Toward Relational Learning with Misinformation

Liang Wu* Jundong Li* Fred Morstatter† Huan Liu*

Abstract

Relational learning has been proposed to cope with the interdependency among linked instances in a network, and it is a fundamental tool to categorize social network users for various tasks. However, the emerging widespread of misinformation in social networks, information that is inaccurate or false, poses novel challenges to utilizing social media data. Malicious users may actively manipulate their content and characteristics, which easily lead to a noisy dataset. Hence, it is intricate for traditional relational learning approaches to deliver an accurate predictive model in the presence of misinformation. In this work, we precisely focus on the problem by proposing a joint framework that simultaneously constructs a relational learning model and mitigates the effect of misinformation by restraining anomalous points. Empirical results on real-world social media data prove the superiority of the proposed approach, Relational Learning with Misinformation (RLM), over traditional approaches on modeling social network users.

1 Introduction

Relational learning (RL) utilizes relationships between instances manifested in a network to improve the predictive performance of various network mining tasks. The triumphant applications of RL have been witnessed in a myriad of domains, such as social networks (*e.g.*, Flicker), language networks (*e.g.*, Wikipedia), and citation networks (*e.g.*, DBLP). The vast amount of social media content, ranging from daily chatter, conversations to information sharing and news reports, together with automatic modeling of the content information, allow for an insight into the public opinion that has been utilized for recommender systems [8], targeted advertising [20], and even predicting the stock price [2] and election results [21].

However, due to emerging challenges brought by malicious social media users, it is increasingly risky to depend on social media data for decision making. Most social media platforms are open to register and easily accessible, which enables malicious users to spread misinformation while easily disguise their accounts. For example, thousands of bot accounts were found to intentionally spread misinformation during the 2016 U.S. election[1]. To complicate the problem, in order to avoid being detected, they copy legitimate content from normal users [24] and farm links with other people [11]. The manipulated content and links camouflage the malicious users that further lead to a polluted dataset on which decision makers may rely to design public policies.

In this work, we precisely focus on the computational challenge brought by emerging misinformation in social media data. Existing efforts in this area mainly focus on a deletion-based way to solve the problem: building a detection model to identify polluted points, removing them from the data, and learning a predictive model with the refined dataset. However, the ground truth data for the malicious users itself can be very difficult to obtain. Hence, a deletion-based method is limited by the availability of additional label information. In real applications, for sake of simplicity, noisy data are often directly used. Therefore, it would be appealing if the negative effect of noisy data instances can be seamlessly mitigated.

The task of learning a predictive model in the presence of misinformation is particularly difficult, if not impossible, especially when we are lacking availability of labels of malicious users. In order to tackle the challenge, we assume that the real performance can be tested on a holdout dataset, and the optimal performance can be achieved by selecting only the unpolluted data instances. Therefore, an optimal set of model coefficients can be achieved by exhausting all possible combinations of instances. Given the size of the selected instance set, the task is a NP-hard problem due to the combinatorial property. Since the size is also a variable and the size of a dataset is usually very large, it is computationally unfeasible to directly search for it. To this end, we propose a novel relational learning method, Relational Learning with Misinformation (RLM), to identify the set of instances in polynomial time.

In order to tackle the computational challenge, we utilize the social network structure to facilitate the

*Computer Science and Engineering, Arizona State University, Tempe, AZ, USA. {wuliang, jundongl, huan.liu}@asu.edu

†Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA. morstatt@usc.edu

[1]https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html

search for optimal coefficients. As revealed in social identity theory [10], the membership of social community is likely to indicate the similar identity shared among all community members, and the community structure is relatively less susceptible to be affected by malicious behaviors. Hence, we propose to model the community structure with an adaptive group Lasso approach to solve the instance selection problem for relational learning. The main contributions of the work are summarized below:

- We study a novel problem in relational learning, *i.e.,* relational learning in the presence of misinformation, and formally define the task;
- Utilize the social network structure to tackle the challenge of computational expense, and propose a novel optimization framework;
- Suggest mathematical formulations to solve the proposed optimization problem efficiently; and
- Conduct extensive experiments to evaluate the proposed framework on real-world online social media datasets against competitive baselines.

The remainder of the paper is structured as follows. In Section 2, we introduce the problem and formally define the computational task. In Section 3, we introduce our proposed framework as well as the optimization method and theoretical analysis. In Section 4, we conduct experiments on real-world social media datasets to evaluate the proposed method. We introduce related work in Section 5, and conclude the paper and present future work in Section 6.

2 Problem Statement

Given a set of social media users, and consider $\mathbf{V} \in \mathbb{R}^{m \times n}$ is the attribute matrix where m is the number of users and n is the number of features, $\mathbf{P} \in \mathbb{R}^{m \times m}$ denotes the adjacency matrix manifested by the social network structure where $\mathbf{P}_{i,j} = 1$ indicates that user i follows j and it equals to 0 otherwise, $\mathbf{t} \in \{0,1\}^m$ is a label vector represents whether a user contains a certain social tag. Given label information for a subset of users $\mathbf{t} \in \{0,1\}^{m_{tr}}$, due to the influence of misinformation, the label vector is noisy and thus there are k instances mislabeled, we aim to predict labels for the rest m_{te} unlabeled users where $m = m_{tr} + m_{te}$. More formally, the problem is stated as follows:

Input

a user-attribute matrix $\boldsymbol{V}$, an adjacency matrix $\boldsymbol{P}$ and the label information $\boldsymbol{y}^{tr} = \{0,1\}^{m_{tr}}$ for a subset of m_{tr} users.

Output

labels of test users, $\boldsymbol{t}^{m_{te}} = \{0,1\}^{m_{te}}$, where m_{te} is the size of testing data.

A mislabeled instance indicates that the label fails to reveal the true identity of the user. In the process of learning, we posit the existence of misinformation and aim to select top k instances that are not mislabeled to build an optimal predictive model. The social labels can be obtained from different sources on different platforms. For example, Flickr users can join different groups and BlogCatalog users are able to subscribe and add tags for themselves. The group memberships and interest tags can be extracted as labels.

3 Proposed Approach: RLM

In order to illustrate our intuition, we illustrate the framework of classic relational learning and the proposed approach in Figure 1. A conventional practice of dealing with social media network data is to construct a classifier with the data matrix extracted from users. Considering potential negative effect brought by misinformation, we argue a model with better accuracy can be obtained by selecting a subset of instances for training. As shown in Figure 1(b), an additional instance selection module is introduced.

Social networking platforms allow users to freely post content information, which reveals the preference and interests of a user and thus could be utilized to characterize the user in relational learning. To this end, a classifier can be constructed by minimizing:

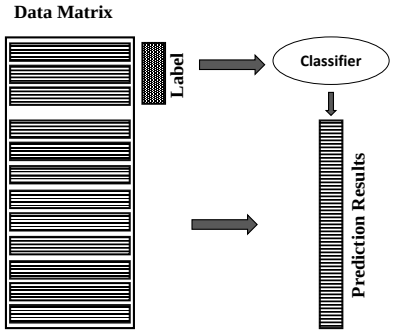
$$\frac{1}{2}||\mathbf{V}\mathbf{w} - \mathbf{y}||_2^2, \tag{3.1}$$

where $\mathbf{V} \in \mathbb{R}^{m \times n}$ is the data matrix, and m is the number of users and n is the number of textual features. Linear regression is adopted here for generality, and $\mathbf{w} \in \mathbb{R}^n$ represents the model coefficients that need to be optimized. $\mathbf{y} \in \mathbb{R}^m$ is the label vector of training data. Throughout the paper, we focus on a binomial classification setting which can be easily extended to the multinomial case.

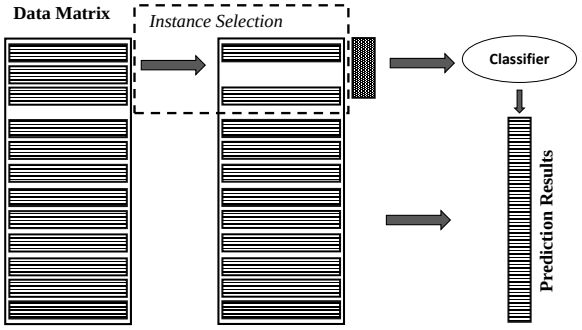
In order to avoid over-fitting, a regularization term is often adopted to control the model complexity. The model can then be formulated as:

$$\frac{1}{2}\min_{\mathbf{w}}||\mathbf{V}\mathbf{w} - \mathbf{y}||_2^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2, \tag{3.2}$$

where λ_1 controls the cutoff between model complexity and accuracy. A larger λ_1 leads to a more simplified model. The formulation achieves an optimal $\mathbf{w}$ through minimizing the training error. Considering the negative effect of misinformation, we introduce to integrate



(a) Relational learning with social media users.



(b) Relational learning with selected instances.

Figure 1: Illustration of comparison between traditional relational learning and the proposed approach with instance selection. A classic relational learning method directly constructs a classifier with available label information; while the proposed framework first removes noise from the label information by actively selecting instances, upon which a classifier is built.

instance selection as,

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{c}} \frac{1}{2}\sum_{i=1}^{m} \mathbf{c}_i(\mathbf{V}_{i,*}\mathbf{w} - \mathbf{y}_i)^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2 \\ &\text{subject to} \quad \sum_i \mathbf{c}_i = k, \mathbf{c} \in \{0,1\}^m, \end{aligned} \tag{3.3}$$

where we introduce an instance selection term $\mathbf{c} \in \{0,1\}^m$ to select k instances to only have influence on the classifier, and k is a predefined budget. Due to the combinatorial nature, it is an *NP-hard* problem which can be difficult to solve. It could also be laborious to find an optimal k. In order to cope with the computational challenge, we try to leverage the social network structures.

On a social networking site, users can be organized by assorted social groups and communities. Since the community structure is often induced from the homophily or proximity relationship between users, it provides a valuable perspective of user profiles [10]. Here, we posit the correlation between social community structure and the information quality that, users belonging to the same group are more likely to provide content of similar quality. The community structure is also more robust to the link farming of malicious users: randomly establishing a link with a legitimate user can be relatively easy, while establishing links with multiple users belonging to the same community can be very difficult.

Next, we define an *index tree* to denote the social community structure for brevity of presentation,

DEFINITION 1. ***Index tree:*** *Let T denote a tree of depth d, where non-leaf nodes represent social communities and leaf nodes are users. Let $T_i = \{G_1^i, G_2^i, \ldots, G_{n_i}^i\}$ denote the nodes on layer i, where $n_0 = 1$ and n_i is the number of nodes on layer i. Given $i < d$, G_j^i represents j^{th} group on the i^{th} layer. $G_1^0 = \{1, 2, \ldots, m\}$ contains indices of all users. In order to maintain a tree structure, nodes should satisfy the following conditions: 1) Nodes on the same layer share no indices with each other ($G_j^i \cap G_k^i = \emptyset, \forall i = 0, \ldots, d, j \neq k, j \leq n_i, k \leq n_i$); 2) Given a non-root node G_j^i, we denote its parent node as G_{j0}^{i-1} ($G_j^i \subseteq G_{j0}^{i-1}, 1 < i \leq d$).*

In order to obtain such a group structure, we select a hierarchical community detection method, namely Louvain [1], where maximum modularity is used to optimize the group structure. The code is available[2].

Given a social community structure, the task of instance selection can boil down to community selection. Though the search space is significantly reduced, exhausting all possible combinations can also be time-consuming. To this end, we further relax the constraint on $\mathbf{c}$ and rewrite the optimization objective in Eq.(3.3),

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{c}} \frac{1}{2}\sum_{i=1}^{m} \mathbf{c}_i(\mathbf{V}_{i,*}\mathbf{w} - \mathbf{y}_i)^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2 + \lambda_2 \sum_{i=0}^{d}\sum_{j=1}^{n_i} ||\mathbf{c}_{G_j^i}||_2 \\ &\text{subject to} \quad \sum_i \mathbf{c}_i = k, \end{aligned} \tag{3.4}$$

where we relax the instance selection vector $\mathbf{c}$ to be a non-binary vector. In order to make the vector

[2]https://perso.uclouvain.be/vincent.blondel/research/louvain.html

more "binary" to align it with the objective of instance selection, we propose to force more entries in $\mathbf{c}$ to be exact 0 or 1. Specifically, we integrate a structured sparsity regularizer $||\mathbf{c}_{G_j^i}||_2$. λ_2 is used to control the extent of sparsity. The adopted sparsity regularizer is a tree-structured group Lasso [9],

$$\sum_{i=0}^{d}\sum_{j=1}^{n_i}||\mathbf{c}_{G_j^i}||_2, \tag{3.5}$$

where an ℓ_2-*norm* is imposed on each member of a group, and an ℓ_1-*norm* is imposed on weights of all groups. This ℓ_{21}-*norm* is iteratively imposed on the social community structure in a bottom-up manner. The combination of ℓ_1- and ℓ_2-*norm* leads to sparse representation of $\mathbf{c}$, while ℓ_1-*norm* determines the organization of sparsity [17]. In particular, imposing ℓ_1-*norm* within each group leads to the inter-group sparsity, *i.e.,* weights of users in some groups are selected to be assigned higher weights, while users in other groups are with lower weights. Therefore, by minimizing the training error, groups that lead to better accuracy are selected by the sparse representation of $\mathbf{c}$.

3.1 Optimization In this section, we introduce how we optimize the problem efficiently. Two variables need to be optimized in Eq.(3.4), $\mathbf{c}$ for instance selection and $\mathbf{w}$ for classifying users. The problem is not jointly convex *w.r.t.* both variables simultaneously. As a conventional practice, we alternatively optimize one variable by fixing the other. The optimization problem boils down to two convex optimization tasks, and we keep iterating over them until convergence.

3.1.1 Instance Selection Here we focus on optimizing $\mathbf{c}$ while keep $\mathbf{w}$ being fixed. Since the squared loss $(\mathbf{V}\mathbf{w}-\mathbf{y}_i)^2$ becomes a constant, we replace it with $\mathbf{p}$, where $\mathbf{p}_i=(\mathbf{V}_{i,*}\mathbf{w}-\mathbf{y}_i)^2$. The objective can then be reformulated as:

$$\begin{aligned}&\min_{\mathbf{c}}\frac{1}{2}\sum_{i=1}^{m}\mathbf{c}_i\mathbf{p}_i+\lambda_2\sum_{i=0}^{d}\sum_{j=1}^{n_i}||\mathbf{c}_{G_j^i}||_2\\ &\text{subject to}\quad \sum_i\mathbf{c}_i=k,\end{aligned} \tag{3.6}$$

where the regularizer $||\mathbf{w}||_2^2$ that is fixed here is also omitted. It is easy to prove that Eq.(3.6) is strongly convex but not directly differentiable, *i.e.*, it is convex and non-smooth with respect to $\mathbf{c}$. In order to find the solution for the optimization problem in Eq.(3.6), we reformulate the problem as follows:

$$\phi_{\lambda_2}(\mathbf{c})=\arg\min_{\mathbf{c}}\frac{1}{2}||\mathbf{c}-\mathbf{x}||^2+\lambda_2\sum_{i=0}^{d}\sum_{j=1}^{n_i}||\mathbf{c}_{G_j^i}||_2, \tag{3.7}$$

where $\mathbf{x}\in\mathbb{R}^m$ and $\mathbf{x}_i=\frac{\mathbf{p}_i^{-1}}{\sum_k^m\mathbf{p}_k^{-1}}$. Therefore, the equality constrained optimization problem is transformed to a Moreau-Yosida regularization problem with the euclidean projection of $\mathbf{c}$ on to a vector $\mathbf{x}$ [13]. The new formulation is continuously differentiable and it admits an analytical solution [16]. Given a proper λ_2, the optimal $\mathbf{c}\in\mathbb{R}^m$ can be obtained in an agglomerative manner, which is shown in Algorithm 1. In the algorithm, the superscript of $\mathbf{c}$ is used to denote the layer of the tree, meaning that the output of the algorithm is $\mathbf{c}^0$. The bisection method can be implemented to find the optimal λ_2. Empirically, λ_2 can be initialized as $\sqrt{\frac{||l'(\mathbf{0})||_2^2}{\sum_{i=0}^{d}n_i}}$, where $l(c)=\frac{1}{2}||\mathbf{c}-\mathbf{x}||^2$. Then we use $\phi_{\lambda_2}(-l'(\mathbf{0}))$ to test whether λ_2 achieves the certain threshold. When $\phi_{\lambda_2}(-l'(\mathbf{0}))=\mathbf{0}$, which means λ_2 is large enough to generate a trivial solution, we start looking for the lower bound as follows:

$$\lambda_2^{(lower)}=\max\{\lambda_2^{(i)}|\lambda_2^{(i)}=\frac{\lambda_2^{(i)}}{2^i},\pi_{\lambda_2^{(i)}}(-l'(\mathbf{0}))\neq\mathbf{0}\} \tag{3.8}$$

otherwise, if $\phi_{\lambda_2}(-l'(\mathbf{0}))\neq\mathbf{0}$, we start looking for the upper bound as follows:

$$\lambda_2^{(upper)}=\min\{\lambda_2^{(i)}|\lambda_2^{(i)}=2^i\lambda_2^{(i)},\pi_{\lambda_2^{(i)}}(-l'(\mathbf{0}))=\mathbf{0}\} \tag{3.9}$$

Algorithm 1 Solution of Moreau-Yosida Regularization

Input: $\{\mathbf{c},G,\lambda_2\}$
Output: $\mathbf{c}^0$.

1: Set $\mathbf{c}^{d+1}=\mathbf{x}$,
2: **for** $i=d$ to 0 **do**:
3: **for** $j=1$ to n_i **do**:
4: Compute:

$$\mathbf{c}_{G_j^i}^i=\begin{cases}\mathbf{0} & if\quad ||\mathbf{c}_{G_j^i}^{i+1}||_2\leq\lambda_2,\\ \frac{||\mathbf{c}_{G_j^i}^{i+1}||_2-\lambda_2}{||\mathbf{c}_{G_j^i}^{i+1}||}\mathbf{c}_{G_j^i}^{i+1} & if\quad ||\mathbf{c}_{G_j^i}^{i+1}||_2>\lambda_2,\end{cases}$$

5: **end for**
6:**end for**

In Algorithm 1, we traverse the tree in an agglomerative manner, *i.e.*, from leaf nodes to the root node. At each node, the ℓ_2-*norm* of the weight $\mathbf{c}$ can be reduced by at most λ_2 as shown in step 4. After the traverse, the analytical solution of $\mathbf{c}$ can be achieved.

3.1.2 Predictor Training When $\mathbf{c}$ is fixed, the problem only depends on $\mathbf{w}$. We reformulate the objective function as follows:

$$\epsilon_{\mathbf{w}} = \frac{1}{2}\sum_{i=1}^{m}\mathbf{c}_i(\mathbf{V}_{i,*}\mathbf{w} - \mathbf{y}_i)^2 + \frac{\lambda_1}{2}||\mathbf{w}||_2^2. \tag{3.10}$$

Therefore, the problem is reduced to an ℓ_2 regularized weighted linear regression problem, which is to minimize the cost $\epsilon_{\mathbf{w}}$. Since social media users and their corresponding contents may be massive, a scalable optimization method is needed. Here we use Stochastic Gradient Descent (SGD) [3]. Since Eq.(3.10) is convex, the corresponding gradient can directly be obtained as:

$$\frac{\partial \epsilon_{\mathbf{w}}}{\partial \mathbf{w}} = \sum_{i=1}^{m}\mathbf{c}_i\mathbf{V}_{i,*}^T(\mathbf{V}_{i,*}\mathbf{w} - \mathbf{y}_i) + \lambda_1\mathbf{w}. \tag{3.11}$$

SGD is scalable since data examples can be updated in parallel [27]. Detailed discussions about the performance can be found in Section 4.

3.2 Time Complexity Analysis Here we analyze the time complexity of the algorithm. The computational costs include computation of $\mathbf{c}$ and $\mathbf{w}$. The computational cost for $\mathbf{c}$ comes from estimating the Moreau-Yosida regularization problem, which takes $\sum_{i=0}^{d}\sum_{j=1}^{n_i}|G_j^i|$. The computation of $\mathbf{w}$ is a standard ℓ_2 regularized regression problem, which can be accelerated with the parallel implementation. The calculation of Louvain method could also speed up and it needs to be done only once as preprocessing [1]. Since the optimization is conducted in an alternative manner and both sub-tasks are convex, both procedures will monotonically decrease. In addition, since the objective function has lower bounds, such as zero, the above iteration converges.

3.3 Convergence Analysis Here we analyze the convergence condition. Since the optimization is conducted in an alternative manner and both sub-tasks are convex, both procedures will monotonically decrease. In addition, since the objective function has lower bounds, such as zero, the above iteration converges.

4 Experiments

RLM is proposed to seamlessly mitigate the negative effect of misinformation in a relational learning method. In this section, we aim to answer two research questions:

- How effective is the proposed method compared with other approaches in terms of classification accuracy?
- In the presence of misinformation, can the proposed RLM identify and downweight the anomalous training instances?

To answer the questions, we conduct experiments on two real-world social media datasets. Next, we will introduce the adopted datasets and experimental settings.

4.1 Datasets We conduct experiments on two real-world social media datasets that are publicly available[3]. Table 1 illustrates some statistics about the two datasets. The users are randomly sampled from the two websites. Assorted features are extracted, such as text and scalar features like age. Following previous work [18, 22], we adopt the user interest tags in BlogCatalog and group memberships in Flickr as labels.

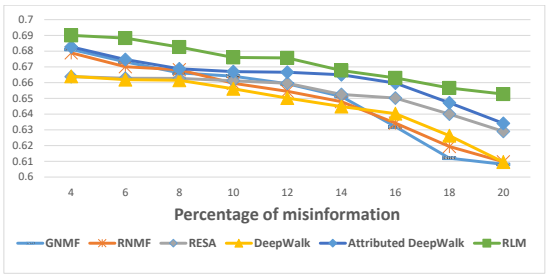
Table 1: The statistics about employed datasets.

	# of Instances	# of Labels	# of features
BlogCatalog	5198	6	8189
Flickr	7575	9	12047

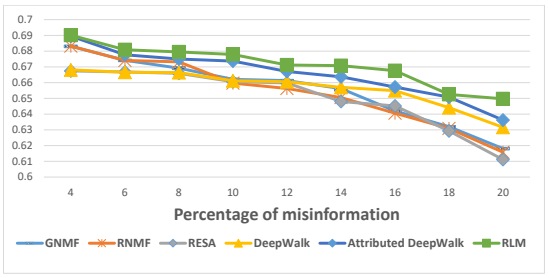
4.2 Baseline Methods and Metrics Our work focuses on classifying instances in a graph. Therefore, we compare with state-of-the-art classification methods with content and network information. We follow experimental settings of graph representation learning approaches by learning a classifier upon the learned dimensions.

- *Graph Regularized NMF*: aims to utilize both content and network information to characterize attributed graph nodes [4]. Based on the assumption of homophily, connected nodes are regularized to be predicted with similar labels. We denote the method as *GNMF*.
- *Robust NMF*: In order to deal with the anomalous instances in a dataset, in the area of robust statistics. We adopt Correntropy Induced Metric Non-Negative Matrix Factorization [6] which extends NMF by incorporating a correntropy induced metric to mitigate the negative effect of non-Gaussian noise. The method is denoted as *RNMF*.
- *Relational Learning with Social Status*: Our previous work that particularly focuses on modeling social network users by integrating social status into the relational learning framework. We denote the approach as *RESA*.

[3] http://socialcomputing.asu.edu/



(a) The *Macro*-F_1 measure of different methods on BlogCatalog data with varying percentage of misinformation.



(b) The *Micro*-F_1 measure of different methods on BlogCatalog data with varying percentage of misinformation.

Figure 2: Comparison of different methods on the BlogCatalog dataset with *Macro*-F_1 and *Micro*-F_1 measures. Additional training instances are randomly selected and flipped with the label.

- *DeepWalk*: is a state-of-the-art graph embedding algorithm that learns distributed representations of social network users, which reports optimal accuracy on the BlogCatalog and Flickr datasets [18].
- *Attributed DeepWalk*: extends DeepWalk by jointly considering the attribute information of graph nodes and reports optimal results among a variety of methods on learning attributed graphs [26].

4.3 Experimental Settings To test the prediction accuracy in terms of both precision and recall, we adopted the F_1-measure to evaluate the performance. Since the adopted dataset contains multiple class labels, and the instance number of different class labels is unbalanced, we adopt *Macro*-F_1 and *Micro*-F_1 to evaluate the performance of different methods.

Macro-F_1 is the arithmetic average of all classes, and it can be formulated as,

$$Macro - F_1 = \frac{1}{|\boldsymbol{T}|}\sum_{t\in\boldsymbol{T}} F_1^t, \tag{4.12}$$

where $\boldsymbol{T}$ is the set of all identity labels and F_1^t is the F_1-measure of task t.

A possible problem of Macro-F_1 is, since the size of different labels varies, the task with fewer instances may be overemphasized. Therefore, *Micro*-F_1 is adopted to mitigate the effect. First, we calculate the micro-averaged precision and recall:

$$Micro - precision = \frac{\#TP}{\#TP + \#FP} \tag{4.13}$$

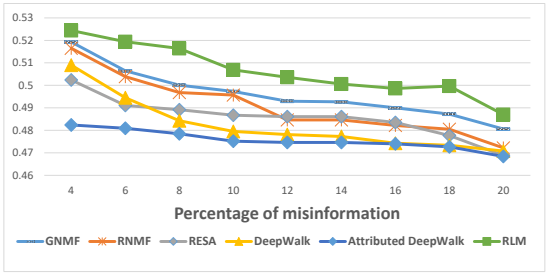
$$Micro - recall = \frac{\#TP}{\#TP + \#FN}, \tag{4.14}$$

where #TP is the number of true positives, #FP is the number of false positives and #FN is the number of false negatives. Then Micro-F_1 is the harmonic average of Micro-precision and Micro-recall. In addition, five-fold cross-validation is adopted for all experiments, and the reported results are the average of all five folds.

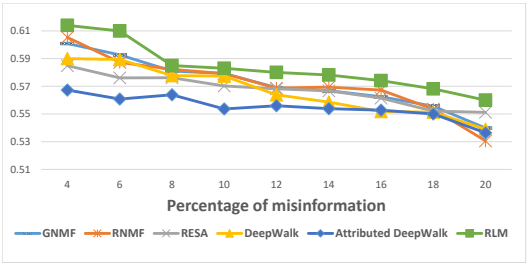
In order to study the effect of misinformation, we randomly select instances in the training set to flip their labels. The classification is conducted in a One versus All (OvA) setting, so flipping the label means changing the label value to the opposite, *i.e.,* 0 to 1 or 1 to 0. Based on the modified training dataset, we learn the classifier and report the experimental results.

4.4 Experiments on BlogCatalog Data The performance of different methods on BlogCatalog dataset with varying percentage of flipped instances, from 4% to 20%, is illustrated in Figure 2. The *x-axis* denotes the percentage of flipped instances, which are randomly sampled from the training set. From the experimental results we draw following observations:

- The proposed approach RLM outperforms all baselines in both settings. The margin between RLM and the runner-up models varies with different percentage of mislabeled data instances.
- The performance of Attributed DeepWalk is the runner-up method in both settings, which implies that both network and content information is useful in modeling a user.
- Since the class distribution of BlogCatalog data is relatively less skewed, the *Macro-* and *Micro*-F_1 results do not show drastic differences.
- DeepWalk has the lowest *Micro-* and *Macro*-F_1 among all six methods. Since DeepWalk investigates only the network information, the result reveals that content information is vital in characterizing social media users.



(a) The *Macro*-F_1 measure of different methods on Flickr data with varying percentage of misinformation.



(b) The *Micro*-F_1 measure of different methods on Flickr data with varying percentage of misinformation.

Figure 3: Comparison of different methods on the Flickr dataset with *Macro*-F_1 and *Micro*-F_1 measures. Additional training instances are randomly selected and flipped with the label.

4.5 Experiments on Flickr Data The performance of different methods on Flickr dataset is illustrated in Figure 3. Based on the experimental results, we draw following observations,

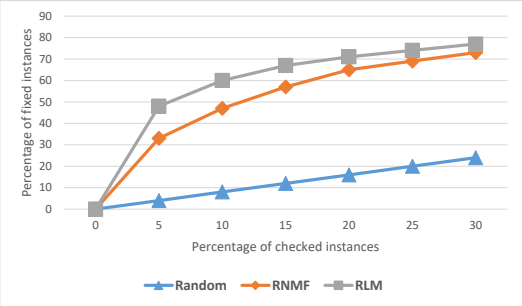
- The proposed RLM achieves the optimal *Macro*-F_1 (Figure 3(a)) and *Micro*-F_1 (Figure 3(b)) on the Flickr dataset.
- Different from the results of BlogCatalog, GNMF is the runner-up for *Macro*-F_1 and RNMF is the runner-up for *Micro*-F_1. Based on the definitions of *Macro*- and *Micro*-F_1, the result indicates that RNMF performs better at a class with more data instances, while GNMF performs relatively better on more *smaller* classes.
- The runner-up method for BlogCatalog, Attributed DeepWalk, is with a relatively low F_1 measure on the dataset of Flickr. The method assumes nodes in the same latent community are more likely to have similar representations. However, since label information for Flickr is the group memberships, it is likely that users form a group without having similar interests or similar content, which contradicts the assumption of Attributed DeepWalk.
- The *Macro*-F_1 measure is generally better than the *Micro*-F_1 measure of all methods. Since we randomly select training instances without considering the class distribution, these minority classes are more vulnerable to the *flipping attacks*.

4.6 Analysis for Instance Selection In this section, we study how well the proposed RLM can identify the mislabeled data instances. We use different methods to select suspicious data instances that are more likely to have been flipped. RLM downweights instances that are more likely to contain misinformation, so the weights are used to rank all data instances in a descending order. RNMF also directly models the negative effect of noisy data points, which is also adopted here. A baseline of Random is also introduced for comparison purposes, which selects instances at random. The results in Figure 4 show that adopting RLM allows us to efficiently find the mislabeled data points without checking too many instances, outperforming the other two baselines.

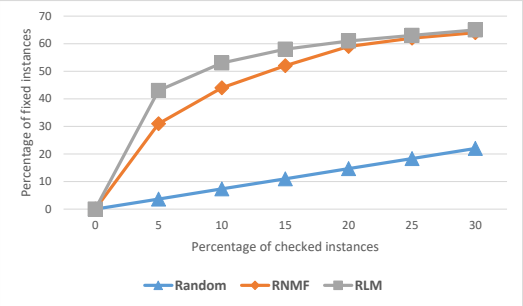
5 Related Work

Relational learning focuses on classification of data instances that are interconnected in a graph. Unlike traditional machine learning tasks where data instances are *i.i.d.*, interconnections between nodes on a graph make the i.i.d. assumption no longer holds. To model the interconnected instances, various methods have been proposed in the area of relational learning. The early research aims to transform the network structure, which is usually represented as an adjacency matrix, to distributed feature vectors, and the graph nodes can be classified [7, 12]. These methods focus on the one-hub friendship, by contrast, the long-distance relationship has been also been studied [25], where latent semantic dimensions are assumed to be underlying connections between users, and thus the connections can be used to project users onto the latent semantic space.

Representation learning has also been widely applied in modeling social network data. The research can be dated back to studies on social dimensions [19], and recent work on network embedding also shows superior accuracy on network clustering and classification [18]. These methods solve the problem through learning a low-rank representation social actors, which can be regarded as extracting features from the network [15, 14]. In order to achieve a higher accuracy, previous studies assumed nodes are not equally weighted, and



(a) The *Macro*-F_1 measure of different methods on Flickr data with varying percentage of misinformation.



(b) The *Micro*-F_1 measure of different methods on Flickr data with varying percentage of misinformation.

Figure 4: Comparison of effectiveness of different methods in identifying mislabeled instances for BlogCatalog and Flickr datasets. Plots show the percentage of mislabeled nodes being fixed by checking instances in training data. RLM ranks data instances with the learned weight in a descending order, RNMF ranks data with the training loss, and we adopt a Random baseline that selects nodes at random.

investigate weights learned from social connections, such as social status [22], social group membership [23]. However, previous studies ignore the potential negative effect brought by misinformation.

In order to cope with the misinformation, it is also common to employ a deletion-based method: first detecting anomalous nodes in the dataset, and training a model by training on the refined dataset. Various methods have been proposed to study the anomalous instances [5], by measuring the anonymity with content and network information. Since malicious users in social media may actively manipulate content and network information to disguise themselves, recent work has been proposed to study how the camouflage can be detected [24, 11]. Deletion-based methods boil down the problem into two stages, and each of them is solved separately. The proposed framework jointly solves the problem in a unified framework. In addition, additional label information for anomalous data is also needed, which can be difficult to annotate and obtain.

Our work is also related to sparse learning and structured Lasso [9], which has been widely applied in feature selection and representation learning. Real-world data is usually of high dimensionality and can be noisy, so sparse learning has been proposed to uncover a descriptive representation of data with a small portion of features being selected. A myriad of sparsity structures has been studied in previous research, such as group-structured and tree-structured. Our work distances from existing sparse learning methods by focusing on selecting data instances in lieu of features.

6 Conclusion and Future Work

The massive amount of social media data allows automatic modeling of users in the social media network. Relational learning, which particularly focuses on interconnected data instances, have been successfully applied in a myriad of applications. An emerging challenge of utilizing social media data is the negative effect brought up by the misinformation. In this work, we precisely focus on the problem of mitigating its harm. In particular, we propose a unified framework that simultaneously selects data instances and learn a relational learning model. In order to allow for efficient optimization, we utilize the social community structure to effectively find groups of instances. We also transform the combinatorial problem into a convex optimization problem with relaxations. Experimental results on real-world datasets show the superiority of the proposed approach over competitive baseline methods. We also conduct experiments to understand how RLM selects and downweights data instances.

There are several promising directions for the problem of relational learning with misinformation. First, this work focuses on misinformation and we produce a dataset by randomly flipping labels of data instances, which is the optimal step we can take with lack of availability of real misinformation dataset. Hence, it would be interesting to collect an organic misinformation dataset and study how the proposed framework performs under different scenarios. Second, in order to scale up the instance selection, we propose to utilize the social community structure, and the community detection is conducted separately with the relational learning. In the future, we are interested in studying how different community detection algorithms influence the performance and investigating possibilities of integrating them with the relational learning approach.

Acknowledgments

We would like to thank anonymous reviewers for their constructive comments. The work is funded, in part, by ONR N00014-16-1-2257 and the Department of Defense under the MINERVA initiative through the ONR N000141310835.

References

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[4] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[6] Liang Du, Xuan Li, and Yi-Dong Shen. Robust nonnegative matrix factorization via half-quadratic minimization. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 201–210. IEEE, 2012.

[7] Hoda Eldardiry and Jennifer Neville. An analysis of how ensembles of collective classifiers improve predictions in graphs. In *CIKM*, pages 225–234, 2012.

[8] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM, 2010.

[9] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[10] Michael A Hogg. Social identity theory. In *Understanding Peace and Conflict Through Social Identity Theory*, pages 3–17. Springer, 2016.

[11] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–904. ACM, 2016.

[12] David Jensen, Jennifer Neville, and Brian Gallagher. Why collective inference improves relational classification. In *KDD*, pages 593–598. ACM, 2004.

[13] Claude Lemaréchal and Claudia Sagastizábal. Practical aspects of the moreau–yosida regularization: Theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):367–385, 1997.

[14] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. Attributed network embedding for learning in a dynamic environment. *arXiv preprint arXiv:1706.01860*, 2017.

[15] Jundong Li, Liang Wu, Osmar R Zaïane, and Huan Liu. Toward personalized relational learning. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 444–452. SIAM, 2017.

[16] Jun Liu and Jieping Ye. Moreau-Yosida regularization for grouped tree structure learning. In *NIPS*, pages 1459–1467, 2010.

[17] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[18] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[19] Lei Tang and Huan Liu. Relational learning via latent social dimensions. In *ACM International Conference on Knowledge Discovery and Data Mining*, 2009.

[20] Catherine E Tucker. Social networks, personalized advertising, and privacy controls. American Marketing Association, 2014.

[21] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185, 2010.

[22] Liang Wu, Xia Hu, and Huan Liu. Relational learning with social status analysis. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 513–522. ACM, 2016.

[23] Liang Wu, Xia Hu, Fred Morstatter, and Huan Liu. Adaptive spammer detection with sparse group modeling. In *ICWSM*, pages 319–326, 2017.

[24] Liang Wu, Xia Hu, Fred Morstatter, and Huan Liu. Detecting camouflaged content polluters. In *ICWSM*, pages 696–699, 2017.

[25] Zhao Xu, Volker Tresp, Shipeng Yu, and Kai Yu. Nonparametric relational learning for social network analysis. In *KDD 2008 Workshop on Social Network Mining and Analysis*, 2008.

[26] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. Network representation learning with rich text information. In *IJCAI*, pages 2111–2117, 2015.

[27] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.