ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

Quantify the intersection of behavior in the work environment

K.Hareesh Kumar, D.Uma Devi

***Abstract*—This is the study of intersection of behaviour to analyze the behaviour of users in the social networking environment. We have so many opportunities and challenges to analyze the behaviour on a large scale. A social-dimension-based approach has been shown effective in addressing the heterogeneity of connections presented in social media. Day by day the clicks are increasing in a particular network. To know the interest and requirement of users we propose an edge-centric clustering scheme to extract sparse social dimensions. Even though there are thousands and lacks of users involved in the social network, the proposed approach can efficiently analyze the even behaviour of the users. This helps to improve the odd areas as well as even areas where the people stick on to their desires.**

***Index Terms*—Collective inference, edge-centric approach, identical activities, social dimensions.**

I. INTRODUCTION

Human intention is the precious source in the present world. As the trends in the technology are increasing, there is a need to know and solve the issues of the visitors who leave their foot prints in their interested areas. We have so many social networking websites. If analysis is made on the interested areas of different people, the united behaviour of the people can be identified. Social media provides sample opportunities to study human interactions and Intersection interests on an unprecedented scale. In this work, we study how networks in social media can help predict some human behaviours and individual preferences. In particular, given the behaviour of some individuals in a network, how can we infer the behaviour of other individuals in the same social network?

This study can help better understand behavioural patterns of users in social media for applications like social advertising and recommendation. Typically, the connections in social media networks are not homogeneous. Different connections are associated with distinctive relations. For example, one user might maintain connections simultaneously to his friends, family, college classmates, and colleagues.

This relationship information, however, is not always fully available in reality. Mostly, we have access to the connectivity information between users, but we have no idea why they are connected to each other. This heterogeneity of connections limits the effectiveness of a commonly used technique — collective inference for network classification. A recent framework based on social dimensions is shown to be effective in addressing this heterogeneity. The framework suggests a novel way of network classification: first, capture

the latent affiliations of actors by extracting social imensions based on network connectivity, and next, apply extant data mining techniques to classification based on the extracted dimensions. In the initial study, modularity maximization [3] was employed to extract social dimensions. The superiority of this framework over other representative relational learning methods has been verified with social media data in the original framework, however, is not scalable to handle networks of colossal sizes because the extracted social dimensions are rather dense. In social media, a network of millions of actors is very common. With a huge number of actors, extracted dense social dimensions cannot even be held in memory, causing a serious computational problem. Sparsifying social dimensions can be effective in eliminating the scalability bottleneck. In this work, we propose an effective edge-centric approach to extract sparse social dimensions [4]. We prove that with our proposed approach, sparsity of social dimensions is guaranteed. Extensive experiments are then conducted with social media data. The framework based on sparse social dimensions, without sacrificing the prediction performance, is capable of efficiently handling real-world networks of millions of actors.

II. INTERSECTION OF INTERESTS

Intersection of interests refers to the common work area in the social network like face book, twitter, orkut etc. The behaviour of users may depend on many parameters. A new user generally searches his friend and communicates with his friends and sometimes he follows the suggestions given by his friends. So by knowingly or unknowingly the behaviour of the user may be influenced by his friends, family and work environment. It generally leads to the developing interests in his friends' fashion. Sometimes friends of friends may influence the user regarding social usage in the network. For example if friend buys a new mobile we generally show interest to buy the same mobile or another different mobile in the same mobile store. The similar type interest will be shown by the people in the social network usage also.

Behaving similar to some persons, having ideas, interests similar to some group of persons can be called as homogeneous behaviour/interests. This can be measured as the tendency of behaving one person similar to another with the help of some parameters. In this world, any person generally likes another person who has similar interests or hobbies or ideas etc. When they meet each other, they show

***Manuscript received Oct 15, 2011*.**

***K.Hareesh Kumar**, Computer Science and Engineering, JNTUK/ Sri Mittapalli Engineering College., Guntur, India, 9392947417;*
***D.Uma Devi**, Computer Science and Engineering, JNTUK/ Sri Mittapalli Engineering College..*

interest to talk about their hobbies and interests which are in common. After time is passed, they will find another person with same interests and hobbies. This community grows like a spider net. They keep on increasing the community and they will generally behave in a similar way with their interests and ideas.

The latest trend in the social network enables us to study intersection of behaviour or interests on a large scale. The interests include connecting a person, clicking an add, joining in a group, marking friends as buddies, becoming a fan of a celebrity, searching for friends, sending gifts for their special days etc. We have a given network with the behavioural outcome of the remaining users within the same network.

Consider, we assume the studied behaviour of one actor can be described with k class labels $\{l_1,l_2,..\ l_k\}$. Each label, l_k, can be either 0 or 1. We can consider that one user might join multiple groups of interests, so $l_i = 1$ denotes that the user subscribes the group i and else $l_i = 0$. A user might be interested in several topics or he can click several types of ads. Here we consider one special case as $l_k=1$, indicating that the studied behaviour can be described by a single label with 1 or 0. For example a user may become fan for a celebrity. If yes, it is indicated by 1, otherwise by 0.

Suppose there are k class labels $L=\{l_1,l_2,\ldots l_k\}$. Given network is G={V,E,L}where V is the vertex set, E is the edge set and Li is sub set of L , are the class labels of a vertex V_i €V and known values of L_i are some subsets of vertices V^S. We need to infer the values of L_i for some remaining vectors.

Then,

L_i values for remaining vertices = L_i values for total vertices – L_i values for some subset of vertices.

This type of calculation shows some problems in the social media and they are rather noisy and heterogeneous.

III. DIMENSIONS FOR THE INTERESTS

People generally want to communicate with the known people. They usually connect to family members, friends, colleages, classmates in the online internet. These relations may useful to determine the targeted behaviour, but not in all times. The relation type information will not available directly in the internet. The social dimensions extracted from the network should satisfy the following properties:

• Informative. The social dimensions should be indicative of affiliations between actors.

• Plural. The same social actor can get involved in multiple affiliations, thus appearing in different social dimensions.

• Continuous. The actors might have different degree of associations to one affiliation. Hence, a continuous value rather than discrete {0, 1} is more favorable.Generally the connection types in the social network are not homogeneous. A direct application of collective inference of label propagation would treat connections in a social network as if they were homogeneous. To recognize the heterogeneity present in the connections, a frame work is needed for intersectional interest learning.

The new framework is composed of two steps.

Social dimension extraction

Discriminative learning

In the first step, latent social dimensions are extracted based on network topology to capture the potential affiliations of users. These extracted social dimensions represent how each user is involved in different relations with people in the network. Consider the following example.

TABLE -1. RELATIONS AND USERS

User	Rel-1	Rel-2	…	Rel-k
1	0	0.7	.	0.2
2	0.3	0.8	.	1
…	…	….	…	…
…	…	….	…	…

The entries in the above table denote the degree of one user involving in an affiliation. The network is converted into features, typical classifiers such as support vector machine and logistic regression can be employed. The discriminative learning procedure determines which social dimension correlates with the targeted behaviour and then assigns proper suitable weights. The primary analysis is that users of the same relationship tend to connect with one another. For example we can consider that police department will communicate with the persons in the same department. Business people generally communicate with the same business people more frequently.

Hence to guess correctly and deduce the users, latent relationships we need to find out the group of people who frequently interact then at random. At the early stage, means at the instantiation of the frame work, FrameDim, a spectral variant of modularity maximization is adopted to extract social dimensions. Social dimensions extracted according to soft clustering are heavily weighted dimensions.

IV. EDGE-CENTRIC VIEW OF UNITS

We have FrameDim to work with social dimensions. But the scalability is limited and the desired outcome is not satisfactory. A network may or may not be strongly dense, but the dimensions are not sparse. Consider an example, we have an employee who connected with two different networks. He just connected with two communities with less number of connections. If the network is very large, he may be connected with lacks of people. Then extracting dimensions from the large network may be in big number and maintaining these many dimensions also a big problem to the persons who are taking care of this work. And the memory required to do this is also a problem. Maintainability is also will be a question in this situation. Hence, we need to minimize the measurable dimensions.

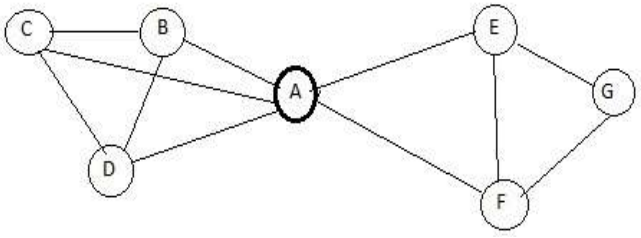


Fig. 1. An employee example

In the above example, A,B,C,..F are the users in a social network. A is a person who connected to two networks. So, we need to develop a table to consider the edge centric view of the network. Here, we can observe that a user may connect with any number of communities. Sometimes it is hard to find his relations with the communities. Node A is connected to both communities. And the nodes in the first and second communities are indirectly connected to the node A.

TABLE -2. EDGE CENTRIC VIEW OF THE RELATIONSHIPS WITH USERS

Users	Details of Edge Partition	
A	1	1
B	1	0
C	1	0
D	1	0
E	0	1
F	0	1
G	0	1

To extract sparse social dimensions, we partition edges rather than nodes into disjoint sets. The edges of those actors with multiple affiliations (e.g., actor 1 in the toy network) are separated into different clusters. Though the partition in the edge view is disjoint, the affiliations in the node-centric view can overlap. Each node can engage in multiple affiliations. In addition, the extracted social dimensions following edge partition are guaranteed to be sparse. This is because the number of one's affiliations is no more than that of her connections. Given a network with m edges and n nodes, if k social dimensions are extracted, then each node vi has no more than min(di, k) non-zero entries in her social dimensions, where di is the degree of node vi. We have the following theorem about the density of extracted social dimensions.

Theorem: Suppose k social dimensions are extracted from a network with m edges and n nodes. The density (proportion of nonzero entries) of the social dimensions based on edge partition is bounded by the following statement.

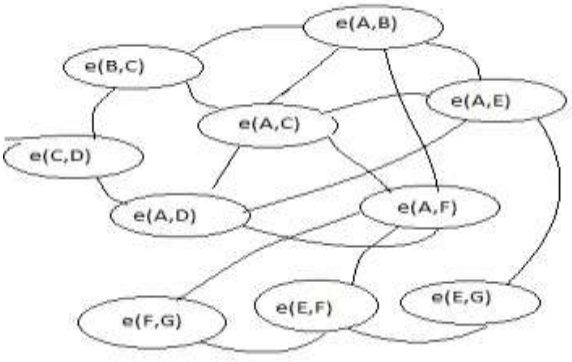
Density is less than or equal to the summation of the $\min(d_i,k)$, where i =1 to n.

We examine a YouTube network1 with 1+ million actors and verify the upper bound of the density. The YouTube network has 1, 128, 499 nodes and 2, 990, 443 edges. Suppose we want to extract 1, 000 dimensions from the network. Since 232 nodes in the network have a degree larger than 1000, the density is upper bounded by (5, 472, 909 + 232 × 1, 000)/(1, 128, 499×1, 000) = 0.51% following the above valid statement.

V. EDGE PARTITION BASED ON GRAPH WITH EDGE CONNECTIONS.

A graph represented communities in a network can be redrawn with respect to edge partition based on edge connections. An edge represents the connectivity of two vertices. Hence an edge based connections are very useful to estimate the relationships of a user with different communities.

This can be expressed in the form of a table as follows which shows the edge information of each node. This



able is very helpful to understand network connections of a particular user.

TABLE -1. EDGE PARTITION BASED ON GRAPH

Edge	A	B	C	D	E	F	G
e(A,B)	1	1	0	0	0	0	0
e(A,C)	1	0	1	0	0	0	0
e(A,D)	1	0	0	1	0	0	0
E(B,C)	0	1	1	0	0	0	0
. . .	.	.	.	.	.	.	.

Still, millions of edges are the norm in a large-scale network. Direct application of some existing k-means implementation cannot handle the problem. E.g., the k-means code provided in the Mat lab package requires
the computation of the similarity matrix between all pairs of data instances, which would exhaust the memory of normal PCs in seconds. Therefore, an implementation with online computation is preferred.

VI.CONCLUSIONS AND FUTUREWORK

Social media provides a virtual social networking environment. The classical IID assumption of data instances is not applicable. Relational learning based on collective inference has been proposed to capture the local dependency of labels between neighbouring nodes. However, it treats the connections within the network homogeneously. In reality, the connections within the same network are often multidimensional. To capture different affiliations among actors in a network, we propose to extract latent social dimensions via modularity maximization. Based on the extracted social features, a discriminative classifier like SVM can be constructed to determine which dimensions are informative for classification. Extensive experiments on social media data demonstrated that our proposed social dimension approach outperforms alternative relational learning methods, especially when the labelled data are few. It is noticed that some relational learning models perform poorly in social media data. This is partly due to the multi-dimensionality of connections and high irregularity of

human interactions as presented in social media. Our approach, by differentiating the connections among social actors, achieves effective learning.

VII. REFERENCES

[1] L. Tang and H. Liu, "Toward predicting collective behavior via social dimension extraction," *IEEE Intelligent Systems*, vol. 25, pp. 19–25, 2010.
[2] ——, "Relational learning via latent social dimensions," in *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009, pp. 817–826.
[3] M. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 74, no. 3, 2006. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.74.036104
[4] L. Tang and H. Liu, "Scalable learning of collective behavior based on sparse social dimensions," in *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*.
New York, NY, USA: ACM, 2009, pp. 1107–1116.
[5] P. Singla and M. Richardson, "Yes, there is a correlation: - from social networks to personal behavior on the web," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 655–664.
[6] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.
[7] A. T. Fiore and J. S. Donath, "Homophily in online dating: when do you like someone like yourself?" in *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM, 2005, pp. 1371–1374.
[8] H. W. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, "Homophily
in the digital world: A LiveJournal case study," *IEEE Internet Computing*, vol. 14, pp. 15–23, 2010.
[9] S. A. Macskassy and F. Provost, "Classification in networked data: A toolkit and a univariate case study," *J. Mach. Learn. Res.*, vol. 8, pp. 935–983, 2007.
[10] X. Zhu, "Semi-supervised learning literature survey," 2006. [Online]. Available: http://pages.cs.wisc.edu/ ~jerryzhu/ pub/ssl survey 12 9 2006.pdf
[11] L. Getoor and B. Taskar, Eds., *Introduction to Statistical Relational
Learning*. The MIT Press, 2007.
[12] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003.
[13] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *SDM*, 2005.
[14] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–352, 2005.
[15] F. Harary and R. Norman, "Some properties of line digraphs," *Rendiconti del Circolo Matematico di Palermo*, vol. 9, no. 2, pp. 161–168, 1960.
[16] T. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," *Physical Review E*, vol. 80, no. 1, p. 16105, 2009.
[17] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multi-scale complexity in networks," 2009. [Online]. Available: http://www.citebase.org/abstract?id=oai: arXiv.org:0903.3178
[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin,
"LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
[19] J. Hopcroft and R. Tarjan, "Algorithm 447: efficient algorithms for graph manipulation," *Commun. ACM*, vol. 16, no. 6, pp. 372–378, 1973.
[20] J. Neville and D. Jensen, "Leveraging relational autocorrelation with latent group models," in *MRDM '05: Proceedings of the 4th international workshop on Multi-relational mining*. New York, NY, USA: ACM, 2005, pp. 49–55.
[21] R.-E. Fan and C.-J. Lin, "A study on threshold selection for multi-label classication," 2007.
[22] L. Tang, S. Rajan, and V. K. Narayanan, "Large scale multilabel
classification via metalabeler," in *WWW '09: Proceedings of the 18th international conference on World wide web*. New York, NY, USA: ACM, 2009, pp. 211–220.
[23] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *AAAI*, 2006.
[24] F. Sebastiani, "Machine learning in automated text categorization,"
ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, 2002.
[25] S. A. Macskassy and F. Provost, "A simple relational classifier,"
in *Proceedings of the Multi-Relational Data Mining Workshop (MRDM) at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.
[26] Z. Xu, V. Tresp, S. Yu, and K. Yu, "Nonparametric relational learning for social network analysis," in *KDD'2008 Workshop on Social Network Mining and Analysis*, 2008.
[27] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
[28] K. Yu, S. Yu, and V. Tresp, "Soft clustering on graphs," in *NIPS*,
2005.
[29] E. Airodi, D. Blei, S. Fienberg, and E. P. Xing, "Mixed membership
stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, 2008.
[30] S. Fortunato, "Community detection in graphs," *Physics Reports*,
vol. 486, no. 3-5, pp. 75–174, 2010.
[31] G. Palla, I. Der´enyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.
[32] H. Shen, X. Cheng, K. Cai, and M. Hu, "Detect overlapping and hierarchical community structure in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706–1712, 2009.
[33] S. Gregory, "An algorithm to find overlapping community structure in networks," in *PKDD*, 2007, pp. 91–102. [Online]. Available: http://www.cs.bris.ac.uk/Publications/ pub master.jsp?id=2000712
[34] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, p. 026113, 2004. [Online]. Available: http://www.citebase. org/abstract?id=oai:arXiv.org:cond-mat/0308217
[35] J. Bentley, "Multidimensional binary search trees used for associative searching," *Comm. ACM*, vol. 18, pp. 509–175, 1975.
[36] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 881–892, 2002.
[37] M. Sato and S. Ishii, "On-line EM algorithm for the normalized gaussian network," *Neural Computation*, 1999.
[38] P. Bradley, U. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases," in *ACM KDD Conference*, 1998.
[39] R. Jin, A. Goswami, and G. Agrawal, "Fast and exact outof-core and distributed k-means clustering," *Knowl. Inf. Syst.*,
vol. 10, no. 1, pp. 17–40, 2006.