공학석사 학위논문

그래프 구조적 특성을 이용한 사회망 그래프 매칭 기법

Heterogeneous Social Network Graph Matching using Structural Features

2017년 2월

서울대학교 대학원

컴퓨터공학부

김 지 영

그래프 구조적 특성을 이용한 사회망 그래프 매칭 기법

Heterogeneous Social Network Graph Matching using Structural Features

지도교수 김 종 권

이 논문을 공학석사 학위논문으로 제출함

2016 년 10 월

서울대학교 대학원
컴퓨터 공학부
김 지 영

김지영의 공학석사 학위논문을 인준함

2016년 12월

위 원 장 : 전 화 숙 (인)
부위원장 : 김 종 권 (인)
위 원 : 권 태 경 (인)

Abstract

Heterogeneous Social Network Graph Matching using Structural Features

Jiyoung Kim
Department of Computer Science & Engineering
The Graduate School
Seoul National University

Social Information from social networks has been used in diverse research areas. Since social networks can provide abundant information, employment of social information commonly regards as the solution of data sparsity problem. In recommender system, for example, numerous researchers uses social information to solve cold start problem, which is that the system cannot draw any inferences for object who has not yet gathered sufficient information. However the information provided by one social network is very limited to surmount data sparsity problem. Graph matching techniques which combines information of heterogeneous social network can be broad and firm base of

other social network related research areas. Nowadays, users are opt to involve in multiple social networks simultaneously. Due to the fact that each social networks offer distinct service function and that data published for research is usually anonymized, there are not sufficient common information among heterogeneous social networks services. However, the graph structure formed by a same user tends to remain similar. In light of above, we propose novel approach to integrate heterogeneous social networks. Differ from other heterogeneous graph matching, we use not only simple in-and-out degree of social networks, but also Jaccard coefficient, Adamic/Adar score, Clustering coefficient, and Page rank to evaluate social status of user. Extensive experiments conducted on multiple real-world data and prove that our proposed method outperforms existed graph matching algorithm.

Keyword : Heterogeneous Social Networks, Graph Matching, Network Structure, Social Network Integration

Student Number : 2015-21235

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social Network has been emerged as prominent research area in both academy and Industry. With the generalization of smartphones and easy access to online services, the usage of Social Networks is drastically increased and the cumulated log data of user's behavior in social networks are exceedingly increased. Using these data, many researchers try to extract valuable information to solve the data sparsity problem of their research. Recommender system, Marketing, Link prediction, and spam detection are some of example areas who use social information actively.

However, a single social networks cannot cover complete information of users. Users use the multiple social networks for different purposes. Therefore, combining all the information from multiple social networks can allow

researchers obtain more abundant information about users and their friends. Zhang [2] used integrated social information from multiple social networks to predict social relation among users and obtain more accurate data of user's interest from them.

Integrating multiple, especially heterogeneous social networks is very challenging due to several reasons. First, data published for research are often pre-anonymized. The attribute information to identify specific users are removed or changed. In other words, the information that represent each user's characteristics are absent. This makes graph matching problem very tough. Second, even when the data which represent user's personal characteristics are exists, user's information in different social networks is very unbalanced. Some social networks contains abundant user's profile data, while the others contains scarce user's profile data. Third, the contents which are created by users have different topics in heterogeneous social networks. Assume a user who is involved in both Flickr and LastFM. The user tends to post photos in Flickr, because Flickr is popular photo-sharing networks, however, the user will have a tendency to share his music tastes in LastFM, because LastFM provides streaming radio services and users are often share their personal music preference. Under these challenging condition, user's social behavior is vital clue to identify themselves in other networks.

To address these challenges, in this paper, we use only structure information of social relation to identify same users in heterogeneous social networks. According to Therefore, we can overcome several challenges such as

data unbalance or topic difference in contents. We also propose a novel user identification algorithm. Named Structural Feature-based User Identification (SFUI).

Chapter 2

Related Work

There are three principal element in Social Networks, profile, content and network structure. Profile is the information made by users to briefly introduce themselves. Content is the object from users to share their thought or fact. There are various form of content including text, pictures, music. Network structure represent the relationship between users in a social network. Therefore, graph matching techniques use these three element as principal means to indicate users in different networks.

2.1. Profile-Based Graph Matching

Almost all social networks require users' basic profile data and some of the profile data is open to public, called public profile attribute. User name, birthday, gender, city are often considered as public profile attribute. So, the public profile data can be most commonly shared information throughout numerous social networks. Researches who focus on these data have assumed that user maintain their profile similar in multiple social networks. There exists considerable amount of researches in this category. [3, 4, 5] used screenname as main attribute. They used username similarity [3] and unsupervised approach to recognize users [4]. While [3, 4] proposed the method to match users in pairwise, Zafarani and Liu matched users in multiple social networks (more than two social networks). For that, they analyzed user behavior reflected on screenname such as the percentage of keys typed using same hand or finger [5]. A username is one of the most powerful contents when it is used sole method. However, it can also have a decisive effect when it is combined with other profile contents. [6, 7, 8] aggregate all the public profiles to obtain same users across the social networks in different types of social networks respectively. According to the results shown in previous works, basic profile information gives strong impact on graph matching. However, the profile data have high probability of being duplicated as social networks getting larger, and being easily impersonated. Therefore, technique of depending on profile data is effective but it has its limitation.

2.2. Content-Based Graph Matching

Content-based graph matching techniques assume that contents created by same user in multiple social networks have similarity in posting time, check-in location, writing styles and content categories. Kong and Zhang, who propose Multiple-Network Anchoring(MNA) algorithm, proves that there exists similarities of user's temporal, spatial, social and content information across multiple social networks [9] .In [10], author used check-in location, activity time pattern, and writing style to search the accounts owned by same user. Along with profile information attributes from content has powerful impact on matching users. However, content information is often very different according to the characteristics of social network services, and some content like location information is often very scarce. And also it's very hard to extract same kind characteristics from contents due to the diversity of content form.

2.3. Structure-Based Graph Matching

As user uses online social networks to maintain and to broaden social relationship with offline friends, the social structure in multiple online social network maintain very similar to that in real-world social relationship. The researchers who used the structure based graph matching techniques uses this graph structure constructed by social networks users as key measure to evaluate user identification. In [11], Narayanan and Shmatikov proposed solely structure based graph matching algorithm called NS, Bartunov et al. used conditional random field with graph structure to graph matching method [12]. Korula and Lattanzi also proposed algorithm using degrees and commonly known users for identifying unknown users [13]. [14] suggested Friend Relationship-Based User Identification (FRUI) algorithm which calculates a score of commonly known friend for all candidate user pairs and select a user pair with top score. While [11, 12, 13, 14] used only structure information, there are considerable amount of researches which used structural information combine with profile and content information. And that joint usage of information may lead to accurate matching results. In [12], Bartunov et al. combine profile with network structure in conditional random field model and get better result. [15] used joint information of profile, content, and structure and proposed energy-based model named COSNET (COnnecting heterogeneous Social NETworks with local and global inconsistency)

We use network structure based graph matching, and achieve robustness from the duplicating profile, data sparsity of common contents. And utilizing

only graph structure enables our method to apply for the (1) Multiple Anonymized Social Networks Alignment(M-NASA) problem, (2) supervised anchor link inference across social networks which is for inferring the anchor links across two social networks with supervised learning model , and (3) de-anonymization problem [20, 21, 22].

Chapter 3

Proposed Method

In this section, we defines related terminologies, and discuss our solution to the graph matching problem

3.1. Terminology

According to the [16], online social network is an online platform that is used by people to build social relations. In addition to making social relationship, online social networks offers people to express themselves through their profile and contents, and to share their personal or career interests, activities. We formally define these concepts below.

Definition 1 (Social Networks) An Social Network is defined as SN ={U,C,E}. U stands for set of Users, C stands for contents U created, E stands for the set of edges which represent relationship between users. In this paper, two social networks A and B will be denoted as $\mathrm{SN_A} = \{\mathrm{U_A}, \mathrm{C_A}, \mathrm{E_A}\}$, $\mathrm{SN_B} = \{\mathrm{U_B}, \mathrm{C_B}, \mathrm{E_B}\}$ respectively. Assume that $\mathrm{SN_A}$ contains $\mathrm{m_A}$ number of users, and $\mathrm{SN_B}$ contains $\mathrm{m_B}$ number of users. We denote these users and set of users as follows. $\mathrm{U_A} = \{U_{A_1}, \cdots, U_{A_i}, \cdots U_{A_{m_A}}\}$, $\mathrm{U_B} = \{U_{B_1}, \cdots, U_{B_j}, \cdots U_{B_{m_B}}\}$

Definition 2 (Matched user pair) If $\mathrm{i-th}$ user in Social Network A, and $\mathrm{j-th}$ user in Social Network are owned by a same person in real life, which is denoted as Ψ, they can be expressed as $\Psi_{\mathrm{A\sim B}}(i,j)$ or $\Psi(\mathrm{U_{A_i}}, \mathrm{U_{B_j}})$.

Definition 3 (Seed user pair) Seed user pairs are Matched user pairs that are given before matching process is executed.

Definition 4 (Adjacent users) Adjacent users are users who have relation with users in Matched users

3.2. Methodology

SFUI assume that user's social relation is maintained similarly across the heterogeneous social networks, and use structure-based graph matching techniques. SFUI has similar format as other structure-based graph matching techniques. From the seed user pairs, it tries to find user pairs with highest possibilities to be owned by same person, iteratively. User matching composed of two steps.

3.2.1 Step1 : Simple Match

In first step, it first choose candidate users, who are related to already matched users, i.e. adjacent users. Among adjacent users in each social networks, it calculates Matching score M. Matching score of two users $\mathbf{U_{A_i}}, \mathbf{U_{B_j}}$ is as follows

$$\mathbf{M_{ij}} = \mathbf{M}\left(\mathbf{U_{A_i}}, \mathbf{U_{B_j}}\right) = \left|\boldsymbol{F}_{\mathbf{A_i}} \cap \boldsymbol{F}_{\mathbf{B_j}}\right| \cdots (\mathbf{1})$$

$\boldsymbol{F}_{\mathbf{A_i}}$, $\boldsymbol{F}_{\mathbf{B_j}}$ denote that identified friends set of $\mathbf{U_{A_i}}$ and $\mathbf{U_{B_j}}$. This matching score which use count of common neighbor can overcome the drawbacks of matching score of NS and JLA which are a ratio of common neighbor and degree.

$$\mathrm{M_{ij}} = \mathrm{M}\left(\mathrm{U_{A_i}}, \mathrm{U_{B_j}}\right) = \frac{c_{in}}{\sqrt{d_{in-B_j}}} + \frac{c_{out}}{\sqrt{d_{out-B_j}}} \cdots (2)$$

The equation above is matching score function of NS. In that equation $\mathrm{c_{In}}$

and c_{out} stand for the number of shared incoming and outgoing neighbors of U_{A_i} and U_{B_j}. d_{in-B_j} and d_{out-B_j} denote the in- and out- degree of U_{B_j}. NS assumes that the same user in different Social Networks has the same amount of in-and-out degree. With this matching scoring method, node with single or small number of degree makes noise in selecting most probable matching user pairs.

$$M_{ij} = M\left(U_{A_i}, U_{B_j}\right) = \frac{2 \times \omega\left(F_{A_i} \cap F_{B_j}\right)}{\omega\left(F_{A_i}\right) + \omega\left(F_{B_j}\right)} \cdots (3)$$

Equation (3) is matching score function of JLA. $\omega(F) = \sum_{v \in F} 1/d(v)$ where $d(v)$ is degree of node v. M_{ij} can get the score range of 0 to 1, when users don't share any common identified friends, score become 0, and when user have same set of shared known friends, score become 1. Similar to the NS score, if there are lots of user pairs who share only one or small number of identified users, incorrect identification is easy to occur.

Basic idea of using ratio like NS and JLA is normalizing the effect of degree. With this algorithm, there is higher chance to better score if node degree is small. In other words, these algorithms consider a common shared friend of user who have small number of friends as more meaningful than that of user who have large number of friends. However, as we mentioned earlier, if there are lots of nodes who share small number of common neighbor and who have small number of friends, the idea of normalization doesn't work. Therefore, by simply counting the number of common identified friends, we can remove the noise of small degrees.

We consider the higher the score a user pair's score the higher probability of users of the user pair belonging to the same individual. Therefore, if there are only **un-contradictory** user pairs in user pair set which has highest score of equation (1), we consider those user pairs as right match. We call un-contradictory if there are not same user in several user pair. For example, $\Psi_{\mathrm{A\sim B}}(i,j)$ and $\Psi_{\mathrm{A\sim B}}(i,k)$ have same matching score, these UMPs are contradictory.

3.2.2 Step2 : Sophisticate Match

Although (1) overcome the drawback of earlier, it can get lots of user pairs with same score. In second step of user matching, we use network structural features to select one of the most probable user pairs. We assume that if a user's social relation is similar in multiple heterogeneous social networks

Jaccard Coefficient : This captures the ratio of shared friends to shared and non-shared friends.

$$\mathrm{JC}\left(\mathrm{U_{A_i}}, \mathrm{U_{B_j}}\right) = \frac{\left|F_{A_i} \cap F_{B_j}\right|}{\left|F_{A_i} \cup F_{B_j}\right|} \cdots (4)$$

Adamic/Adar Score : Similar to the Jaccard coefficient, this also capture the amount of shared friends. However, in this metric, it use $1/\mathrm{d(v)}$ to give higher score to the shared friend who have lower degree

$$\mathrm{AA}\left(\mathrm{U_{A_i}}, \mathrm{U_{B_j}}\right) = \frac{\omega\left(F_{A_i} \cap F_{B_j}\right)}{\omega\left(F_{A_i} \cup F_{B_j}\right)} \cdots (5)$$

Clustering Coefficient : This captures how close neighbors of a node are to being a clique. A user whose friends construct very dense relation in one social network, will also have friends who have lots of relation between themselves in another social network.

$$\mathrm{CC}(\mathrm{U_{A_i}}) = \frac{2\left|\{e_{xy}: U_{A_x}, U_{A_y} \in N_{A_i}, e_{xy} \in E_A\}\right|}{d(U_{A_i})\left(d(U_{A_i})-1\right)} \cdots (6)$$

In equation (6) $\mathrm{e_{xy}}$ is edge between $\mathrm{U_{A_x}}$, $\mathrm{U_{A_y}}$, and Clustering coefficient of $\mathrm{U_{B_j}}$ can be calculated similarly.

Degree : This is another way to represent social status with page rank. While page rank captures the influence of user, degree represent role in social network. In [15], Zhang et al. use degree as social status. They prove that the status of a user in different social networks usually be consistent and top 1% of accounts as “opinion leaders”, the following 10% as “middle class”, and the rest as “ordinary people”. Instead of dividing users’ account in several classes, we use the rank of users’ account directly.

After normalizing each variable from 0 to 1, we combine 5 structural features with parameters, and select the proper one among the user pairs who share maximum number of friends. The equation of combining structural features as bellows.

$$\begin{aligned}\mathrm{SF}\left(\mathrm{U_{A_i}}, \mathrm{U_{B_j}}\right) = {} & \alpha\left(1 - \mathrm{JC}\left(\mathrm{U_{A_i}}, \mathrm{U_{B_j}}\right)\right) + \beta\left(1 - AA\left(\mathrm{U_{A_i}}, \mathrm{U_{B_j}}\right)\right) \\ & + \gamma\left(\mathrm{CC}(\mathrm{U_{A_i}}) - \mathrm{CC}\left(\mathrm{U_{B_j}}\right)\right) + \delta\left(\mathrm{PR}(\mathrm{U_{A_i}}) - \mathrm{PR}\left(\mathrm{U_{B_j}}\right)\right) \\ & + \eta\left(\mathrm{DR}(\mathrm{U_{A_i}}) - \mathrm{DR}\left(\mathrm{U_{B_j}}\right)\right),\end{aligned}$$

When Jaccard coefficient and Adamic/Adar score are higher, the user pair have more possibility to be owned by a same user. On the contrary, Difference of two user's values of other three value which are Clustering coefficient, pagerank and degree should be small, if they are considered to be the same user's account. So, we subtract Jaccard and Adamic/Adar score from 1 to make structural feature score. Among user pairs with highest matching score in first matching step, we choose a user pair with lowest structural score. When there are several user pairs who have lowest structural score, which makes user pair selection problem harder, any of user pairs will be selected and do the second matching step again with user pairs with second highest matching score in first matching step. For simplicity, default parameters of structural feature score are all 1 (i.e. $\alpha = \beta = \gamma = \delta = \eta = 1$)

Chapter 4

Experiment

4.1. Dataset

We conduct experiments on dataset from four popu.ar social network: LiveJouranl, Flickr, LastFM. Myspace. These four dataset is randomly sampled from the data used in [15]. The table below shows the statistics of the original data

Network	# Users	# Relationships
LiveJournal	89,678	3,816,937
Flickr	16,096	489,989
LastFM	58,214	682,219
Myspace	44,469	203,962

Table 1. Data Statistics

LiveJournal: This is free online social network that allows users to share a blog, journal, or diary. The users in this dataset, there are 89,678 users and the number of relation between them is 3.8 million. This data is originally crawled from the website in late 2013.

Flickr : This is one of a popular photo-sharing networks. Users can post and share their photos in this social networks. This dataset is also originally crawled from Flickr website in early 2014 with LiveJournal dataset. The number of users is 16,096 and the number of friend-relationship is 489,989.

LastFM : is a music website. It builds a detailed profile of each user's musical taste with the user's activity history and provides a personalized recommendation. It also offers a streaming music services. This dataset was collected in late 2013 and consists of 58,214 users and 682,219 social relationships.

Myspace : a social networking website that provides users to share music. There are 44,469 users and 203,962 relationships between the users in this data

Ground Truth. It is very challenging to obtain co-users, which are “ground-truth”, of this dataset. To obtain those user pairs we use the linked users’ account to connect the users in [17,18]. Based on user pairs in [18], we tries to connect user pairs through Google profile services, which provides users to integrate different social networks. If a user have two or more different social networks, we use it as ground truth dataset.

4.2. Comparison Method

Considering our proposed method, we use the social networks that only contain social link information, therefore, we compare our method which don't use any other information except link information. For the model which use profile or content, we evaluate accuracy without contents and profile information.

NS : This is the first network structure-based user recognition algorithm across social networks. In their paper they achieve 30.8% user identification in the ground truth only with graph structure. This algorithm calculates the mapping scores of all of unmapped user pairs. They also calculate eccentricity of user pairs. If both mapping score and eccentricity of user pair are higher than thresholds, the user pair is selected, and users in the user pair are considered to be owned by same users. Along with those score, NS also have reverse match process to assure the matching, and this is costly process.

JLA : JLA matches users by comparing the mapped neighbors of each node. This algorithm calculates a network distance of all of unmapped nodes. It is designed for undirected networks. In the empirical study, [12] show that some user pairs can be matched based on solely network structure using JLA.

FRUI: FRUI is state-of-art in matching users using only graph structure by iteration. Like NS and JLA, FRUI start with seed user pair set, and iterate the matching process. In matching, they count shared identified users and use it as matching score like our proposed method. However, unlike our method, for the user pairs who have same number of common identified user pairs, FRUI

simply divide matching score by minimum degree of users in user pair.

SFUI: SFUI is proposed method in this paper, which uses graph structure in diverse aspect. In this algorithm, users commonly known friends are counted and is used as first matching score. For more precise selection of user pairs, it uses numerous graph structure like Jaccard coefficient, Adamic/Adar score, clustering coefficient, Page rank and Degree

4.3. Evaluation Metric

To evaluate the performance of proposed method we use precision, recall and propagation rate.

Precision : is the fraction of the correctly matched user pairs of results and all matched pairs in results. However, since we only knows the user pairs in ground truth, the precision is calculated as the fraction of the number of correctly matched user pairs of results and that of matched user pairs who contains at least one ground truth users.

$$\text{Precision} = \frac{|\#'s\ \ of\ correctly\ matched\ user\ pairs|}{|\#'s\ \ of\ user\ pairs\ who\ contains\ GT\ users\ among\ results\ |}$$

Recall : Similar to precision, we only user pairs related ground truth users for calculating recall. The calculation equation of recall is as bellows.

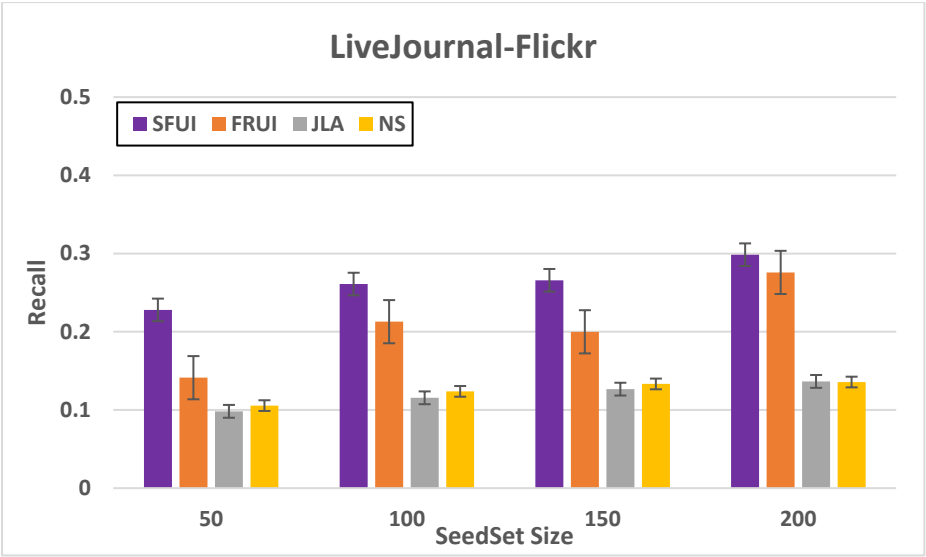
$$\text{Recall} = \frac{|\#'s\ \ of\ correctly\ matched\ user\ pairs|}{|\#'s\ \ of\ \ user\ pairs\ in\ GT\ |}$$

Propagation Rate : One of the key issue in graph matching using iterative method is how far the model propagates the iteration. To evaluate that, we use the ratio of the number of users in the results to the number of total users in dataset.

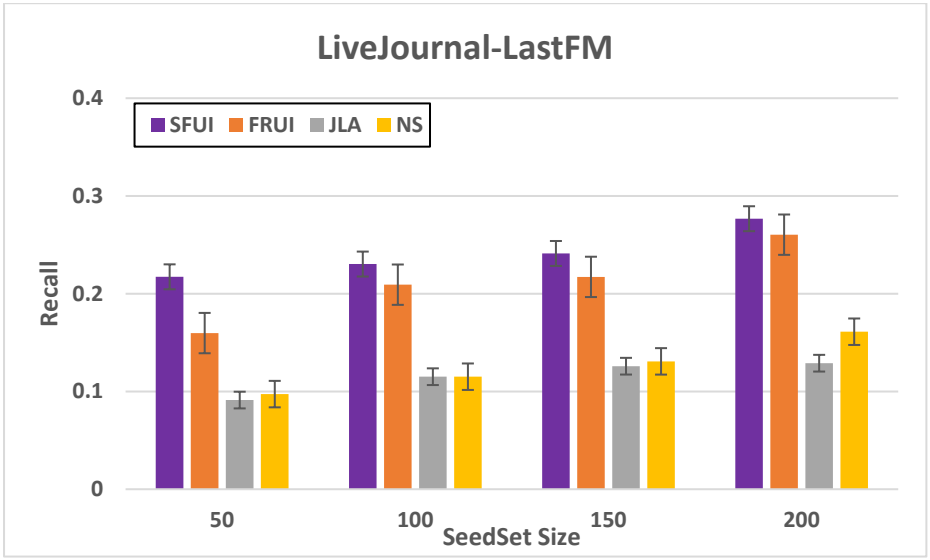
$$\text{Propagtion Rate} = \frac{|\#'s\ \ of\ users\ in\ matching\ result\ \ |}{|\#'s\ \ of\ \ users\ in\ total\ dataset\ |}$$

4.4. Performance Analysis

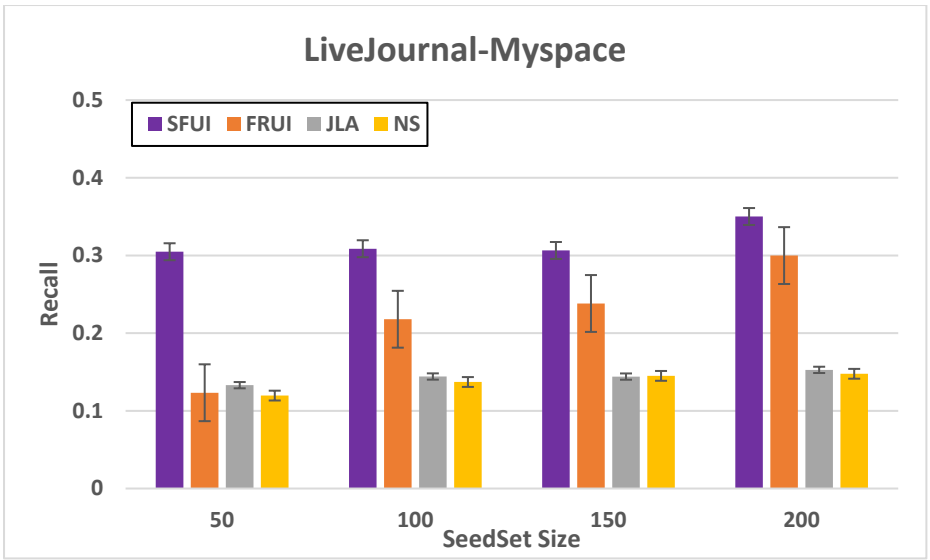
We perform experiment on four data sets which are LiveJournal, Myspace, LastFM, Flickr and get three different matching results which are LiveJournal-Myspace, LiveJournal-LastFM, LiveJournal-Flickr. These results show very similar patterns in terms of recall and precision.



(a) Recall of LiveJournal-Flickr user matching



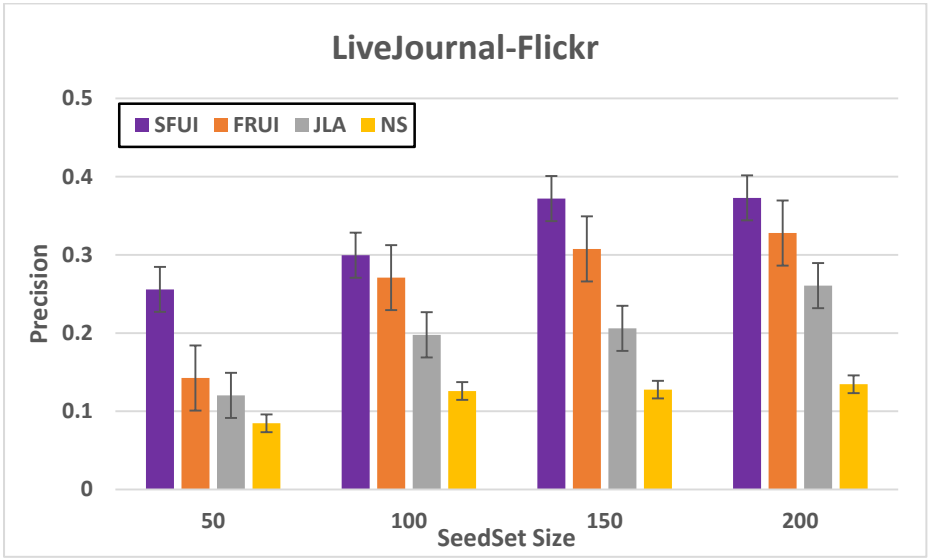
(b) Recall of LiveJournal-LastFM user matching



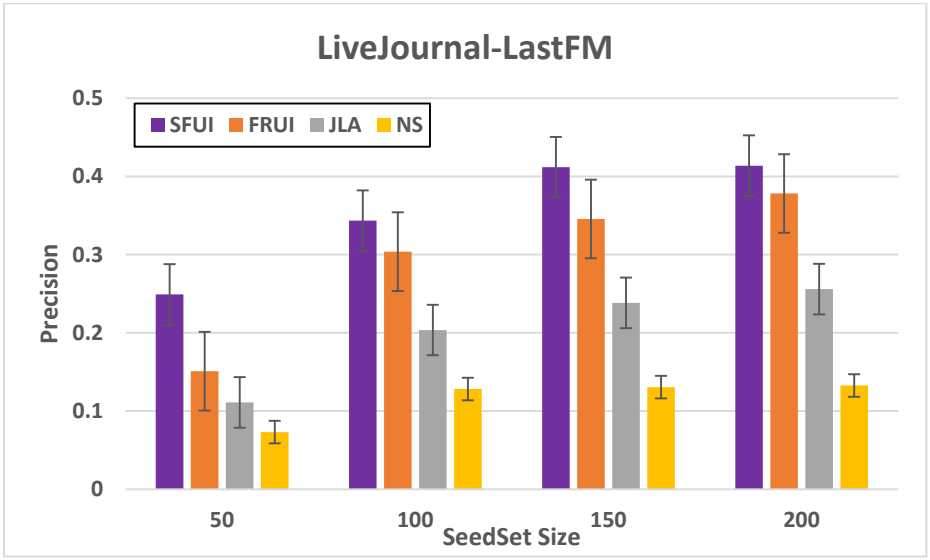
(c) Recall of LiveJournal-Myspace user matching

Figure 1. Recall of various graph matching algorithm

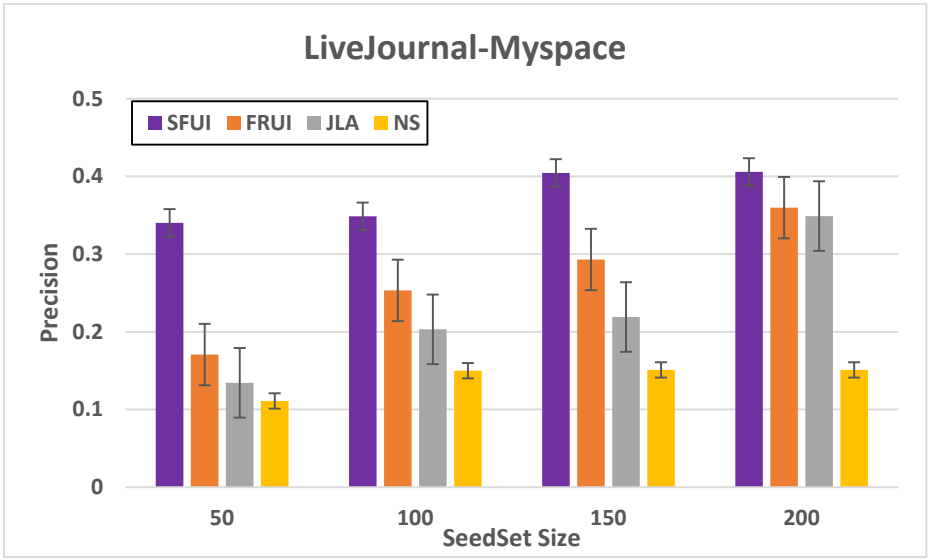
Figure 1 shows that the recall of SFUI, FRUI, JLA, NS using three dataset pairs (LiveJournal-Myspace, LiveJournal-LastFM, LiveJournal-Flickr). Our proposed method outperforms all of other structure based graph matching algorithms. In terms of Precision, SFUI also achieves almost 5 - 30% performance improvement over FRUI which is state-of-art graph matching method. The Precision values are shown in Figure 2



(a) Precision of LiveJournal-Flickr user matching



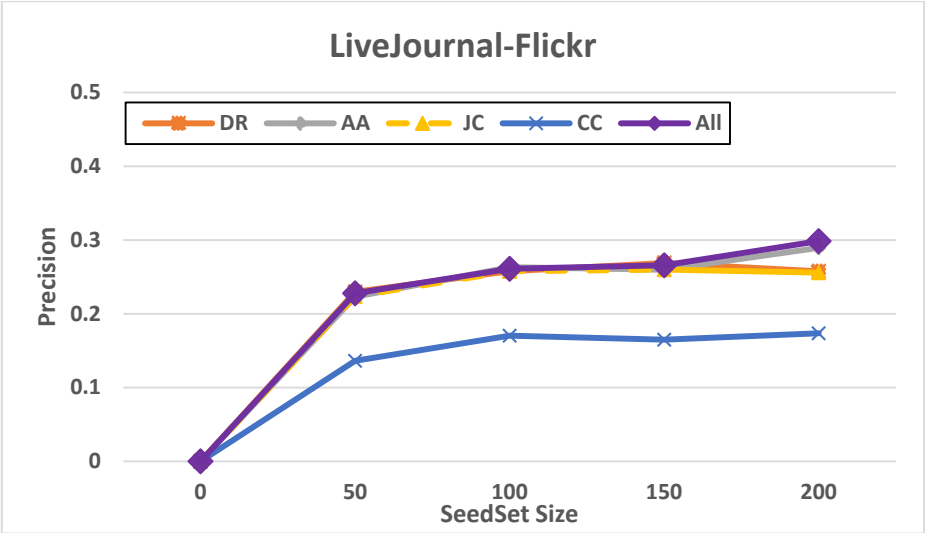
(b) Precision of LiveJournal-LastFM user matching



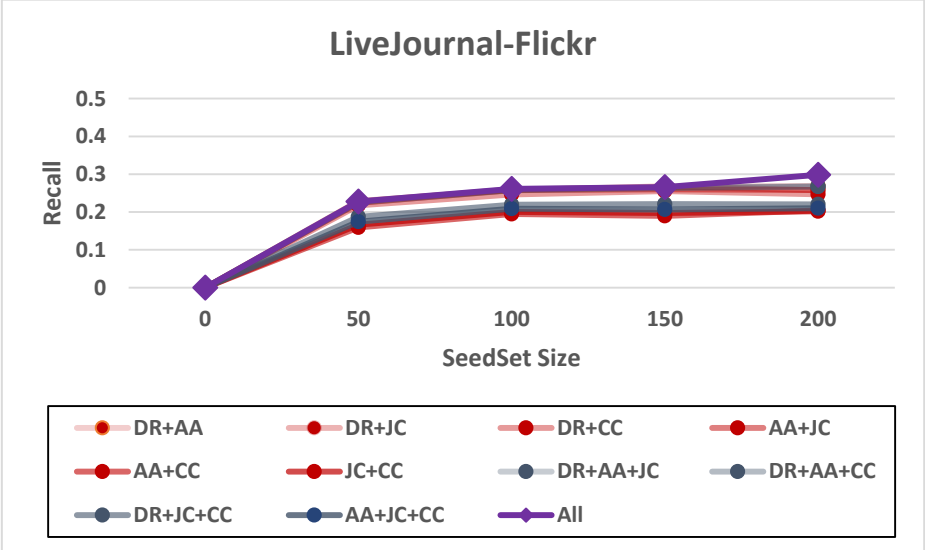
(c) Precision of LiveJournal-Myspace user matching

Figure 2. Precision of various graph matching algorithm

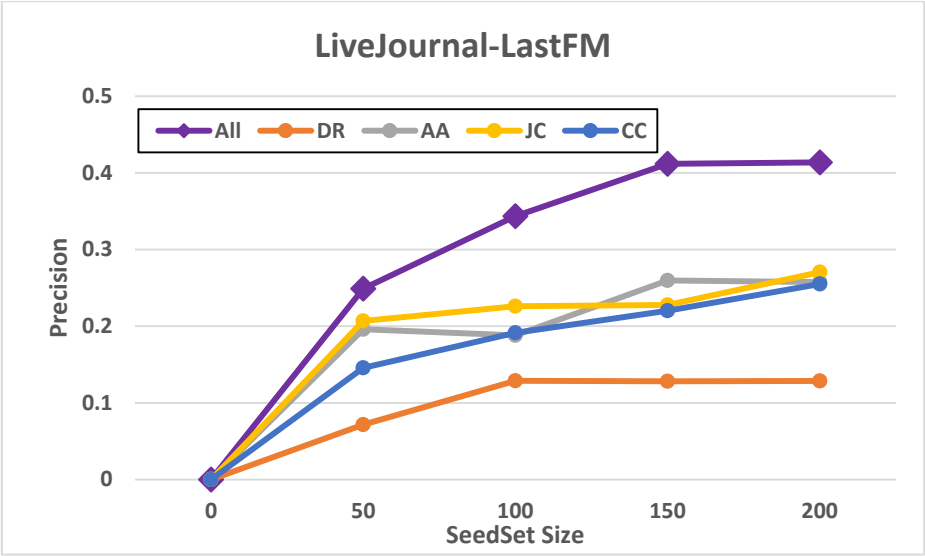
. One of the notable results is that SFUI and FRUI performs much better than JLA and NS. This is due to the difference of way of measuring. While JLA and NS use the ratio using in- and out-degree and the number of commonly known friends, SFUI and FRUI use the number of shared friends directly. Therefore, SFUI and FRUI are not strongly affected by low degree users which can cause noise to the JLA and NS. If the incorrect user pairs are matched in JLA and NS, it affects the matching score of candidate user pairs, and the error become cumulated. So, the effect of noise which usually happens when there are not much matched users is critical to matching accuracy. The results also shows that SFUI obtains more accurate matching results that FRUI. This shows that using simply degree like FRUI, use the combination of more diverse structural features helps to reflect the user’s distinct structural characteristics more precisely.



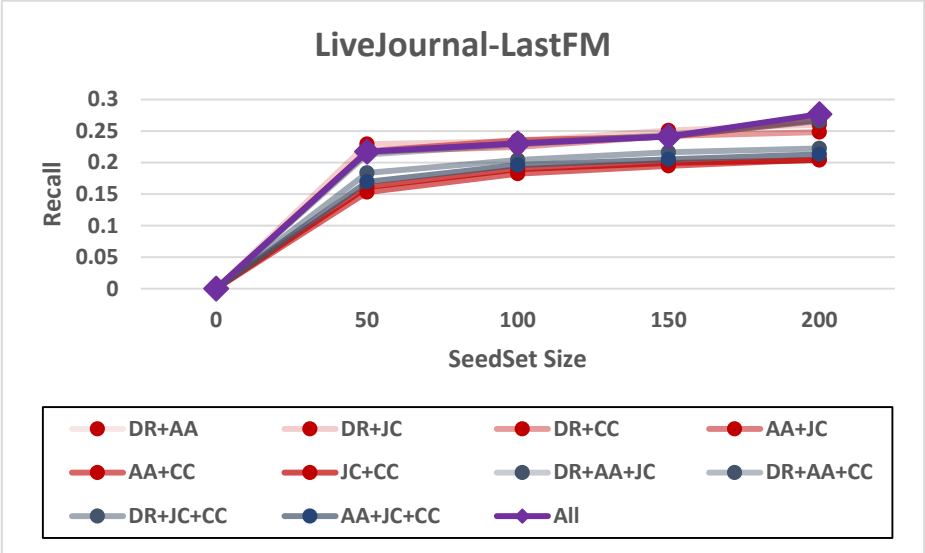
(a)-1 Recall of LiveJournal-Flickr with different features



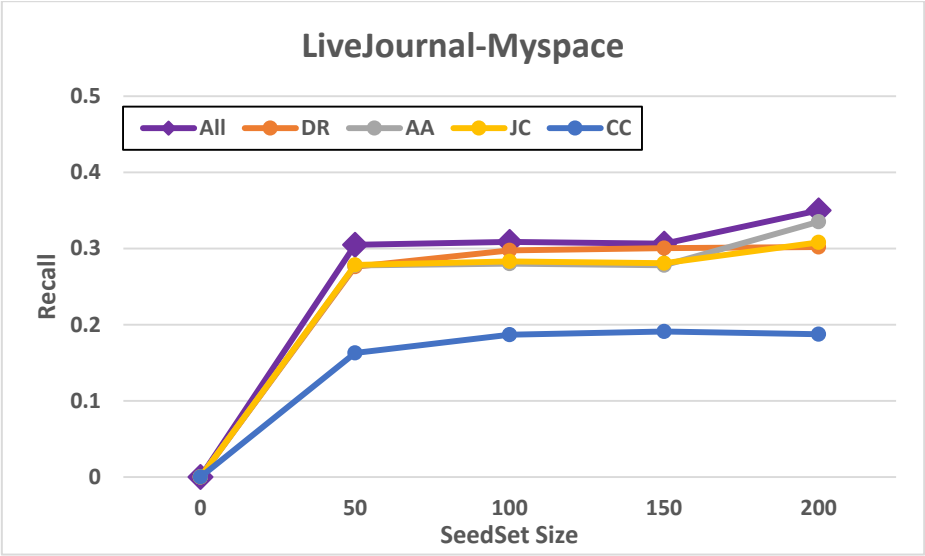
(a)-2 Recall of LiveJournal-Flickr with different feature combinations



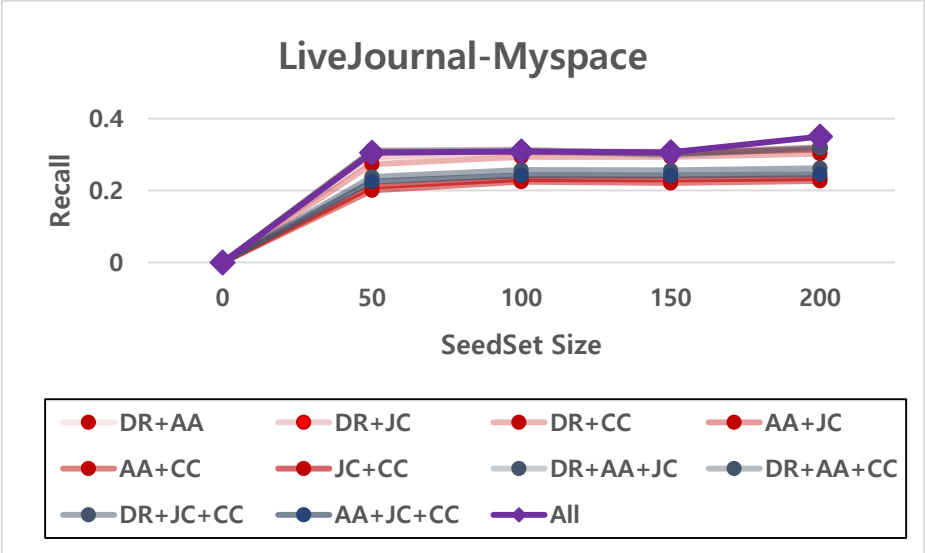
(b)-1 Recall of LiveJournal-LastFM with different features



(b)-2 Recall of LiveJournal-LastFM with different features combinations

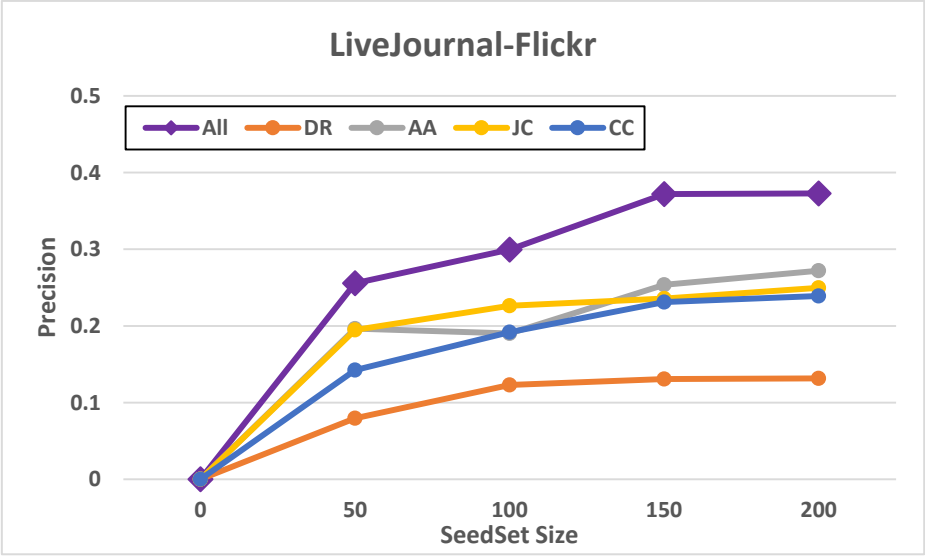


(c)-1 Recall of LiveJournal-Myspace with different features

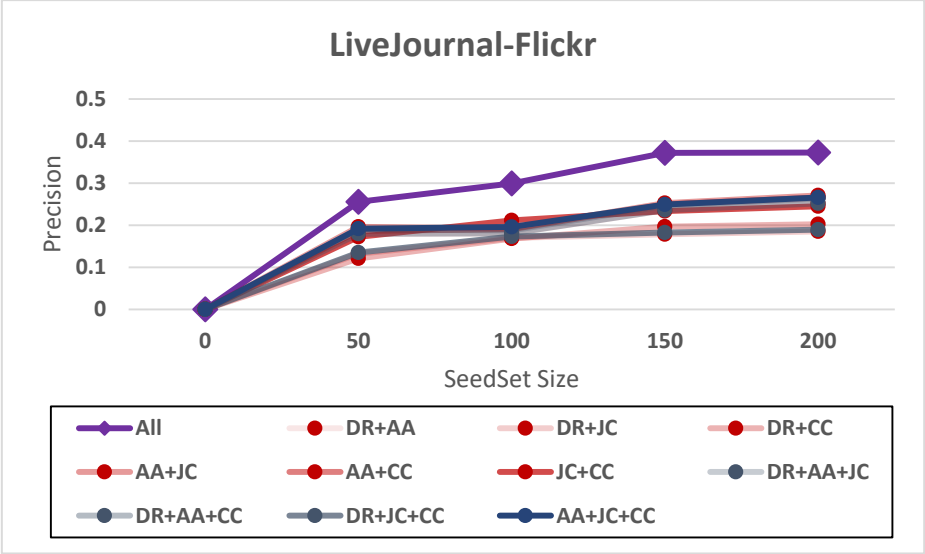


(c)-2 Recall of LiveJournal-Myspace with different feature combinations

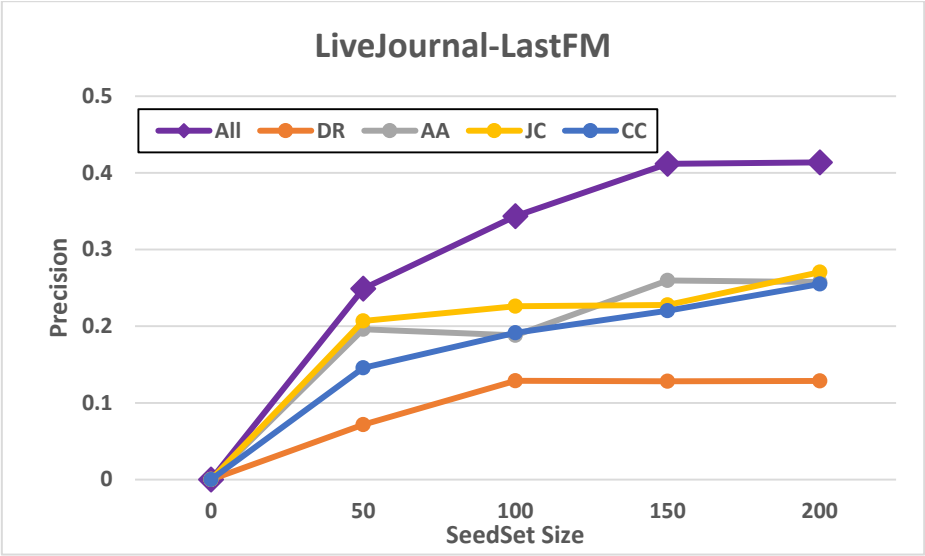
Figure 3. Recall of SFUI with different structural features



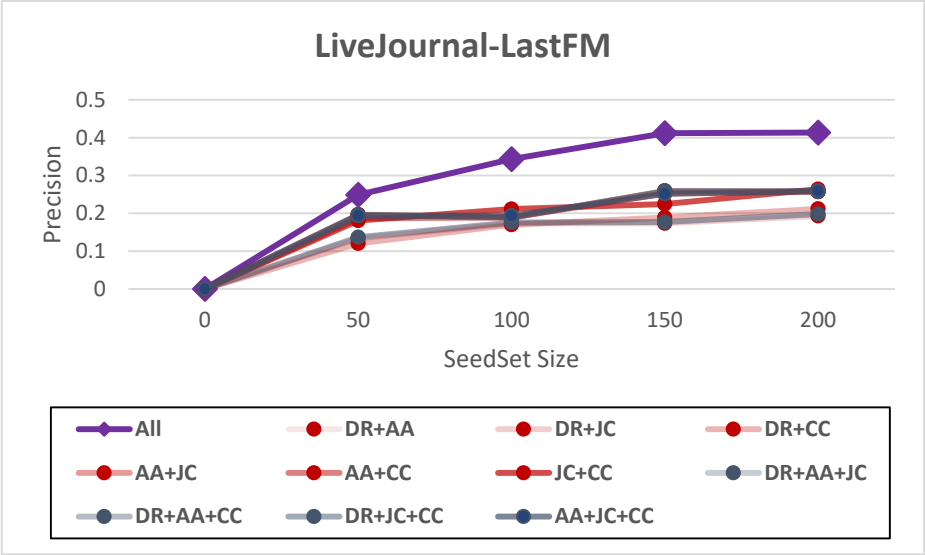
(a)-1 Precision of LiveJournal-Flickr with different features



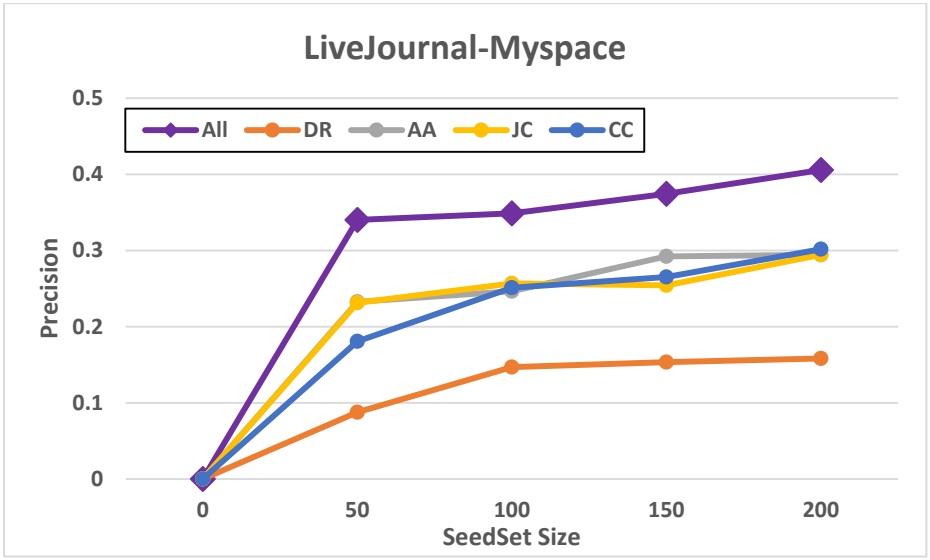
(a)-2 Precision of LiveJournal-Flickr with different features combinations



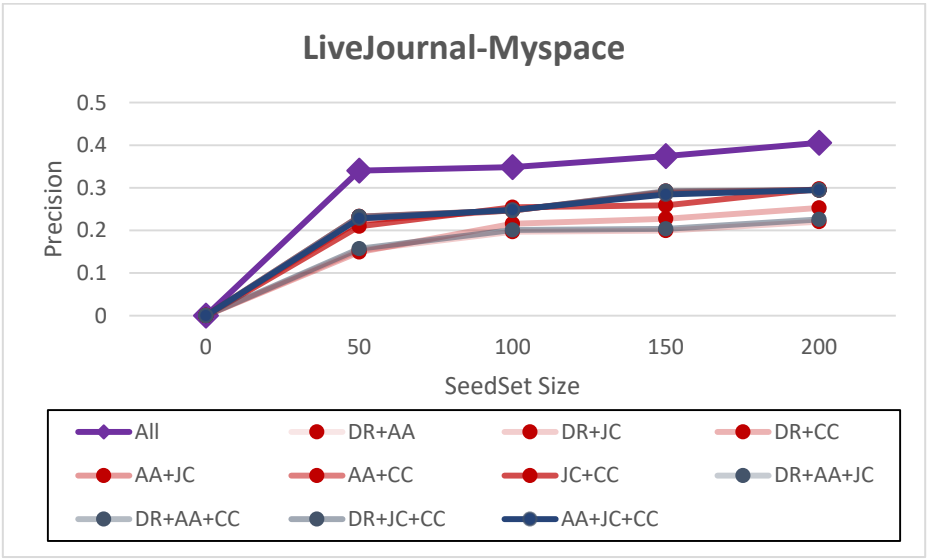
(b) -1 Precision of LiveJournal-LastFM with different features



(b)-2 Precision of LiveJournal-LastFM with different features combinations



(c)-1 Precision of LiveJournal-Myspace with different features



(c)-2 Precision of LiveJournal-Myspace with different features combinations

Figure 4. Precision of SFUI with different structural features

Figure 3 and Figure 4 is recall and precision values when we use different structure feature in second step of matching. Instead of using all the feature, we use only one feature to calculate scores of user pairs who have same number of shared friends. The Recall results of three dataset pairs are shown in (a)-1,(b)-1,(c)-1 of Figure3, and the Precision results are shown in (a)-1, (b)-1, (c)-1 of Figure 4. As we differentiate the measure metric, the results are affected in terms of precision. Especially, Jaccard coefficient, Adamic/Adar score, Clustering coefficient provides similar accuracy regardless of seedset size. Only degree shows notable accuracy difference with other cases. To further exploration, we observe recall and precision of our algorithm with various feature combinations. The recall and precision results are shown in (a)-2, (b)-2, (c)-2 of Figure 3, and in (a)-2, (b)-2, (c)-2 of Figure 4. Through those results, we proves that using combination of various structural feature at the same time helps to reduce the false positive answers and to obtain high accuracy of both recall and precision.

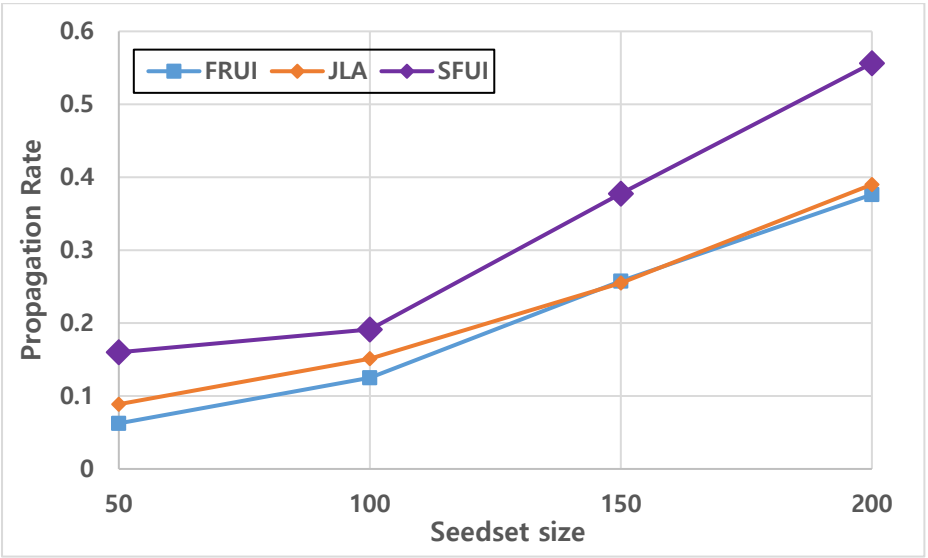


Figure 5. Propagation Rate of various Graph Matching Algorithm

Figure 5 shows the average propagation rate of FRUI, SFUI, and JLA, It proves that FRUI propagtes much further than other method. FRUI shows lowest propagation rate because it uses threshold to obtain accurate matching. I nexperiment FRUI shows much lower accuracy than SFUI when it has relaxed threshold to get good propagation rate.

Chapter 5

Conclusion

In this paper, we study the problem of integration of multiple heterogeneous social networks. We propose a novel graph matching algorithm FRUI to address the problem. FRUI use only structural contents in social networks and achieve applicability to multiple other problems like de-anonymization, and M_NASA problem. In addition, we can obtain noise tolerance. With this algorithm, we also explore diverse graph structural features and prove the effect of these features through experiment. Our experiment is conducted on multiple real world dataset. Through the experiment, the strongness of our algorithm is proved compared to previously suggested algorithms. For the future works, it would be very effective to use other structural features or characteristics that

can explain user identity.

Bibliography

[1] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In CIKM, 2013.

[2] J. Zhang and P. Yu. Integrated anchor and social link predictions across social networks. In IJCAI, 2015..

[3] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, “How unique and traceable are usernames?” n Proc. 11th Int. Conf. Privacy Enhancing Technol., 2011, pp. 1–17.

[4] J. Liu, F. Zhang, X. Song, Y. I. Song, C. Y. Lin, and H. W. Hon, “What’s in a name?: An unsupervised approach to link users across communities,” in Proc. 6th ACM Int. Conf. Web Search Data Mining, 2013, pp. 495–504.

[5] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In KDD’13, pages 41–49, 2013.

[6] F. Abel, E. Herder, G. J. Houben, N. Henze, and D. Krause, “Crosssystem user modeling and personalization on the social web,” User Model. User-Adapted Interaction, vol. 23, pp. 169–209, 2013..

[7] O. De Vel, A. Anderson, M. Corney, and G. Mohay, “Mining e-mail content for author identification forensics,” ACM Sigmod Rec., vol. 30, no. 4, pp. 55–64, 2001.

[8] E. Raad, R. Chbeir, and A. Dipanda, “User profile matching in social networks,” in Proc. 13th Int. Conf. Netw.-Based Inf. Syst.,2010, pp. 297–304.

[9] X. Kong, J. Zhang, and P. S. Yu, “Inferring anchor links across multiple heterogeneous social networks,” in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., 2013, pp. 179–188.

[10] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, “Exploiting innocuous activity for correlating users across sites,” in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 447–458.Y.

[11] A. Narayanan and V. Shmatikov, “De-anonymizing social networks,” in Proc. IEEE 30th Symp. Security Privacy, 2009, pp. 173–187.

[12] S. Bartunov, A. Korshunov, S. Park, W. Ryu, and H. Lee, “Joint link-attribute user identity resolution in online social networks,” in Proc.

6th SNA-KDD Workshop, 2012.

[13] N. Korula and S. Lattanzi, “An efficient reconciliation algorithm for social networks,” arXiv preprint arXiv:1307.1690, 2013.

[14]] X. Zhou, X. Liang, H. Zhang, and Y. Ma. Cross-platform identification of anonymous identical users in multiple social media networks. IEEE Transactions on Knowledge and Data Engineering, 28(2):411–424, 2016.

[15] Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, “COSNET: Connecting heterogeneous social networks with local and global consistency,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 1485–1494.

[16] Wikipedia. (2014). Social networking service [Online]. Available: http://en.wikipedia.org/wiki/Social_networking_service

[17] W. Chen, Z. Liu, X. Sun, and Y. Wang. A game-theoretic framework to identify overlapping communities in social networks. Data Mining and Knowledge Discovery, 21(2):224–240, 2010.

[18] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How unique and traceable are usernames? In Privacy Enhancing Technologies, pages 1–17, 2011.

[19] N. B. Ellison, “Social network sites: Definition, history, and scholarship,” J. Comput.Mediated Commun., vol. 13, no. 1, pp. 210–230, 2007.

[20] M. Hay, G. Miklau, D. Jensen, and D. Towsley, “Resisting structural Identification in anonymized social networks,” in Proc. of the 34th Int. Conf. Very Large Databases, 2008, pp. 102–114.

[21] K. Liu and E. Terzi, “Towards identity anonymization on graphs,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2008, pp. 93–106.

[22] X. Ying and X. Wu, “Randomizing social networks: A spectrum preserving approach,” in Proc. SIAM Int. Conf. Data Mining, 2008, pp. 739–750.

요 약

사회망 네트워크의 정보는 최근 다양한 연구 분야에서 사용되고 있다. 사회망 네트워크가 풍부한 정보를 제공함에 따라, 사회망 네트워크의 정보이용은 데이터 희귀 문제의 해결책으로 여겨지고 있다. 예로, 추천시스템에서는 정보부족으로 인해 어떤 추론도 불가능해 일어나는 문제인 콜드 스타트 문제의 해결책으로, 사회망 네트워크의 정보를 사용하고 있다. 하지만, 사회망 하나가 제공하는 정보는 매우 제한적이어서 정보부족 문제를 해결하기에 부족하다. 그래프 매칭 기법은 서로 다른 성질의 여러 사회망 네트워크를 결합함으로써 사회망 관련 연구의 많은 양의 정확한 기반을 제공할 수 있다. 각각의 사회망 네트워크는 서로 다른 서비스를 제공하고, 익명화된 정보들이 주로 연구에 제공되기 때문에 이종 사회망 서비스간의 공통적인 정보를 찾기는 어렵다. 하지만, 사용자에 의해 형성되는 그래프 구조는 이종 사회망 네트워크에서도 비슷하게 유지된다. 이런 특징에 비춰, 우리는 본 논문에서 이종 사회망 서비스를 통합하는 새로운 방법을 제안하였다. 다른 이종 사회망 통합 방법과는 다르게, 우리는 사회망에서의 단순한 진입차수와 출력차수를 이용하지 않고, 자카드

계수와 아다믹/아다 점수, 사회망 계수, 페이지 랭크 등을 사용자의 사회적 지위를 평가하는데 사용하였다. 우리는 실제데이터를 이용한 실험을 통해 제시한 모델이 이전에 제시되었던 사회망 그래프 통합 방법들에 비해 뛰어난 성능을 보임을 증명 하였다.

주요어 : 이종 사회망 서비스, 그래프 매칭, 그래프 구조 특질, 사회망 통합

학번 : 2015-21235