Comparison of Community Detection Methods for Facebook Ego Network

Yixuan He† Kazuya Uesato† Hayato Yamana†‡

† Graduate School of Fundamental Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku Tokyo, Japan ‡ Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku Tokyo, Japan

‡ National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

E-mail: {pandaxuan, k.uesato, yamana}@yama.info.waseda.ac.jp

Abstract: People tend to build and maintain their friendship relying on SNS nowadays. Thus, the problem of how to organize the social network accurately and automatically has become important to be considered. As this problem draws a lot of attention from both users and companies, there already existed many SNS providing such functions, such as Twitter list, Facebook lists, etc. However, the functions are laborious, means it requires users to manually organize and update their friends' circles or lists every time they add new friends. Instead of using the functions, we may adopt community clustering methods for SNS. In this research, we applied three different common clustering methods to SNAP Facebook dataset and compared the accuracy of these three methods to confirm the adoptability for organizing social network accurately and automatically. Our result shows that LDA performs the best because both of its complexity compared to k-means and of free from setting the number of communities compared to HDP.

Keyword: Facebook, Data Clustering, Data Mining

1. INTRODUCTION

Recently, with the development of computer and capable devices, it is hard to imagine life without the Internet. Also, with the development of the Internet, the new communication media and technologies, SNS, which refers to Social Networking Services, have drawn more and more attention from users and researches.

As one of the most popular and successful Social Network Services, Facebook, undeniably plays an indispensable role in providing a great platform for people and friends not only to share their own feelings, but also to make new friends, maintain old ties, and deeper current relationships. By using Facebook, people can also update their friends' current life, for instance, how other students in the same university think about some classes, how other people in the same fashion club think about the latest trend. Same as real life, usually people can also define and organize their friends into different friend circles.

Therefore, it is common for users to expect that feeds posted by the people in the same group could be organized to show up at the same time. Currently, most of the famous Social Network Service, such as circles on Google+, lists on Facebook and Twitter, etc. provide such functions. However, generally they all either require users to manually categorize their friends into those groups or use a naïve fashion by clustering friends who share some common attributes. Neither works particularly satisfactory. Firstly, it is laborious for users since every time they add some new friends or every time they grow some new common attributes with their old friends, a manually updating of the groups is required. The second works even poorly especially when the profile information is missing or withheld.

In this paper, in order to confirm the adoptability current clustering methods for organizing social network accurately and automatically, we compare the accuracy of different clustering methods, k-means, Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP), using Facebook Data from SNAP, which refers to Stanford Network Analysis Platform. Then, we evaluate the results by using the ground truth provided by SNAP[1]

We choose k-means because we can apply this method with different values of k. Similarly, LDA also requires an assignment of the number of circles, k. However, different from k-means, LDA assigns a document to a mixture of topics. Each document is categorized by one or more topics, for instance, 50% Topic A, 30% Topic B, 20% Topic C, etc. Hence, in most case, LDA gives a better result than k-means because of its complexity compared to k-means. So, in this paper, we want to confirm which method is the best in the three method for determining friends' circles using SNAP Facebook dataset. While the

number of groups must be pre-specified for the two methods above all, the number of k could be automatically inferred from the data when using HDP. In our research, we also use HDP to see how different it is when using LDA and its advanced version, HDP.

In the reminder of the paper, in Section 2, we present the related work in ego network field. And in Section 3 and 4, we describe the methodology employed and the subsequent experimental results, respectively. Section 5 aims to give the evaluation results of our experiments. Finally, we conclude our paper and discuss about the limitation and future work of this paper in the last section.

2. RELATED WORKS

There is a large amount of relevant research based on Facebook and ego network. However, most of the researches focus on either improving the accuracy of community detection by combining existed methods or creating new models, little of them tried to compare the effect of different clustering methods.

Furthermore, due to the difficulty for non-Facebook research to obtain the Facebook data, many research based on Facebook are written by people who work in Facebook. SNAP Facebook Dataset is a dataset provided by Stanford and has not drawn much attention from researchers and thus has been used in a small number of research papers. To the best of our knowledge, our paper is the first paper comparing k-means, LDA and HDP using SNAP Facebook dataset.

2.1. Research related to Facebook

Sungkyu et al.[1] presented how activities on Facebook are associated with the depressive states of users. Their methodology was based on a Web application implemented within Facebook and conducting face-to-face interviews. Their study was novel in some perspectives: both online Facebook data and offline interview data within some time period are used, these data allow them to compare the relevance between the changes of users' depression rate and the changes of Facebook usage. There are also several limitations in their research: most of the respondents are found to be more educated and are disproportionately in male for instance.

In Zubeida's work et al. [2], they presented an analysis of graph search on Facebook in order to highlight how easy it is to access large amount of personal information though Facebook's graph search. Besides, they determined whether the increase of Facebook friends' number has an impact on the increase of the searching results

Lars et al. [3] proposed a method to recognize one user's romantic partner using Facebook data by developing a new measure of tie strength – "dispersion", which implies two people's mutual friends are not well-connected.

In [4], Adrija identified potential social ambassadors of Facebook brand pages of mobile service providers and evaluated the results to be valid by matching with the opinions provided by domain experts. In their future work, they also mentioned the hindrance of data collecting because of the Facebook privacy setting.

2.2. Research related to Ego Network

In R.I.M et al. work [5], they utilized data on frequencies of bi-directional posts to define edges and create ego network using two Facebook datasets and one Twitter dataset. They clustered the dataset into 4 clusters each of which is called "layer." The frequency of contact within each layer varies depending on layers. For example, in the Facebook data case, the contact frequencies are 1.5, 4.3, 10.7 and 27.0 for each cluster. They analyzed the data using two different clustering techniques: k-means and DBSCAN, a density-based clustering technique. They obtained similar results and declared that k-means works good enough if considering the simplicity of k-means.

The work of Jaewon et al. [6] chose 13 different commonly used structural definitions of network communities and examined their robustness and sensitivity, respectively by using a set of 230 large real-world social networks in SNAP. Although their research did a comparison by methods using SNAP data, they did not use any small-sized, anonymized data such as SNAP Facebook dataset.

In another work written by Jaewon and Jure et al. [7], they presented a method called BIGCLAM, which is an example of a bipartite affiliation network model. They successfully dealt with the detection of overlapping, hierarchically nested, as well as non-overlapping communities with relatively higher accuracy.

Julian et al. [8] extracted communities from both the edge structure and node attributes (CESNA). They improved the accuracy of community detection with relatively high speed. Moreover, besides detecting the communities, it can also help to find the interpretation of the detected communities.

3. METHODOLOGY

3.1. Dataset

In this research, SNAP Facebook Dataset is used. This

dataset consists of information of friends' circles from Facebook, for instance, the profiles for each user (features), circles and network. Data is collected from participants who voluntarily provided their Facebook information to a Facebook app. To protect personal information, all the data has been anonymized and the interpretation has all been obscured. The details of this dataset is shown as below:

Tab. 1 Data Statistics of Facebook[1]

Dataset statistics	
Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2647
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

In this Dataset statistics, "nodes" represents the number of friends; "Edges" represents the number of edges in the network; "Nodes in largest WCC" and "Edges in largest WCC" represents the number in the largest weakly connected component of nodes and the number of edges respectively; "Average clustering coefficient" means a measure of the degree to which nodes in a graph tend to cluster together. "Number of triangles" represents the number of triples of connected nodes (considering the network as undirected), "Fraction of closed triangles" represents the number of connected triples of nodes. "Diameter (longest shortest path)" represents the maximum undirected shortest path length and "90-percentile effective diameter" represents the 90-th percentile of undirected shortest path length distribution.[1] They also obtained the ground-truth by hand labeling by volunteer participants.

Thus, we can see that in our dataset, there are 4,039 nodes and 88,234 edges in total. Since this Facebook graph is undirected and consists of friends of one ego node, which makes it possible to reach any nodes from any nodes in the graph, the number of nodes and edges in largest Weakly Connected Component and in Strong Connected Component keep the same to the original value. Besides, due to the small number of participants, the average clustering coefficient is also small. In our research, we use their ground-truth as our ground-truth.

3.2. Applying k-means, LDA and HDP

Both K-means and Latent Dirichlet Allocation (LDA) are unsupervised learning algorithms. It means before clustering, the user should decide a priori value for the parameter K. In our research, in order to simplify the process of evaluation, we set the value of K equals to the value of the ground truth. For example, for ego node 0, if the ground truth of friends' circles is 24, we assign the K equals to 24. Unlike the k-means and LDA, HDP does not require a priori decision of K. Instead, it can automatically assign an appropriate value of K according to the data that is given. In our research, we used sklearn[2], which is a machine-learning library for Python, to calculate the clustering results of k-means. For both LDA and HDP, we use gensim[3], which is a topic-modeling library for Python.

4. RESULTS AND EVALUATION[1]

In this section, we will show the results of clustering. In 4.1, we introduce the evaluation method that we use in our research. In 4.2, we will present the evaluation results.

4.1. Evaluation Method

In this paper, we adopt Purity[9], which is a simple and transparent evaluation measure, as our evaluation method. The calculation of Purity follows two steps. Firstly, we label each cluster as the class in which the most frequent class is included. Second, the accuracy is computed by counting the number of correctly assigned data followed by dividing it by N, which is the total number of the points. For example, we assume that we have clusters as showing below:

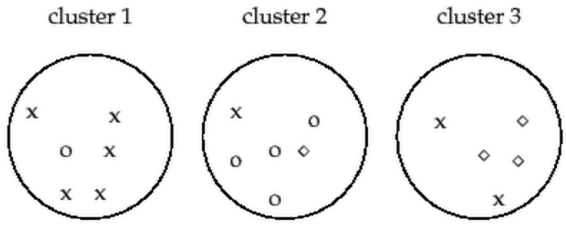

Fig. 1 Example of Purify [9]

It is obvious that, crosses occur the most in cluster 1, circles occur the most in cluster 2 and diamonds occur the most in cluster 3. So, when we evaluate the accuracy of the clustering results, we set cluster 1 as "Cluster Cross", cluster 2 as "Cluster Circle" and cluster 3 as "Cluster

[1] "Stanford Network Analysis Project", Jure Leskovec, http://snap.stanford.edu/index.html, access at Jan 6, 2015

[2] scikit-learn, http://scikit-learn.org/stable/, access at Jan 6, 2015

[3] gensim, https://radimrehurek.com/gensim/, access at Jan 6, 2015

Diamond". The total number of dots is 17, so the Purity can be calculated as (1/17) ×(5+4+3) ≈ 0.71 [9]

The formula can be represented as:

$$\text{purity}(\Omega, Ç) = \frac{1}{N} * \sum_k max_j |\omega_k \wedge c_j| \qquad (1)$$

where $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4 \ldots, \omega_K\}$ is the set of clusters and $C = \{c_1, c_2, c_3, \ldots, c_J\}$ is the set of classes.

4.2. Results and Evaluation

The clustering result of our evaluation is shown in Table 2. For k-means and LDA, since the k is pre-set to be the same as the ground truth, we can compute the accuracy directly. However, for HDP, since the number of communities is automatically assigned, in order to calculate the Purity, we combined small groups together to have a set of large clusters. We did the experiments four times. The results keep changing each time we run the program, which is because for k-means, it uses a randomly chosen starting configuration thus different starting point gives different clustering results, and for LDA and HDP, these two methods use randomness when training and in inference steps thus also returns different results. From the results, we can find that LDA outperforms the other two methods in general. But it does not show a big difference among these three methods. Thus, if considering the simplicity of k-means, we confirm that k-means is good enough for community detection when using small size dataset.

Tab. 2 Evaluation results

Experiment trial	Accuracy (k-means)	Accuracy (LDA)	Accuracy (HDP)
1st time	0.2589	0.2857	0.2679
2nd time	0.2321	0.2500	0.2232
3rd time	0.2411	0.2723	0.2366
4th time	0.2545	0.2813	0.2277

5. CONCLUSION

In this paper, we compared three different commonly used clustering methods using SNAP Facebook dataset. In general, LDA performs the best because of its complexity compared to k-means and the possibility of assigning the number of communities compared to HDP. However, we can also find that there is not a big difference among the results by using these three methods. Thus, due to the simplicity and convenience, we confirm that k-means works well when categorizing, detecting the communities.

There is still some future work that could be done. From the results, we can see that the accuracy is not good enough. There are many perspectives that need to be taken into consideration. Firstly, in our research, we only consider the features of users' profiles information while edge information, which refers to the connections between friends, is also possibility to be utilized. Secondly, in the features files, the rate of the appearance of 1s and 0s is very low. It means, most of the features are 0s. The problem of the rarely occurrence of 1s need to be concerned. Although we chose these three clustering methods because of its popularity of using, it is possible these three methods are not suitable for this kind of small-size, anonymized dataset. Last, in our research, we did not consider overlapping, hierarchically communities.

REFERENCE

[1] Sungkyu Park1 Inyeop Kim2← Sang Won Lee3 Jaehyun Yoo3 Bumseok Jeong3 Meeyoung Cha, Manifestation of Depression and Loneliness on Social Networks: A Case Study of Young Adults on Facebook, Proc. of *CSCW '15, Vancouver, BC, Canada,* 2015

[2] Zubeida Casmod Khan, Thulani Mashiane, An Analysis of Facebook's Graph Search, Proc. of *Information Security for South Africa (ISSA)*, 2014

[3] Lars Backstrom, Jon Kleinberg, Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook, Proc. of *CSCW'14, Baltimore, Maryland, USA,* 2014

[4] Adrija Majumdar, Debashis Saha, Parthasarathi Dasgupta, An analytical method to identify social ambassadors for a mobile service provider's brand page on Facebook, Proc. of *2015 Applications and Innovations in Mobile Computing (AIMoC)*, 2015

[5] R.I.M. Dunbara, Valerio Arnaboldia,b, Marco Contib, Andrea Passarellab, The structure of online social networks mirrors those in the offline world, Social Networks, Vol.43, pp.39–47, 2015

[6] Jaewon Yang, Jure Leskovec, Defining and Evaluating Network Communities based on Ground-truth, *Proc. of 2012 IEEE International Conference on Data Mining (ICDM)*, 2012

[7] Jaewon Yang, Jure Leskovec, Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach, Proc. of *WSDM'13, February 4–8, 2013, Rome, Italy*, 2013

[8] Jaewon Yang, Julian McAuley, Jure Leskovec, Community Detection in Networks with Node Attributes, Proc. of *2013 IEEE 13th International Conference on Data Mining,* 2013

[9] Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008