Review Paper on Scalable Learning of Collective Behavior

Prof. Prafulla Mehar[1], Prof. Vijaya Sawarkar[2]
Department of Information technology[1, 2], DMIETR[1]SSPACE[2]
prafulla.mehar@gmail.com[1] , vijaya_sawarkar01@rediffmail.com[2]

Abstract- The study of collective behavior is to understand how individuals behave in a social network environment. Oceans of data generated by social media like Facebook, Twitter, Flickr and YouTube present opportunities and challenges to studying collective behavior in 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 is adopted to address the heterogeneity of connections presented in social media. However, the networks in social media are normally of colossal size, involving hundreds of thousands or even millions of actors. The scale of 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 social dimension based approach can efficiently handle networks of millions of actors while demonstrating comparable prediction performance as other non-scalable methods.

Index Terms- Facebook, Twitter, Flickr.

1. Introduction

Social media such as Facebook, MySpace, Twitter, Blog-Digg, YouTube and Flickr, facilitate people of allwalks of life to express their thoughts, voice their opinions, and connect to each other anytime and anywhere. For instance, popular content-sharing sites like Del.icio.us, Flickr, and YouTube allow users to upload, tag and comment different types of contents (bookmarks, photos, videos). Users registered at these sites can also become friends, a fan orfollower of others. The prolific and expanded use of social media has turn online interactions into a vital part of human experience. The election of Barack Obama as the President of United States was partially attributed to his smart Internet strategy and access to millions of younger voters through the new social media, such as Facebook. As reported in the New York Times, in response to recent Israeli air strikes in Gaza, young Egyptians mobilized not only in the streets of Cairo, but also through the pages of Facebook.

Owning to social media, rich human interaction information is available. It enables the study of collective behavior in a much larger scale, involving hundreds of thousands or millions of actors. It is gaining increasing attentions across various disciplines including sociology, behavioral science, anthropology, epidemics, economics and marketing business, to name a few. In this work, we study how networks in social media can help predict some sorts of human behavior and individual preference. In particular, given the observation of some individuals' behavior or preference in a network, how to infer the behavior or preference of other individuals in the same social network? This can help understand the behavior patterns presented in social media, as well as other tasks like social networking advertising and recommendation.

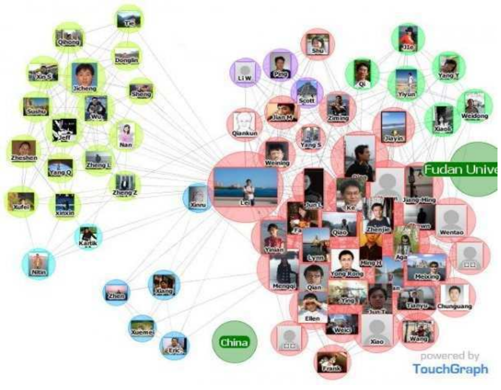
Typically in social media, the connections of the same network are not homogeneous. Different relations are intertwined with different connections. For example, one user can connect to his friends, family, college classmates or colleagues. However, this relation type information is not readily available in reality. This heterogeneity of connections limits the effectiveness of a commonly used technique collective inference for network classification. Recently, a framework based on social dimensions [18] is proposed to address this heterogeneity. This framework suggests extracting social dimensions based on network connectivity to capture the potential affiliations of actors. Based on the extracted dimensions, traditional data mining can be accomplished. In the initial study, modularity maximization [15] is exploited to extract social dimensions. The superiority of this framework over other representative relational learning methods is empirically verified on some social media data [18].

However, the instantiation of the framework with modularity maximization for social dimension extraction is not scalable enough to handle networks of colossal size, as it involves a large-scale eigenvector problem to solve and the corresponding extracted social dimensions are dense. In social media, millions of actors in a network are the norm. With this huge number of actors, the dimensions cannot even be held in memory, causing serious problem about the scalability. To alleviate the problem, social dimensions of sparse representation are preferred. In this work, we propose an effective edge-centric approach to extract sparse social dimensions. We prove that the sparsity of the social dimensions following our proposed approach is guaranteed. Extensive experiments are conducted using social media data. The framework based on sparse social dimensions, without sacrificing the prediction performance, is capable of handling real-world networks of millions of actors in an efficient way.

2. Collective Behavior Learning

The recent boom of social media enables the study of collective behavior in a large scale. Here, behavior can include a broad range of actions: join a group, connect to a person, click on some ad, become interested in certain topics, date with people of certain type, etc. When people are exposed in a social network environment, their behaviors are not independent [6, 22]. That is, their behaviors can be influenced by the behaviors of their friends. This naturally leads to behavior correlation between connected users. This behavior correlation can also be explained by homophily. Homophily [12] is a term coined in 1950s to explain our tendency to link up with one another in ways that confirm rather than test our core beliefs. Essentially, we are more likely to connect to others sharing certain similarity with us. This phenomenon has been observed not only in the real world, but also in online systems [4]. Homophily leads to behavior correlation between connected friends. In other words, friends in a social network tend to behave similarly. Take marketing as an example, if our friends buy something, there's better-than-average chance we'll buy it too.

In this work, we attemt to utilize the behavior correlation presented in a social network to predict the collective behavior in social media. Given a network with behavior information of some actors, how can we infer the behavior outcome of the remaining ones within the same network? Here, we assume the studied behavior of one actor can be described with K class labels {c1, · · · , cK}. For each label,ci can be 0or 1. For instance, one user might join multiple groups of interests, so 1 denotes the user subscribes toone group and 0 otherwise. Likewise, a user can be interestedn several topics simultaneously or click on multiple types of ads. One special case is K = 1. That is, the studied behavior can be described by a single label with 1 and 0 denoting corresponding meanings in its specific context, like whether or not one user voted for Barack Obama in the presidential election.



(i) First item in the second level
(ii) Second item in the second level

Figure 1: Contacts of One User in Facebook

3. Social Dimensions

Connections in social media are not homogeneous. People can connect to their family, colleagues, college classmates, or some buddies met online. Some of these relations are helpful to determine the targeted behavior (labels) but not necessarily always so true. For instance, Figure 1 shows the contacts of the first author on Facebook. The densely-knit group on the right side are mostly his college classmates, while the upper left corner shows his connections at his graduate school. Meanwhile, at the bottom left are some of his high-school friends. While it seems reasonable to infer that his college classmates and friends in graduate school are very likely to be interested in IT gadgets based on the fact that the user is a fan of IT gadget (as most of them are majoring in computer science), it does not make sense to propagate this preference to his high-school friends. In a nutshell, people are involved in different affiliations and connections are emergent results of those affiliations. These affiliations have to be differentiated for behavior prediction. However, the affiliation information is not readily available in social media. Direct application of collective inference[11] or label propagation [24] treats the connections in a social network homogeneously. This is especially problematic when the connections in the network are noisy. To address the heterogeneity presented in connections, we have proposed a framework (SocDim) [18] for collective behavior learning.

The framework SocDim is composed of two steps: 1) social dimension extraction, and 2) discriminative learning. In the first step, latent social dimensions are extracted based on network topology to capture the potential affiliations of actors. These extracted social dimensions represent how each actor is involved in diverse affiliations.

Table 1: Social Dimension Representation

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

One example of the social dimension representation is shown in Table 1. The entries show the degree of one user involving in an affiliation. These social dimensions can be treated as features of actors for the subsequent discriminative learning. Since the network is converted into features, typical classifier such as support vector machine and logistic regression can be employed. The discriminative learning procedure will determine which latent social dimension correlates with the targeted behavior and assign proper weights.

Now let's re-examine the contacts network in Figure 1. One key observation is that when actors are belonging to the same affiliations, they tend to connect to each other as well. It is reasonable to expect people of the same department to interact with each other more frequently. Hence, to infer the latent affiliations, we need to find out a group of people who interact with each other more frequently than random. This boils down to a classical community detection problem. Since each actor can involve in more than one affiliations, a soft clustering scheme is preferred. In the instantiation of the framework SocDim, modularity maximization [15] is adopted to extract social dimensions. The social dimensions correspond to the top eigenvectors of a modularity matrix. It has been empirically shown that this framework outperforms other representative relational learning methods in social media. However, there are several concerns about the scalability of SocDim with modularity maximization: Consequently, it is imperative to develop scalable methods that can handle large-scale networks efficiently without extensive memory requirement. In the next section, we elucidate an edge-centric clustering scheme to extract sparse social dimensions. With the scheme, we can update the social dimensions efficiently when new nodes or new edges arrive in a network.

4. Algorithm-Edgecluster

In this section, we first show one toy example to illustrate the intuition of our proposed edge-centric clustering scheme EdgeCluster, and then present one feasible solution to handle large-scale networks.

4.1 Edges-Centric View

As mentioned earlier, the social dimensions extracted based on modularity maximization are the top eigenvectors of a modularity matrix. Though the network is sparse, the social dimensions become dense, begging for abundant memory space. Let's look at the toy network in Figure 2. The column of modularity maximization in Table 2 shows the top eigenvector of the modularity matrix. Clearly, none of the entries is zero. This becomes a serious problem when the network expands into millions of actors and a reasonable large number of social dimensions need to be extracted. The eigenvector computation is impractical in this case. Hence, it is essential to develop some approach such that the extracted social dimensions are sparse. The social dimensions according to modularity maximization or other soft clustering scheme tend to assign a non-zero score for each actor with respect to each affiliation. However, it seems reasonable that the number of affiliations one user can participate in is upperbounded by the number of connections. Consider one extreme case that an actor has only one connection. It is expected that he is probably active in only one affiliation. It is not necessary to assign a nonzero score for each affiliation. Assuming each connection represents one dominant affiliation, we expect the number of affiliations of one actor is no more than his connections. Instead of directly clustering the nodes of a network into some communities, we can take an edge-centric view, i.e., partitioning the edges into disjoint sets such that each set represents one latent affiliation. 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 Figure 3. Based on the features (connected nodes) of each edge, we can cluster the edges into two sets as in Figure 4, where the dashed edges represent one affiliation, and the remaining edges denote another affiliation. One actor is considered associated with one affiliation as long as any of his connections is assigned to that affiliation. Hence, the disjoint edge clusters in Figure 4 can be converted into the social dimensions as the last two columns for edge-centric clustering in Table 2. Actor 1 is involved in both affiliations under this EdgeCluster scheme. In summary, to extract social dimensions, we cluster edges rather than nodes in a network into disjoint sets. To achieve this, k-means clustering algorithm can be applied. The edges of those actors involving in multiple affiliations (e.g., actor in the toy network) are likely to be separated into different clusters. Even though the partition of edge-centric view is disjoint, the affiliations in the node-centric view can overlap. Each actor can be involved in multiple affiliations. In addition, the social dimensions based on edge-centric clustering are guaranteed to be sparse.

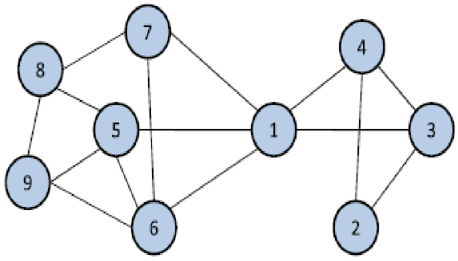


Figure 2: A Toy Example

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 2: Social Dimension(s) of the Toy Example

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
⋮	....								

Figure 3: Edge-Centric View

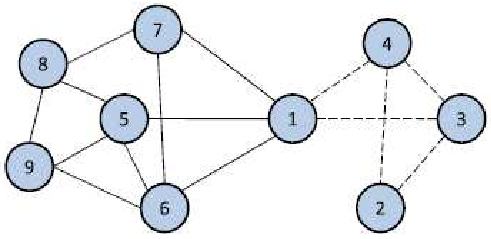


Figure 4: Edge Clusters

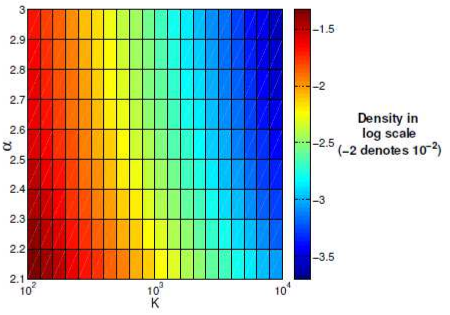


Figure 5: Density Upperbound of Social Dimensions

This is because the affiliations of one actor are no more than the connections he has.

4.2 K-means Variant

As mentioned above, edge-centric clustering essentially treats each edge as one data instance with its ending nodes being eatures. Then a typical k-means clustering algorithm can be applied to find out

disjoint partitions. One concern with this scheme is that the total number of edges might be too huge. Owning 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). This can be verified via the properties of power law distribution.

```
Input: data instances $\{x_i | 1 \le i \le m\}$
       number of clusters $k$
Output: $\{idx_i\}$
1. construct a mapping from features to instances
2. initialize the centroid of cluster $\{C_j | 1 \le j \le k\}$
3. repeat
4.   Reset $\{MaxSim_i\}$, $\{idx_i\}$
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 $sim(i, C_j) > MaxSim_i$
10.        $MaxSim_i = sim(i, C_j)$
11.        $idx_i = j$;
12.  for i=1:m
13.    update centroid $C_{idx_i}$
14. until no change in $idx$ or change of objective $< \epsilon$
```

Figure 6: Algorithm for Scalable K-means variant

k-means variant as in Figure 6 to handle clustering of many edges. We only keep a vector of MaxSim to represent the maximum similarity between one data instance with a centroid. In each iteration, we first identify the set of relevant instances to a centroid, and then compute the similarities of these instances with the centroid. This avoids the iteration over each instance and each centroid, which would cost O(mk) otherwise. Note that the centroid contains one feature (node) if and only if any edge of that node is assigned to the cluster. In effect, most data instances (edge) are associated with few (much less than k) centroids. By taking advantage of the feature-instance mapping, the cluster assignment for all instances (lines 5-11 in Figure 6) can be fulfilled in O(m) time. To compute the new centroid (lines 12-13), it costs O(m) time as well. Hence, each iteration costs O(m) time only. Moreover, the algorithm only requires the feature-instance mapping and network data to reside in main memory, which costs O(m + n) space. Thus, as long as the network data can be held in memory, this clustering algorithm is able to partition the edges into disjoint sets. Later as we show, even for a network with millions of actors, this clustering can be finished in tens of minutes while modularity maximization becomes impractical. As a simple k-means is adopted to extract social dimensions, it is easy to update the social dimensions if the network changes. If a new member joins a network and a new connection emerges, we can simply assign the new edge to the

corresponding clusters. The update of centroids with new arrival of connections is also straightforward. This k-means scheme is especially applicable for dynamic largescale networks

```
Input: network data, labels of some nodes
Output: labels of unlabeled nodes
1. convert network into edge-centric view as in Figure 3
2. perform clustering on edges via algorithm in Figure 6
3. construct social dimensions based on edge clustering
4. build classifier based on labeled nodes' social dimensions
5. use the classifier to predict the labels of unlabeled ones based
```

Figure 7: Scalable Learning of Collective Behavior

5. Conclusion And Futurework

In this work, we examine whether or not we can predict the online behavior of users in social media, given the behavior information of some actors in the network. Since the connections in a social network represent various kinds of relations, a framework based on social dimensions is employed. In the framework, social dimensions are extracted to represent the potential affiliations of actors before discriminative learning. But existing approach to extract social dimensions suffers from the scalability. 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, the extracted social dimensions are warranted to be sparse. Our model based on the sparse social dimensions shows comparable prediction performance as earlier proposed approaches to extract social dimensions. An incomparable advantage of our model is that, it can easily scale to networks with millions of actors while the earlier model fails. This scalable approach offers a viable solution to effective learning of online collective behavior in a large scale.

REFERENCES

[1] J. Bentley. Multidimensional binary search trees used for associative searching. Comm. ACM, 1975.

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

[3] R.-E. Fan and C.-J. Lin. A study on threshold selection for multi-label classification. 2007.

[4] 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, pages 1371–1374, 2005.

[5] L. Getoor and B. Taskar, editors. Introduction to Statistical Relational Learning. The MIT Press, 2007.

[6] M. Hechter. Principles of Group Solidarity. University of California Press, 1988.

[7] R. Jin, A. Goswami, and G. Agrawal. Fast and exact out-of-core and distributed k-means clustering. Knowl. Inf. Syst., 10(1):17–40, 2006.

[8] 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, 24:881–892, 2002.

[9] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In AAAI, 2006.

[10] 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.

[11] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. J. Mach. Learn. Res., 8:935–983, 2007.

[12] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27:415–444, 2001.

[13] 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, pages 49–55, 2005.

[14] M. Newman. Power laws, Pareto distributions and Zipf's law. Contemporary physics, 46(5):323–352, 2005.

[15] M. Newman. Finding community structure in networks using the eigenvectors of matrices. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), 74(3), 2006.

[16] C. Ordonez. Clustering binary data streams with k-means. In DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pages 12–19, 2003.

[17] M. Sato and S. Ishii. On-line em algorithm for the normalized gaussian network. Neural Computation, 1999.

[18] L. Tang and H. Liu. Relational learning via latent social dimensions. In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 817–826, 2009.

[19] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 677–685, 2008.

[20] L. Tang, S. Rajan, and V. K. Narayanan. Large scale multi-label classification via metalabeler. In WWW '09: Proceedings of the 18th international conference on World wide web, pages 211–220, 2009.

[21] 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.

[22] G. L. Zacharias, J. MacMillan, and S. B. V. Hemel, editors. Behavioral Modeling and Simulation: From Individuals to Societies. The National Academies Press, 2008.

[23] X. Zhu. Semi-supervised learning literature survey. 2006.

[24] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In ICML, 2003.