Comparison of Community Identification Techniques for Two-Mode Affiliation Networks Using Wikipedia Data

Fawad Nazir [1,2,3], Hideaki Takeda[2], Aruna Seneviratne[1,3]

[1]Networked Systems
National ICT Australia (NICTA), Sydney, Australia
[2] National Institute of Informatics (NII), Tokyo, Japan
[3]Electrical Engineering & Telecommunications
University of New South Wales (UNSW), Sydney, Australia
Email: {fawad_nazir,takeda}@nii.ac.jp, aruna.seneviratne@nicta.com.au

Abstract. One of the most important questions in social networks is the identification of cohesive subgroups (a.k.a. community identification). These cohesive subgroups are loosely defined as collection of individuals who interact frequently. Once the communities are identified they often reveal interesting properties of the social network members, such as common hobbies, interests, social bindings, occupations etc. Several types of algorithms exist for analysis and identification of cohesive subgroups in one-mode networks that focus on pair-wise ties. However, less attention has been given to identification of cohesive subgroups in two-mode affiliation networks. Two mode affiliation networks focus on ties existing among actors through joint affiliations. Therefore, in this paper we evaluate two cohesive subgroups identification methods i.e. edge betweenness and hierarchical clustering, for two-mode affiliation network using the Wikipedia data. We conclude from our results that edge betweenness technique, when applied to two-mode affiliation network, is a better techniques in terms of the modularity value that means it can generate more strong social communities in terms of social ties. On the other hand this technique is less time efficient as compared to hierarchical clustering.

Keywords: Social Networks, Community Identification, Cohesive subgroups, Edge Betweenness, Hierarchical Clustering, Dendrograms.

1 Introduction

Social networking is an emerging field of research. Social network is a structured representation of the social actors and there interconnections a.k.a. ties [5]. Social networks form social groups or social communities that share interests. These communities on the web are steadily emerging and the demand for forming an on demand social network is immense. Community members profit from being linked to other people sharing common interests, though having widely dispersed residences. Without these online social community portals on the web, people would not be able to find other people sharing the same interest and being available for discussion and collaborations. For example, if a person is searching for specific information, he can look at the interests of people in his social network and get quite relevant references. Therefore in order to fully benefit from social networks, people should be able to identify the community they belong to. In this paper we deal with a topological property of networks, the cohesive subgroups/communities [5][6][13]. The concept of community is common, and it is linked to the classification of objects in categories or subgroups. It is critical to construct efficient procedures and algorithms for the identification of community structure in a generic network. Several types of algorithms exist for analysis and identification of cohesive subgroups in one-mode networks that focus on pair-wise ties [1][2]. However, less attention has been given to identification of cohesive subgroups in two-mode affiliation networks that focus on ties existing among actors through joint affiliations. In this paper we create social network based on the method proposed in [14]. In [14] an edge within a network can represent social interactions, common affiliations, organizational structure, physical proximity etc. Furthermore, we analyze the usefulness of community formation methods in order to identify cohesive subgroups using edge betweenness and hierarchical clustering methods. We evaluate these methods qualitatively using the definition of community i.e. a community is defined as a subset of nodes within the graph such that connections within a community are denser than the connection with rest of the network. In this paper first of all we give a brief introduction to the Wikipedia data structure. Section 3 deals with the methods of identification of cohesive subgroups. This section is followed by section 4 in which we give a comparison between cohesive subgroups identification techniques. In the end we discuss our results and give some concluding remarks.

2 Wikipedia Data

In this section we will discuss about the structure of Wikipedia data. We used Wikipedia data for our analysis. The following is the information about the data that we have used for our analysis:

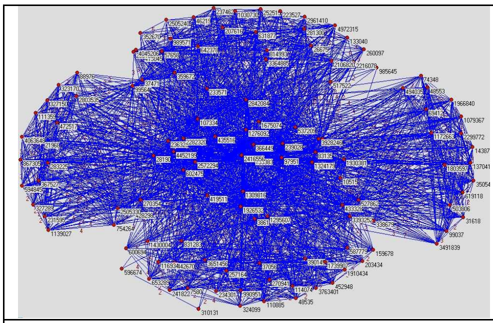
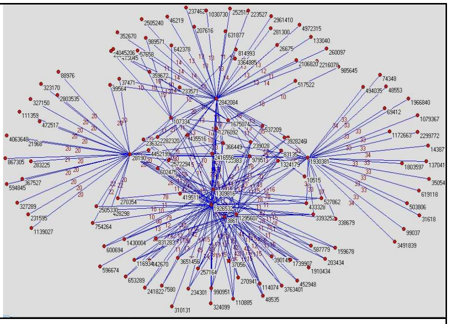
1. Number of articles: 10,218,632
2. Number of users: 65,678
3. Number of revised articles analyzed: 234,357
4. Total number of article revisions studied: 31,135,556
5. Wikipedia dump date: September 08, 2007.

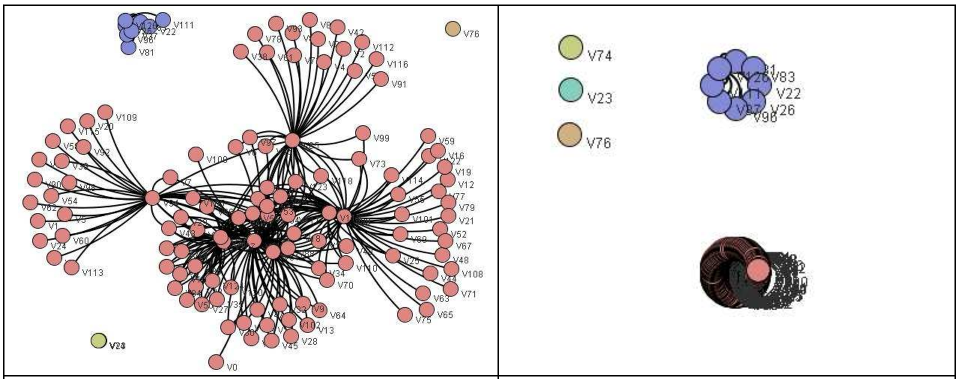
There are 41 Wikipedia tables[12][16]. In this paper we will use only four tables to extract most of our interesting conclusion. The tables that we have used are page table, user table, revision table and categorylink table. Page table is considered to be the core of Wikipedia. It contains the entry of each page in Wikipedia. This table does not contain the page text, it only contain information about the page identity, reference for it in text table (this table contains the page text) and revision table (this table keeps tract of the page revision made by users). User table stores the information about the Wikipedians. Wikipedians are authors/editors of the Wikipedia articles. This table contains information about user identity and user privileges. Next is the revision table, this is the most important table for our social network construction methods. This table holds information about all the edits made to the article by Wikipedians. It keeps track of the article to which edit was made, who made this edit and what time it was done. We use this table to find communities of users according to their article edit patterns. This table forms a baseline for our analysis. The last table which we used in our analysis is categorylink table. This table stores the categories to which a page is associated. This table is used to add the third dimension to our data i.e. Category. Every page is associated with some categories and from this table we can extract the information about what categories a page is associated with.

3 Identification of Cohesive Subgroups

In this section we will discuss different methods of cohesive subgroup identification when applied to Wikipedia data. First of all we explain the methods we used to form social network from Wikipedia. In the next two sections we will apply two methods of cohesive subgroup identification, i.e. Edge Betweenness and Hirarchical Clustering, to this social network and discuss our results. The method that we will use for cohesive subgroup identification is called LS Set. In this method we compare ties within the subgroup to ties outside the subgroup. More formally we can say that if G_s is a sub social group of a social network G, then authors in the sub social group G_s are given by $Author(G_s) = N_s$. Suppose $L \subset N_s$, therefore L is a strong cohesive subgroup of N_s if L has more social ties (Lines) within N_s then outside of N_s. This can be formally written as following and demonstrated in (Figure 5a):

$$\{L \xleftrightarrow{Lines} N_s - L\} > \{L \xleftrightarrow{Lines} N - N_s\}$$

	
(a) Social Network of Complete Data with Link Values.	(b) Social Network of subgraph showing edges greater than 10 with Edge values.



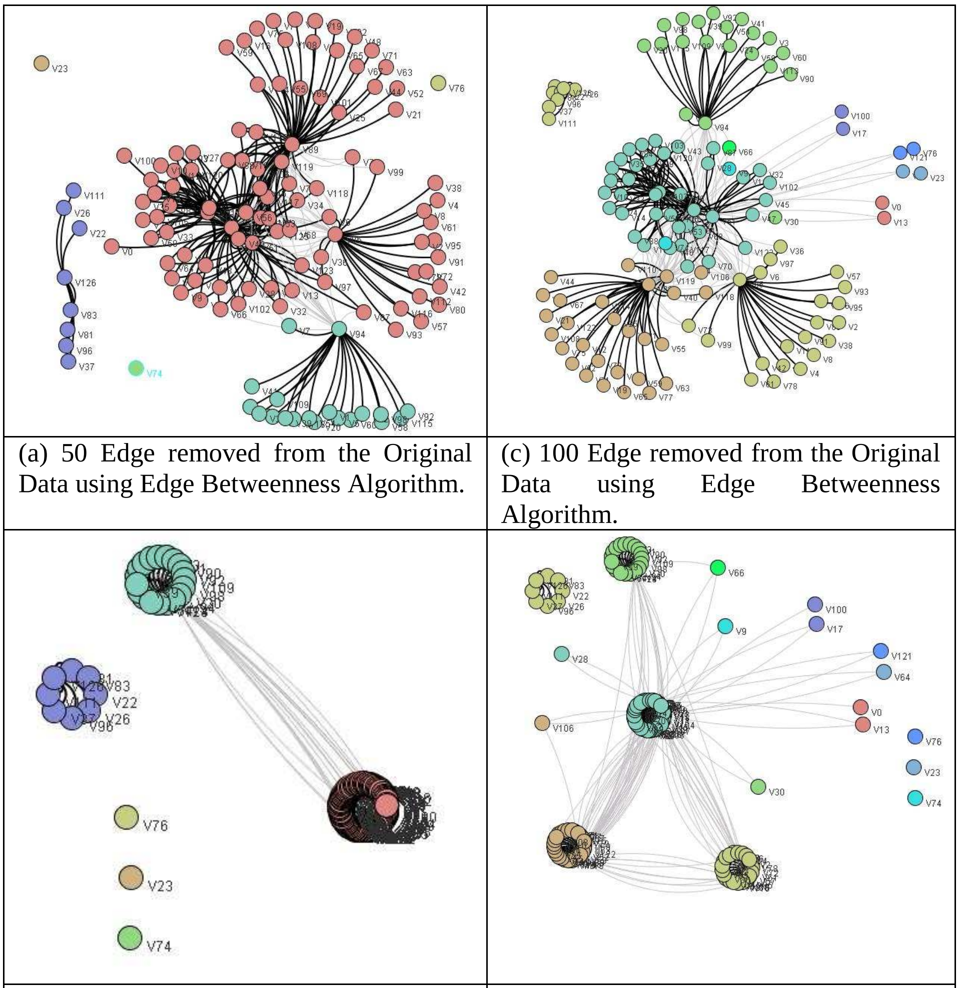
(c) Original Data plotted in JUNG with 0 Edge removed in edge betweenness.

(d) Clustered View of the Original Data plotted in JUNG with 0 Edge removed in edge betweenness.

Figure 1: Social Network Formed from Wikipedia Sports Data.

3.1 Social Network Extraction from Wikipedia

The dataset that we used for our social network extraction is: Sports -> Sports by country -> Cricket by country -> Cricket in Australia -> Australiann first class cricket teams. There are nine categories in this dataset List_of_Tasmanian_representative_cricketers, New_South_Wales_Blues, Prime_Minister's_XI, Queensland_Bulls, Southern_Redbacks, Tasmanian_Tigers, Victorian_Bushrangers, Western_Fury and Western_Warriors. We created this social network based on our tripartite model presented in the paper [14]. In addition to that we also used Principle Component Analysis (PCA) technique to find out similarities and associations between the users. In order to extract the related actors to a category we run an SQL query on our Wikipedia database. In this query we select rev_user, rev_user_text, count(rev_page) from the revision table where rev_page= <<page_id>> and rev_user <> 0 and then we group the data by rev_user. After this extraction we plot the data. These self explanatory graphs are plotted using Pajek[9][11] (Figure 1(a)-1(b)) and JUNG[10] (Figure 1(c)-1(d)).



(a) 50 Edge removed from the Original Data using Edge Betweenness Algorithm.

(c) 100 Edge removed from the Original Data using Edge Betweenness Algorithm.

(b) 50 Edge removed from the Original Data using Edge Betweenness Algorithm. (Clustered View)

(d) 100 Edge removed from the Original Data using Edge Betweenness Algorithm. (Clustered View)

Figure 2: Cohesive Subgroup Formation Using Edge Betweenness Algorithm on Wikipedia Data.

3.2 Cohesive Subgroup Identification using Edge Betweenness

In a dense social network [4] if any two edge nodes have to reach each other they will have to pass through one or more nodes in between the network. Then the edge in the middle that has the most influence of the reach-ability property of the nodes on the edge has the highest betweenness. In this method we remove such edges to form cohesive subgroups from a social network. Consider for example if actor A1 has to reach actor A2, so the shortest path between the two actors is:

$A_1 \leftrightarrow A_2 = \{A_1, A_5, A_6, A_7, A_2\}$. In an another example, actor A10 has to reach actor A11, then the shortest path is: $A_{10} \leftrightarrow A_{11} = \{A_{10}, A_6, A_7, A_{11}\}$. In this example the highest edge betweenness is of the edge $A_6 \leftrightarrow A_7$. Therefore if we remove one edge with the highest betweeness we will get two cohesive subgroups considering A1, A10, A5 and A11, A2 are well connected. Then the two subgroups will be: $G_1 = \{A_1, A_5, A_{6,} A_{11}\}$ and $G_2 = \{A_2, A_7, A_{11}\}$. Now we apply this on our extracted social network from Wikipedia with edges greater than 10 shown in Figure 1(b,c,d). The results of this edge betweenness are shown in the Figure 2.

3.3 Cohesive Subgroup Identification using Hierarchical Clustering

Hierarchical clustering is a data analysis technique that is ideally suited for partitioning actors in cohesive subgroups [17]. Hierarchical clustering groups entities into subsets (communities) that are structurally equivalent. Two vertices in a graph are said to be structurally equivalent if they have identical ties to and from all other actors in the network [3][13]. In this paper we have extended the definition of structural equivalence in order to represent our tripartite social network model [14]. In our paper we denote the structural equivalence of two actors A_1 and A_2 using $A_1 \xleftrightarrow{E} A_2$. Two actors are structurally equivalent if they have both edited the same articles or that have edited the articles belonging to the same categories. More formally we can define as:

$$[A_1 \rightarrow \{I_1, I_2, I_3\}] \xleftrightarrow{E} [A_2 \rightarrow \{I_1, I_2, I_3\} or A_2 \rightarrow C_1]$$
$$Where \quad I_1, I_2, I_3 \subseteq C_1$$

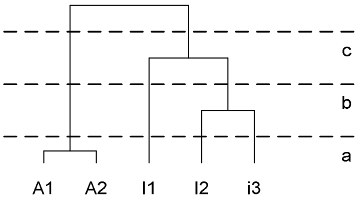


Figure 3: Example Dendrogram of a social Network

Now we will apply our approach to calculate structural equivalence to social network in Figure 1. We have taken a subset of two actors and four instances, therefore total possible connections will be 2*4=8. If we have only have one instance i.e.

I_4 included in both of the actors. Therefore the equivalence value for these actors will be: 0.125. Now we check the structural equivalence using the categories. We have nine categories and two actors so that possible relations are: 9*2 = 18. Therefore the equivalence value for these actors will be: 0.278. Once we have calculated the equivalence value for all the possible pair of actors then we create a density metrics to draw a dendrogram. Then using the dendrogram we create communities as show in Figure 4.

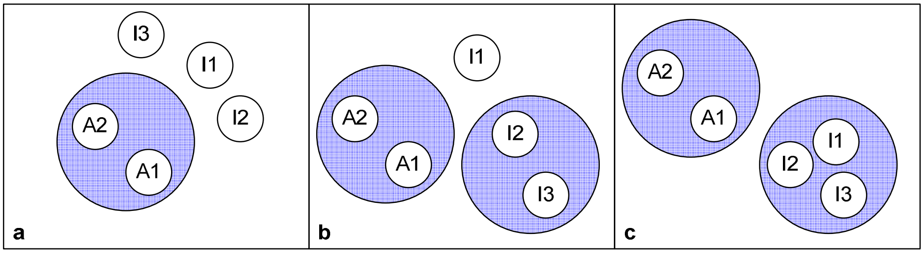
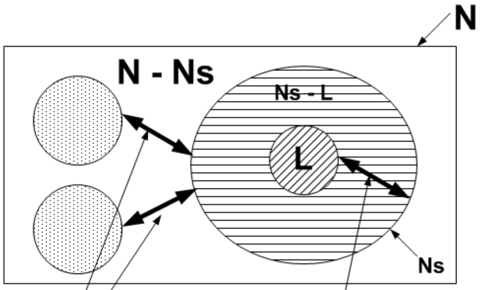


Figure 4: Community Formation from Dendrogram (a) First Level of Dendrogram, (b) Second Level of Dendrogram, (c) Third Level of Dendrogram

4 Comparison of Cohesive Subgroup Identification Techniques

We compare the two cohesive subgroup identification techniques using the concept of modularity [8]. The basic idea is to compare ties within the subgroups to ties outside the subgroup by focusing on the greater frequency of ties among subgroups members compared to the ties from subgroups members to outsiders (Figure 5).



$$\{L \xleftrightarrow{Lines} N_s - L\} > \{L \xleftrightarrow{Lines} N - N_s\}$$

(a) Illustration of Modularity

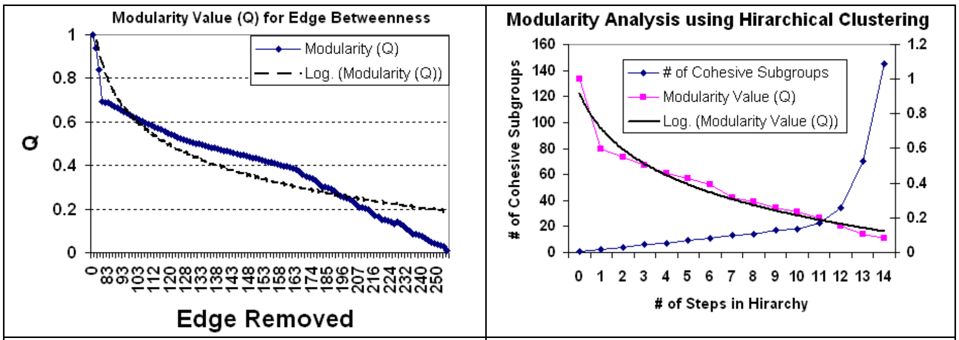
$Author(G_s) = N_s$ $L \subset N_s$ $\{L \xleftrightarrow{Lines} N_s - L\} > \{L \xleftrightarrow{Lines} N - N_s\}$	$\omega = \dfrac{\dfrac{\sum_{i \in N_s} \sum_{j \in N_s} x_{ij}}{g_s(g_s - 1)}}{\dfrac{\sum_{i \in N_s} \sum_{j \notin N_s} x_{ij}}{g_s(g - g_s)}}$
(b) Mathematical Explanation of Fig 5-a	(c) Definition of Modularity[13]

Figure 5: Explanation of the Modularity Concept

In edge betweenness [15] we find out the highest value of Q [8] after removing an edge with the highest betweenness. In the case of hierarchical clustering we find out the highest value of Q at each step of clustering as shown in Figure -3 i.e. Step a,b and c. The formal definition of Q is given in Figure 5(c). Here we can observe the following:

$$\omega = \begin{cases} 1 & No\_Diff \\ >1 & Strong\_Within\_Group \\ <1 & Strong\_Outside\_Group \end{cases}$$

In Figure 5(c) we denote modularity using ω. We can see that if the value of ω is 1, this means that there was no difference after or before applying the clustering algorithm. If the value is greater than 1 this means that the cohesive subgroups are strong and if it’s less than 1 then the cohesive subgroups are weak. Strong cohesive subgroups mean that ties of vertices within a group are stronger than with the vertices outside the group and weak subgroup is the opposite.



(a) Modularity Values using Edge betweenness

(b) Modularity Values using Hierarchical Clustering

Figure 6: Comparison between Edge betweenness & Hierarchical Clustering

In Figure 6(a) and 10(b) we can clearly see that the value of Q is much higher in the case of edge betweenness as compared to hierarchical clustering. On the other hand the time taken in determining cohesive subgroups by edge betweenness is much higher than the time taken by hierarchical clustering. In hierarchical clustering the major time is utilized in finding the structural equivalence of all pairs of actors. Another interesting thing to note in both of these techniques is that number of cohesive subgroups rise nearly exponentially if we increase the edge removal. This creates a relationship between number of cohesive groups and modularity value. They both are directly proportional to each other, as they both rise and fall together. This is apparent in Figure 6(b).

5 Conclusion

Identification of cohesive subgroups is an important research area in social network analysis. Currently in the literature several types of algorithms exist for analysis and identification of cohesive subgroups in one-mode networks. In this paper we have studied identification of cohesive subgroups in two-mode affiliation networks that was not the major focus of previous research. Therefore in this paper we analyzed two techniques for the formation of cohesive subgroups i.e. edge betweenness and hierarchical clustering. We can conclude from our results that edge betweenness is a better techniques in terms of the modularity value that means it can generate more strong social communities. On the other hand this technique is not efficient with time, time efficiency of hierarchical clustering technique is better. We also observed a general conclusion that number of cohesive subgroups rise nearly exponentially if we increase the edge removal or come down from the root level to leaf nodes in hierarchical clustering. This creates a relationship between number of cohesive subgroups and modularity value. They both are directly proportional to each other, as they both rise and fall together. As our future work we will expand our study to different types of social network and see if our findings are applicable to a wide range of social networks as well.

References

1. Michelle Girvan, M. E. J. Newman: Community structure in social and biological networks, PNAS June 11, 2002 vol. 99 no. 12 page 7821–7826 (2002)
2. Aaron Clauset, M. E. J. Newman, Cristopher Moore1: Finding community structure in very large networks, oxford journal. (August 2004)
3. M. E. J. Newman: The mathematics of networks, Technical Report. www-personal.umich.edu/~mejn/papers/palgrave.pdf (2004)
4. M. E. J. Newman: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Physical Review E, Volume 64, 016132, (2001)
5. Scott, J: (2000) Social Network Analysis: A Handbook (Sage, London), 2nd Ed.
6. F. Wu and B. A. Huberman: Finding communities in linear time: A physics approach. Eur. Phys. J. B 38331{338 (2004).
7. M. E. J. Newman, Analysis of weighted networks. Preprint cond-mat/0407503 (2004).
8. M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks. Phys. Rev. E 69,026113 (2004).
9. Pajek, Program for Large Network Analysis: http://vlado.fmf.uni-lj.si/pub/networks/pajek/
10. JUNG, Java Universal Network/Graph Framework: http://jung.sourceforge.net/
11. Wouter de Nooy, Andrej Mrvar, Vladimir Batageli: Exploratory Social Network Analysis with Pajek. Cambridge University Press. (2004)

12. Wikipedia table Schema: http://www.mediawiki.org/wiki/Category:MediaWiki_database_tables
13. Stanley Wasserman, Katherine Faust: Social Network Analysis, Methods and Application, Cambridge University Press (reprint 2007)
14. Fawad Nazir, Hideaki Takeda: Extraction and Analysis of Tripartite Relationships from Wikipedia, 2008 IEEE International Symposium on Technology and Society ,Frederiction, New Brunswick, Canada. (ISTAS 2008),
15. Ulrik Brandes: A Faster Algorithm for Betweenness Centrality, Journal of Mathematical Sociology 25(2):163-177, (2001).
16. Wikipedia database download: http://en.wikipedia.org/wiki/Wikipedia:Database_download
17. Christian Borgs Jennifer Chayes _ Mohammad Mahdian y Amin Saberi, Exploring the Community Structure of Newsgroups, KDD'04, August 22–25, 2004.