2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Network Denoising in Social Media

Huiji Gao*, Xufei Wang†, Jiliang Tang* and Huan Liu*
*Computer Science and Engineering
Arizona State University, Tempe, Arizona 85281
Email: {Huiji.Gao, Jiliang.Tang, Huan Liu}@asu.edu
†LinkedIn Corporation, Mountain View, CA 94043
Email: xwang@linkedin.com

***Abstract*—Social media expands the ways people communicate with each other. On a popular social media website, a user typically has hundreds of contacts (or friends) on average. As a person's social network grows, friend management is increasingly important for effective communications. Often, one can only afford to maintain close friendship in a small scale due to limited time and other resources. In other words, the majority of one's connections are so-so friends and do not hold strong influence on the user. One approach resorts to network denoising, by which unimportant connections are removed as noise. We study the challenges of network denoising in social media and how we can leverage a variety of social media information to denoise the links. We formulate the network denoising task as an optimization problem, and show the efficacy of our network denoising approach and its scalability experimentally in the domain of behavior inference.**

I. Introduction

Social media extends the physical boundary of user relationship to a new level. The total time spent on social media in the U.S. across PC and mobile devices has increased by 37 percent from 88 billion minutes in July 2011 to 121 billion minutes in July 2012 [1]. As reported by the Pew Internet Project in 2012, two-thirds of online adults use social network sites and 83% of teens and young adults are a member of at least one social network [2]. The increased participation in social media brings new challenges for managing friends and studying user behavior among others.

In the world of social media, the differences of time and location disappear, which allows one to have an inordinate number of friends. On average, a Facebook user has 190 friends [3], and a Twitter user has 208 followers[1]. When the circle of one's friendship grows, there is an increasing need for friend management or even "unfriending"[2]. Too many messages displayed on a user's wall prevent effective communication between a user and his close friends. Research by Robin Dunbar [4] indicates that 100 to 150 is the approximate natural group size in which everyone can really know each other, because "our minds are not designed to allow us to have more than a very limited number of people in our social world. The emotional and psychological investments requiring a close relationship are considerable, and the available emotional capital we have is limited."

Though one can have hundreds of online friends, most of them are so-called *"Facebook"* friends, as being a friend does not incur any social capital. When "Facebook" friends become rampant in social media sites, we face a new problem of information overloading[3]. A recent study shows that Twitter users have a very small number of friends compared to the number of followers and followees they declare [5]. In summary, the social network is comprised of valuable friends, casual friends, or event friends who are deemed to be treated differently.

In this paper, we propose to denoise an individual's social networks by *removing* noisy links. The potential benefits include the following. First, managing contacts differently according to the closeness to an individual. For example, the method can be applied automatically to group contacts; and we can apply different privacy levels to groups for information sharing. Second, reducing noisy links in social networks could benefit a wide range of applications, such as behavioral prediction, community detection, influence propagation, viral marketing, etc. However, the identification of noisy links in social media is a challenging task. In social media websites, a connection between two users provides limited information indicating the tie strength, and online information such as profiles can be incomplete. Third, offline behaviors (e.g., communication between two persons) are usually unavailable in social media.

In this work, we attempt to integrate multiple types of social interactions for denoising social networks. The rest of the paper is organized as follows. We define the problem in Section 2 and present the technical details in Section 3. The experimental evaluations are given in Section 4, followed by some related work in Section 5. We conclude the work and future directions in Section 6.

II. Problem Statement

A social network can be represented by a graph. Let $G = (U, E)$ be a social network, with $U = \{u_1, u_2, \ldots, u_n\}$ representing the set of n users, and E the set of connections (links) between users. The social network G is assumed to be undirected, and its adjacency matrix A is defined as

$$A_{ij} = \begin{cases} 1 & u_i \text{ and } u_j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Interaction refers to an activity between a user and another user or an item. We study three types of interactions in this paper: forming connections with other users (Linking),

[1]http://www.beevolve.com/twitter-statistics/#e1
[2]http://www.nytimes.com/2010/10/24/fashion/24Studied.html
[3]http://www.readwriteweb.com/archives/how_many_friends_is_too_many.php

ASONAM 2013, August 25-28, 2013, Niagara Falls, Canada
978-1-4799-1498-2/13/$31.00 ©2013 IEEE

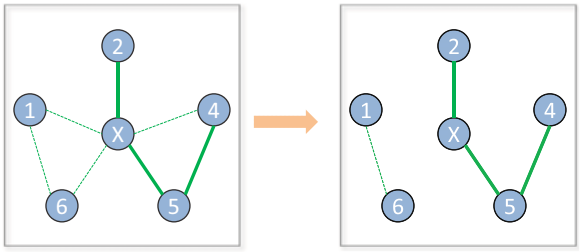


Fig. 1. Network Denoising in X's Neighborhood

subscribing to tags (Tagging), and making comments (Commenting). Linking information is contained in the adjacency matrix $A \in \{0,1\}^{n\times n}$. Tagging matrix $T \in \mathcal{R}^{t\times n}$ represents the subscription relationship between users and tags, where t is the number of unique tags. Commenting matrix $C \in \mathcal{R}^{n\times n}$ represents the number of comments that a user leaves on another user's wall or homepage.

The neighborhood of user u_i is his immediate social network. Thus, the size of the neighborhood is the node degree d_i of u_i. For each user, the i-th column of an interaction matrix corresponds to *user feature* f_i $(1 \le i \le n)$. For example, f_i can represent the neighbors u_i has, the tags u_i subscribes, and users who left comments on his wall, w.r.t Linking, Tagging, and Commenting. *Neighborhood feature* N_i $(1 \le i \le n)$ is a matrix whose j-th column $N_i(:,j)$ corresponds to the j-th neighbor of user i in terms of their *user features*.

The tie strength between two users is related to many factors, such as user similarity, user connectivity, etc. [6]. The homophily principle in social networks suggests that similarity breeds connections [7]. E.g., a higher degree of embeddedness (i.e., common friends) suggests stronger tie strength [8]; like-minded users tend to use similar tags (tagging) [9]. These factors and properties of social media networks can help remove unimportant or random friends. We formulate this network denoising problem as an optimization problem of tie strength estimation as follows:

$$\min_{\mathbf{w}_i \ge 0} \sum_{i=1}^{n} \left(\|N_i \mathbf{w}_i - f_i\|_2^2 + \lambda \|\mathbf{w}_i\|_1 \right), \tag{2}$$

where $\mathbf{w}_i \ge 0$ represents the weight vector (tie strength) between u_i and his social network, with each element $\mathbf{w}_i^j$ indicating the weight between u_i and his friend j. Intuitively, a larger weight between two users means they are more close or they share more similarity. Therefore, $\mathbf{w}_i$ can be utilized for denoising u_i's social network. In this work, we remove the link between a user and his friend for denoising, as long as the learned corresponding weight between them is 0. λ is used to balance the trade-off between the sparsity of tie strength vector and the accuracy of approximation.

The intuitive interpretation of network denoising can be illustrated in Figure 1. It demonstrates the neighborhood of user X before (Left) and after (Right) denoising. In both graphs, solid lines and dashed lines represent strong and weak ties between the users and his social network. After denoising, the three dashed connections are removed, while the two strong ties are kept.

Existing methods of ranking edges include Edge Centrality, PageRank for edge selection [10], etc. However, these methods are inappropriate for network denoising because (1) they are computed on the whole graph, but denoising is essentially a local processing task involving user u_i and his neighbors; (2) a local method is more agile at handling local changes such as localized updating as a social network evolves: users join and leave, links form and dissolve. (3) A global method is sensitive to the structural variation of a graph. For example, removing a small number of links in the graph may significantly affect the global ranking. The proposed approach operates locally, avoiding unnecessary updates. (4) Edge ranking tends to change the structure of a social network after removing lowly ranked edges, resulting in singleton users. The proposed approach removes one's noisy links to compact his social network.

III. Methodology

The tie strength estimation defined in Eq (2) solves a quadratic convex problem with constraints. There are sophisticated methods of solving the constrained least square problems [11], [12], [13]. Solving Eq. (2) can be achieved by solving a series of sub-problems:

$$\begin{aligned} &\min \ \|N_i \mathbf{w}_i - f_i\|_2^2 + \lambda \|\mathbf{w}_i\|_1, \\ &s.t. \ \mathbf{w}_i \ge 0 \end{aligned} \tag{3}$$

There are several efficient solvers for above problem [14], [15], [16], [17]. In the following experiments, we use the state-of-the-art solver developed by Liu et al. [17].

Different types of interactions, e.g., sharing common friends, using similar tags, and commenting can improve the closeness between two users; however, it may contribute differently to the tie strengths between two users. Information obtained from one type of interaction is often incomplete or noisy, such as one's preference in making friends over using tags and commenting. In social media, multiple interactions can be complementary in clarifying user relationships. A novelty of our proposed solution is to enable the integration of different interactions efficiently.

A. Integrating Multiple Interactions

For a given user u_i, we first consider two types of interactions and demonstrate how to integrate them. Let N_i^1 and N_i^2 be the two *neighborhood features* corresponding to two interactions, f_i^1 and f_i^2 be the *user features*, and introducing an all ones vector $e = (1, 1, \ldots, 1)^\top$, the tie strength estimation problem can be rewritten,

$$\min_{\mathbf{w}_i \ge 0} \sum_{i=1}^{n} \left(\alpha_1^2 \|N_i^1 \mathbf{w}_i - f_i^1\|_2^2 + \alpha_2^2 \|N_i^2 \mathbf{w}_i - f_i^2\|_2^2 + \lambda e^\top \mathbf{w}_i \right) \tag{4}$$

Each sub-problem of integrating two interactions can be shown to be equivalent to solve a new constraint least square problem,

$$\begin{aligned} &\alpha_1^2 (N_i^1 \mathbf{w}_i - f_i^1)^\top (N_i^1 \mathbf{w}_i - f_i^1) + \\ &\alpha_2^2 (N_i^2 \mathbf{w}_i - f_i^2)^\top (N_i^2 \mathbf{w}_i - f_i^2) + \lambda e^\top \mathbf{w}_i \\ &= \left\| \begin{pmatrix} \alpha_1 N_i^1 \\ \alpha_2 N_i^2 \end{pmatrix} \mathbf{w}_i - \begin{pmatrix} \alpha_1 f_i^1 \\ \alpha_2 f_i^2 \end{pmatrix} \right\|_2^2 + \lambda e^\top \mathbf{w}_i, \end{aligned} \tag{5}$$

where α_1 and α_2 weighs the importance between the two interactions. Combining all users together, Eq. (4) can be

rewritten as

$$\min_{\mathbf{w}_i \geq 0} \sum_{i=1}^{n} \left(\left\| \begin{pmatrix} \alpha_1 N_i^1 \\ \alpha_2 N_i^2 \end{pmatrix} \mathbf{w}_i - \begin{pmatrix} \alpha_1 f_i^1 \\ \alpha_2 f_i^2 \end{pmatrix} \right\|_2^2 + \lambda e^\top \mathbf{w}_i \right) \tag{6}$$

Eq. (6) can be generalized to include multiple interactions,

$$\min_{\mathbf{w}_i \geq 0} \sum_{i=1}^{n} \left(\|\mathcal{N}_i \mathbf{w}_i - \mathcal{F}_i\|_2^2 + \lambda e^\top \mathbf{w}_i \right), \tag{7}$$

where $\mathcal{N}_i$ and $\mathcal{F}_i$ are obtained by stacking ℓ interactions together, and are given by

$$\begin{aligned} \mathcal{N}_i &= (\alpha_1 {N_i^1}^\top, \alpha_2 {N_i^2}^\top, \ldots, \alpha_\ell {N_i^\ell}^\top)^\top \\ \mathcal{F}_i &= (\alpha_1 {f_i^1}^\top, \alpha_2 {f_i^2}^\top, \ldots, \alpha_\ell {f_i^\ell}^\top)^\top \end{aligned} \tag{8}$$

The weighting parameters (α_1, α_2, …, α_ℓ) can be determined by prior knowledge or cross-validation. However, in this paper, we do not differentiate the importance of different interactions and set them to 1.

Integrating multiple interactions can be interpreted as stacking the neighborhood features and user features, and separately solving a larger least square problem with non-negative and sparsity constraints.

B. Time Complexity Analysis

If the matrices are not sparse, the time complexity of the constrained least square problem defined in Eq. (3) is $O(c\ell_i d_i)$, where c is the largest eigenvalue of matrix $N_i N_i^\top$, and ℓ_i and d_i are the dimensions of matrix N_i [17]. Due to the sparsity of interaction matrices studied in this paper, the time complexity becomes $O(\mu_i)$, where μ_i is the number of non-zero entries in matrix N_i and it has the same order with dimension d_i in social networks [18]. It has been observed that online social networks typically follow a power law degree distribution,

$$p(x) = (1-\alpha)x^{-\alpha}, x \geq x_{min} \geq 1 \tag{9}$$

where α is the exponent of the distribution and its value often falls between 2 and 3 in small world networks [19].

$$\begin{aligned} &\sum_{i=1}^{n} d_i \\ &\approx n \int_{x_{min}}^{\infty} x p(x) dx \\ &= n \cdot \int_{x_{min}}^{\infty} (1-\alpha) x^{1-\alpha} dx \\ &= n \cdot \frac{\alpha - 1}{\alpha - 2} \cdot x_{min}^{2-\alpha} \end{aligned} \tag{10}$$

Discarding constant terms, the time complexity for solving Eq. (2) is $O(n)$. Considering ℓ types of interactions, the total time complexity is given by $O(n\ell)$, which is thus linear with respect to the number of users in a social network.

TABLE I. STATISTICS OF DIFFERENT NETWORK DATA

Dataset	BlogCatalog	Flickr	BlogMI
Users	8,797	8,465	6,069
Edges	290,059	195,847	523,642
Unique Tags	7,418	7,303	5,161
Classes	59	169	70
Density	7.5×10^{-3}	5.5×10^{-3}	2.8×10^{-2}
Avg. Degree	66	46	173

IV. EXPERIMENTAL EVALUATION

We now perform experiments to verify if the proposed approach can achieve its designed purposes. We use behavior inference as evaluation task. We want to see if, after denoising, (1) we can maintain or improve the performance, (2) make the social media networks more compact, and (3) accomplish the same behavior inference task faster. We first introduce the social media datasets and evaluation approach, and then embark on effectiveness and efficiency studies, finally discuss the limitations of denoising.

A. Social Media Datasets

Three datasets from popular social media websites are employed: BlogCatalog, Flickr, and BlogCatalog Multi-Interaction (BlogMI), each containing more than one type of interaction. The first two datasets contain linking and tagging information and are obtained from Tang et al. [20]. The third dataset is crawled to include more interactions. Table I shows the statistics of the three datasets, with detailed descriptions presented below:

BlogCatalog is a blog directory where users can register their blogs under predefined categories. When a new blog is registered, the owner is asked to specify the major category and a sub category in a hierarchical structure, and specify several tags to describe the main topics of the blog. It contains linking and tagging information with $8,797$ users and $7,418$ tags.

Flickr is an image sharing website in which users can specify tags for each photo they upload. Different from BlogCatalog, users can join various groups (e.g. sports clubs, special interest groups, etc) on Flickr. The dataset has $8,465$ users and $7,303$ unique tags with both linking and tagging information. The connections on Flickr are directional and we simply ignore the direction of the edges, i.e., two users are connected if there is a link between them.

BlogMI is crawled to include multiple interactions between users: linking, tagging, and commenting. Commenting refers to when a user leaves comments on another user's wall or homepage. Tags used by less than 10 persons and users who have no tag usage are excluded, we finally obtained a dataset with $6,069$ users and $5,161$ unique tags.

B. Evaluation Approach

Since there is no ground truth about links being noise or not, we verify indirectly: comparing the performance of behavioral inference *before and after* denoising. The hypothesis is that removing unimportant or irrelevant links will not affect the performance of inferring user behaviors. Intuitively, if we can successfully achieve network denoising, we should have similar, if not better, performance with more compact network.

Since the category information on BlogCatalog or user groups on Flickr imply their interests (behaviors as class labels), they are used as ground-truth labels. We first generate two user-user networks based on user friendships. The first one is constructed with the original user friendship links, and the second one is constructed using the links after denoising by removing the link between two users whose corresponding weight $\mathbf{w}_i^j$ is 0. We then extract latent social dimensions [18] for each user on the two networks, respectively. The latent social dimensions are represented as a sparse vector for each user. An entry of the vector is 1 if the user belongs to the corresponding social dimension, 0 otherwise. With the social dimension vectors as features and ground-truth class labels, we validate the performance on each user-user network in a supervised learning fashion, adopting F-Measure - a harmonious mean of precision p and recall r:

$$\text{F-Measure} = \frac{2pr}{p+r} \tag{11}$$

Non-negative Matrix Factorization (NMF) is utilized in experiments to compute social dimensions for simplicity. Since NMF usually converges to local minima, we repeat the experiments 10 times and report the average F-Measure.

C. Behavioral Inference

In this work, to infer a behavior is considered as a classification task. We use a certain percentage (e.g., 10% - 90%) of users (with labels) to train a Linear SVM, and the rest of the users to test. Since only **BlogMI** contains commenting information, for comparison purpose, we first study the use of linking and tagging information for network denoising on the three datasets, and then investigate denoising with multiple interactions (linking, tagging and commenting) on **BlogMI**.

1) Denoising vs No Denoising: Both linking and tagging information can be used for denoising separately and aggregately on the studied datasets. We present the performance w.r.t. different interactions in Figure 2 on dataset BlogCatalog (Left), Flickr (Center), and BlogMI (Right), respectively.

In each figure, the x-axis represents the fraction of users that are used to train and the y-axis represents the F-Measure. Four curves correspond to different denoising schemes: *No Denoising (NoD)* (red, circle), *Denoising with Linking (DLink)* (blue, square), *Denoising with Tagging (DTag)* (cyan, hexagon), and *Denoising with Linking and Tagging (DLinkTag)* (black, diamond).

There are several interesting observations: (1) *DLinkTag* usually achieves the best or no worse performance than other methods in terms of F-Measure; (2) The performance of *DLink* and *DLinkTag* are close in most cases; (3) *DTag* has a big variation across the three datasets; (4) Except *DTag*, the inference performance with denoising is better than that without denoising. The average relative improvement of different denoising schemes compared to *NoD* is shown in Table II. We obtain consistent improvement by applying *DLinkTag* and *DLink* to the studied social networks.

The improved performance in terms of F-Measure confirms our hypothesis that noisy links do exist in the original graphs. Denoising might be useful in tasks related to linkages, such as behavioral inference as shown above. The varied performance improvement ratios suggest that the studied datasets have different noisy levels. Intuitively, the BlogMI dataset is most likely to have more noisy links because of the high nodal degrees, followed by BlogCatalog and Flickr. On the other hand, it is intuitive that denoising is more effective on social networks with a large amount of noisy links. For example, we obtain a larger improvement ratio on BlogMI than BlogCatalog and Flickr, according to Table II.

TABLE II. RELATIVE IMPROVEMENT RATIO OF DIFFERENT DENOISING SCHEMES

Dataset	BlogCatalog	Flickr	BlogMI
DLinkTag (%)	0.91	6.71	16.29
DLink (%)	0.99	6.04	14.94
DTag (%)	-1.55	6.27	7.06

Interestingly, *DTag* is not as effective as *DLink*. Like categories and groups, tags also imply user interests to some extent. The reason that *DTag* does not perform as well as *DLink* is that tagging may impose too strong a constraint on measuring tie strength. Thus, too many links are treated as noisy links, which harms latent social dimension extraction. On the other hand, the performance of *DLinkTag* is slightly better than that of *DLink*, suggesting that linking and tagging information are complementary to some extent.

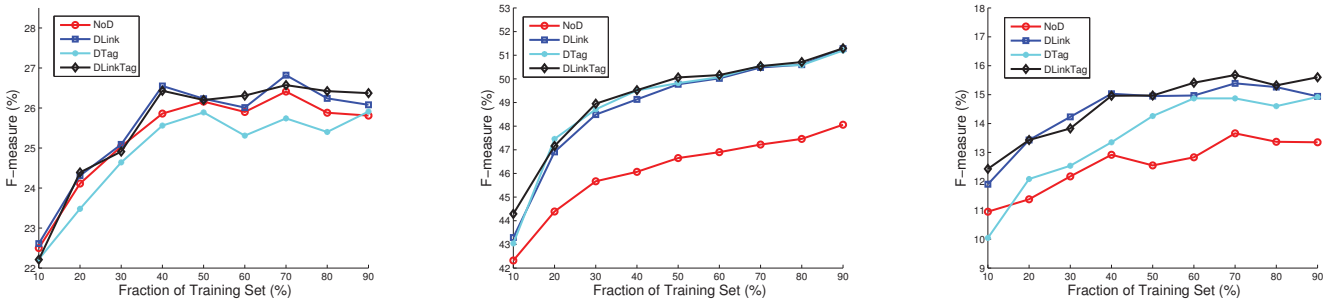
2) Denoising Performance vs λ: The number of links to be treated as noise is controlled by the user specified parameter λ. A larger λ is more likely to keep fewer links. The results based on *DLinkTag* are presented in Figure 3.

In these figures, λ is set to 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, and 0.1 to 1 with an increment of 0.1. Only some of them are presented in the figures. We have several interesting findings. First, on all three datasets, the inference performance increases before peaking, then decreases when λ increases from 10^{-5} to 1. However, on BlogCatalog, the F-measure corresponding to different λs is closer than the other two datasets. Second, the best performance is achieved when λ equals 0.1 and 0.2 on Flickr and BlogMI consistently. Third, the performance between different λs does not differentiate significantly. The reasons will be made clearer in the following sections.

3) Denoising with Multiple Interactions: We study multiple interactions and their impact on behavioral inference. The performance of three interactions and their combinations on BlogMI are presented in Table III. λ is set to 0.2.

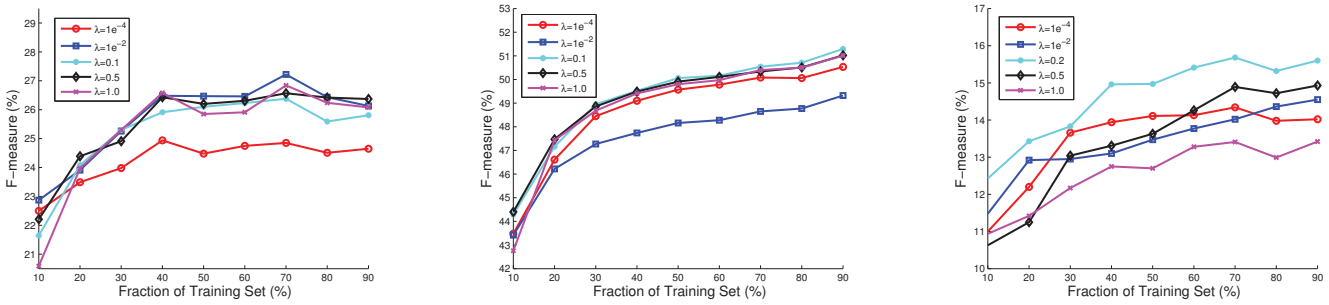
Applying denoising often improves F-Measure. Applying linking, tagging, or commenting and their combinations to denoising consistently improves the performance. However, more interactions do not necessary imply better performance. Denoising with Linking and Tagging performs best in most cases, whereas integrating all three interactions performs reasonably well but not the best on the dataset. On average, integrating two interactions gains 2.74% more than utilizing only one of the interactions, whereas integrating all three interactions is almost equivalent to the peformance of utilizing only one of the interactions.

4) Link Reduction Analysis: We study the link reduction rate for each dataset and show the correlation between the rates and regularization parameter λ. Figure 4 shows the denoised network by *DLinkTag*. The x-axis represents the regularizer λ



(a) BlogCatalog (b) Flickr (c) BlogMI

Fig. 2. Behavioral Inference Performance with Different Denoising Schemes



(a) BlogCatalog (b) Flickr (c) BlogMI

Fig. 3. Behavioral Inference Performance w.r.t. λ

TABLE III. F-MEASURE PERFORMANCE ON BLOGMI DATASET

Proportion of Labeled Users	10%	20%	30%	40%	50%	60%	70%	80%	90%
Linking + Tagging	**12.43**	13.43	13.83	14.96	**14.97**	**15.41**	**15.68**	15.32	**15.60**
No Denoising	10.95	11.38	12.17	12.92	12.55	12.83	13.66	13.37	13.35
Linking	11.90	13.43	**14.23**	**15.03**	14.95	14.97	15.39	15.26	14.94
Tagging	10.04	12.08	12.54	13.35	14.26	14.87	14.87	14.60	14.92
Commenting	11.20	12.85	13.19	14.52	14.50	14.93	15.31	15.12	15.36
Tagging + Commenting	11.62	11.71	13.32	13.80	14.87	15.19	15.46	**15.36**	15.36
Linking + Commenting	12.37	**13.78**	13.94	14.40	14.70	14.86	15.33	15.11	15.49
Linking + Tagging + Commenting	12.20	13.22	13.83	14.22	14.23	14.73	15.16	15.12	14.81

and the y-axis represents the reduction rate. As λ increases from 0.1 to 1.0, more links are expected to be removed, but the reduction rates do not decrease significantly. The result confirms that the selection of λ is not sensitive in the studied datasets. For example, the link reduction rates are around 60%, 37%, and 70% on BlogCatalog, Flickr and BlogMI when λ varies, respectively.

More interestingly, when we take a closer look at the reduction ratios of the networks, even when the regularization term is set to very small (e.g. 10^{-5}), we obtain similar reduction ratio. This suggests that (1) a large portion of users are connected loosely; these users could be so-so friends or they do not interact with each other. (2) In the studied social networks, noisy links are very easy to remove even when the regularization penalty is very small. The results suggest that tuning the regularization parameter λ is not imperative.

Another observation is that the Flickr dataset has a higher link keep ratio than BlogCatalog. It seems that Flickr users share more similarities with their social networks. First, on average, a Flickr user has 178 tags, whereas the BlogCatalog and BlogMI users only have 8 tags. Second, on average,

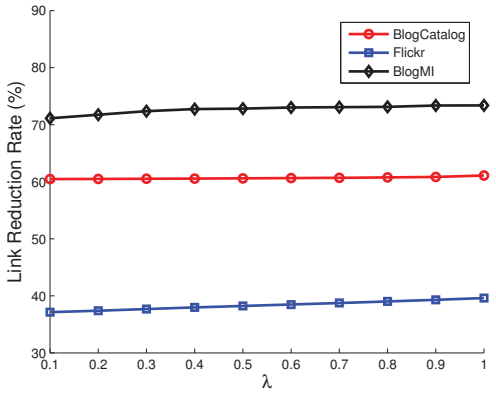


Fig. 4. Link Reduction Rate w.r.t. λ

a Flickr user shares 24.43 tags with his neighbor, but the BlogCatalog and BlogMI users only share 0.27 and 0.24 tags, respectively. Third, Flickr users have fewer friends compared to the users on the other two datasets. It is natural to assume that more friends means more noisy links in a graph, because the number of friends one can have is often upper limited. The

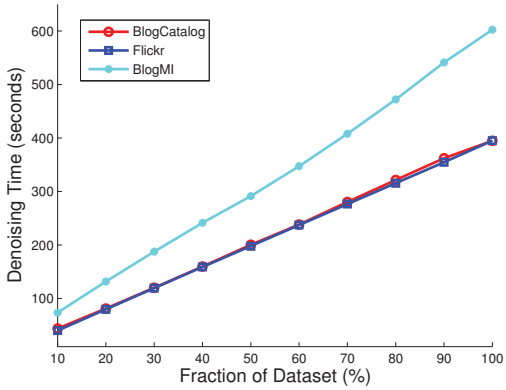


Fig. 5. Scalability of Network Denoising

reduction rates verify above hypothesis.

5) Advantages of Denoising Social Networks: The benefits of denoising social networks are shown empirically. We select the λs that correspond to peak performance on the three datasets and demonstrate statistics of link reduction ratio, relative improvement ratios, and computational time reduction for extracting latent social dimensions. The results are obtained when selecting 90% of the users as training dataset, and 10% of users as test.

First, as shown in Table IV, we gain improvement of inference performance consistently on the three datasets with various denoising schemes. This suggests that denoised social networks are more homogeneous: users have a higher similarity with their social networks.

Second, the links that are removed are typically huge in the studied social networks, especially on BlogCatalog and BlogMI datasets irrespective of denoising schemes. The denoised networks are more compact; thus, further tasks such as data analysis, storage, and visualization can benefit. However, there is a potential risk in the removal of too many connections (over-denoising). For example, denoising with Tagging only keeps 14% of links on BlogCatalog.

Third, the computational time for behavioral inference is reduced due to smaller sized social networks. The time is reduced to one third on BlogCatalog and BlogMI datasets, obtaining no worse performance than behavioral inference without denoising.

D. Scalability

Theoretical analysis shows that the time complexity of network denoising is linear with respect to the size of social networks. We verify it empirically on the studied datasets.

Given a specific ratio (e.g. 50%), we randomly select users from the whole dataset and record the time for network denoising. The process is repeated 10 times and the average elapsed time is reported. As shown in Figure 5, the time spent is increasing linearly when more users are denoised. The time spent on BlogMI is longer because the average node degree in this dataset is larger than the other two. The time required to denoise BlogCatalog and Flickr are very close.

The linear time complexity shows that our method can be scaled up to deal with large scale real world social networks.

E. Statistics before and after Denoising

To verify whether the graph structure is maintained after denoising, we plot the degree distribution for the studied datasets. Figures 6 and 7 demonstrate the degree distribution *before* and *after* denoising the social networks, respectively. In these graphs, the x-axis represents the nodal degree and the y-axis shows the number of users who have the degree values indicated by the x-axis. Apparently, the degree distribution of the corresponding networks are quite similar. A closer look at the denoised networks shows that the singletons only account for 2.9%, 3.3% and 0.02% on BlogCatalog, Flickr, and BlogMI, respectively. The number of singletons increased from 32 to 252, 211 to 279, and 0 to 1 in the three datasets, respectively. This shows that users with a small number of connections are not significantly penalized by applying the denoising procedure. The results reveal that the structure of the social networks are maintained: the majority of users are still connected in a large component and they still connect to their close friends.

The statistics of the networks are presented in Table V. We compute the average nodal degrees, clustering coefficient, and average sharing tags on the networks with and without denoising. As expected, after denoising, the average number of neighbors are reduced to 26, 29, and 49 on BlogCatalog, Flickr, and BlogMI, respectively. The clustering coefficient is also dropped, which suggests that though a certain number of triads are broken, there are still a large number of triads remaining. This result suggests that triads do not imply strong connections. On average, the shared tags between a user and his neighbor is increased 0.13, 4.52, and 0.15.

F. Discussion of limitations on denoising

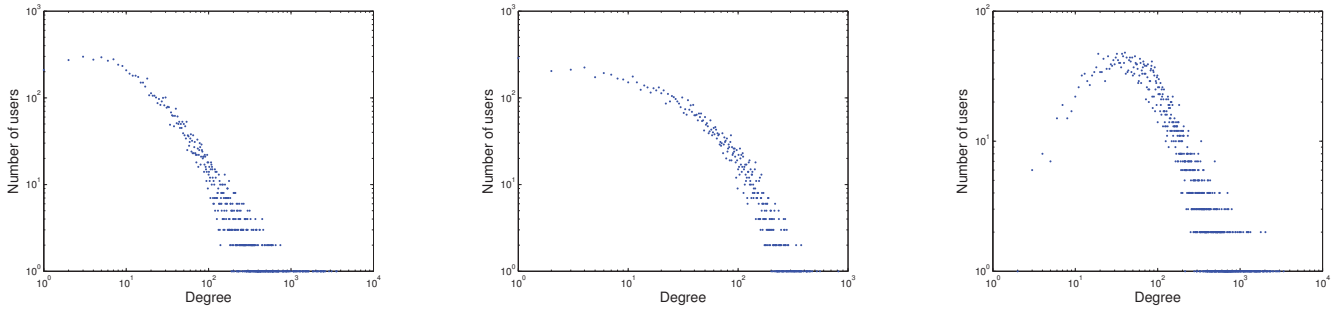
So far we have investigated the effect of denoising on behavioral inference. A more compact network that is easy to organize after denoising could help improve both time-efficiency and performance for certain tasks. However, network denoising is task-oriented. On certain tasks, denoising may not be able to improve the performance, and sometimes may even reduce the performance due to its limitations. One of the major limitations on denoising is "**loss of negative information**". In this work, the "noise user" related to a target user is considered as a user who holds irrelevant or opposite interests to the target user regarding our behavioral inference task. Filtering such users provides a purer environment that could better present homophily effect, which is helpful in behavioral inference. However, correspondingly, negative information (links that connect users with different interests) are lost. This could be analog to a user-user network with trust information (positive information) only while without the observation of distrust information (negative information). Since negative information is important for recommender systems [21], denoising may reduce the performance of tasks like item recommendation.

V. Related Work

Kivran-Swaine et al. [22] studied the impact of structure properties on breaking ties on Twitter. They found that tie strength, social status, and embeddedness are the key factors that influence the breaking of ties. Dyadic reciprocity is an indicator of a strong tie, though it is not clear when more than

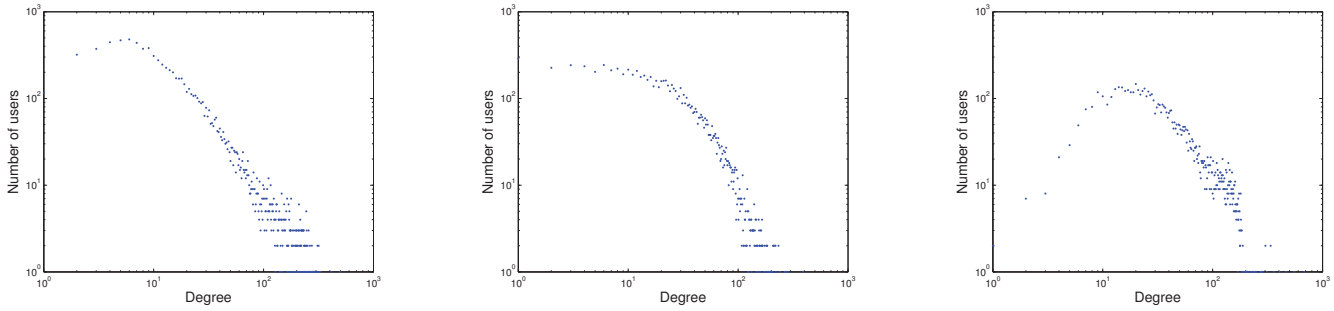
TABLE IV. ADVANTAGES OF DENOISING SOCIAL NETWORKS WHEN 90% OF USERS ARE USED IN TRAINING

Datasets	Methods	F-Measure (%)	Link Reduction	Time (second)
BlogCatalog	No Denoising	25.81	290,059	101.92
	Linking	26.08 (+0.7%)	112,559 (-61.19%)	36.43 (-65.49)
	Tagging	25.91 (+0.4%)	39,775 (-86.29%)	47.44 (-54.48)
	Linking + Tagging	26.37 (+2.2%)	114,337 (-60.58%)	46.65 (-55.27)
Flickr	No Denoising	48.06	195,847	85.62
	Linking	51.28 (+7%)	96,442 (-50.76%)	64.67 (-20.95)
	Tagging	51.20 (+6.53%)	107,247 (-44.24%)	66.75 (-18.87)
	Linking + Tagging	51.29 (+6.72%)	123,130 (-37.13%)	67.07 (-18.55)
BlogMI	No Denoising	13.35	523,642	106.1
	Linking	14.94 (+11.91%)	137,261 (-73.79%)	32.34 (-73.76)
	Tagging	14.92 (+11.76%)	38,575 (-92.63%)	24.26 (-81.84)
	Commenting	15.36 (+15.06%)	49,139 (-90.62%)	54.13 (-51.97)
	Linking + Tagging	15.60 (+16.85%)	147,881 (-71.76%)	35.96 (-70.14)
	Tagging + Commenting	15.36 (+15.06%)	75,027 (-85.67%)	46.68 (-59.42)
	Linking + Commenting	15.49 (+16.03%)	155,004 (-70.40%)	28.77 (-77.33)
	Linking + Tagging + Commenting	14.81 (+10.94%)	141,200 (-73.04%)	32.47 (-73.63)



(a) BlogCatalog (b) Flickr (c) BlogMI

Fig. 6. Degree Distribution before Denoising



(a) BlogCatalog (b) Flickr (c) BlogMI

Fig. 7. Degree Distribution after Denoising

two users are involved. Ties between individuals in denser networks are more likely to be strong; the more common friends a dyad shares, the less likely the edges to be broken. Sibona et al. studied various reasons of unfriending on Facebook and they concluded that "people are most likely to unfriend those who post mundane or inflammatory status updates".

Tie strength refers to the degree of closeness of a relationship. Researchers often consider two types: *strong ties* (close friends) and *weak ties* (acquaintances) [23]. Strong ties tend to exhibit higher similarity between the two subjects and weak ties are passages for conveying novel information [24]. Tie strength can be learned by supervised and unsupervised methods. Gilbert and Karahalios [25] propose to predict the tie strength by linear regression. They found intimacy, intensity, duration, and social distance are the most important factors determining the strength of a relationship. Kahanda and Neville [26] propose to estimate link strength by learning a predictive model leveraging transactional information and show that such features are most influential in estimating tie strength. Xiang et al. [6] present an unsupervised latent variable model to estimate the link strength. The basic assumption is that the strength of a relationship affects the interactions between users: they show improved performance in classification tasks. Although the weight vector defined in Eq. (2) is interpreted as tie strength, the paper differs from former work: (1) the problem studied is different; (2) the motivations and formalizations are also different.

Link prediction is the task of inferring a future connection given the network at current time stamp. Liben-Nowell and Kleinberg [27] formalize the problem and the study on co-

TABLE V. NETWORK STATISTICS BEFORE AND AFTER DENOISING

Measures	Ave. Degree		Clustering Coefficient		Ave. Sharing Tags	
	No Denoising	Denoising	No Denoising	Denoising	No Denoising	Denoising
BlogCatalog	66	26	0.46	0.20	0.27	0.40
Flickr	46	29	0.13	0.12	24.43	28.95
BlogMI	173	49	0.39	0.12	0.24	0.39

authorship networks shows that the future interactions between users can be extracted from network topology alone. Besides the network structure, supervised learning methods leveraging proximity features such as distances, similarity, etc., can produce better performance in social networks and web pages [28], [29]. Scellato et al. [30] applied supervised link prediction methodologies to location-based social networks for friend recommendation with a set of location-based features. Link prediction attempts to *introduce* new links in the future, whereas the task of this work is to *remove* unlikely links in current time stamp. It could be interesting to see how our work can affect the performance of link prediction as future work.

VI. CONCLUSION

Social media allows users to connect to an extraordinary amount of online friends. However, as the user's social network expands, the need for friend management intensifies. In this work, we propose an efficient approach to denoise social networks for friend management by utilizing multiple types of interactions in social media. The advantages of denoising social networks are verified via the performance of behavioral inference, link reduction, and time efficiency. Network denoising not only improves performance but also make social networks more compact for efficient social media mining. Interesting future work can be explored further.As social media grows, more types of interactions will be made available. At least two lines of research can be pursued: integrating all vs. selectively integrating. If we can access two social networking sites at the same time, it is challenging to see if we can take advantage of their different but complementary information.

ACKNOWLEDGMENT

This work is supported, in part, by ONR (N000141010091).

REFERENCES

[1] Nielsen, "State of the media: The social media report 2012," 2012.

[2] J. B. Maeve Duggan, "The demographics of social media users," *Pew Internet & American Life Project*, 2012.

[3] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011.

[4] R. Dunbar, *How Many Friends Does One Person Need? Dunbar's Number and Other Evolutionary Quirks*. Faber and Faber,, 2010.

[5] B. A. Huberman, D. M. Romero, and F. Wu, "social networks that matter twitter under the microscope," *First Monday*, vol. 14, no. 1, 2009.

[6] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[7] L. S.-L. Miller McPherson and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[8] J. L. Martin and K.-T. Yeung, "Persistence of close personal ties over a 12-year period," *Social Networks*, vol. 28, no. 4, pp. 331 – 362, 2006.

[9] X. Wang, L. Tang, H. Gao, and H. Liu, "Discovering overlapping groups in social media," in *the 10th IEEE International Conference on Data Mining series (ICDM 2010)*, Sydney, Australia, December 14 - 17 2010.

[10] Balázs, R. M. J. Csanád Csáji, and V. D. Blondel, "Pagerank optimization by edge selection," *CoRR*, vol. abs/0911.2280, 2009.

[11] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. John Wiley & Sons, 2006.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[13] S. M. Stefanov, "Convex quadratic minimization subject to a linear constraints and box constraints," *Applied Mathematics Research Express*, vol. 18, no. 1, pp. 27 – 48, 2004.

[14] P. H. Calamai, "Projected gradient methods for linearly constrained problems," *Mathematical Programming*, vol. 39, pp. 93 – 116, 1987.

[15] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606 – 617, 2007.

[16] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756 – 2779, 2007.

[17] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP

[18] L. Tang and H. Liu, "Scalable learning of collective behavior based on sparse social dimensions," in *CIKM'09: Proceeding of the 18th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2009, pp. 1107–1116.

[19] M. E. J. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary Physics*, vol. 46, pp. 323–351, 2005.

[20] 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*, 2009, p. 817826.

[21] H. Ma, M. R. Lyu, and I. King, "Learning to recommend with trust and distrust relationships," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 189–196.

[22] F. Kivran-Swaine, P. Govindan, and M. Naaman, "The impact of network structure on breaking ties in online social networks: Unfollowing on twitter," in *CHI*, 2011.

[23] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010, ch. Strong and Weak Ties.

[24] M. Granovetter, "The strength of weak ties," *American Journal of Socialogy*, vol. 78, no. 6, pp. 1360–1380, May 1973.

[25] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the 27th international conference on Human factors in computing systems*, 2009.

[26] I. Kahanda and J. Neville, "Using transactional information to predict link strength in onine social networks," in *Proceedings of Third International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

[27] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of 12th International Conference on Information and Knowledge Management*, 2003.

[28] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications*, 2005.

[29] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in *Advances in Neural Information Processing Systems*, 2003.

[30] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1046–1054.