Representation of K-Cluster Constraint as K-Sat in Social Networking

Rajkumar Jain
Indian Institute of Technology, Indore,
Survey No.113/2-B, Opp. to Veterinary College
AB- Road, Village Harnia Khedi, Mhow (M.P.)

Narendra S. Chaudhari
Indian Institute of Technology, Indore,
Survey No.113/2-B,Opp. to Veterinary College
AB- Road, Village Harnia Khedi, Mhow (M.P.)

ABSTRACT
The information revolution has given birth to Social Networks, which allows structured flow of data, information, and knowledge. Social networks are nodes of individuals, groups, organizations, and related systems that are linked by one or more types of interdependencies. The defining feature of social network analysis is its focus on the structure of relationships. Social network analysis is a set of theories, tools, and processes for better understanding the relationships and structure of a network. Identification of Clusters in Social network is an active area research in artificial intelligence and pattern matching. Adding constraints to clustering improves the performance of a variety of algorithms. Cluster analysis is concerned with the problem of partitioning a given set of entities into homogeneous and well-separated subsets called clusters. Cluster Analysis aims at finding subsets, called clusters, which are homogeneous and/or well separated. Minimum sum of diameters clustering for two clusters can be solved by reduction constraints into the 2-Conjunctive Normal Form statement.

Hansen [4] uses Boolean approach to represent constraint in 2-cluster analysis, Identified constraints are represented in the form of 2-SAT statement. Constraint representation of 3-cluster or more then 3-cluster is not possible using Boolean approach. In our earlier paper [11], an approach was proposed “Belonging approach” using that constraints of 2-Cluster are represented in 2-SAT form. In this paper “Belonging approach” is extended for the representation of constraints in K-cluster. This approach can be used to generate constraints for 3-cluster for any value positive integer value of k. Constraints is generated in the form of K-SAT statement. This paper presents a formulation that find out the constraints in k- cluster based on concept of bonding and bridging in social network.

General Terms
Social Networking Analysis, K-Cluster Analysis, Partitioning, Artificial Intelligence, Constraint Clustering, Pattern Recognition.

Keywords
Must Link Constraint, Can Not Link Constraint, Belonging approach, Bonding, Bridging.

1. INTRODUCTION
Social networks are social communities of the web, connected via electronic mail, websites and web logs, and networking applications such as Twitter, FaceBook, or LinkedIn. In Social network analysis relationships are important. It maps and measures formal and informal relationships to understand what facilitate or impede the knowledge flows that bind interacting units. Social network analysis is somewhat similar to an "organizational x-ray". Social network analysis is a method with increasing application in the social sciences and has been applied in areas as diverse as psychology, health, business organization, and electronic communications. More recently, interest has grown in analysis of leadership networks to sustain and strengthen their relationships within and across groups, organizations, and related systems.

In social networks, "nodes" of the network are people and the "links" are the relationships between people [15]. Nodes are also used to represent events, ideas, objects, or other things. Social network analysis practitioners collect network data, analyses the data and often produce maps or pictures that display the patterns of connections between the nodes of the network. These maps reveal characteristics of the network that help guide participants as they evaluate their network and plan ways to improve their collective ability to identify and achieve shared goals. Constraints provide guidance about the desired partition and make it possible for clustering algorithms to increase their performance.

Wagstaff and Cardie [6] first introduced constraints in the area of data mining research. The introduction of constraints addresses an important problem neatly: the clustering algorithm’s objective function need not capture all the domain expert’s requirements, but user specified constraints can help guide the algorithm to a desirable set partition. Wagstaff and Cardie introduced two instance-level constraints that were termed must-link and cannot-link. In must-link (ML) constraint [7, 9, 10] two instances must be in the same cluster and in cannot-link (CL) constraint [7, 9 10] two instances must be in different clusters. Bonding and bridging [14] are two different kinds of connectivity in social network. Bonding denotes connections in a tightly bind group. Bridging denotes connections to other cluster (see figure 1). In the Social network analysis literature, bonding and bridging are often called "closure" and "brokerage". Analyzing network data to measure bonding and bridging helps to predict important outcomes such as efficiency and innovation: bonding indicates a sense of trusted community where interactions are familiar and efficient; bridging indicates access to new pattern or group.

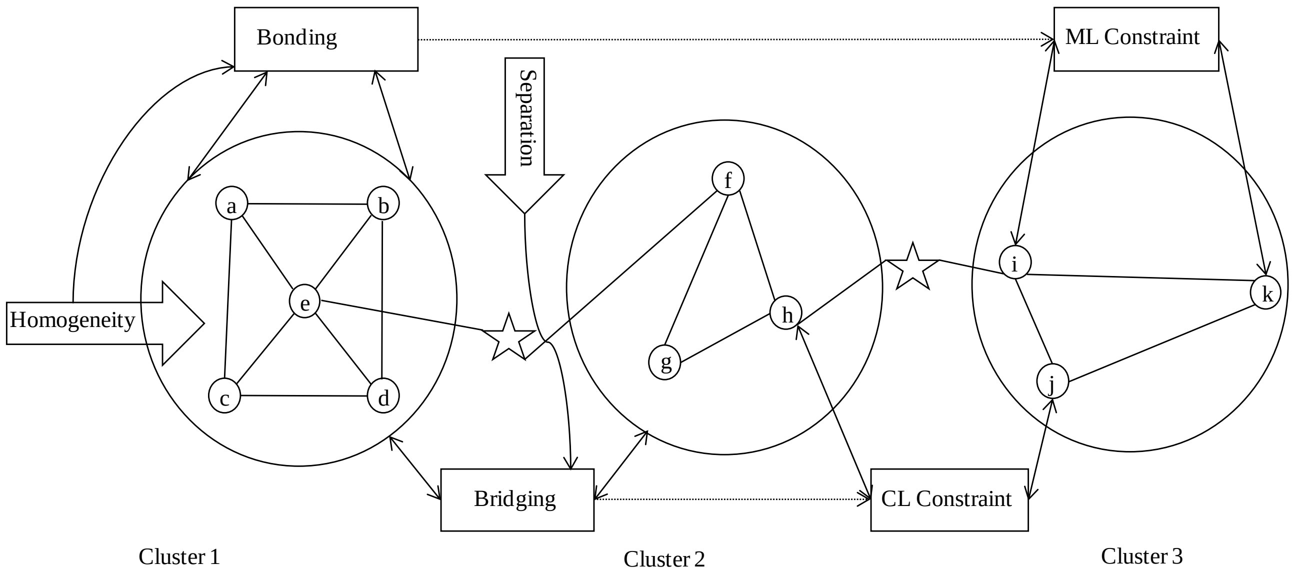

Fig 1: Mapping of Bonding and Bridging to ML & CL Constraints respectively in a Social Network

2. CLUSTER ANALYSIS

2.1 Cluster Analysis Terminology

Cluster analysis aims to partition the entities of a given set into homogeneous and/or well separated classes, called clusters. Clusters are required to be homogeneous and/or well separated. Homogeneity means that entities within the same cluster should resemble one another and separation means that entities in different clusters should differ one from the other. Distance measure [4, 12, 13] determines how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. Entities are partitioned on the basis of dissimilarity values. The split [12] of a cluster is the minimum dissimilarity between any entity in that cluster and any other one outside it. The split of a partition is the minimum of its clusters' splits. The diameter [4, 13] of a cluster is the maximum dissimilarity between any pair of entities of that cluster. The diameter of a partition is the maximum of its clusters' diameters. Minimum Sum of Diameter [4] is Partitioning of the set of subsets, such that the sum of the diameters of the subsets is minimized. Hansen[4] provided an algorithm that solved minimum sum of diameter problem for two clusters that run with time complexity O (n^3logn).

2.2 Clustering Algorithm

There are different types of clustering algorithm [4, 13] Hierarchical algorithms find successive clusters using previously established clusters. Hierarchical algorithms are of two types are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Partitional algorithms determine all clusters at once. Density-based clustering algorithms are devised to discover arbitrary shaped clusters. Subspace clustering methods look for clusters that can only be seen in a particular projection (subspace, manifold) of the data. In Two-way clustering not only the objects are clustered but also the features of the objects

2.3 Cluster Analysis Application

Cluster analysis has applications in statistical data analysis, in machine learning, in data mining, in pattern recognition, in image analysis and in bioinformatics.

3. SATISFIABILITY PROBLEM

3.1 Satisfiability Problem

A Boolean expression is an expression composed of variables, parenthesis and the operators. A formula is said be in conjunctive normal form if a Boolean expression is represented as an expression that is a conjunction of disjunctions, where each disjunction has two arguments that may either be variables or the negations of variables. For example, the following formula is in conjunctive normal form.

$$(a \vee b) \wedge (c \vee b) \wedge (a \vee c)$$

An Expression is Satisfiable if there is some assignments of 0‘s and 1‘s to the variables that gives the expression the value 1. The Satisfiability problem is to determine given a Boolean expression, whether it is Satisfiability [3,5].

3.2 2-Satisfiability (2-SAT)

2-Satisfiability (2-SAT) is the problem of determining the Satisfiability of a formula in conjunctive normal form where each clause is limited to at most two literals.[2]. Aspvall [1] Theorem states that 2-SAT problem can be solved in linear time.

3.3 K-Satisfiability (K-SAT)

K-Satisfiability (K-SAT) is the problem of determining the Satisfiability of a formula in conjunctive normal form where each clause is limited to at most three literals. When the clause size is greater than two, the problem is NP-Complete (Cook 1971). The Cook–Levin theorem [1971] states that the Boolean Satisfiability problem is NP-complete [1].

4. PROBLEM STATEMENT

Let $O = \{O_1, O_2, \ldots, O_N\}$ denote a set of $N = |O|$ entities and $D = \{ d_{ij} / i \leq k \leq N, 1 \leq j \leq N \}$ a set of dissimilarities between pairs of these entities. A dissimilarity d_{ij} is a real number and satisfies to the conditions $d_{ij} \geq 0$, $d_{ii} = 0$, and $d_{ij} = d_{ji}$ for i, j = l, 2, . . . , N.

4.1 K-Cluster Problem Statement

A partition $P_k = \{C_1, C_2, \ldots, C_k\}$ of the entities of O into K clusters is such that no cluster is empty, any pair of clusters has an empty intersection and the union of all clusters is equal to O.

Let Π_K denote the set of all partitions P_K of O into K clusters. We define the diameter of a cluster $C_j \in P_K$ noted d (C_j), as the largest dissimilarity between entities in C_j:

$$d(C_j) = \max_{O_k . O_l \in C_j} d_{kl}$$

The diameter of P_K, noted $d(P_K)$, as the largest of its clusters' diameters:

$$d(P_k) = \min \sum_j d(C_j)$$

Another measure of homogeneity of a partition is sum of its clusters' diameters. This leads to the minimization problem. find P^*_M such that

$$d(P^*_k) = \min_{P_k \in \prod_k} \sum_j d(C_j)$$

We have to find a partition of a given set O of N entities into K non-empty clusters $C_1, C_2, \ldots, C_k$ such that for the given value of $r_1, r_2, \ldots, r_k$ and all constraints are satisfied

4.1.1 Definition 1. The Feasibility Problem[8,10]

Given a set O of data entities, a collection C of constraints, does there exist at least one partition of O into K clusters such that all constraints are satisfied.

4.1.2 Definition 2. The optimization Problem[8,10]

Given a set D of data points, a collection C of constraints, for a given value of K finding partition such that sum of diameter is minimized and all constraints in C are satisfied.

Finding a partition of a given set O of N entities into K non-empty clusters such that $d(C_1) + d(C_2) \ldots\ldots + d(C_k)$ is minimum. Such a partition will be called optimal.

4.1.3 Definition 3 the Feasibility Problem [8, 10]

Given a set D of data points, a collection C of ML and CL constraints on some points in D, upper (K_u) and lower bounds (K_l) on the number of clusters, does there exist at least one partition of D into k clusters such that $K_l \leq K \leq K_u$ and all constraints in C are satisfied?

If the constraints are satisfied then feasible clusters can found at each iteration of clustering under constraints algorithm. The feasibility problem for clustering under ML constraints is in P while clustering under CL only and ML and CL is NP-complete

5. FORMULATION OF CONSTRAINTS IN K-CLUSTERING

Let $O = \{O_1, O_2, \ldots, O_N\}$ denote a set of $N = |O|$ entities and $D = \{ d_{ij} / i \leq k \leq N, 1 \leq j \leq N \}$ a set of dissimilarities between pairs of these entities. A dissimilarity d_{ij} is a real number and satisfies to the conditions $d_{ij} \geq 0$, $d_{ii} = 0$, and $d_{ij} = d_{ji}$ for i, j = l, 2, . . . , N. A partition $P_M = \{C_1, C_2, \ldots, C_k\}$ of the entities of O into K clusters is such that no cluster is empty, any pair of clusters has an empty intersection and the union of all clusters is equal to O.

Minimum sum of diameters clustering problem can be solved by reducing the problem into K-Conjunctive Normal Form or K-SAT.

5.1 Must Link & Can not Link Constraint

Must Link (ML) Constraint: In Must Link Constraint entities O_i and O_j must be in the same cluster. Notation of ML constraint is ML (i, j)

Can Not Link (CL) Constraint: In Can Not Constraint entities O_i and O_j must not be in the same cluster. Notation of CL constraint is CL (i, j).

5.2 Graphical representation of ML & CL Constraints

Consider the following set of constraints:
ML(a,b), ML(b,c), ML(c,d), ML(d,e,), CL(e,f), ML(f,g) ML(g,h), CL(h,i), ML(i,j) & ML(j,k) Constrained Graph for the above constraint is as follows corresponding as follows

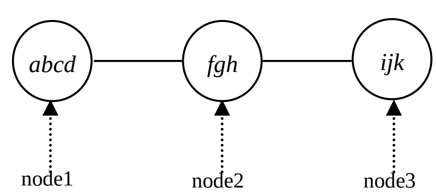

Fig 2: Representation of ML & CL Constraints in a graph

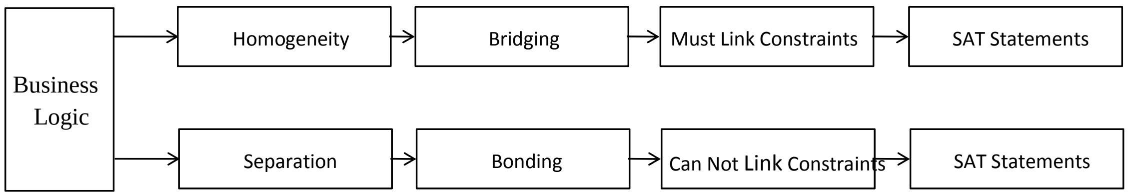

Fig 3: Transformation of Business logic into SAT statements

Compare Figure 1 & Figure 2, node1, node2 and node 3 of figure 2 looks similar to three clusters of figure 1, Cluster 1, Cluster 2 & Cluster 3. Cluster 1 contains entity a, b, c, d, and e; Cluster 2 contains entity f, g and h; Cluster 3 contains only entity i, j and k. It means a social network can be transformed into a mathematical model or in a constrained graph.

One more interesting application feasible constraint graph is that if our network is dynamic(size of network changes at run time) in nature, then if we want to add another entity to an existing feasible network then we can do this by constructing a graph like figure 2. If such a graph (constraint graph like figure 2) is possible it means such a constraint clustering is possible. We can construct a new social network as well as we can check the feasibility of existing social network which is feasible for the given set of constraints or not.

5.3 Transformation of Social Network to SAT Statement

With reference to above section 5.3 and figure 3 it is clear that concept of social networking can be transformed into mathematical model. The transformation process is as follow: Transformation of business logic on the basis of attributes of objects/actor to the properties like homogeneity and separations. These properties homogeneity and separations are transformed into bonding and bridging respectively so they form Social network. These concepts of Social Network are transformed into Must Link & Can Not Link Constraints. These ML & CL constraints are represented in a mathematical form of as a SAT statement.

5.4 Belonging Approach to represent the constraints

Let us consider the set of n entities are partitioned into K cluster C_1, C_2 ,…..,C_k, with diameters r_1, $r_{2,\ldots\ldots,}r_k$ respectively. Assume that $r_1 > r_2 > \ldots\ldots > r_k$.

A new variable type "Belonging Boolean variable" is introduced for the representation of an entity belongs to a particular cluster. So, if a entity O_i belongs to K^{th} Cluster then, it is represented by a Belonging Boolean variable T_{ki}. So there will be K Different Belonging Boolean variable corresponding to K-clusters.

T_{1i}, T_{2i},……, T_{ki} are k Belonging Boolean variables.

Let us associate the following proposition for each entity

Proposition 1. If Entity O_i belongs to Cluster C_k, then associate the rule

$$O_i \ \varepsilon \ C_k \equiv T_{ki}$$

If entity O_k belong to cluster C_1 then $T_{1k} = 1$

5.5 Representation of ML & CL Constraint using belonging approach

Let us consider two Cluster C_1 & C_2 having diameter d_1 & d_2 respectively.

5.5.1 Representation of ML Constraint

In Must Link Constraint entities O_i and O_j must be in the same cluster.

If both entity belong to cluster C_1

$$(T_{1l} \wedge T_{1k}) = 1$$

If both entity belong to cluster C_2

$$(T_{2l} \wedge T_{2k}) = 1$$

5.5.2 Representation of CL Constraint using belonging approach

In Can not Link Constraint entities O_i and O_j must belong to different cluster.

So, if O_i belong to C_1 then O_j belongs to C_2 or if O_i belong to C_2 then O_j belongs to C_1

$$(T_{1l} \wedge T_{2k}) \vee (T_{2l} \wedge T_{1k}) = 1$$

5.6 Constraints for K-Cluster Problem

Constraint can be classified into 3-categories. Total numbers of Possible Constraints are K+1, where K is the number of Cluster.

a) Type 1 Constraint:
If $d_{kl} > r_1$, then O_k and O_l cannot both belong to any of the cluster.
(Only one constraint can possible in this type)

b) Type 2 Constraint:
If $r_j > d_{kl} > r_k$ then

(i) O_k and O_l cannot both belong to the same cluster $C_{j+1}, C_{j+2}, \ldots, C_k$.

(ii) O_k and O_l can both belong to the cluster $C_1, C_{i+1}, \ldots C_j$

Where $2 \le j \le k$
(K-2 constraint can possible in this type)

c) Type 3 Constraint:
If $r_k > d_{kl}$ then there is no restriction, O_k and O_l can belong to any of the cluster.
(Only one constraint can possible in this type)

So total numbers of Possible Constraints type are =
No. of Type 1 Constraint + No. of Type 2 Constraint+ No. of Type 3 Constraint
=> 1 + (K-1) + 1
= > K+1

5.7 Reduction of Constraints into CNF Statement for K-Cluster

All K+1 constraints of section of 4.3 can be reduced as k-SAT statement as follows:

---

a) Constraint 1: If $d_{kl} > r_1$, then O_k and O_l cannot both belong to the same cluster C_1 or C_2 or……. or C_k.

....

if $((O_k \varepsilon C_i) and (O_l \varepsilon C_j)) or .... or ((O_k \varepsilon C_i) and (O_l \varepsilon C_k))$ the n the Boolean formula is:

$$T_{ik} \wedge (T_{1l} \vee T_{jl} \vee ...... \vee T_{kl}) = 1$$

Where i ≠ j and $1 \le i, j \le k$

(i) Example:
if i =1 means

$(O_k \varepsilon C_1) and (O_l \varepsilon\, cluster other then C_1)$ then

$((O_k \varepsilon C_1) and (O_l \varepsilon C_2)) or ((O_k \varepsilon C_1) and (O_l \varepsilon C_3)) or ...... or ((O_k \varepsilon C_1) and (O_l \varepsilon C_k))$

Then the Boolean formula is:

$$T_{1k} \wedge (T_{2l} \vee T_{3l} \vee .... \vee T_{kl}) = 1$$

(ii) Example:
if i = 2 means

$(O_k \varepsilon C_2) and (O_l \varepsilon\, cluster other then C_2)$ then

$((O_k \varepsilon C_2) and (O_l \varepsilon C_1)) or ((O_k \varepsilon C_2) and (O_l \varepsilon C_3)) or .... or ((O_k \varepsilon C_2) and (O_l \varepsilon C_k))$

Then the Boolean formula is:

$$T_{2l} \wedge (T_{1l} \vee T_{3l} \vee .... \vee T_{kl}) = 1$$

Where $1 \le i, j \le k$

....

b) Constraint 2: $r_j > d_{kl} > r_{j+1} > \ldots\ldots > r_k$ then O_k and O_l cannot both belong to the cluster $C_{j+1}, C_{j+2}, \ldots C_k$. and O_k and O_l can both belong to the cluster $C_1, C_{i+1}, \ldots C_j$

....

$((O_k \varepsilon C_{j+1}) and (O_l \varepsilon C_{j+2})) or ((O_k \varepsilon C_{j+1}) and (O_l \varepsilon C_{j+3})) or ...... or ((O_k \varepsilon C_{j+1}) and (O_l \varepsilon C_k))$

Then the Boolean formula is:

$$T_{(j+1)k} \wedge (T_{(j+2)l} \vee .... \vee T_{kl}) = 1$$

$$(T_{1k} \vee T_{1l}) \wedge (T_{2k} \vee T_{2j}) \wedge ...... \wedge (T_{jk} \vee T_{ij}) = 1$$

Where $1 \le j \le k$

(i) Example:
If $r_1 > d_{kl} > i_2$, then O_k and O_l cannot both belong to the cluster C_2 or C_3 or……. or C_k.

$$((O_k \varepsilon C_2) and (O_l \varepsilon C_3)) or .... or ((O_k \varepsilon C_2) and (O_l \varepsilon C_k))$$

$$T_{2k} \wedge (T_{3l} \vee .... \vee T_{kl}) = 1$$

$$(T_{1k} \wedge T_{1l}) = 1$$

(ii) Example:
If $r_1 > r_2 > d_{kl} > r_3$, then O_k and O_l cannot both belong to Cluster C_3 or……. or C_k.

$$((O_k \varepsilon C_3) and (O_l \varepsilon C_4)) or .... or ((O_k \varepsilon C_3) and (O_l \varepsilon C_k))$$

$$T_{3k} \wedge (T_{4l} \vee .... \vee T_{kl}) = 1$$

$$(T_{1k} \wedge T_{1l}) \vee (T_{2k} \wedge T_{2l}) = 1$$

....

c) Constraint 3 : If $r_{3,} > d_{kl}$ then there is no restriction, O_k and O_l can belong to any cluster.

5.8 Checking Satisfiability of Boolean Equation

From the Constraints obtained from section 5.6, a K-SAT instance is generated. These K-SAT instances can be solved by any K-SAT Solver algorithm, to check the Satisfiability of the Boolean expression constructed for some (r_0, r_1, . . ., r_k). If the solution is a feasible solution according to Definition 1 then a K-clustering is possible. There can be more then one feasible solution for the values of (r_0, r_1, r_2). One of the solutions among the feasible solution is optimum solution. according to Definition 2.

6. CONCLUSION

Social Network Analysis is fast growing field data mining. Clustering divides a social network into different classes according to properties or pattern of objects in the social network. Objective of K-Clustering is to partition the entities into k-clusters on the basis of properties or pattern of the entities. In this paper we investigated that how the concept of Social network are transformed into ML and CL constraint and a formulation is proposed which is based on a beautiful and simple concept of "belongingness" to represent constraint in the CNF Statement or K-SAT statements. Feasibility of k-SAT statement can be checked by any SAT solver algorithm.

Boolean approach is a method to identify and represent constraint in 2-cluster analysis which are represented in the form of 2-SAT statement. Hansen [4] uses this approach to find out minimum sum of diameter for two cluster. Boolean approach method is not sufficient to represent the constraint for more then two cluster. Approach of our earlier paper [11], is extended in this paper to represent the constraints in K-cluster. This paper presents a formulation that find out the constraints in k- cluster based on concept of bonding and bridging in social network, Constraints generated are represented in the form of K-SAT statement.

7. REFERENCES

[1] B. Aspvall, M.F. Plass and R.E. Tarjan. A Linear-time Algorithm for Testing the Truth of Certain Quantified Boolean Expressions, IPL, 8, 1979, pp 121-123

[2] A.V. Aho, I.E. Hopcroft and J.D. Ullman: The Design and Analysis of Computer Algorithms, Addison-Wesley, Reading, MA, 1974, pg no. 189- 194

[3] John E. Hopcroft, Jefrry D. Ullman: Introduction to automaton theory, languages and computation. pg no. 324-325

[4] P.Hansen, B. Jaumard: Minimum Sum of Diameters, Journal of Classification 4: 215- 226(1987)

[5] John F. Kohen: An on Line Satisfiability For Conjunctive Normal form Expressions with two literals, Flairs -02, Proceedings,187 – 191

[6] Wagsta, Cardie," Clustering with Instance-level Constraints", Proceedings of the Seventeenth International Conference on Machine Learning, 2000, p. 1103-1110.

[7] Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001) Constrained k-means clustering with background knowledge Proc. of 18th Intl. Conf. on Machine Learning.

[8] I. Davidson and S. S. Ravi, "Identifying and Generating Easy Sets of Constraints For Clustering", Proceedings of the Twenty-First National Conference on artificial Intelligence (AAAI 2006), Boston, MA, July 2006, 6 pages

[9] I. Davidson, S. S. Ravi & Shamis, "A SAT-based Framework for Efficient Constrained Clustering" Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010 94-105.

[10] Davidson and S. S. Ravi, "Using Instance-Level Constraints in Hierarchical Agglomerative Clustering: Theoretical and Empirical Results", Data Mining and Knowledge Discovery, Vol. 18, No. 2, Apr. 2009,

[11] R. K. Jain, N.S. Chaudhari: Identification and Generation of Constraints in Social Network, Proceedings of International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), 2011, IEEE pp no 11-14

[12] Delattre, M., And Hansen, P. (1980): Bicriterion Cluster Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2(4), 277-291.

[13] P.Hansen, B. Jaumard: Cluster Analysis and Mathematical Programming, Mathematical Programming, 79, 1997, pp 191 - 215.

[14] Jiang and Carroll, "Social Capital, Social Network and Identity Bonds: A Reconceptualization" C&T '09 Proceedings of the fourth international conference on Communities and technologies, ACM, pg no 51-60.

[15] Nina Mishra, Robert Schreiber, Isabelle Stanton and Robert E. Tarjan, Clustering Social Networks, Springer Lecture Notes in Computer Science,2007, Volume 4863/2007,56-67