

Role analysis for profile generation in heterogeneous social network

Nishanthi.R

Department of Information Technology,
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai-603110
nisharavi12@gmail.com

Dr.S.Karthika

Department of Information Technology,
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai-603110
skarthika@ssn.edu.in

Abstract— The social network is a powerful data structure allowing the study of relationship information between entities. The existing methods only rely on analyzing homogeneous social network assuming only a single type of node and relation. Though, real world complex networks are not so. Usually it is a heterogeneous social network which assumes a network with different types of nodes and relation. A Multi Relational Network (MRN) is used to describe the nodes and links in the form of a directional semantic graph where each node is related with more than one relationship with others. The profile for each node can be generated based on the relation sequence of each node. To analyze the relation of node in the heterogeneous social network, role analysis is done in the profile creation. The evaluations are conducted on a real-world movie dataset with promising results.

Keywords—profile generation; relation sequence; role analysis; heterogeneous social network;

I. INTRODUCTION

A social network is a graph in nature, where the nodes stand for actors (e.g., director, actor and writer) and the edges between two actors represent their relationships (e.g., spouse of, write script and direct). In Social Network Analysis (SNA), people have proposed different measures for the graph structure to model some general phenomena or to capture some hidden properties, like the well-known small-world phenomena [7]. Analyzing a social network can not only assist experts in understanding the social phenomenon but also help laymen manage their social circles. The goal of such analysis is to enable us to deduce new relations, reveal potential vulnerabilities, and identify an attack before it occurs.

In the context of social network mining, identifying similar nodes can be divided into two categories. The first is to find communities or their structures [10]. A community is a sub-graph containing tensely intra-connected edges within it and loosely inter-connected edges across communities.

The second is to determine the network positions (or social roles) [4] of entities playing similar roles or having close semantics in the network.

Although there are already various successful proposals for social position analysis, most assume there is only single type of nodes and single type of relations in a network. This kind of social network is defined as homogeneous social networks [4]. However, in the real-world different types of objects can be connected through different kinds of relationships, therefore it is natural to define different types of entities and relations in a social network. In this sense, a more universal data structure, termed heterogeneous social network [4] describes the complex relationships (i.e., typed edges) among entities. For example, a heterogeneous movie network shown in Figure 1 takes movies (M), directors (D), writers (W), and actors (A) as nodes, and their corresponding relationships as tuples such as $\langle D, \text{direct}, M_1 \rangle$, $\langle M_1, \text{has actor}, A \rangle$, $\langle M_1, \text{originate from}, M_2 \rangle$, where the letter in the tuple stands for the type of source node, and the second element stands for the type of relations.

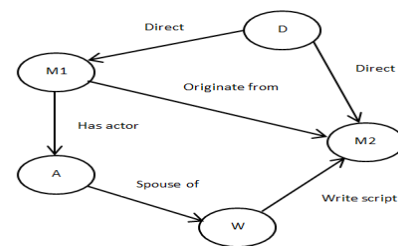


Fig.1. A heterogeneous social network sample for movie domain.

The capital letter of each node stands for its type: M (movie), D (director), A (actor), and W (writer). Besides, there are five relation types, including “write script”, “has actor”, “spouse of”, “direct”, and “originate from” in this example.

A heterogeneous social network in fact contains significantly more semantic information than a homogeneous one. Therefore, applying homogeneous analysis methods to it could lead to a loss of information in the process [20], so it is better to take both the typed relations and the topological information of nodes or links into account.

The remainder of this paper is structured as follows. In Section 2, the various supporting works for profile generation can be provided. In Section 3, the problem statement along with the assumptions made can be illustrated and Section 4 explains about profile generation framework. In Section 5 semantics of nodes in heterogeneous network is modeled. In Section 6, experimental study can be made and in the Section 7 the paper is concluded with the future work.

II. RELATED WORK

There are several works related to preprocessing issues, tasks and detecting communities in a homogeneous social network. The general approach to find dense sub graphs is by partitioning the graph recursively. Recently, researchers have proposed the modularity-based approach [6][8] for detecting communities. The idea behind modularity is to ensure the number of edges across groups is not only small but also smaller than expected. Besides, W. Hwang et al. [9] propose the bridging centrality integrating the global and local features to identify bridges between communities, and then remove some edges from the network by the Bridge-Cut algorithm to form several cohesive sub graphs. SCAN algorithm [15] defines structural similarity as the base to present a density-based structural clustering in a bottom-up manner. V. Satuluri et al. [18] propose utilizing stochastic flows for community detection. Despite their great success in homogeneous networks, none can be easily adopted in the heterogeneous domain.

For heterogeneous social networks, the definition of community can differ from homogeneous social networks. A heterogeneous community does not have to process dense connections within a certain relational graph, but its members might share similar and frequent interactions with communities of other relational graphs. Spectral relational clustering [16] is one of the most well-known approaches to identify communities in a heterogeneous network, which formulates the problem into factorization on multiple matrices. Instead of partitioning the graph regarding patterns of interaction, we aim at grouping nodes based on their roles.

D. Cai et al. [5] address another kind of community detection problem in heterogeneous networks through learning an optimal linear

combination of a user-queried relational structure. More recently, Net Clus algorithm [2] addresses new kind of clusters in a heterogeneous network, where each member in the cluster is a sub-graph with star-shaped schema. However, their solution is restricted to this specific schema and therefore cannot deal with higher-order relational information.

It is about generating the compact summarized representation for a large graph. L. Zou et al. [11] propose summarizing a graph using the topological information of the original homogeneous graph. It is not a trivial matter questioning how their approach can be adopted to heterogeneous graphs. Y. Tian et al. [13] introduce the OLAP-style operations to summarize multi-relational graphs, in which users can apply drill-down and rollup to control summarized resolutions. However, they only use the immediate links of nodes and the high-order relationship information is ignored.

Network visualization aims at efficiently displaying a large network by drawing the structural data with some simple analyses for human explorations. P. Appan et al. [1] summarize key activity patterns of social networks in the temporal domain using a ring-based fashion. L. Singh et al. [19] develop a visual mining program to help people understand the entire multi-mode networks at different abstraction levels, in which the abstraction is performed by merging or dividing among different types of entities.

Shen et al. [3] divide abstraction into structural and semantic parts, and present a visual analytics tool, Onto Vis, where the relations in heterogeneous networks are reduced based on the concept of network ontology. However, all three suffer from insufficiently providing egocentric views to facilitate explorations. Besides, they consider simply links in the one step neighborhood of each node. We argue that high-order topological and relational information should be modeled to produce more meaningful abstraction from diverse aspects through the existing abstraction ideas with the proposed signature profile model.

This refers to the hidden structural backbone of the network in a macro view. Network skeletons preserve various topological properties of the graph, and thus can be regarded as a kind of abstraction. D. Vincent et al. [14] perform transitive reduction, which is an edge-removing operation without losing reachability between any two nodes, on directed graph data. Du et al. [17] build the backbone graph of the super nodes using the minimum spanning tree algorithm, where the amount of overlap serves as the distance between them.

III. PROBLEM STATEMENT

In this paper, we focus on generating the profile for each node involved in heterogeneous social network which is based on the relation sequence of nodes. To analyze the relation, role analysis is done in the profile creation.

We assume that the dataset collected is complete. We also assume that the social network under consideration is a heterogeneous and not a homogeneous one. Based on these assumptions we have developed a Multi Relational Network (MRN), which is used to describe the relation sequence of each node involved in the heterogeneous social network.

IV. FRAMEWORK FOR PROFILE GENERATION

A general framework for modeling the semantics of nodes can be carried out by integrating the high-order relational information with the graph topology. Thus a relational adjacency matrix can be generated to capture the neighborhood information of a node through Multi Relational Network (MRN). By doing this, a heterogeneous social network can be transformed into an abstracted format and many preprocessing tasks can be facilitated.

The overall framework for profile generation by preprocessing the modeling tasks is illustrated in Fig 2.

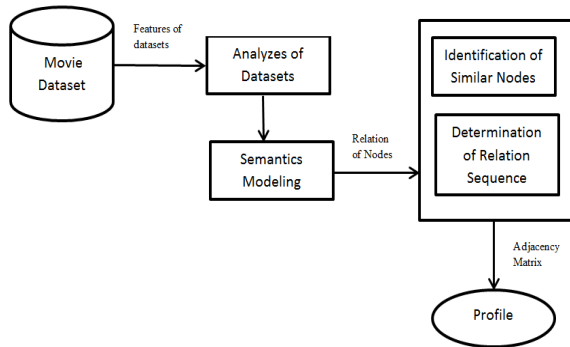


Fig.2. Framework for Preprocessing Tasks

Given a real world movie datasets, in which features of datasets can be analyzed. Based on the features, nodes which play similar roles in the network can be identified. In Semantics modeling, the relation sequence of nodes in the heterogeneous network can be determined.

The complete procedures for our profile generation based on semantics modeling can be elaborated by the following algorithm. Given the heterogeneous social network, the algorithm first progressively attains the relational adjacency matrix (steps 1-4). Then all relations are collected to construct the profiles (steps 5-8). By visiting each horizontal line of the matrix and counting the occurrence times of each node's in the network (steps 7-9), the profile is produced.

Algorithm. Profile Generation

Input: $H=\langle V, E, L \rangle$: a heterogeneous network; k : the step size for relation sequences.

Output: $P(x)$: the Profile for each node x .

```

1: Derive the one-step relational adjacency matrix  $ReAM_1$ .
2:  $RAM = [ReAM_1]$ .
3: for  $step = 2$  to  $k$  do
4:  $ReAM_k = ReAM_{k-1} \cdot ReAM_1$ . // iteratively get  $k$ -step relational adjacency matrices
5:  $P = \text{new int}[n][|Re|]$ . // initialize the profiles
6: for  $x \in V$  do
7:  $P(x) = \text{count}$  and the times of each relations of  $x$  and store the counts into the corresponding cell of  $P$ .
8: end for
9: return:  $P$ 
  
```

V. SEMANTIC MODELING OF HETEROGENEOUS NETWORKS

This section discusses the semantic model for heterogeneous social networks. The fundamental assumption is the information about a node has already been encoded in the form of a heterogeneous social network, and the semantics can be captured and formulated using the surrounding relational structures. It aims at profiling each node, which automatically extracts relational features and measures the relatedness as feature values between the node and the corresponding features.

A heterogeneous network is composed of a topological part and relational part. Each node can be characterized using its neighborhood which consists of a set of directly or indirectly connected nodes and links.

Definition 5.1 (Heterogeneous Social Network). A heterogeneous network $H(V, E, L)$ is a directed labeled graph, where V is a finite set of nodes, L is a finite set of labels, and $E \in V \times L \times V$ is a finite set of edges [20]. Given a notation representing an edge, the source, label, and target map it onto its start vertex, label, and end vertex.

As already mentioned, the role of each node is encoded by its relational neighborhood. This motivates our idea of defining the relational adjacency matrix to capture the direct and indirect relationships between nodes. We start from some basic definitions of relational data structures.

Definition 5.2 (Relation Sequence). A sequence of relations is called a relation sequence (RS). A k -step relation sequence ($k > 0$) is defined as a

sequence of k labeled relations $\langle rx1, rx2, \dots, rxk \rangle$ where each $rxk \in L$.

Definition 5.3 (Relation Sequence Group). The group of relation sequences $\{RS1, RS2, \dots\}$ is called a *relation sequence group* (RSG). Note that RS_i can be of any length, and duplicate elements can exist in an RSG [20]. We can represent the duplicate elements in an RSG using a numerical number before each occurrence of distinct RS. For example, $\{3RS1, 1RS2, \dots\}$ means in this group there are three $RS1$ and one $RS2$.

Definition 5.4 (Relation Sequence Matrix). A relation sequence matrix RSM is defined as an $n \times n$ matrix and each element of the matrix is a relation sequence group. In our model the n stands for the number of nodes in the network.

Then we define the multiplication and summation operations between two relation sequences and two relation sequence groups.

Definition 5.5 (Multiplication on Two Relation Sequences). Given two relation sequences $\langle rx1, rx2, \dots, rx_i \rangle$ and $\langle ry1, ry2, \dots, ry_j \rangle$, their multiplication (denoted by $RS1 * RS2$) is defined as concatenating the second sequence after the first one as, $RS1 * RS2 = \langle rx1, rx2, \dots, rx_i, ry1, ry2, \dots, ry_j \rangle$. This operation is not symmetric.

Definition 5.6 (Summation of Two Relation Sequence Groups). Given two relation sequence groups $RSG_1 = \{RS_{11}, RS_{12}, \dots, RS_{1p}\}$, $RSG_2 = \{RS_{21}, RS_{22}, \dots, RS_{2q}\}$, their summation (denoted by $RSG_1 + RSG_2$) is defined as the group of all elements in every RSGs. That says, $RSG_1 + RSG_2 = \{RS_{11}, \dots, RS_{1p}, RS_{21}, \dots, RS_{2q}\}$.

Since each element in a RSM is a RSG, and we have defined the summation for RSG, the RSMs can be defined as similar to the two numerical matrices.

Definition 5.7 (Relational Adjacency Matrix). The relational adjacency matrix of a given heterogeneous social network H (denoted by $ReAM_1$) is a relation sequence matrix [20] that captures the direct adjacency relationship between any two nodes. That is, given a social network with n nodes, each element in adjacency matrix is the group of direct labeled relations connecting two corresponding nodes. There can be multiple direct connections between two nodes in the network, and a node can also connect to itself.

VI. EXPERIMENTAL STUDY

Experiments on heterogeneous network are generally a challenging issue for social network analysis. The amount of diverse relations makes it a complicated heterogeneous network for humans to analyze. Therefore several diverse experiments have been designed for this model and tasks using both artificial and natural datasets. Hopefully the experiments from different angles can provide a more general explanation about outputs as well as their value.

A. Data Collection

The heterogeneous social network is generated from entities and relations from UCI KDD Archive movie dataset [12]. In this network, there are about 24,000 nodes representing movies (9,097), directors (3,233), actors (10,917), and some other movie-related persons (500) such as producers and writers. There is also 126,926 relations between these nodes. Totally, there are 44 different relation types in the movie network, which can be divided into three groups: relations between people (e.g., spouse and mentor), relations between movies (e.g., remake), and relations between a person and a movie (e.g., director and actor).

B. Results on Heterogeneous Network

In this experiment, relation sequence of actors in movie network can be analyzed. Firstly, few sample actors are determined based on their role played in the heterogeneous social network. Then relation between the actor those who are involved in the movie network (Fig. 4) can be found and created a relation sequence (Fig. 3), relation sequence group (Fig. 5) and relation sequence matrix (Fig. 7) between actors.

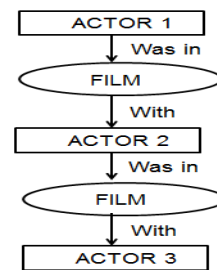


Fig. 3 The template for Relation Sequence (RS)

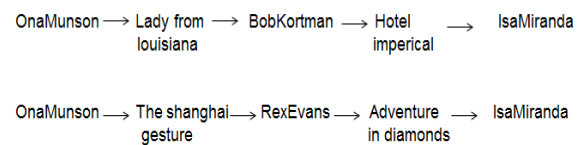


Fig. 4 The sample Relation Sequence (RS)

Relation sequence group is the combination of relation sequences and it can be grouped based on the alphabetical order which can be shown in the Fig.6.

A relation sequence matrix can be formed, where each element of the matrix is a relation sequence group which can be shown in the Fig. 7.

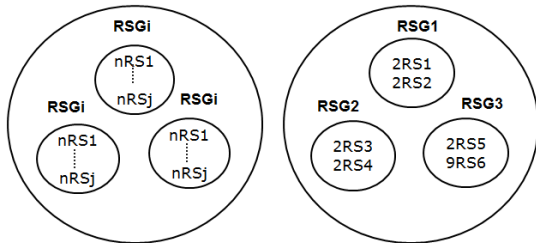


Fig. 5 The template for Relation Sequence Group (RSG)

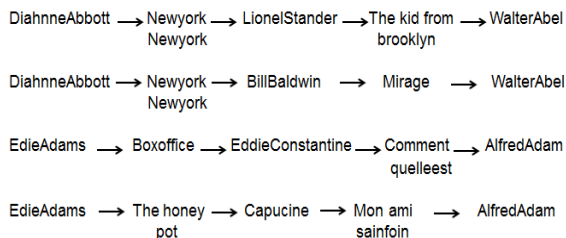


Fig. 6 The sample Relation Sequence Group (RSG)

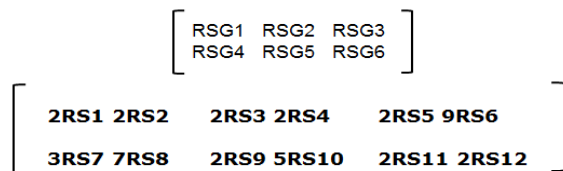


Fig. 7 The sample Relation Sequence Matrix (RSM)

Various operations like multiplication (Fig. 8a), summation (Fig. 8b), are performed on Relation Sequence (RS) and Relation Sequence Group (RSG), in order to generate a 3rd order matrix. It can be shown as follows:

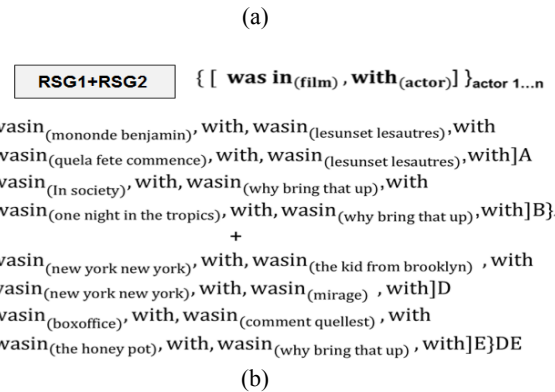
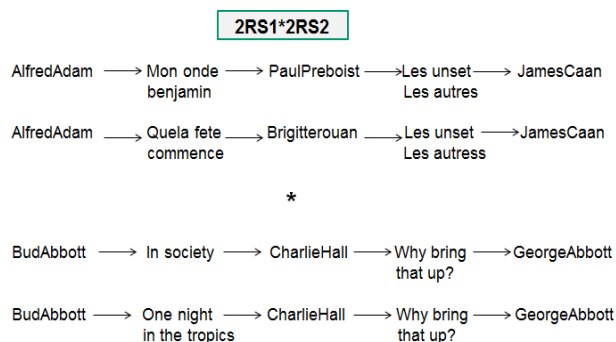


Fig. 8 a) Relation Sequence (RS) Multiplication b) Relation Sequence Group (RSG) Summation

Secondly, the index table (Fig. 9) can be framed based on the relation between the actors and generated the graph where square represents start vertex, circle represents middle vertex and triangle represents end vertex (Fig. 10) based on the value framed in the table which helps to find direct link and indirect link between the actors for the whole movie network and it can be shown in the (Fig. 11).

	A	B	C	D	E	F
1	joss ackland	clinton sundberg	jean adair	kathleen freeman	dan aykroyd	
2	joss ackland	0	1	1	0	0
3	clinton sundberg	0	0	1	0	0
4	jean adair	0	0	0	2	0
5	kathleen freeman	0	0	0	0	2
6	dan aykroyd	1	0	0	0	0

Fig. 9 Sample Index Table

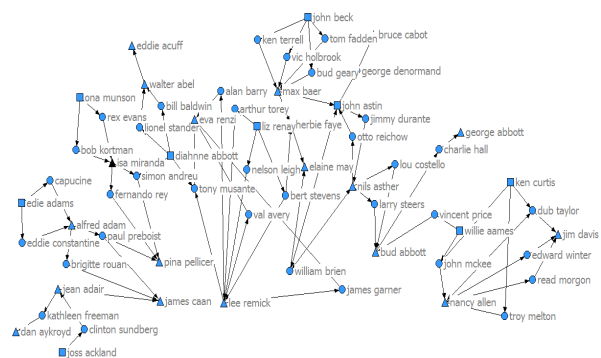


Fig. 10 The sample resulted graph of movie network

	joss ackland	clinton sundberg	dan aykroyd	jean adair	kathleen freeman
joss ackland	0	<1>	<-1>	<1>	0
clinton sundberg	<-1>	0	0	<1>	0
dan aykroyd	<1>	0	0	0	<-1>
jean adair	<-1>	<-1>	0	0	<1>
kathleen freeman	0	0	<1>	<-1>	0

(a)

	joss ackland	clinton sundberg	dan aykroyd	jean adair	kathleen freeman
joss ackland	0	<1,-1>	0	<1,1>	<1,1>
clinton sundberg	<-1,-1>	0	<-1,-1>	<-1,1>	<-1,1>
dan aykroyd	0	<1,1>	0	<1,1>	0
jean adair	<-1,-1>	<-1,1>	<1,1>	0	0
kathleen freeman	<1,1>	<-1,-1>	0	0	0

(b)

Fig. 11 The sample table for direct link and indirect link between the actors for the movie network a) Single Link b) Double Link

Finally, adjacency matrix can be constructed along with the direct and indirect link which can be shown in Fig. 12 in order to facilitate further human analysis. Since it retains critical information and significantly removes uninformative information. Hereby we have taken 10 relation sequences for sample which contains 65 actors. Same actors can have more than 2 sequences.

	joss ackland	clinton sundberg	dan aykroyd	jean adair	kathleen freeman
joss ackland	0	<the thief & the cobbler>	<blues brothers> ¹	<1941> ¹	0
clinton sundberg	<the thief & the cobbler> ¹	0	0	<living in a bigway>	0
dan aykroyd	<1941>	0	0	0	<blues brothers>
jean adair	<living in a bigway> ¹	<the thief & the cobbler> ¹	0	0	<the naked city>
kathleen freeman	0	0	<dragnet>	<the naked city> ¹	0

Fig. 12 The sample adjacency matrix of movie network

The above Fig. 12 shows the various direct links and indirect links between the actors and the relation between the actors will be a film. The inverse of film name represents an indirect relation between two actors. When there is no relation between certain some specific actor it can be represented as zero or null. The results show users utilize significantly less time to reach better-quality results.

VII. CONCLUSION

This paper presents Multi Relational Network (MRN) for knowledge discovery in heterogeneous social networks. Complex information about the graph topology and relational semantics is modeled through an, automatic, and robust mechanism. Here, some of the contribution in the following points:

1) Adjacency matrix can be created by using relation sequences which describes the direct and indirect links between nodes. This network can simultaneously capture the topological and relational semantics of a heterogeneous network. Besides, MRN

is succinct yet powerful, and it is modularized enough to facilitate fast implementation.

2) The experiments on a real-world movie dataset demonstrate that it can analyze the relation sequence of nodes in the heterogeneous network that are otherwise hard to find using existing methods.

3) The outcomes not only demonstrate the usability of network but also show the designed graph from diverse distilling criteria which can assist human analyst in making more accurate, efficient, and confident decisions.

REFERENCES

- [1] P. Appan, H. Sundaram, and B.L. Tseng. Summarization and Visualization of Communication Patterns in a Large-Scale Social Network. In *PAKDD 2006*, 371–379.
- [2] Y. Sun, Y. Yu, and J. Han. Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema. In *KDD 2009*, 797–806.
- [3] Z. Shen, K.L. Ma, and T. Eliassi-Rad. Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction. *IEEE Transactions on Visualization and Computer Graphics*, 12 (6), 1427–1439, 2006.
- [4] S. Wasserman, and K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, UK.
- [5] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Mining Hidden Community in Heterogeneous Social Networks. In *LinkKDD 2005*, 58–65.
- [6] J. Chen, O.R. Zaiane, and R. Goebel. 2009. Detecting Communities in Social Networks Using Max-Min Modularity. In *Proceedings of SIAM International Conference on Data Mining (SDM'09)*, 978–989.
- [7] D.J. Watts, and S.H. Strogatz. 1998. Collective Dynamics of Small-world Networks. *Nature* 393, 440–442.
- [8] M.E.J. Newman, and M. Girvan. Finding and Evaluating Community Structure in Networks. *Physics Review*, E 69, 2004.
- [9] W. Hwang, T. Kim, M. Ramanathan, and A. Zhang. 2008. Bridging Centrality: Graph Mining from Element Level to Group Level. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, 336–344.
- [10] M.E.J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Review*, 45:167–256.
- [11] L. Zou, L. Chen, H. Zhang, Y. Li, and Q. Lou. 2008. Summarization Graph Indexing: Beyond Frequent Structure-Based Approach. In *Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA'08)*, 141–155.
- [12] S. Hettich and S.D. Bay. The UCI KDD Archive. <http://kdd.ics.uci.edu>, University of California, Irvine, Dept. of Information and Computer Science.
- [13] Y. Tian, R.A. Hankins, and J.M. Patel. 2008. Efficient Aggregation for Graph Summarization. In *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*, 567–580.

- [14] D. Vincent, and B. Cecile. 2005. Transitive Reduction for Social Network Analysis and Visualization. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, 128–131.
- [15] X. Xu, N. Yuruk, Z. Feng, and T.A.J. Schweiger. 2007. SCAN: A Structural Clustering Algorithm for Networks. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*, 824–833.
- [16] B. Long, Z.M. Zhang, X. Wu, and P.S. Yu. 2006. Spectral Clustering for Multi-type Relational Data. In *Proceedings of International Conference on Machine Learning (ICML'06)*, 585–592.
- [17] N. Du, B. Wu, and B. Wang. 2007. Backbone Discovery in Social Networks. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, 100–103.
- [18] V. Satuluri, and S. Parthasarathy. 2009. Scalable Graph Clustering Using Stochastic Flows: Application to Community Detection. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, 737–745.
- [19] L. Singh, M. Beard, L. Getoor, and M.B. Blake. 2007. Visual Mining of Multi-Modal Social Networks at Different Abstraction Levels. In *Proceedings of International Conference on Information Visualization (IV'07)*, 672–679.
- [20] Cheng-Te Li and Shou-De Lin. <http://mslab.csie.ntu.edu.tw/~odd/heminetr12.pdf> Technical Report, 2012.