American Journal of Operations Research, 2011, 1, 33-38
doi:10.4236/ajor.2011.12005 Published Online June 2011 (http://www.SciRP.org/journal/ajor)

Extracting a Heterogeneous Social Network of Academic Researchers on the Web Based on Information Retrieved from Multiple Sources

Rasim M. Alguliev, Ramiz M. Aliguliyev, Fadai S. Ganjaliyev
Institute of Information Technology, *Azerbaijan National Academy of Sciences*, *Baku*, *Azerbaijan*
E-mail: *rasim@science.az*, *r.aliguliyev@gmail.com*, *fadaig@yahoo.com*
Received January 25, 2011; *revised March* 7, 2011; *accepted April* 6, 2011

Abstract

The majority of academic researchers present the results of their scientific activity on the Web. This trace can be used to derive useful information of their past, present activity and forecast the future intentions. Hence, social network of academic researchers can be of important value for scientific community. This information can be retrieved from various data source currently available on the Web. From each of them a separate network can be built. In this paper we present a method which can be used to combine multiple single-relational networks into a single network which will combine all relations, hence it will be multi-relational.

Keywords: Multi-Relational Networks, Academic Researchers' Network, Data Source Criteria

1. Introduction

The appearance of web in 1992 put a series of interesting challenges to the researches of social networks. First of the entire web influenced the traditional way of thinking about social networks. Since social network analysis usually was conducted on small group of nodes it appeared to be not so easy to apply the same methods on the Web. In most cases it was nearly impossible to analyze a network of millions of people taking into account that the proper analysis requires construction most of the network. Another hard task was to gather information about a large group of people. Moreover, there can and actually are people who are actors of the network but do not need any generated data which requires specific considerations in this case. These types of networks are much more difficult to study [1].

Social networks on the web have been extracted by retrieving relationships between entities all automatically derived from multilingual news [2]. Social networks also have been extracted from log files of online shared workspaces [3]. Another method is used to extract biographical information of historical persons from multiple unstructured sources on the Web [4]. Extraction of social networks has also been conducted via Internet and Networked Sensing [5]. A social network extraction system from the Web was designed named Referral Web [6]. Mika developed a system for extracting social networks from the Web, named Flink [7]. A system which extracts social network from user's inbox was developed as well [8]. Matsuo developed a system Polyphonet which is used for extracting and analyzing a social network of academic figures [9]. Tang *et al.*, demonstrate a method for extracting an academic social network [10]. Some authors address extraction of multi-relational networks [11]. How a social network can be extracted from email communication of users is also addressed [12]. If two users exchanged more than some N number of emails then ad edge is invented between them. In order to assign ranks to users two statistics are measured. First is based on the hypothesis that if two users communicate more, then they should exchange more emails. So the first statistics is simply the number of emails a user has sent or received. The second statistics is the average response time. After the social network has been built then all cliques are obtained by means of the method described in [13]. Users ranked based on the degree centrality, betweenness centrality and also on the number of cliques a user is contained within, the size of the clique and the weights of the clique edges.

One of the most popular searches on the Web is searching people related to some other person. For example, people need to find other people who have published more papers in a specific topic or to find the most

famous actor in a certain area. Authors of [14] propose an approach to resolve the problem of people search sharing similar interests by representing a person with the aid of the user's Web site content. Some studies have extracted networks from FOAF documents [15-17]. Entity ranking results' differentiation is demonstrated in [18]. Community mining, which is one of primary research areas in SNA, is also addressed in some works [19-21].

Although not little work done on detecting social networks on the Web most of them assume that there is only one relationship between entities in the graph. In fact, in most communities nodes are connected by more than one relation. Moreover, each relation has a particular role in a particular task. When supposing that in a given community only one relation exists between people much of valuable information can be lost. Besides, many of these algorithms are not suitable for large-scale networks and concentrate on small networks.

Academic researchers may have relations of different ki-nds: they may have co-authored one or more papers, may participate in the same conference, may be members of the same scientific centre or might have taken part in the same project and etc. Academic researchers' social network extraction is another interesting topic on the Web [9,10,19]. The major issue with the work done on extracting social network of academic researchers on the Web is that they do not embrace the entire academic network.

2. Constructing a Heterogenous Social Network of Academic Researchers from Multiple Sources

The richness of information on the Web gives a base to assume that information for one and the same set of entities obtained from multiple different sources will rarely provide similar picture. Say, we have built a social network of given named of entities based on information from one data source only. In such a case it may happen, for example, that those two actors have no relation in common. But having built the same network, in terms of actors, from a different source we may see that those two actors indeed posses a relation. This is turn means that the more data sources we use to build final network one the more exact results we shall receive as a result.

In our case we consider undirected networks only. We assume that the entities are given beforehand. Although the number of data sources and actors can be any in the method described in the paper, suppose we have three data sources s_1, s_2 and s_3 from which we can obtain network data for the given five named entities; also assume we are given five named entities a_1, a_2, a_3, a_4, and a_5.

Assume from three different data sources we have built three different social networks shown below.

Table 1 represents the edge weights for each of the networks.

Table 1. Edge weight values for networks from the three data sources.

Pair	Data source		
	s_1	s_2	s_3
w_{12}	0,2	0,0	0,2
w_{13}	0,1	0,5	0
w_{14}	0,3	0	0
w_{15}	0,3	0,8	0
w_{23}	0	0,7	0,3
w_{24}	0	0,4	0,1
w_{25}	0	0	0
w_{34}	0	0,0	0,8
w_{35}	0	0	0
w_{45}	0	0,9	0,9

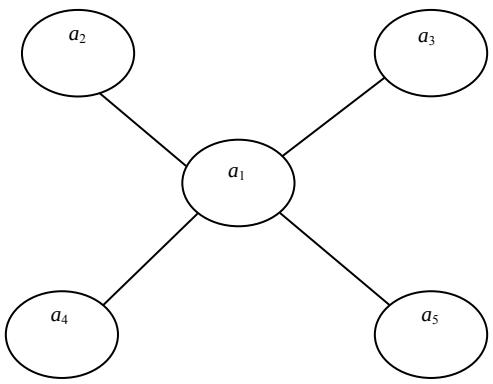

Figure 1. The social network of the actors from the 1st data source.

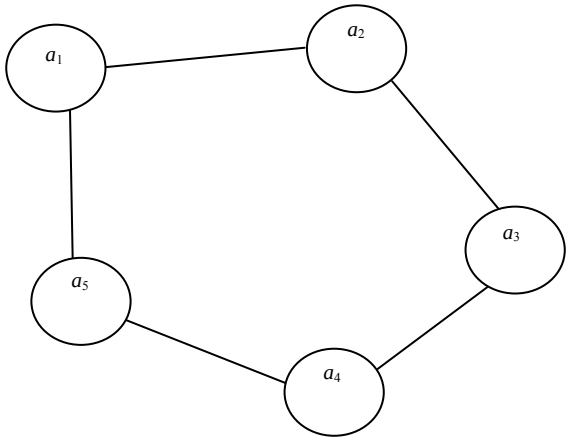

Figure 2. The social network of the actors from the 2nd data source.

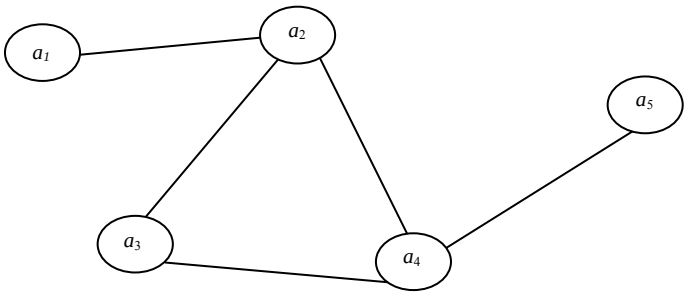

Figure 3. The social network of the actors from the 3rd data source.

After we have built three single-relational networks our objective is to combine them into a single network with the aid of the method described in [22]. Here a multi-relational network is built. The edge weight between any two entities in this network is the sum of the weights of the corresponding entities in each of the single-relational networks multiplied by some coefficient which shows the importance degree of the data source. In our case the single-relational networks will be those depicted in **Figures 1**, **2** and **3**. Sum of these coefficients equals unit. For this case in terms of formulas the basic rule will have the following form:

$$w^{ij} = z_1 w_1^{ij} + z_2 w_2^{ij} + z_3 w_3^{ij} \quad (1)$$

where w^{ij} is the weight of an edge between the actors i and j of the resultant heterogeneous network, z_k is the coefficient which shows the importance degree of the corresponding data source k, and w_1^{ij}, w_2^{ij}, and w_3^{ij} are edge weights between actors i and j from the three source correspondingly. Moreover, $z_1 + z_2 + z_3 = 1$.

3. Computing Data Source Weights

At this step we have built a single multi-relational network which combines multiple networks into it. We have also given a formula with the aid of which we can find weights of the resultant network edges. The problem now comes to finding the unknowing coefficients for the data sources mentioned in formula (1). These coefficients will tell us "how much out of each of the networks" we need to include in the resultant network.

In order to find these unknown values we shall use the method presented in [23]. In here a method of finding the best alternative out of possible multiple choices is presented. It's supposes that each choice is associated with multiple criterion. Namely, it's assumed that on given set of choices $S = \{s_1, s_2, ..., s_n\}$ the set of criterion $C = \{c_1, c_2, ..., c_m\}$ is given. Each criterion is represented as the following fuzzy set.

$$c_j = \left\{ \frac{w_1^{(j)}}{s_1}, \frac{w_2^{(j)}}{s_2}, \ldots, \frac{w_n^{(j)}}{s_n} \right\}, j = 1, 2, \cdots, m.$$

Elements $w_i^{(j)}$ are numbers in the interval [0,1] which can be considered as weights of choices with respect to criteria c_j sum of which by each criterion equals unit *i.e.*

$$w_1^{(j)} + w_2^{(j)} + \ldots + w_n^{(j)} = 1, j = 1, 2, \cdots, m.$$

According to Bellman-Zadeh principle the best alternative S_{opt} should be searched in the intersection of fuzzy-criterion sets $D = c_1 \cap c_2 \cap \cdots \cap c_m$. As author states the best choice S_{opt} is the one $S_{opt} \in D$ with maximal weight.

$$w(s_{opt}) = \max_{i=1,2,\ldots,n} \min \left\{ w_i^{(1)}, w_1^{(2)}, \cdots + w_i^{(m)} \right\}$$

Since in the theory of fuzzy sets one can use the substitution $\cap \to \min$, from which in turn that the set of good solutions can be represented as follows.

$$D = \left\{ \frac{\min\left\{w_1^{(1)}, \ldots, w_1^{(m)}\right\}}{S_1}, \frac{\min\left\{w_2^{(1)}, \ldots, w_2^{(m)}\right\}}{S_2}, \ldots, \frac{\min\left\{w_n^{(1)}, \ldots, w_n^{(m)}\right\}}{S_n} \right\}.$$

In order to find these weights the author uses the ranks of the choices which are larger if reliability of the choice is so. Hence, the following is true.

$$\frac{w_1}{r_1} = \frac{w_2}{r_2} = \ldots = \frac{w_l}{r_l} = \cdots \frac{w_n}{r_n}.$$

If S_l is the worst choice (by criterion c_j) with weight w_l and rank r_l, then the last correlation can be written as follows.

$$w_1 = r_1 \frac{w_l}{r_l}, w_2 = r_2 \frac{w_l}{r_l}, \ldots, w_n = r_n \frac{w_l}{r_l},$$

which in turn can be represented as following formula:

$$w_l = \frac{1}{\frac{r_1}{r_l} + \frac{r_2}{r_l} + \ldots + \frac{r_n}{r_l}} = \frac{1}{\sum_{i=1} \frac{r_i}{r_l}},$$

Since the sum of weights of choice equals unit by each criterion. Here, as also the author mentions, the ration r_i/r_j is taken from Saati 9 - scale, in which this ration equals one of the numbers in the interval [1,8] depending on how much alternative s_i is better than s_j Using

the last formula for w_l one can calculate можно weights of choices with the aid of ratio of the rank of the choice to the rank of the worst choice.

Therefore, in order to use this method we need to consider one more condition: the networks from these data sources should possess some criteria. In the method presented here we suppose that the networks have the criteria: average tie strength, density, average distance. **Table 2** shows values for each of these criteria for each of the networks.

We proceed according to the method mentioned.

1) Criteria: average tie strength (ATS). The worst alternative is s_1. The weight of the worst alternative is.

$$z_1 = \frac{1}{\frac{r_1}{r_1}+\frac{r_2}{r_1}+\frac{r_3}{r_1}} = \frac{1}{1+5+3} = \frac{1}{9}$$

Weights of the other two alternatives are.

$$z_2 = \frac{r_2}{r_1} z_1 = \frac{5}{9}, \quad z_3 = \frac{r_3}{r_1} z_1 = \frac{1}{3}.$$

2) Criteria: density (D). The worst alternative is s_2. The weight of the worst alternative is.

$$z_2 = \frac{1}{\frac{r_2}{r_2}+\frac{r_1}{r_2}+\frac{r_3}{r_2}} = \frac{1}{1+3+5} = \frac{1}{5}.$$

Weights of the other two alternatives are.

$$z_1 = \frac{r_1}{r_2} z_2 = \frac{1}{3}, \quad z_3 = \frac{r_3}{r_2} z_2 = \frac{5}{9}.$$

3) Criteria: average distance (ADS). The worst alternative is $s_2(s_1)$. The weight of the worst alternative is

$$z_2 = \frac{1}{\frac{r_2}{r_2}+\frac{r_1}{r_2}+\frac{r_3}{r_2}} = \frac{1}{1+1+3} = \frac{1}{5}.$$

Weights of the other two alternatives are

$$z_1 = \frac{r_1}{r_2} z_2 = \frac{1}{5}, \quad z_3 = \frac{r_3}{r_2} z_2 = \frac{3}{5}.$$

In the calculations above r_i is a rank of the alternative s_i and ratio r_i/r_j is taken from the very Table (Saaty1 - 9 scale) in the method and shows how much the alternative i is better than the alternative j with respect to a certain criteria.

Table 2. Criteria values for the three networks.

Network	Criteria value		
	Average tie strength	Density	Average distance
s_1	0,225	0,66	0,46
s_2	0,660	0,48	0,46
s_3	0,460	0,985	0,712

According to the mentioned method the received weights of the alternatives for different criteria allows us to represent the criteria as fuzzy sets as shown next.

$$ATS = \left\{ \frac{1/9}{s_1}, \frac{5/9}{s_2}, \frac{1/3}{s_3} \right\}$$

$$D = \left\{ \frac{1/3}{s_1}, \frac{1/9}{s_2}, \frac{5/9}{s_3} \right\}$$

$$ADS = \left\{ \frac{1/5}{s_1}, \frac{1/5}{s_2}, \frac{3/5}{s_3} \right\}$$

Choosing the maximum of the minimum of these values for each of the sources correspondingly will give us information about which alternative is the best in respect to these criteria *i.e.* the following set.

$$R = \left\{ \frac{1/9}{s_1}, \frac{1/9}{s_2}, \frac{1/3}{s_3} \right\}$$

This shows us which source is the best, which is worse and which is the worst. The better is the one with higher value of the numerator in the fraction.

As mentioned our objective is not to find out which source is the most reliable but how to integrate data from these sources into a single one from which we can build a single resultant network. We could consider the received values for alternatives as the coefficients we search but this would contradict to our requirement of $z_1 + z_2 + z_3 = 1$. So to have this condition we normalize values in the resultant set of alternative weights *R*. Hence we receive that $z_1 = 1/5$, $z_2 = 1/5$, and $z_3 = 3/5$ which in turn satisfies the condition that the sum of the unknown coefficients must be equal unity.

Further we use formula (1) to calculate edge weights of the resultant network. As a result we receive the network which follows with edge weight values shown in **Table 3**.

Computing values for each of the criteria for this network we receive the results from **Table 4**.

From the procedures accomplished above we can mention some interesting facts. Actors a_2 and a_4 do not have any direct relations in common in the 1st and 2nd networks but does have a relation in the 3rd network. So they do in the resultant network. Similarly, a_4 and a_5 do not have a direct relationship in the 1st network but have relations in the 2nd and 3rd networks. Our method showed that they have a relation in the resultant network as well. This gives us a base to assume that in this sense our method does not undergo information loss Another important point is that we can clearly see that $z_3 = 1/3$, *i.e.* the 3rd network should be the best in the sense that

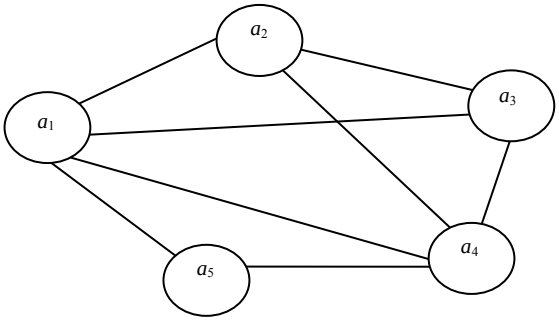

Figure 4. The resultant multi-relational network.

Table 3. Edge weights of the resultant multi-relational network.

Pair	Value
w_{12}	0,16
w_{13}	0,12
w_{14}	0,06
w_{15}	0,22
w_{23}	0,32
w_{24}	0,14
w_{25}	0,00
w_{34}	1,2
w_{35}	0,00
w_{45}	0,36

Table 4. Criteria value for the resultant multi-relational network.

Average tie strength	Density	Average distance
0,32	0,72	0,198

it should combine all optimal values of the criteria. Indeed, for the 2nd criteria the 3rd network is best, for the 1st criteria the network is neither worst nor the best, and finally, for the 3rd criteria the network is the worst which does not correspond to the received value of the coefficient.

4. Conclusions

Different kinds of social networks of academic researchers can be retrieved from information available on the Web. Most of currently existing methods deals with only specific network types of researchers. Ignoring any of the networks derived in this way may result in valuable information loss. We presented a method which can be used to create a single heterogeneous network out of multiple homogeneous networks of academic researchers. In the method presented none of the networks is ignored completely. In the future, we might be able to show some applications of the method described.

5. References

[1] J. Golbeck, "Web-based Social Networks: A Survey and Future Directions," Technical Report, Citeseer, 2005.

[2] B. Poliqean, H. Tanev and M. Atkinson, "Extracting and Learning Social Networks out of Multilingual News," *Proceedings of the Social Networks and Application Tools Workshop*, Skalica, September 2008, pp. 19-21.

[3] P. Nasirifard, V. Peristeras, C, Hayes and S. Decker, "Extracting and Utilizing Social Networks from Log Files of Shared Workspaces," *Proceedings of the* 10*th IFIP Working Conference on Virtual Enterprises*, Thessaloniki, October 2009, pp. 7-9.

[4] G. Geleijnse and J. Korst, "Creating a Dead Poets Society: Extracting a Social Network of Historical Persons from the Web," *Proceedings of the* 6*th International and the* 2*nd Asian Conference on Asian Semantic Web Conference*, Busan, Vol. 4825, October 2007, pp. 155-168.

[5] T. Nishimura and Y. Matsuo, "A Method of Social Network Extraction via Internet and Networked Sensing," *Proceedings of the* 3*rd International Conference on Networked Sensing Systems*, Chicago, May-June 2006.

[6] H. Kautz, B. Selman and M. Shah, "The Hidden Web," *AI Magazine*, Vol. 18, 1997, pp. 27-35.

[7] P. Mika, "Flink: Semantic Web Technology for Extraction and Analysis of Social Networks," *Journal of Web Semantics*, Vol. 3, No. 2, October 2005, pp. 211-223. doi:10.1016/j.websem.2005.05.006

[8] A. Culotta, R. Bekkerman and A. McCallum, "Extracting Social Networks and Contact Information from Email and the Web," *Proceedings of the* 1*st Conference on Email and Anti-Spam*, California, July 2004.

[9] Y. Matsuo, J. Mori and M. Hamasaki, "POLYPHONET: An Advanced Social Network Extraction System," *Proceedings of the* 15*th International Conference on World Wide Web*, Edinburgh, May 2006, pp. 397-406.

[10] J. Tang, D. Zang and L. Yao, "Social Network Extraction of Academic Researchers," *Proceedings of the* 7*th IEEE International Conference on Data Mining*, Omaha, 28-31 October 2007, pp. 292-301.

[11] V. Stroele, J. Oliveira, G. Zimbrao and J. M. Souza, "Mining and Analyzing Multi-relational Social Networks," *Proceedings of the* 12*th International Conference on Computational Science and Engineering*, Vancouver, Vol. 4, 29-31 August 2009, pp. 711-716. doi:10.1109/CSE.2009.69

[12] R. Rowe, G. Creamer, S. Hershkop and S. J. Stoflo, "Automated Social Hierarchy Detection Through Email and Network Analysis," *Proceedings of the* 13*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, 12-15 August 2007, pp. 109-117.

[13] C. Bron and J. Kerbosch, "Finding All Cliques of an Undirected Graph," *Communications of the ACM*, Vol. 16, No. 9, Septermber 1973, pp. 575-577.

[14] Q. Li and Y. B. Wu, "People Search: Searching People Sharing Similar Interests from the Web," *Journal of the*

American Society for Information Science and Technology, Vol. 59, No. 1, January 2008, pp. 111-125. doi:10.1002/asi.20736

[15] T. Finin, L. Ding and L. Zou, "Social Networking on the Semantic Web," *The Learning Organization*, Vol. 12, No. 5, December 2005, pp. 418-435. doi:10.1108/09696470510611384

[16] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A. Sheth, I. B. Arpinar, A. Joshi and T. Finin, "Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection," *Proceedings of the* 15*th International World Wide Web Conference*, Edinburgh, 23-26 May 2006, pp. 407-416. doi:10.1145/1135777.1135838

[17] J. Goldbeck and M. Rothstein, "Linking Social Networks on the Web with FOAF," *Proceedings of the* 23*rd National Conference on Artificial Intelligence*, Chicago, Vol. 2, 13-17 July 2008, pp. 1138-1143.

[18] R. Alguliev, R. Aliguliyev and F. Ganjaliyev, "Investigation the Role of Similarity Measure and Ranking Algorithm in Mining Social Network," *Journal of Information Science*, Vol. 37, No. 3, March 2011, pp. 229-234. doi:10.1177/0165551511400946

[19] N. Du, B. Wu, X. Pei, B. Wang and L. Xu, "Community Detection in Large-Scale Social Networks," *Proceedings of the* 8*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, August 2007, pp. 16-25.

[20] J. Baumes, M. Goldberg, M. Magdon and W. Wallace, "Discovering Hidden Groups in Communication Networks," *Proceedings of the* 2*nd NSF/NIJ Symposium on Intelligence and Security Informatics*, Tucson, July 2004. doi:10.1007/978-3-540-25952-7_28

[21] D. Cai, Z. Shao, X. He, X. Yan and J. Han, "Mining Hidden Community in Heterogeneous Social Networks," *Proceeings of the* 3*rd International Workshop on Link Discovery*, Chicago, 21-24 August 2005, pp. 58-65.

[22] F. Ganjaliyev, "Building a Heterogeneous Social Network of Academic Researchers," *Proceedings of the* 3*rd International Conference of Problems of Cybernetics and Informatics*, Baku, 6-8 Septermber 2010, pp. 179-182.

[23] P. A. Rotshtein, "Fuzzy Multicriteria Choice among Alternatives: Worst-Case Approach," *Journal of Computer and Systems Sciences International*, Vol. 25, No. 9, September 2010, pp. 948-957.

