Time Varying Community Structure extraction scheme for Temporally Valid Multi Dimensional Large Scale Social Networks

B.Sindhuja
M.Tech,CSE Department,
G.Narayanamma Institute of Technology and Science,Hyderabad,INDIA.

Dr.K. Venu Gopala Rao
Professor,CSE Department,
G.Narayanamma Institute of Technology and Science,Hyderabad,INDIA.

Abstract---Complex systems being accepting increasing consideration by the scientific network, moreover because of the accessibility of significant system data from various domains. One concern analyzed up to now in complex system evaluation is Community Discovery, i.e. the recognition of group of nodes largely associated, or highly associated. On the other hand, one element of these systems has been ignored so far: significant systems are usually temporally applicable multidimensional, i.e. temporally applicable numerous associations may reside around every two nodes, either to show various types of temporally applicable connections, or to associate nodes by various values of the similar kind of tie. In this perspective, the concern of Community Discovery should be changed, considering temporally applicable multidimensionality. In this document, we try to accomplish, by determining the concern in the temporally valid multidimensional perspective, also by providing another new measure allowed to define the communities discovered. We then incorporate a complete framework for choosing and defining values of the similar kind of temporally valid tie. Our studies on real life temporally valid multidimensional systems maintain the strategy suggested in this document, also open the option for a new category of algorithms, targeted at acquiring the multifaceted complexity of associations concerning nodes in a system.

Index Terms: Communities, Structure Extraction, Temporal Networks, and Seed items.

1. Introduction

Determined by real-world situations like social communities, technology systems, the Web, biological systems, etc, in the past years, wide [3], multidisciplinary, as well as significant analysis has been dedicated to the retrieval of non trivial information from these networks [13,4,5]. Anticipating potential links concerning the actors of a system, detecting as well as exploring the distribution of data among them [1,2], mining consistent patterns of users' routines, are just a few instances of the purpose in the area of Complex Network Analysis which contains, among everyone, mathematicians, physicians, computer scientists, sociologists, economists as well as biologists [9,10].

The information at the perspective of this area of analysis is enormous [4], heterogeneous, also semantically productive, as well as this enables determining various properties also dealings of the actors associated in a system. One essential process at the perspective of Complex Network Analysis will be Community Discovery, [8,9,10] i.e., the revelation of set of nodes largely associated, or highly connected. There are various strategies capable of determine communities in systems [15], affording revealing hierarchical associations, effective nodes in networks [14], or just set of nodes that reveal certain properties or routines. To do so, the associations involving the nodes of a system are presented at the target of research [18], as they perform a vital role in the analysis of the system structure, development, and tendencies.

Present, a lot of the work completed in the literary works is reduced to a quite easy understanding of these associations, centering only on either two nodes are associated or not. In the real life, anyhow [19,16], this will be not continually sufficient to model every obtainable data, specifically if the actors are customers, with their numerous desires, their complex actions, also their multifaceted connections. A most complex evaluation of this component could assist all the strategies basing their effectiveness on the information of the design of a system.

ISSN: 2278 – 7798

To this plan [20], in this document we handle temporally valid multidimensional systems, i.e. communities in which several associations may occur around a set of nodes [2,7], exhibiting numerous connections (i.e., dimensions) around them. Temporally valid multidimensionality in significant systems can be shown by either various kinds of associations (two persons can be linked as they are colleagues, friends, they play with each other in a team, etc), or various quantitative values of one particular connection (co-authorship around two authors can manifest in many countless years, for instance) [1,2,3]. We may describe around explicit or implicit specifications, the former remaining connections explicitly fixed by the nodes (friendship, for instance), while the alternative being connections inferred by the expert, that can connect two nodes based on their likeness or other concepts (two users can be passively connected if they published a content on the equivalent topic).

2. Related work

In this situation, we propose the issue of temporally valid multidimensional Community Discovery [11,12], i.e. the issue of revealing communities of actors in temporally valid multidimensional systems. We determine an approach of temporally valid multidimensional network, also we propose a new estimate targeted at evaluating the temporally valid multidimensional attributes of the networks revealed. We then provide a system for locating as well as characterizing temporally valid multidimensional networks also we reveal the outcomes attained by using that system on real-world communities [17], providing a few instances of appealing temporally valid multidimensional networks discover in various situations: movie collaborations as well as terrorist assaults.

Our major factor is then: we propose as well as formally determine the issue of temporally valid multidimensional network discovery [19]; we propose a estimate for characterizing the networks revealed; we develop a structure for handling the released issue with a connection of existing strategies also our newly presented aspects [4]; we execute a report on real systems, revealing the outcomes acquired and the description of the networks.

There are numerous experiments on network discovery, in numerous fields of analysis: computer science, physics, sociology, a lot more. Many of them could be collected based on the characterization of network they utilize.

One probability is determining a network as a group of nodes using a large density of connections concerning them [6], while there are actually sparse associations among various networks. The documents performing with this description depend upon knowledge theoric concepts or on the strategy of modularity, and if a function described to discover the ratio around intra- as well as inter-community amounts of edges. Modularity will be extensively analyzed as well as extended in numerous works [12]: one among them will be a greedy optimization capable of measure up to systems with enormous amounts of edges.

Other works depend on many statistical attributes of the graph. In a structure for the discovery of overlapping networks, i.e. networks permitting the vertices to stay about one network, is provided.

The other class of techniques depend on the propagation inside system of a description or a specific classification of structure (normally a clique). The initial strategy is recognized for becoming a quasi linear remedy for the issue; the next one permits finding overlapping networks.

One algorithm which attempts to optimize quality as well as quantity procedures on its outcomes will be InfoMap, an arbitrary walk-based algorithm. An appearing novel issue description might be revealed in, wherein author's condition that network discovery algorithms might not cluster nodes but ends, focusing the purpose of the connection located in a network.

Considering 2009, temporally valid multidimensionality has established to be considered into account in the network discovery issue. To the ideal of our understanding, the primary techniques are two. In the authors prolong the description of modularity to blend to the temporally valid multidimensional situation, which they label "multislice". In the authors generate a machine learning process which finds the available various latent measurements concerning the organizations in the system also uses them as attributes for the classification algorithm. It is significant to note in which both techniques do not choose any description of "temporally valid multidimensional community", neither they define and evaluate the networks discovered also their temporally valid multidimensional framework: their primary constraint is to basically determine a technique for handling temporally valid multidimensional systems, getting mono dimensional networks as result.

ISSN: 2278 – 7798

3. FINDING AND CHARACTERIZING TEMPORALLYVALID MULTI DIMENSIONAL COMMUNITIES

In this segment, following a design for temporally valid multidimensional systems, we determine temporally valid multidimensional networks, a measure targeted at characterizing them, also the issue handled in this document.

A. A model for temporally valid multidimensional networks

We utilize a *multigraph* to design a temporally valid multidimensional system also its attributes. With regard to efficiency, in our design we just reveal undirected multigraphs also as we don't choose node labels, henceforth we utilize *edge-labeled undirected multigraphs,* represented by a triple G = (V, E, D) where: V is a group of nodes; D is a group of labels; E is a group of labeled ends, i.e. the group of triples (u, v, d) where u, $v \in V$ will be nodes as well as $d \in D$ will be a label. And, we utilize the word *dimension* to signify *label*, also we state that a node *belongs to* or *appears in* a provided dimension *d* if there will be one edge labeled with d adjacent to it. We even state that an edge *belongs to* or *appears in* a dimension *d* if its label will be *d*. We suppose that provide a pair of nodes u, $v \in V$ also a label $d \in D$ just one edge (u, v, d) can exist. Hence, every pair of nodes in G may be associated by essentially |D| probale edges.

B. Temporally valid multidimensional Community

The literary works on network discovery provides a significant amount of assorted descriptions of network. Incorporating temporally valid multidimensionality to the issue prospects to an especially opinable approach of temporally valid multidimensional network. We begin with a high-level conceivable description, also then we consider to incorporate additional semantic to it.

Definition 1 (Temporally valid multidimensional Community): A *temporally valid multidimensional community* is a set of nodes densely associated in a temporally valid multidimensional community.

Since we notice, whereas in a mono dimensional community the density of a network relates unambiguously to the proportion around the quantity of edges concerning the nodes as well as the quantity of every available edges, the temporally valid multidimensional environment provides an extra degree of freedom (i.e., the various dimensions). Choose Figure 1: in (a) we come with a network whose density mainly counts by the connection offered by one dimension; in (b) we come with another scenario, while both the dimensions tend to be causing the density of the network. Could the two be regarded similar or may we ascertain concerning them? Being answer this query, we determine a measure, p, targeted at characterizing temporally valid multidimensional networks. To render it doable to evaluate its values involving various communities, we generate it choose values in [0,1]. In the appropriate we utilize this notation: c will be a temporally valid multidimensional network; d will be a dimension in D; P will be set of pairs (u, v) associated by at least single dimension in the community; $\overline{\overline{P}}$ will be the set of pairs associated by at least two dimensions p_c will be the subset of P listed in c; $\overline{\overline{P}}_c \subseteq \overline{\overline{P}}$ will be the subset of $\overline{\overline{P}}$ incorporating just pairs in c.

C. Redundancy ρ

The strategy we determine is known as redundancy; also it captures the occurrence in which a group of nodes that represent a network in a dimension usually represent a network also in different dimensions. We may notice this estimate as a basic indicator of the redundancy of the associations: the additional dimensions associate every set of nodes inside a network, the increased the redundancy should be. We may then determine the redundancy ρ by counting the number of pairs come with redundant associations, normalizing with the theoretical optimum.

$$\rho_c = \sum_{(\mathrm{u,v}) \in \overline{\overline{P}}_c} \frac{\left|\{d : \Im(\mathrm{u,v,d}) \in \mathrm{E}\right|}{\left|D\right| \times \left|P_c\right|}$$

With the assistance of Figure 1 we notice how ρ provides values in [0,1]: in 1(b), every pair of nodes will be associated in just one dimension, then $\left|\overline{\overline{P}}\right| = 0$ also the numerator will be equal to zero; in 1(c), all the node pairs tend to be associated in all the dimensions of D, that will be similar to the amount of associated pairs $|P_c|$ multiplied by the amount of system dimensions | D| (the denominator), making ρ = 1. We notice that p is undefined for networks generated by a single node, where $|P_c|$ = 0 also then the denominator will be equal to zero. To this kind of networks, anyhow, the redundancy measure will not significant, therefore we may ignore this case.

ISSN: 2278 – 7798

D. Problem definition

We could now prepare the issue under analysis:

Problem 1 (MCD): Provided a temporally valid multidimensional system G, choose the complete group of temporally valid multidimensional systems C, also characterize every $c \in C$ based on ρ

In this offer, we are considering in developing that seed-based network discovery strategy for a temporally valid multidimensional system. There are numerous data exploration as well as data retrieval programs wherein there are many dimensions/entities in systems. In web social media, networked information comprises of several dimensions/entities including tags, pics, comments also stories. We are considering to discover a set of users who connect considerably on such media entities. In a co-citation system, we are considering to discover a set of authors who cite/collaborate together (or a group of documents which are associated with one another) considerably on publication details in titles, abstracts, also keywords as several dimensions/entities in the system.

Fig. 1a reveals an instance of an educational publication system where many aspects are described to documents, also every document is connected with multiple keywords, authors as well as time frame. In this system, there tend to be five dimensions/entities. Anyhow, items in three dimensions (author, keyword as well as paper) tend to be temporally associated among them, also items in two dimensions (paper as well as concept) tend to be temporally connected to one another. We may utilize a tensor as well as a matrix

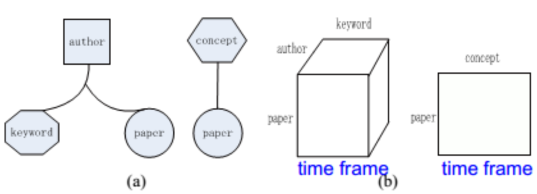

Fig. 1. (a) An example of an academic publication network. (b) The representation of the multi-dimensional network in (a): a tensor is used to represent the interactions among items in three dimensions/entities: author, keyword and paper, and a matrix is used to represent the interactions between items in two dimensions/entities: concept and paper.

to describe their connections, find Fig. 1b. For the comprehensive description, we can reveal in Section 3. Multiple connections concerning entities/dimensions might be included and analyzed to determine useful as well as significant community framework in that temporally valid multidimensional system.

The primary objective of this document is to suggest an algorithm, Time varying MultiComm, to determine a seed-based temporally valid network framework in a multi-dimensional system in a way that the included items of the entities around the network connect considerably, also subsequently, they are not extremely determined by the items away from network. In our proposition, a network is designed beginning with a seed comprising several items of the entities considered to be contributing in a worthwhile network. Provided the seed item, we iteratively adjoin emerging items by assessing the affinity around the items to construct a network in the system. When there are several connections concerning the items from various dimensions/entities in a temporally valid multidimensional system, the primary concern is strategy to assess the affinity around the two items in the equivalent kind of entity (from the equivalent dimension/entity) or in various sorts of entities (from various dimensions/entities). For instance, in Fig. 3, the affinity around a paper “A” as well as a paper “B” (the similar kind of entity), also the affinity around a paper “A” as well as a keyword “C” (various kinds of entities) tend to be needed to assess and determine the papers “A” as well as “B” or the paper “A” also the keyword “C” to put collectively in a network. However, we require a standard to assess a high standard of produced networks by the suggested algorithm, thereby we analyze a local modularity estimate of a network in a multi-dimensional system. Studies according to synthetic as well as real-world information recommend that the suggested system is capable to discover a network efficiently. Empirical results have revealed that the efficiency of the suggested algorithm is adept in reliability than the other evaluation algorithms in discovering networks.

4. Social network with temporal validity

The authors provided in a prominent science journals showing site that comprised of 2310 authors of the extend of 6 to 12 topics. Every associate was provided with a threshold amount of journals with a unique identifier. The developed procedure displays the nearness of every two authors inside the provided temporal threshold number, which is regarding a topic redundancy. The information is obtained from any of science publications indexing sites like DOI, Scopus or DBLP. The program has a temporal resolution threshold, to ensure closeness connections are recognized over successive temporal resolution threshold time periods. The experimental information

ISSN: 2278 – 7798

are hence normally symbolized as a temporal social network. The information is preprocessed which discards some records using data excellence factors, so in these we can work with a temporal system with N~2310 nodes.

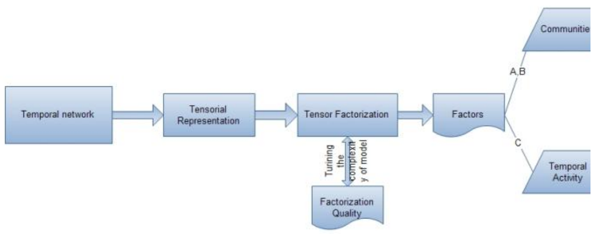

Fig 2: Shows the architecture for temporal networks

Network structure of the public: No private data is connected with the chosen nodes. Anyhow, every identifier is connected to the subject the author is associated to, to ensure we come with a ground reality for the networks that determine the network structure of the examined population.

The temporal community dataset we utilize consists of 2310 authors also their connection with a temporal quality threshold provided. The routine of topics as well as the article information that we utilize as a ground reality for the activity timelines, although, is described on a coarser temporal level. Therefore for the current analysis we total the information with various time periods. In our studies, these assorted levels of aggregation using the temporal level of the activity timelines is discovered.

5. Uncovering Latent Structures by Tensor Factorization

The tensor $\mathrm{T} \in \mathbb{R}^{N \times N \times S}$, where N is the amount of nodes of the system also S the amount of system snapshots, encodes either the topological as well as temporal data on the system under analysis. Finding structures that can equate to networks or correlated activity models involves the recognition as well as extraction of lower-dimensional aspects. To this conclude, we utilize tensor factorization strategies, i.e., we select to describe the tensor as an appropriate supplement of lower-dimensional aspects. This may be attained through the alleged canonical decomposition (canonical polyadic decomposition, CP). CP in 3 dimensions targets at writing a tensor $\mathrm{T} \in \mathbb{R}^{N \times N \times S}$ in a factorized manner:

$$t_{ijk} = \sum_{r=1}^{R_T} a_{ir} b_{jr} c_{kr}$$

5.1 Factorization Methodology

In the current situation, every rank-1 tensor, which we henceforth label element, refers to a group of nodes as their activities are associated. The focus here just to discover an appropriate factorization, but instead to estimate the tensor with an amount of elements lesser than the rank of the initial tensor. This kind of an approximation concerning the tensor is similar to reducing the variation between T as well as [[A,B,C]] (PARAFAC decomposition),

$$\min_{A,B,C} \left\| \Gamma - A, B, C \right\|_F^2$$

where A,B,C correspondingly have dimensions $N \times R, N \times R$ as well as SxR, $S \times R, \mathrm{R} < R_\Gamma$

Handling this issue quantities to choosing the R rank-1 tensors that really estimated the tensor T. The number R of elements is selected on the perspective of the required level of information: a low amount of elements only results in the strongest frameworks, probably ignoring significant features, however utilizing a elevated amount of elements deals with the chances of overfitting noise. Selecting R quantities to an optimization issue wherein we obtain the amount of elements that preferred describe the framework of the tensor without outlining the conceivable noise of the information. In this regard, the tensor factorization technique is equivalent to network discovery strategies where the amount of networks is resolved a priori: the range of elements we select to estimate the tensor is the amount of networks or activity models we acquire.

5.2 Assessing the Quality of Factorization: How to Control a Multi-scale Method

Enhancing the amount R of elements enables us to describe progressively structure of the temporal system. Although, as the amount of elements improves, we choose from under appropriate to overfitting these frameworks, i.e., we deal with the ordinary trade-off around approximating elaborate structures as well as overfitting them, possibly getting noise. This is a attribute of any fundamentally multi-scale technique, also it is significant to handle it by developing and utilizing quality metrics for the

ISSN: 2278 – 7798

acquired decompositions that may assist their usage in the framework of a certain research concern or software. Determine that we don't intend at establishing an ‘‘optimal’’ amount of elements, but instead at evaluating the standard of a decomposition acquired for a provided selection of R. Here we render usage of an ordinary metric known as ‘‘core consistency”, which we quickly illustrate in the subsequent.

We observe that the tensor decomposition of Eq. 1 may be made as

$$t_{ijk} = \sum_{m=1}^{R_T} \sum_{n=1}^{R_T} \sum_{p=1}^{R_T} g^1_{mnp} a_{im} b_{jn} c_{kp}$$

Where $g^1_{ijk} = \delta_{ij}\delta_{kj}\delta_{ik}$ is the unit super diagonal tensor.

5.3 Interpreting the Factors: Community Structure and Activity Patterns

The component matrices A, B, C each have R columns, all of them related to one element. As we utilized non-negative tensor factorization, every entry of these matrices tend to be non-negative. In the certain case of an undirected system, in common A = B. The components a_{ir} of matrix A connect every element r to the nodes *i* it spans, i.e., they illustrate network framework of the original system, with the matrix records giving weights for the account of nodes to these networks. In the instance of a directed system, both A as well as B are required to encode the framework of the system, i.e., the concept of node account to an element gets a directed connection. The matrix component air defines the weight of the arriving account of node i to element r, and a_{ir} defines the weight of the outbound account of node *i* to element *r*. To both situations, directed as well as undirected systems, the components c_{kr} of matrix C, alternatively, connect every element *r* to the time periods *k* it spans, also the matrix values for a provided element show the activity level of that element as a function of time (index k), i.e., its temporal activity routine.

We comment that specific nodes may be users of various elements, with various weights. This will be, non-negative factorization of the temporal system tensor may commonly obtain overlapping networks. This mirrors the outcomes of the research by Yang as well as Leskovec, wherein non-negative matrix factorization would be revealed efficient at discovering densely overlapping and non-overlapping networks in static systems. Likewise, non-negative tensor factorization enables to pull non-overlapping temporal networks, densely overlapping networks, also multi-scale network framework.

When noted earlier, the factor C results in the temporal procedure of every element (network), in spite of the node structure of the element. We determine the activity stage of every network r at a provided period k (time index) by utilizing both the details on the temporal procedure of the element (from C) also the memberships durability of every node (from A) in that element. The durability of an element r at period k is hence described as

$$s_{kr} = c_{kr} \sum_{i=1}^{N} a_{ir}$$

where r= 1,... ,R also k= 1,... ,S. Observe that the procedure of an element during time may be extremely inconsistent, i.e., it is doable to obtain frameworks (elements) that have temporally-disjoint activity locations. This will be a result of the concept that nonnegative tensor factorization records simply constructive elements of the original system tensor also doesn't depend or impose at all issues of temporal continuity on the recognized frameworks.

6. Results & Discussion

We examined the usage of developed non-negative tensor factorization strategies for the discovery of the temporally appropriate network activity framework of temporal systems. The strategy we suggest is fundamentally temporal as well as enables to concurrently determining temporally valid system communities collectively with their activity designs overtime. Provided the lack of commonly recognized benchmarks for discovering the temporally valid network framework of temporal systems, we assessed the technique by centering on the certain instance of an experimental temporal social system in which we come with a ground reality that enables us to temporally verify the frameworks we identify in provisions of recognized groups as well as recognized activity routines. The instance we examine is a time-varying social system assessed of wearable detectors: this will be a specifically rich as well as complicated dataset, as it attributes complex frameworks at several scales, both topologically as well as temporally, overlapping networks, also in basic designs developing from the social as well as organizational framework of the situations.

ISSN: 2278 – 7798

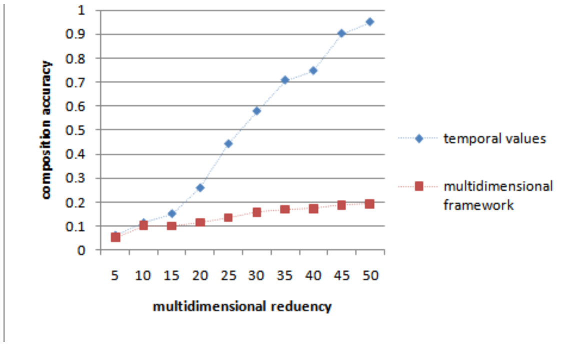

Fig . 2. The cumulative distribution for temporally valid ρ

7. CONCLUSIONS

We provide dealt with the issue of temporally valid network detection, used to the situation of temporally valid multidimensional systems. We come with provided a achievable description of temporally valid multidimensional network also supplied an estimate targeted at define the temporally valid networks discovered. On this perspective, we have developed a structure for locating as well as characterizing temporally valid multidimensional networks, which can be according to a mapping with temporally valid multidimensional to mono dimensional system, on the program of established mono dimensional temporally valid network detection algorithms to it, on the recovering of the actually located temporally valid multidimensional framework of the networks, basically the characterization of them using the ρ measures.

REFERENCES

[1] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multi-scale complexity in networks. Nature, 2010.

[2] Fabr´ ıcio Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virg´ ılioA. F. Almeida. Characterizing user behavior in online social networks.In Internet Measurement Conference, pages 49–62, 2009.

[3] Michele Berlingerio, Michele Coscia, and Fosca Giannotti. Mining thetemporal dimension of the information propagation. In IDA, pages 237–248, 2009.

[4] Bringmann Bjoern, Berlingerio Michele, Bonchi Francesco, and GionisArisitdes. Learning and predicting the evolution of social networks.IEEE Intelligent Systems, 25(4):26–35, 2010.[5] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, andEtienne Lefebvre. Fast unfolding of communities in large networks.J.STAT.MECH., page P10008, 2008.

[6] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Findingcommunity structure in very large networks. Physical Review E,70:066111, 2004.

[7] Diane J. Cook, Aaron S. Crandall, Geetika Singla, and Brian Thomas.Detection of social interaction in smart spaces. Cybernetics and Systems,41(2):90–104, 2010.

[8] S. Fortunato and M. Barthelemy. Resolution limit in communitydetection. Proceedings of the National Academy of Science, 104:36–41, January 2007.

[9] Santo Fortunato. Community detection in graphs. Physics Reports,486(3-5):75 – 174, 2010.

[10] Steve Gregory. Finding overlapping communities using disjoint community detection algorithms. In Complex Networks: CompleNet 2009,pages 47–61. Springer-Verlag, May 2009.

[11] Jure Leskovec, Kevin J. Lang, and Michael W. Mahoney. Empiricalcomparison of algorithms for network community detection. In WWW,pages 631–640, 2010.

[12] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela.Community structure in time-dependent, multiscale, and multiplex networks. Science 328, 876, 2010.

[13] David L. Nowell and Jon Kleinberg. The link prediction problem forsocial networks. In CIKM ’03, pages 556–559, New York, NY, USA,2003. ACM.

[14] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncoveringthe overlapping community structure of complex networks in nature andsociety. Nature, 435(7043):814–818, June 2005.

[15] Spiros Papadimitriou, Jimeng Sun, Christos Faloutsos, and Philip S. Yu.Hierarchical, parameter-free community discovery. In ECML PKDD ’08,pages 170–187, Berlin, Heidelberg, 2008. Springer-Verlag.

[16] Pascal Pons and Matthieu Latapy. Computing communities in largenetworks using random walks. In Computer and Information Sciences -ISCIS 2005, volume 3733, chapter 31, pages 284–293. Springer BerlinHeidelberg, Berlin, Heidelberg, 2005.

ISSN: 2278 – 7798

[17] Usha Nandini Raghavan, Reka Albert, and Soundar Kumara. Near lineartime algorithm to detect community structures in large-scale networks.Physical Review E, 76:036106, 2007.

[18] M. Rosvall and C. T. Bergstrom. Maps of random walks on complexnetworks reveal community structure. PNAS, 105:1118–1123, January2008.

[19] Lei Tang and Huan Liu. Scalable learning of collective behavior basedon sparse social dimensions. In CIKM, 2009.

[20] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure patternmining. ICDM ’02, pages 721–. IEEE Computer Society, 2002

About the authors**--**

B.sindhuja completed B.Tech in CSE from Vishnu Institute Of Technology , Bhimavaram in 2012. She is Pursuing M.Tech ,CSE in G.Narayanamma Institute of Technology and Science, Hyderabad. Her Research interests include Data mining and Big Data .

Dr. K. Venugopala Rao He was a recipient of **Academic Excellence Award** by GNITS in 2009. He is also recipient of **The Best Engineering college Teacher(Computer Science & Engineering)** Andhra Pradesh State Award by ISTE(Indian society for Technical Education),A.P Section for the year 2013. National level in 2014, and having 25 years of teaching experience. He has published 75 papers in International and National Journals and Conferences. 2 PhD’s awarded under his guidance .He is presently guiding 12Ph.D students.

ISSN: 2278 – 7798