Social Networking And Its Collective Behavior

G.Anitha,G.Bindu Madhavi
CSE dept.,Anurag Group Of Institutions,Ghatkesar,501301

Abstract

This paper is to understand how people behave in a public press environment. Collective behaviour refers to how individuals behave when they are exposed in a social network. This collective behaviour gives the opportunity to predict online behaviours of users in a network, given the behaviour information of some actors in the network. However, the systems in the network are normally of heavy size, involving many. The range of these systems entails scalable learning of models for combined actions forecast. we propose an edge-centric clustering scheme to extract rare public measurements. With sparse social dimensions, the proposed approach can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other non-scalable methods.we develop charts user/group,user/month in this social networks.

Keywords— ***Collective Behaviour, Social Network, Social-Dimensions, Scalable Learning***.

1. Introduction

This study of collective behavior is to understand how individuals behave in a social networking environment. Oceans of data generated by social media like Facebook, Twitter, Flicker, and YouTube present opportunities and challenges to study collective behavior on a large scale. In this work, we aim to learn to predict collective behavior in social media. In particular, given information about some individuals, how can we infer the behavior of unobserved individuals in the same network? A social-dimension-based approach has been shown effective in addressing the heterogeneity of connections presented in social media. However, the networks in social media are normally of colossal size, involving hundreds of thousands of actors. The scale of these networks entails scalable learning of models for collective behavior prediction.

To address the scalability issue, we propose an edge-centric clustering scheme to extract sparse social dimensions. With sparse social dimensions, the proposed approach can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other non-scalable methods.Social media facilitate people of all walks of life to connect to each other.

In the initial study, modularity maximization is exploited to extract social dimensions.With huge number of actors, the dimensions cannot even be held in memory.In this work, we propose an effective edge-centric approach to extract sparse social dimensions. The advancement in computing and communication technologies enables people to get together and share information in innovative ways. Social networking sites (a recent phenomenon) empower people of different ages and backgrounds with new forms of collaboration, communication, and collective intelligence.

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. We prove that with our proposed approach, sparsity of social dimensions is guaranteed.

Fudan Univ
China
powered by TouchGraph

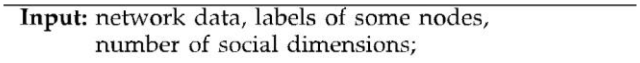
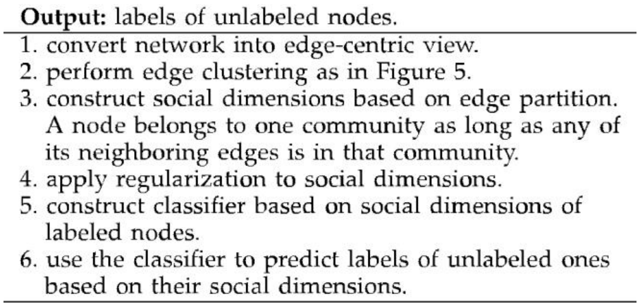
Figure 1: Contacts of One User in Facebook

II. Material and Methodology

Collective Behavior

Collective behavior refers to the behaviors of individuals in a social networking environment, but it is not simply the aggregation of individual behaviors. In a connected environment, individuals' behaviors tend to be interde- pendent, influenced by the behavior of friends. This naturally leads to behavior correlation between connected users . Take marketing as an example: if our friends buy something, there is a better than average chance that we will buy it, too. This behavior correlation can also be explained by homophily . Homophily is a term coined in the 1950s to explain our tendency to link with one another in ways that confirm, rather than test, our core beliefs. Essentially, we are more likely to connect to others who share certain similarities with us. This phenomenon has been observed not only in the many processes of a physical world, but also in online systems .

Homophily results in behavior correlations between connected friends. In other words, friends in a social network tend to behave similarly. The recent boom of social media enables us to study collective behavior on a large scale.

```
Input: network data, labels of some nodes,
       number of social dimensions;
Output: labels of unlabeled nodes.
1. convert network into edge-centric view.
2. perform edge clustering as in Figure 5.
3. construct social dimensions based on edge partition.
   A node belongs to one community as long as any of
   its neighboring edges is in that community.
4. apply regularization to social dimensions.
5. construct classifier based on social dimensions of
   labeled nodes.
6. use the classifier to predict labels of unlabeled ones
   based on their social dimensions.
```

Figure. 2. Algorithm for learning of collective behavior.

Social Dimensions

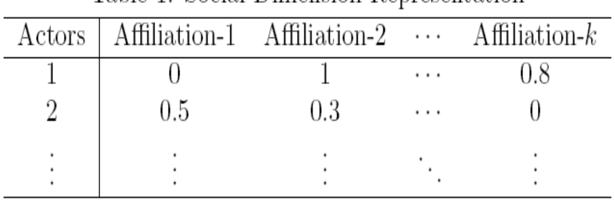
Connections in social media are not homogeneous. People can connect to their family, colleagues, college classmates, or buddies met online. Some relations are helpful in determining a targeted behavior (category) while others are not. This relation-type information, however, is often not readily available in social media. A direct application of collective inference or label propagation would treat connections in a social network as if they were homogeneous. To address the heterogeneity present in connections, a framework (SocioDim) has been proposed for collective behavior learning.

Sparse Social Dimensions

We implement an edge centric view basing on information available of a user by using following methodologies and then regularize it to observe efficient results.

Table 1 shows how an affiliation is represent

Table 1: Social Dimension Representation



Actors	Affiliation-1	Affiliation-2	···	Affiliation-k
1	0	1	···	0.8
2	0.5	0.3	···	0
⋮	⋮	⋮	⋱	⋮

Communities In An Edge Centric View

Though SocioDim with soft clustering for social dimension extraction demonstrated promising results, its scalability is limited. A network may be sparse (i.e., the density of connectivity is very low), whereas the extracted social dimensions are not sparse.

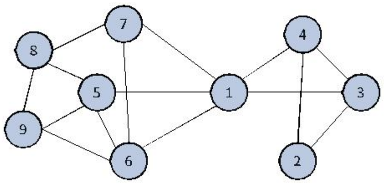
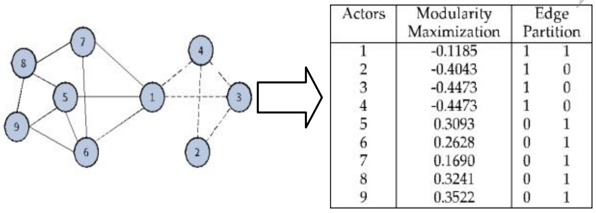


Fig 3. Toy example

Let's look at the toy network with two communities in Figure 3.Its social dimensions following modularity maximization are shown in Table 1 . Clearly, none of the entries is zero. When a network expands into millions of actors, a reasonably large number of social dimensions need to be extracted. The corresponding memory requirement hinders both the extraction of social dimensions and the subsequent discriminative learning. Hence, it is imperative to develop some other approach so that the extracted social dimensions are sparse.



Actors	Modularity Maximization	Edge Partition	
1	-0.1185	1	1
2	-0.4043	1	0
3	-0.4473	1	0
4	-0.4473	1	0
5	0.3093	0	1
6	0.2628	0	1
7	0.1690	0	1
8	0.3241	0	1
9	0.3522	0	1

Fig 4: edge clusters representing social dimensions

The disjoint edge clusters in Figure 4 can be converted into the representation of social dimensions as shown in the last two columns in above table , where an entry is 1(0) if an actor is (not) involved in that corresponding social dimension.

To extract sparse social dimensions, we partition edges rather than nodes into disjoint sets. The edges of those actors with multiple affiliations are separated into different clusters. 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. We have a theorem that finds the density of extracted social dimension

Algorithm for Learning of Collective Behavior

Input: network data, labels of some nodes, number of social dimensions;

Output: labels of unlabeled nodes.

1. Convert network into edge-centric view.

2. Perform edge clustering .

3. Construct social dimensions based on edge partition node belongs to one community as long as any of its neighboring edges is in that community.

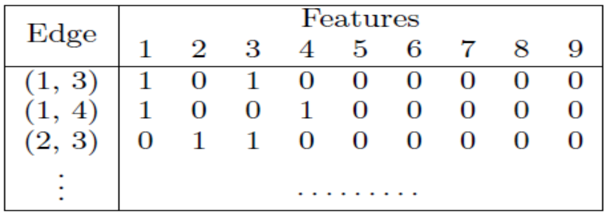
4. Apply regularization to social dimensions.

5. Construct classifier based on social dimensions of labeled nodes.

6. Use the classifier to predict labels of unlabeled ones based on their social dimensions.

Edge Partition Via Clustering Edge Instances:

In order to partition edges into disjoint sets, we treat edges as data instances with their terminal nodes as features. For instance, we can treat each edge in the toy network in Figure 3 as one instance, and the nodes that define edges as features. This results in a typical feature-based data format as in Table 2. Then, a typical clustering algorithm like k-means clustering can be applied to find disjoint partitions.

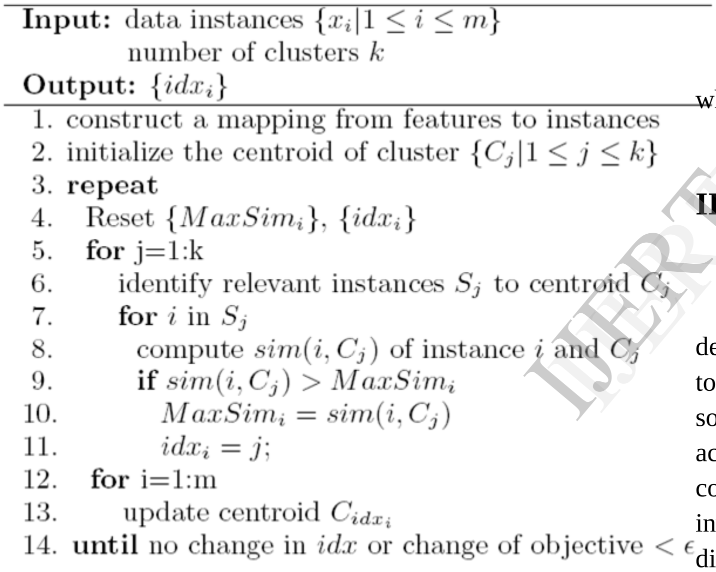


Edge	Features								
	1	2	3	4	5	6	7	8	9
(1, 3)	1	0	1	0	0	0	0	0	0
(1, 4)	1	0	0	1	0	0	0	0	0
(2, 3)	0	1	1	0	0	0	0	0	0
⋮	....								

TABLE 2: Edge Instances of the Toy Network in Fig. 3

The algorithm k-means variant is used to perform edge clustering

The algorithm helps to cluster the edges which are in the network.



```
Input: data instances $\{x_i | 1 \le i \le m\}$
       number of clusters $k$
Output: $\{idx_i\}$
construct a mapping from features to instances
initialize the centroid of cluster $\{C_j | 1 \le j \le k\}$
repeat
  Reset $\{MaxSim_i\}$, $\{idx_i\}$
  for j=1:k
    identify relevant instances $S_j$ to centroid $C_j$
    for $i$ in $S_j$
      compute $sim(i, C_j)$ of instance $i$ and $C_j$
      if $sim(i, C_j) > MaxSim_i$
        $MaxSim_i = sim(i, C_j)$
        $idx_i = j$;
  for i=1:m
    update centroid $C_{idx_i}$
until no change in $idx$ or change of objective $< \epsilon$
```

Fig 5 : k-means Variant

One concern with this scheme is that the total number of edges might be too huge. Owing to the power law distribution of node degrees presented in social networks, the total number of edges is normally linear, rather than square, with respect to the number of nodes in the network.That is, m = O(n) as stated in the following theorem.The total number of edges is usually linear, rather than quadratic, with respect to the number of nodes in the network with a power law distribution. Power law distribution in large-scale social networks

$$p(x) = Cx^{-\alpha}, \quad x \ge 1$$

Where α is the exponent of the power law distribution.

Theorem 1:. The total number of edges is usually linear, rather than quadratic, with respect to the number of nodes in the network with a power law distribution. In particular, the expected number of edges is given as

$$E[m] = \frac{n}{2}\frac{\alpha - 1}{\alpha - 2},$$

where α is the exponent of the power law distribution.

III. Results and Discussion

It is well known that actors in a network demonstrate correlated behaviors. In this work, we aim to predict the outcome of collective behavior given a social network and the behavioral information of some actors. In particular, we explore scalable learning of collective behavior when millions of actors are involved in the network. Our approach follows a social-dimension-based learning framework. Social dimensions are extracted to represent the potential affilia-tions of actors before discriminative learning occurs. As existing approaches to extract social dimensions suffer from scalability, it is imperative to address the scalability issue. We propose an edge-centric clustering scheme to extract social dimensions and a scalable k-means variant to handle edge clustering. Essentially, each edge is treated as one data instance, and the connected nodes are the corresponding features. Then, the proposed k-means clustering algorithm can be applied to partition the edges into disjoint sets, with each set representing one possible affiliation. With this edge-centric view, we show that the extracted social dimensions are guaranteed to be sparse. This model, based on the sparse social dimensions, shows comparable prediction perfor-mance with earlier social dimension approaches. An incomparable advantage of our model is that it easily scales to handle networks with millions of actors while the earlier models fail. This scalable approach offers a viable solution to effective learning of online collective behavior on a large scale.

In social media, multiple modes of actors can be involved in the same network, resulting in a multimode network . For instance, in YouTube, users, videos, tags, and comments are intertwined with each other in coexistence. Extending the edge-centric clustering scheme to address this object heterogeneity can be a promising future direction. Since the proposed EdgeCluster model is sensitive to the number of social dimensions as shown in the experiment, further research is needed to determine a suitable dimensionality automatically. It is also interesting to mine other behavioral features (e.g., user activities and temporal-spatial informa-tion) from social media, and integrate them with social networking information to improve prediction performance.

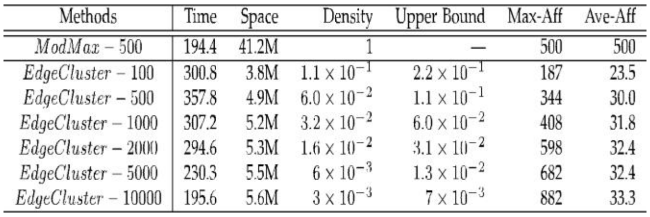
We also study how the performance varies with dimensionality. Finally, concrete examples of extracted social dimensions are given.

A. Prediction Performance

The prediction performance on all data is shown in Tables 3. The entries in bold face denote the best performance in each column. Obviously, EdgeCluster is the winner most of the time. Edge-centric clustering shows comparable performance to modularity maximization on BlogCatalog network, yet it outperforms ModMax on Flickr. ModMax on YouTube is not applicable due to the scalability constraint. Clearly, with sparse social dimensions, we are able to achieve comparable performance as that of dense social dimensions. But the benefit in terms of scalability will be tremendous as discussed in the next section.

The NodeCluster scheme forces each actor to be involved in only one affiliation, yielding inferior performance compared with EdgeCluster. BiComponents, similar to EdgeCluster, also separates edges into disjoint sets, which in turn deliver a sparse representation of social dimensions.

However, BiComponents yields a poor performance. This is because BiComponents outputs highly imbalanced commu-nities. In short, BiComponents is very efficient and scalable. However, it fails to extract informative social dimensions for classification.We note that the prediction performance on the studied social media data is around 20-30 percent for F1 measure. This is partly due to the large number of distinctive labels in the data.



Methods	Time	Space	Density	Upper Bound	Max-Aff	Ave-Aff
$ModMax - 500$	194.4	41.2M	1	—	500	500
$EdgeCluster - 100$	300.8	3.8M	1.1×10^{-1}	2.2×10^{-1}	187	23.5
$EdgeCluster - 500$	357.8	4.9M	6.0×10^{-2}	1.1×10^{-1}	344	30.0
$EdgeCluster - 1000$	307.2	5.2M	3.2×10^{-2}	6.0×10^{-2}	408	31.8
$EdgeCluster - 2000$	294.6	5.3M	1.6×10^{-2}	3.1×10^{-2}	598	32.4
$EdgeCluster - 5000$	230.3	5.5M	6×10^{-3}	1.3×10^{-2}	682	32.4
$EdgeCluster - 10000$	195.6	5.6M	3×10^{-3}	7×10^{-3}	882	33.3

TABLE 3: Sparsity Comparison on BlogCatalog Data with 10,312 Nodes

A. Scalability Study

The social dimensions constructed according to edge-centric clustering are guar-anteed to be sparse because the density is upper bounded by a small value. Here, we examine how sparse the social dimensions are in practice. We also study how the computation time (with a Core2Duo E8400 CPU and 4 GB memory) varies with the number of edge clusters

B. Chart Generation for User/Group:

Two data sets reports are used to examine our proposed model for collective behavior learning. The first data set is acquired from user interest, the second from concerning behaviour we study whether or not a user visits a group of interest, then generates chart based on the user visit group in the month. The below chart contains comm. unities Vs users

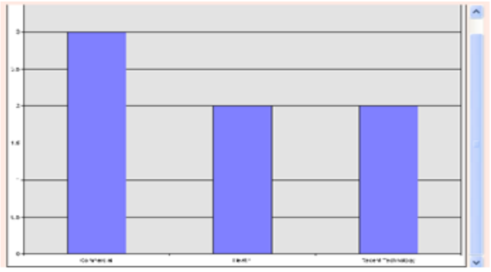


Fig 6:chart generated for user/group

C. **Chart Generation for Group/Month:**

Two data sets reported in are used to examine our proposed model for collective behavior learning. The first data set is acquired from user interest, the second from concerning behavior; we study whether or not a user visits a group of interest. Then generates chart the based on the user visit group in the month.

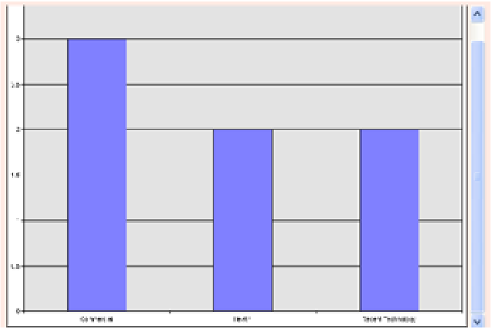


Fig 7: chart generated for group/month

IV. Conclusion

Our approach follows a social-dimension based learning framework. Social dimensions are extracted to represent the potential affiliations of actors before discriminative learning occurs. As existing approaches to extract social dimensions suffer from scalability, it is imperative to address the scalability issue. We propose an edge-centric clustering scheme to extract social dimensions and a scalable k-means variant to handle edge clustering. This model, based on the sparse social dimensions, shows comparable prediction performance with earlier social dimension approaches.

It is also interesting to mine other behavioral features (e.g., user activities and temporal spatial information) from social media, and integrate them with social networking information to improve prediction performance.

It is well known that actors in a network demonstrate correlated behaviors. In this work, we aim to predict the outcome of collective behavior given a social network and the behavioral information of some actors. In particular, we explore scalable learning of collective behavior when millions of actors are involved in the network. Our approach follows a social-dimension based learning framework. Social dimensions are extracted to represent the potential affiliations of actors before discriminative learning occurs.

As existing approaches to extract social dimensions suffer from scalability, it is imperative to address the scalability issue. We propose an edge-centric clustering scheme to extract social dimensions and a scalable k-means variant to handle edge clustering. Essentially, each edge is treated as one data instance, and the connected nodes are the corresponding features. Then, the proposed k-means clustering algorithm can be applied to partition the edges into disjoint sets, with each set representing one possible affiliation. With this edge-centric view, we show that the extracted social dimensions are guaranteed to be sparse.

This model, based on the sparse social dimensions, shows comparable prediction performance with earlier social dimension approaches. An incomparable advantage of our model is that it easily scales to handle networks with millions of actors while the earlier models fail. This scalable approach offers a viable solution to effective learning of online collective behavior on a large scale. In social media, multiple modes of actors can be involved in the same network, resulting in a multimode network. For instance, in YouTube, users, videos, tags, and comments are intertwined with each other in co-existence. Extending the edge-centric clustering scheme to address this object heterogeneity can be a promising future direction. Since the proposed *Edge Cluster* model is sensitive to the number of social dimensions as shown in the experiment, further research is needed to determine a suitable dimensionality automatically. It is also interesting to mine other behavioral features (e.g., user activities and temporal spatial information) from social media, and integrate them with social networking information to improve prediction performance.

V.Acknowledgement

It is my privilege and pleasure to express my profound sense of respect, gratitude and indebtedness to my guide Mrs. G.Bindu Madhavi, Asst Professor, Department of Computer Science and Engineering ANURAG GROUP OF INSTITUTIONS formerly CVSR College of Engineering

References

[1] M.Mcpherson, L.smith-Lovin, and J.M. Cook, "Birds of a feather: Homophily in social network, "Annual review of Sociology, vol.27, pp.415-444, 2001.
[2] H.W.Lauw,J.C.Shafer, R.Agrawal, and A.Ntoulas "Homophile in the digital world: A live Journal case study,"IEEE Internet Computing, vol.14 ,pp. 15-23, 2010
[3] S.A. Macskassy and F.Provost,"Classification innetworked data:A tool kit and a univariate case study,"J.Mach.Learn Res., Vol.8,pp. 935-983,2007.
[4] X.Zhu ," Semi - supervised learning literature survey,"2006[Online].Available:http://Pages.cs.wisc.edu/jerryzhu/pub/ssl survey 12 9 2006.pdf
[5] X.zhu,Z.Ghahramani and J.lafferty,"semi- supervised learning using Gaussian fields and harmonic functions, in ICML,2003.
[6] J. Neville and D. Jensen, "Leveraging Relational Autocorrelation with Latent Group Models," MRDM '05: Proc. Fourth Int'l Workshop Multi-Relational Mining, pp. 49-55, 2005.
[7] R.-E. Fan and C.-J. Lin, "A Study on Threshold Selection for Multi-Label Classification," technical report, 2007.
[8] L. Tang, S. Rajan, and V.K. Narayanan, "Large Scale Multi-Label Classification via Metalabeler," WWW '09: Proc. 18th Int'l Conf. World Wide Web, pp. 211-220, 2009.
[9] Y. Liu, R. Jin, and L. Yang, "Semi-Supervised Multi-Label Learning by Constrained Non-Negative Matrix Factorization," Proc. Nat'l Conf. Artificial Intelligence (AAAI), 2006.
[10] F. Sebastiani, "Machine Learning in Automated Text Categoriza-tion," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
[11]S.A. Macskassy and F. Provost, "A Simple Relational Classifier,"

Sites Referred:

http://www.sourcefordgde.com

http://www.networkcomputing.com/

http://www.ieee.org

http://www.computer.org/publications/dlib

http://www.ceur-ws.org/Vol-90/

http://www.microsoft.com/isapi/redir.dll?prd=ie&pver=6&ar=msnhome