NEAT: A Label Noise-resistant Complementary Item Recommender System with Trustworthy Evaluation

Luyi Ma, Jianpeng Xu, Jason H.D. Cho, Evren Korpeoglu, Sushant Kumar, Kannan Achan
Walmart Global Tech, Sunnyvale, CA
{luyi.ma, jianpeng.xu, jason.cho, EKorpeoglu, sushant.kumar, kannan.achan}@walmart.com

***Abstract*—The *complementary item recommender system* (CIRS) recommends the complementary items for a given query item. Existing CIRS models consider the item co-purchase signal as a proxy of the complementary relationship, due to the lack of human-curated labels from the huge transaction records. These methods represent items in a complementary embedding space and model the complementary relationship as a point estimation of the similarity between items vectors. However, co-purchased items are not necessarily complementary to each other. For example, customers may frequently purchase *bananas* and *bottle water* within the same transaction, but these two items are not complementary. Hence, using co-purchase signals directly as labels will aggravate the model performance. On the other hand, model evaluation will not be trustworthy if the labels for evaluation are not reflecting the true complementary relatedness. To address the above challenges from noisy labeling of the co-purchase data, we model the co-purchases of two items as a Gaussian distribution, where the mean denotes the co-purchases from the complementary relatedness, and covariance denotes the co-purchases from the noise. To do so, we represent each item as a Gaussian embedding and parameterize the Gaussian distribution of co-purchases by the means and covariances from item Gaussian embedding. To reduce the impact of the noisy labels during evaluation, we propose an independence test-based method to generate a trustworthy label set with certain confidence. Our extensive experiments on both the publicly available dataset and the large-scale real-world dataset justify the effectiveness of our proposed model in complementary item recommendations compared with the state-of-the-art models.**

***Index Terms*—Recommender System, Gaussian embedding, Complementary Item Recommendation**

I. Introduction

Item recommendation tasks in e-commerce industry are essential for improving user experiences by recommending related items to a query item. Different types of recommender systems have been proposed to address use cases under various aspects of the *relatedness*, such as substitutional items (SI) recommendation and complementary items (CI) recommendation [1] [2] [3]. In economics, a complementary item is a type of items whose appeal increases with the popularity of its complement. Therefore, complementary items usually have higher chances to be purchased together to complete the same shopping goal. For example, `Shampoo` and `Conditioner` are complementary to each other in order to fulfill the needs of *shower supplies*; similarly, `TV` and `TV Mount` are also complementary items for *TV entertainment* purposes. While SI recommendations have been extensively studied in the past [4] [5], complementary item recommender systems (CIRS) become increasingly important as they provide the customers with the opportunities to explore and interact with items that are complementary with what they have been interested in, and hence complete the customers' shopping journey by suggesting purchasing those items together.

Although the complementary relationship between items seems well-defined, it is impossible to gain the ground truth of the complementary relationship for all item pairs from the catalogue. To mitigate the labeling challenge, a common practice is to indicate the complementary relationship using the co-purchase signal of two items [1] [2] [6] [3]. These CIRS models usually represent each item as an embedding vector under co-purchase space, and the similarity between the items in the latent space reflects the frequency of co-purchases, and hence the complementary relatedness under their assumptions.

However, co-purchased items are not necessarily complementary to each other. For example, certain popular items can appear in many transactions and hence be co-purchased frequently with items that are not complementary. Simply removing these popular items from all recommendations will hurt the results for item pairs with real complementary relations and decrease the business metrics (e.g., Gross Merchandise Value) of the recommender systems. Recently, Hao et al. proposed to annotate the co-viewed but not co-purchased item pairs as the negative labels and consider the co-purchased but not co-viewed item pairs as positive labels for learning [7]. However, co-view data are noisy by themselves as well. Cleaning noisy labels with another noisy data source is not trustworthy in general. Identifying and cleaning co-purchased non-complementary items is not feasible due to the lack of ground truths. Hence, it is challenging to learn the real complementary relationships between items pairs and evaluate the recommendation results with the noisy labels.

To address the noisy label issue during training, we assume that the co-purchases of items are composed by two components: (a) co-purchases motivated by the true complementary relationships, and (b) co-purchases from other motivations (say, the noise). We directly model component (a) by the similarities or distances of item embeddings under the complementary space, and component (b) by the variance around (a). Hence, the co-purchase data can be assumed as a Gaussian distribution, where the mean is the co-purchases from the true complementary, and the variance is the co-purchases from the noise. To achieve this, instead of representing items as item embeddings under point estimation, we employ Gaussian embeddings [8] with a mean vector and a covariance matrix to as item representations. The Gaussian distribution of the co-purchase data can be naturally parameterized by the item Gaussian embeddings and fit into the noisy co-purchase data by optimizing the *expected likelihood* [9] between Gaussian embeddings. To this end, works such as [1] [2] [6] [3] are special cases of this assumption which assume that all co-purchases occur under complementary relationships. They represent each item as a vector in the embedding space and the co-purchases or complementary relationships are calculated by the similarities between item embeddings.

To address the noisy label issue during evaluation, we follow the definition of complementary items and develop an independence test-based method to surface the item pairs with more complementarity as positive labels for evaluation. Given a pair of co-purchased items, we treat the purchase of the individual item as a binary random variable and study the difference between observed co-purchase frequency and the expected co-purchase frequency under the independence assumption via Chi-squared independence test [10]. Based on the definition of complementary items in economics, the purchases of them should be dependent and the observed co-purchase frequency should be larger then the expected independent co-purchase frequency due to the synergy effect between complementary items. A set of co-purchase labels could be generated for evaluation by providing a predefined p-value, which controls the certainty of the label selection from the noisy observation. Although it is promising to use the selected label as the ground truth labels for training as well, the coverage of this set over the item catalogue is very limited and hence not feasible to be generalized for training purpose.

In summary, we developed a label ***N***oise-r***E***sist***A***n***T*** CIRS model named **NEAT**, which learns the complementary relationship by Gaussian embedding representation. In order to accurately evaluate the model performance, we created a trustworthy label set with controllable confidence via an independence test. Extensive experiments are conducted on the publicly available *Instacart* dataset and a real-world large-scale dataset collected from www.walmart.com. The results demonstrated the effectiveness of **NEAT** in modeling complementary relationships from co-purchase data, and the superior performance over the state-of-the-art models in CIRS.

The rest of the paper is structured in the following: the related work on CIRS is discussed in Section II. Section III describes the details of the proposed method **NEAT** and Section IV presents the trustworthy label creation for evaluation. Experiment settings and results are reported in Section V. In the end, we conclude the paper in Section VI.

II. Related Work

A. Embedding-based Complementary Item Recommendations

Embedding-based CIRS are most popular in recent work of CIRS. They treat each item as a vector in the embedding space and estimate the complementary relationship based on the distance between item vectors. The first embedding-based method for CIRS was proposed in [11], which models the co-purchase of items by the similarity between the embeddings of the co-purchased items under the effective training paradigm of Skip-gram with Negative Sample (SGNS) [12]. Wan et. al. extended [11] by incorporating user behavior into the modeling of the item-level complementary relationship with the user and item embeddings learned jointly [6]. Besides of modeling the item embeddings with co-purchase data using SGNS, co-purchase data are also represented as item graphs in [1] [2] [3] and identifying the complementary relationships between items are treated as the link prediction tasks between item nodes. They use the co-purchase records as labels for link predictions based on the distance between item embeddings. To further improve the complementary recommendations, different types of auxiliary data are incorporated into the modeling. Multi-modal data of items such as item descriptions and images are also included in [13] to learn the multi-modal representations of items. The distance between vector embeddings of two co-purchased items in each modal's embedding space is minimized for complementarity measurement. Xu et al. in [14] considers the last l purchased items as the context to learn the attention-based encoder and represents the complementary relationship via the inner product between encoded item embeddings. This work is reduced to the **Item2Vec** model in [11] without the context of the last l items in the history. Although the **P-companion** model in [7] pre-processes the co-purchase labels by removing co-view data and leverage the product-type information to improve the diversity, it still models the complementary relationship via the distance between item embeddings without addressing the noise in the data by parameters. Despite of various auxiliary information such as graph structure, multi-modal data source, shopping context and product taxonomy, all these models are trying to build item embeddings in co-purchase space, and model the co-purchase data using the similarity or distance between the co-purchased items. Hence, these models will suffer from the noisy labels for learning complementary relations.

In addition to the item-level complementary recommendation models, many in-basket recommendation models try to address the complete-the-basekt tasks. For example, **BasConv** [15] leverages the heterogeneous graph embeddings to perform the in-basket recommendations; multiple intents in the same basket are modeled in [16] to the in-basket recommendation. Although these models capture the co-purchase pattern in the same basket to complete the basket, items in the same basket might not be complementary, for instance, a basket containing both grocery shopping and the household shopping. Even the items with the same shopping intent like grocery shopping might not be complementary. The goal of the in-basket recommendations focuses on completing the basket which cover various types of recommendations such as re-purchase, popular items and user preference in addition to co-purchases and complementary items. Hence, in-basket recommendation is out of the scope of the discussion in this paper.

B. Gaussian Embedding in Recommender Systems

Gaussian embedding [8] has been applied in recommender system in recent years, e.g., Gaussian embeddings for collaborative filtering [17] and convolutional Gaussian embeddings for personalized item recommendations [18]. They mainly use the Gaussian embeddings to address the different confidences of user/item representations introduced by the lack of user/item information or contradictions between user/item behaviors (e.g., item ratings and reviews by users). However, these methods were not designed for CIRS and hence not applicable to address the unique challenges from CIRS.

III. Label Noise-resistant Complementary Item Recommender Systems

In this section, we first define the co-purchase records from transactions and then go through the details of modeling item-level complementary relationship as well as item representation for recommendations.

A. Co-purchase Records from Transactions

Let v denote an item from the item set $\mathcal{V}$ and b denote a transaction (a set of purchased items) from the transactions set $\mathcal{B}$ where $b = \{v_1, v_2, ...\}$. A tuple (v_i, v_j), $v_i \neq v_j$, from the same transaction b can be considered as a pair of co-purchased items (i.e., a co-purchase record). To further distinguish the role of co-purchased items during training, inference and evaluation, we treat the first item in an item pair (v_i, v_j) as the query item q and the second item as the recommendation of q.

B. Learning Item-level Complementary Relationship

Learning the complementary relationship with the co-purchase data as labels could suffer from the label noisy, as co-purchased items are not necessarily complementary items. Previous CIRS models simply treat the co-purchased items as the positive label of complementary relationships and fit them by the distance between item embeddings in the embedding space. Formally, given a pair of co-purchased items (q, v), they try to maximize the density of a particular normal distribution at zero: $\mathcal{N}\left(0; \boldsymbol{\mu}_q - \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_{zero}\right) \propto -\left(\boldsymbol{\mu}_q - \boldsymbol{\mu}_v\right)^T \left(\boldsymbol{\mu}_q - \boldsymbol{\mu}_v\right)$, where $\boldsymbol{\Sigma}_{zero}$ is zero and $\boldsymbol{\mu}$ is the item embeddings, to bring the item embedding $(\boldsymbol{\mu}_q, \boldsymbol{\mu}_v)$ closer in the embedding space. Because these models do not consider the noise in the co-purchase labels, the distance between item embeddings is hardly reflecting the complementary relationship, even it might be a good approximation for co-purchases.

To address the label noise issue for learning complementary relationship, as aforementioned, we model the co-purchase data as a Gaussian distribution, where the mean is the co-purchases from the true complementary, and the variance is the co-purchase from the noise. In order to do so, we consider each item $v \in \mathcal{V}$ as a Gaussian embedding $\mathcal{N}\left(x; \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right)$, where $\boldsymbol{\mu}_v \in \mathbb{R}^d$ is the mean vector and $\boldsymbol{\Sigma}_v \in \mathbb{R}^{d \times d}$ is the covariacne matrix in the d-dimensional embedding space, which models the variation in the co-purchase behavior of v. While the inner product between vectors of two items is used to model their complementary relationship from the co-purchase record in the literature [6] [11], we compute the *expected likelihood* [9] as the inner product of two Gaussian embeddings [8] to parameterize the Gaussian distribution of complementary relationships. Given an item pair (q, v), the *expected likelihood* between their Gaussian embeddings is defined in Equation 1, which is the probability density of a Gaussian distribution at zero, $\mathcal{N}\left(0; \boldsymbol{\mu}_q - \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_v\right)$.

$$\begin{aligned} E(q, r) &= \int_{x \in \mathbb{R}^d} \mathcal{N}\left(x; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right) \mathcal{N}\left(x; \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right) dx \\ &= \mathcal{N}\left(0; \boldsymbol{\mu}_q - \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_v\right) \end{aligned} \tag{1}$$

Hence, $\mathcal{N}\left(x; \boldsymbol{\mu}_q - \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_v\right)$ denotes the Gaussian distribution of the co-purchase data between (q, v), where the mean is the difference between two items mean vectors in complementary space and the covariance matrix combines the variance of each individual items. The probability density at zero, $\mathcal{N}\left(0; \boldsymbol{\mu}_q - \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_v\right)$, represents the likelihood of observing a co-purchase record of (q, v) when considering both their complementary relationship ($\boldsymbol{\mu}_q - \boldsymbol{\mu}_v$) and variations of purchase behaviors ($\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_v$).

To illustrate the benefit of representing both co-purchase data and items embeddings as Gaussian distributions, let's consider an example from our daily shopping: `milk`, `cereal` and `chips`, in a 1-dimensional embedding space in Figure 1. In Figure 1-(a), `milk` has the largest variance among three items because it is usually a must-buy for many customers and very likely to be co-purchased with other items without complementary relationships. `Cereal` has the smallest variance due to its stable co-purchase behavior with *milk*. The variance of `chips` is intermediate because it has some stable combinations such as `chips dips` while users might also buy them individually as a snack before checkout, which makes it variance relatively larger. In Figure 1-(b), we show the Gaussian distribution of their complementary relationship and highlight the their probability density at zero by the point A for (`milk`, `chips`) and the point B for (`milk`, `cereal`) when `milk` serves as the query item. Because the difference of variances, the Gaussian distribution of the co-purchase for (`milk`, `cereal`) shows less variance than that for (`milk`, `chips`). Even though the observed co-purchase records between `milk` and `chips` might be more than those between `milk` and `cereal`, our model can still capture the correct order of complementary relationships by comparing $|\mu_{milk} - \mu_{cereal}|$ and $|\mu_{milk} - \mu_{chips}|$. However, previous methods using item embeddings might result differently due to the directly fit for co-purchase frequency.

We follow the paradigm of Skip-gram with Negative Sampling (SGNS) and generate the negative sample v' which is

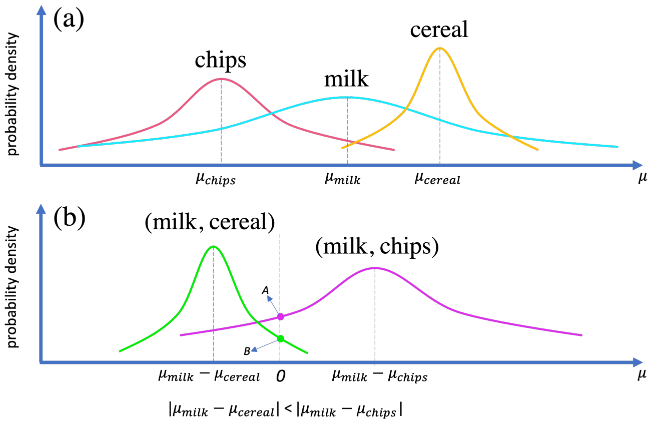

Fig. 1. (a) Examples of Gaussian embeddings (1-D) of items `milk`, `cereal` and `chips`. (b) Visualization (1-D) of $\mathcal{N}\left(0; \boldsymbol{\mu}_{milk} - \boldsymbol{\mu}_{chips}, \boldsymbol{\Sigma}_{milk} + \boldsymbol{\Sigma}_{chips}\right)$ (the point A) and $\mathcal{N}\left(0; \boldsymbol{\mu}_{milk} - \boldsymbol{\mu}_{cereal}, \boldsymbol{\Sigma}_{milk} + \boldsymbol{\Sigma}_{cereal}\right)$ (the point B) when `milk` serves as an query item. While likelihood of observing (`milk`, `chips`) could be larger than that of observing (`milk`, `cereal`) in the noisy co-purchase records ($A > B$), the correct complementary relationship between `milk` and `cereal` is captured by the distance between $\boldsymbol{\mu}_{milk}$ and $\boldsymbol{\mu}_{cereal}$.

not co-purchased with q. Following the margin-based loss [8], we construct a max-margin loss function with the margin γ in Equation 2:

$$\mathcal{L}_{item}(q, v, v') = \max(0, \gamma - \log E(q, v) + \log E(q, v')) \quad (2)$$

where

$$\begin{aligned} \log E(q, v) = &-\frac{1}{2} \log \det \left(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_v\right) - \frac{d}{2} \log(2\pi) \\ &- \frac{1}{2} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_v)^T \left(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_v\right)^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_v) \end{aligned}$$

C. Connection to Existing User-Item-level CIRS models

Existing works on CIRS such as [6] show that user information will help improve the learning of item-to-item relationship in a collaborative way by introducing the user embedding to the **Item2Vec** [11] model. Our model can be extended easily with user information. Specially, we adapt the advantage of modeling the cohesion of each (*item, item, user*) triplet and modify the BPR loss [19] to model user-item relationship by minimizing the loss function 3 and 4, where $\sigma(\cdot)$ is the sigmoid function and q', v' represent the negative samples that are not purchased. These loss functions can be combined together with $\mathcal{L}_{item}(q, v, v')$ to form a new loss function $\mathcal{L}_{item}(q, v, v'|u) = \mathcal{L}_{item}(q, v, v') + \mathcal{L}_{BPR}(u, q, q') + \mathcal{L}_{BPR}(u, v, v')$.

$$\mathcal{L}_{BPR}(u, q, q') = 1 - \sigma\left(\boldsymbol{\theta}_u^T \boldsymbol{\mu}_q - \boldsymbol{\theta}_u^T \boldsymbol{\mu}_{q'}\right) \quad (3)$$

$$\mathcal{L}_{BPR}(u, v, v') = 1 - \sigma\left(\boldsymbol{\theta}_u^T \boldsymbol{\mu}_v - \boldsymbol{\theta}_u^T \boldsymbol{\mu}_{v'}\right) \quad (4)$$

D. Optimization

Depending on whether we consider $\mathcal{L}_{item}(q, v, v')$ or $\mathcal{L}_{item}(q, v, v'|u)$ for a given co-purchase record (q, v), the final objective function $\mathcal{L}$ can be written in Equation 5, where S denotes the sampled records for training and $\mathcal{L}_{item}$ could be $\mathcal{L}_{item}(q, v, v')$ or $\mathcal{L}_{item}(q, v, v'|u)$. We optimize $\mathcal{L}$ by mini-batch Stochastic Gradient Descent.

$$\mathcal{L} = \sum_{(q,v,v',u) \in S} \mathcal{L}_{item} \quad (5)$$

E. Complementary Item Recommendation

To recommend complementary items, we extract the item Gaussian embeddings and treat the mean vector of each item as its representation under complementary relation. To mitigate the impact of the vector magnitude when computing the distance between mean vectors for ranking and comparison, we follow **Item2Vec** [11] and **Triple2Vec** [6] and use the cosine similarity between two items' mean vectors to represent the relevance of the complementary relationship.

IV. Trustworthy Evaluation

Although we have addressed the label noise issue in the modeling step by considering the co-purchase data as a Gaussian distribution with item Gaussian embeddings, label noise will impact the evaluation accuracy as well for result reporting purpose. In this section, a trustworthy evaluation is developed to exam the models with high quality labels generated from an independence test-based method. Note that this evaluation does not require extra information (item description, co-view data, etc.) for creating the high quality labels.

Inspired from the definition of complementary items, we treat the purchase of an individual item v as a random variable from a Bernoulli distribution $Y_v \sim$ **Bernoulli**(p_v), and study the independence between two items' purchase to surface the item pairs which are co-purchased dependently. Pearson's chi-squared test is suitable for this task, as it can assess whether observations consisting of measures on two variables, expressed in a contingency table, are independent of each other [10]. Given two co-purchased items v_i and v_j, we define the 2-by-2 contingency table (Table I) for the observations of the purchase event between v_i and v_j with the 1 degree of freedom. Let N denote the total number of observed co-purchase records in the evaluation dataset. $\mathcal{F}_{v_i}$ ($\mathcal{F}_{v_j}$) represents the frequency of co-purchases including the item v_i (v_j) and O_i represents the observed frequency of different purchase events defined in Table I. Typically, O_1 represents the observed co-purchases of (v_i, v_j). Following the definition of $\mathcal{F}_{v_i}$ and $\mathcal{F}_{v_j}$, we can compute that $O_2 = \mathcal{F}_{v_j} - O_1$, $O_3 = \mathcal{F}_{v_i} - O_1$ and $O_4 = N - O_1 - O_2 - O_3 = N - \mathcal{F}_{v_i} - \mathcal{F}_{v_j} + O_1$.

Without any knowledge of item complementary relationships, we assume that each pair of co-purchased items, v_i and v_j, are independent (the null hypothesis in our test H_0). The alternative hypothesis H_a is that they are purchased dependently. We can compute the estimated frequency for each purchase event E_i based on the independence assumption by Table II. Following the Chi-squared test, we can compute the value of the Chi-squared statistics $\mathcal{X}^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$ which is used to determine the significance (p-value) by

TABLE I
2-BY-2 CONTINGENCY TABLE OF 4 DIFFERENT PURCHASE EVENTS BETWEEN v_i AND v_j.

	$Y_{v_i} = 1$	$Y_{v_i} = 0$	**SUM**
$Y_{v_j} = 1$	O_1 = frequency of observed co-purchase (v_i, v_j)	O_2 = frequency of observed co-purchase of v_j with all items$\backslash v_i$	$\mathcal{F}_{v_j}$
$Y_{v_j} = 0$	O_3 = frequency of observed co-purchase of v_i with all items$\backslash v_j$	O_4 = frequency of observed co-purchase w/o (v_i, v_j)	N - $\mathcal{F}_{v_j}$
SUM	$\mathcal{F}_{v_i}$	N - $\mathcal{F}_{v_i}$	N

comparing to a Chi-squared distribution with one degree of freedom. Item pairs which pass the Chi-squared test mean that their co-purchase are dependent.

Further more, we need to determine if the dependency of a co-purchased item pair is positive or negative. To achieve this, we require that the observed co-purchase frequency of an item pair should be larger than the expected frequency under independence assumption, $O_1 > E_1$, if a co-purchased item pair has a positive dependency. With a predefined p-value for the statistic significance, we can create the high quality co-purchase labels for evaluations. For clarity, we denote the item pairs which pass the Chi-squared test and $O_1 > E_1$ as the **positively-dependent item pairs** and the item pairs which pass the Chi-squared test and $O_1 <= E_1$ as the **negatively-dependent item pairs** in the rest of our paper. We summarize the algorithm of generating the trustworthy labels for evaluation in Algorithm 1.

Algorithm 1 Trustworthy Label Generation for Evaluation

```
Require: a transaction set $\mathcal{B}$, an empty hashtable $\Psi$, $\mathcal{X}^2$ threshold $t_{\mathcal{X}^2}$ for a p-value;
Ensure:
for each transaction $b$ in $\mathcal{B}$ do
    sample co-purchase item pairs $(v_i, v_j)$ from each transaction $b \in \mathcal{B}$, $v_i \neq v_j$;
    compute the frequency of purchasing $(v_i, v_j)$ together and store the frequency in $\Psi$, i.e., $\Psi[(v_i, v_j)]$ represents the co-purchase frequency of $(v_i, v_j)$;
end for
set $N = \sum_{(v_i,v_j)} \Psi[(v_i, v_j)]$;
set $\mathcal{F}_{v_i} = \sum_{(v_k,v_j),v_k=v_i} \Psi[(v_k, v_j)]$;
set $\mathcal{F}_{v_j} = \sum_{(v_i,v_k),v_k=v_j} \Psi[(v_i, v_k)]$;
for each $(v_i, v_j)$ stored in $\Psi$ do
    compute the 2-by-2 contingency table by $\Psi[(v_i, v_j)]$, $N, \mathcal{F}_{v_i}$ and $\mathcal{F}_{v_j}$ based on Table I;
    compute the table of expected value based on Table II;
    compute $\mathcal{X}^2_{(v_i,v_j)} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$
    if $\mathcal{X}^2_{(v_i,v_j)} > t_{\mathcal{X}^2}$ and $O_1 > E_1$ then
        mark $(v_i, v_j)$ as a qualified co-purchase label for evaluation
    end if
end for
```

TABLE II
EXPECTED VALUE OF 4 DIFFERENT PURCHASE EVENTS BETWEEN v_i AND v_j.

	$Y_{v_i} = 1$	$Y_{v_i} = 0$
$Y_{v_j} = 1$	$E_1 = \frac{\mathcal{F}_{v_i} \cdot \mathcal{F}_{v_j}}{N}$	$E_2 = \frac{(N - \mathcal{F}_{v_i}) \cdot \mathcal{F}_{v_j}}{N}$
$Y_{v_j} = 0$	$E_3 = \frac{\mathcal{F}_{v_i} \cdot (N - \mathcal{F}_{v_j})}{N}$	$E_4 = \frac{(N - \mathcal{F}_{v_i}) \cdot (N - \mathcal{F}_{v_j})}{N}$

V. EXPERIMENTS

In this section, we study **NEAT** by comparing it with the state-of-the-art baselines on the real-world datasets.

A. Dataset

For the publicly available dataset of raw transactions, we consider *Instacart* dataset (INS) published by [20]. The date of each order in this dateset is not provided but the sequence of transactions by each user is available. Items in each transaction are sorted by their purchase orders and the item-types are also provided in *Instacart* by the aisles. *Instacart* dataset has 134 aisles from 21 departments and 3.3 million transactions, which is small compared with the real-world applications with more item-types and larger volume of transactions. We use the default train (INS-T) and test (INS-E) split provided by *Instacart* dataset. To further study the model performance, we collect a proprietary dataset (WMT) with a larger scale from *Walmart* e-Commerce platform (www.walmart.com) following the same format of *Instacart*, where the sequence order of transactions are kept and the order of purchases in the same sequence is also preserved. For WMT dataset, we randomly sample 15.2 million transactions from the past 6-month history data and keep the latest 1.2 million transactions as our test dataset (WMT-E). The rest of 14 million transactions are used for training (WMT-T). Similar to INS dataset, we collect the item categories based on the taxonomy of *Walmart* platform. Co-purchase records are created from INS and WMT dataset respectively to serve model training and label generation for evaluation. Table III summarizes the statistics of the INS and WMT datasets.

B. Label Generation for Training and Evaluation

We follow the steps in III-A to collect all the co-purchase records for training from the training set. To improve the quality of labels for model training, we remove labels selected in the previous steps where two items are from the same aisles (for INS dataset) or the same category (for WMT dataset) to remove similar items. This is similar to the strategy used in [7] when co-view data are not available because items from the same aisle or category are similar and are likely to be co-viewed for substitution.

For evaluation, we create the trustworthy labels following Section IV under different p-value = $\{0.05, 0.01, 0.001\}$. We conduct the experiments on these unique labels for evaluation. As we mentioned previously, even these labels are with high quality, it is not practical to be used for training purpose, due to the limited coverage in item space. Table IV summarizes the number of unique labels of the INS and WMT datasets.

TABLE III
DATA DESCRIPTION OF WMT AND INS DATASET

	INS-T	INS-E	WMT-T	WMT-E
# Transactions	3,214,874	131,209	∼14 m	∼1.2 m
# Items	49,677	–	∼0.1 m	–
# Categories/Aisles	134	–	∼850	–
# Users	206,209	131,209	∼0.7 m	∼0.4 m

TABLE IV
NUMBER OF LABELS IN INS-E AND WMT-E

	p-value = 0.05	p-value = 0.01	p-value = 0.001
INS-E	7,961	6,077	4,752
WMT-E	78,719	53,119	36,251

C. Experiment Setup

*1) **Baselines**:* To evaluate the effectiveness of applying Gaussian distribution on co-purchase data and item embeddings, we compare **NEAT** with the following state-of-the-art baselines:

- **Collaborative Filtering (CF)** [21]: an item recommendation model which factorizes the user-item.
- **Bayesian Personalized Ranking (BPRMF)** [19]: an item recommendation model which factorizes the user-item implicit feedback from raw transactions by approximately optimizing the AUC ranking metric.
- **Item2Vec** [11]: the first model that learns vector representations of items via SGNS and optimizes the similarity between item vectors for co-purchase data. It can be used to model item complementarity by considering co-purchase records of item pairs as input. As aforementioned, most of the CIRS models can be viewed as **Item2Vec** plus auxiliary information such as graph, context and multi-modal data source. In our work, we focus on modeling item complementary relationship rather than the advantage of incorporating such auxiliary information into models. Hence, we choose this model as the baseline to represent other item-to-item CIRS models for a fair comparison.
- **Triple2Vec** [6]: this state-of-the-art model learns vector representations of item and user, and considers the triplet interaction between a user and her/his co-purchased item pair for complementarity. It can be viewed as an extension of **Item2Vec** with user embeddings.

Besides, we also consider two popularity-based baselines: **Popular Item (Pop)** and **Popular Co-purchase (PopCo)**. In **Pop**, the complementary item recommendations for the query item are the most popular items globally. In **PopCo**, we take the query item's popular co-purchased items as the complementary item recommendations.

*2) **NEAT** Variants:* Depending on whether incorporating the user-item level collaborative learning into the model, we develop two variants of our model:

- **NEAT** : This model is trained by optimizing $\mathcal{L}$ with $\mathcal{L}_{item} = \mathcal{L}_{item}(q, v, v')$ to model the item-level complementary co-purchase signals.
- **NEAT**+bpr: In addition to the item-level complementary signals, this model is trained by optimizing $\mathcal{L}$ with $\mathcal{L}_{item} = \mathcal{L}_{item}(q, v, v'|u)$ (see section III-C) to further model the user-item level collaborative learning for complementary signals.

*3) **Implementation Details**:* For simplicity, we set the covariance matrix in the **NEAT** model to be spherical. The margin γ in Equation 2 is set to be 0.5 for the computation of Hit-Rate (HR) and Normalized Discounted Cumulative Gain (NDCG). We applied the following settings for all models in the experiments, unless it is specified: the dimension of the item embeddings are set to be 100, the window size for sampling co-purchased items is set to be 5, and all models are trained for 5 epochs. For **Item2Vec**, **Triple2Vec** and our model, the batch size is 128, with the initial learning rate of 0.05 and the mini-batch Stochastic Gradient Descent (SGD) optimizer. We follow the skip-gram training paradigm and set the number of negative sample to 5 during training.

D. Study on Labels for Evaluation

1) Label Quality: In this section, we present the study of the trustworthy label generation method and show its effectiveness. There are three major concerns of data labeling: coverage, consistency and accuracy.

Coverage: a good data labeling method should have enough coverage on the representative patterns of the dataset. In our case, the label generation should show a good coverage of different item categories and departments instead of being biased to few item categories. To illustrate the coverage of our label generation method, we focus on the department level without the loss of generality and readability and compute the distribution of labels over different departments for the INS-E dataset in Table V. Compared with the distribution of total co-purchase records from the INS-E dataset, our labels show similar distributions over all departments. We notice that the Pets department is not covered by our labels. This is because most of the raw co-purchase records with pet-related items also consist of non-pet-related items like grocery in the INS-E dataset, which are not complementary. The label distribution over departments indicates that our method is not biased to a certain department and covers complementary signals of item purchase behaviors under various departments.

Consistency: by the design of our label generation method, the percentage of complementary labels should increase as the p-value decreases. To show such a consistency, we plot the distribution of $\mathcal{X}^2$ statistics for each item pair which passes the test for a given p-value for both positively dependent item pairs ($O_1 > E_1$) and negatively dependent item pairs ($O_1 <= E_1$) in Figure 2. We see that while most of the labels (both negative and positive) are with $\mathcal{X}^2$ statistics between 0 and 99, the percentage of positively dependent item pairs with higher $\mathcal{X}^2$ statistics has a larger lift as the p-value decreases compared with the negatively dependent item pairs. Because we use the positively dependent item pairs as the labels for evaluation, this consistency between the increase of more complementary labels and the decrease of p-value indicates that raising the

TABLE V
DISTRIBUTION OF LABELS OVER DEPARTMENTS OF INS-E DATASET

Department	Total	$p = 0.05$	$p = 0.01$	$p = 0.001$
alcohol	0.370%	0.339%	0.411%	0.505%
babies	1.451%	0.364%	0.378%	0.337%
bakery	4.462%	4.459%	4.048%	3.788%
beverages	9.437%	8.504%	8.672%	9.007%
breakfast	2.796%	1.306%	1.349%	1.410%
bulk	0.100%	0.100%	0.115%	0.126%
canned goods	4.001%	2.927%	2.781%	2.399%
dairy eggs	17.101%	23.678%	22.100%	20.623%
deli	3.594%	2.487%	2.403%	2.273%
dry goods pasta	3.559%	1.859%	1.678%	1.599%
frozen	8.715%	3.944%	3.966%	4.167%
household	3.016%	0.867%	0.889%	0.989%
international	1.138%	0.603%	0.675%	0.526%
meat seafood	2.474%	2.613%	2.271%	2.041%
missing	0.759%	1.017%	1.152%	1.305%
other	0.168%	0.038%	0.049%	0.042%
pantry	6.720%	2.927%	2.946%	2.736%
personal care	1.772%	0.113%	0.115%	0.147%
pets	0.484%	0.000%	0.000%	0.000%
produce	17.369%	30.511%	31.019%	31.587%
snacks	10.512%	11.343%	12.983%	14.394%
SUM	6249077	7961	6077	4752

TABLE VI
POSITIVELY-DEPENDENT ITEM PAIR, INS DATASET WITH P-VALUE = 0.001

Query Item	**Co-purchased Item**	$\mathcal{X}^2$
Beef Hot Dogs	Classic Hot Dog Buns	5084.536
Everything Bagels	Whipped Cream Cheese	85.501
Thin & Light Tortilla Chips	Medium Salsa Roja	239.958
Eggo Homestyle Waffles	Original Syrup	170.825
Cherrios Honey Nut (cereal)	Reduced Fat 2% Milk	62.804
Green Curry Paste	Organic Coconut Milk	51.005
Plain Mini Bagels	Philadelphia Cream Cheese Spread	33.513
Stand 'n Stuff Taco Shells	Original Taco Seasoning Mix	20.774
Snack Bags (food storage)	Sandwich Bags (food storage)	106.078
Fabric Softener Dryer Sheet	Tall Kitchen Bag With Febreze Odor Shield	1414.015

TABLE VII
NEGATIVELY-DEPENDENT ITEM PAIR, INS DATASET WITH P-VALUE = 0.001

Query Item	**Co-purchased Item**	$\mathcal{X}^2$
Organic Sea Salt Roasted Seaweed Snacks	Banana	108.817
Free & Clear Unscented Baby Wipes	Large Lemon	61.033
Naturals Savory Turkey Breakfast Sausage	Strawberries	18.558
Gluten Free Whole Grain Bread	Large Lemon	52.104
Eggo Homestyle Waffles	Organic Cucumber	42.681
Naturals Chicken Nuggets	Organic Avocado	60.222
Cheerios Honey Nut (cereal)	Jalapeno Peppers	33.853
Everything Bagels	Organic Strawberries	35.512
Taco Seasoning	Organic Raspberries	53.735
Laundry Detergent Free & Clear	Banana	16.293

TABLE VIII
TOP-20 GLOBALLY POPULAR ITEMS, INS DATASET

Rank	**Item**
1	Banana
2	Bag of Organic Bananas
3	Organic Strawberries
4	Organic Baby Spinach
5	Organic Hass Avocado
6	Organic Avocado
7	Large Lemon
8	Strawberries
9	Limes
10	Organic Whole Milk
11	Organic Raspberries
12	Organic Yellow Onion
13	Organic Garlic
14	Organic Zucchini
15	Organic Blueberries
16	Cucumber Kirby
17	Organic Fuji Apple
18	Organic Lemon
19	Apple Honeycrisp Organic
20	Organic Grape Tomatoes

significance level by p-value can further concentrate the co-purchase labels with positively dependence (complementary relationships).

Accuracy: we provide a case study of the positively dependent item pairs (our labels) and the negatively dependent item pairs to show that our model can provide more accurate labels for evaluation in Table VI and VII. Note that the chi-squared statistics $\mathcal{X}^2$ should be not smaller than 10.83 for p-value = 0.001. Both positive and negative item pairs show large enough chi-squared statistics. While the positive labels are showing clear complementary relationships, e.g., syrup for waffle, hot dog buns for hot dog, and kitchen bag for laundry-related items for household, the negative labels reflect the noise in the co-purchase records even though they pass the Chi-squared test. Most of the co-purchased items in the negative labels are fruits like Banana, which are the popular items in the INS dataset. See examples of top-20 popular items in the INS dataset in Table VIII. The comparison between the positive labels and the negative labels indicates that our label generation method can surface more complementary labels while suppressing the noise in the co-purchase records.

E. Evaluation on Item-level Co-purchase Data

Evaluation metrics: We mainly focus on HitRate (HR@K) and NDCG@K of evaluation. Given the query item q, we consider the top-K recommendations R_q has a hit on the test co-purchase record (q, v) if $v \in R_q$: HR@$K = \begin{cases} 1, & \text{if } v \in R_q \\ 0, & \text{otherwise} \end{cases}$. For NDCG@$K$, we consider the binary relevance score and define it as NDCG@$K = \begin{cases} \frac{1}{\log_2(1+rank_v)}, & \text{if } v \in R_q \\ 0, & \text{otherwise} \end{cases}$.

To evaluate the ability to surface complementary recommendations from the noisy co-purchase data, we firstly generate the recall set by taking the top-K most co-purchased items for

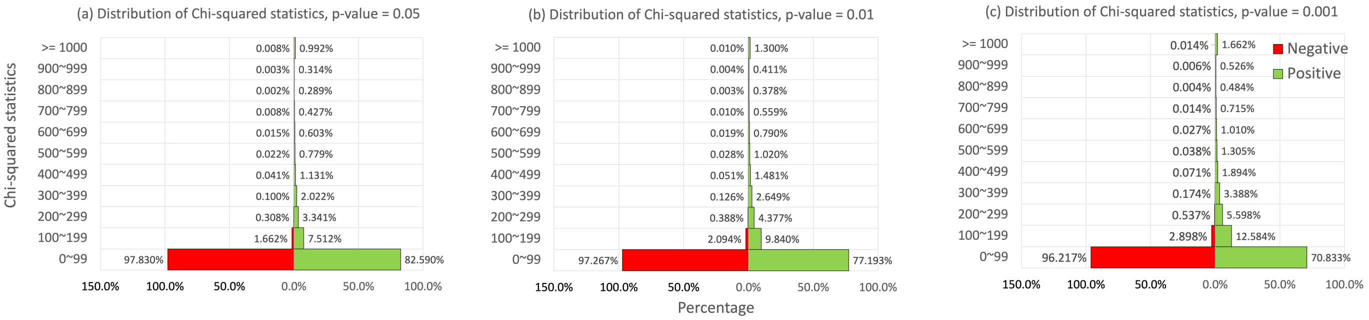

Fig. 2. Distribution of $\mathcal{X}^2$ of both positively dependent labels (in green) and negatively dependent labels (in red) with p-value = $\{0.05, 0.01, 0.001\}$.

the query item in the training data, rather than a sampled item set in which each ground truth item in the test set is paired with a few (e.g., 100) randomly sampled negative items [22] [23] [24] [25]. We report the average score over the co-purchase records for HR@K and NDCG@K, $K = \{1, 3, 5, 10, 20\}$.

Results: We summarize the results of HR@K and NDCG@K for INS dataset (in Tables IX-XI) and WMT dataset (in Tables XII-XIV). The best performance for each metric is highlighted in bold. **Pop** shows zero HR@K and NDCG@K when K is small. As aforementioned, popular items are involved in many co-purchase records which are not motivated by complementary relationships. After removing irrelevant co-purchase records from the dataset by the trustworthy label generation, **Pop** is less likely to hit a complementary co-purchase. **Popco** still achieves reasonable performance on all metrics because it captures the noisy item-to-item complementary relationship via ranking the co-purchased items by their co-purchase frequency with the query item. **Item2Vec** and **Triple2Vec** outperform the frequency-based baselines due to the advantage of item vector representation. Our model further improves the performance on both HR and NDCG compared with frequency-based baselines and the vector-based baselines. The results indicate the advantage of modeling the label noise in the co-purchase distribution.

*F. Ablation Study of **NEAT***

Our model can be extended with user embeddings to model the complementary relationship from the user-item-level co-purchase data. To study the extensibility of our model and the influence of involving user embeddings, we compute HR@K and NDCG@K for **NEAT** and **NEAT**+bpr, $K = \{1, 3, 5, 10, 20\}$. The results are summarized in Tables IX-XIV. We can see that both **NEAT** and **NEAT**+bpr perform similarly but **NEAT**+bpr outperforms **NEAT** in most cases when: (1) K becomes larger or (2) number of items increases from INS dataset to WMT dataset. This indicates that including user-item-level signals improves the model performance especially when the number of items is large.

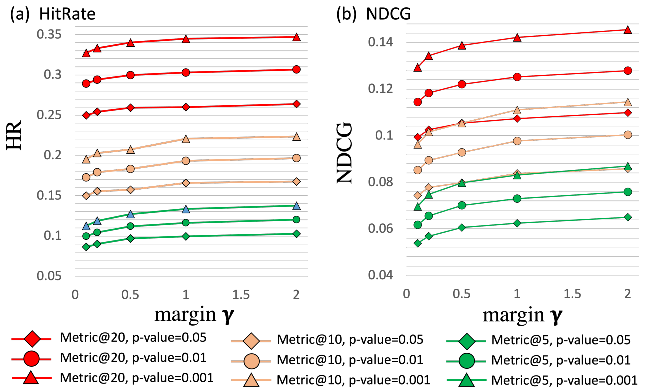

Fig. 3. Analysis of the margin γ on three label sets, p-value = $\{0.05, 0.01, 0.001\}$, for metric@$\{5, 10, 20\}$ of Hit-Rate and NDCG on INS dataset.

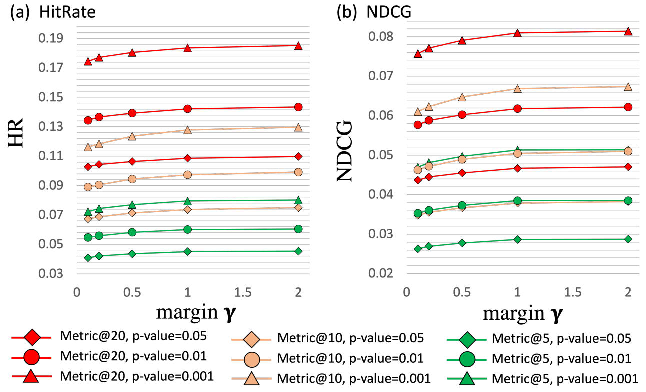

Fig. 4. Analysis of the margin γ on three label sets, p-value = $\{0.05, 0.01, 0.001\}$, for metric@$\{5, 10, 20\}$ of Hit-Rate and NDCG on WMT dataset.

G. Sensitivity Analysis of the Margin γ

We conduct experiments on **NEAT** with different margins $\gamma = \{0.1, 0.2, 0.5, 1.0, 2.0\}$ on the three label sets of INS dataset and WMT dataset respectively. We report HR@K and NDCG@K for evaluation with $K = \{5, 10, 20\}$ and summarize the results in Figure 3 and 4. The results indicate that the model is in favor of a larger margin.

H. Case Study: Item Representation as a Distribution

To test whether the Gaussian embedding of items could capture the variation of items in their co-purchase, we fo-

TABLE IX
INS LABELS, P-VALUE = 0.05

	HR@1	NDCG@1	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20
Pop	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0014	0.0005	0.0035	0.0010
PopCo	0.0122	0.0122	0.0437	0.0303	0.0734	0.0425	0.1334	0.0617	0.2168	0.0826
CF	0.0087	0.0087	0.0245	0.0176	0.0396	0.0238	0.0765	0.0355	0.1516	0.0543
BPRMF	0.0067	0.0067	0.0225	0.0155	0.0368	0.0214	0.0720	0.0326	0.1467	0.0512
Item2Vec	0.0196	0.0196	0.0484	0.0360	0.0746	0.0468	0.1271	0.0636	0.2231	0.0876
Triple2Vec	0.0221	0.0221	0.0541	0.0403	0.0813	0.0514	0.1325	0.0678	0.2110	0.0874
NEAT	**0.0252**	**0.0252**	**0.0633**	**0.0468**	**0.0970**	**0.0606**	0.1574	0.0798	**0.2593**	**0.1054**
NEAT+bpr	0.0249	0.0249	0.0628	0.0464	0.0927	0.0586	**0.1628**	**0.0811**	0.2591	0.1053

TABLE X
INS LABELS, P-VALUE = 0.01

	HR@1	NDCG@1	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20
Pop	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0015	0.0005	0.0023	0.0007
PopCo	0.0155	0.0155	0.0541	0.0378	0.0900	0.0526	0.1593	0.0747	0.2587	0.0996
CF	0.0100	0.0100	0.0276	0.0200	0.0443	0.0268	0.0849	0.0397	0.1711	0.0612
BPRMF	0.0076	0.0076	0.0262	0.0180	0.0415	0.0243	0.0819	0.0372	0.1654	0.0580
Item2Vec	0.0230	0.0230	0.0559	0.0418	0.0859	0.0541	0.1450	0.0729	0.2549	0.1004
Triple2Vec	0.0253	0.0253	0.0635	0.0472	0.0931	0.0593	0.1502	0.0775	0.2391	0.0998
NEAT	**0.0293**	**0.0293**	**0.0734**	**0.0543**	**0.1121**	**0.0701**	0.1833	0.0928	0.2998	0.1221
NEAT+bpr	0.0286	0.0286	0.0732	0.0540	0.1084	0.0684	**0.1899**	**0.0945**	**0.3011**	**0.1224**

TABLE XI
INS LABELS, P-VALUE = 0.001

	HR@1	NDCG@1	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20
Pop	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0013	0.0004	0.0023	0.0007
PopCo	0.0189	0.0189	0.0640	0.0450	0.1048	0.0618	0.1833	0.0869	0.2963	0.1152
CF	0.0112	0.0112	0.0316	0.0227	0.0499	0.0302	0.0943	0.0443	0.1896	0.0681
BPRMF	0.0082	0.0082	0.0286	0.0197	0.0452	0.0266	0.0922	0.0415	0.1841	0.0645
Item2Vec	0.0265	0.0265	0.0623	0.0468	0.0962	0.0607	0.1616	0.0816	0.2870	0.1129
Triple2Vec	0.0276	0.0276	0.0711	0.0525	0.1040	0.0659	0.1681	0.0864	0.2668	0.1111
NEAT	0.0335	0.0335	**0.0835**	**0.0619**	**0.1273**	**0.0798**	0.2075	0.1054	0.3403	0.1388
NEAT+bpr	**0.0341**	**0.0341**	0.0823	0.0613	0.1227	0.0778	**0.2163**	**0.1078**	**0.3424**	**0.1395**

TABLE XII
WMT LABELS, P-VALUE = 0.05

	HR@1	NDCG@1	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20
Pop	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0001	0.0014	0.0004
PopCo	0.0069	0.0069	0.0207	0.0148	0.0310	0.0190	0.0506	0.0253	0.0803	0.0328
CF	0.0033	0.0033	0.0076	0.0058	0.0105	0.0070	0.0193	0.0098	0.0451	0.0162
BPRMF	0.0042	0.0042	0.0108	0.0080	0.0164	0.0103	0.0276	0.0139	0.0505	0.0196
Item2Vec	0.0082	0.0082	0.0200	0.0149	0.0298	0.0189	0.0504	0.0256	0.0818	0.0335
Triple2Vec	0.0087	0.0087	0.0210	0.0158	0.0294	0.0192	0.0438	0.0239	0.0615	0.0283
NEAT	0.0120	0.0120	0.0292	0.0219	0.0437	0.0278	0.0715	0.0367	0.1065	0.0455
NEAT+bpr	**0.0121**	**0.0121**	**0.0298**	**0.0221**	**0.0437**	**0.0278**	**0.0717**	**0.0368**	**0.1074**	**0.0459**

TABLE XIII
WMT LABELS, P-VALUE = 0.01

	HR@1	NDCG@1	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20
Pop	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0013	0.0003
PopCo	0.0099	0.0099	0.0291	0.0208	0.0432	0.0266	0.0695	0.0351	0.1080	0.0448
CF	0.0042	0.0042	0.0098	0.0074	0.0134	0.0089	0.0244	0.0124	0.0574	0.0206
BPRMF	0.0055	0.0055	0.0141	0.0104	0.0212	0.0133	0.0359	0.0180	0.0648	0.0252
Item2Vec	0.0110	0.0110	0.0261	0.0196	0.0388	0.0248	0.0649	0.0332	0.1059	0.0435
Triple2Vec	0.0117	0.0117	0.0273	0.0207	0.0379	0.0250	0.0563	0.0310	0.0786	0.0366
NEAT	0.0165	0.0165	0.0393	0.0295	**0.0583**	**0.0373**	**0.0945**	**0.0490**	0.1393	0.0603
NEAT+bpr	**0.0165**	**0.0165**	**0.0401**	**0.0299**	0.0582	0.0373	0.0944	0.0490	**0.1402**	**0.0606**

cus on three items, Whole Milk, Cereal and Organic Tortilla Chips in INS dataset, and study the relationship between item Gaussian embeddings and complementary relationships when Whole Milk becomes the query item. On one hand, the cosine similarity of $\boldsymbol{\mu}_{\texttt{Whole Milk}}$ and $\boldsymbol{\mu}_{\texttt{Cereal}}$ is larger then that of $\boldsymbol{\mu}_{\texttt{Whole Milk}}$ and $\boldsymbol{\mu}_{\texttt{Organic Tortilla Chips}}$,

TABLE XIV
WMT LABELS, P-VALUE = 0.001

	HR@1	NDCG@1	HR@3	NDCG@3	HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20
Pop	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0012	0.0003
PopCo	0.0140	0.0140	0.0407	0.0293	0.0599	0.0371	0.0948	0.0484	0.1437	0.0607
CF	0.0055	0.0055	0.0129	0.0097	0.0173	0.0115	0.0313	0.0160	0.0730	0.0263
BPRMF	0.0069	0.0069	0.0177	0.0130	0.0271	0.0168	0.0461	0.0229	0.0831	0.0322
Item2Vec	0.0145	0.0145	0.0342	0.0257	0.0504	0.0324	0.0837	0.0431	0.1365	0.0563
Triple2Vec	0.0156	0.0156	0.0356	0.0271	0.0491	0.0327	0.0725	0.0402	0.1010	0.0474
NEAT	**0.0226**	**0.0226**	0.0524	0.0396	**0.0771**	**0.0498**	0.1237	0.0648	0.1806	0.0792
NEAT+bpr	0.0223	0.0223	**0.0532**	**0.0399**	0.0771	0.0497	**0.1241**	**0.0649**	**0.1815**	**0.0794**

which aligns with the expectation of stronger complementary relationship between Whole Milk and Cereal. On the other hand, the query item Whole Milk which has higher popularity than Cereal and Organic Tortilla Chips in INS dataset also shows higher variation (indicated by the determinant of the spherical covariance matrix). In particular, the $\det(\mathbf{\Sigma}_{\texttt{Whole Milk}})$ is 30 times larger than $\det(\mathbf{\Sigma}_{\texttt{Cereal}})$ and is 547 times larger than $\det(\mathbf{\Sigma}_{\texttt{Organic Tortilla Chips}})$. This also aligns with our expectation of their variation since Whole milk (35633 purchases) is more popular than Cereal (12184 purchases) and Organic Tortilla Chips (13776 purchases) in INS dataset and hence more likely to form irrelevant co-purchases.

VI. CONCLUSIONS

In this paper, we proposed a label noise-resistant complementary item recommendation model named **NEAT** to address the label noise issue for complementary item recommendation when the co-purchase data are used as labels. **NEAT** learns the item representations as Gaussian embeddings, and assumes the co-purchase data as a Gaussian distribution, where the mean is the co-purchases from the true complementary relation, and the variance is the co-purchases from the noise. In addition, we developed a trustworthy label generation method for model evaluation to alleviate the impact of noisy labels in evaluation step. We performed extensive experiments on two real-world datasets and the results show the effectiveness of the proposed method over state-of-the-art models.

REFERENCES

[1] Y. Liu, Y. Gu, Z. Ding, J. Gao, Z. Guo, Y. Bao, and W. Yan, "Decoupled graph convolution network for inferring substitutable and complementary items," in *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, Eds. ACM, 2020, pp. 2621–2628. [Online]. Available: https://doi.org/10.1145/3340531.3412695

[2] J. J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. ACM, 2015, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2783258.2783381

[3] Z. Wang, Z. Jiang, Z. Ren, J. Tang, and D. Yin, "A path-constrained framework for discriminating substitutable and complementary products in e-commerce," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, Y. Chang, C. Zhai, Y. Liu, and Y. Maarek, Eds. ACM, 2018, pp. 619–627. [Online]. Available: https://doi.org/10.1145/3159652.3159710

[4] T. Chen, H. Yin, G. Ye, Z. Huang, Y. Wang, and M. Wang, "Try this instead: Personalized and interpretable substitute recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020, pp. 891–900. [Online]. Available: https://doi.org/10.1145/3397271.3401042

[5] S. Zhang, H. Yin, Q. Wang, T. Chen, H. Chen, and Q. V. H. Nguyen, "Inferring substitutable products with deep network embedding," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 4306–4312. [Online]. Available: https://doi.org/10.24963/ijcai.2019/598

[6] M. Wan, D. Wang, J. Liu, P. Bennett, and J. J. McAuley, "Representing and recommending shopping baskets with complementarity, compatibility and loyalty," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. ACM, 2018, pp. 1133–1142. [Online]. Available: https://doi.org/10.1145/3269206.3271786

[7] J. Hao, T. Zhao, J. Li, X. L. Dong, C. Faloutsos, Y. Sun, and W. Wang, "P-companion: A principled framework for diversified complementary product recommendation," in *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, Eds. ACM, 2020, pp. 2517–2524. [Online]. Available: https://doi.org/10.1145/3340531.3412732

[8] L. Vilnis and A. McCallum, "Word representations via gaussian embedding," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6623

[9] T. Jebara, R. Kondor, and A. G. Howard, "Probability product kernels," *J. Mach. Learn. Res.*, vol. 5, pp. 819–844, 2004. [Online]. Available: http://jmlr.org/papers/volume5/jebara04a/jebara04a.pdf

[10] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[11] O. Barkan and N. Koenigstein, "ITEM2VEC: neural item embedding for collaborative filtering," in *26th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2016, Vietri sul Mare, Salerno, Italy, September 13-16, 2016*. IEEE, 2016, pp. 1–6. [Online]. Available: https://doi.org/10.1109/MLSP.2016.7738886

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

[13] Y. Zhang, H. Lu, W. Niu, and J. Caverlee, "Quality-aware neural complementary item recommendation," in *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, S. Pera, M. D. Ekstrand, X. Amatriain, and J. O'Donovan, Eds. ACM, 2018, pp. 77–85. [Online]. Available: https://doi.org/10.1145/3240323.3240368

[14] D. Xu, C. Ruan, E. Körpeoglu, S. Kumar, and

K. Achan, “Product knowledge graph embedding for e-commerce,” *CoRR*, vol. abs/1911.12481, 2019. [Online]. Available: http://arxiv.org/abs/1911.12481

[15] Z. Liu, M. Wan, S. Guo, K. Achan, and P. S. Yu, “Basconv: Aggregating heterogeneous interactions for basket recommendation with graph convolutional neural network,” in *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020, Cincinnati, Ohio, USA, May 7-9, 2020*, C. Demeniconi and N. V. Chawla, Eds. SIAM, 2020, pp. 64–72. [Online]. Available: https://doi.org/10.1137/1.9781611976236.8

[16] Z. Liu, X. Li, Z. Fan, S. Guo, K. Achan, and P. S. Yu, “Basket recommendation with multi-intent translation graph neural network,” *CoRR*, vol. abs/2010.11419, 2020. [Online]. Available: https://arxiv.org/abs/2010.11419

[17] L. D. Santos, B. Piwowarski, and P. Gallinari, “Gaussian embeddings for collaborative filtering,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. ACM, 2017, pp. 1065–1068. [Online]. Available: https://doi.org/10.1145/3077136.3080722

[18] J. Jiang, D. Yang, Y. Xiao, and C. Shen, “Convolutional gaussian embeddings for personalized recommendation with uncertainty,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 2642–2648. [Online]. Available: https://doi.org/10.24963/ijcai.2019/367

[19] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “BPR: bayesian personalized ranking from implicit feedback,” *CoRR*, vol. abs/1205.2618, 2012. [Online]. Available: http://arxiv.org/abs/1205.2618

[20] Instacart, “Instacart market basket analysis,” Accessed on Dec. 2017. [Online]. Available: https://www.kaggle.com/c/instacart-market-basket-analysis/data

[21] Y. Hu, Y. Koren, and C. Volinsky, “Collaborative filtering for implicit feedback datasets,” in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 2008, pp. 263–272. [Online]. Available: https://doi.org/10.1109/ICDM.2008.22

[22] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. Chua, “Neural collaborative filtering,” in *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, Eds. ACM, 2017, pp. 173–182. [Online]. Available: https://doi.org/10.1145/3038912.3052569

[23] W. Kang and J. J. McAuley, “Self-attentive sequential recommendation,” in *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*. IEEE Computer Society, 2018, pp. 197–206. [Online]. Available: https://doi.org/10.1109/ICDM.2018.00035

[24] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, “Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, and J. X. Yu, Eds. ACM, 2019, pp. 1441–1450. [Online]. Available: https://doi.org/10.1145/3357384.3357895

[25] J. Tang and K. Wang, “Personalized top-n sequential recommendation via convolutional sequence embedding,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, Y. Chang, C. Zhai, Y. Liu, and Y. Maarek, Eds. ACM, 2018, pp. 565–573. [Online]. Available: https://doi.org/10.1145/3159652.3159656