International Journal of Emerging Trends in Information & Knowledge Management
Vol. 4, Issue 1 – 2020
ISSN: 2581-3579

Gathering of Consumer Actions Base on the Uncovered Social Dimensions

Dr. Y. Dasaratha Rami Reddy[1], Mr. G. Sreenivasa Reddy[2]

[1]*Associate Professor, BVSR Engineering College.*
[2]*Associate Professor, CBIT Engineering College.*

Abstract

A collection of user behaviors approach has been shown effectively in addressing the heterogeneous connections among all the social networks. in every social network it involves in multi actors and huge size of data. To make this effective we can propose a model to collect the behavior of users. To make this very effective in addressing the data we use edge-centric clustering scheme to extract the sparse social dimensions. This can handle thousand of networks and the users who involved in it. It can compare the performance to the other dimension mechanisms.

Keywords: Heterogeneous Connections, Dimension Mechanism, Collective Behavior, Social Dimensions, Sparse Dimensions.

Introduction

The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. There are multiple phenomena referred to by this name in domains such as numerical analysis, sampling, combinatorics, machine learning, data mining, and databases. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. Also organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data however all objects appear to be sparse and dissimilar in many ways which prevents common data organization strategies from being efficient. 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 it isshown to be effective in addressing this heterogeneity.

International Journal of Emerging Trends in Information & Knowledge Management
Vol. 4, Issue 1 – 2020
ISSN: 2581-3579

The framework suggests a novel way of network classification: first, capture the latent affiliations of actors by extracting social dimensions based on network connectivity, and next, apply extant data mining techniquesto 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. 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.

Collective Behaviors

It 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 [7], [8]. 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. Here, behaviors include a broad range of actions: joining a group,connecting to a person, clicking on an ad, becoming interested in certain topics, dating people of a certain type, etc. In this work, we attempt to leverage the behavior correlation presented in a social network in order to predict collective behavior in social media. Given a network with the behavioral informa-tion of some actors, how can we infer the behavioral outcome of the remaining actors withinthe same network? Here, we assume the studied behavior of one actor can be described with K class labels fc$_1$ . . . c$_K$ g. Each label, c$_i$, can be 0 or 1. For instance, one user mightjoin multiple groups of interest, denotes that the user subscribes to group i, otherwise. Likewise, a user can be interested in several topics simultaneously, or click on multiple types of ads. One special case is K ¼ 1, indicatingthat the studied behavior can be described by a single label with 1 and 0. For example, if the event is the presidential election, 1 or 0 indicates whether or not a voter voted for Barack Obama. The problem we study can be described formally as follows.

Table 1.Social Dimension Representation

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

It should be noted that this problem shares the same spiritas within-network classification [9]. It can also be considered as a special case of semi-supervised learning [10] or relational learning [11] where objects are connectedwithin a network. Some of these methods, if applied directly to social media, yield only limited success [2]. This is because connections in social media are rather noisy and heterogeneous. In the next section, we will discuss the connection heterogeneity in social media, review the concept of social dimension, and anatomize the scalability limitations of the earlier model proposed in [2], which provides a compelling motivation for this work.

Social Dimensions

We change intellectual gears when we confront Blumer's final form of collective behavior, the social movement. He identifies several types of these, among which are active social movements such as the French Revolution andexpressive ones such as Alcoholics Anonymous. An active movement tries to change society; an expressive one triesto change its own members. The social movement is the form of collective behavior which satisfies least well the first definition of it which was offered at the beginning of this article. These episodes are less fluid than the other forms, and do not changes as often as other forms do. Furthermore, as can be seen in the history of the labor movement and many religious sects, a social movement may begin as collective behavior but over time become firmly established as a social institution. Connections in social media are not homogeneous. People can connect to their family, colleagues, college classmates, or buddies metonline. Some relations are helpful in determining a targetedbehavior (category) while others are not. This relation-type information, however, is often not readily available in social media. A direct application of collective inference [9] or label propagation [12] would treat connections in a social network as if they were homogeneous. To address the heterogeneity present in connections, a frame-work (SocioDim) [2] has been proposed for collective behavior learning.

Social dimensions extracted according to soft clustering, such as modularity maximization and probabilistic methods, are dense. Suppose there are 1 million actors in a network and 1,000 dimensions are extracted. If standard double precision numbers are used, holding the full matrix alone requires 1 M _ 1 K _ 8 ¼ 8 G memory. This large- size dense matrix poses thorny challenges for the extraction of social dimensions as well as subsequent discrimi-native learning. A key observation is that actors ofthe same affiliation tend to connect with each other. For instance, it is reason-able to expect people of the same department to interact with each other more frequently. Hence, to infer actors' latent affiliations, we need to find out a group of people who interact with each other more frequently than at random. This boils down to a classic community detection problem. Since each actor can get involved in more than one affiliation, a soft clustering scheme is preferred.

International Journal of Emerging Trends in Information & Knowledge Management
Vol. 4, Issue 1 – 2020
ISSN: 2581-3579

Example:

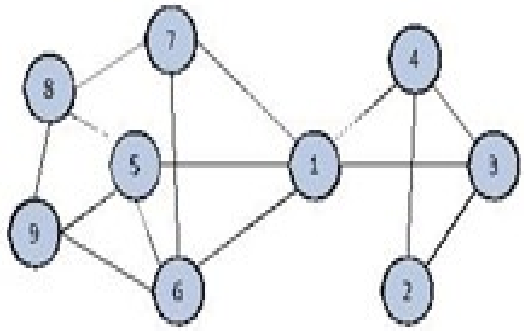

Figure 1.

PageRank scores, can be done very efficiently, computing thousands of eigenvectors or even more for a mega-scale network becomes a daunting task. Networks in social media tend to evolve, with new members joining and new connections occurring between existing members each day. This dynamic nature of networks entails an efficient updateof the model for collective behavior prediction. Efficient online updates of eigenvectors with expanding matricesremain a challenge. Consequently, it is imperative to develop scalable methods that can handle large-scale networks efficiently without extensive memory require- ments. Next, we elucidate on an edge-centric clustering scheme to extract sparse social dimensions. With such a scheme, we can also update the social dimensions efficiently when new nodes or new edges arrive.

Sparse Social Dimensions

In this section, we first show one toy example to illustrate the intuition of communities in an "edge" view and then present potential solutions to extract sparse social dimensions.

A. 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, It seems reasonable to state that the number of affiliations one user can participate in is upper bounded by his 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 of the many other affiliations. Assuming each connection represents one involved affiliation, we can expect the number of affiliations an actor has is no more than that of his connections. Rather than defining a community as a set of nodes, we redefine it as a set of edges. Thus, communities can be identified by partitioning edges of a network into disjoint sets.

International Journal of Emerging Trends in Information & Knowledge Management
Vol. 4, Issue 1 – 2020
ISSN: 2581-3579

Table 2.Social Dimension Table

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

Theorem 1: 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. Moreover, for many real-world networks whose node degree follows a power law distribution.

B. Edge Partition via Line Graph Partition

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 [15]. We will show that this is not a practical solution. 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 Fig. 4. For instance, eð1; 3Þ and eð2; 3Þ are connected in the line graph as they share one terminal node 3. Given a network, graph partition algorithms can be applied to its corresponding line graph. The set of communities in the line graph corresponds to a disjoint edge partition in the original graph. Recently, such a scheme has been used to detect overlapping communities [16], [17]. It is, however, prohibitive to construct a line graph for a mega-scale network. We notice that edges connecting to the same node in the original network form a clique in the corresponding line graph. For example, edges eð1; 3Þ, eð2; 3Þ, and eð3; 4Þ are all neighboring edges of node 3 in Fig. 1. Hence, they are adjacent to each other in the line graph in Fig. 4, forming a clique. This property leads to many more edges in a line graph than in the original network.

Theorem3: Let denote the exponent of a power law degree distribution for a given network, n the size of the network, and M the number of connections in its corresponding line graph. The divergence of the expectation tells us that as we go to larger and larger data sets, our estimate of number of connections with respect to the original network size will increase without bound. As we have mentioned, the majority of large-scale networks follow a power law with _ lying between 2 and 3. Consequently, the number of connections in a line graph can be extremely huge. Take the aforementioned YouTube network as an example again. It contains approximately 1 million nodes and 3 million links. Its resultant line graph will contain 4,436, 252,282 connections, too large to be loaded into memory. Recall that the original motivation to use sparse social dimensions is to address the scalability concern. Now, we end up dealing with a much larger sized.

International Journal of Emerging Trends in Information & Knowledge Management
Vol. 4, Issue 1 – 2020
ISSN: 2581-3579

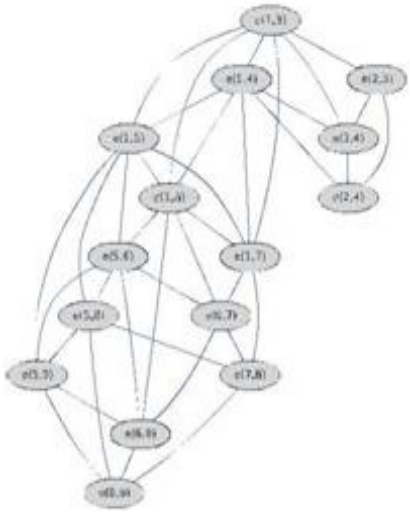

Figure 2.Each node in the line graph corresponds to an edge in the original graph

C. Edge Partition via Clustering Edge Instances

In order to partition edges into disjoint sets, we treatedges as data instances with their terminal nodes as features. For instance, we can treat each edge in the toy network in Fig1 as one instance, and the nodes that define edges as features. This results in a typical feature-based data format as in Table3. Then, a typical clustering algorithm like k-means clustering can be applied to find disjoint partitions. One concern with this scheme is that thetotal 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 ¼ OðnÞ as stated in the following theorem.

Theorem 4: 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 Still, millions of edges are the norm in a large-scale network. Direct application of some existing k-means implementation cannot handle the problem. For example, the k-means codeprovided in the Matlab package requires the computationof 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. On the other hand, the data of edge instances are quite sparse and structured. As each edge connects two nodes, the corresponding edge instance has exactly only two nonzero features as shown in Table 3. This sparsity can help accelerate the clustering process if exploited wisely. We conjecture that the centroids of k- means should also be feature sparse. Often, only a small portion of the data instances share features with the centroid. Hence, we just need to compute the similarity of the centroids with their relevant instances. In order to efficiently identify instances relevant to one centroid, we build a mapping from features (nodes) to instances (edges) beforehand. Once we have the mapping, we can easilyidentify the relevant instances by checking the nonzero features of the centroid.

International Journal of Emerging Trends in Information & Knowledge Management
Vol. 4, Issue 1 – 2020
ISSN: 2581-3579

D. 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.

Table 3.Statistics of Social Media Data

Data	BlogCatalog	Flickr	YouTube
Categories	39	195	47
Nodes (n)	10, 312	80, 513	1, 138, 499
Links (m)	333, 983	5, 899, 882	2, 990, 443
Network Density	6.3×10^{-3}	1.8×10^{-3}	4.6×10^{-6}
Maximum Degree	3, 992	5, 706	28, 754
Average Degree	65	146	5

In summary, we conduct the following steps to learn a model for collective behavior. Given a network (say, Fig. 1), we take the edge-centric view of the network data (Table 3) and partition the edges into disjoint sets (Fig. 2).

Based on the edge clustering, social dimensions can be constructed (TableII). Certain regularizations can be applied. Then, discriminative learning and prediction can be accomplished by considering these social dimensions asfeatures. The detailed algorithm is summarized in Fig. 6.

Experiment Setup

In this section, we present the data collected from social media for evaluation and the baseline methods for comparison.

A. Social Media Data

Two data sets reported in [2] are used to examine our proposed model for collective behavior learning. The first data set is acquired from BlogCatalog,3 the second from a popular photo sharing site Flickr.4 Concerning behavior, following [2], we study whether or not a user joins a group of interest. Since the BlogCatalog data do not have this group information, we use blogger interests as the behavior labels. Both data sets are publicly available at the first author's homepage. To examine scalability, we also include a mega-scale network5 crawled from YouTube.6 We remove those nodes without connections and select theinterest groups with 500 þ subscribers. Some statistics of the three data sets can be found in Table 4.

B. Baseline Methods

As we discussed in Section 4.2, constructing a line graphis prohibitive for large-scale networks. Hence, the line- graph approach is not included for comparison. Alternatively, the edge-centric

clustering (or EdgeCluster) in Section 4.2 is used to extract social dimensions on all data sets. We adopt cosine similarity while performing the clustering. Based on cross validation, the dimensionality is set to 5,000, 10,000, and 1,000 for BlogCatalog, Flickr, and YouTube, respectively. A linear SVM classifier [18] isthen exploited for discriminative learning. Another related approach to finding edge partitions is bi-connected components [19]. Bi-connected components of a graph are the maximal subsets of vertices such that the removal of a vertex from a particular component will not disconnect the component. Essentially, any two nodes in a bi-connected component are connected by at least two.

Node Cluster: Note that our prediction problem is essentially multi-label. It is empirically shown that thresholding can affect the final prediction performance drastically [21], [22]. For evaluation purposes, we assume the number of labels of unobserved nodes is already known, and check whether the top-ranking predicted labelsmatch with the actual labels. Such a scheme has been adopted for other multilabel evaluation works [23]. We randomly sample a portion of nodes as labeled and report the average performance of 10 runs in terms of Micro-F1 and Macro-F1 [22].

Experiment Results

In this section, we first examine how prediction performances vary with social dimensions extracted following different approaches. Then, we verify thesparsity of social dimensions and its implication for scalability. We also study how the performance varies withdimensionality. Finally, concrete examples of extracted social dimensions are given.

Table 4.Sparsity Comparison on BlogCatalog Datawith 10,312 Nodes

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 5.Sparsity Comparison on Flickr Data with 80,513 Nodes

Methods	Time	Space	Density	Upper Bound	Max-Aff	Ave-Aff
ModMax – 500	2.2×10^{3}	322.1M	1	–	500	500.0
EdgeCluster – 200	1.2×10^{4}	31.0M	1.2×10^{-1}	3.9×10^{-1}	156	24.1
EdgeCluster – 500	1.3×10^{4}	44.8M	7.0×10^{-2}	2.2×10^{-1}	352	34.8
EdgeCluster – 1000	1.6×10^{4}	57.3M	4.5×10^{-2}	1.3×10^{-1}	619	44.5
EdgeCluster – 2000	2.2×10^{4}	70.1M	2.7×10^{-2}	7.2×10^{-2}	986	54.4
EdgeCluster – 5000	2.6×10^{4}	84.7M	1.3×10^{-2}	2.9×10^{-2}	1405	65.7
EdgeCluster – 10000	1.9×10^{4}	91.4M	7×10^{-3}	1.5×10^{-2}	1673	70.9

International Journal of Emerging Trends in Information & Knowledge Management
Vol. 4, Issue 1 – 2020
ISSN: 2581-3579

A. Scalability Study

As we have introduced in Theorem 1, 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. The computation time, the memory footprint of social dimensions, their density, and other related statistics on all three data sets are reported in Tables 8, 9, and 10. Concerning the time complexity, it is interesting that computing the top eigenvectors of a modularity matrix is actually quite efficient as long as there is no memory concern. This is observed on the Flickr data.

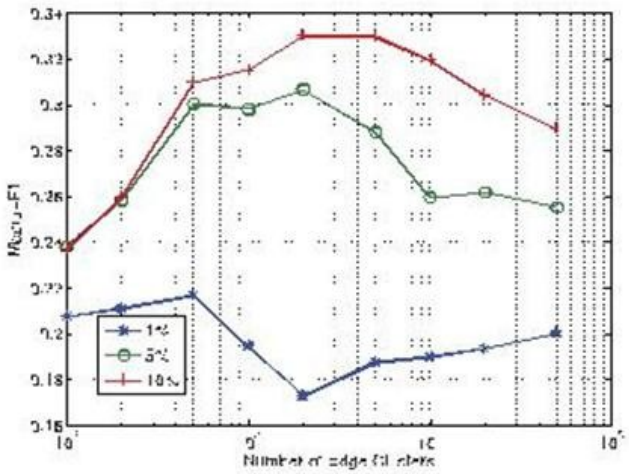

Figure 4.

However, when the network scales to millions of nodes (YouTube), modularity maximization becomes difficult (though an iterative method or distributed computation can be used) due to its excessive memory requirement. On the contrary, the EdgeCluster method can still work efficiently as shown in Table 10. The computation time of EdgeCluster for YouTube is much smaller than for Flickr, because the YouTube network is extremely sparse. The number of edges and the average degree in YouTube are smaller than those in Flickr, as shown in Table 4. Another observation is that the computation time of EdgeCluster does not change much with varying numbers of clusters. No matter how many clusters exist, the computa-tion time of EdgeCluster is of the same order. This is due to the efficacy of the proposed k-means variant in Fig. 5. In the algorithm, we do not iterate over each cluster and each centroid to do the cluster assignment, but exploit the sparsity of edge-centric data to compute only the similarity of a centroid and those relevant instances. This, in effect, makes the computational cost independent of the number of edge clusters.

B. Sensitivity Study

Our proposed EdgeCluster model requires users to specify the number of social dimensions

(edge clusters). One question which remains to be answered is how sensitive the performance is with respect to the parameter. We examine all three data sets, but find no strong pattern to determine optimal dimensionality. Due to the space limit, we include only one case here. Fig. 7 shows the Macro-F1 performance change on YouTube data. The performance, unfortunately, is sensitive to the number of edge clusters. It thus remains a challenge to determine the parameter automatically.

C. Regularization Effect

In this section, we study how EdgeCluster performance varies with different regularization schemes. Three differentregularizations are implemented: regularization based on community size, node affiliation, and both (we first apply community size regularization, then node affiliation). The results without any regularization are also included as a baseline for comparison. In our experiments, thecommunity size regularization penalty function is set to (log(jCijÞÞ2, and node affiliation regularization normalizes the summation of each node's community membership to 1 As shown in Fig. 8, regularization on both community size and node affiliation consistently boosts the performance on Flickr data. We also observe that node affiliation regularization significantly improves performance. It seems like community-size reg-ularization is not as important as the node-affiliation regularization. Similar trends are observed in the other two data sets. We must point out that, when both community-size regularization and node affiliation are applied, it is indeed quite similar to the tf-idf weighting scheme for representation of documents [24]. Such a weighting scheme actually can lead to a quite different performance in dealing with social dimensions extracted from anetwork.

Related Work

Classification with networked instances are known as within-network classification [9], or a special case of relational learning [11]. The data instances in a network arenot independently identically distributed (i.i.d.) as in conventional data mining. To capture the correlation between labels of neighboring data objects, typically a Markov dependency assumption is assumed. That is, the label of one node depends on the labels (or attributes) of its neighbors. Normally, a relational classifier is constructed based on the relational features of labeled data, and then an iterative process is required to determine the class labels for the unlabeled data. The class label or the class membership is updated for each node while the labels of itsneighbors are fixed However, a network tends to present heterogeneous relations, and the Markov assumption can only capture the local dependency. Hence, researchers propose to model network connections or class labels based on latent groups [20], [26]. A similar idea is also adopted in [2] to differentiate heterogeneous relations in a network by extracting social dimensions to represent the potential affiliations of actors in a network. The authors suggest using the community membership of a soft clustering scheme as social dimensions.

The extracted social dimensions are treated as features, and a support vector machine based on

that can be constructed for classification. It has been shown that the proposed social dimension approach significantly out performs representative methods based on collective inference. There are various approaches to conduct soft clustering for a graph. Some are based on matrix factorization, like spectral clustering [27] and modularity maximization [3]. Probabilistic methods are also developed [28], [29]. Please refer to [30] for a comprehensive survey. A disadvantage with soft clustering is that the resultant social dimensions.

Conclusions and Future Work

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.

References

1. L. Tang and H. Liu, "Toward Predicting Collective Behavior via Social Dimension Extraction," IEEE Intelligent Systems, vol. 25, no. 4, pp. 19-25, July/Aug.2010.
2. L. Tang and H. Liu, "Relational Learning via Latent Social Dimensions," KDD '09: Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 817-826, 2009.
3. M. Newman, "Finding Community Structure in Networks Using the Eigenvectors of Matrices," Physical Rev. E (Statistical, Nonlinear, and Soft Matter Physics), vol. 74, no. 3, p. 036104, http://dx.doi.org/10.1103/PhysRevE.74.036104, 2006.
4. L. Tang and H. Liu, "Scalable Learning of Collective Behavior Based on Sparse Social Dimensions," CIKM '09:Proc. 18th ACM Conf. Information and Knowledge Management, pp. 1107-1116, 2009.
5. P. Singla and M. Richardson, "Yes, There Is a Correlation: - From Social Networks to Personal Behavior on the Web," WWW '08: Proc. 17th Int'l Conf. World Wide Web, pp. 655-664, 2008.
6. M. McPherson, L. Smith-Lovin, and J.M. Cook, "Birds of a Feather: Homophily in Social Networks," Ann. Rev. of Sociology, vol. 27, pp. 415-444, 2001.

