Feature Selection for Social Media Data

JILIANG TANG, Arizona State University
HUAN LIU, Arizona State University

Feature selection is widely used in preparing high-dimensional data for effective data mining. The explosive popularity of social media produces massive and high-dimensional data at an unprecedented rate, presenting new challenges to feature selection. Social media data consists of (1) traditional high-dimensional, attribute-value data such as posts, tweets, comments, and images, and (2) linked data that provides social context for posts and describes the relationships between social media users as well as who generates the posts, etc. The nature of social media also determines that its data is massive, noisy, and incomplete, which exacerbates the already challenging problem of feature selection. In this paper, we study a novel feature selection problem of selecting features for social media data with its social context. In detail, we illustrate the differences between attribute-value data and social media data, investigate if linked data can be exploited in a new feature selection framework by taking advantage of social science theories. We design and conduct experiments on datasets from real-world social media websites and the empirical results demonstrate that the proposed framework can significantly improve the performance of feature selection. Further experiments are conducted to evaluate the effects of user-user and user-post relationships manifested in linked data on feature selection, and discuss some research issues for future work.

Categories and Subject Descriptors: 1.5.2 [**Pattern Recognition**]: Design Methodology—*Feature evaluation and Selection*

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Feature Selection, Social Media Data, Social Context

ACM Reference Format:
Tang, J., and Liu, H., 2012. Feature Selection for Social Media Data. TKDD 9, 4, Article 39 (June 2012), 27 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

In recent years, myriads of social media services are emerging that attract more and more people to participate in online social activities. Users in social media have dual roles both passive content consumers and active content producers. The explosion of social media produces massive user generated content at an unprecedented rate. For example, 250 million tweets are sent to Twitter [1] per day, which is more than 200 percent growth in a year[2]; more than 3,000 photos are uploaded to Flickr[3] per minute and

[1]http://www.twitter.com
[2]http://techcrunch.com/2011/06/30/twitter-3200-million-tweets/
[3]http://www.flickr.com/

An earlier version of this paper was published at SDM2012 with the title " Feature Selection with Linked Data in Social Media"
Author's addresses: J. Tang and H. Liu, Computer Science and Engineering, Arizona State University, Tempe, AZ 85281

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

more than 153 million blogs are posted per year[4]. One characteristic of such data is high-dimensional such as there are tens of thousands of terms in tweets or pixels for photos in Flickr. This brings about challenges to traditional data mining tasks such as classification and clustering due to the curse of dimensionality. Feature selection is proven to be an effective way to handle high-dimensional data for efficient data mining [Guyon et al. 2002; Liu and Motoda 2008]. Feature selection aims to select relevant features from the high-dimensional data for a compact and accurate data representation. It can alleviate the curse of dimensionality, speed up the learning process, and improve the generalization capability of a learning model [Liu and Yu 2005].

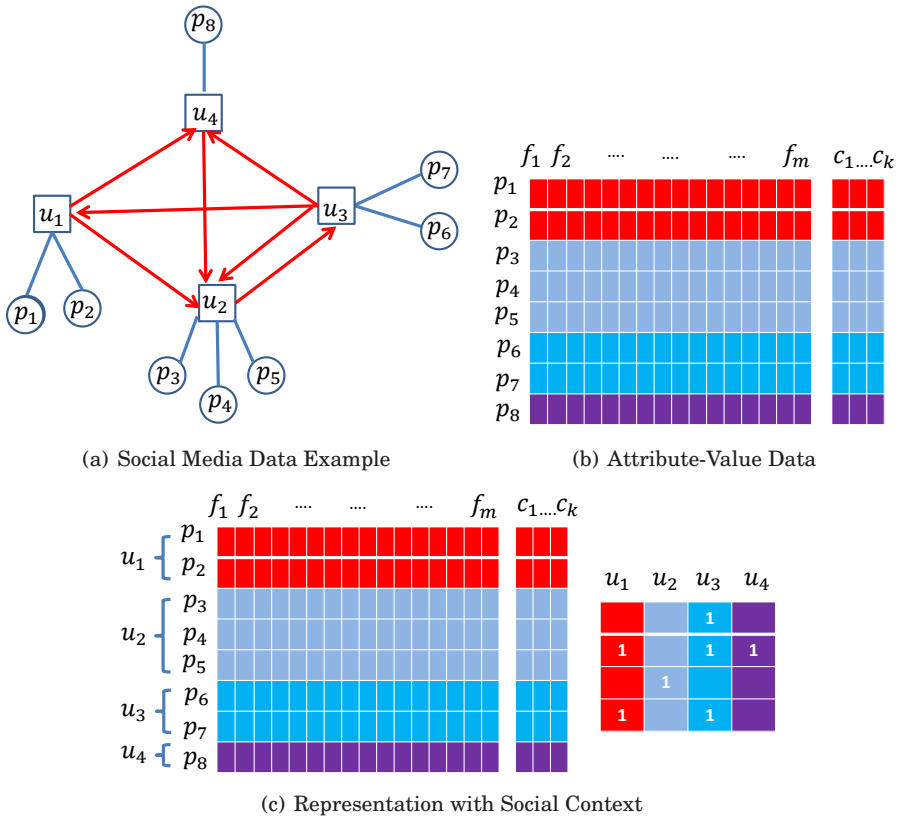
Without loss of generality, Figure 1 illustrates a simple but typical example of social media data with its two data representations. There are two types of objects in Figure 1(a): users ($u_1, ..., u_4$) and their posts (p_1, ..., p_8). Note that we use posts here in a loose way to cover all types of user generated content in social media such as tweets in Twitter, blogs in blog websites, or photos in Flickr. Users in social media have two typical behaviors: (1) following other users (e.g., u_1 follows u_2 and u_4); (2) creating some posts (e.g., user u_1 has two posts p_1 and p_2). For example, Twitter users can send tweets as well as follow other users. Figure 1(b) is a conventional representation of attribute-value data: rows are posts (e.g., tweet) and columns are features for posts (e.g., terms in tweets). Its similarity with social media data stops here. In the context of social media, there is additional information in the form of linked data such as who generates the posts (user-post relations) and who follows whom (user-user relations), named social context of posts in this paper, as shown in Figure 1(c)[5]. Social media data is patently distinct from traditional attribute-value data, posing both challenges and opportunities for feature selection.

The vast majority of existing feature selection algorithms were designed for "flat" data (e.g., Figure 1(b)) containing uniform entities (or attribute-value data points) that are typically assumed to be independent and identically distributed (*i.i.d.*). Actually, *i.i.d.* assumption is among the most enduring and deeply buried assumptions of traditional machine learning methods [Jensen and Neville 2002; Taskar et al. 2003]. However, social media data apparently does not follow independent and identically distributed assumption since its data instances are inherently linked through social context (e.g., Figure 1(c)). On the other hand, social media data provides extra information beyond attribute value in the form of links. In general, there are correlations between linked instances. For example, blogs (tweets or photos) from the same user or two linked users are more likely to have similar topics. The availability of link information enables advanced research about feature selection.

The unique properties of social media data present both challenges and opportunities for feature selection. In this paper, we investigate issues of feature selection for social media data as illustrated in Figure 1(c). Specifically, we would like to develop a novel framework performing feature selection on posts (e.g., tweets, blogs, or photos) in the context of social media with link information between user and user or between user and posts. Traditional feature selection methods are unequipped for social media data in terms of taking advantage of the additional information in linked data. We proceed to study two fundamental problems: (1) *relation extraction* - what are distinctive relations (or correlations) that can be extracted from linked data, and (2) *mathematical representation* - how to represent these correlations and integrate them in a state-of-the-art feature selection formulation. Providing answers to the two

[4]http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/

[5]For some types of social media data, there are hyperlinks among posts such as blogs and tweets, which can not be applied to other types such as photos in Flickr. However, data in social media usually has social context.



(a) Social Media Data Example

(b) Attribute-Value Data

(c) Representation with Social Context

Fig. 1. Typical Social Media Data and Its Two Representations

problems, we propose a novel feature selection framework (LinkedFS) for social media data. The main contributions of this paper are summarized next.

— Identify the need for feature selection in social media and propose to exploit social correlation theories and linked data in formulating the new problem of feature selection for social media data;
— Show that various relations can be extracted from linked data guided by social correlation theories and provide a way to capture link information;
— Propose a framework for social media feature selection (LinkedFS) that integrates conventional feature selection with extracted relations;
— Develop various measures for tie strength prediction for LinkedFS due to the low cost of link formation resulting in strong links and weak links mixed together; and
— Evaluate LinkedFS systematically using real-world social media data to verify if different types of relations improve the performance of feature selection.

The rest of this paper is organized as follows. The problem of feature selection with linked data in social media is formally defined in Section 2. A new feature selection framework, LinkedFS, is introduced in Section 3 based on social correlation theories. In section 4, we investigate how to incorporate tie strength into LinkedFS. Empirical evaluation is presented in Section 5 with discussions. The related work is reviewed in Section 6. The conclusion and future work are presented in Section 7.

2. PROBLEM STATEMENT

We first give the notations to be used in this paper. Scalars are denoted by lower-case letters $(a, b, \ldots; \alpha, \beta, \ldots)$, vectors are written as lower-case bold letters $(\mathbf{a}, \mathbf{b}, \ldots)$, and matrices correspond to bold upper-case letters $(\mathbf{A}, \mathbf{B}, \ldots)$. $\mathbf{A}(i, j)$ is the entry at the i^{th} row and j^{th} column of the matrix $\mathbf{A}$, $\mathbf{A}(i, :)$ is the i^{th} row of $\mathbf{A}$ and $\mathbf{A}(:, j)$ is the j^{th} column of $\mathbf{A}$.

Let $\mathbf{p} = \{p_1, p_2, \ldots, p_N\}$ be the post set (e.g., p_1 to p_8) where N is the number of posts. Let $\mathbf{f} = \{f_1, f_2, \ldots, f_m\}$ denote the feature set where m is the number of features. For each post p_i, $\mathbf{f}_i \in \mathbb{R}^m$ are the set of feature values where $\mathbf{f}_i(j)$ is the frequency of f_j used by p_i. $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_\mathbf{N}\} \in \mathbb{R}^{m \times N}$ denotes the whole dataset $\mathbf{p}$. We assume that the subset $\{p_1, p_2, \ldots, p_{N_l}\}$ is the labeled data where N_l is the number of labeled posts. Then $\mathbf{X} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{\mathbf{N_l}}\} \in \mathbb{R}^{m \times N_l}$ is the matrix for labeled data. Let $\mathbf{c} = \{c_1, c_2, \ldots, c_k\}$ denote the class label set where k is the number of classes. $\mathbf{Y} \in \mathbb{R}^{N_l \times k}$ is the class label matrix for labeled data where $\mathbf{Y}(i, j) = 1$ if p_i is labeled as c_j, otherwise zero.

Except the conventional representation, social media data has social context in the form of links such as who creates the posts and who follows who. Let $\mathbf{u} = \{u_1, u_2, \ldots, u_n\}$ be the user set (e.g., u_1 to u_4) where n is the number of users. $\mathbf{F}_i$ denotes the set of posts from user u_i (e.g., $\mathbf{F}_1 = \{p_1, p_2\}$). We also model the user-user following relationships as a graph with adjacency matrix $\mathbf{S}$, where $\mathbf{S}(i, j) = 1$ if there is a following relationship from u_j to u_i and zero otherwise (e.g., $\mathbf{S}(:, 1) = [0, 1, 0, 1]^\top$). Let $\mathbf{P} \in \mathbb{R}^{n \times N}$ denote user-post relationships where $\mathbf{P}(i, j) = 1$ if p_j is posted by u_i, zero otherwise (e.g., $\mathbf{P}(1, :) = [1, 1, 0, 0, 0, 0, 0, 0]$).

With the notations defined above, our problem with linked data is stated as:
Given labeled data $\mathbf{X}$ *and its label indicator matrix* $\mathbf{Y}$*, the whole dataset* $\mathbf{F}$*, its social context (or social correlations) including user-user following relationships* $\mathbf{S}$ *and user-post relationships* $\mathbf{P}$*, we aim to develop a framework to select a subset* K *of most relevant features,* $\mathbf{f}'$*, from the* m *original features,* $\mathbf{f}$*, on the dataset* $\mathbf{F}$ *with its social context* $\mathbf{S}$ *and* $\mathbf{P}$, which can be formally described as,

$$\{(\mathbf{X}, \mathbf{Y}, \mathbf{F}, \mathbf{P}, \mathbf{S}); \mathbf{f}\} \rightarrow \{\mathbf{f}'\}, \tag{1}$$

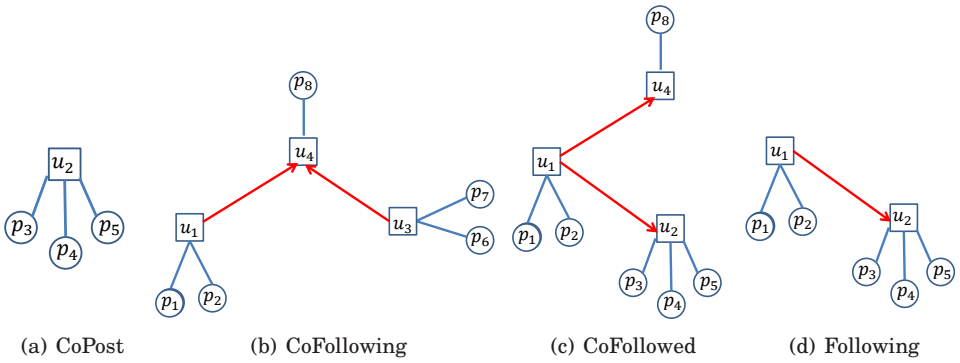
while traditional supervised feature selection aims to select a subset features from m features based on $\{\mathbf{X}, \mathbf{Y}\}$, stated as,

$$\{(\mathbf{X}, \mathbf{Y}), \mathbf{f}\} \rightarrow \{\mathbf{f}'\}. \tag{2}$$

The proposed problem is substantially distinct from that of traditional feature selection: (1) data studied in our problem is embedded in the social context hence naturally correlated while traditional feature selection works with "flat" data assumed to be independent and identically distributed; (2) in addition to capturing the attribute-value part as traditional feature selection methods do, our problem further investigates how to exploit social context for solving feature selection problems.

3. A NEW FRAMEWORK - LINKEDFS

Recall the two fundamental problems for feature selection on social media data: (1) relation extraction, and (2) mathematical representation. Their associated challenges are: (a) *what are different types of relations among data instances from their social context and how to capture them*, and (b) *how to model these relations for feature selection*. In this section, we discuss how to capture relations from linked data guided by social correlation theories, propose a framework (LinkedFS) of social media data that naturally integrates different relations into a state-of-the-art formulation of feature selection, and turn the integrated formulations to an optimization problem with convergence analysis when developing its corresponding feature selection algorithm.



(a) CoPost (b) CoFollowing (c) CoFollowed (d) Following

Fig. 2. Different Types of Relations Extracted from Social Correlations among Social Media Data

3.1. Extracting Various Relations

Examining Figure 1(a), we can find four basic types of relations for the linked data from its social context as shown in Figure 2. They are (a) CoPost from user-post relations: a user (u_2) can have multiple posts (p_3, p_4, and p_5), (b) CoFollowing from user-user following relations: two users (u_1 and u_3) follow a third user (u_4), (c) CoFollowed from user-user following relations: two users (u_2 and u_4) are followed by a third user (u_1), and (d) Following from user-user following relations: a user (u_1) follows another user (u_2). Social correlation theories such as homophily [McPherson et al. 2001] and social influence [Marsden and Friedkin 1993] can be helpful to explain what these relations suggest. Homophily indicates that people with similar interests are more likely to be linked, and social influence reveals that people that are linked are more likely to have similar interests. Based on these theories that define social correlations among linked instances, we turn the four types of relations into four corresponding hypotheses that can affect feature selection with linked data.

CoPost Hypothesis: This hypothesis assumes that posts by the same user (e.g., $\{p_3, p_4, p_5\}$, in Figure 2(a)) are of similar topics. In other words, the posts from the same user are more similar, in terms of topics (say, "sports", "music"), than randomly selected posts.

CoFollowing Hypothesis: This hypothesis suggests that if two users follow the same user (e.g., u_1 and u_3 follow u_4 as in Figure 2(b)), their posts, $\{p_1, p_2\}$ and $\{p_6, p_7\}$, are likely in similar topics. Its counterpart in citation analysis is bibliographic coupling [Morris 2005]: if two papers cite a paper, they are more similar than other papers that do not share references.

CoFollowed Hypothesis: This says that if two users are followed by the same user, their posts are similar in topics. For example, in Figure 2(c), both users u_2 and u_4 are followed by user u_1, and then their posts $\{p_3, p_4, p_5\}$ and $\{p_8\}$ are of more similar topics. It is similar to the co-citation relation [Morris 2005] in citation analysis: if two papers are cited by the same paper, they are more similar than other paper that are not.

Following Hypothesis: This hypothesis assumes that one user follows another (e.g., u_1 follows u_2 in Figure 2(d)) because u_1 shares u_2's interests. Thus, their posts (e.g., $\{p_1, p_2\}$ and $\{p_3, p_4, p_5\}$) are more likely similar in terms of topics.

Next, we elaborate how the above four hypotheses can be modeled into a feature selection formulation in our effort to create a new framework for feature selection with linked data.

3.2. Modeling Hypotheses

We first introduce a representative feature selection method for attribute-value data based on $\ell_{2,1}$-norm regularization [Liu et al. 2009], which selects features across data points with joint sparsity [Ding et al. 2006].

$$\min_{\mathbf{W}} \quad \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \alpha \|\mathbf{W}\|_{2,1} \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix and the parameter α controls the sparseness of $\mathbf{W}$ in rows. $\mathbf{W} \in \mathbb{R}^{m \times k}$ and $\|\mathbf{W}\|_{2,1}$ is the $\ell_{2,1}$-norm of $\mathbf{W}$, which is defined as follows:

$$\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{k} \mathbf{W}^2(i,j)} = \sum_{i=1}^{m} \|\mathbf{W}(i,:)\|_2 \tag{4}$$

This formulation in Eq. (3) is a supervised feature selection method where data instances are assumed to be independent. We now discuss how one can jointly incorporate $\ell_{2,1}$ minimization and different types of relations in feature selection with linked data.

CoPost Relation: To integrate this hypothesis into Eq. (3), we propose to add a regularization term that enforces the hypothesis that the class labels (i.e., topics in this paper) of posts by the same user are similar. Thus, feature selection with the CoPost hypothesis can be formulated as the following optimization problem.

$$\begin{aligned} \min_{\mathbf{W}} \quad & \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \alpha \|\mathbf{W}\|_{2,1} \\ & + \beta \sum_{u \in \mathbf{u}} \sum_{\mathbf{f}_i, \mathbf{f}_j \in \mathbf{F}_u} \|T(\mathbf{f}_i) - T(\mathbf{f}_j)\|_2^2 \end{aligned} \tag{5}$$

β adjusts the contribution from the CoPost relation. Let $\mathbf{A}$ be the co-post matrix, which is defined as $\mathbf{A}(i,j) = 1$ if post p_i and post p_j are posted by the same user, and $\mathbf{A}(i,j) = 0$ otherwise. $\mathbf{A}$ can be obtained from the user-post relationship matrix $\mathbf{P}$, i.e., $\mathbf{A} = \mathbf{P}^\top \mathbf{P}$. Let $T(\mathbf{f}_i) : \mathbb{R}^m \to \mathbb{R}^k$ be the function to predict the labels of the post p_i, i.e., $T(\mathbf{f}_i) = \mathbf{W}^\top \mathbf{f}_i$. $\mathbf{L_A} = \mathbf{D_A} - \mathbf{A}$ is the Laplacian matrix, and $\mathbf{D_A}$ is a diagonal matrix with $\mathbf{D_A}(i,i) = \sum_j \mathbf{A}(j,i)$.

THEOREM 3.1. *The formulation in Eq (5) is equivalent to the following optimization problem:*

$$\min_{\mathbf{W}} tr(\mathbf{W}^\top \mathbf{B} \mathbf{W} - 2\mathbf{E}\mathbf{W}) + \alpha \|\mathbf{W}\|_{2,1} \tag{6}$$

where $tr(\cdot)$ is the trace of a matrix. $\mathbf{B}$ *and* $\mathbf{E}$ *are defined as follows:*

$$\begin{aligned} \mathbf{B} &= \mathbf{X}\mathbf{X}^\top + \beta \mathbf{F} \mathbf{L_A} \mathbf{F}^\top \\ \mathbf{E} &= \mathbf{Y}^\top \mathbf{X}^\top \end{aligned} \tag{7}$$

PROOF. It is easy to verify that the first part of Eq (5) can be written as:

$$\|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 = tr(\mathbf{W}^\top \mathbf{X}\mathbf{X}^\top \mathbf{W} - 2\mathbf{Y}^\top \mathbf{X}^\top \mathbf{W} + \mathbf{Y}^\top \mathbf{Y}) \tag{8}$$

With the definition of $\mathbf{F}$ and $\mathbf{L_A}$, the last regularization constraint of Eq (5) can be written as:

$$\begin{aligned}
&\sum_{u\in\mathbf{u}}\sum_{\mathbf{f}_i,\mathbf{f}_j\in\mathbf{F}_u}\|T(\mathbf{f}_i)-T(\mathbf{f}_j)\|_2^2\\
&=\sum_i\sum_j\mathbf{A}(i,j)\|\mathbf{W}^\top\mathbf{f}_i-\mathbf{W}^\top\mathbf{f}_j\|_2^2\\
&=\sum_k\sum_i\sum_j\Big(\mathbf{W}(k,:)\mathbf{A}(i,j)\mathbf{f}_i-\mathbf{W}(:,k)\mathbf{A}(i,j)\mathbf{f}_j\Big),\\
&=\sum_k\mathbf{W}(k,:)\mathbf{F}(\mathbf{D}_A-\mathbf{A})\mathbf{F}^\top\mathbf{W}(k,:),\\
&=tr(\mathbf{W}^\top\mathbf{F}\mathbf{L_A}\mathbf{F}^\top\mathbf{W})
\end{aligned}\tag{9}$$

Since $\mathbf{Y}^\top\mathbf{Y}$ is constant, the object function is equivalent to:

$$\begin{aligned}
&tr(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}-2\mathbf{Y}^\top\mathbf{X}^\top\mathbf{W}+\mathbf{Y}^\top\mathbf{Y})+\beta tr(\mathbf{W}^\top\mathbf{F}\mathbf{L_A}\mathbf{F}^\top\mathbf{W})\\
&=tr(\mathbf{W}^\top(\mathbf{X}\mathbf{X}^\top+\beta\mathbf{F}\mathbf{L_A}\mathbf{F}^\top)\mathbf{W}-2\mathbf{Y}^\top\mathbf{X}^\top\mathbf{W})\\
&=tr(\mathbf{W}^\top\mathbf{B}\mathbf{W}-2\mathbf{E}\mathbf{W})
\end{aligned}\tag{10}$$

which completes the proof. □

CoFollowing, CoFollowed and Following relations are extracted from user-user following relations, indicating the correlations of the interests of two users based on their posts. For example, the CoFollowing relation suggests two cofollowing users are more likely to have similar interests in terms of the topics of their posts. To model these hypotheses, we have to formally define users' topic interests. With the definition of $T(\mathbf{f}_i)$, the topic interests for u_k, $\hat{T}(u_k)$, is defined as follows:

$$\hat{T}(u_k)=\frac{\sum_{\mathbf{f}_i\in\mathbf{F}_k}T(\mathbf{f}_i)}{|\mathbf{F}_k|}=\frac{\sum_{\mathbf{f}_i\in\mathbf{F}_k}\mathbf{W}^\top\mathbf{f}_i}{|\mathbf{F}_k|}\tag{11}$$

where $|\mathbf{F}_k|$ denotes the size of $\mathbf{F}_k$. One user's topic interests is defined as the distribution of their posts' topics. With the definition of users' topic interests, we are allowed to model the CoFollowing, CoFollowed and Following hypotheses for feature selection.

CoFollowing Relation: For this hypothesis, we add a regularization term into Eq. (3), reflecting the constraint that two co-following users have similar interested topics. The feature selection formulation with CoFollowing hypothesis is below:

$$\begin{aligned}
min_{\mathbf{W}}\ \ &\|\mathbf{X}^\top\mathbf{W}-\mathbf{Y}\|_F^2+\alpha\|\mathbf{W}\|_{2,1}\\
&+\beta\sum_{u_k\in\mathbf{u}}\sum_{u_i,u_j\in\mathbf{N}_k}\|\hat{T}(u_i)-\hat{T}(u_j)\|_2^2
\end{aligned}\tag{12}$$

where $\mathbf{N}_k$ is the set of users who follow u_k. Let $\mathbf{FI}$ be the co-following matrix where $\mathbf{FI}(i,j)=1$ if u_i and u_j are following at least one other person (e.g., u_k). $\mathbf{FI}$ can be obtained from the adjacency matrix $\mathbf{S}$, i.e., $\mathbf{FI}=sign(\mathbf{S}^\top\mathbf{S})$ where the function $sign(x)=1$ if $x>0$ and 0 otherwise.

Let $\mathbf{H}\in\mathbb{R}^{N\times n}$ be an indicator matrix where $\mathbf{H}(i,j)=\frac{1}{|\mathbf{F}_j|}$ if u_j is the author of p_i. Let $\mathbf{L_{FI}}$ be the Laplacian matrix defined on $\mathbf{FI}$. We can develop the following theorem for the CoFollowing relation.

THEOREM 3.2. *The formulation in Eq (12) is equivalent to the following optimization problem:*

$$\min_{\mathbf{W}} tr(\mathbf{W}^\top \mathbf{B}\mathbf{W} - 2\mathbf{E}\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \tag{13}$$

where $\mathbf{B}$ *and* $\mathbf{E}$ *are defined as follows:*

$$\begin{aligned}\mathbf{B} &= \mathbf{X}\mathbf{X}^\top + \beta\mathbf{F}\mathbf{H}\mathbf{L}_{\mathbf{FI}}\mathbf{H}^\top\mathbf{F}^\top \\ \mathbf{E} &= \mathbf{Y}^\top\mathbf{X}^\top\end{aligned} \tag{14}$$

PROOF. We can see that the first part of Eq (5) is the same as that of Eq (12). For this part, the proof process is similar to that of the CoPost hypothesis. The last regularization constraint of Eq (12) can be written as:

$$\begin{aligned}&\sum_{u_k\in\mathbf{u}}\sum_{u_i,u_j\in\mathbf{F}_k}\|\hat{T}(u_i) - \hat{T}(u_j)\|_2^2 \\ &= \sum_{i,j}\mathbf{FI}(i,j)\|\mathbf{W}^\top\mathbf{F}\mathbf{H}(:,i) - \mathbf{W}^\top\mathbf{F}\mathbf{H}(:,j)\|_2^2 \\ &= tr(\mathbf{W}^\top\mathbf{F}\mathbf{H}\mathbf{L}_{\mathbf{FI}}\mathbf{H}^\top\mathbf{F}^\top\mathbf{W})\end{aligned} \tag{15}$$

Then the object function can be converted into:

$$\begin{aligned}&tr(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W} - 2\mathbf{Y}^\top\mathbf{X}^\top\mathbf{W} + \mathbf{Y}^\top\mathbf{Y}) \\ &\quad + \beta tr(\mathbf{W}^\top\mathbf{F}\mathbf{H}\mathbf{L}_{\mathbf{FI}}\mathbf{H}^\top\mathbf{F}^\top\mathbf{W}) \\ &= tr(\mathbf{W}^\top(\mathbf{X}\mathbf{X}^\top + \beta\mathbf{F}\mathbf{H}\mathbf{L}_{\mathbf{FI}}\mathbf{H}^\top\mathbf{F}^\top)\mathbf{W} - 2\mathbf{Y}^\top\mathbf{X}^\top\mathbf{W}) \\ &= tr(\mathbf{W}^\top\mathbf{B}\mathbf{W} - 2\mathbf{E}\mathbf{W})\end{aligned} \tag{16}$$

which completes the proof. □

Following a similar approach to the CoFolllowing relation, we can develop similar theorems for CoFollowed and Following relations. We leave the detailed theorems for CoFollowed relation and Following relation in Appendix **A** and Appendix **B**, respectively.

In this paper, we focus on the effect of each hypothesis on feature selection and will discuss the combination of multiple hypotheses into the same formulation to capture multi-faceted relations later. Closely examining the optimization problems for these four hypotheses, we can see that the LinkedFS framework is tantamount to solving the following optimization problem.

$$\min_{\mathbf{W}} tr(\mathbf{W}^\top\mathbf{B}\mathbf{W} - 2\mathbf{E}\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \tag{17}$$

Next we will present an optimization formulation to solve this problem and give convergence analysis.

3.3. An Optimal Solution for LinkedFS

In this section, inspired by [Nie et al. 2010], we give a new approach to solve the optimization problem shown in Eq. (17). The Lagrangian function of the problem is:

$$\mathcal{L}(\mathbf{W}) = tr(\mathbf{W}^\top\mathbf{B}\mathbf{W} - 2\mathbf{E}\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \tag{18}$$

Taking the derivative of $\mathcal{L}(\mathbf{W})$,

$$\frac{\partial\mathcal{L}(\mathbf{W})}{\partial\mathbf{W}} = 2\mathbf{B}\mathbf{W} - 2\mathbf{E}^\top + 2\alpha\mathbf{D}_{\mathbf{W}}\mathbf{W} \tag{19}$$

where $\mathbf{D_W}$ is a diagonal matrix with the i-th diagonal element as[6]:

$$\mathbf{D_W}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2} \tag{20}$$

All matrices $\mathbf{B}$ defined above are semi-positive definite matrices and therefore $\mathbf{B} + \alpha\mathbf{D_W}$ is a positive definite matrix. Then setting the derivative to zero, we have:

$$\mathbf{W} = (\mathbf{B} + \alpha\mathbf{D_W})^{-1}\mathbf{E}^\top \tag{21}$$

$\mathbf{D_W}$ is dependent to $\mathbf{W}$ and we proposed an iterative algorithm to obtain the solution $\mathbf{W}$. The detailed optimization method for LinkedFS is shown in Algorithm 1. We next verify that Algorithm 1 converges to the optimal $\mathbf{W}$, beginning with the following two lemmas.

ALGORITHM 1: LinkedFS

Input: $\{\mathbf{F}, \mathbf{X}, \mathbf{Y}, \mathbf{S}, \mathbf{P}\}$ and the number of features expected to select, K;
Output: K most relevant features

1: Construct $\mathbf{E}$ and $\mathbf{B}$ according to the hypothesis you choose;
2: Set $t = 0$ and initialize $\mathbf{D}_{\mathbf{W}_t}$ as an identity matrix;
3: **while** Not convergent **do**
4: Calculate $\mathbf{W}_{t+1} = (\mathbf{B} + \alpha\mathbf{D}_{\mathbf{W}_t})^{-1}\mathbf{E}^\top$;
5: Update the diagonal matrix $\mathbf{D}_{\mathbf{W}_{t+1}}$, where the i-th diagonal element is $\frac{1}{2\|\mathbf{W}_{t+1}(i,:)\|_2}$;
6: $t = t + 1$;
7: **end while**
8: Sort each feature according to $\|\mathbf{W}(i,:)\|_2$ in descending order and select the top-K ranked ones.

LEMMA 3.3. *For any non-zero constants x and y, the following inequality holds [Nie et al. 2010].*

$$\sqrt{x} - \frac{x}{2\sqrt{y}} \leq \sqrt{y} - \frac{y}{2\sqrt{y}} \tag{22}$$

PROOF. The detailed proof is similar to that in [Nie et al. 2010]. □

LEMMA 3.4. *The following inequality holds provided that $\mathbf{w}_t^i|_{i=1}^r$ are non-zero vectors, where r is an arbitrary number [Nie et al. 2010].*

$$\begin{aligned}\sum_i \|\mathbf{w}_{t+1}^i\|_2 - \sum_i \frac{\|\mathbf{w}_{t+1}^i\|_2}{2\|\mathbf{w}_t^i\|_2} \\ \leq \sum_i \|\mathbf{w}_t^i\|_2 - \sum_i \frac{\|\mathbf{w}_t^i\|_2^2}{2\|\mathbf{w}_t^i\|_2}\end{aligned} \tag{23}$$

PROOF. Substitute x and y in Eq. (22) by $\|\mathbf{w}_{t+1}^i\|_2$ and $\|\mathbf{w}_t^i\|_2^2$, respectively, we can see that the following inequality holds for any i.

$$\|\mathbf{w}_{t+1}^i\|_2 - \frac{\|\mathbf{w}_{t+1}^i\|_2}{2\|\mathbf{w}_t^i\|_2} \leq \|\mathbf{w}_t^i\|_2 - \frac{\|\mathbf{w}_t^i\|_2^2}{2\|\mathbf{w}_t^i\|_2} \tag{24}$$

[6]Theoretically, $\|\mathbf{W}(i,:)\|_2$ can be zero, then we can regularize $\mathbf{D_W}(i,i)$ as $\mathbf{D_W}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2+\epsilon}$, where ϵ is a very small constant. It is easy to see that when $\epsilon \to 0$, then $\mathbf{D_W}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2+\epsilon}$ approximates $\mathbf{D_W}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2}$

Summing Eq. (24) over i, we see that Eq. (23) holds. □

According to Lemma (3.4), we have the following theorem:

THEOREM 3.5. *At each iteration of Algorithm 1, the value of the objective function in Eq. (17) monotonically decreases.*

PROOF. It can be easily verified that $\mathbf{W}_{t+1}$ in line 4 of Algorithm 1 is the solution to the following problem,

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} tr(\mathbf{W}^\top(\mathbf{B} + \alpha\mathbf{D}_{\mathbf{W}_t})\mathbf{W} - 2\mathbf{E}\mathbf{W}),$$

which indicates that,

$$\begin{aligned} &tr(\mathbf{W}_{t+1}^\top(\mathbf{B} + \alpha\mathbf{D}_{\mathbf{W}_t})\mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) \\ &\quad \leq tr(\mathbf{W}_t^\top(\mathbf{B} + \alpha\mathbf{D}_{\mathbf{W}_t})\mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t) \end{aligned}$$

That is to say,

$$\begin{aligned} &tr(\mathbf{W}_{t+1}^\top\mathbf{B}\mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) + \alpha\sum_i \frac{\|\mathbf{W}_{t+1}(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2} \\ &\leq tr(\mathbf{W}_t^\top\mathbf{B}\mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t) + \alpha\sum_i \frac{\|\mathbf{W}_t(i,:)\|_2^2}{2\|\mathbf{W}_t^(i,:)\|_2} \end{aligned}$$

Then we have the following inequality,

$$\begin{aligned} &tr(\mathbf{W}_{t+1}^\top\mathbf{B}\mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) + \alpha\sum_i \|\mathbf{W}_{t+1}(i,:)\|_2 \\ &\quad - \alpha(\sum_i \|\mathbf{W}_{t+1}(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_{t+1}(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2}) \\ &\leq tr(\mathbf{W}_t^\top\mathbf{B}\mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t) + \alpha\sum_i \|\mathbf{W}_t(i,:)\|_2 \\ &\quad - \alpha(\sum_i \|\mathbf{W}_t(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_t(i,:)\|_2^2}{2\|\mathbf{W}_t^(i,:)\|_2}) \end{aligned}$$

Meanwhile, according to Lemma 3.4, we have,

$$\begin{aligned} &\sum_i \|\mathbf{W}_{t+1}(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_{t+1}(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2} \\ &\leq \sum_i \|\mathbf{W}_t(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_t(i,:)\|_2^2}{2\|\mathbf{W}_t^(i,:)\|_2} \end{aligned}$$

Therefore, we have the following inequality:

$$\begin{aligned} &tr(\mathbf{W}_{t+1}^\top\mathbf{B}\mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) + \alpha\|\mathbf{W}_{t+1}\|_{2,1} \\ &\leq tr(\mathbf{W}_t^\top\mathbf{B}\mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t) + \alpha\|\mathbf{W}_t\|_{2,1} \end{aligned}$$

which indicates that the objective function of Eq. (17) monotonically decreases using the updating rules in Algorithm 1. Since $\mathbf{B} + \alpha\mathbf{D}_{\mathbf{W}_t}$ is a positive definite matrix, the iterative approach in Algorithm 1 converges to an optimal solution, which completes the proof. □

Time Complexity: there are three main operations in Algorithm 1 and the time complexity of each operation can be computed as:

— The time complexity of matrix multiplications when constructing $\mathbf{B}$ is $O(mN)$ since the matrix representations of social media data, $\mathbf{X}$ and $\mathbf{F}$, are very sparse [Tang and Liu 2009].
— The major operation of constructing $\mathbf{E}$ is $\mathbf{Y}^\top\mathbf{X}^\top$, whose time complexity is about $O(n)$ since $\mathbf{X}$ is very sparse [Tang and Liu 2009].
— The most time-consuming operation is to update $\mathbf{W}$ as $\mathbf{W} = (\mathbf{B} + \alpha\mathbf{D_W})^{-1}\mathbf{E}^\top$. According to [Nie et al. 2010], this operation can be efficiently obtained through solving the linear equation:

$$(\mathbf{B} + \alpha\mathbf{D_W})\mathbf{W} = \mathbf{E}^\top \tag{25}$$

the time complexity is $O(m^2k)$.

In summary, the overall time complexity of Algorithm 1 is $\#iterations * O(m^2k + mN)$.

4. TIE STRENGTH FOR LinkedFS

Social media greatly extends the physical boundary of user relationship and allows one to establish relationships with an inordinate number of users. For example, on average, a Facebook user has more than 130 friends [7], and a Twitter user has more than 126 followers [8]. Social correlation theories such as homophily and social influence suggest that the user-user relationship is an important social context and indirectly contains rich information about posts. Among the four basic types of relations extracted by LinkedFS, CoFollowing, CoFollowed and Following are from user-user following relations, represented by a binary matrix $\mathbf{S}$ where $\mathbf{S}(i, j) = 1$ if u_j is followed by u_i, zero otherwise. The adjacency matrix with binary values provides only a coarse representation of user-user relations. Actually the low cost of tie formation can lead to networks with heterogeneous strengths (e.g., weak ties and strong ties mixed together) [Xiang et al. 2010]. Since users with strong ties are likely to exhibit greater topic similarity than those with weak ties, equally treating all relationships will increase the level of noise in the learned models and likely lead to degradation in learning performance [Xiang et al. 2010]. In this section, we first introduce various measures to predict tie strength for user-user following relations, and then investigate how to consider tie strength for LinkedFS with CoFollowing, CoFollowed, and Following relations that involve user-user relations.

Predicting the strength of links has already been applied to a wide range of online applications such as community detection [Tang et al. 2011], relational learning [Macskassy and Provost 2007] and location prediction [Gao et al. 2012]. Tie strength prediction differs from traditional link prediction. The former focuses on modeling the strength of existing links rather than link existence. It aims to infer a continuous-valued strength for relations. Recalling the example of social media data in Figure 1(c), the binary value representation of user-user following relations will be converted into a continuous-valued representation after tie strength prediction, as illustrated in Figure 3. As the major concentration of this section is to how to incorporate tie strength prediction into the proposed framework, LinkedFS, we investigate various measures and finally select four representative types of measures following the basic ideas in [Tang et al. 2012; Xiang et al. 2010; Kahanda and Neville 2009], i.e., structural measure (SM), content measure (CM), interaction measure (IM), and hybrid measure (HM).

[7] http://www.facebook.com/press/info.php?statistics
[8] http://www.guardian.co.uk/technology/blog/2009/jun/29/twitter-users-average-api-traffic

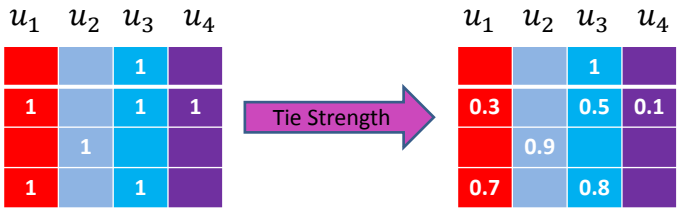


Fig. 3. Tie Strength Prediction

Structural measure: This is used to measure how close two users are in social networks and the intuition behind this measure is that if two users are close in their social network, they are more likely to form a strong tie. We choose three representative quantities in this work as below,

— *Normalized Common Follower (NCF).* For the following relation $u_i \rightarrow u_j$, the normalized common follower is formally defined as,

$$NCF(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{\Gamma(i)}, \tag{26}$$

where $\Gamma(i) = \{x | x \rightarrow u_i\}$ is the set of followers of u_i.

— *Jaccard's coefficient.* Jaccard's coefficient is defined as the number of shared followers of two users divided by the total number of their unique followers. Specifically, for the following relation $u_i \rightarrow u_j$, Jaccard's coefficient is formally stated as,

$$J(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \tag{27}$$

— *Katz Score (KS).* For the following relation $u_i \rightarrow u_j$, this sums all possible paths from u_i to u_j with exponential damping by length to weight short paths more heavily as,

$$KS(i,j) = \sum_{\ell=1}^{\infty} \beta^{\ell} |path_{i,j}^{\ell}|, \tag{28}$$

where $path_{i,j}^{\ell}$ is the set of paths from u_i to u_j with length ℓ (damping factor β is typically set to 0.05).

The Katz score for all pairs of nodes $\mathbf{K} \in \mathbb{R}^{n \times n}$ can be obtained in matrix form through the user-user following relation matrix $\mathbf{S}$ as,

$$\mathbf{K} = \sum_{\ell=1}^{\infty} \beta^{\ell} \mathbf{S}^{\ell} = (\mathbf{I}_n - \beta \mathbf{S})^{-1} - \mathbf{I}_n, \tag{29}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Content measure: Homophily is one of the most important reasons that people form ties with others and it is likely that the stronger the tie, the higher the similarity of their posts. The similarity of user generated content is an important indicator for tie strength [Tang et al. 2012]. Assume that $\mathbf{c}_i$ is the vector support model (VSM) of the posts from u_i as,

$$\mathbf{c}_i = \frac{1}{|\mathbf{F}_i|} \sum_{\mathbf{f}_j \in \mathbf{F}_i} \mathbf{f}_j, \tag{30}$$

where $|\cdot|$ denotes the size of a set. For the following relation $u_i \rightarrow u_j$, the content similarity (CS) is the cosin similarity of $\mathbf{c}_i$ and $\mathbf{c}_j$, formally defined as,

$$CS(u_i, u_j) = \frac{\langle \mathbf{c}_i, \mathbf{c}_j \rangle}{\|\mathbf{c}_i\| \|\mathbf{c}_j\|}, \tag{31}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors and $\|\cdot\|$ is the ℓ_2-norm of a vector.

Interaction measure: a user in social media can have hundreds of ties while a very small number of them are strong. Users with strong ties are likely to communicate more frequently than these with weak ties. For example, in the case of Twitter, users with strong ties are more likely to retweet, write comment on and reply to the tweets posted by their close followers. Interaction activities provide additional information about social networks and attract more attention in recent years. Tie strength prediction is one of the most important applications of interaction activities [Xiang et al. 2010; Kahanda and Neville 2009].

Let $\mathcal{I} = \{I_1, I_2, \ldots, I_M\}$ be the set of interaction types where M is the number of types of interaction. $I_k \in \mathbb{R}^{n \times n}$ denotes the k-th type of interaction and entities in I_k represents the interaction frequency. For example, $I_k(i, j)$ is the interaction frequencies between u_i and u_j. Therefore for each following relation $u_i \rightarrow u_j$, interaction measure is defined as,

$$IM(i, j) = \frac{\sum_{k=1}^{m} I_k(i, j)}{\sum_{j=1}^{n} \sum_{k=1}^{m} S(i, j) I_k(i, j)}, \tag{32}$$

the molecular of Eq. (32) is the total interaction frequencies between u_i and u_j while the denominator is the total interaction frequency of u_i.

Hybrid measure: Hybrid measure comprehensively considers the combinations of structural measure, content measure and interaction measure. In this work, we employ a linear combination of these three measures and for the following relation $u_i \rightarrow u_j$, the hybrid measure is defined as,

$$HM(i, j) = \theta_1 SS(i, j) + \theta_2 CS(i, j) + (1 - \theta_1 - \theta_2) IM(i, j), \tag{33}$$

where $0 \leq \theta_1 \leq 1, \;\; 0 \leq \theta_2 \leq 1, \;\; 0 \leq \theta_1 + \theta_2 \leq 1$.

The relations between hybrid measure and other three measures are stated as: (1) when $\theta_1 = 1$, hybrid measure is boiled down to structural measure; (2) when $\theta_2 = 1$, hybrid measure is equivalent to content similarity; and (3) when $\theta_1 = 0$ and $\theta_2 = 0$, hybrid measure is exactly interaction measure.

After slightly modifying the original framework in Algorithm 1, we can obtain a variant of LinkedFS, incorporating tie strength in Algorithm 2. The differences between Algorithm 1 and Algorithm 2 are: (1) In addition to the inputs of LinkedFS, LinkedFSTS has an extra input, i.e., the user interaction activities; and (2) LinkedFSTS refines the binary adjacency matrix to a continued-value matrix through tie strength prediction.

ALGORITHM 2: LinkedFS with Tie Strength Prediction

Input: $\{\mathbf{F}, \mathbf{X}, \mathbf{Y}, \mathbf{S}, \mathbf{P}, \mathcal{I}\}$ and the number of features expected to select, K;
Output: K most relevant features

1: Refine the adjacency matrix $\mathbf{S}$ via tie strength predictor you choose;
2: LinkedFS($\mathbf{F}$, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{S}$, $\mathbf{P}$);

Table I. Statistics of the Datasets

	BlogCatalog	Digg
# Posts	7,877	9,934
# Original Features	84,233	12,596
# Features after *TFIDF*	13,050	6,544
# Classes	14	15
# Users	2,242	2,561
# Following Relations	55,356	41,544
Ave # Posts	3.5134	3.8790
Max # Followers	820	472
Min # Followers	1	1
Network Density	0.0110	0.0063
Clustering Coefficient	0.3288	0.2461

5. EXPERIMENTS

In this section, we present the experiment details to verify the effectiveness of the proposed framework, LinkedFS. After introducing social media data used in experiments, we first confirm whether linked data contains more additional information than data randomly put together, then study how different relational information affects feature selection performance and how different measures of tie strength affect the performance of LinkedFS. In particular, we investigate how feature selection performance changes with different factors such as number of selected features, the amount of labeled data, which relational hypothesis impacts the performance most, and the relationships between the factors.

5.1. Social Media Data

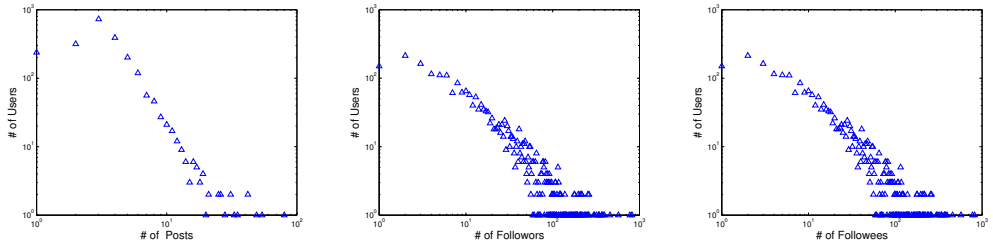
Two real-world social media datasets are collected to evaluate LinkedFS, i.e., Digg and BlogCatalog. For both datasets, we have posts and their social contextual information such as user-post relationships and user-user relationships.

Digg: Digg[9] is a popular social news aggregator that allows users to submit, digg, reply to and comment on stories. It also allows users to create social networks by designating other users as friends and tracking friends' activities. We obtain this dataset from [Lin et al. 2009]. The following relationships form a directed graph and the topics of stories are considered as the class labels. In addition to the content and social context of posts, we have three types of user interaction activities, i.e., digging, replying and commenting, which are used in the interaction measure of tie strength in our experiment.

BlogCatalog: BlogCatalog[10] is a blog directory where users can register their blogs under predefined categories, which is used as class labels of blogs in our work. This dataset is obtained from [Wang et al. 2010]. The "following" relationships in BlogCatalog form an undirected graph, which means the CoFollowing and CoFollowed relationships in this dataset are the same.

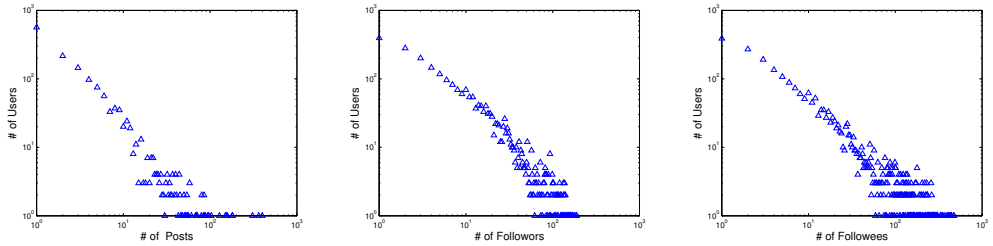
The posts are preprocessed for stop-word removal and stemming. Obviously irrelevant features (terms) are also removed using *TFIDF*, a common practice in information retrieval and text mining [Wu et al. 2008]. We compute the number of posts, followers and followees for each user and the distributions are shown in Figure 4 and Figure 5 for BlogCatalog and Digg, respectively, suggesting power-law distributions. Some other statistics of these datasets are shown in Table I.

[9]http://www.digg.com
[10]http://www.blogcatalog.com



(a) Posts (b) Followers (c) Followees

Fig. 4. Distributions of Posts, Followers and Followees in BlogCatalog



(a) Posts (b) Followers (c) Followees

Fig. 5. Distributions of Posts, Followers and Followees in Digg

5.2. Preliminary Verification

Before conducting extensive experiments for feature selection, we validate if it is worthy of doing so. For the four relations, we form a null hypothesis for each: there is no difference between relational data and random data. If the null hypothesis is rejected, we then proceed to perform extensive experiments for feature selection. The difference is measured by a topic distance (T_{dist}) defined next.

Let $\mathbf{c}_i$ be the class vector for post p_i, where $\mathbf{c}_i(j) = 1$ if p_i belongs to the class c_j, $\mathbf{c}_i(j) = 0$ otherwise. The topic distance between two posts, p_i and p_j, is defined as the distance between their class vectors, $\mathbf{c}_i$ and $\mathbf{c}_j$:

$$T^p_{dist}(p_i, p_j) = \|\mathbf{c}_i - \mathbf{c}_j\|_2. \tag{34}$$

For each post p_i, we construct two vectors $\mathbf{cp}_t(i)$ and $\mathbf{cp}_r(i)$: the former by calculating the average T^p_{dist} between p_i and other posts from the same user, and the latter by calculating the average T^p_{dist} between p_i and randomly chosen posts from other users. The number of randomly chosen posts is the same as the number of co-posts of p_i in the CoPost relation.

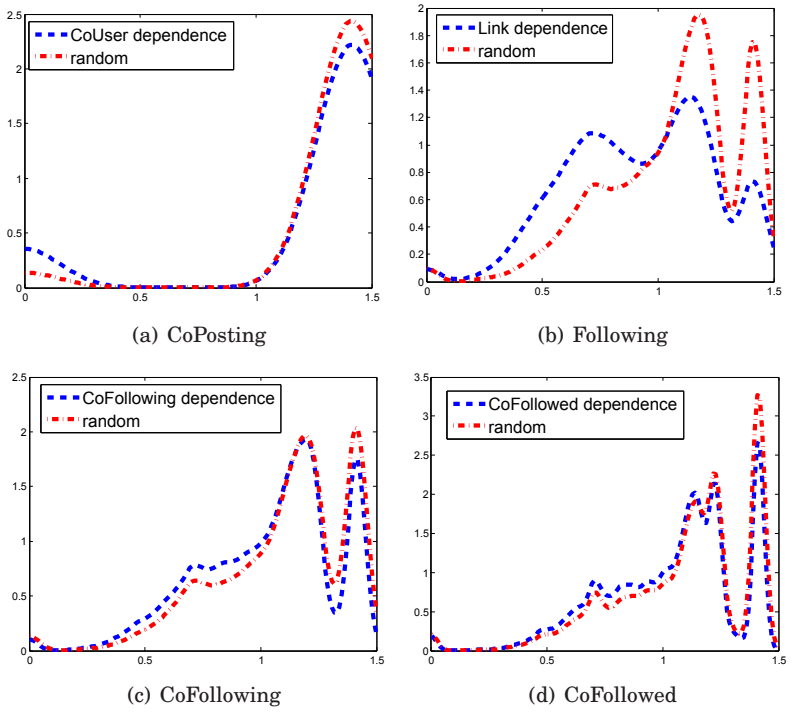
The topic distance between two users is defined as:

$$T^u_{dist}(u_i, u_j) = \|\bar{T}^*(u_i) - \bar{T}^*(u_j)\|_2; \tag{35}$$

where $\bar{T}^*(u_i) \in \mathbb{R}^k$, $\bar{T}^*(u_i) = \frac{\sum_{p_j \in \mathbf{F}_i} \mathbf{c}_j}{|\mathbf{F}_i|}$ is the topic interest distribution of u_i, and its j^{th} element represents the probability of u_i being interested in the j^{th} class. In the same spirit of constructing $\{\mathbf{cp}_t, \mathbf{cp}_r\}$, for each user u_i, we construct $\{\mathbf{cfi}_t(i), \mathbf{cfi}_r(i)\}$, $\{\mathbf{cfe}_t(i), \mathbf{cfe}_r(i)\}$ and $\{\mathbf{fi}_t(i), \mathbf{fi}_r(i)\}$ by calculating their average $T^u_{dist}(u_i, u_j)$ according to CoFollowing, CoFollowed, and Following relations, respectively. The average topic

Table II. Average T_{dist} for pairs with relations vs randomly chosen pairs

Datasets	$\mathbf{cp}_r$	$\mathbf{cp}_t$	$\mathbf{cfi}_r$	$\mathbf{cfi}_t$	$\mathbf{cfe}_r$	$\mathbf{cfe}_t$	$\mathbf{fi}_r$	$\mathbf{fi}_t$
Digg	1.3391	1.0189	1.0834	0.8333	1.0808	0.8345	1.0613	0.9181
BlogCatalog	0.9290	0.6296	0.9893	0.9089	0.9893	0.9089	1.0817	0.8938



(a) CoPosting (b) Following

(c) CoFollowing (d) CoFollowed

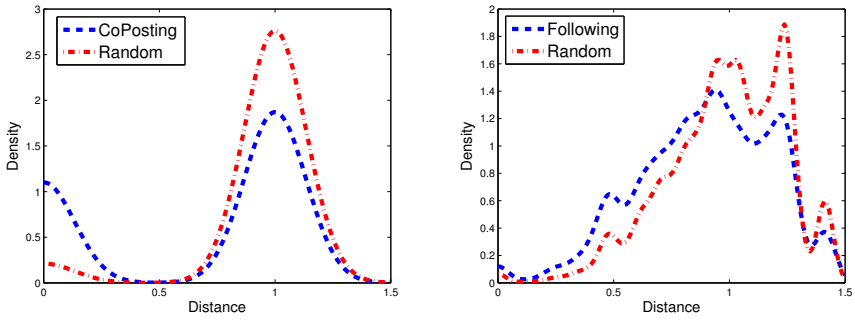
Fig. 6. Density Estimates of T_{dist} for Dependence Hypotheses in Digg

distance is shown in Table II. We note that the average topic distance for pairs with relations is much smaller than that for pairs randomly chosen.

For a visual comparison, in Figures 6 and 7, we plot the Kernel-smoothing density estimations for these pairs of T_{dist} for Digg and BlogCatalog, respectively. These curves show similar patterns: for smaller T_{dist}, the curves for pairs with relations are above those for pairs without, while for larger T_{dist}, they are below, which means that T_{dist} for pairs with relations is more concentrated in smaller topic distance. These results support our hypotheses: posts with CoPost relations are more likely to have similar class labels, and users with CoFollowing, CoFollowed, or Following are more likely to have similar interests in terms of the topics of their posts.

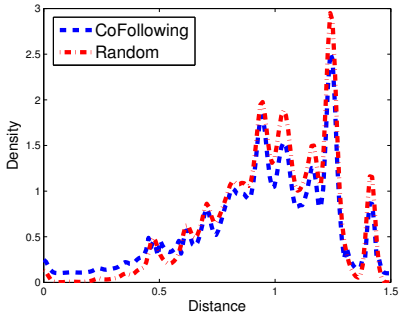
With the four pairs of vectors, we also perform a two-sample t-test on each pair. The null hypothesis, H_0, is that there is no difference between the pair; the alternative hypothesis, H_1, is that the average topic distance following a relation is less than that not. For example, for the CoPost relation, H_0: $\mathbf{cp}_t = \mathbf{cp}_r$, and H_1: $\mathbf{cp}_t < \mathbf{cp}_r$. The t-test results, p-values, are shown in Table III[11]. The star ($*$) next to the p-value means

[11]We use the "ttest2" function from Matlab, which reports p-value as 0 if p-value is too small, i.e., exceeding the decimal places one allows. In our work, we use "<1.00e-14" when Matlab reports p-value as 0, which indicates that it is significant.

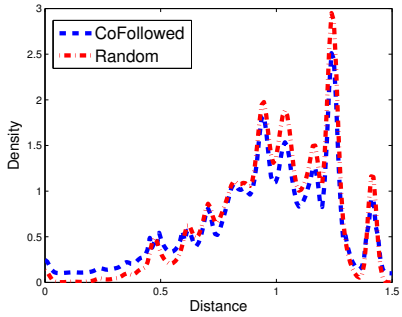


(a) CoPosting

(b) Following



(c) CoFollowing



(d) CoFollowed

Fig. 7. Density Estimates of T_{dist} for Dependence Hypotheses in BlogCatalog

Table III. Statistics of T_{dist} to Support Relation Hypotheses ($\alpha = 0.01$)

	Digg	BlogCatalog
CoPost	<1.00e-14*	<1.00e-14*
CoFollowing	<1.00e-14*	2.80e-8*
CoFollowed	<1.00e-14*	1.23e-8*
Following	<1.00e-14*	1.23e-8*

that there is strong evidence ($p < 0.01$) to reject the null hypothesis. We observe that p-values for all four pairs are close to zero on both datasets. Hence, there is strong evidence to reject the null hypothesis.

The evidence from average topic distance, a visual comparison, and t-test supports that these relations are not random patterns. With preliminary verification, we now check how they help feature selection.

5.3. Quality of Selected Features and Determining Factors

For both datasets, we randomly and evenly (50-50) split data into training data, $\mathcal{T}$ and test data, $\mathcal{U}$. Following [Nie et al. 2010; Zhao et al. 2010], feature quality is assessed via classification performance. If a feature subset is more relevant to the target concept, a classifier trained with the subset should achieve better accuracy [Zhao and Liu 2007]. Linear SVM [Fan et al. 2008] is used for classification. As a common practice, the parameters in feature selection algorithms and SVM are tuned via cross-validation. Since the performance of supervised learning improves with the number of labelled data, we

Table IV. Classification Accuracy of Different Feature Selection Algorithms in Digg

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
$\mathcal{T}_5$	50	45.45	44.50	46.33	45.27	**58.82**	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	**59.43**	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	**63.32**
	300	54.04	55.24	56.54	59.27	65.30	60.40	59.93	**66.19**
$\mathcal{T}_{25}$	50	49.91	50.08	51.54	56.02	**58.90**	57.76	57.01	**58.90**
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	**65.02**
	200	59.97	57.37	60.07	64.36	**67.33**	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	**69.52**	65.46	65.01	67.95
$\mathcal{T}_{50}$	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	**60.77**
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	**65.74**
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	**71.32**
	300	61.47	62.35	64.77	69.58	71.86	71.40	70.50	**72.65**
$\mathcal{T}_{100}$	50	51.74	56.06	55.94	58.08	**61.51**	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	**65.65**
	200	60.49	62.78	65.18	66.87	**69.75**	67.40	67.00	67.31
	300	62.97	66.35	67.12	69.27	**73.01**	70.99	69.50	72.64

fix test data $\mathcal{U}$ and sub-sample from $\mathcal{T}$ to generate training sets of different sizes, $\{\mathcal{T}_5, \mathcal{T}_{25}, \mathcal{T}_{50}, \mathcal{T}_{100}\}$, corresponding to $\{5\%, 25\%, 50\%, 100\%\}$ of $\mathcal{T}$, respectively. Another factor affecting learning performance is the number of features, still an open problem for feature selection [Yang et al. 2011]. Usually, other things being equal, the fewer features, the better. We vary the numbers of selected features as $\{50, 100, 200, 300\}$.

Four representative supervised feature selection algorithms are chosen as baseline methods: ttest (TT), Information Gain (IG), FisherScore (FS) [Duda et al. 2001], and Joint $\ell_{2,1}$-Norms (RFS) [Nie et al. 2010][12] where RFS applies $\ell_{2,1}$ for both loss function and regularization, and FisherScore selects features by assigning similar values to the samples from the same class and different values to samples from different classes. We compare the four baseline methods with four methods based on LinkedFS, i.e., CoPost (CP), CoFollowing (CFI), CoFollowed (CFE), and Following (FI). The results are shown in Tables IV and V for Digg and BlogCatalog, respectively. Since the Following relations in BlogCatalog are undirected, CFI is equivalent to CFE, having the same performance as shown in Table V.

General trends.. As seen in Tables IV and V, the performance of all the methods improves with increasing amount of labeled data. More often than not, with more features selected, the performance also improves. On both datasets, TT, IF, and FS perform comparably and RFS performs best. RFS selects features in batch mode and considers feature correlation. It is consistent with what was suggested in [Yang et al. 2011; Zhao et al. 2010], that it is better to analyze instances and features jointly for feature selection.

Comparison with baselines.. Our proposed methods, CP, CFI, CFE, and FI, consistently outperform all baseline methods on both datasets. Comparing with the best performance of baseline methods, the relative improvement of our methods is obtained and then averaged over different numbers of features. The results are given in Table VI. It is clear that CP and FI achieve better performance than CFI and CFE. That is, CoPost and Following hypotheses hold more strongly than CoFollowing and CoFollowed in our studied datasets.

[12]We obtain the code for TT, IG and FS from featureselection.asu.edu, and RFS from the first author's webpage(http://sites.google.com/site/feipingnie)

Table V. Classification Accuracy of Different Feature Selection Algorithms in BlogCatalog

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFE	CFI	FI
$\mathcal{T}_5$	50	46.54	40.96	41.31	46.16	**53.37**	53.01	53.01	52.84
	100	46.77	43.08	43.02	48.81	53.44	52.48	52.48	**53.82**
	200	46.84	44.06	45.66	50.77	55.94	53.61	53.61	**57.30**
	300	46.91	44.59	43.93	52.73	57.22	55.13	55.13	**60.02**
$\mathcal{T}_{25}$	50	48.13	40.58	45.44	47.60	**53.40**	53.24	53.24	52.79
	100	48.42	41.94	46.34	51.47	**57.02**	53.62	53.62	56.57
	200	48.05	43.45	53.07	53.64	58.83	55.81	55.81	**60.50**
	300	47.44	42.32	54.58	60.29	**65.56**	61.00	61.00	63.67
$\mathcal{T}_{50}$	50	48.66	52.21	48.23	52.51	56.22	53.47	53.47	**56.97**
	100	49.11	51.61	50.72	55.38	**59.32**	56.00	56.00	57.43
	200	48.43	51.54	53.74	62.02	**68.08**	63.58	63.58	65.66
	300	48.20	52.21	53.67	61.78	**70.95**	63.75	63.75	68.76
$\mathcal{T}_{100}$	50	50.54	54.33	52.39	54.55	**58.34**	55.31	55.31	55.92
	100	50.32	53.89	52.99	57.11	60.45	58.20	58.20	**65.51**
	200	50.77	54.02	54.80	66.33	**70.81**	63.11	63.11	68.31
	300	49.03	54.45	56.84	63.26	69.06	65.40	65.40	**69.89**

Table VI. Classification Accuracy Improvement of the Proposed Methods

	Improvement in Digg(%)			
Datasets	CP	CFI	CFE	FI
$\mathcal{T}_5$	+14.54	+7.01	+4.69	+15.25
$\mathcal{T}_{25}$	+4.59	+1.59	0	+4.02
$\mathcal{T}_{50}$	+7.19	+3.92	+1.05	+8.48
$\mathcal{T}_{100}$	+4.90	+3.15	+1.63	+4.64
	Improvement in BlogCatalog(%)			
Datasets	CP	CFI	CFE	FI
$\mathcal{T}_5$	+10.71	+7.89	+7.89	+12.62
$\mathcal{T}_{25}$	+10.04	+5.00	+5.00	+9.50
$\mathcal{T}_{50}$	+9.70	+2.16	+2.16	+7.34
$\mathcal{T}_{100}$	+7.18	+0.46	+0.46	+7.67

Relationships between amounts of labeled data and types of relations.. Table VI also says that our methods work more effectively when using small amounts of labeled data. For example, in Digg, CP is better than the best baseline by 14.54% with $\mathcal{T}_5$, but only by 4.9% with $\mathcal{T}_{100}$. In Tables IV and V, if we select, for instance, 50 features, the performance using linked data with $\mathcal{T}_5$ is comparable with that without linked with $\mathcal{T}_{100}$. In other words, linked data compensates the shortage of labeled data. The finding has its practical significance as in social media; it is not easy to obtain labeled data but there is often abundant linked data.

5.4. Impact of Tie Strength on LinkedFS

In this subsection, we investigate the impact of tie strength on LinkedFS. Since Co-Following(CFI), CoFollowed(CFE) and Following(FI) are extracted from user-user relations, therefore we study how tie strength affects their performance. Through the above study in section 5.3, most of the time, LinkedFS achieves the best performance when the number of selected features is 300 thus we fix the number of selected features to 300 in the following experiments. To save space, we only report the results in $\mathcal{T}_5$ and $\mathcal{T}_{100}$ since we have similar observations on $\mathcal{T}_{25}$ and $\mathcal{T}_{50}$. The results are demonstrated in Tables VII and VIII for BlogCatalog and Digg, respectively.

The experimental settings of hybrid measure are different for BlogCatalog and Digg. For BlogCatalog, the interaction information is not available thus we vary θ_1 from 0 to 1 with an incremental step of 0.1 and $\theta_2 = 1 - \theta_1$. For Digg, to systematically evaluate

Table VII. Impact of tie strength on LinkedFS in BlogCatalog. Note that "Imp" denotes improvement over performance without tie strength.

Measures		$\mathcal{T}_5$			$\mathcal{T}_{100}$		
		CFI (Imp)	CFE (Imp)	FI (Imp)	CFI (Imp)	CFE (Imp)	FI (Imp)
Without Tie Strength		55.13	55.13	60.02	65.40	65.40	69.89
Structural Measure	NCF	57.50 (+4.30)	57.50 (+4.30)	60.59 (+0.95)	67.92 (+3.85	67.92 (+3.85)	70.41 (+0.74)
	Jaccard	57.79 (+4.83)	57.79 (+4.83)	60.84 (+1.37)	68.27 (+4.39)	68.27 (+4.39)	70.39 (+0.72)
	KS	57.93 (+5.08)	57.93 (+5.08)	61.08 (+1.77)	68.72 (+5.08)	68.72 (+5.08)	70.65 (+1.09)
Content Measure		58.88 (+6.80)	58.88 (+6.80)	62.32 (+3.83)	69.68 (+6.54)	69.68 (+6.54)	71.77 (+2.69)
Hybrid Measure	$\theta_1 = 0.2$,$\theta_2 = 0.8$	59.17 (+7.33)	59.17 (+7.33)	62.74 (+4.53)	70.01 (+7.05)	70.01 (+7.05)	72.34 (+3.51)

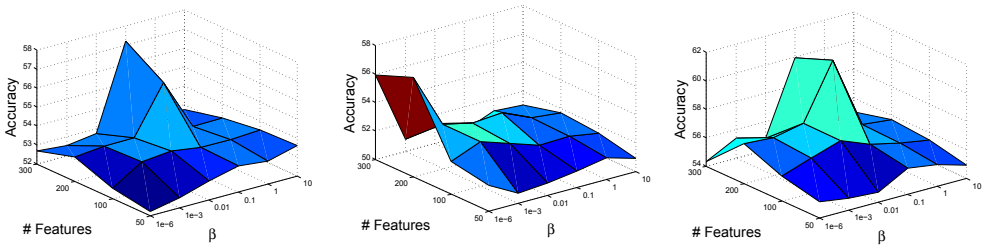
Table VIII. Impact of tie strength on LinkedFS in Digg. Note that "Imp" denotes improvement over performance without tie strength.

Measures		$\mathcal{T}_5$			$\mathcal{T}_{100}$		
		CFI (Imp)	CFE (Imp)	FI (Imp)	CFI (Imp)	CFE (Imp)	FI (Imp)
Without Tie Strength		60.40	59.93	66.19	70.99	69.50	72.74
Structural Measure	NCF	62.51 (+3.49)	62.50 (+4.29)	66.92 (+1.10)	71.38 (+0.55)	70.99 (+2.14)	73.14 (+0.55)
	Jaccard	62.62 (+3.68)	62.38 (+4.09)	67.29 (+1.66)	71.27 (+0.39)	71.11 (+2.32)	72.95 (+0.29)
	KS	62.97 (+4.26)	62.43 (+4.17)	67.23 (+1.57)	71.51 (+0.73)	71.29 (+2.58)	73.27 (+0.73)
Content Measure		63.91 (+5.81)	63.20 (+5.46)	68.04 (+2.80)	72.21 (+1.72)	71.90 (+3.45)	73.98 (+1.70)
Interaction Measure		64.54 (+6.85)	64.01 (+6.81)	68.57 (+3.60)	72.84 (+2.61)	72.59 (+4.45)	74.25 (+2.08)
Hybrid Measure	$\theta_1 = 0.3$,$\theta_2 = 0.7$	64.31 (+6.47)	63.78 (+6.42)	68.48 (+3.46)	72.59 (+2.25)	72.05 (+3.67)	74.44 (+2.34)
	$\theta_1 = 0.1$,$\theta_2 = 0$	64.62 (+6.99)	64.34 (+7.36)	68.75 (+3.87)	73.04 (+2.89)	72.88 (+4.86)	74.89 (+2.96)
	$\theta_1 = 0$,$\theta_2 = 0.4$	65.20 (+7.95)	64.74 (+8.03)	69.18 (+4.52)	73.87 (+4.06)	73.59 (+5.88)	75.40 (+3.66)
	$\theta_1 = 0.1$,$\theta_2 = 0.4$	65.31 (+8.13)	64.84 (+8.19)	69.23 (+4.59)	74.07 (+4.34)	73.84 (+6.24)	75.77 (+4.17)

different combinations of structural measure, content measure and interaction measure, we vary both θ_1 and θ_2 from 0 to 1 with an incremental step of 0.1 and report: (1) the best performance when combining structural measure and content measure, i.e., $\theta_1 + \theta_2 = 1$; (2) the best performance when combining structural measure and interaction measure, i.e., $\theta_2 = 0$; (3) the best performance when combining content measure and interaction measure, i.e., $\theta_1 = 0$; and (4) the best performance when combining these three measures.

We have the following observations,

— LinkedFS with tie strength consistently outperforms itself without. For example, on average, LinkedFS with hybrid measure gains 6.15% relative improvement in BlogCatalog. These results support that LinkedFS benefits from incorporating tie strength in user-user following relations.
— CoFollowing relation and CoFollowed relation obtain more improvement than Following relation although Following relation still gets the best performance. For example,



(a) CoPost (b) CoFollowing (c) Following

Fig. 8. Performance Variation of Our Methods in $\mathcal{T}_5$ from BlogCatalog Dataset

on average, CoFollowing gains more than 5% relative improvement while Following improves less than 3% in Digg.

— Among three quantities of structural measure, KS always performs best. Jaccard and NCF are defined on users' local networks, i.e., users and their followers (local information), while KS is computed on the whole social network (global information). These results indicate that global information is more important than local information for structural measure in terms of tie strength prediction. Among the three measures of tie strength, interaction measure gains the best performance. It demonstrates that users with more interactions are likely to have similar interests.

— In BlogCatalog, hybrid measure always performs best. The best performance is achieved when $\theta_1 = 0.2$ and $\theta_2 = 0.8$, controlling the contributions from structural measure and content measure. A large weight is given to content measure, indicating that content measure is important in terms of tie strength prediction. It also suggests that an appropriate combination of these two measures is crucial to achieving higher accuracy.

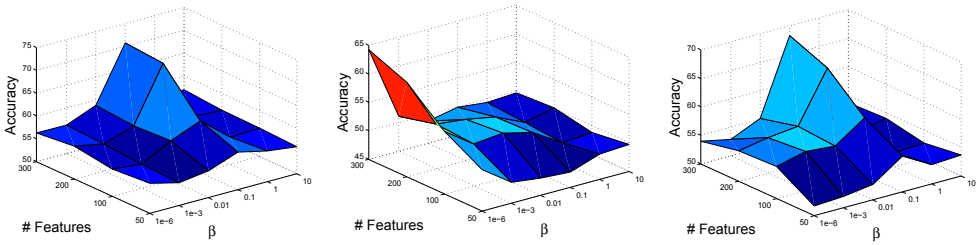
— In Digg, properly combining measures can consistently improve performance. The best performance is achieved when combining all three measures, suggesting that structural measure, content measure, and interaction measure provide complementary information to each other. We also note that the best performance of the combination of all three measures is very close to that of combination of content measure and interaction measure, indicating that structure contains limited extra information beyond the combination of content measure and interaction measure for tie strength prediction.

These observations demonstrate that considering heterogeneous strengths of user-user following relations can improve the performance of LinkedFS.

5.5. Effects of β and Numbers of Selected Features

An important parameter in LinkedFS is β that determines the impact of a relation on feature selection. A high value indicates the importance of this relation, or that the corresponding hypothesis holds strongly. Another important parameter is the number of selected features. In this subsection, we investigate how β and the number of selected features affect the performance of LinkedFS. In particular, we study how the performance of CP, CFI, CFE, and FI varies with β and the number of selected features. We vary β as $\{1e{-}6, 1e{-}3, 0.1, 1, 10\}$ and the number of selected features are varied as $\{50, 100, 200, 300\}$.

To save space, we only show the results in $\mathcal{T}_5$ and $\mathcal{T}_{50}$ of BlogCatalong in Figure 8 and Figure 9, respectively since we have similar observations with other settings. Since CFI and CFE are equivalent for BlogCatalog, there are only three plots for CP,



(a) CoPost (b) CoFollowing (c) Following

Fig. 9. Performance Variation of Our Methods in $\mathcal{T}_{50}$ from BlogCatalog Dataset

CFI, and FI. With the increase of β, most of the time, the performance first increases, reaches its peak value and then decreases. CP and FI achieve the peak performance with $\beta = 0.1$, and CFI with $\beta = 1e$-6. CoPost and Following hypotheses hold more strongly than CoFollowing and CoFollowed hypotheses in the two datasets. Among the two parameters, performance is relatively more sensitive to the number of selected features. As pointed in [Yang et al. 2011], how to determine the number of selected features is still an open problem.

5.6. Combinations of Multiple Relations

The focus of this paper is to study the effect of each hypothesis (CoPost, CoFollowing, CoFollowed, and Following) on feature selection. However, it may be interesting in discussing some results on how the combinations of hypotheses would affect the feature selection results. The formulation of combining four types of hypotheses is presented as

$$\begin{aligned}\min_{\mathbf{W}} \quad & \|\mathbf{X}^\top\mathbf{W} - \mathbf{Y}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta_{cp}\,\Omega(CoPost) \\ & + \beta_{fi}\,\Omega(CoFollowing) + \beta_{fe}\,\Omega(CoFollowed) + \beta_f\,\Omega(Following) \end{aligned} \tag{36}$$

where $\Omega(CoPost)$, $\Omega(CoFollowing)$, $\Omega(CoFollowed)$, and $\Omega(Following)$ denote the regularization terms for CoPost, CoFollowing, CoFollowed, and Following hypothesis, whose contributions are controlled by β_{cp}, β_{fi}, β_{fe}, and β_f, respectively. We can study the effects of combinations of hypotheses on the results by investigating these four parameters - β_{cp}, β_{fi}, β_{fe}, and β_f. For example, by setting two of them to zero and others to nonzero, the resulting frameworks are these with combinations of two hypotheses. We only show the results with 50 selected features on 5% and 50% training data in the BlogCatalog dataset since we have similar observation with other settings. The results are demonstrated in Table IX. Note that "Optimal" in the table denotes a non-zero optimal value of the corresponding parameter, which is obtained via cross-validation.

We make following observations

— Combining multiple hypotheses can obtain better performance. Note that sometimes the combination can lead to the same performance due to the fact that in BlogCatalog, the user-user following graph is undirected, and CoFollowing hypothesis is the same as CoFollowing hypothesis.
— combining CoPost hypothesis with CoFollowing, CoFollowed or Following hypotheses always outperforms other combinations. CoPost hypothesis is extracted from user-post relations, while CoFollowing, CoFollowed and Following hypotheses are from user-user following relations. These results support that user-post relations contain complementary information to user-user following relations. Combinations with hy-

Table IX. The Performance of Frameworks with Combinations of Multiple Hypotheses in BlogCatalog; Note that "Optimal" in the table denotes a non-zero optimal value of the corresponding parameter.

β_{cp}	β_{fi}	β_{fe}	β_f	BlogCatalog(5%)	BlogCatalog(50%)
Optimal	Optimal	0	0	56.45	58.01
Optimal	0	Optimal	0	56.45	58.01
Optimal	0	0	Optimal	55.81	57.74
0	Optimal	Optimal	0	53.01	53.47
0	Optimal	0	Optimal	53.19	57.11
0	0	Optimal	Optimal	53.19	57.11
Optimal	Optimal	Optimal	0	56.45	58.01
Optimal	Optimal	0	Optimal	57.01	58.63
Optimal	0	Optimal	Optimal	57.01	58.63
0	Optimal	Optimal	Optimal	53.19	53.61
Optimal	Optimal	Optimal	Optimal	57.01	58.63

Table X. Comparison of Different Classifiers in Digg.

Datasets	LS	GLS	SVM	$\mathbf{W}_{cp}$	$\mathbf{W}_{fi}$	$\mathbf{W}_{fe}$	$\mathbf{W}_f$
$\mathcal{T}_5$	57.82	60.25	59.63	65.54	62.95	62.43	67.02
$\mathcal{T}_{25}$	63.27	67.00	66.72	69.96	68.49	68.73	69.79
$\mathcal{T}_{50}$	65.93	70.54	68.85	72.57	70.99	71.24	73.00
$\mathcal{T}_{100}$	67.77	71.51	69.95	73.42	72.15	71.89	73.39

potheses from different sources perform better than these with hypotheses from the same sources.

5.7. Applying LinkedFS

LinkedFS can be applied to many applications of social media data such as classification, clustering and visualization. Since LinkedFS can learn a linear classifier $\mathbf{W}$, the immediate application of LinkedFS is to apply $\mathbf{W}$ to classify unseen posts.

We choose representative classifiers as baseline methods including least square(LS), least square with group lasso (GLS) and SVM. $\mathbf{W}_{cp}$, $\mathbf{W}_{fi}$, $\mathbf{W}_{fe}$, and $\mathbf{W}_f$ denote the classifiers learned with CoPost, CoFollowing, CoFollowed and Following hypothesis, respectively. The comparison results in Digg are shown in Table X.

We note that with the increase of training data, the performance of all classifier trends to increase. GLS outperforms LS and SVM. The major reason is that GLS embeds feature selection in the learning process and reduces the effect of curse of dimensionality. Classifiers learned by LinkedFS always obtain better performance than baseline methods. The key reasons are two-fold - (1) LinkedFS embeds feature selection in the classifier learning process; (2) LinkedFS exploits link information, which is especially important when the labeled data is small.

6. RELATED WORK

Feature selection methods fall into three categories, i.e., the filter model, the wrapper model and embedded model [Liu and Yu 2005]. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. Widely used filter-type feature selection methods include t-test, Information Gain, ReliefF and its multi-class extension ReliefF [Robnik-Šikonja and Kononenko 2003], mRmR [Peng et al. 2005], LaplacianScore [He et al. 2006] and its extensions [Zhao and Liu 2007]. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance. Methods in that category include supervised algorithms [Guyon et al. 2002; Duda et al. 2001] and unsupervised algorithms [Roth and Lange 2004; Dy and Brodley 2004; Con-

stantinopoulos et al. 2006]. However, these methods are usually computationally expensive [Liu and Yu 2005] and may not be able to be applied on large scale data mining problems. For the embedded model, the procedure of feature selection is embedded directly in the training process [Cawley and Talbot 2006; Cawley et al. 2006].

Recently sparsity regularization such as $\ell_{2,1}$ of a matrix in dimensionality reduction has been widely investigated and also applied into feature selection studies. The $\ell_{2,1}$ of a matrix is first introduced in [Ding et al. 2006] as rotational invariant ℓ_1 norm. A similar model for $\ell_{2,1}$-norm regularization is proposed in [Argyriou et al. 2007; Liu et al. 2009] to couple feature selection across tasks. Nie et al. [Nie et al. 2010] introduced a robust feature selection method emphasizing joint $\ell_{2,1}$-norm minimization on both loss function and regularization. The $\ell_{2,1}$-norm based loss function is robust to outliers in data points and $\ell_{2,1}$-norm regularization selects features across all data points with joint sparsity. Zhao et al. [Zhao et al. 2010] proposes a spectral feature selection algorithm based on a sparse multi-output regression with a $\ell_{2,1}$ norm constraint, which can do well in both selecting relevant features and removing redundancy. Yang et al. [Yang et al. 2011] proposed a joint framework for unsupervised feature selection which incorporates discriminative analysis and $\ell_{2,1}$-norm minimization. Different from existing unsupervised feature selection algorithms, this proposed algorithm selects the most discriminative feature subset from the whole feature set in batch mode.

The methods above focus on attribute-value data that is independent and identically distributed. There are recent developments that try to address relational data. In [Jensen and Neville 2002], the authors propose the problem of *relational feature selection*. Relational features are different from traditional features. A relational feature is, as an example in [Jensen and Neville 2002], $Max(Age(Y)) > 65\ where\ Movie(x),\ Y = \{y|ActedIn(x, y)\}$ where $ActedIn$ is a relation that connects two objects x and y. Relational feature selection identifies a particular relation that links a single object to a set of other objects. Feature selection with linked data (or LinkedFS) still selects traditional features. Since LinkedFS involves more than one type (or source) of data such as user-post relationships and user-user relationships, it is related to *multi-source feature selection* (MSFS) [Zhao and Liu] with the following differences: (1) sources in MSFS are different views of the same objects while additional sources in LinkedFS are different types of relations; and (2) MSFS and LinkedFS take different approaches to data of different sources: MSFS linearly combines multiple sources to a single source before applying single source feature selection, and LinkedFS considers a relation as a constraint.

LinkedFS uses link information as auxiliary information in the learning process. In this sense, our work is related to algorithms using "Hints" as auxiliary information in learning. A good review of "Hints" can be found in [Abu-Mostafa 1995]. However, LinkedFS is very different from hints. First, LinkedFS incorporates link information by extracting four types of relations based on social theories, while hints will be represented and incorporated into the objective function as virtual examples. Second, link information in LinkedFS indicates the correlations among the training samples, while hints are the auxiliary information about the target function.

LinkedFS adds a regularization term to incorporate link information, which is related to Tikhonov regularization [Golub et al. 1999]. We will use LinkedFS with CoPost relation to illustrate their connections, which can be written as,

$$\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \beta \mathbf{F}\mathbf{L}_\mathbf{A}\mathbf{F}^\top + \alpha \|\mathbf{W}\|_{2,1}. \tag{37}$$

Since $\mathbf{L_A}$ is the Laplacian matrix, we can perform eigen-decomposition on $\mathbf{L_A}$ as $\mathbf{L_A} = \mathbf{U}\Sigma\mathbf{U}^\top$. Then $\mathbf{F}\mathbf{L_A}\mathbf{F}^\top$ can be written as $\mathbf{F}\mathbf{L_A}\mathbf{F}^\top = \mathbf{F}\mathbf{U}\Sigma^{\frac{1}{2}}(\mathbf{F}\mathbf{U}\Sigma^{\frac{1}{2}})^\top$. By setting

$\Gamma = \mathbf{F}\mathbf{U}\Sigma^{\frac{1}{2}}$, Eq. (37) can be rewritten as

$$\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \beta\|\Gamma\mathbf{W}\| + \alpha\|\mathbf{W}\|_{2,1}. \quad (38)$$

where Eq. (38) is a special case of Tikhonov regularization by setting $\alpha = 0$.

7. CONCLUSIONS

Social media data differs from traditional data used in data mining. It presents new challenges to feature selection. In this work, we suggest to research a novel problem - feature selection for linked social media data. In particular, we extract four types of relations from linked data and propose a novel framework (LinkedFS) that integrates relational constraint into a state-of-the-art feature selection formulation. We further show that an optimal solution can be developed, and tie strength prediction can be incorporated into LinkedFS. Extensive experiments are conducted to show its efficacy and the relationships among several factors intrinsic to feature selection: numbers of selected features, percentages of labeled data, and importance of four types of relations in performance improvement.

This work aims to show the effectiveness of using linked data for feature selection. Our future work will extended to study the combination of relations in a general model that can efficiently determine their contributions to feature selection, exploring additional and relevant information hidden in social media, and develop an open-source platform for collaborative research in this challenging new direction of feature selection.

ACKNOWLEDGMENTS

The work is, in part, supported by NSF grants #0812551 and IIS-1217466.

REFERENCES

ABU-MOSTAFA, Y. S. 1995. Hints. *Neural Computation 7,* 4, 639–671.

ARGYRIOU, A., EVGENIOU, T., AND PONTIL, M. 2007. Multi-task feature learning. *NIPS 19*, 41.

CAWLEY, G. AND TALBOT, N. 2006. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics 22,* 19, 2348.

CAWLEY, G., TALBOT, N., AND GIROLAMI, M. 2006. Sparse multinomial logistic regression via bayesian l1 regularisation. In *NIPS*. Vol. 19. 209.

CONSTANTINOPOULOS, C., TITSIAS, M., AND LIKAS, A. 2006. Bayesian feature and model selection for gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1013–1018.

DING, C., ZHOU, D., HE, X., AND ZHA, H. 2006. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 281–288.

DUDA, R., HART, P., STORK, D., ET AL. 2001. *Pattern classification*. Vol. 2. wiley New York.

DY, J. AND BRODLEY, C. 2004. Feature selection for unsupervised learning. *The Journal of Machine Learning Research 5*, 845–889.

FAN, R., CHANG, K., HSIEH, C., WANG, X., AND LIN, C. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research 9*, 1871–1874.

GAO, H., TANG, J., AND LIU, H. 2012. Exploring social-historical ties on location-based social networks. In *Sixth International AAAI Conference on Weblogs and Social Media*.

GOLUB, G. H., HANSEN, P. C., AND O'LEARY, D. P. 1999. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications 21,* 1, 185–194.

GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. 2002. Gene selection for cancer classification using support vector machines. *Machine learning 46,* 1, 389–422.

HE, X., CAI, D., AND NIYOGI, P. 2006. Laplacian score for feature selection. *NIPS 18*, 507.

JENSEN, D. AND NEVILLE, J. 2002. Linkage and autocorrelation cause feature selection bias in relational learning. In *In Proceedings of the 19th International Conference on Machine Learning*. Citeseer, 259–266.

KAHANDA, I. AND NEVILLE, J. 2009. Using transactional information to predict link strength in online social networks. In *Proceedings of the Third International Conference on Weblogs and Social Media*.

LIN, Y., SUN, J., CASTRO, P., KONURU, R., SUNDARAM, H., AND KELLIHER, A. 2009. Metafac: community discovery via relational hypergraph factorization. In *ACM SIGKDD*. ACM, 527–536.

LIU, H. AND MOTODA, H. 2008. *Computational methods of feature selection*. Chapman & Hall.

LIU, H. AND YU, L. 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering 17,* 4, 491.

LIU, J., JI, S., AND YE, J. 2009. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 339–348.

MACSKASSY, S. AND PROVOST, F. 2007. Classification in networked data: A toolkit and a univariate case study. *The Journal of Machine Learning Research 8*, 935–983.

MARSDEN, P. AND FRIEDKIN, N. 1993. Network studies of social influence. *Sociological Methods and Research 22,* 1, 127–151.

MCPHERSON, M., LOVIN, L. S., AND COOK, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27,* 1, 415–444.

MORRIS, S. 2005. Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology 56,* 12, 1250–1273.

NIE, F., HUANG, H., CAI, X., AND DING, C. 2010. Efficient and robust feature selection via joint l21-norms minimization. NIPS.

PENG, H., LONG, F., AND DING, C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 1226–1238.

ROBNIK-ŠIKONJA, M. AND KONONENKO, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning 53,* 1, 23–69.

ROTH, V. AND LANGE, T. 2004. Feature selection in clustering problems. *NIPS 16*, 473–480.

TANG, J., GAO, H., AND LIU, H. 2012. mtrust: Discerning multi-faceted trust in a connected world. In *the 5th ACM International Conference on Web Search and Data Mining*.

TANG, J., WANG, X., AND LIU, H. 2011. Integrating social media data for community detection. In *MSM-MUSE*.

TANG, L. AND LIU, H. 2009. Scalable learning of collective behavior based on sparse social dimensions. In *Proceeding of the 18th ACM conference on Information and knowledge management*. ACM, 1107–1116.

TASKAR, B., ABBEEL, P., WONG, M., AND KOLLER, D. 2003. Label and link prediction in relational data. In *Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data*. Citeseer.

WANG, X., TANG, L., GAO, H., AND LIU, H. 2010. Discovering overlapping groups in social media. In *2010 IEEE International Conference on Data Mining*. IEEE, 569–578.

WU, H., LUK, R., WONG, K., AND KWOK, K. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS) 26,* 3, 1–37.

XIANG, R., NEVILLE, J., AND ROGATI, M. 2010. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web*. ACM, 981–990.

YANG, Y., SHEN, H., MA, Z., HUANG, Z., AND ZHOU, X. 2011. L21-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*.

ZHAO, Z. AND LIU, H. Multi-source feature selection via geometry-dependent covariance analysis. In *Journal of Machine Learning Research, Workshop and Conference Proceedings*. Vol. 4. Citeseer, 36–47.

ZHAO, Z. AND LIU, H. 2007. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*. ACM, 1151–1157.

ZHAO, Z., WANG, L., AND LIU, H. 2010. Efficient spectral feature selection with minimum redundancy. In *Proceedings of the Twenty-4th AAAI Conference on Artificial Intelligence (AAAI)*.

Appendix A.

In this appendix, we give the detailed theorem about how to incorporate CoFollowed relations for feature selection.

CoFollowed Relation: Let **FE** be the CoFollowed matrix where $\mathbf{FE}(i, j) = 1$ if u_i and u_j are followed by at least one other person u_k. **FE** can be obtained from the

adjacency matrix $\mathbf{S}$, i.e., $\mathbf{FE} = sign(\mathbf{SS}^\top)$. Let $\mathbf{L_{FE}}$ be the Laplacian matrix defined on $\mathbf{FE}$.

THEOREM 7.1. *The formulation for CoFollowed relation is equivalent to the following optimization problem:*

$$\min_{\mathbf{W}} tr(\mathbf{W}^\top \mathbf{B}\mathbf{W} - 2\mathbf{E}\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \tag{39}$$

where $\mathbf{B}$ *and* $\mathbf{E}$ *are defined as follows:*

$$\begin{aligned} \mathbf{B} &= \mathbf{X}\mathbf{X}^\top + \beta\mathbf{F}\mathbf{H}\mathbf{L_{FE}}\mathbf{H}^\top\mathbf{F}^\top \\ \mathbf{E} &= \mathbf{Y}^\top\mathbf{X}^\top \end{aligned} \tag{40}$$

PROOF. The proof process is similar to that for CoFollowing relation thus we ignore the proof to save space. □

Appendix B.

In this appendix, we develop the theorem providing details about how to take advantage of Following relation for feature selection.

Following Relation: Let $\mathbf{L_S}$ be the Laplacian matrix defined on the user-user following relations $\mathbf{S}$.

THEOREM 7.2. *The formulation for Following relation is equivalent to the following optimization problem:*

$$\min_{\mathbf{W}} tr(\mathbf{W}^\top \mathbf{B}\mathbf{W} - 2\mathbf{E}\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \tag{41}$$

where $\mathbf{B}$ *and* $\mathbf{E}$ *are defined as follows:*

$$\begin{aligned} \mathbf{B} &= \mathbf{X}\mathbf{X}^\top + \beta\mathbf{F}\mathbf{H}\mathbf{L_S}\mathbf{H}^\top\mathbf{F}^\top \\ \mathbf{E} &= \mathbf{Y}^\top\mathbf{X}^\top \end{aligned} \tag{42}$$

PROOF. The proof process is similar to that for CoFollowing relation thus we ignore the proof to save space. □