accuracy	F1-score
Cosine	0.001692	0.822581	0.122967	0.003378
Jaccard	0.009278	0.717742	0.855325	0.018318
Common	0.009278	0.717742	0.855325	0.018318
Adamic	0.007819	0.475806	0.885322	0.015385

5.3 A Small Dataset from Facebook

This Facebook dataset differs from the Facebook data which we used in Sect. 5.1.1. This is the smallest dataset among the presented data in Table 5 with the biggest average degrees. Table 6 illustrates the confusion matrices for the small dataset from Facebook where all measures give plenty false alarms. The best TP result is from cosine similarity. Jaccard and common neighbor similarity measures demonstrate the same results in all rates. According to Table 7, the best recall is from cosine similarity but at the same time, the worst results in precision which leads to the worst F1-score.

5.4 LastFM

This dataset contains a social network from LastFM social networking platform. Table 8 illustrates the confusion matrices for LastFM dataset where all measures give plenty of false alarms. The best TP result is from cosine similarity. Jaccard and common neighbor similarity measures demonstrate the same results in all prediction rates. According to Table 9, the best recall result is from cosine similarity with the worst results in precision which leads to the worst F1-score.

Table 8 Illustration of prediction rates of LastFM dataset

	Cosine
(a) <i>Cosine rates</i>	
tp	406
fn	168
tn	384,640
fp	1,424,806
	Jaccard
(b) <i>Jaccard rates</i>	
tp	176
fn	398
tn	1,569,014
fp	124,157
	Common
(c) <i>Common neighbor rates</i>	
tp	176
fn	398
tn	1,569,014
fp	124,157
	Adamic
(d) <i>Adamic–Adar rates</i>	
tp	136
fn	438
tn	1,588,450
fp	104,017

Table 9 Evaluating similarity measures for LastFM dataset

	Precision	Recall	Accuracy	F1-score
Cosine	0.000285	0.707317	0.212730	0.000570
Jaccard	0.001416	0.306620	0.926462	0.002818
Common	0.001416	0.306620	0.926462	0.002818
Adamic	0.001306	0.236934	0.938303	0.002597

5.5 GrQc

General relativity and quantum cosmology (GR-QC) collaboration network is from the e-print arXiv and covers scientific collaborations among authors with papers submitted to General Relativity and Quantum Cosmology category.

Table 10 illustrates the confusion matrix for GrQc dataset. As shown in the table, again, all measures are prone to false alarms. According to Table 11, the best recall is given by cosine similarity and the worst results in precision as well as F1-score.

Table 10 Illustration of similarity rates for GrQc dataset

	Cosine
(a) <i>Cosine rates</i>	
tp	472
fn	479
tn	3,018,531
fp	5,461,423
	Jaccard
(b) <i>Jaccard rates</i>	
tp	321
fn	630
tn	7,853,685
fp	29,385
	Common
(c) <i>Common neighbor rates</i>	
tp	321
fn	630
tn	7,853,685
fp	29,385
	Adamic
(d) <i>Adamic–Adar rates</i>	
tp	147
fn	804
tn	7,861,336
fp	21,129

Table 11 Evaluating measures table for GrQc dataset

	Precision	Recall	Accuracy	F1-score
Cosine	0.000086	0.496320	0.355977	0.000173
Jaccard	0.010806	0.337539	0.996193	0.020941
Common	0.010806	0.337539	0.996193	0.020941
Adamic	0.006909	0.154574	0.997218	0.013227

5.6 HepTh

High energy physics theory (HEP-TH) citation graph is from the e-print arXiv. The average degree of this dataset is the smallest among all the presented datasets in Table 5. Table 12 illustrates the confusion matrices for HepTh dataset. According to Table 13, the best recall is given by cosine similarity.

5.7 CondMat

Condense matter physics (CondMat) collaboration network is from the e-print arXiv and covers scientific collaborations between papers submitted to Condense Matter category. This

Table 12 Illustration of various similarity rates for HepTh dataset

	Cosine
(a) <i>Cosine rates</i>	
tp	1073
fn	869
tn	10,804,371
fp	21,680,233
	Jaccard
(b) <i>Jaccard rates</i>	
tp	681
fn	1261
tn	29,960,801
fp	82,874
	Common
(c) <i>Common neighbor rates</i>	
tp	681
fn	1261
tn	29,960,801
fp	82,874
	Adamic
(d) <i>Adamic–Adar rates</i>	
tp	319
fn	1623
tn	29,988,844
fp	53,063

Table 13 Evaluating measures table for HepTh dataset

	Precision	Recall	Accuracy	F1-score
Cosine	0.000049	0.552523	0.332613	0.000099
Jaccard	0.008150	0.350669	0.997200	0.015930
Common	0.008150	0.350669	0.997200	0.015930
Adamic	0.005976	0.164264	0.998180	0.011532

is the biggest dataset among the presented dataset in Table 5. Table 14 illustrates the confusion matrices for CondMat dataset. According to Table 15, the best recall is given by cosine similarity.

Table 14 Illustration of similarity rates for CondMat dataset

	Cosine
(a) <i>Cosine rates</i>	
tp	3817
fn	2323
tn	59,017,890
fp	190,840,644
	Jaccard
(b) <i>Jaccard rates</i>	
tp	2820
fn	3320
tn	230,958,602
fp	627,368
	Common
(c) <i>Common neighbor rates</i>	
tp	2820
fn	3320
tn	230,958,602
fp	627,368
	Adamic
(d) <i>Adamic–Adar rates</i>	
tp	1610
fn	4530
tn	231,080,071
fp	499,409

Table 15 Evaluating measures table for CondMat dataset

	Precision	Recall	Accuracy	F1-score
Cosine	0.000020	0.621661	0.236215	0.000040
Jaccard	0.004475	0.459283	0.997277	0.008863
Common	0.004475	0.459283	0.997277	0.008863
Adamic	0.003213	0.262215	0.997824	0.006349

Table 16 Evaluating measures table for VK (deleting approach)

	Cosine	Jaccard	Adamic–Adar
Precision	0.0004111	0.0053608	0.0289651
Recall	0.0428219	0.0403065	0.0016984
Accuracy	0.0085148	0.9190919	0.8309117
F1 score	0.0008144	0.0094629	0.0032087

Bold value indicates the best or the worst result

Based on the performance results that we showed for five different datasets, we can observe how different measures behave on different datasets. Generally, what we want to have is high TP and low FP. In reality, however, when TP grows FP also grows. The same applies to TN and FN. Therefore, at least for these five datasets that we examined here, cosine similarity is not a good metric for link prediction problems since it results in high false positives. Although it has the highest TP rate among the four similarity measures under examination, it is only at the cost of high FP. We carefully suggest that, according to the results, Jaccard similarity is the most successful measure among the four, since it stably shows high TP rate with low FP as well as high TN and low FN. Adamic–Adar similarity sometimes has better performance than Jaccard similarity, but it is not as stable since sometimes it shows poor TP rate.

6 Results

Obtained results implies that it is possible to predict future connections at the price of false positives. Also the increasing size of a network (number of nodes) implies the exponential increase of TN. To overcome such shortcomings, we improvised new approaches.

6.1 Next Steps of Prediction on VK

The next step of our predictions using VK dataset was the assumption that not all user are “useful” and we should exclude them from our dataset. The assumption of “Usefulness” concerns that some profiles are not profiles of an ordinary user, but a company profile or a famous person (like singers), some profiles are not may be companies or popular people and in both cases they do not need to have recommendations otherwise the mean and the median of users in our dataset are about 240 friends also based on Dunbar’s number [2]; therefore, we delete users who have more than 500 users in their friend list.

The dataset contains 13,951 users, after excluding users who have more than 500 friends. To make predictions, we deleted 30% of the edges.

Table 16 illustrates evaluations of similarity measures: precision, recall, accuracy and F1-score. We found that Jaccard similarity has the best values in precision and F1-score.

6.1.1 Wall Analysis

As described in Sect. 2.1, any user of the VK social network may repost some information and these posts are attached to his wall. The posts have id of user or group, which shows

Table 17 Evaluating measures table for VK (real data comparisons)

	Cosine	Jaccard	Adamic–Adar
Precision	0.0000038	0.0000297	0.0001549
Recall	0.0001232	0.0001102	0.0000779
Accuracy	0.0030727	0.8596805	0.9532686
F1 score	0.0000073	0.0000468	0.0001037

where the repost was made from. Combining gathered frequency of posts and reposted places (ids where the repost was made from), we introduced a similarity measure of user A and user B:

$$\text{Wall Similarity} = \log(I(A) + R(B))$$

where I (shared ids) is number of common groups of two users

where R (shared reposts) is number of common posts of two users

(14)

6.1.2 Real Data

We compare VK dataset with 6966 users which was collected on December 22, 2016, with 1,506,714 edges and the dataset with the same users which was collected on January 24, 2017, which contains 1,523,040 edges. Table 17 illustrates the obtained results.

6.2 Proposed Measure

In this section, we present the new approaches to overcome the shortcomings shown before.

1. According to the evaluation further we pay attention only to TP, FP and FN rates, without using precision, recall, F1-score and accuracy.
2. Also we try to limit the number of predictions and select only the highest values of similarity to predict future connections.
3. In Sect. 6.1.1, we describe results of the approach using users’ walls to find similarity between users.

6.3 Evaluation of Other Approaches

Tables 18, 19 and 20 illustrate results of the similarity measures when we limited the number of predictions by selecting the top 2% from the best similarity indexes for 1000 randomly selected users.

We omit common neighbor similarity here since it shows similar results with Jaccard index and cosine. Instead we add two other measures, i.e., second neighbor similarity in Eq. (5) and shortest path length in Eq. (6).

Table 18 Illustration of various rates (data from January)

	Jaccard
(a) <i>Jaccard rates</i>	
tp	0
fn	534
fp	1006
	Adamic
(b) <i>Adamic–Adar rates</i>	
tp	0
fn	534
fp	1006
	Second
(c) <i>Second neighbor rates</i>	
tp	0
fn	4608
fp	1006
	Second
(d) <i>Shortest path rates</i>	
tp	0
fn	4800
fp	1006

Table 19 Illustration of various rates (data from February 2017)

	Jaccard
(a) <i>Jaccard rates</i>	
tp	0
fn	534
fp	1252
	Adamic
(b) <i>Adamic–Adar rates</i>	
tp	0
fn	534
fp	1252
	Second
(c) <i>Second neighbor rates</i>	
tp	0
fn	4608
fp	1252
	Shortest
(d) <i>Shortest path rates</i>	
tp	0
fn	4800
fp	1252

The main outcome of the results is that TP is equal to zero which implies that we predict very small number of potential relationships.

6.3.1 Top 25

In this experiment, we select 25% from the best similarity indexes for 1000 randomly selected users. Results are summarized in Tables 21, 22 and 23. The values for TP have increased, but we still have big false predictions.

6.4 Comparison of Second Neighbor Approach

In this section, we have used Facebook dataset which is previously presented. We have deleted 30% edges and then make prediction for 1000 randomly selected users. Table 24 illustrates that the proposed second neighbor measure shows the best TP and lowest FP with smaller FN compared to the shortest path measure.

7 Conclusion and Future Work

There exist many different types of link prediction approaches, and they can be roughly categorized into learning-based methods and proximity-based methods. In our research, we considered proximity-based methods, namely, cosine similarity, common neighbor, Jaccard similarity, Adamic–Adar index, second neighbor and shortest path. All of them are structural proximity-based methods which does not consider nodal information. We have thoroughly tested the measures on different datasets. Many existing studies make prediction only for randomly deleted edges due to lack of real datasets. We collected our own snapshots of graphs from VK social network within 4 months. To facilitate the collected dataset for prediction, we filtered users that have too many links and also maintained the same users for each snapshot since we predict links but not new users in this study.

The results do not conclude which approach is superior in most of the cases. The prediction method to use for real networks should be chosen with heuristics given the nature of the network. In this study, we showed that there is a great difference with the same measures in VK and Facebook datasets.

In [4], authors discuss on the problem of current link prediction approaches. We have pointed out that evaluating link predictions based on deleted edges is not suitable considering the dynamics of social networks. In addition, links are not only added but also deleted over time. Since this fact also affects the topological similarity measures, it is not effective to predict future links only based on the current snapshot of the network. We plan to work

Table 20 Illustration of various rates (data from March 2017)

	Jaccard
(a) <i>Jaccard rates</i>	
tp	0
fn	534
fp	1567
	Adamic
(b) <i>Adamic–Adar rates</i>	
tp	0
fn	534
fp	1567
	Second
(c) <i>Second neighbor rates</i>	
tp	0
fn	4608
fp	1567
	Shortest
(d) <i>Shortest path rates</i>	
tp	0
fn	4800
fp	1567

Table 21 Illustration of various rates (data from January 2017)

	Jaccard
(a) <i>Jaccard rates</i>	
tp	2
fn	24,336
fp	397
	Adamic
(b) <i>Adamic–Adar rates</i>	
tp	2
fn	24,336
fp	397
	Second
(c) <i>Second Neighbor rates</i>	
tp	2
fn	211,130
fp	397
	Shortest
(d) <i>Shortest path rates</i>	
tp	2
fn	219,836
fp	397

Table 22 Illustration of various rates (data from February 2017)

	Jaccard
(a) <i>Jaccard rates</i>	
tp	3
fn	24,335
fp	459
	Adamic
(b) <i>Adamic–Adar rates</i>	
tp	3
fn	24,335
fp	459
	Second
(c) <i>Second neighbor rates</i>	
tp	3
fn	211,129
fp	459
	Shortest
(d) <i>Shortest path rates</i>	
tp	3
fn	219,835
fp	459

Table 23 Illustration of various rates for 25% top predictions (data from March 2017)

	Jaccard
(a) <i>Jaccard rates</i>	
tp	4
fn	24,334
fp	535
	Adamic
(b) <i>Adamic–Adar rates</i>	
tp	4
fn	24,334
fp	535
	Second
(c) <i>Second neighbor rates</i>	
tp	4
fn	211,128
fp	535
	Shortest
(d) <i>Shortest path rates</i>	
tp	4
fn	219,834
fp	535

Table 24 Illustration of various rates (data from the Facebook dataset)

	Jaccard
(a) <i>Jaccard rates</i>	
tp	1776
fn	105,520
fp	10,212
	Adamic
(b) <i>Adamic–Adar rates</i>	
tp	1776
fn	105,520
fp	10,212
	Second
(c) <i>Second neighbor rates</i>	
tp	1791
fn	545,865
fp	10,197
	Shortest
(d) <i>Shortest path rates</i>	
tp	1791
fn	922,919
fp	10,197

Bold values indicate the best or the worst results

toward dynamics of networks to predict links with less false positives.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Chaney AJB, Blei DM, Eliassi-Rad T (2015) A probabilistic model for using social networks in personalized item recommendation. In: Proceedings of the 9th ACM conference on recommender systems, RecSys 2015, Vienna, Austria, Sept 16–20, pp 43–50
- Dunbar RIM (1992) Neocortex size as a constraint on group size in primates. *J Hum Evolut* 22(6):469–493
- Hussain R, Nawaz W, Lee JY, Son J, Seo JT (2016) A hybrid trust management framework for vehicular social networks. In: International conference on computational social networks. Springer, pp 214–225
- Junuthula RR, Xu KS, Devabhaktuni VK (2016) Evaluating link prediction accuracy in dynamic networks with added and removed edges. In: 2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom) (BDCloud-SocialCom-SustainCom), pp 377–384
- Kooti F, Lerman K, Aiello LM, Grbovic M, Djuric N, Radosavljevic V (2015) Portrait of an online shopper: understanding and predicting consumer behavior. *CoRR*, abs/1512.04912
- Lebedev A, Lee JY, Rivera V, Mazzara M (2017) Link prediction using top-k shortest distances, LNCS. Springer
- Lee JY, Oh JC (2017) Agent perspective social networks: distributed second degree estimation. *Encyclopedia of social network analysis and mining*, pp 1–12
- Lee JY (2014) Reputation computation in social networks and its applications
- Lee JY, Duan Y, Oh JC, Du W, Blair H, Wang L, Jin X (2011) Automatic reputation computation through document analysis: a social network approach. In: 2011 international conference on advances in social networks analysis and mining (ASONAM), pp 559–560. IEEE
- Lee JY, Duan Y, Oh JC, Du W, Blair H, Wang L, Jin X (2012) Social network based reputation computation and document classification. *J UCS* 18(4):532–553
- Lee JY, Lopatin K, Hussain R, Nawaz W (2017) Evolution of friendship: a case study of mobiclique. In: Proceedings of the computing frontiers conference. ACM, pp 267–270
- Lee JY, Oh JC (2014) Estimating the degrees of neighboring nodes in online social networks. In: International conference on principles and practice of multi-agent systems. Springer, pp 42–56
- Li Z (Lionel), Fang X, Sheng ORL (2015) A survey of link recommendation for social networks: methods, theoretical foundations, and future research directions. *CoRR*, abs/1511.01868
- Solonets S, Drobny V, Rivera V, Lee JY (2017) Introducing ADeegree: anonymisation of social networks through constraint programming, LNCS. Springer
- Szczepanski PL, Barcz AS, Michalak TP, Rahwan T (2015) The game–theoretic interaction index on social networks with applications to link prediction and community detection. In: Proceedings of the twenty-fourth IJCAI 2015, Buenos Aires, Argentina, July 25–31, pp 638–644
- Tigunova A, Lee JY, Nobari S (2015) Location prediction via social contents and behaviors: location-aware behavioral lda. In: 2015 IEEE international conference on data mining workshop (ICDMW). IEEE, pp 1131–1135
- Wu L, Ge Y, Liu Q, Chen E, Long B, Huang Z (2016) Modeling users’ preferences and social links in social networking services: a joint-evolving perspective. In: Proceedings of the thirtieth AAAI, Feb 12–17, Phoenix, AZ, USA, pp 279–286
- Zhang J, Tang J, Li J, Liu Y, Xing C (2015) Who influenced you? predicting retweet via social influence locality. *TKDD* 9(3):25