Collaboration Based Multi-Label Propagation for Fraud Detection

Haobo Wang[1,2], **Zhao Li**[2*], **Jiaming Huang**[2], **Pengrui Hui**[2], **Weiwei Liu**[3*], **Tianlei Hu**[1] and **Gang Chen**[1]
[1]Key Lab of Intelligent Computing Based Big Data of Zhejiang Province, Zhejiang University
[2]Alibaba Group, Hangzhou, China
[3]School of Computer Science, Wuhan University
wanghaobo@zju.edu.cn, {lizhao.lz, jimmy.hjm, pengrui.hpr}@alibaba-inc.com, liuweiwei863@gmail.com, {htl, cg}@zju.edu.cn

Abstract

Detecting fraud users, who fraudulently promote certain target items, is a challenging issue faced by e-commerce platforms. Generally, many fraud users have different spam behaviors simultaneously, e.g. spam transactions, clicks, reviews and so on. Existing solutions have two main limitations: 1) the correlations among multiple spam behaviors are neglected; 2) large-scale computations are intractable when dealing with an enormous user set. To remedy these problems, this work proposes a collaboration based multi-label propagation (CMLP) algorithm. We first introduce a generic version that involves collaboration technique to exploit label correlations. Specifically, it breaks the final prediction into two parts: 1) its own prediction part; 2) the prediction of others, i.e. collaborative part. Then, to accelerate it on large-scale e-commerce data, we propose a heterogeneous graph based variant that detects communities on the user-item graph directly. Both theoretical analysis and empirical results clearly validate the effectiveness and scalability of our proposals.

1 Introduction

The rapid growth of information technologies enables billions of people to shop online. E-commerce platforms connect customers with factories, stores, and third-party merchants, providing them a convenient, reliable and fast manner of shopping. Meanwhile, e-commerce has brought huge economic benefits to society. For instance, in the fiscal year of 2018, eBay GMV (gross merchandise volume) is reported to reach US $95 billion[1] and Alibaba GMV is reported to reach US $673 billion[2].

Most e-commerce platforms calculate a ranking index and a reputation factor for sellers using the number of actions, e.g. clicks, purchases, and reviews. In general, items with a better ranking index will be listed in the front of the search results,

Service	Description
Spam Search & Clicking	Spammers use specific query to complete a search link and click target items, attemping to promote Click through Rate (CTR) and the number of clicks.
Spam Add-to-Cart	Spammers seek a specific item or service in a fraudulent way, then add target item into online shopping cart. Their purpose is to fake Add-to-Cart factor and receive over exposures.
Spam Transactions	Spammers are asked to make certain transactions in a specified manner and charge fraudulent merchants a certain amount of label cost. This behavior attempts to hack ranking mechanisms.
Spam Product Reviews	Spammers evaluate products with serious bias. Usually they put unreal reviews aiming to mislead consumers'decisions.
Two-day Task	First add target item into online shopping cart, and then create a spam transaction on the next day
...	...

Figure 1: Some services that a malicious service platform provides. The dishonest merchants can freely select different combinations of these services, e.g. **Two-day Task**.

and buyers prefer those items with a good reputation. Regular ways to boost these measurements include providing high-quality items, good services, and advertising, which usually take much effort. It motivates some malicious merchants to promote their items by spam actions. Such dishonest behaviors lead to serious consequences: 1) for the customers, they are misled to purchase items seemingly only good on numbers over quality; 2) for regular sellers, their incomes are directly affected, which causes unfair competition; 3) for the e-commerce platforms, it increases the difficulty of recognizing good sellers and decreases the advertising revenues. As demonstrated in [Tian *et al.*, 2015], the malicious promotion has caused hundreds of millions of dollars loss worldwide.

To deal with this problem, e-commerce platforms usually treat the fraud detection task as a binary classification problem, i.e., detecting fraud ones from an extremely huge number of users. For example, some algorithms [Li *et al.*, 2014; Tian *et al.*, 2015; Vlasselaer *et al.*, 2017] propagate the label information on the user adjacency graph to discover the abnormalities. However, in reality, a fraud user may perform

*Zhao Li and Weiwei Liu are corresponding authors.
[1]https://investors.ebayinc.com/overview/default.aspx
[2]https://www.alibabagroup.com/cn/ir/earnings_2018.php

multiple species of fraud operations simultaneously, such as spam transactions, clicks, reviews and so on. As illustrated in Figure 1, the third-party fraud platforms will provide some combinational fraudulent strategies. Hence, a simple binary classification model is less powerful to deal with such rich labeling information.

To bridge this gap, we propose to assign each user multiple fraud labels. Moreover, since there are hundreds of millions of users, fully-annotation is impossible and only a few labeled data is available. Hence, it is formalized to a semi-supervised multi-label learning (SSML) problem. A straightforward solution is binary relevance [Zhang and Zhou, 2014], that decomposes the original problem to a series of semi-supervised binary classification tasks. Despite its simplicity and computational efficiency, it ignores the correlations among the labels and the predictive performance is limited. Utilizing off-the-shelf SSML methods [Kong *et al.*, 2013; Wang and Tsotsos, 2016; Tan *et al.*, 2017] seems to be an appealing strategy. However, most of them are either incompatible for e-commerce tasks or incapable of handling instance-level large-scale datasets.

Consequently, this paper proposes a novel collaboration based multi-label propagation method (CMLP) for large-scale SSML fraud detection task. First of all, we present a generic version that only propagates the independent label information and recovers the original labels by collaboration. To accommodate the large-scale e-commerce data, we employ the user-item interaction matrix as a bipartite graph to detect the communities and propose a scalable heterogeneous graph propagation variant of CMLP (H-CMLP). Furthermore, we rigorously prove that the resultant algorithm actually propagates information using Adamic-Adar weight. Extensive experiments demonstrate that the proposed method not only outperforms on ordinary multi-label datasets, but is effective and scalable on large-scale e-commerce dataset.

The main contributions of this work include: 1) A generic semi-supervised multi-label learning method CMLP is proposed. It integrates the collaboration and label propagation techniques to fully utilize the label dependencies and latent data distribution. 2) To handle e-commerce data, we accelerate CMLP by propagating information on the user-item graph directly. 3) Both theoretical analysis and empirical results verify the efficiency and effectiveness of our method.

2 Related Work

2.1 Fraud Detection

Malicious promotion is one of the major threats faced by e-commerce platforms. Dishonest merchants usually employ many fraud users to create unreal transactions, clicks, reviews and so on. To find out these fraud users, various detection algorithms have been proposed. One of the most popular approaches [Li *et al.*, 2014; Tian *et al.*, 2015; Tseng *et al.*, 2015; Hu *et al.*, 2017; Vlasselaer *et al.*, 2017] focuses on graph-based fraud detection. Other detecting techniques include mixture model based methods [Bahrololum and Khaleghi, 2008], deep neural network models [Guo *et al.*, 2019] and so on. However, they cannot be directly applied to our problem for two reasons. First, many of them are either dedicated to one specific platform or designed for other application domains, e.g. click fraud [Li *et al.*, 2014], phone call fraud [Tseng *et al.*, 2015] and so on. Second, existing fraud detection methods can handle single-label classification problems only.

2.2 Semi-Supervised Multi-Label Learning

Multi-label learning (MLL) [Chen and Lin, 2012; Zhang and Zhou, 2014; Liu and Tsang, 2017; Liu *et al.*, 2019] assumes that each data example is associated with multiple labels simultaneously. Most existing MLL methods focus on a full-supervised setting. However, in real-world applications, it is expensive and difficult to obtain precisely annotated data. Therefore, semi-supervised multi-label learning (SSML) has significantly attracted the attention of researchers. To learn from both unlabeled and multi-labeled data, some algorithms [Zhan and Zhang, 2017] involve co-training technique in SSML, but they work well only when the conditional independence assumption holds. Another practical solution relies on utilizing the graph structure of data. For instance, manifold regularization based approaches [Jing *et al.*, 2015; Tan *et al.*, 2017] explore the topological structure of data; label propagation based methods [Kang *et al.*, 2006; Kong *et al.*, 2013] aggregate the neighbor information iteratively. Although these graph-based methods are promising on ordinary multi-label data, they are time-consuming and impractical in e-commerce fraud detection task. Specifically, manifold regularization algorithms usually require complex optimization techniques, and thus are demanding in large-scale datasets. Moreover, all of them are designed to run on the user-user adjacency graph, which is usually dense and large.

Note that that there are also some works study large-scale MLL problems [Bhatia *et al.*, 2015]. However, they focus on the eXtreme Multi-Label learning (XML) setting, i.e. there are extremely many labels. We concentrate on the instance-level large-scale problem, because a malicious service platform usually provides only dozens of kinds of services.

3 Proposed Method

In this section, we first introduce a generic solution to SSML, i.e. collaboration based multi-label propagation algorithm (CMLP). Then, to make it feasible in large-scale e-commerce data, we accelerate it by propagating label information on a heterogeneous graph directly.

3.1 Collaboration Based Multi-Label Propagation

We denote the instance matrix by $\boldsymbol{X} \in R^{n\times p}$. The target matrix of labeled data is denoted by $\boldsymbol{Y} \in \{-1,+1\}^{l\times q}$ ($q \ll n$). Given an undirected graph $\mathcal{G} = \langle E, V, \boldsymbol{W}\rangle$, vanilla label propagation method iteratively updates the labels of one certain node by aggregating its neighbor's label information. Here E, V are edge, vertex sets, respectively. $\boldsymbol{W} = [w_{ij}]_{n\times n}$ is a non-negative weight matrix. Let $\boldsymbol{P} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ be the propagation matrix by normalizing the columns of $\boldsymbol{W}$, where $\boldsymbol{D} = \mathrm{diag}[d_1, d_2, ..., d_n]$ is a diagonal matrix with $d_i = \sum_{j=1}^{n} w_{ij}$. Denote the model output by $\boldsymbol{F} = \left[\begin{smallmatrix}\boldsymbol{F}^l \\ \boldsymbol{F}^u\end{smallmatrix}\right] \in R^{n\times q}$, where $\boldsymbol{F}^l \in R^{l\times q}$. According to

[Zhou *et al.*, 2003], on t-th round, the updating procedure of vanilla label propagation algorithm is as follows,

$$\boldsymbol{F}_{t+1} = ((1-\beta-\beta\mu)\boldsymbol{I} + \beta\boldsymbol{P})\boldsymbol{F}_t + \beta\mu\tilde{\boldsymbol{Y}} \tag{1}$$

where $\tilde{\boldsymbol{Y}} = \left[\begin{smallmatrix}\boldsymbol{Y}\\ \boldsymbol{0}\end{smallmatrix}\right]$ and $\boldsymbol{F}_0 = \tilde{\boldsymbol{Y}}$. Here β is the learning rate and μ is a regularization parameter.

In multi-label setting, such a simple strategy fails to utilize the label correlations and achieves degenerated performance. To cope with this limitation, CMLP makes a collaborative assumption [Feng *et al.*, 2019] that *the prediction for an individual label consists of two parts: its own prediction and the prediction of others*. We involve a correlation matrix $\boldsymbol{R} = [r_{ij}]_{q\times q}$ where r_{ij} reflects the contribution of i-th label to j-th label and $r_{ii} = 0$, i.e. there is no collaboration from the label itself. Suppose the desired method gives an ordinary prediction $f(\boldsymbol{X})$, the final prediction is made by,

$$\hat{\boldsymbol{Y}} = (1-\alpha)f(\boldsymbol{X}) + \alpha f(\boldsymbol{X})\boldsymbol{R} \tag{2}$$

That is, the final output absorbs the prediction of other labels in a collaborative fashion. By regarding the ground-truth of labeled data as final prediction, CMLP estimates $\boldsymbol{R}$ by,

$$\begin{aligned} \min_{\boldsymbol{r}_j} & ||((1-\alpha)\boldsymbol{y}_j + \alpha\boldsymbol{Y}\boldsymbol{r}_j) - \boldsymbol{y}_j||^2 + \gamma||\boldsymbol{r}_j||^2 \\ & s.t. \quad r_{jj} = 0 \end{aligned} \tag{3}$$

where $\boldsymbol{y}_j$, $\boldsymbol{r}_j$ denote the j-th columns of $\boldsymbol{Y}$, $\boldsymbol{R}$. Here α is the collaboration degree, a tradeoff parameter between the original prediction and the collaborative prediction. γ is the regularization parameter. By simple reformulation, it can be transformed to a standard ridge regression problem and therefore can be efficiently solved.

Recall the main drawbacks of existing multi-label propagation algorithms that they are either time-consuming or incapable of exploiting label dependencies. To tackle these problems, we propose an effective method that propagates only the independent part of labels instead of original label information. By assigning an intermediate variable $\boldsymbol{Z} \in R^{l\times q}$ to the labeled instances, CMLP aims to optimize the following objective,

$$\begin{aligned} \mathcal{L}(\boldsymbol{F},\boldsymbol{Z}) =& \frac{1}{2}(\sum_{i,j=1}^{n} w_{ij}||\frac{1}{\sqrt{d_{ii}}}\boldsymbol{f}^i - \frac{1}{\sqrt{d_{jj}}}\boldsymbol{f}^j||^2) \\ &+ \frac{1}{2}\mu||\boldsymbol{F}^l - \boldsymbol{Z}||_{\mathcal{F}}^2 + \frac{1}{2}\lambda||\boldsymbol{Z}\boldsymbol{Q} - \boldsymbol{Y}||_{\mathcal{F}}^2 \end{aligned} \tag{4}$$

where $\boldsymbol{Q} = (1-\alpha)\boldsymbol{I} + \alpha\boldsymbol{R}$, μ and λ are trade-off parameters. $\boldsymbol{f}^i$ is the i-th row vector of $\boldsymbol{F}$. $||\cdot||_{\mathcal{F}}$ is the Frobenius norm. To be more specific, CMLP simultaneously solves two subproblems: 1) propagating independent label information $\boldsymbol{Z}$ on the graph; 2) fitting the final predictions using correlation matrix $\boldsymbol{R}$. Such an objective not only encourages the mutuality of these two subproblems, but enables our method to be highly scalable.

In this work, we initialize the variables by $\boldsymbol{F}_0 = \left[\begin{smallmatrix}\boldsymbol{Y}\\ \boldsymbol{0}\end{smallmatrix}\right]$ and $\boldsymbol{Z}_0 = \boldsymbol{Y}$. Since Eq. (4) is a biconvex problem, it can be solved in an alternating way.

Updating F With Fixed Z

When $\boldsymbol{Z}$ is fixed, the remaining items constitute an objective of soft label propagation algorithm with independent label $\boldsymbol{Z}$. We first derive the gradient by,

$$\frac{\partial\mathcal{L}(\boldsymbol{F};\boldsymbol{Z}_t)}{\partial\boldsymbol{F}} = \boldsymbol{F} - \boldsymbol{P}\boldsymbol{F} + \mu(\boldsymbol{F} - \tilde{\boldsymbol{Z}}_t) \tag{5}$$

where $\tilde{\boldsymbol{Z}}_t = \left[\begin{smallmatrix}\boldsymbol{Z}_t\\ \boldsymbol{0}\end{smallmatrix}\right]$. Then we perform a gradient descent step with learning rate β for $\boldsymbol{F}$ such that,

$$\begin{aligned} \boldsymbol{F}_{t+1} &= \boldsymbol{F}_t - \beta(\boldsymbol{F}_t - \boldsymbol{P}\boldsymbol{F}_t + \mu(\boldsymbol{F}_t - \tilde{\boldsymbol{Z}}_t)) \\ &= ((1-\beta-\beta\mu)\boldsymbol{I} + \beta\boldsymbol{P})\boldsymbol{F}_t + \beta\mu\tilde{\boldsymbol{Z}}_t \end{aligned} \tag{6}$$

Updating Z With Fixed F

With $\boldsymbol{F}$ being fixed, the optimization problem reduces to,

$$\min_{\boldsymbol{Z}} \frac{1}{2}\mu||\boldsymbol{F}^l - \boldsymbol{Z}||_{\mathcal{F}}^2 + \frac{1}{2}\lambda||\boldsymbol{Z}\boldsymbol{Q} - \boldsymbol{Y}||_{\mathcal{F}}^2 \tag{7}$$

A closed-form solution can be obtained by setting the gradient to zero,

$$\boldsymbol{Z}_{t+1} = (\boldsymbol{F}_{t+1}^l + \eta\boldsymbol{Y}\boldsymbol{Q}^\top)(\boldsymbol{I} + \eta\boldsymbol{Q}\boldsymbol{Q}^\top)^{-1}, \quad \eta = \frac{\lambda}{\mu} \tag{8}$$

When the iterations end, we have to transform the output to final prediction $\hat{\boldsymbol{Y}}^u = \Psi(\boldsymbol{F}^u\boldsymbol{Q})$. Here Ψ is a post-process operator, consisting of normalization and binarizing.

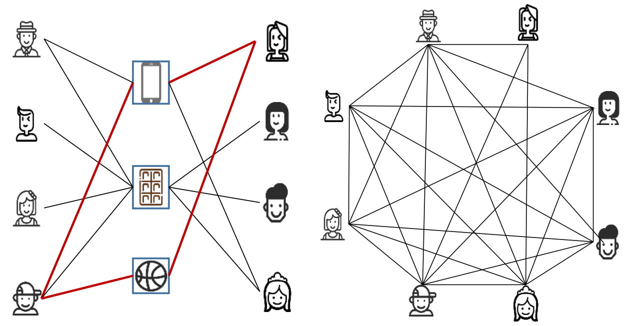
Time Complexity

The main computation of our method lies in the iteratively propagating labels on the graph. In Eq. (6), the matrix product of $\boldsymbol{P}$ and $\boldsymbol{F}_t$ dominates and takes $\mathcal{O}(qn^2)$, which is the same as vanilla LP method. As for Eq. (8), $\eta\boldsymbol{Y}\boldsymbol{Q}^\top$ and $(I + \eta\boldsymbol{Q}\boldsymbol{Q}^\top)^{-1}$ can be pre-computed with only few computation. Thus, updating $\boldsymbol{Z}$ requires $\mathcal{O}(lq^2)$ and is much cheaper than updating $\boldsymbol{F}$. In this paper, we do not consider extreme classification setting, and thus $q \ll n$. We conclude that our algorithm is as fast as the original LP method with time complexity $\mathcal{O}(qn^2)$ in each iteration.

Compared to other state-of-the-art multi-label propagation algorithms, our method is more scalable. For instance, CLP [Kang *et al.*, 2006] solves a constrained optimization problem in each iteration, which is infeasible in large-scale semi-supervised MLL problems. Our updating procedure only requires simple matrix operations and can be highly parallelized to achieve good scalability.

3.2 Heterogeneous Graph Propagation

Yet, we proposed a generic correlation-aware propagation algorithm for ordinary SSML tasks. In effect, the graph can be collected in a variety of ways, e.g. k-nearest neighbor adjacency graph [Wang and Tsotsos, 2016], webpage links [Wu *et al.*, 2014; Wu *et al.*, 2015] and so on. In the e-commerce fraud detection scenario, our goal is to determine whether a user has some fraud behaviors. In general, malicious merchants will hire many fraud users to buy the same items. Therefore, a natural choice is to synthesize a user-user mapping (U-U graph) from the user-item bipartite graph (U-I graph), i.e. two users are connected to each other if they are interested in the



(a) User-Item Graph (b) User-User Graph

Figure 2: An example of User-Item graph and the corresponding User-User graph. The U-I graph has 12 edges and the U-U graph has 24 edges. The popular item *chocolate* is linked to seven users, and thus the U-U graph contains a complete subgraph with seven nodes. The red lines in U-I graph represent fraud behaviours.

same item. Though this strategy avoids building a graph explicitly, which is usually time-consuming, we observe that it gives a really dense graph in practice (Figure 2). In many e-commerce platforms, a popular item can be connected to millions of users. Thus, U-U graph may contain many complete subgraphs and the number of edges grows dramatically. As we discussed above, the main computation cost of CMLP lies in the propagation procedure, i.e. aggregating label information from adjacent nodes. Thus, it will be demanding in large-scale datasets.

To alleviate this problem, we present a heterogeneous graph propagation (H-CMLP) approach, which detects communities on U-I graph directly. For simplicity, we concentrate on the propagation part in Eq. (6), i.e. $\boldsymbol{A}_{t+1} = \boldsymbol{P}\boldsymbol{F}_t$. The calculation of $\boldsymbol{A}_{t+1}$ can be divided into two steps. In the first step, for each item, we aggregate information from its neighbor users. Denote the temporary label vector for an item e_k by $\boldsymbol{s}^k$. On t-th round, it is updated by,

$$\boldsymbol{s}_t^k = \sum_{j\in\mathcal{N}(e_k)} \boldsymbol{f}_t^j \tag{9}$$

where $\mathcal{N}(e_k)$ is the index set of users that connect to e_k, and $\boldsymbol{f}_t^j$ is the j-th row vector in $\boldsymbol{F}_t$. In the second step, H-CMLP propagates the labels back to the users. Formally, for i-th user u_i, its propagation part $\boldsymbol{a}_{t+1}^i$ is calculated by,

$$\boldsymbol{a}_{t+1}^i = \frac{\sum_{k\in\mathcal{N}(u_i)} \frac{\boldsymbol{s}_t^k - \boldsymbol{f}_t^i}{\log(|\mathcal{N}(e_k)|)}}{\sum_{k\in\mathcal{N}(u_i)} \frac{|\mathcal{N}(e_k)|-1}{\log(|\mathcal{N}(e_k)|)}} \tag{10}$$

Here $|\cdot|$ denotes the capacity of a set and $\log(\cdot)$ is natural logarithm.

Lemma 1. *The updating rule defined by Eq. (9) and Eq. (10) is equivalent to propagating information on U-U graph using Adamic-Adar weight.*

Proof. Consider a U-U graph based LP method that adopts Adamic-Adar [Adamic and Adar, 2003; Benson and Kleinberg, 2019], a common measure in graph mining problems, to represent the weight between two users u_i, u_j,

$$w_{ij} = \sum_{k\in\mathcal{N}(u_i)\cap\mathcal{N}(u_j)} \frac{1}{\log(|\mathcal{N}(e_k)|)} \tag{11}$$

In words, we expect that more sharing items lead to larger weight, while more popular item results in smaller contribution. The propagation part is produced as follows,

$$\boldsymbol{a}_{t+1}^i = \frac{\sum_{j\in\Phi(i)} w_{ij}\boldsymbol{f}_t^j}{\sum_{j\in\Phi(i)} w_{ij}}, \quad \Phi(i) = \{j|\mathcal{N}(u_i)\cap\mathcal{N}(u_j) \neq \emptyset\} \tag{12}$$

Define $H_{g,h}(i,j,k) = \sum_{j\in\Phi(i)}\sum_{k\in\mathcal{N}(u_i)\cap\mathcal{N}(u_j)} h(k)g(j)$ with given function h and g. Namely, once there is an item e_k shared by u_i and u_j, it contributes $h(k)g(j)$ to $H_{g,h}(i,j,k)$. Actually, we can reformulate it to $H_{g,h}(i,j,k) = \sum_{k\in\mathcal{N}(u_i)} h(k)\sum_{j\in\mathcal{N}(e_k),j\neq i} g(j)$. Note that in Eq. (12), the numerator is a typical H function with $h(k) = (\log(|\mathcal{N}(e_k)|)^{-1}$ and $g(j) = \boldsymbol{f}_t^j$. The denominator is also an H function with a same h and $g(j) = 1$. We conclude that Eq. (12) can be rewritten to,

$$\boldsymbol{a}_{t+1}^i = \frac{\sum_{k\in\mathcal{N}(u_i)} \frac{\sum_{j\in\mathcal{N}(e_k),j\neq i}\boldsymbol{f}_t^j}{\log(|\mathcal{N}(e_k)|)}}{\sum_{k\in\mathcal{N}(u_i)} \frac{(|\mathcal{N}(e_k)|-1)}{\log(|\mathcal{N}(e_k)|)}} \tag{13}$$

We observe that $\sum_{j\in\mathcal{N}(e_k),j\neq i}\boldsymbol{f}_t^j = \boldsymbol{s}_t^k - \boldsymbol{f}_t^i$. Hence, Eq. (12) is equivalent to Eq. (10) and the lemma is proved. □

Time Complexity

According to Eq. (12), original CMLP takes $\mathcal{O}(2q|E|)$ time, i.e. summing up the degree number of all the nodes, to update the labels of each node. When the U-U graph is dense, e.g. many users are interested in some popular items, $|E|$ can be extremely large. However, in heterogeneous graph setting, each propagation step travels all the edges on U-I graph. Thus, the time complexity is $\mathcal{O}(2q|\hat{E}|)$, where $\hat{E}$ is the edge set of U-I graph. As illustrated in Figure 2, $|\hat{E}|$ is much smaller because user-item interactions are usually sparse.

4 Experiments

4.1 Ordinary Multi-Label Data

To show the effectiveness of our collaboration technique, we test CMLP on some ordinary multi-label datasets.

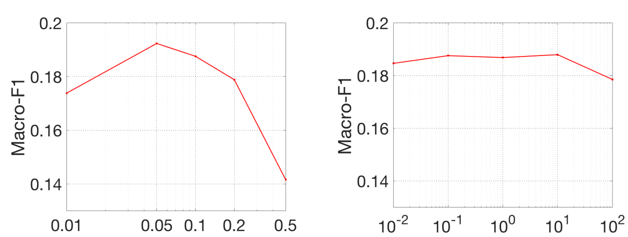
Datasets

We choose four real-world multi-label datasets from different task domains: 1) Medical [Pestian *et al.*, 2007]: a text dataset contains clinical free texts, each of which is with 45 ICD-9-CM labels, from CCHMC Department of Radiology. 2) Image [Wang *et al.*, 2019]: a collection of $2,000$ images that are annotated by 5 labels. 3) Slashdot [Read *et al.*, 2009]: a web text dataset collects $3,782$ technology-related news from 22 categories. 4) Eurlex-sm [Loza Menc'ia and Fürnkranz, 2008]: a large text dataset contains $19,348$ legal documents about European Union law, having 201 subject matters tags.

All the datasets are randomly partitioned to 5% labeled data and 95% unlabeled data. In this paper, we focus on

Dataset	Ranking Loss↓					
	CMLP	Vanilla-LP	SMILE	TRAM	CPLST	DeepFraud
Image	**.2569**±.0123	.2741±.0088	.4391±.0230	.2823±.0220	.3018±.0133	.5917±.0434
Medical	**.1158**±.0205	.1814±.0280	.2064±.0194	.1974±.0120	.1747±.0196	.6836±.0302
Eurlex-sm	**.0454**±.0013	.0544±.0033	.1002±.0045	.1528±.0033	.1791±.0045	.5472±.0064
Slashdot	**.1691**±.0034	.1914±.0038	.2558±.0045	.1860±.0057	.2088±.0082	.8327±.0130
Dataset	Example-F1↑					
	CMLP	Vanilla-LP	SMILE	TRAM	CPLST	DeepFraud
Image	**.5240**±.0154	.5082±.0078	.3966±.0187	.4430±.0190	.4320±.0211	.4187±.0434
Medical	**.4946**±.0129	.4617±.0180	.2860±.0353	.4119±.0270	.4032±.0370	.3204±.0274
Eurlex-sm	**.5946**±.0056	.5864±.0063	.3599±.0091	.2688±.0068	.2448±.0067	.4700±.0065
Slashdot	**.3537**±.0160	.3355±.0218	.2467±.0089	.3423±.0202	.2414±.0185	.1688±.0132
Dataset	Macro-F1↑					
	CMLP	Vanilla-LP	SMILE	TRAM	CPLST	DeepFraud
Image	**.5199**±.0135	.5024±.0086	.3291±.0288	.3996±.0275	.4638±.0116	.4618±.0592
Medical	**.1444**±.0072	.1392±.0179	.0416±.0092	.0649±.0299	.0838±.0085	.0923±.0136
Eurlex-sm	.2866±.0090	.2705±.0072	.1226±.0092	**.3517**±.0027	.0249±.0009	.1951±.0076
Slashdot	**.1875**±.0152	.1719±.0053	.1319±.0045	.1585±.0151	.0926±.0145	.0963±.0056
Dataset	Micro-F1↑					
	CMLP	Vanilla-LP	SMILE	TRAM	CPLST	DeepFraud
Image	**.5181**±.0188	.5016±.0082	.4075±.0202	.4367±.0178	.4651±.0216	.4760±.0412
Medical	.4854±.0141	.4565±.0224	.2822±.0352	.4115±.0172	**.5250**±.0156	.4287±.0229
Eurlex-sm	**.5670**±.0128	.5622±.0090	.3553±.0125	.3808±.0120	.3032±.0100	.5595±.0044
Slashdot	.3133±.0083	.3042±.0177	.2227±.0056	**.3342**±.0201	.3177±.0142	.2462±.0163

Table 1: Transductive performance comparison on ordinary multi-label datasets. The best ones are in bold.



(a) α (b) λ

Figure 3: Performance of CMLP changes as parameters α and λ change on Slashdot dataset.

the transductive setting. We conduct these experiments for 5 times and the mean metric values with standard deviations are provided. In this work, we use four popular multi-label evaluation metrics [Zhang and Zhou, 2014], i.e. Ranking Loss, Example-F1, Macro-F1, and Micro-F1.

Benchmarks

The baselines include three state-of-the-art SSML methods and two supervised MLL algorithms: 1) **Vanilla-LP** [Zhou *et al.*, 2003]: Vanilla label propagation algorithm utilizes the smooth property between labeled and unlabeled data to design a classifying function. It can also be regarded as a binary relevance model with LP model as the base classifier. 2) **SMILE** [Tan *et al.*, 2017]: SMILE explores second-order label correlations by estimating a label correlation matrix from labeled data. Then it trains a linear multi-label classifier with manifold regularizer. 3) **TRAM** [Kong *et al.*, 2013]: TRAM is a popular SSML algorithm that uses a hard label strategy to find out communities on the graph. 4) **CPLST** [Chen and Lin, 2012]: CPLST is a supervised MLL method that combines the concepts of principal component analysis and canonical correlation analysis. 5) **DeepFraud** [Guo *et al.*, 2019]: DeepFraud is a deep neural network based fraud detection algorithm, which can deal with single-label problems only. Hence, we decompose the SSML problem into a set of single-label classification tasks to run this method.

Inspired by [Wang *et al.*, 2019], we build a k-NN adjacency graph and the weight matrix is learned by reconstructing each instance using its neighbors. When building graphs, k is set as 20. For our methods, γ is selected from $\{0.1, 1, 10, 100\}$. α is chosen from $\{0.01, 0.05, 0.1, 0.2, 0.5\}$. β, λ and μ are empirically fixed to 0.1. For Vanilla-LP, β and μ are the same as CMLP. The parameter α and s in SMILE are both set as 0.5. For DeepFraud, we apply a three layer neural network with ReLU activation. The hidden size is set as 128. The learning rate and regularization parameter are set as 0.001 and 0.5. Other parameters of the baselines are set to their recommended values in their papers.

Results

Table 1 lists the transductive results on all the datasets. Figure 3 reports the parameter sensitivity of α and λ. From the empirical results, we conclude that: 1) CMLP generally achieves the best performance. For instance, on Image dataset, in terms of Example-F1, Macro-F1 and Micro-F1, CMLP im-

Method	Ranking Loss↓	Example-F1↑	Macro-F1↑	Micro-F1↑
H-CMLP	**.0419**±.0001	**.3828**±.0026	**.3220**±.0015	**.5243**±.0071
HV-LP	.0424±.0001	.3616±.0064	.3078±.0102	.5131±.0064
H-TRAM	.0515±.0013	.2801±.0039	.1884±.0021	.4610±.0077
DeepFraud	.0445±.0009	.3769±.0046	.1713±.0081	.5071±.0060

Table 2: Transductive performance comparison of three graph-based algorithms on Taobao-FUD dataset. The best ones are in bold.

Property	♯User	♯Item	♯U-I Edges	♯U-U Edges	♯Label
Value†	72M	80 M	877 M	> 100 B	4

† M, B represent million, billion.

Table 3: The characteristics of Taobao-FUD dataset.

Method \ Capacity	1K	5K	10K
H-CMLP	50.8	52.5	53.8
Vanilla-LP	234.2	501.0	2712.8

Table 4: Average time on an iteration (in seconds) of H-CMLP and Vanilla-LP on different subsets.

proves the best results of the baselines by 3.1%, 3.5% and 3.3%. 2) Vanilla-LP and TRAM underperform our method, because they neglect the correlations among labels. 3) The effectiveness of SMILE is limited since it exploits the second-order correlations only. 4) CPLST and DeepFraud are inferior to other methods because they ignore the distribution information of unlabeled data. 5) α has a significant influence on the final performance. Empirically, relatively small α leads to good performance. Moreover, CMLP is relatively stable with varying values of λ.

4.2 E-commerce Fraud Data

Experimental Setting

The other experiment is conducted on a large-scale e-commerce dataset Taobao-FUD. Taobao-FUD collects an interaction matrix from more than 72 million users and 80 million items. Each user is equipped with four binary labels {*Transaction, Cart, Click, Review*}, indicating whether the users have corresponding fraud behaviors on some items. If we construct a U-U graph from Taobao-FUD, the number of edges can be larger than 100 billion. Obviously, mining on such an extremely large graph is impossible. To verify the time efficiency of H-CMLP, we sample some user subsets with capacities of $\{1\,\mathrm{K}, 5\,\mathrm{K}, 10\,\mathrm{K}\}$ ($1\,\mathrm{K} = 1000$) to construct three subgraphs and report the running time of H-CMLP and Vanilla-LP. Then, to compare the predictive performance, we implemented heterogeneous versions of Vanilla-LP (HV-LP) and TRAM (H-TRAM). We randomly select 5% examples as labeled data and the rest are used to evaluate the transductive performance. The statistical information of Taobao-FUD can be found in Table 3.

Note that SMILE and CPLST are incapable of dealing with instance-level large-scale data. Thus, we choose Vanilla-LP, TRAM, and DeepFraud as the benchmarks. Nevertheless, there are only graph-structured data in Taobao-FUD dataset, and DeepFraud cannot be applied directly. Therefore, we firstly extracts user embeddings using GraphSAGE on the U-I Graph and then feed them into DeepFraud model. The parameter setup and evaluation metrics are the same as last subsection. The computations are performed on MaxCompute platform, a fast, distributed and fully hosted GB/TB/PB level data warehouse solution. We use three computation instances for time comparison and 3000 instances for performance comparison.

Results

Table 2 reports the transductive performance of all the methods on Taobao-FUD. From the results, we observe that H-CMLP is the most successful method. In terms of Micro-F1, Macro-F1, and Example-F1, H-CMLP improves the best results of the baselines by 1.6%, 4.6%, 2.2% respectively. The results demonstrate that the proposed method can effectively address the fraud detection task because our collaboration technique can sufficiently exploit the label correlations.

Table 4 shows the average running time on each iteration of H-CMLP and Vanilla-LP on three subgraphs. For H-CMLP, the running time has no obvious fluctuation. The reason is the size of U-I graph is really small, and the main computation is the scheduling procedure of MaxCompute. For Vanilla-LP, the U-U graph is much larger, and thus its running time grows quickly. By these observations, we conclude that our method can efficiently address the multi-label fraud detection task.

5 Conclusion

The huge success of online shopping motivates dishonest sellers to illegally promote their reputations. In general, these malicious merchants will hire many fraud users to create different kinds of spam actions concurrently. To tackle this problem, we treat the fraud user detection task as semi-supervised multi-label learning (SSML) problem and propose a scalable graph-based SSML algorithm. We first introduce a generic version that decomposes the final prediction to an independent part and a collaborative part. Then, to accommodate large-scale e-commerce data, we accelerate it by detecting communities on the user-item interaction graph directly. Extensive experiments demonstrate that our method can effectively and efficiently handle both ordinary SSML problems and fraud detection task.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61976161) and Key R&D Program of Zhejiang Province (Grant No. 2020C01024).

References

[Adamic and Adar, 2003] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[Bahrololum and Khaleghi, 2008] M. Bahrololum and M. Khaleghi. Anomaly intrusion detection system using hierarchical gaussian mixture model. In *International Journal of Computer Science and Network Security*, 2008.

[Benson and Kleinberg, 2019] Austin R. Benson and Jon M. Kleinberg. Link prediction in networks with core-fringe data. In *WWW*, pages 94–104, 2019.

[Bhatia *et al.*, 2015] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In *NeurIPS*, pages 730–738, 2015.

[Chen and Lin, 2012] Yao-Nan Chen and Hsuan-Tien Lin. Feature-aware label space dimension reduction for multi-label classification. In *NeurIPS*, pages 1538–1546, 2012.

[Feng *et al.*, 2019] Lei Feng, Bo An, and Shuo He. Collaboration based multi-label learning. In *AAAI*, pages 3550–3557, 2019.

[Guo *et al.*, 2019] Qingyu Guo, Zhao Li, Bo An, Pengrui Hui, Jiaming Huang, Long Zhang, and Mengchen Zhao. Securing the deep fraud detector in large-scale e-commerce platform via adversarial machine learning approach. In *WWW*, pages 616–626, 2019.

[Hu *et al.*, 2017] Jinlong Hu, Junjie Liang, and Shoubin Dong. ibgp: A bipartite graph propagation approach for mobile advertising fraud detection. *Mobile Information Systems*, 2017:6412521:1–6412521:12, 2017.

[Jing *et al.*, 2015] Liping Jing, Liu Yang, Jian Yu, and Michael K. Ng. Semi-supervised low-rank mapping learning for multi-label classification. In *CVPR*, pages 1483–1491, 2015.

[Kang *et al.*, 2006] Feng Kang, Rong Jin, and Rahul Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, pages 1719–1726, 2006.

[Kong *et al.*, 2013] Xiangnan Kong, Michael K. Ng, and Zhi-Hua Zhou. Transductive multilabel learning via label set propagation. *IEEE Trans. Knowl. Data Eng.*, 25(3):704–719, 2013.

[Li *et al.*, 2014] Xin Li, Min Zhang, Yiqun Liu, Shaoping Ma, Yijiang Jin, and Liyun Ru. Search engine click spam detection based on bipartite graph propagation. In *WSDM*, pages 93–102, 2014.

[Liu and Tsang, 2017] Weiwei Liu and Ivor W. Tsang. Making decision trees feasible in ultrahigh feature and label dimensions. *J. Mach. Learn. Res.*, 18:81:1–81:36, 2017.

[Liu *et al.*, 2019] Weiwei Liu, Donna Xu, Ivor W. Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):408–422, 2019.

[Loza Menc'ia and Fürnkranz, 2008] Eneldo Loza Menc'ia and Johannes Fürnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *ECML/PKDD*, pages 50–65, 2008.

[Pestian *et al.*, 2007] John P. Pestian, Chris Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In *BioNLP@ACL*, pages 97–104, 2007.

[Read *et al.*, 2009] Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *ECML/PKDD*, pages 254–269, 2009.

[Tan *et al.*, 2017] Qiaoyu Tan, Yanming Yu, Guoxian Yu, and Jun Wang. Semi-supervised multi-label classification using incomplete label information. *Neurocomputing*, 260:192–202, 2017.

[Tian *et al.*, 2015] Tian Tian, Jun Zhu, Fen Xia, Xin Zhuang, and Tong Zhang. Crowd fraud detection in internet advertising. In *WWW*, pages 1100–1110, 2015.

[Tseng *et al.*, 2015] Vincent S. Tseng, Jia-Ching Ying, Che-Wei Huang, Yimin Kao, and Kuan-Ta Chen. Fraudetector: A graph-mining-based framework for fraudulent phone call detection. In *SIGKDD*, pages 2157–2166, 2015.

[Vlasselaer *et al.*, 2017] Véronique Van Vlasselaer, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. Gotcha! network-based fraud detection for social security fraud. *Management Science*, 63(9):3090–3110, 2017.

[Wang and Tsotsos, 2016] Bo Wang and John K. Tsotsos. Dynamic label propagation for semi-supervised multi-class multi-label classification. *Pattern Recognition*, 52:75–84, 2016.

[Wang *et al.*, 2019] Haobo Wang, Weiwei Liu, Yang Zhao, Chen Zhang, Tianlei Hu, and Gang Chen. Discriminative and correlative partial multi-label learning. In *IJCAI*, pages 3691–3697, 2019.

[Wu *et al.*, 2014] Jia Wu, Xingquan Zhu, Chengqi Zhang, and Philip S. Yu. Bag constrained structure pattern mining for multi-graph classification. *IEEE Trans. Knowl. Data Eng.*, 26(10):2382–2396, 2014.

[Wu *et al.*, 2015] Jia Wu, Shirui Pan, Xingquan Zhu, and Zhihua Cai. Boosting for multi-graph classification. *IEEE Trans. Cybernetics*, 45(3):430–443, 2015.

[Zhan and Zhang, 2017] Wang Zhan and Min-Ling Zhang. Inductive semi-supervised multi-label learning with co-training. In *SIGKDD*, pages 1305–1314, 2017.

[Zhang and Zhou, 2014] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014.

[Zhou *et al.*, 2003] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NeurIPS*, pages 321–328, 2003.