Toward Predicting Collective Behavior via Social Dimension Extraction

Lei Tang and Huan Liu, *Arizona State University*

The social-dimension-based learning framework (SocioDim) can help predict online behaviors of social media users given a network and the behavior information of some actors in the network.

Social media such as Facebook, MySpace, Twitter, Digg, YouTube, and Flickr help people from all walks of life express their thoughts, voice their opinions, and connect to each other anytime and anywhere. Popular content-sharing sites such as Delicious, Flickr, and YouTube let users upload, tag, and comment on different types of content (including bookmarks, photos, and videos). Users registered at these sites can also become friends, fans, or followers of others.

The prolific and expanded use of social media has turned online interactions into a vital part of the human experience. Barack Obama's successful US presidential election campaign, for example, has been partially attributed to his smart Internet strategy and access to millions of younger voters through social media such as Facebook, the popular social networking site currently claiming 400 million active users. The large population actively involved in social media also provides great opportunities for businesses. Dell, one of the top PC companies, has reportedly "earned $3 million in revenue directly through [the social networking and microblogging service] Twitter since 2007" (see http://bits.blogs.nytimes.com/2009/06/12/dell-has-earned-3-million-from-twitter).

Concomitant with the opportunities provided by rocketing online traffic in social media are the challenges of user and customer profiling, accurate user matching at different domains, as well as effective social networking advertising, marketing, and recommendations. For example, the *New York Times* has reported that the banner ads displayed to members of social media received little attention because most of them were irrelevant (http://www.nytimes.com/2008/12/14/business/media/14digi.html). On the other hand, social networking sites can only collect limited user profile information, either because of privacy issues or because users decline to share information. If we can leverage the abundant network data accessible in social media wisely, the situation of social networking advertising might be improved significantly.

In this article, we examine how to predict the online behavior of social media users given the behavior information of some actors in the network. We can connect many social media tasks to the problem of collective behavior prediction. Because the connections in a social network represent

1541-1672/09/$26.00 © 2010 IEEE
Published by the IEEE Computer Society

Related Work in Collective Behavior Prediction

The collective behavior prediction problem is relevant to within-network classification[1] when data instances are presented in a network format. In the case of social learning, the data instances are not independently, identically distributed as in conventional data mining. To capture the correlation between labels of neighboring data instances, typically a Markov dependency 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, which then requires an iterative process 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 its neighbors are fixed. This process is repeated until the label inconsistency between neighboring nodes is minimized. Research has shown that a simple weighted-vote relational neighborhood (wvRN) classifier works reasonably well on some benchmark relational data and is recommended as a baseline for comparison.[1]

Most relational classifiers, following the Markov assumption, only capture the local dependency. To handle the long-distance correlation, the latent group model[2] and the nonparametric infinite hidden relational model[3] assume Bayesian generative models such that the link (and actor attributes) are generated based on the actors' latent cluster membership. However, the model intricacy and high computational cost for inference associated with these models hinder their application to large-scale networks. Hence, a clustering algorithm is applied first to find the hard cluster membership of each actor, and then the latent group variables are fixed for later inference.[2] Because each actor is assigned to only one latent affiliation, it does not capture the multitude of affiliation association required in social learning.

References

1. S.A. Macskassy and F. Provost, "Classification in Networked Data: A Toolkit and a Univariate Case Study," *J. Machine Learning Research*, vol. 8, no. 5, 2007, pp. 935–983.
2. J. Neville and D. Jensen, "Leveraging Relational Autocorrelation with Latent Group Models," *Proc. 4th Int'l Workshop Multi-relational Mining* (MRDM 05), ACM Press, 2005, pp. 49–55.
3. Z. Xu, V. Tresp, S. Yu, and K. Yu, "Nonparametric Relational Learning for Social Network Analysis," *Proc. KDD Workshop on Social Network Mining and Analysis*, ACM Press, 2008.

various kinds of relations, we introduce SocioDim, a social-dimension-based learning framework. Using SocioDim, we can extract social dimensions that represent the latent affiliations associated with actors and then apply supervised learning to determine which dimensions are informative for behavior prediction. The framework is especially suitable for large-scale networks, paving the way for collective behavior study in many real-world applications.

Collective Behavior

We can generalize the social network advertising problem to the study of collective behavior. Different types of behavior can include a broad range of actions, such as joining a group, connecting to a person, clicking on an ad, becoming interested in certain topics, or dating certain types of people. Collective behavior refers to behaviors of individuals in a social network environment, but it is not simply the aggregation of individual behaviors. In a connected environment, individuals' behaviors tend to be interdependent, influenced by the behavior of friends. This naturally leads to behavior correlation between connected users.

Such collective behavior correlation can also be explained by *homophily*,[1] a term coined in the 1950s to explain our tendency to link up with one another in ways that confirm rather than test our core beliefs. In other words, we are more likely to connect to others sharing certain similarities with us, and similar people tend to become friends, leading to similar behavior between connected egos in a social network. This phenomenon has been observed in both the real world and online environments. For example, if our friends buy something, there is a better-than-average chance that we'll buy it, too.

Because a social network provides valuable information concerning actor behaviors, it is natural to ask how we can use the behavior correlation presented in a social network to predict collective behavior. That is, given a social network with behavior information of some actors, how can we infer the behavior outcome of the remaining actors within the same network?

This problem assumes that we can observe the behaviors of some individuals so that social learning is attainable. The amount of information that we can collect in reality depends on tasks. For instance, if we want to know whether a user will click on an ad, we can collect this information when the ad is displayed to the user. To determine behavior concerning voting for a presidential candidate, we can collect some voluntary responses using online surveys. With such behavior information, we can unravel the collective behavior by exploiting the network connectivity between actors. (See the "Related Work in Collective Behavior Prediction" sidebar for previous work in this area.)

Heterogeneous Relations in Social Networks

To understand collective behavior, researchers in social science and behavioral studies have used the *threshold*

Table 1. Social dimension representation.

Actors	ASU	Fudan	Sanzhong
Lei	1	1	1
Actor 1	1	0	0
⋮	⋮	⋮	⋮

model,[2] in which an actor adopts one action when the number of his friends taking an action exceeds a certain threshold. In his seminal work, Thomas C. Schelling used a variant of this threshold model to show that a small preference for one's neighbors to be of the same color could lead to total race segregation.[3] Similarly, the machine learning community has adopted collective inference[4] to make predictions about collective behavior. It assumes that the behavior of one actor depends on that of her friends. For prediction, collective inference is required to find an equilibrium status to minimize the inconsistency between connected actors. This is normally done by iteratively updating the possible behavior output of one actor while fixing the behavior output (or attributes) of his connected friends in the network. Researchers have shown that approaches that consider this network connectivity for behavior prediction outperform those that do not.

However, connections in social media are often not homogeneous. The heterogeneity presented in network connectivities can hinder the success of collective inference. People can connect to their family, colleagues, college classmates, and online buddies. Some of these heterogeneous relations might be helpful in determining a targeted behavior, while others are not. For instance, the Facebook contacts of one user Lei might consist of postgraduate friends he met at Arizona State University (ASU), his undergraduate classmates at Fudan University, and some high school friends in Sanzhong. While it seems reasonable to infer that his friends at ASU are likely to attend a football game, based on the fact that he is going to watch an ASU football game, it does not make sense to propagate this preference to his high school friends or undergraduate classmates. Directly applying collective inference to these kinds of networks does not differentiate heterogeneous connections, thus making it risky for prediction of collective behavior.

Moreover, online social networks tend to be noisier than those in the physical world because it is much easier for users to connect online. Some users have thousands of online friends whereas this is hardly true in reality. For instance, one Flickr user connects to more than 19,000 contacts (see http://www.flickr.com/people/22711787@N00). In such cases, it is likely that only a small portion of them can influence the actor's behavior. Therefore, it is helpful to differentiate people's relations for behavior prediction.

It is often a luxury to have detailed relation information, although some sites like LinkedIn and Facebook ask people how they know each other when they become connected. Most of the time, people decline to share such detailed information, resulting in a social network between users without explicit information about pairwise relation types. Even if the pairwise relation information is available, it is not necessarily relevant or refined enough to help determine the behaviors of connected users. For example, knowing two actors are college classmates does not necessarily help predict how they will vote for a presidential candidate.

Therefore, collective behavior prediction must address two challenges:

- Without relation-type information, is it possible to differentiate relations based on network connectivity?
- If relations are differentiated, how can we determine whether a relation can help behavior prediction?

Social Dimensions

Differentiating pairwise relations based on network connectivity alone is by no means an easy task. Alternatively, we can look at actors' *social dimensions*,[5] which represent the relations associated with actors, with each dimension denoting one relation. If two actors a_i and a_j are connected because of relationship R, both a_i and a_j should have a nonzero entry in the social dimension that represents R. In the context of our previous Facebook example, we can characterize the relations between Lei and his friends by three affiliations: Arizona State University (ASU), Fudan University (Fudan), and high school (Sanzhong). Table 1 shows the actors' corresponding social dimensions. If one actor belongs to one affiliation, then the corresponding entry is nonzero. Because Lei is an ASU student, his social dimension includes a nonzero entry for the ASU dimension to capture the relationship between him and his ASU friends.

Social dimensions capture prominent interaction patterns presented in a network. Because of the multifaceted nature of human social life, one actor is likely to be involved in multiple social dimensions; the table shows three different relations for Lei.

SocioDim Framework

The social dimensions shown in Table 1 are constructed based on

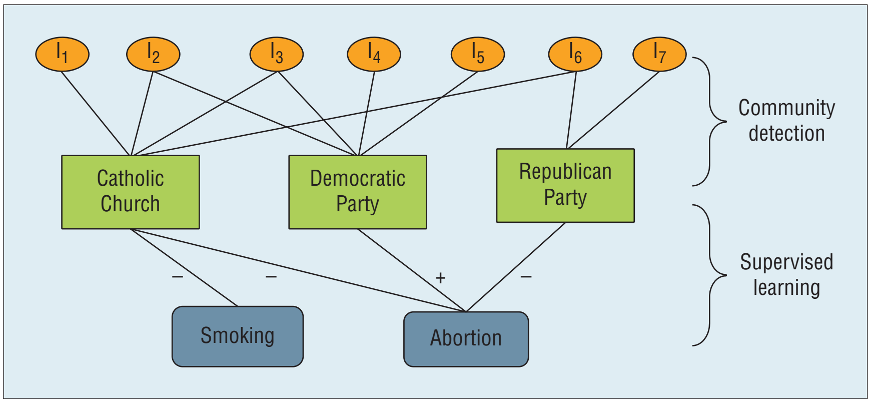

Figure 1. Underlying collective behavior model for the SocioDim framework. The orange circles denote individuals, the green rectangles denote affiliations, and the blue blocks at the bottom denote behaviors.

explicit relation information. Without knowing such true relationship, how can we extract latent social dimensions? One key observation is that actors of the same relation tend to connect to each other. For instance, Lei's friends at ASU tend to interact with each other as well. Hence, to infer a latent social dimension, we need to find a group of people who interact with each other more frequently than at random. This boils down to a classical community-detection problem. Each actor can be assigned to multiple communities.

After we extract the social dimensions, we consider them as normal features and combine them with the behavioral information to conduct supervised learning. Different tasks might represent the user behavior in different ways. In certain cases, we can represent the behavior output by labels—for instance, whether a user joins a group, likes a product, or votes for a presidential candidate. In some other cases, we might represent the behavior output more properly using continuous numbers, such as the probability that a user clicks on an ad and the frequency that a user visits an interest group. Depending on the behavior representation (discrete or continuous values), we can use a classifier or a regression learner. This supervised learning is critical because it will determine which dimensions are relevant to the target behavior and assign proper weights to different social dimensions.

Hence, we can apply the SocioDim framework to handle the network heterogeneity.[5] It consists of two steps, each of which addresses one challenge sketched in the previous section:

- Extract meaningful social dimensions based on network connectivity via community detection.
- Determine relevant social dimensions through supervised learning.

Prediction is straightforward once a learned model is ready, because the social dimensions have been calculated for all actors. Applying the constructed model to the actors' social dimensions without behavior information, we obtain the behavior predictions.

The SocioDim framework basically assumes that the affiliation membership of actors determines their behavior (see Figure 1). Individuals are associated with different affiliations in varying degrees (with line thickness indicating the degree of association) and distinctive affiliations regulate the member behavior differently. For instance, the Catholic Church opposes smoking and abortion, while the Democratic Party does not oppose abortion. Some affiliations might have no influence over certain behavior, however, such as the Democratic and Republican Parties over smoking. The final behavior output of individuals depends on the affiliation regularization and individual associations. The first step of our proposed SocioDim framework essentially finds out the individual associations, and the second step learns the affiliation regularization by assigning weights to different affiliations.

Social Dimension Extraction

We proposed the SocioDim framework to address the relation heterogeneity presented in social networks. Thus, a sensible method for social dimension extraction becomes critical to its success. We can categorize existing methods to extract social dimensions into node and edge views.

Node-view methods concentrate on clustering network nodes into communities. As we have mentioned, the extraction of social dimensions boils down to a community-detection task, and each actor can be assigned multiple affiliations. Many existing community-detection methods, with the aim of partitioning network nodes into disjointed sets, do not satisfy this requirement. Alternatively, a soft-clustering scheme is preferred. Hence, we can apply variants of spectral clustering, modularity maximization, non-negative matrix factorization, or block models. One representative example of node-view methods is modularity maximization.[6] The top eigenvectors of a modularity matrix can be used as the social dimensions.[5]

Suppose we are given a toy network, as in Figure 2, of which there are nine actors, with each circle representing one affiliation. For k affiliations, typically at least $k - 1$ social dimensions are required. Table 2 shows the top social dimensions based on modularity maximization of the toy example. The actors with negative values belong to one affiliation, and actor 1 and those with positive values belong to the other affiliation. Note that actor 1 is involved in both affiliations. Hence, actor 1's value is in between (close to 0). This social dimension does not explicitly state the association(s), but it presents the degree of associations for all actors.

Edge-view methods concentrate on clustering edges of a network into communities.[7] An edge resides in only one affiliation, although a node can be involved in multiple affiliations. For instance, in Figure 2, actor 1 participates in both affiliations, but his connections are well separated. Hence, instead of directly clustering the nodes of a network into some communities, we can take an edge-centric view—that is, partitioning the edges into disjointed sets such that each set represents one latent affiliation (as in Figure 2). In the figure, the red edges represent one affiliation and the green ones denote the other.

Table 2 shows how we can convert the resultant edge partition into a social dimension representation. An actor is involved in one affiliation as long as any of his connections are involved in that affiliation. For instance, actor 1 has connections engaged in both affiliations: connection (1, 7) is in the red set, and connection (1, 4) is in the green one. Thus, actor 1 has nonzero entries for both affiliations, as Table 2 shows. On the contrary, with all of its connections residing in the green set, actor 4 has only one nonzero entry in its corresponding social dimension. This naturally leads to sparse social dimensions (see Table 2).

A node-view method such as modularity maximization yields nonzero values for all the entries, resulting in a dense representation. By contrast, the social dimensions based on edge-view methods are guaranteed to be sparse. One consequence of this edge partition is that the number of affiliations is bounded by the number of connections one actor has. If one actor has d connections, his affiliations are no more than d. In the extreme case, if one actor has only one connection, this actor can engage in only one affiliation. Owing to the power law distribution presented in large-scale networks,[8] a large portion of nodes in a network would bear a low degree. Hence, the resultant social dimensions would be sparse.

Both node- and edge-view methods can be applied to extract social dimensions. The key difference between these methods is that the node view defines a community as a set of nodes with each node assigned to multiple communities, while an edge view defines a community as a set of edges with each edge assigned to only one community. One method is not better than the other; it depends on the network data, applications, and approaches being used.

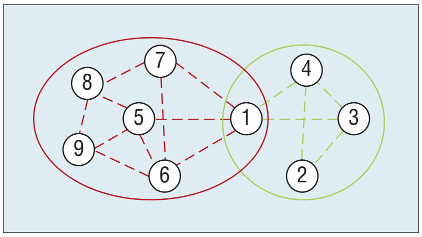

Figure 2. A network of two communities. Actor 1 is affiliated with both the red and green communities.

Comparative Study

The SocioDim framework has many advantages over collective inference. We studied behaviors on three representative social media sites to yield empirical results. In particular, we crawled social networks on the blog directory BlogCatalog (http://www.blogcatalog.com), Flickr, and YouTube. User interests or subscribed interest groups were deemed as behavior labels. We used F1, the harmonic mean of precision and recall, to evaluate predictions. That is, let y and $\hat{y} \in \{0,1\}^n$ denote the true labels and predictions, respectively. Precision (P), Recall (R), and F1-measure ($F1$) are defined as

$$P = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} \hat{y}_i},$$

$$R = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i},$$

$$F1 = \frac{2PR}{P+R}$$

Figure 3 and Table 3 report the average performance over a multitude

Table 2. Social dimensions of the toy example.

Actors	Node-centric clustering	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

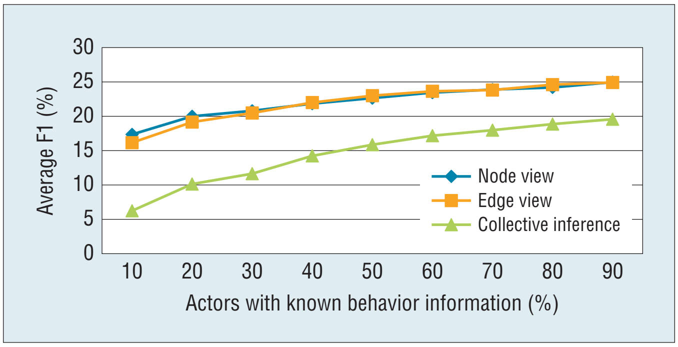

Figure 3. SocioDim's performance on the BlogCatalog network with 10,312 actors. Node view and edge view denote SocioDim with modularity maximization[5] and edge-centric clustering[7] for social dimension extraction. Collective inference represents the weighted-vote relational neighbor (wvRN) method.[4]

of behaviors for each site. These results illustrate the benefits of the SocioDim framework.

By handling heterogeneity, SocioDim can outperform collective inference considerably, especially when the social network is sparse and little information on user behavior is available.[5] Figure 3 shows the performance of representative methods of node-view, edge-view, and collective inference, respectively. The SocioDim framework, with social dimension extraction either in node or edge view, indicates that differentiating connections between actors does help with behavior prediction.

With a proper method to extract social dimensions, we can develop a scalable instantiation of the framework in terms of both time and space complexity. Table 3 shows that, with a normal PC, SocioDim with social dimension extraction in edge view can handle a YouTube network of more than 1 million users in approximately 10 minutes and keep the extracted social dimensions extremely sparse, occupying only a 40-Mbyte memory space. On the other hand, when there is no memory constraint, node-view methods cost less time, as the Flickr data shows. The computation time of the edge-view method on the YouTube network is much smaller than on Flickr, because although it has fewer nodes, Flickr has more edges in the network. The node-view method—which involves an eigenvector computation problem—is proportional to the number of nodes, whereas the edge-view method is proportional to the number of edges.

The SocioDim framework essentially converts a network into features, offering a simple mechanism to seamlessly integrate two seemingly orthogonal kinds of information: social networks and actor features. When some actor features (such as user profiles and blog content) are also available, we can combine these features with the extracted social dimensions before subsequent supervised learning. Our results show that this kind of integration can boost the performance of relying on either type of information alone.[5]

In addition, SocioDim enables code reuse and saves human effort in practical deployment. SocioDim consists of two steps: community detection and supervised learning. Many algorithms have been developed and numerous existing software packages can be plugged in instantaneously.

Table 3. Scalability comparison of different methods for social dimension extraction.[7]

500 dimensions	Extraction methods	Flickr (80,000 actors, 6 million links)	YouTube (1.1 million actors, 3 million links)
Memory footprint*	Node view	322.1 Mbytes	4.6 Gbytes
	Edge view	44.8 Mbytes	39.9 Mbytes
Computation time**	Node view	40 minutes	—
	Edge view	3.6 hours	10 minutes

*Memory footprint is the size of the extracted social dimensions.

**Computation time refers to the time to compute social dimensions.

The SocioDim framework demonstrates promising results toward predicting collective behavior. However, many challenges require further research. For example, networks in social media are continually evolving, with new members joining a network and new connections established between existing members each day. This dynamic nature of networks entails efficient update of the model for collective behavior prediction. It is also intriguing to consider temporal fluctuation into the problem of collective behavior prediction. We expect that along with the SocioDim framework, more research work would emerge in the near future.

Acknowledgments

This work is supported in part by Air Force Office of Scientific Research (AFOSR) and Office of Naval Research (ONR). We thank the anonymous reviewers wholeheartedly for their expert opinions and constructive suggestions.

References

1. M. McPherson, L. Smith-Lovin, and J.M. Cook, "Birds of a Feather: Homophily in Social Networks," *Ann. Rev. of Sociology*, vol. 27, 2001, pp. 415–444.

2. M. Granovetter, "Threshold Models of Collective Behavior," *Am. J. Sociology*, vol. 83, no. 6, 1978, p. 1420.
3. T.C. Schelling, "Dynamic Models of Segregation," *J. Mathematical Sociology*, vol. 1, 1971, pp. 143–186.
4. S.A. Macskassy and F. Provost, "Classification in Networked Data: A Toolkit and a Univariate Case Study," *J. Machine Learning Research*, vol. 8, no. 5, 2007, pp. 935–983.
5. L. Tang and H. Liu, "Relational Learning via Latent Social Dimensions," *Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (KDD 09), ACM Press, 2009, pp. 817–826.
6. 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, 2006.
7. L. Tang and H. Liu, "Scalable Learning of Collective Behavior Based on Sparse Social Dimensions," *Proc. 18th ACM Conf. Information and Knowledge Management* (CIKM 09), ACM Press, 2009, pp. 1107–1116.
8. M. Newman, "Power Laws, Pareto Distributions and Zipf's Law," *Contemporary Physics*, vol. 46, no. 5, 2005, pp. 323–352.

THE AUTHORS

Lei Tang is a doctoral candidate in computer science and engineering at Arizona State University. His research interests are in social computing and data mining, in particular, relational learning with heterogeneous networks, group evolution, profiling and influence modeling, and collective behavior modeling and prediction in social media. Tang has a BS in computer science from Fudan University. He is a member of the ACM and IEEE. Contact him at l.tang@asu.edu.

Huan Liu is a professor of computer science and engineering at Arizona State University. His research interests are in data and Web mining, machine learning, social computing, and artificial intelligence, investigating problems that arise in many real-world applications with high-dimensional data of disparate forms such as social computing, modeling group interaction, text categorization, biomarker identification, and text and Web mining. Liu has a PhD in computer science from the University of Southern California. He is a member of the AAAI, ACM, American Society for Engineering Education (ASEE), and IEEE. Contact him at huan.liu@asu.edu.