

# Understanding and Promoting Micro-Finance Activities in Kiva.org

Jaegul Choo  
Georgia Institute of  
Technology  
jaegul.choo@cc.gatech.edu

Changhyun Lee  
Georgia Institute of  
Technology  
clee407@gatech.edu

Daniel Lee  
Georgia Tech Research  
Institute  
daniel.lee@gtri.gatech.edu

Hongyuan Zha  
Georgia Institute of  
Technology  
zha@cc.gatech.edu

Haesun Park  
Georgia Institute of  
Technology  
hpark@cc.gatech.edu

## ABSTRACT

Non-profit Micro-finance organizations provide loaning opportunities to eradicate poverty by financially equipping impoverished, yet skilled entrepreneurs who are in desperate need of an institution that lends to those who have little. Kiva.org, a widely-used crowd-funded micro-financial service, provides researchers with an extensive amount of publicly available data containing a rich set of heterogeneous information regarding micro-financial transactions. Our objective in this paper is to identify the key factors that encourage people to make micro-financing donations, and ultimately, to keep them actively involved. In our contribution to further promote a healthy micro-finance ecosystem, we detail our personalized loan recommendation system which we formulate as a supervised learning problem where we try to predict how likely a given lender will fund a new loan. We construct the features for each data item by utilizing the available connectivity relationships in order to integrate all the available Kiva data sources. For those lenders with no such relationships, e.g., first-time lenders, we propose a novel method of feature construction by computing joint nonnegative matrix factorizations. Utilizing gradient boosting tree methods, a state-of-the-art prediction model, we are able to achieve up to 0.92 AUC (area under the curve) value, which shows the potential of our methods for practical deployment. Finally, we point out several interesting phenomena on lenders' social behaviors in micro-finance activities.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information filtering; I.2.6 [Artificial Intelligence]: Learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
WSDM '14, February 24–28, 2014, New York, New York, USA.  
Copyright 2014 ACM 978-1-4503-2351-2/14/02 ...\$15.00.  
<http://dx.doi.org/10.1145/2556195.2556253>.

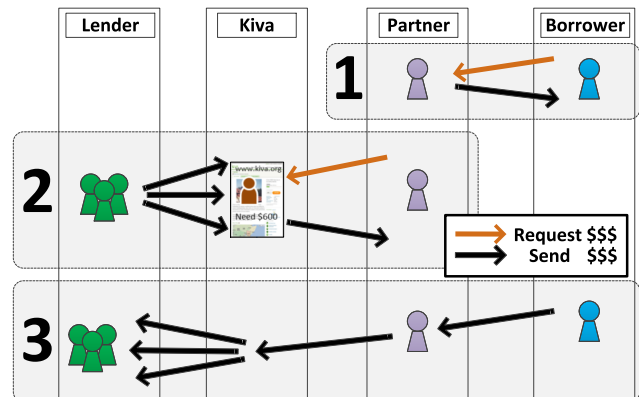


Figure 1: An overview of how Kiva works. 1. A borrower requests a loan to a (field) partner, and a loan is disbursed. 2. The partner uploads a loan request to Kiva, and lenders fund the loan. 3. The borrower makes repayments through the partner, and Kiva then repays the lenders. They can make another loan, donate to Kiva, or withdraw the money to their PayPal account.

## Keywords

Recommender systems; cold-start problem; microfinance; crowdfunding; joint matrix factorization; gradient boosting tree; heterogeneous data

## 1. INTRODUCTION

Kiva was founded by Matt Flannery and Jessica Jackley who based their concept on the inspiration of Muhammad Yunus' lecture on the Grameen Bank. The Grameen Bank, which won the Nobel Peace Prize in 2006 for its impact in helping the impoverished, was founded by Yunus in 1977 to address the lack of practical credit available to the under-utilized, yet skillful entrepreneurs in impoverished countries [32]. In Yunus' Book, he documented how he came up with the concept by noticing that the very poor could barely sustain themselves, let alone work their trade, since many times the poor were taking loans to buy the materials, only to sell their finished product back as repayment. In response, Yunus began his credit-loaning program which

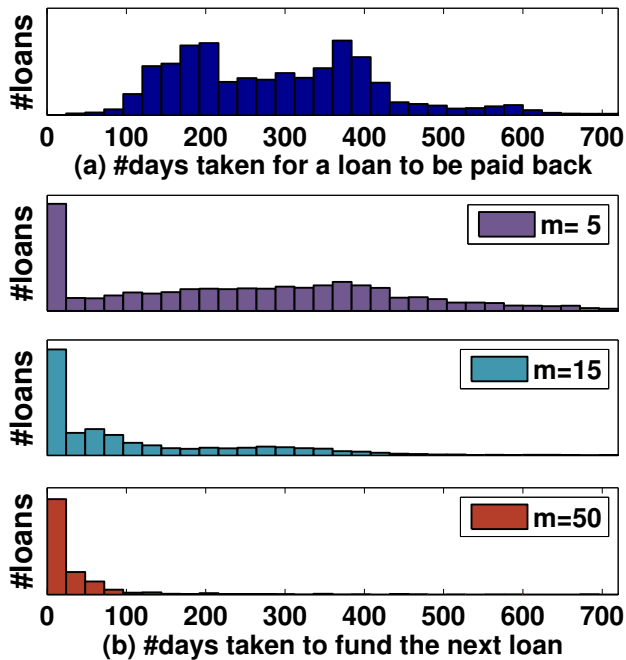


Figure 2: Temporal lending patterns for different lender groups with a specific lending count  $m$

provided loans without collateral and interest, and with an easy repayment plan. There are multitudes of success stories in Yunus’ book as well as on the Kiva blog<sup>1</sup> that portray how micro-financing has given opportunity to change the lives of the borrowers, their businesses, and their local areas.

Since its inception in 2005, Kiva and its generous lenders now impact the lives of courageous, hardworking borrowers across 72 countries. Kiva<sup>2</sup> is a non-profit micro-financial organization which acts as an intermediary service to provide people with the opportunity to lend money to underprivileged entrepreneurs in developing countries. Kiva’s lending model is based on a crowd-funding model in which any individual can fund a particular loan by contributing to a loan individually or as a part of a lender team. The Kiva loan process is summarized in Fig. 1.<sup>3</sup>

Open public access to Kiva’s data, provided through daily snapshots and an API, is a part of Kiva’s charitable initiative to provide the working poor with an infrastructure that Kiva hopes will encourage life-changing lending. This level of transparency lies at the core of Kiva’s successful growth as Matt Flannery puts it: “Transparency in this next period will be our best weapon against the challenges of growth. This model thrives on information, not marketing” [13].

Kiva data contain a wealthy set of heterogeneous information about lenders, loans, lender teams, borrowers, and field partners. As of June 2013, the publicly available Kiva data set contained over 1,100,000 lenders, 500,000 loans, and 150,000 journal entries for over 4,000,000 transactions that resulted in over 400,000,000 USD of loans issued. There are also multiple types of many-to-many relationships between each of the data entities. For example, lenders may be a part of multiple lender teams while lenders may choose any number of loans to participate in. Furthermore, borrowers

may optionally update their progress to their field partner for entry into the Kiva website as a journal entry. This data set includes geospatial, temporal, and free-text data along with a variety of other numerical and categorical information, consequently forming a fascinating set of data for many data mining and social media researchers.

**Loan recommendation and diverse lender behaviors.** Kiva as a non-profit organization encourages lending by promoting the idea that those in need can create better lives for themselves and their families when given the opportunity, i.e., capital. Thus, one can naturally realize that lenders, who are also regarded as donors due to the lack of any interest or reward they receive in return for their loan, are a pivotal component to the Kiva model. Consequently, one of the keys to a healthy Kiva ecosystem relies on keeping their lenders interested in continuing in their generous donations. This is where active recommendation can play a major role by matching the lender with loans that they would be sincerely interested in.

In addition, what makes loan recommendation an interesting problem is the diversity of lenders’ behaviors. How do lenders differ in their lending behaviors and what are the major factors to drive these differences? Fig. 2 displays an example showing temporal lending patterns. For a particular loan to be fully paid, it usually takes from a half to a full year (Fig. 2(a)). In case of passive lenders with a small number of lending experiences (a smaller  $m$  in Fig. 2(b)), the time taken between two consecutive lending activities show a relatively high correlation with the time required for a loan to be paid, compared to the other cases. This behavior is most likely explained by the notion that some passive lenders participate in another loan when their initial loan is paid back, rather than contributing more money of their own. However, active lenders with more lending experiences continue their lending activities mainly within a short time interval, as shown in a strong peak with almost no tail in the examples with a larger  $m$  in Fig. 2(b).

**Challenges in loan recommendation.** The problem of loan recommendation presents various challenges compared to other traditional recommendation problems.

The first is the *transient nature* of loans. Standard recommendation techniques based on collaborative filtering primarily utilize other similar users’ ratings or preferences on the items for recommendation. The key notion is that the items being recommended such as books or movies are persistent and reusable, i.e., an item (or a copy thereof) can serve many users. Loans, on the other hand, are transient and a particular loan can only serve a single borrower. More importantly, loans are only available for a short amount of time until the loan request is fully met, often leaving little or no information available to utilize from previous lenders.

The second challenge is the *binary rating* structure. Most rating systems are composed of a multi-grade set of ratings from which a user can select, yet in Kiva, the only information available similar to a rating is whether or not s/he funded the loan.<sup>4</sup> Furthermore, the fact that the funding did not happen may not necessarily mean that s/he did not like it. This challenge is often found in other settings where the recommendation relies only on previous purchases, viewing of item pages, etc. Such limited information and ambigu-

<sup>1</sup><http://pages.kiva.org/kivablog>

<sup>2</sup><http://www.kiva.org/>

<sup>3</sup><http://www.kiva.org/about/how>

<sup>4</sup>Individual loan amount could be utilized similarly to rating information, but such information is not available from Kiva API for lender privacy.

ity require more than just standard collaborative filtering approaches.

Finally, another challenge is the *heterogeneity* of data. The Kiva data set comprises a variety of intertwined entities giving rise to a rich set of heterogeneous information. Merging and fusing this diverse set of information in a unified predictive framework for loan recommendation presents a non-trivial problem.

**Overview of proposed approaches.** In order to better handle these challenges and deeply analyze various lending patterns among Kiva users, we propose a supervised learning approach to tackle this unique loan recommendation problem. That is, we formulate it as a binary classification/regression problem, where, given a lender and loan pair, the trained model computes the score that represents the likelihood of funding. In order to train our model with all the available information, we propose two main feature generation methods: (1) *graph-based feature integration* (for lenders with previous loans) (2) *feature alignment via joint nonnegative matrix factorization* (for lenders with no previous loans). The former provides us with a general framework for incorporating all the available heterogeneous information to represent a lender-loan pair. On the other hand, the latter alleviates the lack of information for newcomers, which is a well-known issue referred to as the cold-start problem in many recommendation applications.

Utilizing the proposed approaches along with a gradient boosting tree, a state-of-the-art learner model, we achieve a practically useful level of performance up to around 0.92 AUC (area under the curve) value. Furthermore, we present in-depth analysis of the resulting model and its output, revealing various interesting knowledge about lenders' social behaviors in micro-finance activities.

The rest of this paper is organized as follows. Section 2 describes our basic preprocessing steps to handle the heterogeneity of Kiva data; in addition, we have made the post-processed data readily available on the web for other researchers. Section 3 describes our main approaches for loan recommendation, and Section 4 presents the prediction performances as well as various findings from our analysis. Section 5 discusses related work. Finally, Section 6 concludes the paper and discusses future work.

## 2. BASIC DATA REPRESENTATION

The Kiva data set is composed of various entities, each of which has its own set of rich information including unstructured data (e.g., text, image, and video) as well as structured data (e.g., geo-spatial, numerical, categorical, and ordinal data). Lender entities contain basic web profile data, i.e., profile image, registration timestamp, location, loan count, and other fields, in addition to links to various types of entities. For example, a lender will have links to loans that s/he has funded and to any number of lender teams with which s/he is affiliated. Field partners manage loans within their local region, while borrowers request loans from their local field partner in respect to their lack of access to a computer with internet access.

Kiva provides a recent snapshot of its data set in JSON and XML formats,<sup>5</sup>. For our work, we used a 2.9 GB JSON snapshot which was collected on 5/31/2013. We preprocessed it to obtain the numerical representations of each

available field. Particularly, the preprocessing of temporal, categorical, and textual fields all required a nontrivial amount of work. For temporal data, such as the loan's *posting date* and lender's *sign-up date*, we converted it to a serial date number using Matlab's *datenum* function, which represents the whole and fractional number of days from a fixed preset date of January 0 in year 0000. For categorical data, such as a lender's *gender* and a loan's *country code*, we used a dummy encoding scheme which converts a variable with  $m$  categories into an  $m$ -dimensional binary vector where only the values in the corresponding categories are set to ones.

Finally, for textual data, we encoded each textual field separately as a bag-of-words vector where an individual dimension corresponds to a unique word. Afterwards we reduced the dimensionality using nonnegative matrix factorization<sup>6</sup> (NMF) [21, 19] to 100 for each textual field. We performed dimension reduction for two reasons. First, although the encoded representations may be in sparse format, the entire dimension easily amounts up to the hundreds of thousands requiring enormous computational time in learning our prediction model. Second, the reduced dimensions, which are composed of a group of words, are more semantically meaningful than individual term dimensions, and thus, they can be versatile for both good prediction performance and data/model understanding [10, 30]. The reduced dimension was set to 100 because larger values did not improve the prediction performances reported in Section 4.

As a final preprocessing step we created mappings between entities from the different tables. For example, a lender entity found in the table containing metadata for lenders may have a different identifier in another table about the lender-loan graph, and even worse, it may exist in only one table, meaning that some information about it will be completely missing. The mappings we created allow these issues to be handled with ease. We made the processed formatted data as Matlab files available at <http://fodava.gatech.edu/processed-kiva-data>.

## 3. METHODOLOGY

In this section, we describe our methodology for promoting non-profit micro-finance activities in Kiva. We formulate this task as a binary classification/regression problem. That is, we consider a pair  $(u, l)$  of a lender  $u$ <sup>7</sup> and a loan  $l$  as an individual data item, and given such a pair, we intend to predict how likely s/he will fund the loan, which we denote as  $f(u, l)$ . The associated label is set to 1 if funding occurred for the pair and 0 otherwise. Once the learner model is trained based on a set of data items along with these labels, it can then predict the likelihood of funding for any given lender-loan pair. Such a capability is broadly applicable in various loan recommendation problems. For example, it allows one to identify the best matching lender for a particular loan by solving  $\arg \max_u f(u, l)$  for a fixed  $l$  as well as the most appropriate loan to recommend given a particular lender by solving  $\arg \max_l f(u, l)$  for a fixed  $u$ .

In this approach, the key procedure affecting the overall performance is feature generation, i.e., how we characterize and represent a particular lender-loan pair. This is especially challenging considering the complexity of the Kiva data set which involves heterogeneous entities, such as bor-

<sup>5</sup><http://build.kiva.org/docs/data/snapshots>

<sup>6</sup><http://www.cc.gatech.edu/~hpark/nmfsoftware.php>

<sup>7</sup>We use an acronym  $u$  by viewing a lender as a kiva 'u' ser.

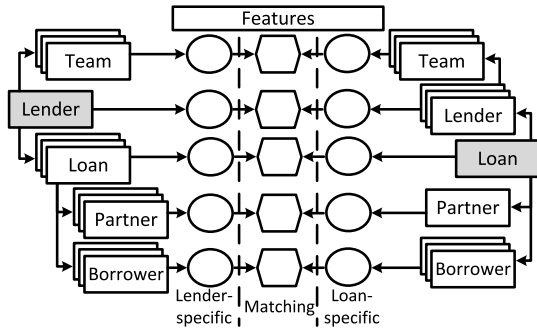


Figure 3: A graph-based feature integration for a lender-loan pair (grey-colored).

rowers, field partners, loans, lenders, and lender teams, with their own various set of information and complex relationships among them. To properly handle this issue, we act appropriately for two situations split by whether or not a lender has had previous funding experiences. In the following, we present our feature generation procedure for each case in detail.

### 3.1 Graph-based Feature Integration

When information about previous funding experiences of a particular lender is available, we utilize relationship links between different entities to take into account all the information available from the linked entities. As summarized in Fig. 3, given a lender-loan pair  $(u, l)$ , we first retrieve all the linked entities from both the lender and the loan. Specifically, a lender  $u$  will contain links to the list of teams s/he is affiliated with, loans s/he funded previously, and partners and borrowers his/her previous loans were associated with. Similarly, a loan will contain the links to the associated partner and the lists of borrowers, lenders (excluding the lender of interest), and lender teams that lenders are affiliated with.

**Lender- and loan-specific features.** Each entity type, e.g., the  $i$ -th type among a borrower, a partner, a loan, a lender, and a lender team, composes the entity-type-wise feature (column) vectors,  $v_i^u$  and  $v_i^l$ , to represent a lender  $u$  and a loan  $l$ , respectively, which, in turn, form a lender-specific feature vector  $v^u = [v_1^u \ \dots \ v_5^u]^T$  and a loan-specific one  $v^l = [v_1^l \ \dots \ v_5^l]^T$  (circles in Fig. 3).

In this process, one issue is that we may have a variable number of linked entities of the same type. For instance, one lender may have funded four loans in the past, yet another may have funded fifteen. To maintain a fixed number of dimensions for  $v_i^u$  (or  $v_i^l$ ) given a variable number of entities, we aggregate them into a single set of features by adding up all the feature vectors of individual entities. Suppose the  $i$ -th entity type is a loan and a lender  $u$  is associated with a set of entities (loans)  $\{(e_i^u)_j : j = 1, \dots, n\}$  where an entity  $(e_i^u)_j$  is represented as a feature vector  $(v_i^u)_j$ . The feature vector  $v_i^u$  (of the  $i$ -th entity type) for  $u$  is represented as

$$v_i^u = \sum_j (v_i^u)_j. \quad (1)$$

For example, a loan’s *requested amount* (in dollars) will correspond to the summation of the values from multiple loans, a single value indicating a total *requested amount*. For categorical variables, such as a lender’s gender which is represented as a binary vector in two dimensions, after summing up the feature vectors of lenders for a particular loan,

the values corresponding to the two dimensions become the number of male and females lenders, respectively. The same idea can also be applied to textual features, which are *non-negative* representations computed by NMF.

In addition, even if there are no links to entities of a particular entity type, e.g., no associated loans for a particular lender, Eq. (1) still holds since it will produce an equal-dimensional feature vector containing all zeros.

**Lender-loan matching features.** We have described how we generate lender- and loan-specific features by including information from each of the linked entities. We note that although the resulting data include links to heterogeneous entity types, both a lender and a loan now have counterparts generated from the same entity type, which can be directly compared with each other. In other words, both lenders and loans will have all the feature sets associated with borrowers, field partners, loans, lenders, and lender teams. Intuitively, if the entities from a lender side and a loan side are similar, our predictor  $f(u, l)$  should give a high score about the likelihood of funding. To leverage this in our feature representation, we generate an additional set of features  $v^{ul}$  that indicate how well the entities of the same type matches in an individual feature level. To this end, we compute the product of individual features referring to them as lender-loan matching features (hexagons in Fig. 3), i.e.,  $v^{ul} = v^u \circ v^l$ , where  $\circ$  represents an element-wise product. Given the nonnegativity of  $v^u$  and  $v^l$ ,  $v^{ul}$  indicates how strongly the values of a particular dimension are represented in ‘both’ the lender and the loan sides; this can be considered as the degree of matching at an individual feature level.

These matching features, which are originally the second-order terms of existing features, may be inherently utilized in nonlinear or kernel models, but they are potentially critical information to many other models such as linear models and other tree-based models that deal with only one variable at a time, as will be described in Section 4.1.

**Temporal features.** Inspired from the analysis discussed earlier in Section 1, we generate additional features using temporal information about a lender and a loan. Available temporal information includes a lender’s *member\_since* and a loan’s *posted\_date*, *funded\_date*, and *paid\_date*. By considering the relative time differences between a loan  $l$  and the most recent loan,  $l_r$ , that a lender funded in the past, we construct six temporal features having the form of  $x - y$  where  $x$  is one of  $l$ ’s *posted\_date* and *funded\_date* and  $y$  is one of  $l_r$ ’s *posted\_date*, *funded\_date*, and *paid\_date*. These features basically reflect the temporal patterns of consecutive lending activities.

### 3.2 Feature Alignment via Joint Nonnegative Matrix Factorization

**Cold-start problem.** The feature generation procedure described previously is quite general and flexible when incorporating all the information from each of the heterogeneous entities, but the main limitation of this approach arises when little or no relationship link between a lender and/or a loan exists. Although details may differ, this problem, which is often referred to as a cold-start problem, is common in many recommendation applications. For instance, suppose a new Kiva user considers funding a loan for the first time and we would like to recommend the most appropriate loan they would be likely to fund. It is very likely that they may not have any connections with lender teams, previous loans, and

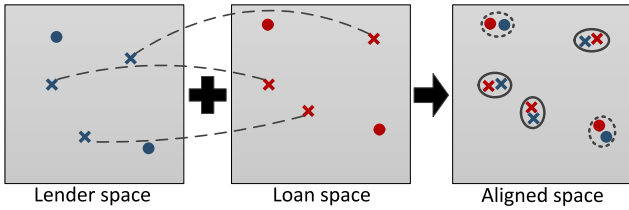


Figure 4: An overview of how joint NMF works. Given a high-dimensional space of lenders’ and loans’ textual data (‘ $\times$ ’-marked) along with their linked information (dashed lines), joint NMF generates a common aligned space where linked data points are closely placed. First-time lenders and fresh loans (‘ $\bullet$ ’-marked) are then mapped to the aligned space so that the resulting representations reveal their hidden linked relationships (dotted ellipses).

accordingly, any partners or borrowers. On the other hand, suppose a new loan webpage has just launched on the Kiva website and it currently has not secured a lender. In this scenario we would not have any available links to lenders and their lender teams that can be utilized in the feature generation process on the loan’s side. These cold-start problems make our loan recommendation task challenging since a number of feature blocks depicted in Fig. 3 would be zero vectors, leaving little information useful for recommendation.

**How joint NMF works.** As a way to alleviate this problem, we propose a novel feature generation approach based on joint nonnegative matrix factorization (NMF) for a first-time lender and a fresh loan that have no available link information. As shown in Fig. 4, the main idea behind this approach is to transform the features generated from heterogeneous sources, one of which comes from a lender’s side and the other from a loan’s side, into a common space where the vectors representing a lender and a loan with which it is linked can be placed close to each other. Once we obtain the vector representations of a lender and a loan in the resulting common space, one can also easily generate the corresponding lender-loan matching features which would play a significant role in estimating the likelihood of funding.

**Input matrices for joint NMF.** To begin, we start with textual fields, e.g., a lender’s *loan\_because*, a lender’s *occupational info*, which a lender fills out when signing up at Kiva.org, and a loan’s *loan\_description*. As described in Section (2), each of these textual fields is initially represented as a bag-of-words vector based on its own vocabulary. Note that the vocabulary set of a particular textual field is independent of that of any other, making each of them represented in a separate space.

Now, we form two term-document matrices  $A_u$  and  $A_l$  using the textual field from a lender and a loan, respectively. That is,  $A_u$  encodes either a lender’s *loan\_because* or *occupational info* while  $A_l$  encodes a loan’s *loan\_description*. Additionally, we assume the columns of  $A_u$  and  $A_l$  are aligned based on the linked relationships between lenders and loans. For example, the first column of  $A_u$  and that of  $A_l$  represent a lender and a loan, respectively, that have a link. Following this assumption, we exclude those lenders and loans that have no links when forming  $A_u$  and  $A_l$ . When a particular loan, i.e., a column of  $A_l$ , has links to multiple lenders, we sum up the textual vectors of the corresponding lenders and put this single vector in the corresponding column of  $A_u$ . In this manner, we maintain a one-to-one mapping between the columns of  $A_u$  and  $A_l$ .

**Formulation.** Given the two matrices  $A_u \in \mathbb{R}_+^{m_u \times n}$  and  $A_l \in \mathbb{R}_+^{m_l \times n}$ , an integer  $k$ , and a parameter  $\alpha$ , joint NMF solves

$$\min_{W_u, H_u, W_l, H_l} \left\| A_u - W_u H_u^T \right\|_F^2 + \left\| A_l - W_l H_l^T \right\|_F^2 + \alpha \|H_u - H_l\|_F^2, \quad (2)$$

where  $W_u \in \mathbb{R}_+^{m_u \times k}$ ,  $H_u \in \mathbb{R}_+^{n \times k}$ ,  $W_l \in \mathbb{R}_+^{m_l \times k}$ ,  $H_l \in \mathbb{R}_+^{n \times k}$  are nonnegative factors. In the above equation, the first and the second term correspond to standard NMF formulations, but at the same time, the third term enforces  $H_u$  and  $H_l$  to be close to each other. As a result, the rows of  $H_u$  and  $H_l$  can be considered as new vector representations in a common  $k$ -dimensional space where the linked lender and loan vectors are closely placed.

**Joint-NMF features for a first-time lender and a fresh loan.** Up to now, we have computed joint NMF using the textual information of lenders and loans by using their linked relationships, leading to a common space where these relationships are revealed. However, we still need to represent a first-time lender and a fresh loan to a joint-NMF space. To achieve this task, we utilize the resulting factor matrices  $W_u$  and  $W_l$ , which provide a mapping for an arbitrary bag-of-words representation in an original space to the joint-NMF space. In detail, given  $a_u \in \mathbb{R}_+^{m_u \times 1}$  and  $a_l \in \mathbb{R}_+^{m_l \times 1}$  corresponding to a first-time lender and a fresh loan, respectively, we solve the following nonnegativity-constrained least squares problem,

$$\min_{h_u \geq 0} \|a_u - W_u h_u^T\|_2 \quad \text{and} \quad \min_{h_l \geq 0} \|a_l - W_l h_l^T\|_2 \quad (3)$$

where  $h_u^T \in \mathbb{R}_+^{k \times 1}$  and  $h_l^T \in \mathbb{R}_+^{k \times 1}$  are our new representations in the joint-NMF space, i.e., joint-NMF features.

These joint-NMF features mainly have two advantages. First, even though a first-time lender and a fresh loan have no explicit links, one can expect their joint-NMF features that generated in this way to better reveal their proximity owing to the learnt factor matrices  $W_u$  and  $W_l$ . Second, since they are considered to be in the common space, we can now generate their lender-loan matching features in a similar way presented in the previous subsection.

## 4. EXPERIMENTS AND FINDINGS

In this section, we present our experiments and analysis on two loan recommendation cases depending on whether a lender of interest has previous funding history.

### 4.1 Experimental Setup

**Learner.** Considering the heterogeneity of our data and the complexity of the problem, it is crucial to use the most suitable and powerful prediction model to date. To this end, we have chosen a gradient boosting tree (GBtree)<sup>8</sup> [17, 14]. A GBtree is an ensemble method where an individual learner is a decision tree [6].

The reason for choosing a GBtree for our problem is as follows: First of all, an ensemble method is known for its superior generalization capability for unseen data. More importantly, a decision tree, our base learner, uses one variable at each node when it is trained/constructed as well as when it

<sup>8</sup>The GBtree implementation we used is available at <https://sites.google.com/site/carlosbecker/resources/gradient-boosting-boosted-trees>

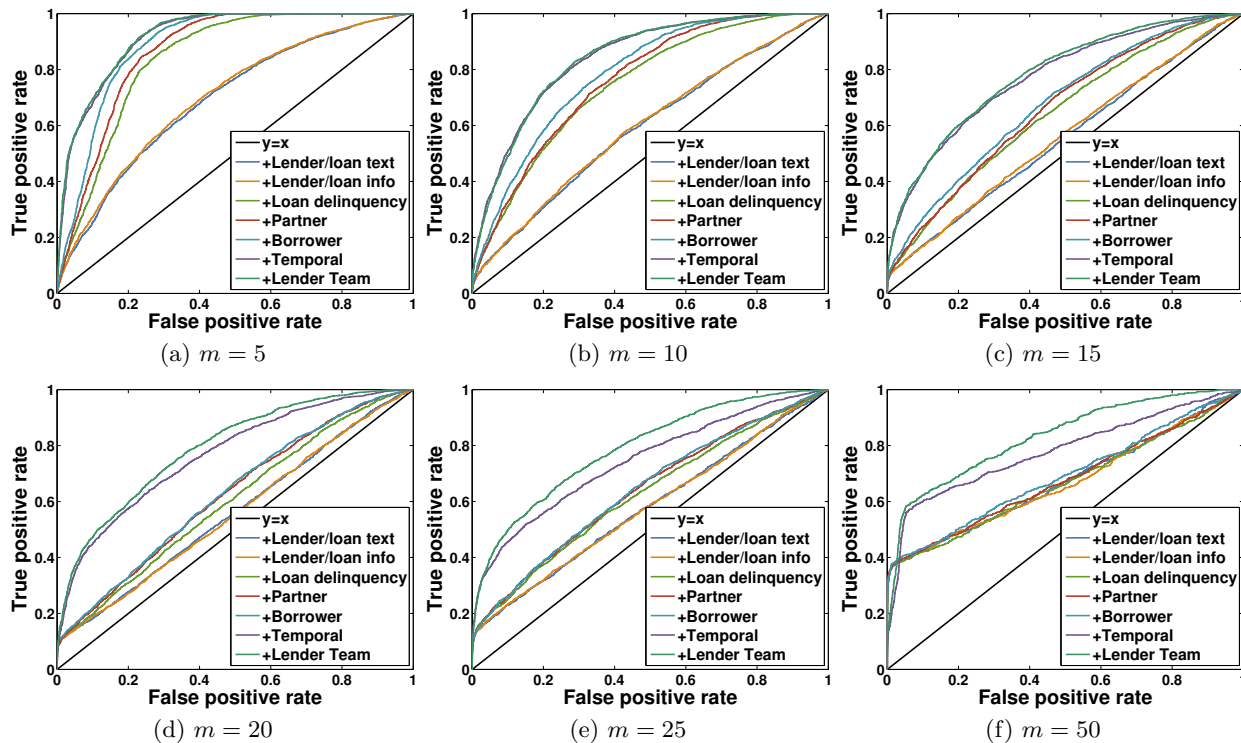


Figure 5: The ROC curve results for different lender groups with various numbers of previous loans  $m$ .

is applied to test data. This characteristic prevents us from worrying about heterogeneity in the features we generated. The downside to other learners, such as logistic regression and support vector machines, is that heterogeneous features have to be normalized via, say, standardization of their distributions, which transforms each feature to have zero mean and unit variance. Such normalization does not always make sense for binary and integer features, and furthermore it removes the nonnegativity of our feature representation that offers intuitive interpretation of them.

**Lender groups and data selection.** As previously highlighted, it is important to handle different user behaviors properly. Therefore, we first selected lenders that have a specific lending count  $m$ , where we varied  $m$  from 5 to 50, indicating the degree of how actively lenders participated in loans. Then, we conducted our experiments separately on each of these lender groups. We felt that lenders within this range of  $m$  contained the set of lenders not too active nor too passive, and thus we expect them to be more significantly influenced when given a recommendation for an appropriate loan.

Next, we formed a lender-by-loan adjacency matrix where only the components whose corresponding lenders funded the corresponding loans are set to 1 and 0 otherwise. From this graph, we randomly selected 5,000 positive (1-valued components) and 5,000 negative (0-valued components) samples and generated their feature vectors, as described in Section 3.<sup>9</sup> These samples are then used as our training and test sets under a 10-fold cross-validation setup.

<sup>9</sup>Note that we used balanced data sets in terms of positive vs. negative samples while original data are severely unbalanced. However, the ROC-based performance measure does not depend on the balancedness [12].

**Feature groups.** For lenders with funding history, we utilized various features presented in Section 3.1 and constructed several feature groups as follows:

- (1) Loan/lender text (600 dimensions): Textual features from a lender’s *loan\_because* and a loan’s *loan\_description*, whose dimension is reduced by NMF (Section 2).
- (2) Loan/lender info (183 dimensions): Features from a lender’s and a loan’s various fields.
- (3) Loan delinquency (13 dimensions): Features indicating how many previous loans for a lender have been non-paid or delinquent.
- (4) Partner (33 dimensions): Features about field partners.
- (5) Borrower (12 dimensions): Features about borrowers, e.g., a borrower’s *gender* and *pictured*.
- (6) Temporal (6 dimensions): Time differences between a new loan and a lender’s most recently funded loan (Section 3.1).
- (7) Lender team (15 dimensions): Features about lender teams a lender is associated with.

Using this structure, a lender-loan pair, which is our data item, is represented as an 862-dimensional vector.

For lenders without funding history, many of these features are not available. Thus, the two sets of joint-NMF features described in Section 3.2 were mainly used: those generated from aligning (1-a) a lender’s *loan\_because* versus a loan’s *loan\_description* (300-dimensional) and (1-b) a lender’s *occupational\_info* versus a loan’s *loan\_description* (300-dimensional), respectively. Next, we included (2) loan/lender info (61-dimensional), (3) partner (11-dimensional), and (4) borrower (4-dimensional) information.

**Performance measure.** Although our experimental setting is a binary classification, the desired capability from learning the function  $f(u, l)$  by a GBtree is to compute the likelihood of funding, which allows us to rank the most ap-

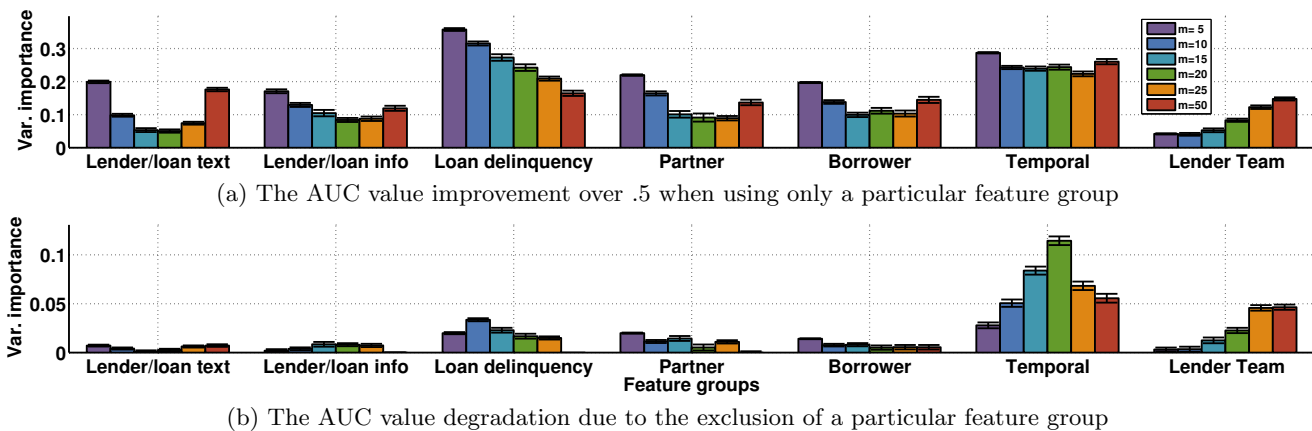


Figure 6: The analysis on the variable importance

Table 1: The cumulative AUC value in Fig. 5

	The number of previous loans $m$					
	5	10	15	20	25	50
Lender/loan text	.6938	.5930	.5524	.5594	.5788	.6730
Lender/loan info	.7010	.5974	.5601	.5572	.5793	.6679
Loan delinquency	.8416	.7453	.6438	.6000	.6265	.6691
Partner	.8646	.7610	.6600	.6222	.6391	.6778
Borrower	.8879	.7852	.6760	.6275	.6415	.6909
Temporal	.9179	.8415	.7736	.7675	.7449	.7802
Team	.9209	.8420	.7839	.7923	.7900	.8318

appropriate loans for a particular lender as well as the most appropriate lenders for a particular loan. Therefore, we are interested in the quality in terms of the resulting ranking of a given test set of lender-loan pairs, rather than the classification accuracy. In this respect, we report a receiver operating characteristic (ROC) curve and its area under the curve (AUC) value, which measures how much higher positive samples are ranked than negative samples.

## 4.2 Predictive Performance

### 4.2.1 Lenders with available funding history

**Overall performance.** In cases where previous funding information of a lender is available, we gradually incorporated additional features described in Section 4.1 for different lender groups. The performance results are shown by the ROC curves in Fig. 5 along with their AUC values summarized in Table 1. The best AUC values ranged from .78 to .92, which is a significant improvement over a baseline value of .5. These results were generally achieved only when using all the features available, indicating the advantage of our feature integration framework. Among different lender groups, lenders with  $15 \leq m \leq 25$  were the most difficult in predicting their likely loans to fund while lenders with a lower or higher  $m$  were relatively easier.

**Analysis on feature groups.** The analysis on the variable importance of each feature group, as shown in Fig. 6, reveals various interesting knowledge about micro-finance activities, as follows:

**(1) The relative time with respect to the last funded loan plays an important role.** Temporal features, which contain elapsed time information since the most recently funded loan, e.g., when it was posted and/or when it was re-

paid, consistently improve the performance by a non-trivial amount for all cases.

**(2) Loan delinquencies discourage passive lenders although they do not impact active lenders as much.**

The performance increase due to the loan delinquency features is substantial for lender groups with  $m \leq 15$ , but that increase drops significantly for lender groups with  $m = 50$ . Our further investigation showed these features were negatively correlated with the labels. For example, when  $m = 5$ , only 36% of the lenders who previously experienced loan delinquency had positive labels while 53% of the lenders without such experiences had positive labels. On the other hand, when  $m = 50$ , these two ratios were 49.7% and 50.1%, respectively, showing almost no correlation.

**(3) Lender teams exhibit greater influence on active lenders.** The performance due to the inclusion of lender team features improves as  $m$  increases. We conjecture that it is partly because passive lenders did not join teams yet. In fact, we found that the average number of teams of each lender with  $m = 50$  was .72 while that with  $m = 5$  was only .25. In addition, from Figs. 5(e)(f), these features pull up the ROC curve mainly at the false positive rate value (an  $x$  axis) from .4 to .7. This indicates that they are helpful in correctly classifying those somewhat ambiguous lender-loan pairs.

### 4.2.2 First-time lenders and fresh loans

**A baseline approach.** To evaluate the effectiveness of joint-NMF features, we designed a baseline approach to compare our method against, as follows. In the baseline approach, each pair of textual fields, (a lender’s *loan\_because*, a loan’s *loan\_description*) and (a lender’s *occupational\_info*, a loan’s *loan\_description*), has been aggregated into a single document corpus, which is encoded as a list of bag-of-words vectors based on a common vocabulary set. Next, we applied standard NMF in order to obtain their reduced-dimensional vectors. Note that, similar to the joint NMF approach, the resulting vector representations of lenders’ and loans’ textual data exist in a common space. Nonetheless, the main difference is that joint NMF utilizes additional link information and enforces linked lenders and loans to be close to each other in the common space (Fig. 4).

**Performance comparison.** For first-time lenders and fresh loans, Fig. 7 shows the comparisons in terms of AUC measures between the joint NMF and the baseline approaches. For each case, joint NMF was computed based on a differ-

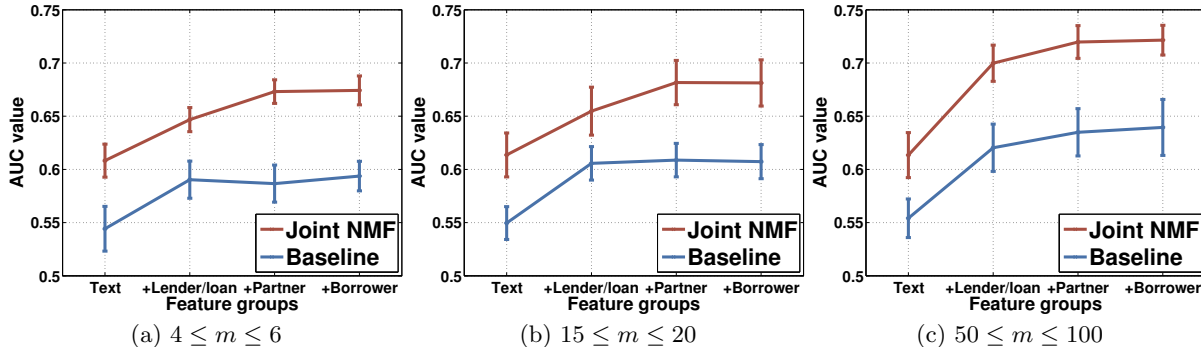


Figure 7: The AUC values for first-time lenders and fresh loans when training joint NMF with lenders with various numbers of previous loans  $m$  and their associated loans.

Table 2: The representative keywords of two topic pairs aligned by joint NMF

Topic 1	
a lender’s <i>occupational_info</i>	a loan’s <i>loan_description</i>
teacher, preschool, math, librarian, school	children, school, family, married, husband
Topic 2	
a lender’s <i>occupational_info</i>	a loan’s <i>loan_description</i>
student, mba, college, graduate, university	business, activities, entrepreneur, revenue

ent lender group and its associated loans depending on the value range of  $m$ . Note that all the training/test data in the supervised learning experiments have been selected only from first-time lenders and fresh loans and that loan delinquency and temporal features were excluded since they are not available for first-time lenders.

In all the results, the joint-NMF approach shows significant improvement over the base line approach, indicating that joint-NMF features are clearly helpful in revealing hidden links between first-time lender and fresh-loans. Combined with other features available, the best AUC result, which is about .72, was found when using the active lender group with  $50 \leq m \leq 100$ . This observation is somewhat counter-intuitive since first-time lenders would be expected to have similar behaviors to those of passive lenders with a smaller value of  $m$ . However, it can still be explained in a sense that active lenders likely provide detailed information about themselves in a lender’s textual fields, which would have provided joint NMF with vital clues in learning the mapping between lenders and loans.

**Aligned topics.** The qualitative analysis of the resulting mapping of joint NMF suggests in-depth understanding of lending behaviors. Table 2 shows the examples of aligned topics between a lender’s *occupational\_info* and a loan’s *loan\_description*. These representative keywords were obtained as the most highly weighted terms in the corresponding columns of the two matrices  $W_l$  and  $W_u$  in Eqs. (2) and (3).

Both topics in Table 2 are related to lenders with school-oriented occupations. Lenders in Topic 1 are shown to have professional jobs in a education environment, such as teachers and librarians, while those in Topic 2 mainly consist of students. By examining the associated topic keywords in a loan’ *loan\_description*, one can see that the former group tends to participate in family-related loans, e.g., helping chil-

dren go to school and supporting a family and/or a husband. On the contrary, the latter group (students) likes to lend to entrepreneurs with a particular business such as running a restaurant.

### 4.3 Further Discussions

Our analysis on loan recommendation and lending behaviors suggests several important directions that Kiva should take to promote micro-financial activities.

First, as seen from the significant importance of temporal features, performing loan recommendation at a right time would be crucial in keeping lenders actively involved. As shown in Fig. 2, Kiva can give recommendation (1) soon after a lender funded a loan as well as (2) when one’s previous loans have been paid back. Otherwise, people tend to gradually lose interest in micro-finance activities as time goes on.

Second, Kiva should help lenders, especially passive or novice lenders, avoid potentially risky loans. From our analysis, non-paid and/or delinquent loans seem to be the major cause for passive lenders to stop their lending activities, and thus it would be important to lead them to loans with a high chance of repayment.

Finally, in order to secure active lenders as much as possible, Kiva should encourage passive lenders to join teams since lender teams seem to be one of the driving factors for active lenders.

## 5. RELATED WORK

In this section, we mainly discuss related work about (1) recommender systems (relevant to Section 3.1), (2) manifold alignment (relevant to Section 3.2), and (3) analysis on micro-financial activities.

**Recommender systems.** Basically, a recommender system, an active information filtering system [5], aims at estimating the so-called utility function for a given item and a user, which is analogous to our funding likelihood function  $f(u, l)$ . A recommender system typically fall within two methods: *content-based methods* which match users to products by matching a user’s profile to the product’s characteristics and *collaborative filtering methods* which recommend products that other users with similar preferences have chosen in the past [26, 1]. Numerous studies on recommender systems have focused on collaborative filtering approaches. These methods are generally categorized according to whether they are memory-based and model-based.



For a comprehensive summary of collaborative filtering techniques, the reader is referred to the survey articles [28, 1].

Due to the discussed challenges in Section 1, which make collaborative filtering methods inapplicable, our work partly follows the content-based approach in that the proposed lender-specific features can be viewed as a user’s profile while the loan-specific features represent the product content. However, the typical content-based approach, mainly originating from information retrieval literature [4], focuses only on textual information. In order to integrate all the other information available, our approach extends it in the context of *ad-hoc* information retrieval [24], which throws various information as features and trains a learner model for predicting a relevance score of an item. These types of approaches are widely applicable in various novel applications including online dating systems [11].

**Manifold alignment.** This area has been actively studied recently in the context of image analysis [20] and cross-lingual information retrieval [29, 8, 9]. The problem setting is generally similar to that of our feature alignment where, given the different vector representations and/or relationships of the corresponding items, their new embedding in a common space is computed. Recently, from the perspective of multi-relational learning from multiple graphs or sources, several advanced methods based on joint matrix factorization have been proposed [33, 27]. In addition, a joint NMF-based approach has been proposed for multi-view clustering problems [22]. However, most of these methods focus on the best representations of existing data items while our proposed approach focuses on a generalization capability, i.e., embedding of unseen data into a common space so that their hidden correspondences are properly revealed.

**Analysis on micro-financial activities.** Previous work related to the complex micro-finance lending behavioral patterns and Kiva’s now-integral role in the crowd-sourced micro-financing movement have looked at the effects of the internet on micro-financing [7] and other peer-to-peer lending transactions [3]. Studies on micro-finance decision-making have discovered that lenders favor lending opportunities not only to entities similar to themselves but also to individuals in situations that trigger an emotional reaction [2, 15].

Kiva-related findings have suggested bias in lending decisions by showing that particular borrower features generate a higher level of attraction from the wider lending audience. In particular, women and more physically attractive individuals inherit a greater chance of securing charitable loan support, at least from lenders that constitute the set of first-time and lesser-active lenders [18]. Other studies on Kiva have observed the nature of lending behavior by correlating the impact of group dynamics to lending participation [16, 23]. These studies provide a basis for our work in which we extend similar decision-making processes through automation to support our lender-loan recommendation system.

All these studies have analyzed Kiva’s data in a number of ways, yet there is a lack of research that has utilized statistical numerical analysis approaches which are closer to our body of work. One such study manually defined a set of categories about the motivation of lending and applied machine learning techniques to train automatic text classifiers using a lender’s *loan\_because* field [23]. It also incorporated several simple features such as the loan count and team affiliations in performing regression on lending frequency and amount. This work revealed various interesting knowledge

about lending behavior, but the used information and techniques are relatively limited compared to our work.

To the best of our knowledge, *our work is the first in-depth study to directly tackle the loan recommendation problem by incorporating all the heterogeneous information available from Kiva*. As seen in Section 4, *we achieve performance viable for practical application and reveal significant finding about lending behavior that has not been discussed in any previous other work*.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel application of loan recommendation in the non-profit micro-finance sector. Starting with an extensive data set from Kiva, a famous micro-finance service, we tackled the problem using a supervised learning approach. In order to represent any given lender-loan pair as a feature vector, which is a key procedure in this approach, we proposed two main methodologies: (1) graph-based feature integration to flexibly incorporate all the heterogeneous information available and (2) feature alignment via joint NMF to enhance the limited information of first-time lenders and fresh loans. Based on the proposed approaches combined with a gradient boosting tree, a state-of-the-art prediction model, we achieved up to .92 AUC value. Furthermore, we presented interesting phenomena about micro-financing behaviors of Kiva lenders from temporal and social aspects.

The importance of our work and the information-rich nature of the Kiva data open up various future research possibilities. We describe a few of them in the following.

**Selecting negative instances.** Although we found our experiments showed consistent results over multiple runs of different sets of random samples, it would be beneficial to choose negative samples with more care. That is, not all negative examples are truly negative. For example, a lender may not have funded a particular loan simply because he did not know about it but not because he decided not to fund it. Advanced techniques such as the one-class type approach [25] and the one leveraging the context of user-system interactions [31], which tackle these issues in other recommendation applications, could be adopted in our work.

**Fraud detection.** As seen in Section 4, non-paid and delinquent loans significantly impact further lending activities of novice lenders, and thus, it is critical to detect potentially fraudulent loans and discourage lenders from lending them. A fraud loan detection problem can be formulated and solved in a similar way to the proposed methods in this paper. Eventually, integrating the resulting potential fraud score to our feature representation will increase the loan recommendation performance even further.

## 7. ACKNOWLEDGMENTS

This work was supported in part by NSF IIS-1116886, NSF CCF-0808863, NSFC 61129001, and DARPA XDATA grant FA8750-12-2-0309. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of funding agencies. We also thank anonymous reviewers for their insightful comments and suggestions.

## 8. REFERENCES

- [1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(6):734–749, 2005.
- [2] J. Andreoni. Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477, 1990.
- [3] A. Ashta and D. Assadi. Do social cause and social technology meet? impact of web 2.0 technologies on peer-to-peer lending transactions. *Cahiers du CEREN*, 29:177–192, 2009.
- [4] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. Addison-Wesley, 1999.
- [5] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12):29–38, 1992.
- [6] L. Breiman. *Classification and regression trees*. CRC press, 1993.
- [7] T. Bruett. Cows, kiva, and prosper. com: How disintermediation and the internet are changing microfinance. *Community Development Investment Review*, 3(2):44–50, 2007.
- [8] P. A. Chew, B. W. Bader, T. G. Kolda, and A. Abdelali. Cross-language information retrieval using parafac2. In *Proc. the 13th ACM international conference on Knowledge discovery and data mining (SIGKDD)*, pages 143–152, 2007.
- [9] J. Choo, S. Bohn, G. Nakamura, A. White, and H. Park. Heterogeneous data fusion via space alignment using nonmetric multidimensional scaling. In *Proc. the SIAM International Conference on Data Mining (SDM)*, pages 177–188, 2012.
- [10] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41:391–407, 1990.
- [11] F. Diaz, D. Metzler, and S. Amer-Yahia. Relevance and ranking in online dating systems. In *Proc. the 33rd international ACM conference on Research and development in information retrieval (SIGIR)*, pages 66–73, 2010.
- [12] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, 31:1–38, 2004.
- [13] M. Flannery. Kiva and the birth of person-to-person microfinance. *Innovations*, 2(1-2):31–56, 2007.
- [14] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.
- [15] J. Galak, D. Small, and A. T. Stephen. Micro-finance decision making: A field study of prosocial lending. *Journal of Marketing Research*, 48(SPL):S130–S137, 2011.
- [16] S. Hartley. Kiva. org: Crowd-sourced microfinance & cooperation in group lending. 2010.
- [17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
- [18] C. Jenq, J. Pan, and W. Theseira. What do donors discriminate on? evidence from kiva.org. 2012.
- [19] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.
- [20] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28:1784–1797, 2006.
- [21] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
- [22] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. the SIAM International Conference on Data Mining (SDM)*, pages 252–260, 2013.
- [23] Y. Liu, R. Chen, Y. Chen, Q. Mei, and S. Salib. I loan because...: Understanding motivations for pro-social lending. In *Proc. the 5th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 503–512, 2012.
- [24] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.
- [25] R. Pan and M. Scholz. Mind the gaps: weighting the unknown in large-scale one-class collaborative filtering. In *Proc. the 15th ACM international conference on Knowledge discovery and data mining (SIGKDD)*, pages 667–676, 2009.
- [26] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *Proc. the 2nd ACM conference on Electronic commerce*, pages 158–167, 2000.
- [27] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proc. the 14th ACM international conference on Knowledge discovery and data mining (SIGKDD)*, pages 650–658, 2008.
- [28] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4:2–4:2, 2009.
- [29] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *Proc. the 25th International Conference on Machine Learning (ICML)*, pages 1120–1127, 2008.
- [30] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. the 26th international ACM conference on Research and development in information retrieval (SIGIR)*, pages 267–273, 2003.
- [31] S.-H. Yang, B. Long, A. J. Smola, H. Zha, and Z. Zheng. Collaborative competitive filtering: learning recommender using context of user choice. In *Proc. the 34th international ACM conference on Research and development in Information Retrieval (SIGIR)*, pages 295–304, 2011.
- [32] M. Yunus. *Banker to the Poor*. Penguin Books India, 1998.
- [33] D. Zhou, S. Zhu, K. Yu, X. Song, B. L. Tseng, H. Zha, and C. L. Giles. Learning multiple graphs for document recommendations. In *Proc. the 17th international conference on World Wide Web (WWW)*, pages 141–150, 2008.