Discovering Distinct Behavior using Sparse SocioDim

Mohammed Imran[#1], Dr. S Murali Krishna[#2]
#1 Computer Science and Engineering, M.Tech, Madanapalle Institute of Technology Science(MITS), Madanapalle – 517325, Chittoor District, Andhra Pradesh, India,
#2 Department of Computer Science and Engineering, Professor and HOD, Madanapalle Institute of Technology Science(MITS), Madanapalle – 517325, Chittoor District, Andhra Pradesh, India,

ABSTRACT

Collective behavior refers to the behaviors of individuals in a social networking environment, but it is not simply the aggregation of individual behaviors. This study of collective behavior is to understand how individuals behave in a social networking environment. 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.

In social media, multiple modes of actors can be involved in the same network, resulting in a multimode network. Since the proposed Edge Cluster model is sensitive to the number of social dimensions, 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.

Key words: Collective behavior, scalable learning, social dimensions, affiliations, edge clustering, k-means variant.

Corresponding Author: Mohammed Imran

INTRODUCTION

Oceans of data generated by social media like Facebook, Twitter, Flickr, and YouTube presents 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.

In this work, we study how networks in social media can help predict some human behaviors and individual preferences. In particular, given the behavior of some individuals in a network, how can we infer the behavior of other individuals in the same social network? This study can help better understand behavioral 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 as shown in Fig 1.

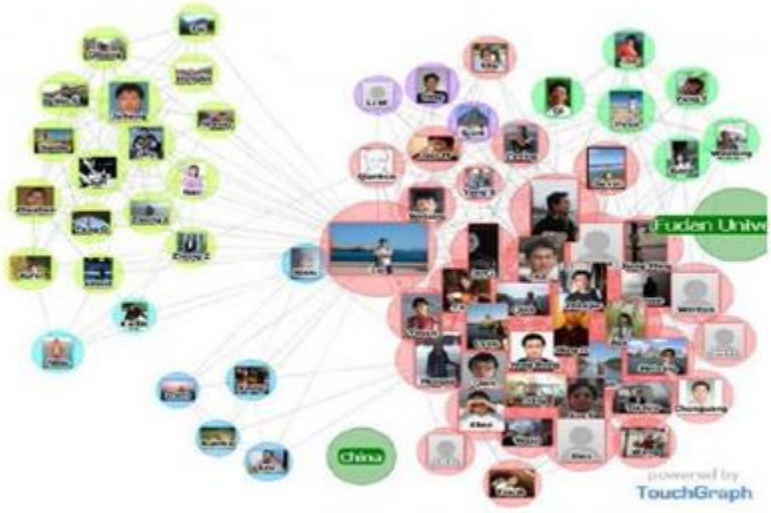

Fig 1: Contacts of one user in a Social Network

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. This study of collective behavior is to understand how individuals behave in a social networking environment. 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.

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. Figure 2 shows how an affiliation is represented.

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

Figure 2: Social Dimensions Representation

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.

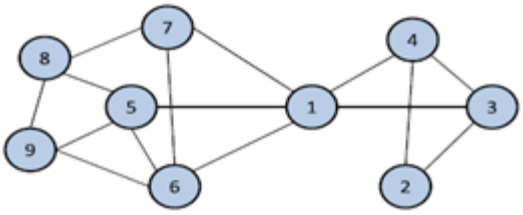

Figure 3: Toy Example

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-Centric Clustering	
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

Table 1: Social Dimensions of the Toy Example

An actor is considered associated with one affiliation if one of his connections is assigned to that affiliation. For instance, the two communities in Figure 3 can be represented by two edge sets in Figure 4 as shown below where the dashed edges represent one affiliation, and the remaining edges denote the second affiliation.

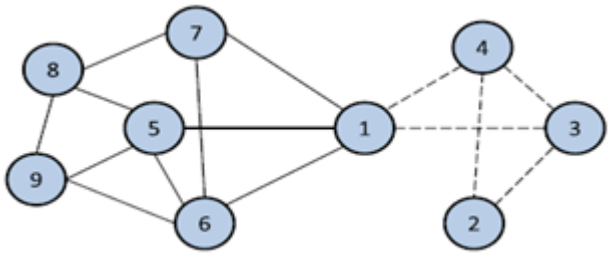

Figure 4: Edge Clusters

The disjoint edge clusters in Figure 3 can be converted into the representation of social dimensions as shown in the last two columns in Table 1, 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

$$density \leq \frac{\sum_{i=1}^{n} \min(d_i, k)}{nk} = \frac{\sum_{\{i|d_i \leq k\}} d_i + \sum_{\{i|d_i > k\}} k}{nk}$$

Where k is number of social dimensions to be extracted, m is number of edges, n is number of nodes and d_i is the degree of node. Moreover, for many real-world networks whose node degree follows a power law distribution, the upper bound in above equation can be approximated as follows:

$$\frac{\alpha - 1}{\alpha - 2}\frac{1}{k} - \left(\frac{\alpha - 1}{\alpha - 2} - 1\right) k^{-\alpha+1}$$

Where $\alpha > 2$ is the exponent of the power law distribution.

EDGE PARTITION VIA LINE GRAPH

In order to partition edges into disjoint sets, one way is to look at the "dual" view of a network, i.e., the line graph. In a line graph L(G), each node corresponds to an edge in the original network G, and edges in the line graph represent the adjacency between two edges in the original graph. The line graph of the toy example is shown in Figure 5.

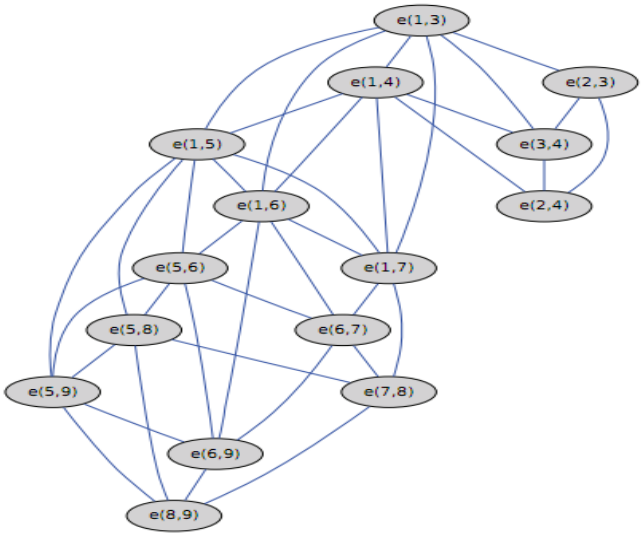

Figure 5: The line graph of the toy example

For instance, e(1, 3) and e(2, 3) are connected in the line graph as they share one terminal node 3. Each node in the line graph corresponds to an edge in the original graph. To increase many more edges in the graph we use following equation

$$N = m, \quad M \geq m\left(\frac{2m}{n} - 1\right)$$

Where n denotes the number of nodes, m denotes number of connections in a network. N & M denotes number of nodes and connections in its line graph respectively.

With respect to power law degree distribution theorem, the lower bound in above equation can never be achieved. The number of connections in a line graph grows without a constant bound with respect to the size of a given network.

$$E[2M/n] \begin{cases} diverges & \text{if } 2 < \alpha <= 3 \\ = \frac{\alpha-1}{(\alpha-3)(\alpha-2)} & \text{if } \alpha > 3 \end{cases}$$

Where α denote the exponent of a power law degree distribution, n the size of the network, and M the number of connections in its corresponding line graph.

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 2 as one instance, and the nodes that define edges as features. This results in a typical feature-based data format as in Table 3. 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
e(1, 3)	1	0	1	0	0	0	0	0	0
e(1, 4)	1	0	0	1	0	0	0	0	0
e(2, 3)	0	1	1	0	0	0	0	0	0
.	.....								
.	.....								
.	....								

Table 3: Edge Centric View

K-Means clustering can be implemented abiding following algorithm

Input: data instances $\{x_i | 1 \leq i \leq m\}$, number of clusters k
Output: $\{idx_i\}$
Procedure:

1. construct a mapping from features to instances
2. initialize the centroid of cluster $\{C_j | 1 \leq j \leq k\}$
3. repeat
4. Reset $\{MaxS_i m_i\}$, $\{idx_i\}$
5. for j=1:k
6. identify relevant instances S_j to centroid C_j
7. for i in S_j
8. compute sim(i,C_j) of instance i and C_j
9. if sim(i,C_j)>MaxSimi
10. MaxSimi = sim(i,C_j)
11. idx_i = j;
12. for i=1:m

13. update centroid $Cidx_i$
14. until change of objective value < ε

Algorithm of Scalable *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. 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.

The detailed algorithm is summarized.

Input: network data, labels of some nodes, number of social dimensions;
Output: labels of unlabeled nodes.
Procedure:

1. Convert network into edge-centric view.
2. Perform edge clustering as in above algorithm.
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.

Algorithm for Learning of Collective Behavior

REGULARIZATION ON COMMUNITIES

The extracted social dimensions are treated as features of nodes. Conventional supervised learning can be conducted. In order to handle large-scale data with high dimensionality and vast numbers of instances, we adopt a linear SVM, which can be finished in linear time generally; the larger a community is, the weaker the connections within the community are. Hence, we would like to build an SVM relying more on communities of smaller sizes by modifying the typical SVM objective function.

CONCLUSION

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. 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 information) from social media, and integrate them with social networking information to improve prediction performance.

REFERENCE

[1] Lei Tang, Xufei Wang, and Huan Liu, Scalable Learning of Collective Behavior, 2012.

[2] 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.

[3] M. Newman, "Power laws, Pareto distributions and Zipf's law," Contemporary physics, vol. 46, no. 5, pp. 323–352, 2005.

[4] T. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," Physical Review E, vol. 80, no. 1, p.16105, 2009

[5] 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

[6] L. Tang and H. Liu, "Toward predicting collective behavior via social dimension extraction," IEEE Intelligent Systems, vol. 25,pp. 19–25, 2010

[7] M. Newman, "Finding community structure in networks using the eigenvectors of atrices," Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), vol. 74, no. 3, 2006.

[7] X. Zhu, "Semi-supervised learning literature survey," 2006.

[8] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, 2002.

[9] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, no. 3-5, pp. 75-174, 2010.

[10] J. Bentley, "Multidimensional binary search trees used for associative searching," Comm. ACM, vol. 18, pp. 509–175, 1975.

[11] 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.

[12] P. Bradley, U. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases," in ACM KDD Conference, 1998.