Pattern Mining on Ego-Centric Networks of Friendship Networks

Nobuhiro Inuzuka, Shin Takeuchi, and Hiroshi Matsushima

Nagoya Institute of Technology,
Gokiso-cho Showa, Nagoya 466-8555, Japan

inuzuka@nitech.ac.jp, takeuchi@nous.nitech.ac.jp,
matsushima@nous.nitech.ac.jp

Abstract. The paper proposes a procedure to analyse local patterns of connectivity among people in social networks using the idea of ego-centric network. The ego-centric networks of every nodes are transformed into normalized representation and classified into patterns. The procedure can be applied to large dataset by giving it in SQL code. We applied the procedure to friendship networks and demonstrated distinguished properties compared to other networks. We found out that friendship network contains large variety of patterns.

1 Introduction

In this paper we concern data mining about patterns of social networks surrounding individuals in a group of people. The social network is a field developed in studies of the sociology. It has been actively investigated since 1970's in order to understand features and roles of individual and gropes through relationship among people and structure of networks about such as friendships and influence[1, 6, 7]. Then the area of network science is evolved as a science to understand people or entities which are influenced each other by revealing characters of the structure of the connection. Network science helps to analyse roles of persons in organisations and business trades among companies. It is also an interesting topic to apply the methodology of network analysis for online data, such as link analysis of world wide web, activity in social network service, and online shopping services.

Most of social network studies analyse networks and entities in the networks from the global points of view, that is, using some measure through whole networks, such as by the property of power law and the cluster coefficient. On the other hand we may bring other ideas from data mining technique. Pattern mining, i.e., enumeration of frequently appeared patterns, in networks is a straight forward ideas there. Sociologists discuss local behaviours, that is, detailed behaviours of individuals with surrounding areas of the individuals in social networks. While such detailed analysis helps to develop knowledge about entities in the networks, it is not suitable for large scale networks.

In order to overcome the difficulty we use the idea of frequent pattern analysis and the idea of ego-centric network. An ego-centric network is a network among people who is directly connected to a particular individual. Ego-centric networks are used to discusses roles of people in influence relationship among people[5, 7]. In this paper we give a procedure to classify ego-centric networks and apply it to friendship networks. By the application we demonstrate a significant characteristic of friendship networks comparing with other networks which share other global measures.

2 Ego-centric networks

Let V a set of entities (individuals) and consider an undirected graph $G = (V, E)$, where $E \subseteq V \times V$ where an undirected edge $e = (v_1, v_2) \in E$ represents some influence between v_1 and v_2, typically we understand the influence as friendship.

In sociology an individual in question is called *ego* and other individuals *alters*. An *ego-centric network* of an ego is a local network consisting of the ego and alters who are directly connected to the ego. When we use the terminology of graph theory we can define it in precise. An ego-centric network is a subgraph induced from the set of nodes consisting the ego and alters. Technically it is represented by a triple including the ego as defined bellow.

Definition 1 *For an undirected graph $G = (V, E)$ and an ego $c \in G$, an* ego-centric network *of G with respect to c is a triple $G_c = (c, V_c, E_c)$, where $V_c = \{v \in V \mid (c, v) \in E\} \cup \{c\}$ and $E_c = \{(u, v) \in E \mid u \in V_c \wedge v \in V_c\}$.*

In order to count patterns of ego-centric networks in whole network we define isomorphism to match two networks.

Definition 2 *Two ego-centric networks $G_1 = (v_1, V_1, E_1)$ and $G_2 = (v_2, V_2, E_2)$ are* isomorphic *when there is a bijection f from V_1 to V_2 which satisfies that $f(v_1) = v_2$ and $\forall v, w \in V_1$, $(v, w) \in E_1$ if and only if $(f(v), f(w)) \in E_2$.*

Definition 3 *The* frequency *of an ego-centric network G_c in $G = (V, E)$ is the number of nodes in V whose ego-centric networks are isomorphic to G_c.*

3 Classification of patterns

An ego-centric network of an ego represents a pattern of relationship where the ego is laid. We give a procedure to enumerate and classify patterns of ego-centric networks or nodes in the whole network.

It is shown that the procedure to know two networks are isomorphic is hard problem we only know a procedure of exponential complexity with respect to the number of nodes. The problem are in the class NP but it does not become clear that the problem is in the NP-complete nor P.

We gave a procedure for enumeration. In order to see isomorphic matching we need to give a canonical form of graphs, which must be unique for all isomorphic graphs. For this purpose we give a number or a rank to each node which is given by only structural property.

The idea is to give a rank to a node by the connection to neighbouring nodes and also by neighbours of the neighbours, and so on. This can be constructed by iterative procedure which renew ranks of nodes by deepening to see neighbours.

The procedure is described as follows.

1. Sort and give rank to nodes by its degree in descending order, the number of nodes connected to the node. Give the same rank to nodes with the same degree.
2. Obtain a list of ranks of neighbouring nodes for each node. We do not include the ego in the list. Ranks in each list are put in the order of rank.
3. Sort and renew the ranks of nodes by the lists of neighbours' ranks in the lexicographic order.
4. If the renewal does not change the ranks it stops. Otherwise continue the procedure from step 2.

Table 1 demonstrates this procedure using the graph in Fig. 1. The graph has seven nodes including the ego A. Step 1 of the procedure sorts nodes by their degrees and gives ranks 1, 2, 2, 2, 5, 5 and 7 to A, D, E, F, B, C and G, respectively. Step 2 makes lists of ranks of neighbours of nodes. For example node D has non-ego neighbours B, E and F and so a list $\langle 5, 2, 2\rangle$ is obtained. The nodes in the list sorted as $\langle 2, 2, 5\rangle$. These results are describes in the first round column in the table.

By this procedure the nodes B and C are discerned by their lists $\langle 2, 7\rangle$ and $\langle 2, 2\rangle$. and are given different ranks. By sorting the neighbours' rank lists in the lexicographic order the rank of nodes are renewed as shown in the rank of second round column. Then again neighbours' rank lists are described using the renewed rank as in second round column. Here the node D is discerned from E and F and is given different rank. The renewed ranks are in the node column of third round. Again neighbours' rank lists are given using the renewed ranks as in third round column. The lists are different from the second round but it does not make any renewal for ranks and then the procedure terminates.

Table 1. An example of the procedure to give normal rank to networks.

		first round			second round			third round		
node	degree	node	rank	neighbours	node	rank	neighbours	node	rank	neighbours
A	6	A	1	2 2 2 5 5 7	A	1	2 2 2 5 **6** 7	A	1	2 2 **4** 5 6 7
B	3	D	2	2 2 5	D	2	2 2 **6**	E	2	2 **4** 5
C	3	E	2	2 2 5	E	2	2 2 5	F	2	2 **4** 5
D	4	F	2	2 2 5	F	2	2 2 5	D	**4**	2 2 6
E	4	B	5	2 7	C	5	2 2	C	5	2 2
F	4	C	5	2 2	B	**6**	2 7	B	6	**4** 7
G	2	G	7	5	G	7	**6**	G	7	6

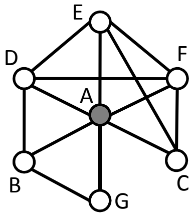


Fig. 1. An example of an ego centric network which is used for illustration in Table 1. Shadowed node is the ego.

After the procedure of normalizing rank, we give unique identifier to each node. The ranks given by the procedure may be identical for some nodes and then we can not use the ranks as identifiers. We give identifiers arbitrarily as they do not violate the order of ranks. For the example, the ranks of A to G was 1, 2, 2, 4, 5, 6, and 7 and then we give identifiers them 1, 2, **3**, 4, 5, 6, and 7. Using these identifiers the example ego-centric network is represented in the matrix as in Table 2. When two ego-centric networks are isomorphic if and only if they have the same representation in the matrix form.

This procedure can easily described in an SQL code. We gave an SQL code in order to process large network data as in database system.

Table 2. Representation of the graph in Fig.1.

	1	2	3	4	5	6	7
1	0	1	1	1	1	1	1
2	1	0	1	1	1	0	0
3	1	1	0	1	1	0	0
4	1	1	1	0	0	1	0
5	1	1	1	0	0	0	0
6	1	0	0	1	0	0	1
7	1	0	0	0	0	1	0

4 Experiments and comparison models

We applied the procedure of pattern classification to friendship networks. The friendship networks were the network constructed using class attendance records to lectures in Nagoya Institute Technology, the records which are collected in 2007 for all undergraduate students in 2007. The institute collects attendance records to manage classes and to analyse relationship between student behaviours and their scores.

Inuzuka et al.[2] gave a procedure to predict friendship relation from the attendance records based on the conjecture that friend two students likely act together and so the time differences between attendance time of the two students to classes are short. The prediction results relatively high accuracy and so we use the friendship network produced from the prediction for our purpose. Matsushima et al.[4] studied network properties of the friendship networks predicted in [2] and showed that the friendship network shares the power law of degrees of nodes, high cluster coefficient, and short average path among nodes (small world property) with other many social networks.

We only used networks for freshmen (first year students) and sophomores (second year students). Freshman network includes 931 nodes and sophomore network includes 939 nodes. 2950 pairs and 3403 pairs of freshmen and sophomores, respectively, are predicted to have friendship relation, which corresponds to edges in network. We do not take the data for higher grade students because they have different style of curriculum.

We prepared two types of networks, random networks and a network generated by node deactivation model[3] for comparison. We generated a random network by randomly choosing 2950 edges for 931 nodes from all possible edges. The numbers of edges and nodes are the same as the network for freshmen.

The node deactivation model (DA-model in short) can generate networks which posses the three properties shared with social networks. DA-model generates network as follows. In the procedure each node keeps a flag information of active or inactive.

1. Take a complete graph with k nodes as an initial network. A complete graph is a graph in which every pair of nodes are connected. The k becomes the average degree of the network resulted. Let the flag of the all nodes active.
2. Add a node to the network and connect the node to all active nodes. Let the flag of the new node active.
3. Choose a node from active nodes in the probability proportional to the inverse of their degrees. Then let the flag of the node inactive.
4. Continue Step 2 till necessary number of nodes are added.

By the above procedure we obtain a network of average degree of k. We can only choose an integer average degree. In order to adjust to the average degree to the average degree 3.4 of friendship network of freshmen, we need another step, that is, after we a network of average degree 4 was obtained, edges were cut randomly to adjust the number of edges to friendship network. As a result the numbers of edges and nodes are the same as the network for freshmen.

5 Experimental results

We obtained 397 and 486 patterns of ego-centric networks from freshmen and sophomores friendship networks, respectively. On the other hand, 33 to 37 patterns from random networks and 227 patterns from networks generated from DA-model are appeared. Fig. 2 shows the relation between the frequent patterns

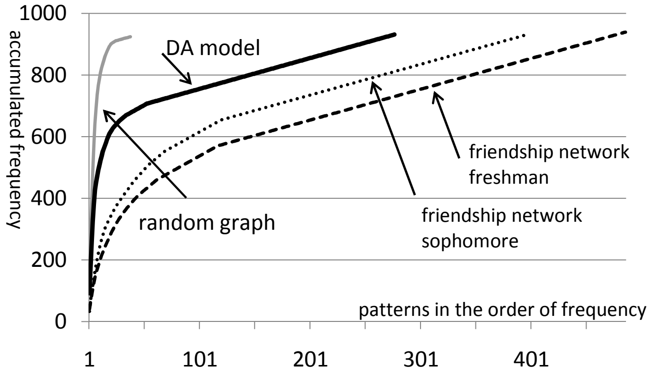


Fig. 2. Relation between frequent patterns and coverage in the networks.

and the coverage over the whole networks. The value of 10 on X-axis shows the tenth frequent patterns and the value of Y-axis is the accumulated number of nodes whose ego-centric networks are isomorphic to some of patterns before the tenth frequent pattern.

While in random network and DA-model's network, small number of frequent patterns cover a large part of networks, it is not true in friendship networks. The 20 most frequent patterns cover 98% of nodes in the random network and those patterns in DA-model's cover 66%. In friendship networks 36% and 30% of nodes are covered by the 20 most frequent patterns for freshmen and sophomore, respectively.

Three graphs in Fig. 3 show relation between the degrees of egos in ego-centric network patterns and other measures. The graph (a) is the relation between the degree of egos and the number of different ego-centric patterns with the degree. The friendship networks have large peeks in 6 to 9. This means friendship network has large variety of patterns especially in gropes of 6 to 9 people. DA model's has also a peek in 6 to 8 but the peek is smaller. Random network has very small number of patterns.

The graph (b) shows the degree distribution in the networks, i.e., the relation between the degree of egos and the number of total number of nodes which have egos with the degree. Friendship networks have similar distribution to the random networks.

The graph (c) shows the the relation between the degree of egos and the average frequency of patterns whose egos have the degree. We can observe that each patterns in friendship network covers only small parts compared to random networks and also DA-model's.

Fig. 4 shows the ten most frequent patterns appeared in the networks. Interesting difference among networks are appeared. Most of frequent patterns in random network is star graphs. Complete graph or dense clusters which can be

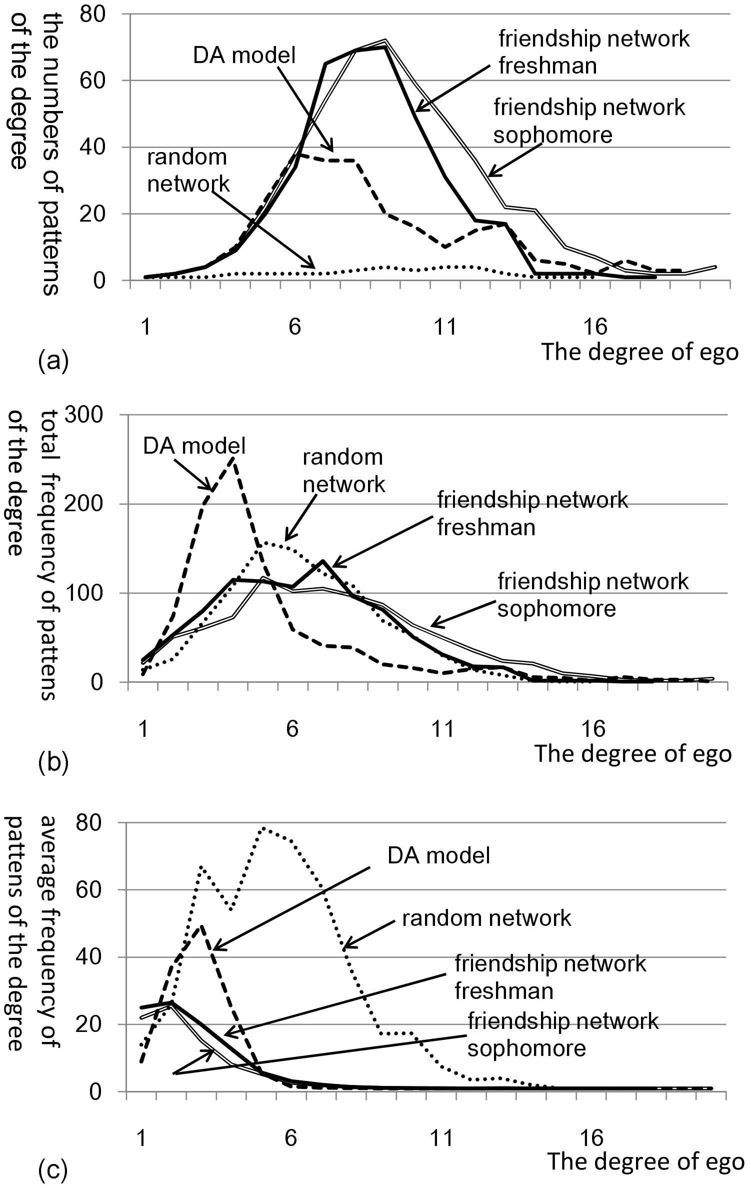
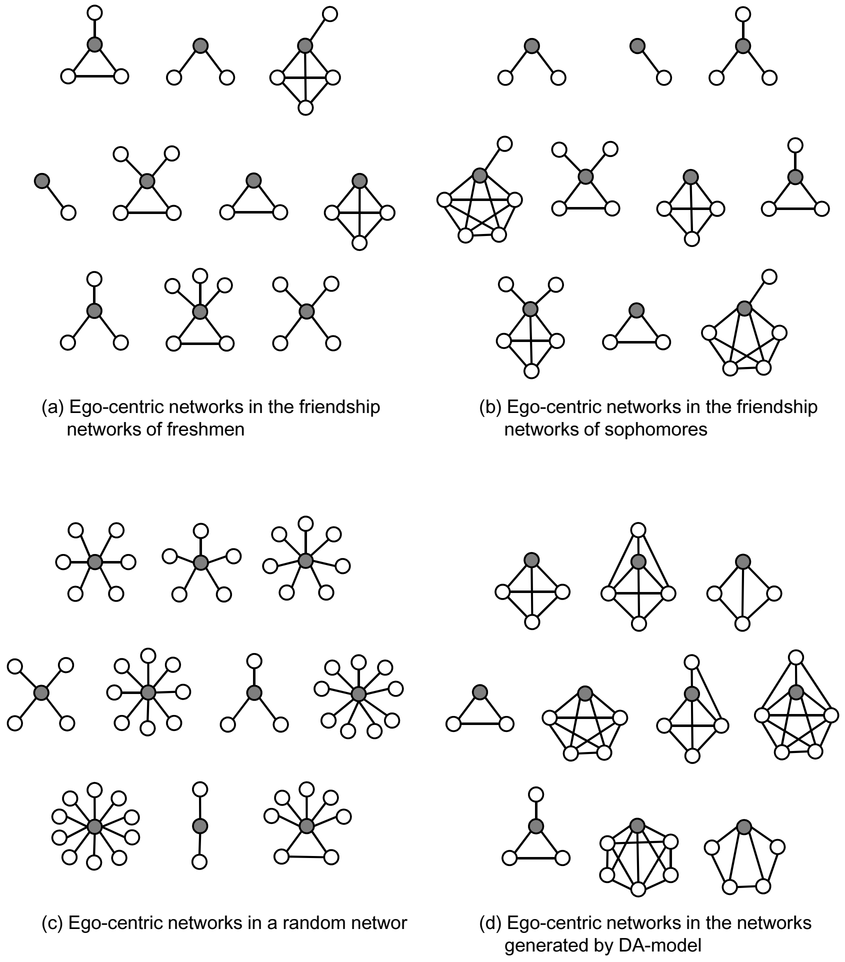


Fig. 3. Relation degrees of ego in patterns and (a) the number of patterns, (b) the total frequency of patterns and (c) the average frequency of patterns with the degree.



(a) Ego-centric networks in the friendship networks of freshmen

(b) Ego-centric networks in the friendship networks of sophomores

(c) Ego-centric networks in a random networ

(d) Ego-centric networks in the networks generated by DA-model

Fig. 4. The ten most frequent patterns appeared in the networks. The shadowed nodes are egos. In each span, the upper left pattern is the first frequent, the upper middle is the second, and so on.

seen as a graphs lacked small number of edges from complete graphs are appeared in DA-model's network. On the other hand, we find in friendship networks other characters. In many frequent patterns in friendship networks, an ego plays a role of bridge between clusters and other nodes. We note that the nodes not in the cluster are not necessarily isolated but may connect to other clusters, because we observe only ego-centric patterns.

6 Conclusion

We developed a classification procedure for ego-centric network and applied it to friendship networks. As a result, we observed unrevealed characteristics in friendship networks by this analysis. We observed that friendship networks contains very large variety of patterns in their ego-centric networks. It is remarkable that the friendship networks have distinguished property from DA model's network even though friendship networks and DA-model's share the properties of social networks, i.e., the high cluster coefficient and the power law. The mechanism to constitute friendship relation may cause the rich variety of patterns, while we left the further study of evolution of patterns in friendship networks. Association between patterns and the attributes of individuals are also interesting topic in the future works.

References

1. A.-L. Barabási and R.Albert. Emergence of Scaling in Random Networks. Science, 286, 509, 1999.
2. N.Inuzuka, T.Nakano and K.Shimomura. Friendship Analysis Using Attendance Records to University Lecture Classes. Proc. IASK Int'l Conf. Teaching and Learning, 478-486, 2008.
3. K.Klemm and V.M.Eguiluz. Highly clustered scale-free networks, Physical Review E, 65, 036123, 2002.
4. H.Matsushima, S.Kadosaka, S.Yamamoto and N.Inuzuka. Analysis of Friendship Network Using Attendance Records to Lecture Classes, 30th Sunbelt Conf., (Tech. Rep., Inuzuka labo.), 2010.
5. M.E.J.Newman. Ego-centered networks and the ripple effect -or- Why all your friends are wired. Social Network 25, 83-95, 2003.
6. M.E.J.Newman and M.Girvan. Finding and evaluating community structure in networks, PHYSICAL REVIEW E 69, 026113, 2004.
7. S.Wasserman and K.Faust. Social Network Analysis, Cambridge U. Press, 1994.