Toward Collective Behavior Prediction via Social Dimension Extraction

Lei Tang and Huan Liu, Arizona State University

Abstract

Collective behavior refers to how individuals behave when they are exposed in a social network environment. In this article, we examine how we can predict online behaviors of users in a network, given the behavior information of some actors in the network. Many social media tasks can be connected to the problem of collective behavior prediction. Since connections in a social network represent various kinds of relations, a social-learning framework based on social dimensions is introduced. This framework suggests extracting social dimensions that represent the latent affiliations associated with actors, and then applying supervised learning to determine which dimensions are informative for behavior prediction. It demonstrates many advantages, especially suitable for large-scale networks, paving the way for the study of collective behavior in many real-world applications.

Keywords: Behavior Prediction, Collective Behavior, Social Dimensions, Social Media, Edge-Centric Clustering, Node-Centric Clustering, Collective Inference

Collective Behavior

Social media such as Facebook, MySpace, Twitter, BlogCatalog, Digg, YouTube and Flickr, facilitate people of all walks 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 (e.g., bookmarks, photos, videos). Users registered at these sites can also become friends, a fan or follower of others. The prolific and expanded use of social media has turned 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, a popular social networking site claiming to attract 400 million active users up to date[1]. The large population actively involved in social media also provides great opportunities for business. One of the top

[1] http://www.facebook.com/press/info.php?statistics

PC companies Dell said that "the company had earned $3 million in revenue directly through Twitter since 2007"[2], where Twitter is a social networking and micro-blogging service that enables its users to send and read short messages.

Concomitant with the opportunities indicated by the rocketing online traffic in social media are the challenges for user/customer profiling, accurate user matching at different domains, recommendation as well as effective advertising and marketing. Take social networking advertising as an example. Currently, advertising in social media has encountered many challenges[3]. A recent study[4] from the research firm IDC suggested that "just 57% of all users of social networks clicked on an ad in the last year, and only 11% of those clicks lead to a purchase". Note that some social networking sites can only collect very limited user profile information, either due to the privacy issue or because the user declines to share information. On the contrary, a social network such as the friendship network in Facebook, the contact network in Flickr or YouTube, and the follower network in Twitter, is accessible. If one can leverage a small portion of user information and the network data wisely, the situation might improve significantly.

The aforementioned social network advertizing problem can be generalized to the study of *collective behavior*. Here, behavior can include a broad range of actions: joining a group, connecting to a person, clicking on some ad, becoming interested in certain topics, dating with people of certain type, etc. **Collective behavior** refers to behaviors of individuals who are exposed in a social network environment.

Collective behavior is not simply the aggregation of individuals' behavior. In a connected environment, behaviors of individuals tend to be interdependent. That is, one's behavior can be influenced by the behavior of his/her friends. This naturally leads to *behavior correlation* between connected users. Such collective behavior correlation can also be explained by *homophily*. Homophily [3] 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 environments. In other words, similar people tend to become friends, leading to similar behavior between connected egos in a social network. Take marketing as an example. If our friends buy something, there is a better-than-average chance that we'll buy it too.

Since a social network provides valuable information concerning actor behaviors, one natural question is how we can utilize the behavior correlation presented in a social network to predict collective behavior. In particular, the *collective behavior prediction* problem can be stated as follows:

> **Given a social network with behavior information of some actors, how can we infer the behavior outcome of the re-**

[2] http://bits.blogs.nytimes.com/2009/06/12/dell-has-earned-3-million-from-twitter/
[3] http://www.nytimes.com/2008/12/14/business/media/14digi.html
[4] http://www.nytimes.com/2008/12/01/technology/internet/01facebook.html

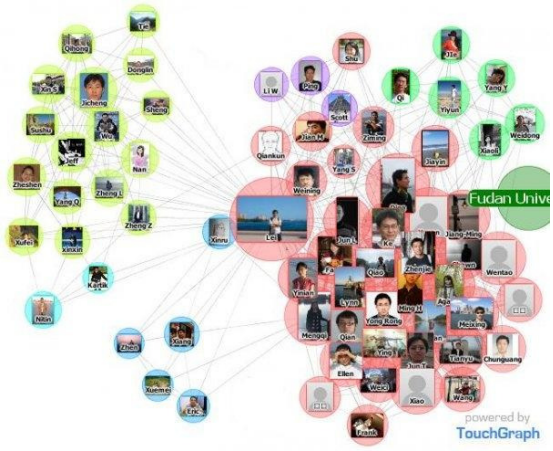

Figure 1: Contacts of One User in Facebook

maining ones within the same network?

This problem assumes that the behaviors of some individuals are observed so that social learning is attainable. This portion of information can be collected in reality depending on tasks. For instance, if the behavior is about whether a user clicks on an ad, this information can be collected when the ad is displayed to the user. For another kind of behavior concerning voting for a presidential candidate, some voluntary responses can be collected through sending out online surveys. Given some behavior information, the collective behavior can be unraveled by exploiting the network connectivity between actors.

Heterogeneous Relations in Social Networks

To understand collective behavior, one classical model well studied in social science and behavioral study is the threshold model [1], in which an actor adopts one action when the number of his friends taking an action exceeds a certain threshold. Indeed, Schelling in his seminal work [7] 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. A similar idea, *collective inference* [2], is adopted in machine learning community to make predictions about collective behavior. It assumes that the behavior of one actor is dependent upon that of his friends. For prediction, *collective inference* is required to find an equilibrium status such that the inconsistency between connected actors is minimized. 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. It has been shown that considering this network connectivity for behavior prediction outperforms 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, or some buddies met online. Some of these relations are helpful in determining the targeted behavior, but not necessarily always so. For instance, Figure 1 shows the contacts of the first author on Facebook. The densely-knit group on the right side consists of mostly his college classmates at Fudan University, while the upper left corner shows his connections in his graduate school (Arizona State University). Meanwhile, at the bottom left are some of his high-school friends in Sanzhong. While it seems reasonable to infer that his friends at ASU is likely to participate in a football game, based on the fact that the user is going to watch an ASU football game, it does not make sense to propagate this preference to his high-school friends or college classmates. A social network can consist of heterogeneous relations. Directly applying collective inference to this kind of networks does not differentiate these connections, thus becoming risky for prediction of collective behavior.

Moreover, online social networks tend to be more noisy than those in the physical world, as it is much easier for users to get connected online. It is not surprising that some users have thousands of online friends whereas this is hardly true in reality. For instance, one user in Flickr connects to more than 19,000 contacts[5]. Among so many friends, it might be the case that only a small portion of them can influence the actor's behavior. In summary, **people are involved in different relations and it is helpful to differentiate these relations for behavior prediction**.

It is often a luxury to have detailed relation information, though some sites like LinkedIn and Facebook do 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 type. 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 help much for the behavior prediction of voting for a presidential candidate.

The above concerns pose the following two challenges to be addressed for collective behavior prediction:

- **Without information of relation type, 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?**

Table 1: Social Dimension Representation

Actors	ASU	Fudan	Sanzhong
Lei	1	1	1
$Actor_1$	1	0	0
⋮	⋮	⋮	⋱

Social Dimensions

Differentiating pairwise relations based on network connectivity alone is by no means an easy task. Alternatively, we can look at *social dimensions* [8] of actors. **Social dimensions** are introduced to represent the relations associated with actors, with each dimension denoting one relation. Suppose two actors a_i and a_j are connected because of relationship R, both a_i and a_j should have a non-zero entry in the social dimension which represents R. Let us revisit the example in Figure 1. The relations between the user and his friends can be characterized by three affiliations: Arizona State University (ASU), Fudan University (Fudan), and a high school (Sanzhong). The corresponding social dimensions of actors in Figure 1 are shown in Table 1. In the table, if one actor belongs to one affiliation, then the corresponding entry is non-zero. Since Lei is a student ASU, his social dimension includes a non-zero entry for the ASU dimension to capture the relationship of his ASU friends and him.

Social dimensions capture prominent interaction patterns presented in a network. Note that one actor is very likely to be involved in multiple different social dimensions (e.g., Lei participates in 3 different relations in the table). This is consistent with multi-facet nature of human social life that one is likely to be involved in distinctive relations with different people.

SocioDim Framework

The social dimensions shown in Table 1 are constructed based on the explicit information of relations. In reality, without knowing 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 as well. For instance, as shown in Figure 1, the friends of Lei at ASU tend to interact with each other as well. Hence, to infer a latent social dimension, 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. A requirement is that one actor is allowed to 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,

[5] http://www.flickr.com/people/22711787@N00

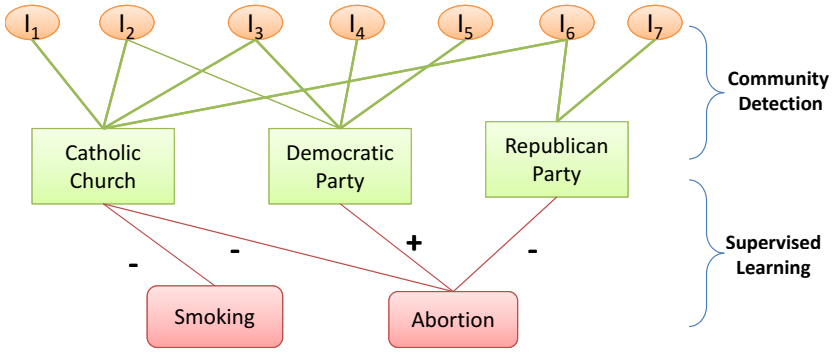

Figure 2: Underlying Collective Behavior Model for SocioDim framework

whether a user joins a group, whether he likes a product, whether he votes for a presidential candidate. In some other cases, it might be true that the behavior output is represented more properly using continuous numbers, like 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), a classifier or a regression learner can be used. This supervised learning is critical as it will determine which dimensions are relevant to the target behavior and assign proper weights to different social dimensions.

In summary, a social-dimension based learning framework SocioDim [8] can be applied to handle the network heterogeneity. It consists of two steps, with each addressing 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, since the social dimensions have been calculated for all actors. Applying the constructed model to the social dimensions of the actors without behavior information, we obtain the behavior predictions.

This SocioDim framework basically assumes the affiliation membership of actors determines one's behavior. This can be visualized more clearly in an example in Figure 2. The circles in orange denote individuals, the green rectangles affiliations and the red blocks at the bottom behaviors. 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, Catholic Church opposes smoking and abortion while Democratic Party supports abortion. Note that some affiliations might have no influence over certain behavior such as Democratic Party and Republican Party over smoking. The final behavior output of individuals depends on the affiliation regularization and individual associations. The first step of our

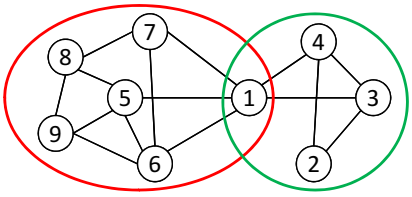

Figure 3: A Toy Example

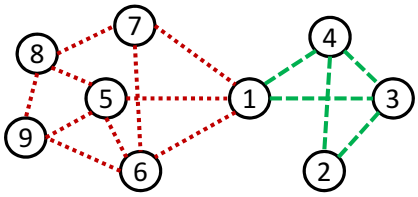

Figure 4: Edge Clusters

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

SocioDim framework is proposed to address the relation heterogeneity presented in social networks. Thus, a sensible method for social dimension extraction becomes critical to its success. Briefly, existing methods to extract social dimensions can be categorized into node-view and edge-view.

- **Node-view methods** concentrate on clustering nodes of a network into communities. As we have mentioned, the extraction of social dimensions boils down to a community detection task. The requirement is that one actor should be allowed to be assigned to multiple affiliations. Many existent community detection methods, with the aim of partitioning the nodes of a network into disjoint sets, do not satisfy this requirement. Instead, a soft clustering scheme is preferred. Hence, variants of spectral clustering, modularity maximization, non-negative matrix factorization or block models can be applied.

  One representative example of node-view methods is modularity maximization [6]. The top eigenvectors of a modularity matrix are used as the social dimensions in [8]. Suppose we are given a toy network as in Figure 3, of which there are 9 actors, with each circle representing one affiliation. For k affiliations, typically at least $k-1$ social dimensions are required. The top social dimension based on modularity maximization of the toy example is shown in Table 2. The actors of negative values belong to one affiliation, and actor 1 and those actors with positive values belonging 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 state explicitly about the association, but presents degree of associations for all actors.

- **Edge-view methods** concentrate on clustering edges of a network into communities. One representative edge-view method is proposed in [9]. The critical observation is that an edge resides in only one affiliation, though a node can be involved in multiple affiliations. For instance, in

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

Table 2: Social Dimensions of the Toy Example

Figure 3, actor 1 participates in both affiliations, but his connections are well separated, either in the red affiliation or in the green one. Hence, 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 (as shown in Figure 4). In the figure, the red edges represent one affiliation and the green ones denote the other. We can convert the resultant edge partition into social dimension representation as shown in Table 2. 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) in the green one. Thus, actor 1 has non-zero entries for both affiliations as shown in the table. Actor 4, on the contrary, with all its connections residing in the green set, has only one non-zero entry in its corresponding social dimension. This naturally leads to sparse social dimensions as shown in Table 2. By contrast, a node-view method like modularity maximization yields non-zero values for all the entries, resulting in dense representation.

In addition, 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. Owning to the power law distribution [5] presented in large-scale networks, a large portion of nodes in a network would bear a low degree. Hence, the resultant social dimensions would be sparse. To give a concrete example, we examine a YouTube network [9] with more than 1 million actors and verify the upperbound of the density. The YouTube network has $1,128,499$ nodes and $2,990,443$ edges. Suppose we want to extract $1,000$ dimensions from the network. Since 232 nodes have degrees larger than 1000, and the remaining nodes have degrees totaling $5,472,909$, the density of extracted social dimensions is

Table 3: Difference between Node-View and Edge-View Methods

	Objects under focus	Community Assignment
Node-View Methods	Nodes	multi-assignment
Edge-View Methods	Edges	single-assignment

upperbounded by $(5,472,909+232\times1,000)/(1,128,499\times1,000) = 0.51\%$. Based on our proposed edge clustering method in [9], the true density of the extracted social dimensions is $0.23\% < 0.51\%$.

Both node-view and edge-view methods can be applied to extract social dimensions. Table 3 lists their key differences. It is not obvious that one is better than the other. It depends on the network data, applications, and approaches being used. Next, we will show some empirical results to demonstrate the potential of the SocioDim framework.

Comparative Study

The SocioDim framework has many advantages over collective inference. Below we show some empirical results by studying behaviors on three representative social media sites. In particular, we crawl social networks in BlogCatalog[6], Flickr[7], and YouTube[8], respectively. BlogCatalog is a blog directory, Flickr is a popular photo sharing site and YouTube is well known as a video sharing platform. User interests or subscribed interest groups are deemed as behavior labels. F1, the harmonic mean of precision and recall[9], is employed to evaluate predictions. The average performance over a multitude of behaviors are reported at each site.

- **Reusable**. SocioDim is composed of two parts: community detection and supervised learning. Both are well-studied. Many algorithms have been developed and numerous existing software packages can be plugged in instantaneously, enabling code reuse and saving many human efforts for practical deployment.

- **Accurate**. By handling heterogeneity, SocioDim is suitable to be applied to online networks collected from social media to predict collective behavior. It has been shown to outperform collective inference considerably, especially when the social network is quite sparse and the behavior information of users is little [8]. For instance, Figure 5 shows the performance

[6]http://www.blogcatalog.com/

[7]http://www.flickr.com/

[8]http://www.youtube.com/

[9]Let $\mathbf{y}$, $\hat{\mathbf{y}} \in \{0,1\}^n$ denote the true labels and the predictions, respectively. Precision (P), Recall (R) and F1-measure (F1) are defined as $P = \frac{\sum_{i=1}^n y_i \hat{y}_i}{\sum_{i=1}^n \hat{y}_i}$, $R = \frac{\sum_{i=1}^n y_i \hat{y}_i}{\sum_{i=1}^n y_i}$, $F1 = \frac{2PR}{P+R}$.

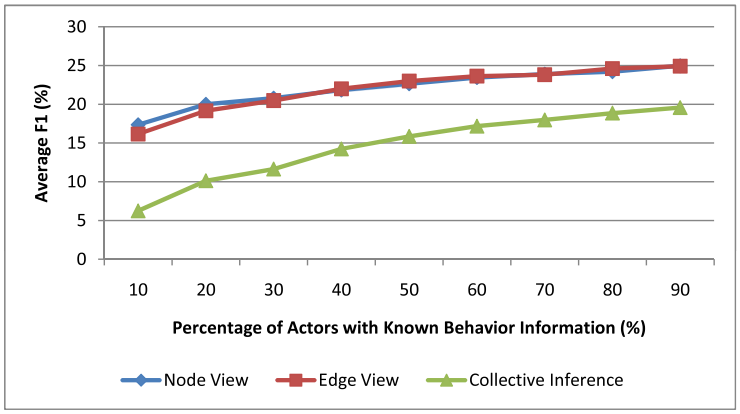

Figure 5: Performance on BlogCatalog network with 10, 312 actors [9]. Node View and Edge View denote SocioDim with modularity maximization [8] and proposed edge-centric clustering in [9], respectively for social dimension extraction. Collective Inference represents the wvRN method recommended in [2].

Table 4: Computation Time of SocioDim with Node-View Social Dimension Extraction versus Collective Inference on Flickr network of 80,513 actors, measured by seconds on Core2Duo 3G CPU. Collective inference does not show up in Pre-processing and Training as the method works like a lazy learner, which does not require training.

Percentage of Actors with Behavior Information		2%	4%	6%	8%	10%
Pre-processing	SocioDim	2857				
Training Time	SocioDim	46.5	91.5	134.1	162.6	211.2
Test Time	SocioDim	2.4	2.4	2.3	2.3	2.3
	CollectiveInf	1387	1084	740	588	470

of representative methods of node-view, edge-view and collective inference, respectively. SocioDim framework, with social dimension extraction either in node view or in edge view, outperforms collective inference substantially, indicating that differentiating connections between actors does help for behavior prediction.

- **Efficient**. A key difference of SocioDim framework from collective inference is that it is very efficient for prediction by trading off more time in network pre-processing and training as shown in Table 4. Collective inference typically requires many scans of the whole network for prediction while SocioDim accomplishes the task in one shot. SocioDim is more suitable for online applications as a majority of them emphasize more on prompt response for predictions.

Table 5: Scalability comparison of different methods for social dimension extraction [9]. Memory footprint is the size of the extracted social dimensions, and computation time refers to the time to compute social dimensions.

500 Social Dimensions	Extraction Methods	Flickr 80K actors 6M links	YouTube 1.1M actors 3M links
Memory Footprint	Node-View	322.1MB	4.6GB
	Edge-View	44.8MB	39.9MB
Computation Time	Node-View	40 mins	—
	Edge-View	3.6 hours	10mins

- **Scalable**. With a proper method to extract social dimensions, scalable instantiation of the framework can be developed in terms of both time and space complexity. For instance, with a normal PC, SocioDim with social dimension extraction in edge view, is able to handle a YouTube network of more than 1 million users in approximately 10 minutes and keep the extracted social dimensions extremely sparse, occupying only 40 megabyte memory space as shown in Table 5.

  On the other hand, when there is no memory constraint, node-view methods costs less time. This is observable on the Flickr data. The computation time of edge-view method on YouTube network is much smaller than on Flickr, because Flickr, though with fewer nodes, 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.

- **Generalizable**. The SocioDim framework essentially converts a network into features, offering a simple mechanism to seamlessly integrate two seemingly orthogonal information: social networks and actor features. It might be the case that some actor features (e.g., user profiles, blog content) are also available. These features can be combined with the extracted social dimensions before subsequent supervised learning. This kind of integration has been shown to boost the performance of relying on either type of information alone as shown in Figure 6.

Challenges and Future Work

In the previous sections, we have introduced the problem of collective behavior prediction, covered a social learning framework based on social dimensions, discussed two categories of methods for social dimension extraction and showed some potential advantages of the framework. However, many challenges are still there and need further research. Below, we elaborate some interesting directions.

Extraction of actor information: In social learning, the structural information of social networks alone is a weak indicator of user behavior. In all our

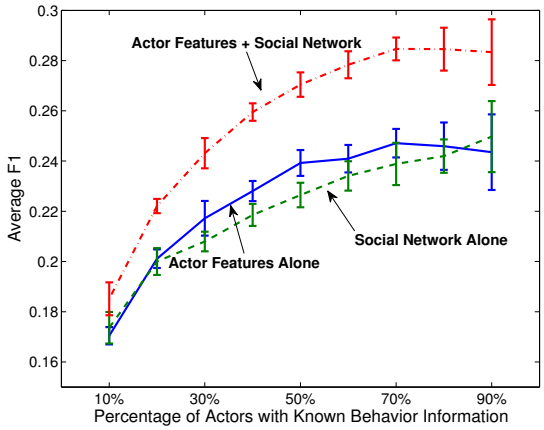

Figure 6: Averaged performance of network information with actor features on BlogCatalog data [8]. The social network is the blogger friendship network, and the actors features are the content of 5 most recent blog posts of bloggers. SocioDim framework provides a simple mechanism to combine social networks with actor features, leading to a substantial improvement over the performance of relying on either type of information alone [8].

experiment results, it is noticed that based on network information alone, the collective behavior prediction performance is far from satisfactory (with 20-30% F1-measure). But when combined with some other actors features like blog contents, the performance can be improved. Thus, a reliable collective behavior prediction system should include more information concerning actors. What kind of information can we collect to help behavior prediction? For instance, in Flickr and YouTube, users are allowed to upload tags and comments concerning shared resources. In Twitter, the tweets can be informative of potential user behavior. In certain cases, users are involved in 3-mode interactions like user-tag-resource. How can we extend the SocioDim framework to handle this kind of higher-order interactions? Is there any more effective method other than simple juxtaposition to integrate social dimensions and actor features?

Hybrid approach to social dimension extraction: In edge view methods, one fundamental assumption is that each edge belongs to only one affiliation. But it is well known that some weak ties are likely to bridge two different communities. We expect that a hybrid approach of node-view and edge-view methods might help locate those cross-community nodes and edges, so that a more meaningful social dimension representation can be extracted from a network. When a network becomes more and more heterogeneous, it becomes difficult to learn a clean social dimension representation from the network. Filtering out those irrelevant dimensions to the target behavior during extraction can also be important.

Efficient dynamic update: Networks in social media are evolving incessantly, 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 quite intriguing to consider temporal fluctuation into the problem of collective behavior prediction.

Scalability: We have shown that the edge-view method can handle mega-scale networks, but it is still memory-based. That is, everything is loaded into memory so the social dimension extraction can be finished efficiently. In reality, the network size could be so large that the network data cannot even be held in memory. It remains a challenging task to develop disk-based method to handle networks of extreme scale.

Model selection: The SocioDim framework requires users to specify the number of social dimensions to extract from a network. Our empirical experience indicated that the optimal dimensionality depends on the network size, network density, as well as number of users with known behavior. It would be practically useful to develop an automatic process such that the framework can determine the optimal number of social dimensions.

In this article, we examine how we can predict the online behavior of users in social media, given the behavior information of some actors in the network. Many social media tasks can be connected to the problem of collective behavior prediction. Since the connections in a social network represent various kinds of relations, a framework based on social dimensions is introduced. This framework suggests to 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. It demonstrates many potential amenities, and is especially suitable to be applied to large-scale networks, paving the way for collective behavior study in many real-world applications. We expect that along with this direction, more research work would emerge to address the many aforementioned challenges in the near future.

Related Work

The problem of collective behavior prediction is relevant to within-network classification [2], a classification problem when data instances are presented in a network format. In this case of social learning, the data instances are not independently identically distributed (i.i.d.) as in conventional data mining. To capture the correlation between labels of neighboring data instances, 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 its neighbors are fixed. This process is repeated until the label inconsistency between neighboring nodes is minimized. It is shown in [2]

that a simple weighted vote relational neighborhood classifier works reasonably well on some benchmark relational data and is recommended as a baseline for comparison.

Most relational classifiers, following the Markov assumption, capture the local dependency only. To handle the long-distance correlation, the latent group model [4], and the nonparametric infinite hidden relational model [10] assume Bayesian generative models such that the link (and actor attributes) are generated based on the actors' latent cluster membership. These models and social dimensions [8] pursue the same fundamental goal to capture the latent affiliations of actors. But the model intricacy and high computational cost for inference associated with the aforementioned models hinder their application to large-scale networks. Hence, Neville and Jensen [4] propose to use clustering algorithm to find the hard cluster membership of each actor first, and then fix the latent group variables for later inference. As each actor is assigned to only one latent affiliation, it does not capture the multitude of affiliation association required in social learning.

Acknowledgments

This work is, in part, supported by AFOSR and ONR. We thank the anonymous reviewers wholeheartedly for their expert opinions and constructive suggestions.

The Authors

Lei Tang is a PhD 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. He was awarded ASU GPSA Research Grant, SDM Doctoral Student Forum Fellowship, Student Travel Awards and Scholarships in various conferences. He is a member of ACM, IEEE and AIS. 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/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/web mining. He received his PhD in Computer Science at University of Southern California. His research has been sponsored by NSF, NASA, AFOSR, and ONR, among others. His well-cited publications include books, book chapters, encyclopedia entries as well as conference and journal papers. He serves on journal editorial boards and numerous conference program committees, and is a founding organizer of the International Workshop Series on

Social Computing, Behavioral Modeling, and Prediction (http://sbp.asu.edu/) in Phoenix, AZ (SBP08 and SBP09). His professional memberships include AAAI, ACM, ASEE, and IEEE. Contact him at Huan.Liu@asu.edu.

Contact Information

Lei Tang
Computer Science and Engineering
Arizona State University
PO Box 878809
Tempe, AZ 85287-8809
Phone: (480)727-7808
Fax: (480)965-2751
Email: L.Tang@asu.edu

Huan Liu
Computer Science and Engineering
Arizona State University
PO Box 878809
Tempe, AZ 85287-8809
Phone: (480)727-7349
Fax: (480)965-2751
Email: Huan.Liu@asu.edu

References

[1] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420, 1978.

[2] 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.

[3] 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.

[4] 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, New York, NY, USA, 2005. ACM.

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

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

[7] T. C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1:143–186, 1971.

[8] 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, New York, NY, USA, 2009. ACM.

[9] 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*, pages 1107–1116, New York, NY, USA, 2009. ACM.

[10] 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.