

# 浙 江 大 学

## 本 科 生 毕 业 论 文 开 题 报 告



学生姓名: 朱梦莹

学生学号: 3100102796

指导教师: 郑小林

年级与专业: 计算机科学与技术

所在学院: 计算机学院

一、题目：    P2P 借贷平台中的信用评估模型    

## 二、指导教师对开题报告、外文翻译和文献综述的具体要求:

1 文献综述要求查阅互联网金融与大数据分析方面国内外文献 10 篇以上，其中外文文献不少于 5 篇。在掌握文献内容基础上对现有研究工作进行分类，分析各类工作的优缺点。要求文献综述字数在 3000 字以上。

2 外文翻译要求选择和 P2P 借贷的信用评估相关, 且较经典的外文文献。翻译过程中对出现的专业词汇要逐一掌握, 理解文中介绍系统、技术和方法等的具体原理。要求翻译后的文档语言简练、通顺, 描述准确。要求外文翻译字数在 3000 字以上。

3 开题报告要求从现有工作的不足或具体应用需求中导出毕业论文的立题依据，确定 P2P 借贷的信用评估方面的研究方案，提炼可能的研究内容，大致确定毕业论文各组成部分、相应关键技术及技术路线，制定毕业论文时间安排。要求开题报告字数在 3500 字以上。

指导教师（签名）\_\_\_\_\_

年 月 日

毕业论文开题报告、外文翻译和文献综述考核

导师对开题报告、外文翻译和文献综述评语及成绩评定：

成绩比例	开题报告 占（20%）	外文翻译 占（10%）	文献综述 占（10%）
分 值			

导师签名\_\_\_\_\_

年 月 日

答辩小组对开题报告、外文翻译和文献综述评语及成绩评定：

成绩比例	开题报告 占（20%）	外文翻译 占（10%）	文献综述 占（10%）
分 值			

开题报告答辩小组负责人（签名）\_\_\_\_\_

年 月 日

## 目 录

本科毕业论文开题报告 .....	1
1. 课题背景 .....	1
2. 目标和任务 .....	2
3. 可行性分析 .....	3
4. 研究方案和关键技术考虑 .....	5
5. 预期研究结果 .....	11
6. 进度计划 .....	12
本科毕业论文文献综述 .....	13
本科毕业论文外文翻译 .....	32

## 本科毕业论文开题报告

### 1. 课题背景

互联网金融是现今又一热点话题，在 2013 年 10 月 30 日互联网金融全球峰会 IFC1000 在北京召开，主题为“大金融 大数据 大战略”。并且当今互联网金融的主要热点是电子支付、互联网借贷和网络理财，在这些热点话题的背后都少不了数据的支撑。大数据是互联网金融中不可缺少的一个核心组成部分。因为金融没有类似实物的物理生产、仓储、物流等过程，但其本身是数据的生产、仓储、挖掘、传输、分析和集成。所以大数据对于金融而言，相比其他行业，无疑是有更巨大的影响力。

小额借贷为中小企业的发展提供了重要资金支持。P2P 借贷 (People-to-People lending, 也成为 Peer-to-Peer lending, P2P lending) 是一种以互联网为载体的民间借贷形式，借贷双方通过没有任何金融机构中介的网络平台进行无担保的借贷行为。它是一个兼具传统金融与互联网金融特点的新方式。作为民间借贷的一种变体，P2P 借贷有助于小型企业和个人获得资金。

2005 年以来，在线 P2P 借贷在许多国家，包括美国、加拿大、英国、日本、意大利和中国，以不同的形式经历了快速增长时期。一些在线 P2P 贷款平台是出于慈善目的，旨在收集和提供资金给在贫困中的人，而另一些则有商业目的，有意给借款人和贷款人提供便利。最成功的在线平台是英国的 Zopa，美国的 Prosper 和 Kiva。例如，在 2006 年成立的 Prosper 在 2009 年成功完成 1.7 亿美元的贷款 (Lin et al., 2012)。到 2011 年 4 月为止，总部设在旧金山的非盈利组织 Kiva 通过其平台获得的贷款的总额已经达到 2.05 亿美元 (Kiva, 2011)。

尽管起步较晚，在线 P2P 借贷在中国也有相当大的发展。例如拍拍贷 (PPDai.com)、宜信 (CreditEasy.com)、齐放 (Qifang.com) 等网站也具有一定的影响力。国内最大的 P2P 借贷网站拍拍贷成立于 2007 年，在一年半的时间里就积累了超过 8 万名用户，使得这种金融投资新式迅速在国内发展。在三年内，CreditEasy.com 已经在北京和其他 15 个城市从成千上万的私人投资者中吸收近 1 亿美元，变成全国性的 P2P 借贷平台。

然而 P2P 借贷在蓬勃发展的同时，也相应地产生了许多问题。例如借贷资金的安全仍会存在隐患，借款人的信息可信度如何保证，一笔贷款产生后缺乏行之有效的对于还款的监管手段。这其中，如何评估 P2P 借贷过程中的个人信用，建立一个良好有效的 P2P 借贷信用体系至关重要。

P2P 借贷的安全问题，本质上就是借贷过程中的信息不对称问题。由于多数 P2P 借贷都是信用贷款，所以具有更大的风险，而网络交易本身就存在的匿名性和虚拟性使得这一问题更加严重。借款人的机会主义行为所带来的信息不对称和不信任使得借款人和贷款人存在匹配效率低下的问题。因此，大多数借款人只有一次申请贷款，然后就退出了 P2P 借贷平台(Collier & Hampshire, 2010)。研究表明，不止技术因素，而且心理因素也可能影响借款人和贷款人的行为。因此，分析和识别成功先例的借贷行为对在线 P2P 贷款的健康发展是有意义的。

基于上述背景，我将从信用角度来分析这些 P2P 借贷网站产生的真实数据，找到影响借贷的一些关键因素，分析这些因素如何对借款人和还款人产生影响。最终，希望能得出一个合理的信用评估模型。

## 2. 目标和任务

通过前期文献调研，我们发现，虽然在线 P2P 贷款在美国和中国有不同的操作模式，但“硬”和“软”信用信息在这两个国家都可能影响贷款的结果。

在线 P2P 借贷仍在起步阶段，这个领域的学术研究是相当有限的。在信号理论和社会资本理论的基础上，目前的研究已经从信息不对称的角度审视了借贷行为，但仅限于从 Prosper.com 获得的数据的分析。因此，目前的研究成果应该被推广到其他平台。此外，由于研究结果存在差异，它有利于从不同的角度研究整个国家的现象。

鉴于当前研究的局限性，需要进一步的理解在线 P2P 贷款的信用特点，我准备在未来对以下三个方向做出研究：(1) 对比中美不同 P2P 借贷平台的信用模型；(2) 分析社会资本在 P2P 借贷中起的作用；(3) 基于 P2P 借贷平台产生的数据构建信用模型。

### 1) 对比中美不同 P2P 借贷平台的信用模型

P2P 借贷平台的成功很大程度上取决于其创新的商业模式。深入检查不同的在线 P2P 借贷平台的商业模式不仅有助于我们更好地理解在线借贷的本质，而且还提供了

洞察这些平台的改进和新业务模式的设计的机会。

## 2) 分析社会资本在 P2P 借贷中起的作用

在文献阅读的过程中,可以发现社会资本在网络 P2P 借贷的过程中起着复杂而重要的作用,由于这些文献得出的结论不甚相同,所以我希望通过自己的研究在这方面获得一些新的发现和进展。

## 3) 基于 P2P 借贷平台产生的数据构建信用模型

通过分析不同 P2P 借贷平台的数据,我们可以获得很多有趣的结论。而基于这些结论,我将尝试着以一些现有的信用模型为基础,构建出一个 P2P 借贷的信用模型。在这其中,我们将会加入社会资本的因素,如果有可能,也会加入非结构化信息。在这个前提背景之下,本项目大致有这么几项工作:

- 1) **获取原始数据。**这些数据主要是基于 Prosper 和拍拍贷网站产生的用户借贷数据,采取的手段主要是下载数据和抓取数据。
- 2) **数据预处理。**由于获取的原始数据结构复杂,噪音严重,我们将进行去噪音化,并且提取关键项的数据,并输入到数据库中进行保存。
- 3) **确定聚类个数。**无论是 K-Means 聚类还是高斯混合聚类,都需要确定聚类的 K 值,我们将采用 a cluster validation method 来自动估计聚类的个数。
- 4) **高斯混合聚类。**得到 K 值后,我们将进行结合 K-Means 和高斯混合聚类,对数据进行聚类,并且找到影响 P2P 借贷的关键因素。
- 5) **可视化分析。**对聚类结果我们将进行可视化分析。
- 6) **构建信用模型。**基于现有的互联网信用模型,我们将尝试构建一种新的在线 P2P 借贷的信用模型。
- 7) **模型对比与验证**

以上每一步都会用到一些算法和技术,本项目将研究将这些算法和技术整合到 P2P 借贷信用模型中来的可行方法,希望能够通过研究和实践提出能够构建 P2P 借贷信用模型的理论方法。

## 3. 可行性分析

在线 P2P 借贷平台的迅速兴起于发展。这就为本项目的这项工作带来了广泛的数

据来源。并且随着经济的发展，人们对于基于社会资本例如社交网络资源的信用模型将会越来越依赖，这种基于在线社交的 P2P 小额贷款将会成为与传统银行借贷并驾齐驱的金融投资方式。但是，如何评估 P2P 借贷过程中的个人信用，对建立一个良好有效的 P2P 借贷信用体系至关重要，这也关系到在线 P2P 借贷的安全问题。因此，如果可以基于现有的在线 P2P 借贷平台的用户数据，构建一个有效的 P2P 信用模型，无疑是非常必要的。下面，我从几个方面来进行这个系统的可行性分析：

- 1) **环境可行性。**随着时代和技术的进步，社交网络与金融借贷相结合已经成了必要的趋势。比如 Prosper、拍拍贷等中外在线 P2P 借贷平台都拥有自己的社交网络，而大型社交网络例如 facebook、微信等也在推行自己的 P2P 借贷应用。因此，外部的环境完全可以为本项目的系统提供保障。
- 2) **经济可行性。**我们获取 P2P 借贷平台上的数据仅用于研究，而非出于商业目的。在研究的过程之中，本项目的系统大部分是用软件来实现的，也不需要投入大量资金用于购买昂贵的外围设备。因此，在经济上，这个研究是完全可行的。
- 3) **政策可行性。**最近互联网金融与大数据分析已经成了一个研究热点，而在线 P2P 借贷平台在国家政策的指导下，迅速兴起，成为新型创新互联网金融企业，这也促使国家金融投资市场的方式呈现多样化趋势。
- 4) **技术可行性。**国内外 P2P 借贷平台在信用体系上都有成熟的技术，例如 Prosper 通过关联第三方信用机构给出自己的信用评级，并且在借贷 listing 中提供了借款人的其他个人信息包括金融、社会信息等；而拍拍贷将借款人的历史信用情况以及个人信息综合得出一个自己平台的信用分数。本项目的系统基于的技术都是成熟可用的。比如在获取数据来源的过程中使用了爬虫与 XML 格式解析的手段；在分析数据时才有高斯混合聚类的算法。而在构建信用模型时采用 Canonical Discriminant Analysis 分析“硬”和“软”信用信息的几种因素的权重比例。这些都是可以实现的算法。

综上所述，本项目在环境、经济、政策以及技术方面都是可行的。有了这四个方面的支撑，本项目的实现也不会有太多的问题。



## 4. 研究方案和关键技术考虑

### 研究方案

主要研究方案是基于中美两个 P2P Lending 网站 Prosper 和拍拍贷的数据，对其进行数据挖掘，得到关于信用评估的一些结论，最后建立一个信用评估模型，具体目标已经在上文中阐述。

在具体的研究方法方面，首先要查找相关的文献资料，了解在线 P2P 借贷的研究进展和现有的互联网信用模型。这些资料不仅可以帮助了解整个信用模型的组成部分和构成，更能够从中吸取前人的经验教训。然后会选取若干文献，关于要做的项目所要用到的技术都要重点关注。这里最主要使用到的技术是用聚类的方法获得主要影响 P2P 借贷的因素以及根据 CDA 的方法构建信用模型。

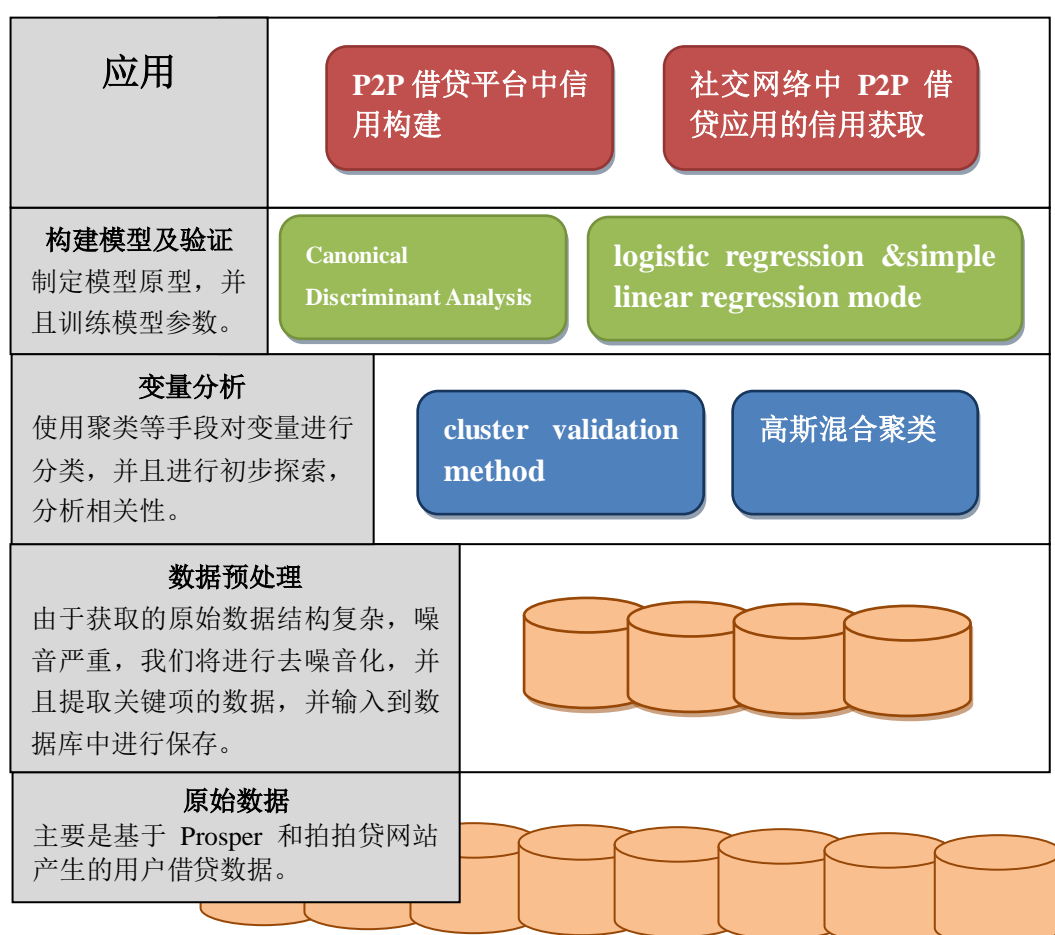


图 1：系统框架图

整个方案的难点也就集中在这两块。所以首先会查找阅读关于这两方面的文

献资料。其中关于获得主要影响 P2P 借贷的因素会使用高斯混合聚类方法。具体算法将在关键技术中给出。

另一块就是根据 CDA 的方法构建信用模型，具体算法也将在关键技术中给出。最后，我们来得到“硬”和“软”信用信息的几种因素的权重比例。然后使用 logistic regression 和 simple linear regression mode 来得到获得贷款时“硬”和“软”信用信息的几种因素的权重，误差，t 值，p 值。

## 关键技术

### 1) 聚类个数估计 The Clustering Number Discrimination Method

对于数据聚类，无论是 Kmeans 还是 Gaussian Mixture Models 首先需要知道聚类的个数。诚然，对于少量数据，可以根据数据分散的结果大致估计聚类的个数，但是在对大量数据聚类的时候，我们就需要一个可以自动估计聚类个数的方法。

这里，我们使用 a cluster validation method 来自动估计聚类的个数。Cluster validation (或者 stability based approach) 是一种基本方法来解决 model order identification (或者 cluster number estimation) 问题(Lange, et al., 2002; Levine and Domany, 2001)。该方法的前提是，如果 model order 是相同的真值，然后从数据中估计的聚类结构是稳定的重采样；否则，它是更可能是采样数据的 artifact。

表 1: Sense number estimation procedure for word sense discrimination.

1	Set lower bound $Kmin$ and upper bound $Kmax$ for sense number $k$ ;
2	Set $k = Kmin$ ;
3	Conduct the cluster validation process presented in Table 2 to evaluate the merit of $k$ ;
4	Record $k$ and the value of $Mk$ ;
5	Set $k = k + 1$ . If $k \leq Kmax$ , go to step 3, otherwise go to step 6;
6	Choose the value $\hat{k}$ that maximizes $Mk$ , where $\hat{k}$ is the estimated sense number.

表 1 给出了 the cluster number estimation 方法的程序流程。在本文中，我们设置  $Kmin$  为 2,  $Kmax$  为 10。The evaluation function  $Mk$  (在表 3 中给出) 将会得到最佳的聚类个数  $k$ 。在本文中， $q$  被设置为 20。聚类的方法将采用 Kmeans。

(Levine and Domany, 2001)给出了 Table 2 中的函数  $M(C^\mu, C)$ :

$$M(C^\mu, C) = \frac{\sum_{i,j} 1\{C_{i,j}^\mu = C_{i,j} = 1, d_i \in D^\mu, d_j \in D^\mu\}}{\sum_{i,j} 1\{C_{i,j} = 1, d_i \in D^\mu, d_j \in D^\mu\}}$$

其中 $D^\mu$ 是所有数据中采样 size 为 $\alpha|D|$ 的子集， $C^\mu$ 和 $C$ 是对 $D$ 和 $D^\mu$ 使用聚类算法得到的 $|D| \times |D|$ 关联矩阵， $0 \leq \alpha \leq 1$ （在本文中被设置为 0.9）。关联矩阵 $C$ 被定义为：当 $d_i$ 和 $d_j$ 属于同一个聚类时， $C_{i,j} = 1$ ；否则 $C_{i,j} = 0$ 。 $C^\mu$ 的定义也是类似的。

表 2: The cluster validation method for evaluation of values of cluster number  $k$ .

<b>Function: Cluster Validation(<math>k, D, q</math>)</b>	
<b>Input: cluster number <math>k</math>, data set <math>D</math>, and sampling frequency <math>q</math>;</b>	
<b>Output: the score of the merit of <math>k</math>;</b>	
Perform clustering analysis using Kmeans on data set $D$ with $k$ as input;	
Construct connectivity matrix $C_k$ based on above clustering solution on $D$ ;	
Use a random predictor $\rho_k$ to assign uniformly drawn labels to instances in $D$ ;	
Construct connectivity matrix $C_{\rho_k}$ using above clustering solution on $D$ ;	
For $\mu = 1$ to $q$ do	
1	Randomly sample a subset ( $D^\mu$ ) with size $\alpha D $ from $D$ , $0 \leq \alpha \leq 1$ ;
2	Perform clustering analysis using Kmeans on ( $D^\mu$ ) with $k$ as input;
3	Construct connectivity matrix $C_k^\mu$ using above clustering solution on ( $D^\mu$ );
4	Use $\rho_k$ to assign uniformly drawn labels to instances in ( $D^\mu$ );
5	Construct connectivity matrix $C_{\rho_k}^\mu$ using above clustering solution on ( $D^\mu$ );
Endfor	
Evaluate the merit of $k$ using following objective function:	
$M_k = \frac{1}{q} \sum_{\mu} M(C_k^\mu, C_k) - \frac{1}{q} \sum_{\mu} M(C_{\rho_k}^\mu, C_{\rho_k})$	
where $M(C^\mu, C)$ is given by equation (1);	
Return $Mk$ ;	

$M(C^\mu, C)$ 度量了数据对在 $D$ 和 $D^\mu$ 两种数据集中被划分到同一聚类的可能性。很显然， $0 \leq M \leq 1$ 。直观地说，如果 $k$ 是一个最佳的聚类个数，聚类结果在采样的数据集 $D^\mu$ 和全部数据集 $D$ 应该是最相似的，也就是说 $M$ 的值会上升。

在我们的算法中，我们使用表 3 的 6 中的等式来正则化 $M_k$ 。之所以正则化 $M_k$ ，

是因为当  $k$  增加的时候,  $M_k$  是递减的。因此, 可以避免被选择的聚类个数  $k$  偏小。

## 2) 混合高斯模型 Gaussian Mixture Models

每个 Gaussian Mixture Model 由  $K$  个 Gaussian 分布组成, 每个 Gaussian 称为一个 “Component”, 这些 Component 线性加成在一起就组成了 GMM 的概率密度函数:

$$p(x) = \sum_{k=1}^K p(k)p(x|k) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

现在假设我们有  $N$  个数据点, 并假设它们服从某个分布 (记作  $p(x)$ ), 现在要确定里面的一些参数的值, 例如, 在 GMM 中, 我们就需要确定  $\pi_k$ 、 $\mu_k$  和  $\Sigma_k$  这些参数。我们的想法是, 找到这样一组参数, 它所确定的概率分布生成这些给定的数据点的概率最大, 而这个概率实际上就等于  $\prod_{i=1}^N p(x_i)$ , 我们把这个乘积称作似然函数 (Likelihood Function)。通常单个点的概率都很小, 许多很小的数字相乘起来在计算机里很容易造成浮点数下溢, 因此我们通常会对其取对数, 把乘积变成加和  $\sum_{i=1}^N \log p(x_i)$ , 得到 log-likelihood function。接下来我们只要将这个函数最大化 (通常的做法是求导并令导数等于零, 然后解方程), 亦即找到这样一组参数值, 它让似然函数取得最大值, 我们就认为这是最合适的参数, 这样就完成了参数估计的过程。

- 
- 1 估计数据由每个 Component 生成的概率: 对于每个数据  $x_i$  来说, 它由第  $k$  个 Component 生成的概率为

$$\gamma(i, k) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

由于式子里的  $\mu_k$  和  $\Sigma_k$  也是需要我们估计的值, 我们采用迭代法, 在计算  $\gamma(i, k)$  的时候我们假定  $\mu_k$  和  $\Sigma_k$  均已知, 我们将取上一次迭代所得的值 (或者初始值)。

- 2 估计每个 Component 的参数: 现在我们假设上一步中得到的  $\gamma(i, k)$  就是正确的 “数据  $x_i$  由 Component  $k$  生成的概率”, 亦可以当做该 Component 在生成这个数据上所做的贡献, 或者说, 我们可以看作  $x_i$  这个值其中有  $\gamma(i, k)x_i$  这部分是由 Component  $k$  所生成的。集中考虑所有的数据点, 现在实际上可以看

作 *Component* 生成了  $\gamma(1,k)x_1, \dots, \gamma(N,k)x_N$  这些点。由于每个 *Component* 都是一个标准的 Gaussian 分布，可以很容易分布求出最大似然所对应的参数值：

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^N \gamma(i,k)x_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^N \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T$$

其中  $N_k = \sum_{i=1}^N \gamma(i,k)$ ，并且  $\pi_k$  也顺理成章地可以估计为  $N_k/N$ 。

### 3 重复迭代前面两步，直到似然函数的值收敛为止。

表 3: GMM 算法流程

我们可以看到 GMM 和 K-means 的迭代求解法其实非常相似，因此也有和 K-means 同样的问题——并不能保证总是能取到全局最优，如果运气比较差，取到不好的初始值，就有可能得到很差的结果。对于 K-means 的情况，我们通常是重复一定次数然后取最好的结果，不过 GMM 每一次迭代的计算量比 K-means 要大许多，我们的做法是先用 K-means（已经重复并取最优值了）得到一个粗略的结果，然后将其作为初值（只要将 K-means 所得的 centroids 传入 GMM 函数即可），再用 GMM 进行细致迭代。

在表 6-13 中，每一列都代表着一个高斯分布。每一个聚类的几率在第一行，其余的行为每个高斯分布的特征向量。每个特征向量中的值都被颜色标记来表示对借贷者的吸引程度。其中，绿色表示具有吸引力，红色表示借贷者投资存在风险。颜色的深度是根据图表中的最大值和最小值。最后在表的下方会有符号  $\checkmark$  和 X，其中  $\checkmark$  表示成功的信用需求，X 表示失败的信用需求。在 prosper 数据中，我们认为 funded 的比例超过 80% 为成功；同样在拍拍贷数据中，我们认为 FundingPrb 的比例超过 80% 为成功。

### 3) Canonical Discriminant Analysis

CDA 的方法被用来确定信用模型中“硬”和“软”信用信息的几种因素的权重比例。

典型判别分析的基本思想类似于主成分分析，通过数据的降维技术，找到能区分各类别的变量的线性组合。典型判别分析（Fisher 判别分析）方法的本质即

为确定该线性判别函数。该判别函数为如下的线性函数：

$$f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \cdots + u_p X_{pkm}$$

其中， $f_{km}$  是第  $k$  组第  $m$  个样例的 canonical discriminant 函数；变量  $X_{ikm}$  为第  $k$  组样例  $m$  的识别变量  $X_i$ ， $u_i$  为待求解的判别函数的系数，其具体的计算主要基于以下的原则：同一类中的变量的差异最小，而不同类中变量的差异最大。

由于 CDA 和 Canonical Correlations 相似，所以会产生多个 discriminant functions。第一个函数是以最大限度的组间差异建立的；当样本均值向量不位于同一条直线上，而是位于  $p$  维样本空间内的一个平面上时，我们需要在建构第二个 canonical function，这个函数与第一个 canonical function 正交，也是以最大限度的组间差异建立的。构建的 canonical discriminant function 越多，Canonical Discriminant analysis 的优越性就越下降。

为了得到 canonical discriminant 函数，首先我们需要构建一组 Sums of Squares and Cross Products (SSCP) 矩阵。

- A total covariance matrix

$$t_{ij} = \sum_{k=1}^g \sum_{m=1}^{n_k} (X_{ikm} - X_{i..})(X_{jkm} - X_{j..})$$

其中  $g$  表示组群数量， $n_k$  表示第  $k$  组中样例的数量， $X_{ikm}$  表示第  $k$  组第  $m$  个样例中第  $i$  个变量的值， $X_{i..}$  表示所有样例变量  $i$  的平均值。

- A within group covariance matrix

$$w_{ij} = \sum_{k=1}^g \sum_{m=1}^{n_k} (X_{ikm} - X_{ik.})(X_{jkm} - X_{jk.})$$

其中  $g$  表示组群数量， $n_k$  表示第  $k$  组中样例的数量， $X_{ikm}$  表示第  $k$  组第  $m$  个样例中第  $i$  个变量的值， $X_{ik.}$  表示第  $k$  组样例变量  $i$  的平均值。

- A between group covariance matrix

$$\mathbf{B} = \mathbf{T} - \mathbf{W}$$

当我们已知  $\mathbf{W}$  和  $\mathbf{T}$  矩阵，我们可以通过上述式子得出  $\mathbf{B}$  矩阵。

一旦我们拥有  $\mathbf{B}$  和  $\mathbf{W}$  矩阵，我们可以通过满足以下公式得出结果向量 ( $\mathbf{v}_i$ ):

$$\sum b_{1i} v_i = \lambda \sum w_{1i} v_i$$

$$\sum b_{2i}v_i = \lambda \sum w_{2i}v_i$$

.

.

.

$$\sum b_{pi}v_i = \lambda \sum w_{pi}v_i$$

其中 $v_i$ 的平方和为 1。

第二步，在确定 $\lambda$ 和 $v_i$ 后，就可以得出判别函数的权重系数：

$$u_i = v_i \sqrt{n - g}$$

$$u_0 = - \sum_{i=1}^p u_i X_{i..}$$

## 5. 预期研究结果

本项目准备根据以上提出的技术手段，在以下四点研究方向做出研究：

### 1) 对比中美不同 P2P 借贷平台的信用模型

P2P 借贷平台的成功很大程度上取决于其创新的商业模式。深入检查不同的在线 P2P 借贷平台的商业模式不仅有助于我们更好地理解在线借贷的本质，而且还提供了洞察这些平台的改进和新业务模式的设计的机会。

### 2) 分析社会资本在 P2P 借贷中起的作用

在文献阅读的过程中，可以发现社会资本在网络 P2P 借贷的过程中起着复杂而重要的作用，由于这些文献得出的结论不甚相同，所以我希望通过自己的研究在这方面获得一些新的发现和进展。

### 3) 基于 P2P 借贷平台产生的数据构建信用模型

通过分析不同 P2P 借贷平台的数据，我们可以获得很多有趣的结论。而基于这些结论，我将尝试着以一些现有的信用模型为基础，构建出一个 P2P 借贷的信用模型。在这其中，我们将会加入社会资本的因素，如果有可能，也会加入非结构化信息。

### 4) 模型的对比与验证

在构建信用模型后，我们将使用这个模型，对现有数据进行模拟验证。并且与现有的信用模型进行比较。

## 6. 进度计划

根据研究方法及论文大纲，将论文的进度安排如下：

时间	进度安排
11月-2月	查找和阅读相关文献及资料，了解课题背景，实现部分关键算法。
3月3日-3月15日	详细阅读相关文献，确定研究方案和关键技术。
3月16日-3月23日	撰写和修改开题报告、文献综述，完成外文翻译。
3月24日-4月5日	预处理 P2P Lending 网站相关数据，分析数据构成。
4月6日-4月20日	实现论文中需要的关键算法，并对数据进行分析，得出结论。
4月21日-5月1日	确定论文初稿。
5月2日-5月15日	进行论文修改，并确定终稿。



## 本科毕业论文文献综述

### 前言

信息技术(IT)的进步导致了电子市场的快速增长(Malone et al., 1987)[39]。电子市场最令人印象深刻的特性之一就是它能够消除或减少对传统中间商角色的依赖,使得产品/服务提供商和最终客户可以直接联系(Patsuris,1998)[41]。尤其这样使得在线点对点(P2P)贷款成为互联网用户一种新型融资模式。

P2P 借贷是指无担保贷款,借贷双方通过没有任何金融机构中介的网络平台进行借贷行为 ( Lin et al., 2009a[34]; Collier & Hampshire, 2010[13]; Bachmann et al., 2011[4])。

作为应用信息技术在金融领域和 Web 2.0 的革命(Iyer et al., 2009[27]; Lin et al., 2009b[35]), P2P 借贷能有效地促进信息发布和搜索,并提供所有必要的功能来完成交易(Brown, 2008[10]; Herzenstein et al., 2008[25])。

这种创新的借贷模型拥有以下优点:(1)借款人和贷款人可以轻松在网络平台上发布、搜索信息,并且以较低的交易成本完成交易(Lin, 2009[36]; Lin et al., 2009a[34]; Lin et al., 2009b[35]) ; (2)低交易成本使得非常小的贷款(例如小额贷款)变得可行;(3)多笔小额贷款可能汇集在一起,组成一个需要大量资金的大型的基金项目,以此规避风险;(4)通过查看网上认证信息和在社交网络上搜索信息,贷款人可以收集更多关于借款人的信用记录的信息,来减轻借款人和贷款人之间的信息不对称,减少贷款风险,并使得贷款范围超出了传统的熟人圈子。

网络 P2P 借贷主要针对小额贷款,这不仅为小企业或个人提供创始资金,还为他们提供短期流动资金(Johnson et al., 2010[28]; Wang et al., 2009[46])。尽管 P2P 借贷平台扮演着金融机构的角色,把借款人和贷款人联系起来,但它的利润不是从借款和贷款中获得,而是在交易过程中抽取一定的佣金(Lin, 2009[36])。这个有效的可持续的贷款方式——在线 P2P 贷款已经收到来自学术界和企业越来越多的关注(Brown, 2008[10]; Galloway, 2009[18]; Lin, 2009[36]; Bachmann et al., 2011[4])。

自 2005 年以来,在线 P2P 借贷在许多国家,包括美国、加拿大、英国、日本、意大利和中国,以不同的形式经历了快速增长时期。一些在线 P2P 贷款平台是出于慈

善目的，旨在收集和提供资金给在贫困中的人，而另一些则有商业目的，有意给借款人和贷款人提供便利。最成功的在线平台是英国的 Zopa，美国的 Prosper 和 Kiva。例如，在 2006 年成立的 Prosper 在 2009 年成功完成 1.7 亿美元的贷款(Lin et al., 2009b[35])。到 2011 年 4 月为止，总部设在旧金山的非盈利组织 Kiva 通过其平台获得的贷款的总额已经达到 2.05 亿美元 (Kiva,2011)[30]。

尽管起步较晚，在线 P2P 借贷在中国也有相当大的发展。例如拍拍贷(PPDai.com)、宜信(CreditEasy.com)、齐放(Qifang.com)等网站也具有一定的影响力。国内最大的 P2P 借贷网站拍拍贷成立于 2007 年，在一年半的时间里就积累了超过 8 万名用户，使得这种金融投资新式迅速在国内发展。在三年内，CreditEasy.com 已经在北京和其他 15 个城市从成千上万的私人投资者中吸收近 1 亿美元，变成全国性的 P2P 借贷平台。

尽管 P2P 借贷市场快速增长，但它仍处于婴儿阶段。事实上，只有少数在线平台能够在激烈的竞争中生存和发展(Lin, 2009[36])。即使是 Prosper.com，也只有少于 10% 的借款人能够成功地获得贷款，而许多贷款人并无法发现潜在借款人。

也有人认为借款人的机会主义行为所带来的信息不对称和不信任使得借款人和贷款人存在匹配效率低下的问题。因此,大多数借款人只有一次申请贷款，然后就退出了 P2P 借贷平台(Collier & Hampshire,2010[13])。研究表明，不止技术因素，而且心理因素也可能影响借款人和贷款人的行为。因此，分析和识别成功先例的借贷行为对在线 P2P 贷款的健康发展是有意义的。

为了促进在线 P2P 借贷市场的健康发展，从业人员和研究人员都需要解决以下问题：(1)在线 P2P 贷款成功的关键因素是什么?(2)使用或放弃使用在线 P2P 贷款的决定因素是什么?(3)具体是什么原因使得贷款人不打算兑现借款人的贷款请求?(4)什么可能导致不良贷款，从而对借款人的信用产生负面影响?(5) 基于社交网络提供的信用信息，我们如何评估贷款的风险?

然而，这些问题并没得到足够的反思。为了解决这些问题，我们研究了世界上两个最大的经济体：美国和中国的在线 P2P 借贷的表现。两国的在线借贷都经过了快速发展，但前者代表了发达国家，而后者代表发展中国家。本文的其余部分构成如下：首先介绍了 P2P 借贷的定义和模型，影响 P2P 借贷的因素，然后提出我的结论与讨论未来的研究方向。

# 1 P2P 借贷市场

个人贷款的概念并不是一个新的商业模式，而是一种传统的方式即在没有任何中介的情况下以个人名义借钱的人(Everett, 2008[15]; Herrero- Lopez, 2009[24])。是什么使转移到互联网平台的在线 P2P 借款成为一个新现象。

## 1.1 P2P 借贷平台

在 2005 年第一个借贷平台 Zopa 成立于欧洲(英国)。此后，各种形式的借贷平台开始形成。Garman et al. (2008)[20]阐述了在世界范围内现有的 24 个借贷平台，仅在美国就有 12 个，在 P2P-Banking.com 的博客中称 2010 年全世界 33 个不同的借贷平台。

在美国第一个借贷平台于 2006 年 2 月成立(prosper.com)。Smava(smava.de)，德国第一个 P2P 借贷公司，成立于 2007 年 2 月。如今大多数现有的借贷平台工作在国家层面上，因为不同国家有不同的法律要求(Berger & Gleisner, 2009[7])。下面的表显示了现有的主要借贷平台：

在线 P2P 借贷平台存在不同类型。他们基本上可以分为两种类型：商业和非商业(Ashta & Assadi, 2009)[3]。而商业平台一般仅限于国内市场，非商业性平台通常在全球范围内运作。两种平台之间的主要区别是贷款人的总体意图和他的期望回报。贷方使用商业平台希望在一个合理的风险下获得利益。在非商业平台贷款他们愿意承担的风险但并不一定要获得利益。这里贷款人更希望“捐赠”小额贷款给世界上经济欠发达的地区。

## 1.2 借款人与贷款人

在线 P2P 借贷是一个双面的市场，与传统的银行系统并没有太多区别(Klaft, 2008)[31]。贷款人和借款人是平台的所有活动的主要目标群体。

因此大部分的研究都集中在这些利益相关者和影响贷款成功的决定因素上(Freedman & G.Z. Jin, 2008[17]; Iyer et al., 2009[27])。贷款人在给定的风险水平下尽可能寻求有利可图的投资机，借款人在一定的违约风险下寻找不同的资金来源。P2P 网站作为中介机构，将这些人组织在一起。他们试图匹配双方的期望。借款人和贷款人有时参与一些可以提现他们共同利益的组织和社区(M. E. Greiner & Wang, 2009[23]; Herrero-Lopez, 2009[24])。

### 1.3 监管机构、合作银行、信用机构

作为一个(小)金融市场的一部分,P2P 借贷也要受到不同国家的不同的监管限制。根据国家规定,有现有的银行合作伙伴是主要要求。几篇文章提到了银行介入的必要性(Galloway,2009[18]),但这主要是为了促进借贷的过程。这个过程还包括确认借款人信用评级的信用部门或其他外部监测机构的参与。这些系统的确认和识别也因国家而异,导致全球研究在这一领域不适用。

### 1.4 借贷流程

一些平台直接连接贷款人和借款人,而其他平台通过第三方(通常是银行)连接。在线 P2P 借贷平台在借款人的利率设置方式上略有不同。例如 prosper.com 使用拍卖竞价(Galloway,2009[18]),借款人可以设定一个他们愿意支付的最高利率。在有限的时间内(prosper 上是 14 天)贷款人可以不断提交他们愿意出的金额和他们接受的最低利率。甚至当贷款金额满足借款人所需金额后,贷款人仍然可以提交 bid 来降低其他贷款人提供的最低利率,增加自己愿意出的贷款金额。在这种情况下,bids 的总金额已经比需要的贷款金额更多,那么那些给出最低利率的 bid 将会被采纳。然后所有借款人会收到此时的最高的利率和它的贷款金额,即使最低利率已经变小。

其他网站,就像德国的 smava.de 平台,根据借款人的特点(金融和人口特征)计算贷款请求的利率。竞标过程结束后,贷款已经被产生,所以进一步的报价不会影响产生的利率(Collier & R. Hampshire, 2010[13])。

如果贷款过程产生一个成功的贷款资助,一些平台像 prosper.com 实现了借款人的支付能力的核查,包括稳定收入的确认,然后才授予借款人的贷款,并且最终启动还款过程(S. Garman, R. Hampshire, et al., 2008[20])。

在线 P2P 借贷平台的收入是通过服务费获得的,他们从借款人和贷款人处获得(Klafft,2008)[31]。大多平台会收取一定比例的贷款资金来让借款人关闭交易,以及后期违约的费用。借款人通常需要支付基于放贷金额一定比例的服务费。

## 2 影响 P2P 借贷的因素

在传统借贷的背景下,金融机构,如商业银行,充当了交易中介的角色。这些银行以较低的利率吸收存款,然后以更高的利率向客户发放贷款。由于银行使用了复杂的风险评估机制,并且了解更多的借款人信息,他们可以在贷款过程中更有效地缓解

信息不对称。相比之下，在网络 P2P 借贷的环境下，贷款人很难对借款人获得全面的信息，导致信息不对称的问题十分严重(Lin et al., 2009a[34])。因此，大多数研究在线 P2P 借贷都集中在以缓解借款人和贷款人之间的信息不对称，从而在贷款过程中减少风险为目的，包括(1)在线 P2P 贷款的经营模式；(2) “硬信用信息”对贷款结果的影响，如个人信息；(3) “软信用信息”对贷款结果的影响

## 2.1 影响贷款的信贷“硬信用信息”

“硬信用信息”指的是可以准确量化、容易存储、可以有效传播的信用信息。在 P2P 借贷中，信用信息包括借款人的信用背景，如借款人的负债收入比、信用评级、过去获得的信贷金额和借款人持有信用卡的数量(Lin, 2009[36]; Lin et al., 2009a[34]; Lin et al., 2009b[35])。

国外由于信用制度相对健全，信用信息容易获取，因此国外大多数 P2P 借贷平台都会要求借款人提供自己的财务情况，并且以此作为判断借款人信用的主要指标。典型的财务特征包括：信用评级，每月的详细收入和支出，房子所有权以及债务收入比等。这些信息往往由收集个人信息和财务数据的外部评级机构给出。还有一些平台如 Prosper 还提供借款人额外的财务信息如他们当前信用卡的额度或银行卡的利用率(Klaft, 2008[31])。

在线 P2P 借贷，因为贷款人无法获得有关借款人的详细信息，贷款人必须依赖于可用来判断借款人的信誉的信号，并相应地做出贷款决定。研究表明，存在两个在贷款的决策中发挥着举足轻重作用的特征：获得信号的成本和信号难度评估(Collier & Hampshire, 2010[13])。在 P2P 贷款中，借款人的个人信息和贷款 listing 上的信息被认为是评估借款人的可信度的重要的信号，用来评估借款人的违约风险和设置利率(Collier & Hampshire, 2010[13]; Lin, 2009[36])。

### 借款者的信用评级

Klaft(2008)研究表明，P2P 网络借贷的规则其实非常类似于传统的银行系统。他对 Prosper 平台上的数据进行分析，证明对借款利率影响最大的因素为借款人的信用评级，而借款人的债务收入比的影响虽然显著，但是影响却小得多。其他信息如经过核实的借款人的银行帐户或经过验证的借款人是否自有房产对借款利率几乎没有影响。然而令人吃惊的是，当研究的因变量变成借款是否成功时，借款人的银行帐户的存在与否却是决定借款能否成功的最重要因素，甚至连借款人的信用评级的影响力

也只能排到第二名，由于借款人的信用评级是一项更加复杂的，包含了银行账户信息的变量，以上结论非常难以解释。Klaft(2008)[31]同时也指出，信用评级较差的，在传统银行渠道无法贷到款的借款人，在 P2P 借贷平台上也仍然不太可能成功借到款。他的分析数据表明，借款人的信用评级为 HR 的列表占 Prosper 上所以借款列表的 57.4%，但是其中只有 5.5% 借款成功，而信用评级为 AA 的借款人的借款成功率却高达 54%[6]。

通过分析从 Prosper.com 收集的数据，Lin (2009)[36]发现信用评级较低的贷款请求不太可能被资助，并且这样的贷款请求更有可能违约或以很高的利率结束。更有趣的是，Lin 等人(2009b)[35]发现信用卡利用率曲线也贷款结果产生影响：当信用卡使用在中低水平的借款人的信用水平高，信用卡的高利用率会导致获得资金的概率降低，只能增加利率从而产生高、易受冲击的风险。

更深入的研究(Iyer et al., 2009[27])发现借款人的违约率、负债收入比和在过去六个月贷款请求的数量对贷款人的决定有显著负面影响。虽然在中国没有结论性的结果关于贷款信用评级对贷款结果的影响，但陈(2012)表面，信用评级在 Ppdai.com 对获得贷款的概率产生了部分影响，而利率是决定因素。然而，违约率较高的借款人信用水平要低得多。

### 借款人的财务信息

Freedman & Jin (2008)[16]在他们的研究中发现，Prosper 平台上的平均借款成功率从 2005 年 11 月-2006 年 6 月之间的 8.51% 上升到至 2007 年 3 月-2008 年 7 月期间的 10.14%。他们对此现象的解释是，借贷平台要求借款人更多的提供自己的财务信息，使得借款人借贷成功率得以上升（2007 年 2 月 12 日，Prosper 增加了更多的需要借款人填写的财务信息，如尽可能地要求借款人报告其当前的收入，就业状况和职业）。

### 借贷信息

研究表明，贷款的成功率与利率负相关。在实践过程中，借款人必须权衡这两个因素之间的关系。此外，贷款的规模与降低成功率和提高利率有关；因此，借款人可以通过支付更高的利率和 / 或减少贷款规模增加贷款的成功率 (Collier & Hampshire, 2010[13])。贷款请求可能会在一个固定格式的表格中被列出，一旦总 Bids 金额达到借款人所期望的总额，借款人就可以结束这笔交易或者继续开放窗口使得贷

款人可以继续降低利率，直到在指定时间即使完全达到资助的要求。

研究表明，交易的进度会影响贷款人的判断：快要达到贷款目标的借贷有更多机会获得资助，而不是更高的利率。然而，Lin 等人(2009b)[35]和 Puro 等人(2010)[44]指出交易进度在对违约率的影响上并没有显著差异。此外，贷款的目的也会对贷款人的决定产生影响：商业贷款比债务合并贷款获得成功的几率高，并且可以得到更高的利率(Wang et al., 2009[46])。Collier and Hampshire (2010[13])发现，贷款规模、借款人的财务状况(例如负债收入比)和交易进度都对利率产生影响。

研究也显示，贷款人会使用一些主观的、标准化的结论获得借款人的信用情况。例如，愿意支付最高利率的借款人贷款人是一种潜在的积极信号(Iyer et al., 2009)[27]。在中国的借贷网站，信息不对称被发现可以用社会特征和朋友间的信任来缓和，这是对贷款意愿至关重要的(Chen et al., 2012[12])。

### 人口特征因素

研究表明，借款人的人口特征因素，如种族、性别和年龄，可能对贷款人的贷款意愿产生影响(Ashta & Assadi, 2009[3]; Berger & Gleisner, 2007[6]; Kumar, 2007[33])。例如 Ravina (2007) [45]对出借人和借款人之间相似性如何影响借贷行为进行了分析。结果显示双方之间的相似性对促成借贷行为的发生具有很强的正影响作用。与借款人在同一个城市，属于同一种族或者仅仅是性别相同，都会增加潜在出借人借出资金的可能性。

Pope & Sydnor (2008)[42]表明，非裔美国人借贷成功的可能性要比那些具有相近信用评级的白人低 25%到 34%，同时非裔美国人贷款的利率比白人贷款的贷款利率高 0.6%和 0.8%。然而就预期回报率而言，非洲裔美国人的贷款要明显差于白人的贷款的。因此，非裔美国人较高的借款年利率并不足以弥补其更高的违约概率。Herzenstein 等 (2008) [25]也证实，非裔美国人确实比其他种族的人得到资助的几率更小。Ravina (2007) [45]认为，种族歧视主要体现在被歧视种族的借款人必须支付更高的借款利率才能获得贷款。根据她的研究，在都得到借款的情况下，非裔美国人要比类似的白人借款人多付 1.39-1.46 个百分点的利率[8]。

Pope & Sydnor (2008)[42]分析了影响借款人的年龄对于借款成功率的影响，文章得出的结论是：同 35-60 岁的群体相比，35 岁以下的人借款成功的可能性要比其高 0.4-0.9 个百分点；而 60 岁及以上的人要比 35-60 岁群体借款成功的可能性低 1.1

和 2.3 个百分点[9]。

Pope & Sydnor (2008)[42]的实证研究发现，单身女性比类似条件的男性要少支付 0.4 个百分点的利率，尽管就贷款的预期收益率而言，单身女性的比男性少约 2 个百分点[9]。Barasinska(2009)[5]研究了出借人的性别是否与预期收益率和贷款的风险相关。令她吃惊的是，她发现，女性出借人比男性出借人的风险规避意识更差。

女性出借人的更多的选择了利率较低和信用评级较低的借款人进行投标。Barasinska 认为这些结果可能的解释是，女性出借人在借出资金时，比男性出借人更容易被利他动机和同情心理驱动，而把钱借给愿意承受利率较低的借款人。

总之，在这一领域的研究仍然有限。目前的研究主要是使用从 Prosper.com 收集的数据集(如 Lin et al., 2009a[34])，但这很难将他们的结论普遍化。不同的借贷平台可能采用不同的信贷“硬信用信息”，以及“硬信用信息”对贷款结果的影响需要进一步研究在其他网站的上下文中被研究。

## 2.2 社会特征和“软信用信息”

与“硬信用信息”如信用评分或借款人的财务状况相比，“软信用信息”是指借款人的模糊的无法量化的信息。在 P2P 借贷中，软信用信息可能来自借款人的社交网络，例如网络上的“朋友”，网络上的“团体”，借款列表中添加的照片等(Collier & Hampshire, 2010[13]; Iyer et al., 2009[27]; Krumme & Herrero, 2009[32]; Lopez, 2009[37])。

在线 P2P 借贷平台不仅披露了借款人的个人贷款信息，还提供借款人的社交信息。使用 Web 2.0 技术，涉及 P2P 贷款的贷款人可以很容易地从借款人的社交网络中获取软信用信息(Lin, 2009[36])。小额信贷理论表明，社交网络可以帮助减少贷款过程中的信息不对称，并且可以激励借款人偿还贷款(Krumme & Herrero 2009[32])。社交网络的作用也适用于在线 P2P 借贷的上下文(Lin et al., 2009b[35])。

社会资本对充分获得贷款起着积极的影响，可以减少借款人的可能获得的利率，并且对信用评级较低的借款人产生越来越大的影响(M. E. Greiner & Wang, 2009[23])。根据 Herrero-Lopez(2009) [24]的研究表明，当金融功能并不足以构建一个成功的贷款请求时，培养社交功能可以增加贷款的机会。

朋友



“朋友”实际上是通过 P2P 借贷平台联系的人们。他们代表一个一对一的链接从自身到其他借款人或贷款人。这种关系通常是基于家庭、友谊或先前的交易。这种联系公开激励属于借款人第二或更高维度的社交网络中的贷款人基于间接信任给予 bid(Herrero-Lopez,2009[24])。Freedman &Jin (2008)[17]发现有贷款推荐人或是 bid 是由借款人的朋友提交的,很少会拖欠还款,并且有更高的回报率。他们的结论是,借款人的朋友能更好地识别风险和拥有可信赖性,因为他们拥有更多的额外信息。而且他们认为社交网络的监控为还款提供了更强大的动力。

## 照片

Klaft(2008)[31]指出,利率在有没有照片的情况下几乎是相同的。他认为照片不是主要影响贷款成功的因素,但却扮演一个次要角色。

相反,Ravina(2007)[45]的研究表明,在相同情况下,美丽的借款人更容易获得贷款,并且比一个长相普通的借款人的利率少 81 个百分点。Pope &Sydnor(2008)[42]发现,Prosper 市场对没有照片或者照片中的人不开心的借款人的 listing 比较消极。

## 组织与社区

大部分 P2P-Lending 平台都允许用户形成特殊的社区。如果组织存在积极的向导和激励,那么组织可以清除一些信息障碍(Freedman &G.Z. Jin, 2008)[17]。例如成为一个被信任的组织的成员在 Prosper 可以获得更高概率的贷款请求(Herrero-Lopez,2009[24])。但加入一个可信组织并不能完全保证获得贷款,这仍然需要一个合理的报价。

Berger (2009)[7]和 Greiner 与 Wang (2009)[23]发现,仅仅加入一个组织就可以显著减少资金的贷款利率。Greiner &Wang (2009)[23]表明,借款人是一个组织的成员比不是具有略高的还款率和较低的违约率。Greiner 与 Wang (2009) [23], Herrero-Lopez (2009)[24] 和 Freedman 与 Jin (2008)[17]声称,拥有社交的贷款比没有朋友关系或不是组织成员的贷款更有可能获得资助。

组织领导的职责在论文中受到争议。尽管组织领导单独推荐的某笔借贷交易会增加它被资助的可能性(Kumar, 2007[33]),但它不会影响贷款的利率本身(Berger &Gleisner, 2009[7])。研究表明,只有积极竞标的组长和其他成员可以减少贷款的利率(Berger &Gleisner, 2009[7]; Collier & Hampshire, 2010[13])。Freedman 和 Jin (2008)[17]不同意这些研究结果,他们表明只有组长的推荐和组长的投标结合在一起才会增加的

贷款利率，这意味着一个组长的报价实际上是被视为一个负面的信号。他们提出的证据表明，一些组织产生的负面信号效应和组织领导者的报价是由于组长的反向激励奖励。直到 2007 年第 4 季度，Prosper 才开始支付每个在他们组织内部成功资助贷款的组织领导人 12 美元。在研究 2008 年或更新的数据时，这种消极的信号效应没有被提到了。Berger & Gleisner(2009) [7]; Kumar(2007)[33]的研究一致表明，组织领导人的报价与贷款的违约率是无关的。这表明他们的报价并不适合作为积极的信号效应诱导贷款人降低利率。惊人的是组织评分，这是 Prosper 基于组织过去的表现和平均违约率产生的，很少或没有影响利率(Berger & Gleisner, 2009[7]; Collier & Hampshire, 2010[13])。Berger & Gleisner (2009) [7]以及 Collier & Hampshire (2010) [13]表明，组织的规模增加会导致较低的利率。

然而 Freedman & Jin (2008) [17]的研究结果却与以上研究相矛盾。他们认为，组织规模越大，利率越高，贷款人的收益率越低。相同地他们是在 2007 年早些时候 Prosper 的情况下，可能导致这个结论的事实是组织领导人人为试图增加组织大小，而没有勤奋的筛查成员。根据 Freedman & Jin (2008)[17]和 Greiner & Wang (2009) [23]的文章，借款人和贷款人属于同一组织的比例对利率比组织规模有更大的影响。贷款人比例越高，利率越低。

如果在成为组员关系之前，审查借款人的个人信息是强制性，借款人的贷款利率在其他条件不变的情况下会变小(Berger & Gleisner, 2009[7]; Collier & Hampshire, 2010[13]; Greiner & Wang, 2009[23])。强制性的审查对信用等级较低的借款人产生最重要的影响，特别是 D,E 和 HR 等级的借款人(Berger & Gleisner, 2009[7])。这些借款人是最有可能在团体组织中的(Freedman & G.Z. Jin, 2008[17])，因为他们最大得益于这些组织产生的社会资本(Collier & Hampshire, 2010[13])。有趣的是，Klein (2008)[51]表明，组织领导人选择不选择那些拥有良好特征的借款人到他们组织。他解释说这种短期激励的非理性行为是由 2007 年年底前的 Prosper 诱导的。

## 社会特征对借贷的积极影响

### 减少信息不对称

信息不对称理论是指在市场经济活动中，各类人员对有关信息的了解是有差异的；掌握信息比较充分的人员，往往处于比较有利的地位，而信息贫乏的人员，则处于比较不利的地位。该理论认为：市场中卖方比买方更了解有关商品的各种信息；掌握更

多信息的一方可以通过向信息贫乏的一方传递可靠信息而在市场中获益；买卖双方中拥有信息较少的一方会努力从另一方获取信息。

社交网络的关系维度可以降低交易过程中的信息不对称。使用从 Prosper.com 收集的数据，Lin 等人(2009a[34];2009b[35];2009[36])研究了社交网络在提高贷款成功率和降低贷款利率中起的作用，他们发现社交网络可以有效地降低交易过程中的信息不对称。在线社交网络在减少信息不对称，提高信用评级方面起了重要作用(Everett, 2008[15])。研究进一步表明，成为一个受信任组织的成员能提高贷款的成功率，同时也能帮助信用评分较低的人们以负担得起的利率获得贷款资助(Lopez et al., 2009[37])。在信息技术的帮助下，社交网络可以向贷款人发送有价值的信号，但信号的有效性取决于他们的可靠性和可验证性(Krumme &Herrero 2009[32])。

### 降低借贷风险

社会资本一般指与社交网络相关联的资源。个体的社会资本可以从他的社交网络，包括朋友和同事(Burt, 1992[11])，或社交网络中的组员(Portes, 2000[43])来评估。在 P2P 借贷中，社会资本作为软信用信息的主要来源可能会影响贷款的成功率和利率。Wang 等人(2009)[46]发现，社会资本越多的借款人，拥有更大的机会可以获得他们的贷款资助，并且可以得到更低的利率。当借款人的信用评分较低时，贷款人需要更多的信息来进一步评估借款人的信誉以降低贷款风险。

在这种情况下，贷款人依赖于借款人的贷款的社交网络来获取信息产生自己的决策。Collier and Hampshire (2010) [13]研究了基于信号理论社交网络对贷款行为的影响，并发现通过与其他成员频繁交流，经常积极参与组织事物和为他人贷款担保，借款人可以向贷款人发出强烈的诚信信号。Lin 等人(2009b)[35]发现，社会资本产生“信息外部性”，可以用来促进在线交易，减少贷款的利率。关于 P2P 贷款在中国背景下的出版刊物是有限的。Chen 等人(2012)[12]发现，贷款结构和关系型社会资本都是有影响力的信任因素，这样的关系可以弥补信息不对称。

而建立一个自组织的团体可以帮助在线 P2P 借贷市场有效运作，实证研究揭示了这种混合结果：允许该组织领导人在完成贷款后获得奖励(如成功贷款费用)是有害的(Hildebrand et al., 2010[26])。Freedman 和 Jin (2008)[17]发现，组织贷款的投资回报率明显低于非组织贷款。当贷款需要朋友的推荐，贷款违约率相对较低，但利率可能会增加。

## 降低违约率

关于社会资本对贷款违约的影响,研究也提出了一个混合的描述(Karlan, 2007[29]; Krumme &Herrero, 2009[32])。朋友数量被发现不是预测违约的因素, Lin(2009) [36]分析了, 借款人社交网络的朋友中存在贷款人, 则平均违约率降低了 9%。如果借款人的朋友参与借贷, 则违约的可能性进一步降低; 尤其当一个朋友的报价成功, 会导致借款人在朋友的压力下偿还贷款。Berger 和 Gleisner(2007[6];2009[7])提供了进一步的证据支持社会资本对违约率的影响是基于组织成员。从信息不对称的角度, Berger 解释说, 组织领导人可能比一般贷款人有更多的关于借款人的私人信息, 以便组织领导人更能够选择正确的借款人和对借款人产生更大的压力迫使借款人偿还贷款。事实上, 组织领导人实际上是在交易过程中扮演着贷款中介的角色。

然而, 一些研究则反映了相反的结果对于社会资本的影响。Greiner and Wang (2009) [23]发现, 尽管贷款人在决策过程中将考虑社会资本, 但社会资本在减少借款人的违约率方面没有显著的影响, 他们解释说社会资本的力量在预测违约率很弱(只有 0.3%)。此外, 通过分析来自 Prosper.com 的贷款人 6 个月的二次借贷过程中贷款人、借款人和偿还贷款的数据, Kumar (2007)[33]表明, 信用评级和帐户验证度可以降低贷款违约的概率, 而贷款规模与违约率呈正相关。有趣的是, 影响利率和风险的某些因素, 如债务收入比、房屋所有权和组织领导人的认可, 证明对违约率无显著影响。

总之, 有在社交网络促进在线 P2P 贷款方面有着广泛的研究。然而, 目前的研究仍受到一些局限性: 1)目前的研究主要集中于从 Prosper.com 获得的数据, 而没有从贷款行为和理论方面研究; 2)当前研究主要关注商业平台如 Prosper.com, 没有考虑很多其他成功的 P2P 贷款在线平台, 他们可能采用不同的商业模式与工作机制; 3)由于各个国家社会和文化环境不同, 社交网络可以在不同的国家扮演不同的角色, 所以研究不同的借贷平台在不同的社会和文化环境中的作用也是必要的; 4)当前的研究只考虑一些来自社交网络容易量化的信息, 这并包括来自其他用户的评价和贷款的文本描述, 这是理解贷款行为的重要因素。

## 3 信用模型研究和应用现状

随着互联网的发展, 人们已经越来越认可互联网信任体系的重要性, 因此学术界和企业界都对互联网信任体系投入了大量的研究和实践工作。

### 3.1 信任的概念和特点

信用是建立在人与人信任关系的基础上的，所以我们有必要清楚信任的本质。信任是一个复杂的概念，它是由个人价值观、态度、心情及情绪、个人魅力交互作用的结果。虽然很多的学者认识到信任的重要性，并给出了他们对信任的定义，但是这些定义都是在他们的研究领域中提出来的，因此没有一个完整、准确、统一的定义。在研究信任的过程中往往会简化信任的概念，本文中用的信任采用 Gambetta(2000)[19]提出的信任定义：

信任是指个体 A 主观上认为另一个个体 B 会执行某一动作的可能性。在某种情况下，个体 A 愿意依赖他认为相对安全的事物，即使这个决定可能带来严重的后果。

信任和安全有很多的相似性，但是两者之间不能完全划等号。安全一般用于保护系统和数据不受怀有恶意的或没有经过授权的用户侵害，它的效果是使系统和数据看起来更加可靠。而信任是比安全更为复杂的概念，用户会在并非 100%安全的情况下采取适当的行为，也就是说用户必须在有风险的情况下依赖他人来作出决策。

根据信任的定义和 Abdul-Rahman(2000)[1]等人的思想，我们将信任的特点总结为以下几点：

- 信任具有主观性。不同的用户对同一个实体给出的信任评价可能不同，甚至有很大的差异，这是因为每个人对信用的判别标准往往会随着他们性格、个人经历的差异而各不相同。在不同环境下，用户对同一个信任值也有不同的处理结果。例如用户在借贷过程中，借贷金额越大时，用户对信任值的要求就越严格。
- 信任是动态可变的。用户之间的信任不是常量，而是随着用户行为和相互之间的交互活动动态改变的。用户的一个新行为可能会增加或降低其他用户对它的信任值。
- 信任关系是非对称的。A 信任 B，并不代表 B 信任 A，即使 A、B 相互信任，他们对对方的信任程度也不一定相同。
- 信任是有条件传递的。A 信任 B，B 信任 C，但是不能由此推断出 A 信任 C。这是因为在现实中，信任从来不是绝对的，而且信任在传递的过程中会不断减弱。信任传递链中还存在消极信任，它会在信任的传递过程中起反效应。例如，A 不信任 B，B 不信任 C，这这种情况下，A 信任 C 是合理的，因为 A

使用了“我的敌人的敌人是我的朋友”这一原则。而从 A 的角度来看, A 可能会认为 C 连 B 都不会信任他, 所以 A 也不信任 C。

- 信任是可以组合的。A 信任 B, B 信任 E, 同时 A 也信任 C, C 信任 D, D 信任 E, 那么 A 是否信任 E 呢? 在这种情况下, A 必须综合考虑这两条信任链来决定是否要信任 E。信任的可组合性是计算信任的又一个重要特点。
- 信任和上下文环境相关。人们只会在某一个或几个特定的领域信任别人。例如 A 信任 B 能归还商业贷款, 但是不信任 B 能偿还负债贷款, A 对 B 的信任域是“商业贷款”。因此任何抛开上下文环境研究信任的行为都是没有意义的。信任的这个特点对信任的传递也有一定的影响, 一条有效的信任传递路径的必要条件是信任链中所有的用户的信任域相同。
- 信任是可以量化的。尽管许多经济学家深信信任不是一个可计算概念, 并认为信任的非计算模型对研究这种关系具有重大的意义, 但是这样的模型应该是社会学和心理学的范畴。信任的量化模型很多, 最常见的模型包括用离散等级或 0 到 1 之间的实数表示信任。
- 信任具有模糊性。信任是一个模糊的概念, “信任”和“不信任”之间没有一个明确的界限。

### 3.2 基于信誉的信用模型

根据 Bonattiet al.(2007)[8]的分类, 目前存在基于策略和基于信誉两种确定信任的方式。

策略描述了获得信任的必要条件, 并且也可以指示出在满足一定条件下的行为和结果。基于策略的方法频繁涉及证书的交换和验证, 这些是由一个实体发布的信息(有时通过使用数字签名), 并可能说明另一实体的品质或特点。

信誉是根据与实体交往或者进行观察-包括与评估者直接接触(个人经验)、根据他人的报道(建议或第三方验证)的历史评估。这些历史评估的合并方式可以是多样的, 并且当我们使用第三方实体的信息时可能产生信任的传递问题。

基于信誉的信任研究首先起源于电子商务领域, 其中最经典的是 Amazon[2]和 eBay[14]网上的信任模型。一般情况下, 用户需要借助他人的经验对其他用户做出信任评估, 而获得其他信息的一种方式是通过咨询一个第三方权威代理, 该代理具有你所需要的交往信息, 这种系统成为集中式的信誉系统。但是在互联网中几乎是找不到这样

的第三方权威代理的。因此现在大多数的模型都是采用分布式的信誉管理方法，即用户不依赖于一个集中式的权威来得到其他用户的信誉，而是通过获取声望信息，由用户自己来进行信任评估。Yu 和 Singh(2003)[49]提出了一种分布式的解决方案，用户可以通过其他用户获得相应的信誉信息，结合这些信息来做出信任决策。

Wang(2003)[47]等人提出了基于贝叶斯网络进行信任计算的方法。贝叶斯网络的理论基础是贝叶斯规则，它利用统计学方法来描述不同属性之间概率大小的关系网络。在该模型中，信任根据信任对象的不同被划分为对自己的信任、对服务提供者的信任、对推荐者的信任和对所在团队的信任。

信任可以通过信任网络的连通关系进行计算，Golbeck(2004)[21]在他的 TrustMail 应用中描述了如何利用信任网络来计算用户之间的信任值。通过广度优先搜索的方式，从起始节点向目标节点逐步扩张开来，在搜索过程中，信任值低的用户提供的评价信息将会被忽略掉。最后通过平均所有信任评价值得到两个节点之间的信任情况。

在推荐系统中，往往也会进行信任的研究。Ziegler(2007)[50]等人提出利用信任值和用户相似度之间的关联可以加强推荐系统的有效性和准确性。基于同样的思路，John O'Donovan(2005)[40]提出了将信用模型和传统的协同过滤框架结合的几种方法，使得推荐预测具有更高的准确性。Hao Ma(2009)[38]等人则通过将社会网络信任模型引入了协同过滤算法中，从而解决了协同过滤算法中进行矩阵运算时，数据稀疏带来的问题。

## 4 结论和未来的研究方向

我们发现，虽然在线 P2P 贷款在美国和中国有不同的操作模式，但“硬”和“软”信用信息在这两个国家都可能影响贷款的结果。本文可以提高我们理解在线 P2P 借贷，并为进一步的研究提供方向。

在线 P2P 借贷仍在婴儿阶段，这个领域的学术研究是相当有限的。在信号理论和社会资本理论的基础上，目前的研究已经从信息不对称的角度审视了借贷行为，但仅限于从 Prosper.com 获得的数据的分析。因此，目前的研究成果应该被推广到其他平台。此外，由于研究结果存在差异，它有利于从不同的角度研究整个国家的现象。

鉴于当前研究的局限性，需要进一步的理解在线 P2P 贷款的信用特点，我准备在未来对以下三个方向做出研究：（1）对比中美不同 P2P 借贷平台的信用模型；（2）分析社会资本在 P2P 借贷中起的作用；（3）基于 P2P 借贷平台产生的数据构建信用模型。

## 对比中美不同 P2P 借贷平台的信用模型

P2P 借贷平台的成功很大程度上取决于其创新的商业模式。深入检查不同的在线 P2P 借贷平台的商业模式不仅有助于我们更好地理解在线借贷的本质，而且还提供了洞察这些平台的改进和新业务模式的设计的机会。

## 分析社会资本在 P2P 借贷中起的作用

在文献阅读的过程中，可以发现社会资本在网络 P2P 借贷的过程中起着复杂而重要的作用，由于这些文献得出的结论不甚相同，所以我希望通过自己的研究在这方面获得一些新的发现和进展。

## 基于 P2P 借贷平台产生的数据构建信用模型

通过分析不同 P2P 借贷平台的数据，我们可以获得很多有趣的结论。而基于这些结论，我将尝试着以一些现有的信用模型为基础，构建出一个 P2P 借贷的信用模型。在这其中，我们将会加入社会资本的因素，如果有可能，也会加入非结构化信息。

## 参考文献：

1. Abdul-Rahman, A., & Hailes, S. (2000, January). Supporting trust in virtual communities. In *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on* (pp. 9-pp). IEEE.
2. Amazon.com. <http://www.amazon.com>.
3. Ashta, A., & Assadi, D. (2009). An analysis of European online micro-lending websites. *Cahiers du CEREN*, 29, 147-160.
4. Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., ... & Funk, B. (2011). Online Peer-to-Peer Lending--A Literature Review. *Journal of Internet Banking & Commerce*, 16(2).
5. Barasinska, N. (2009). The role of gender in lending business: evidence from an online market for peer-to-peer lending. *The New York Times*, 217266, 1-25.
6. Berger, S. C., & Gleisner, F. (2007). Electronic marketplaces and intermediation: An empirical investigation of an online p2p lending marketplace.
7. Berger, S., & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *BuR Business Research Journal*, 2(1).
8. Bonatti, P., Duma, C., Olmedilla, D., & Shahmehri, N. (2007). An integration of reputation-based and policy-based trust management. *networks*, 2(14), 10.
9. Briceno Ortega, A. C., & Bell, F. (2008). Online social lending: Borrower-generated content.



10. Brown, C.M. (2008). Is peer-to-peer lending right for you? *Black Enterprise* (39:2), pp. 146-146.
11. Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard university press.
12. Chen, D.Y. (2012). Is online peer-to-peer lending market effective? A study on herding behavior in China, Working Paper (School of Management, Fuzhou University).
13. Collier, B. C., & Hampshire, R. (2010, February). Sending mixed signals: multilevel reputation effects in peer-to-peer lending markets. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 197-206). ACM.
14. eBay.com. <http://www.ebay.com>.
15. Everett, C. (2008). Group membership, relationship banking and loan default risk: the case of online social lending. *Relationship Banking and Loan Default Risk: The Case of Online Social Lending* (March 15, 2010).
16. Freedman, S., & Jin, G. (2008). *Dynamic learning and selection: the early years of prosper. com*. Working Paper.
17. Freedman, S., & Jin, G. Z. (2008). *Do social networks solve information problems for peer-to-peer lending? evidence from prosper. com* (No. 08-43). com .NET Institute Working Paper.
18. Galloway, I. (2009). *Peer-to-peer lending and community development finance* (No. 2009-06). Federal Reserve Bank of San Francisco.
19. Gambetta, D. (2000). Can we trust trust. *Trust: Making and breaking cooperative relations*, 213-237.
20. Garman, S. R., Hampshire, R. C., & Krishnan, R. (2008). Person-to-person lending: The pursuit of (more) competitive credit markets.
21. Golbeck, J., & Hendler, J. (2004). Accuracy of metrics for inferring trust and reputation in semantic web-based social networks. In *Engineering knowledge in the age of the semantic web* (pp. 116-131). Springer Berlin Heidelberg.
22. Gonzalez, L., Komarova, Y., Kabadayi, S., & Werner, F. (2012). When Can a Photo Increase Credit?: The Impact of Lender and Borrower Beauty, Gender and Age in Online Peer-to-Peer Loans.
23. Greiner, M. E., & Wang, H. (2009). The role of social capital in people-to-people lending marketplaces.
24. Herrero-Lopez, S. (2009, June). Social interactions in P2P lending. In *Proceedings of*

- the 3rd Workshop on Social Network Mining and Analysis* (p. 3). ACM.
25. Herzenstein, M., Andrews, R. L., & Dholakia, U. M. (2008). The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities. papers. ssrn. com.
  26. Hildebrand, T., Puri, M., & Rocholl, J. (2010). Skin in the game: The incentive structure in online social lending.
  27. Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2009). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending?.
  28. Johnson, S., Ashta, A., & Assadi, D. (2010). Online or Offline?: The Rise of “Peer-to-Peer” Lending in Microfinance. *Journal of Electronic Commerce in Organizations (JECO)*, 8(3), 26-37.
  29. Karlan, D. S. (2007). Social connections and group banking\*. *The Economic Journal*, 117(517), F52-F84.
  30. Kiva. (2011). Latest statistics. Available at <http://www.kiva.org/about/facts> (Accessed on 8/15/2011).
  31. Klafft, M. (2008, March). Peer to peer lending: auctioning microcredits over the internet. In *Proceedings of the International Conference on Information Systems, Technology and Management*, A. Agarwal, R. Khurana, eds., IMT, Dubai.
  32. Krumme, K. A., & Herrero, S. (2009, August). Lending behavior and community structure in an online peer-to-peer economic network. In *Computational Science and Engineering, 2009. CSE'09. International Conference on* (Vol. 4, pp. 613-618). IEEE.
  33. Kumar, S. (2007). Bank of one: empirical analysis of peer-to-peer financial marketplaces.
  34. Lin, M., Prabhala, N., & Viswanathan, S. (2009a). Social networks as signaling mechanisms: Evidence from online peer-to-peer lending. *WISE 2009*.
  35. Lin, M., Prabhala, N. R., & Viswanathan, S. (2009b). Judging borrowers by the company they keep: Social networks and adverse selection in online peer-to-peer lending. *SSRN eLibrary*.
  36. Lin, M. (2009). Peer-to-peer lending: An empirical study.
  37. Lopez, S. H., Pao, A. S. Y., & Bhattacharya, R. (2009). The effects of social interactions on P2P lending. *MAS Final Project*, 1-24.
  38. Ma, H., King, I., & Lyu, M. R. (2009, July). Learning to recommend with social trust

- p>ensemble. In
- Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*
- (pp. 203-210). ACM.
39. Malone, T. W., Yates, J., & Benjamin, R. I. (1987). Electronic markets and electronic hierarchies. *Communications of the ACM*, 30(6), 484-497.
  40. O'Donovan, J., & Smyth, B. (2005, January). Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces* (pp. 167-174). ACM.
  41. Patsuris, P. (1998). Cut out the middleman, *Forbes*
  42. Pope, D. G., & Sydnor, J. R. (2011). What's in a Picture? Evidence of Discrimination from Prosper. com. *Journal of Human Resources*, 46(1), 53-92.
  43. Portes, A. (2000). Social capital: Its origins and applications in modern sociology. LESSER, Eric L. *Knowledge and Social Capital*. Boston: Butterworth-Heinemann, 43-67.
  44. Puro, L., Teich, J. E., Wallenius, H., & Wallenius, J. (2010). Borrower decision aid for people-to-people lending. *Decision Support Systems*, 49(1), 52-60.
  45. Ravina, E. (2008, May). Beauty, personal characteristics and trust in credit markets. In *American Law & Economics Association Annual Meetings* (p. 67). bepress.
  46. Wang, H., Greiner, M., & Aronson, J. E. (2009). People-to-people lending: the emerging e-commerce transformation of a financial market. In *Value Creation in E-Business Management* (pp. 182-195). Springer Berlin Heidelberg.
  47. Wang, Y., & Vassileva, J. (2003, October). Bayesian network-based trust model. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on* (pp. 372-378). IEEE.
  48. Weiss, G. N., Pelger, K., & Horsch, A. (2010). *Mitigating adverse selection in P2P lending: empirical evidence from Prosper*.
  49. Yu, B., & Singh, M. P. (2003, July). Detecting deception in reputation management. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems* (pp. 73-80). ACM.
  50. Ziegler, C. N., & Golbeck, J. (2007). Investigating interactions of trust and interest similarity. *Decision Support Systems*, 43(2), 460-475.
  51. Klein, T. (2010). *Microfinance 2.0: Group Formation and Repayment Performance in Online Lending Platforms During the US Credit Crunch*. Ibidem-Verlag.

## 本科毕业论文外文翻译

### **Social Interactions in P2P Lending**

Sergio Herrero-Lopez

MIT

77 Massachusetts Avenue

Cambridge, MA 02139-4307

+1-617-417-6055

[sherrero@mit.edu](mailto:sherrero@mit.edu)

The 3rd SNA-KDD Workshop '09 ( SNA-KDD'09), June 28, 2009 , Paris,  
France .

### **在P2P借贷中的社交行为**

塞尔吉奥·雷罗·洛佩兹

第三届SNA-KDD研讨会'09（SNA-KDD'09），2009年6月28日，巴黎，法国。

## 中文翻译译稿

### 摘要

以信用借贷的形式获得资本需要银行有能够衡量偿还给定回报的风险的能力。在过去,借贷需要凭借知名的组织或需要所需的贷款抵押来担保。在现代,金融机构提供贷款给那些满足资格测试的个人。在孟加拉的格莱珉银行已经证明,小贫困社区受益于“小额贷款”这种金融创新形式,这使得一些非银行担保的企业家可以优先从事个体经营项目。在线 P2P 借贷被考虑作为一种小额贷款的进化,并且反映了网络社区的应用原则。互联网企业例如 Prosper.com, Zopa 或 Lendingclub.com,提供了一种借款人和贷款人交流的途径,并定义了社会群体中的一部分社交关系。

本文衡量了社交关系对借贷行为中产生的风险的影响力;并且特别关注了一对一和一对多关系的影响。结果表明,当金融特征不足以构建差异化成功的信贷时,培养社交功能可以增加贷款获得资助的机会。对于这个任务,基于模型的聚类方法被应用于由 Prosper.com 提供的实际的 P2P 贷款数据。

### 分类和主题描述符

H.2.8[数据库管理]:数据库应用程序——数据挖掘。

J.4[社会和行为科学]:社会学

### 一般术语

经济学, 人为因素

### 关键词

社交网络、集群、高斯混合

### 1.介绍

在 Prosper.com 上 P2P 贷款是基于一个在线的逆向拍卖;个人请求分为两种:一种是借钱,以借款人的角色,另一种是购买贷款,以贷款人的角色。借款人设置他们需要借的金额和他们可以支付的对于这一笔贷款的最高利率;而贷款人列出他们可以贷出的部分金额和他们希望收到的最低利率。P2P 借贷的用户可以成为借款人、贷款人或者两者。主要和传统银行产业的区别是贷款人不仅可以得到借款人的个人金融信息,也可以基于个人的社会特征评估借贷操作的风险。

对整个借贷行为的管理是由 P2P 借贷企业负责的。他们收集和显示用户行为所产生的 listings 和 bids，并且提供可以让用户建立联系和建立网络组织的社交网络引擎。

本文的目标不仅仅是揭示隐藏在 P2P 借贷行为中的用户行为模式，也证实或反驳了通常假定的一些显而易见的行为。本次工作集中使用了几种数据聚类技术来发现和分析底层 P2P 借贷数据的分布模式。

本文的构成如下：第二节总结了以前对 P2P 借贷的相关工作。第三节介绍了 Prosper 数据集的数据结构，并且描述了应用先前进行的信息分析的结果和数据预处理的情况。这一节也呈现了 P2P 借贷网站的基本社交特征。而关于 P2P 借贷的数据聚类技术将在第四节中详细描述。本研究的结论在第五节进行阐述。最后，第六节提出了未来工作计划，来进一步了解 P2P 借贷这一新兴金融服务。

## 2.相关工作

社会资本在 P2P 借贷的经济价值没有在学术研究中被广泛探讨。迄今为止,大多数分析都是在经济背景下完成的。

Knack 和 Keefer [1]提出目前的证据表明，社会资本确实是衡量经济表现的很重要的因素。

Hulme 和 Wright [2]从多个角度提供了对社会贷款主题广泛而深入的处理。

[3]研究了社会网络能否解决对等贷款信息不对称问题，揭示了有组织贷款的预计回报率比那些由于贷款人的学习和消除组长奖励的无组织贷款要低。

Herzenstein 和 Andrews[5]探究了 Prosper 上获得贷款的主要因素。作者发现借款人的金融实力和发布 listing 后的努力使相比人口因素更能决定贷款能否成功的主要原因。

Sydnor 和 Pope 的研究通过观察贷款人如何应对信号特征，如种族、年龄、性别等，提出了借贷过程中存在歧视。

J.Ryan, K.Reuk 和 C.Wang [7]利用真实数据分析了变量之间的回归模型和关联。他们研究的目的是衡量每个独立的金融和社会特征的相对关系，并确定在 listing 成功转换成一笔贷款的过程中它们的影响力。他们的研究表明，金融特征是决定因素。

对于最新的研究进展，我们发现没有研究是完全从模式识别角度对 P2P 借贷的数据进行数据挖掘的。这就是为什么我们认为对一个真正的 P2P 借贷数据集的分析可能是有意义的方式，可以极大地影响这一新兴金融服务的发展轨迹。这项研究的重点是

以 P2P 借贷业务模型为前提，找到真实数据中的集中点，并将这些集中点与社会网络的特点结合。

### 3.P2P 借贷数据

Prosper 网站的数据库是一个可用的让研究者对其商业模式理解的数据库。尽管不同的 P2P 借贷网站中社交特征可能使用了不同的命名约定，但在概念上他们有着类似的底层细节。因此，任何来源于分析 Prosper 数据库产生的结论都有可能是广义耳朵 P2P 借贷行业的结论。

Prosper 关系模型是由 9 个表组成：*Bid*, *Category*, *Credit Profile*, *Listing*, *Loan*, *Group*, *Loan Performance*, *Marketplace* 和 *Member*，如图 1 所示。对这次研究，*Member*, *Group*, *Listing*, *Bid* 被采用是为了构造一个可以代表借贷过程的模型。为了方便地理解不同元素，一个简要描述在下面给出：

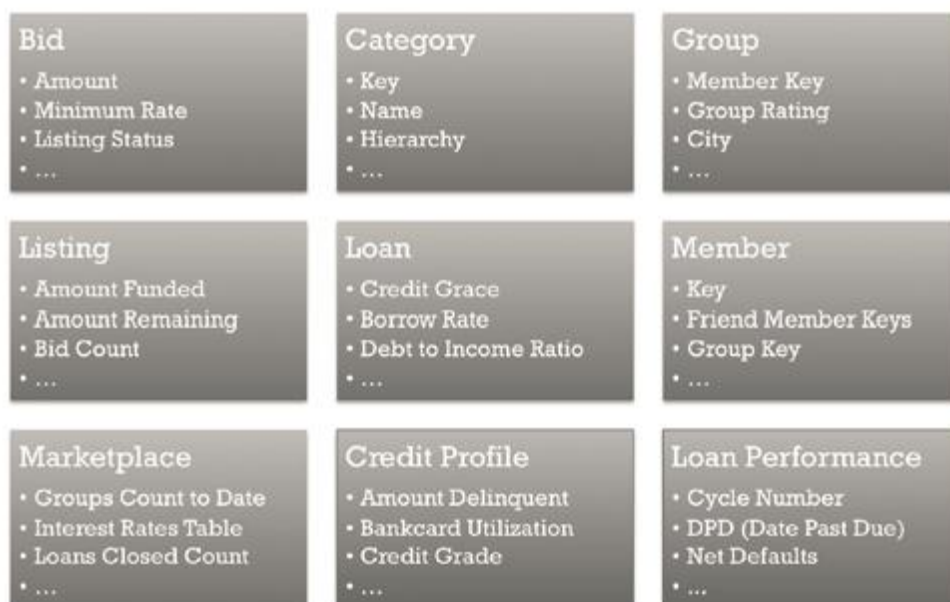


图 1：P2P 借贷关系模型

•用户是 P2P 借贷网站的注册用户。用户可以有一个或多个角色，这些角色确定用户允许执行哪些行为。拥有共同兴趣或关系的用户的集合可以加入一个组织。

•组织是相同社区的用户，有着共同的利益。组织由一个领导管理，负责接受/拒绝申请者，以及组织成员之间奖励分配。每个组织根据成员的贷款表现被 Prosper 评分。为了避免误解，有必要澄清一下组织的声誉是由 P2P 借贷网站对市场活动进行评估，而不是基于一部分成员的信贷情况。因此，信用评分差的成员也可以成为市场上评价最高的社区的一部分，只要他们被接受。因此，高评价是由具有良好的信用档案

的成员组成这种说法是一个误区。一旦被承认，成员通过自己的表现给组织带来积极或消极的表现。只有当组织存在历史活动或历史记录以后他们才会被评分，并且评分范围是 1 至 5。

- 当一个贷款人对于某个由借款人创建的 listing 愿意借给借款人钱时 Bid 会被创建。Bid 是由指定的金额和贷款人希望接收的最低利率组成。为了使得 listing 成为一笔贷款，bids 需要赢得竞卖。

- 借款人创建 listing 来描述自己的情况和他们想要借钱的原因。如果 listing 在规定时间内收到贷款人给的足够的 bids，它将成为一笔贷款。

### 3.1 在 P2P 借贷中的社交

本研究重点是理解在 P2P 借贷交易中的社交行为。就像其他在线网络引擎，P2P 借贷市场遵循社交网络的基本性质，并且根据两种现象来行为：社会影响[9]和选择[10, 11]。首先，社会影响就是一个人将想法通过社交来扩散给另一个人。第二，选择现象是为什么人们倾向于社区形式与和相似的人建立关系。这两个现象通过以下常见的部分存在于点对点借贷：

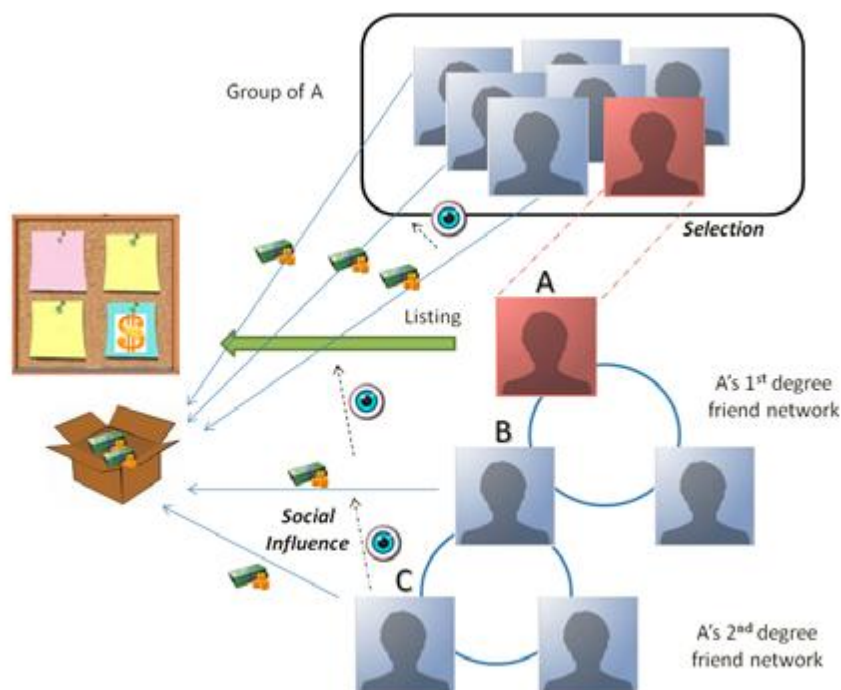


图 2：P2P 贷款中的社会影响和选择现象

- “朋友”：它代表一个一对一的链接从用户到其他借款人或贷款人。这通常是基于家庭、友谊、在 P2P 借贷中的上下文、以前的交易历史来建立的关系。这可以激励



处于借款人的二度社交网络中的贷款人基于间接新人给予 bids。例如在图 2 中，A 发布 listing 并得到来自朋友 B 的 bids，通过社会影响，朋友 B 的朋友 C 虽然不是 A 的朋友，但也会对 A 的 listing 给予 bids。

- “组织”：用户可以组成一个社区。组织成员互相帮助，该组织的评分取决于成员的表现。组织是由选择现象产生的——一个人倾向于相信那些与他们有相似之处的人。组织成员之间的信任不仅有助于创造成功的 listing，而且借款人还会产生来自同辈的压力迫使他们在还款中拥有适合的表现。组织是由领导人管理，他负责将借款人带入 P2P 借贷网站，维护组织在网站中的存在，并募集或分享奖励。借款人是组织的成员可以经常获得更好的利率，因为贷款人往往对属于受信任组织里的借款人更有信心。文[12]对社交网络中组织的形成进行了研究。

- “推荐”：用户被允许获得之前借贷交易中的用户的反馈和评价。这个公众反馈可能通过社会影响力会改变朋友或组织成员对该用户的印象。

P2P 贷款中的社会影响和选择现象如图 2 所示。

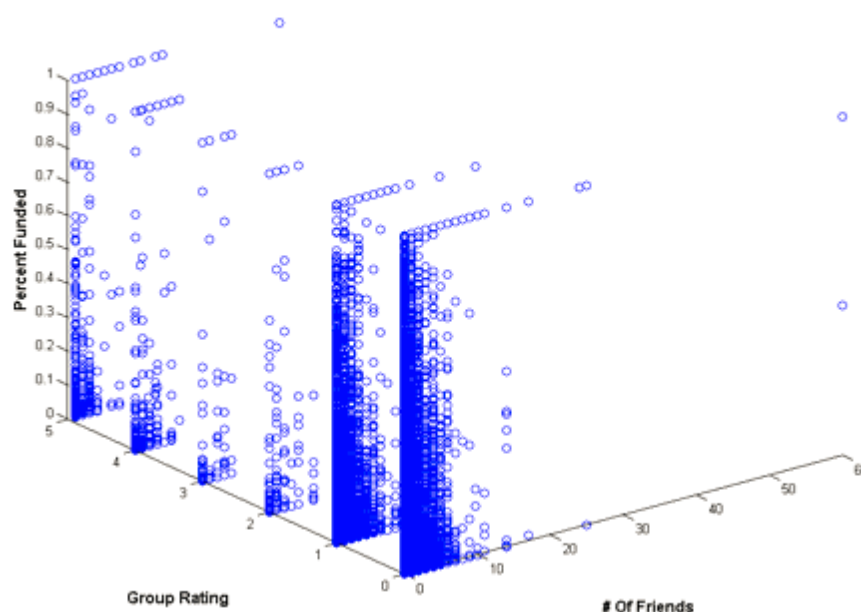


图 3：社交网络利用率

图 3 显示了 listing 通过借款人所属的组织评分和借款人的朋友数量设法收集分类的百分比。组织评级是一个由 Prosper 规定的定量评价，用来衡量组织中成员的贷款表现。等级 0 是最糟糕的，通常分配给任何人都可以加入并且具有很高的人口密度的开

放的组织；而等级 5 是分配不对外开放只有很少并且互相了解的成员的组织。大部分用户集中在低贷款百分比的那些评分只有 0 和 1 的组织中，并且他们几乎没有朋友。简单的统计数据表明，社交网络功能并不常使用，因为 88.35% 的用户没有朋友并且 77.19% 的用户属于组织评分为 0 的组织。

### 3.2 数据预处理

Prosper 数据集包含自 2005 年 11 月成立以来，所有的交易记录和用户数据。这是一个相当大数量的信息，包含约（到 2008 年 12 月为止）6,000,000 条 bids，900,000 名用户，4000 个组织和 350,000 条 listings。为了方便数据的分析，数据集被过滤为只包含所有在 2007 年创建的 listings，为这些 listings 创建的 bids，创建过这些 bids 或 listings 的用户，和最后，这些用户隶属的组织。过滤导致数据集进行了相当大的精简：1,500,00 条 bids，74,000 名用户，2,000 个组织和 93,000 条 listings。减少数据集是基于 listings 的创建日期，所以并不会缺少整个数据集中用户间行为的信息。

为了分析 P2P 借贷数据，我们有必要删除一些不必要的特征，然后编码的一些非数值化的特征。与时间、地理位置和个人描述相关的特征在这项研究中被忽略了，而金融特征，例如信用等级，将原来从 AA 到 E 的分级，转换为数值。同样，组织评分（0-5 星）和房屋所有权(没有/有)也编码成数值。我们认为描述性特性可能对贷款人评估贷款风险存在一定帮助。然而，定量分析一个描述性的特征在本次研究中不会采用。

有先前的研究对于 Prosper 的数据集进行过特征选择和分析变量之间的回归模型和关联[7]。在本文中，不同的特征选择算法被用于证实或反驳这些结果和揭示新发现。接下来，一个简洁的特征分析的描述在下面提供：

- 最大借款利率：如果 listing 变成贷款，借款人愿意支付的最大利率。
- 最后借款利率：如果 listing 变成贷款，借款人需要支付的最终利率。
- 信用评级：借款人的信用等级。它的范围从 7(最好)到 1(最差)。
- 债务收入比：借款人的债务与收入的比率。
- 房主：表明借款人是否拥有自己的住房。取值：真/假。
- #推荐：申请贷款的借款人获得的积极推荐的数量。
- # 直接朋友：直接是申请贷款的借款人的朋友的数量。
- #组织成员：借款人隶属的组织中用户的数量。

- 组织评分：基于组织历史借贷情况由 Prosper 给出的组织评分。范围是 5(最好)到 0(最差)。

- 总数量#bids：在竞卖过程中 listing 接收到的 bid 的数量。

- 投资百分比：获得资助的金额占据借款人需要的总金额的百分比。投资百分比为 90%或以上被认为是“成功 listing”，其余的则是“失败 listing”。

#### 4. P2P 借贷数据聚类

在统计模式识别中，聚类可以使用基于模型的方法，如高斯混合模型，或启发式方法，如 K-Means。利用基于模型的方法的优点是：形式上完成聚类的数量的选择或给定模型的验证。

在这种情况下，任意复杂的概率密度函数(pdf)需要被建模。众所周知，混合模型可以表示复杂条件分类 pdfs，因此，混合模型被认为是一个适合这个问题的模型。

传统的方法试图使用 expectation-maximization 算法(EM)[15],[16],[17]让数据适应有限混合模型，并且当收敛到最大似然(ML)估计时停止。不幸的是，由于贪婪的 EM 算法，如果算法在特征空间的边界收敛，不同的初始化会对结果产生显著影响。

通常,选择混合模型的数量会产生问题：选择太多的模型会使结果对于数据过拟合；而选择太少的模型将不能代表适合的分布。

为了避免这两个问题，M. Figueredo and A. Jain[14]提出的聚类技术被用于这次研究。为了方便我们采用的混合模型是高斯混合模型。这种技术使用了 EM 算法的变体并且具有一些显著的优点：

- 它比标准 EM 对初始化不敏感，并且可以避免在特征空间的边界收敛，避免取到局部极大值。

- 选择了最优聚类数量来权衡数据和模型的拟合（对数似然）和混合的复杂性（高斯模型的个数）。

第 4.1 小节给出执行高斯混合模型(GMM)技术产生的特征向量的结果。协方差矩阵为简单起见被省略了，但可按照客户要求定制。

为了比较的结果，这种聚类技术与基于启发式方法——K-Means 同样应用于相同的数据集，其中 K-Means 的 K 值等于高斯混合的模型数量。K-Means 最初没有被考虑是由于其没有能力选择正确的聚类数量。获得的结果与 GMM 结果一致，可以得到相同的结论。由于和 GMM 相似，K-Means 也可以被本文采用作为聚类的方案。

## 4.1 高斯混合模型

高斯混合模型的应用遵循[14]中所描述的方法。为了方便解释该算法生成的聚类，计算约束高斯模型需要有对角化但不同的协方差矩阵。

初始化方法遵循高熵方法称为随机开始[17]。这种技术将使算法显示自退火行为[18]。这个初始化方法的性能在[17]中被报告。

在表 4-8 中，每一列都代表着一个高斯分布。每一个聚类的几率在第一行，其余行为每个高斯分布的特征向量。每个特征向量中的值都被颜色标记来表示对借贷者的吸引程度。其中，绿色表示具有吸引力，红色表示借贷者投资存在风险。颜色的深度是根据图表中的最大值和最小值。最后在表的下方会有符号  $\checkmark$  和  $\times$ ，其中  $\checkmark$  表示成功的信用需求， $\times$  表示失败的信用需求。在 Prosper 数据中，我们认为资助的比例超过 90% 为成功。

最初，聚类算法应用于整个 2007 数据集旨在区分两组：“成功的 listings”和“失败的 listings”。然后通过结果，基于获得的结论来划分子数据集，然后重新进行聚类，这样就可以得到最符合这些结果的特征。

### 4.1.1 在2007年整个数据集的聚类

聚类算法发现三个聚类，如图 4 所示，其中两个代表“失败 listings”占据 90% 的数据，其他则为“成功 listings”，占了 10%。简单查看每个聚类结果就可以发现，隶属于一个评分高的组织和低债务收入比是吸引更多的 bids 的决定因素，并且这样最终可以得到完全资助。

P	%10	%44	%46
Max Borrower Ratio	0.20	0.24	0.12
Final Borrower Ratio	0.17	0.24	0.12
Credit Grade (7-0)	3.77	6.16	5.67
Debt To Income Ratio	0.34	0.55	0.55
Is Home Owner? (1-0)	0.49	0.30	0.35
# Endorsements	0.66	0.53	0.31
# 1st degree Friends	0.79	0.61	0.39
Group Rating (5.0)	3.89	1.04	1.01
# Bids	149.54	3.76	2.30
Percent Funded	%95	%3	%2





图 4：整个 2007 年数据集的聚类

除此之外，可以看出作为一个有房产的借款者，或是有推荐或朋友是一种后台。有趣的是，一些贷款请求，即使他们提供很有吸引力的利率或由信用等级高的人(见聚类 2)发布，都不能有效地吸引贷款者。

从聚类结果考虑，组织评分和 bids 的数量是定义聚类的决定性变量。因此，这两个变量被认为是最好的用于将整个数据集分成子集的变量。按照 bids 的数量将整个数据集分为“积极 Listings”的子数据集和“消极 Listings”的子数据集；而根据组织评分将数据集分为“可信组织的 Listings”和“开放组织的 Listings”。接下来,聚类算法将应用于每一个子集，找到自己的特点。

#### 4.1.2 “积极的Listings”的聚类

2007 年对所有 listings 产生的 bids 平均数量为 17。因此，整个数据集只保留那些 bids 数量高于平均水平的数据，从而在 P2P 贷款上下文中创造出积极的 Listings 子数据集。

聚类算法发现 6 个聚类,如图 5 所示。大约 82%的收到超过 17 个 bids 的 Listings 成功。组织评分在成功的聚类（3、4、5、6）和失败聚类（1、2）的突出区别被显示是成功的决定因素。具体来说，聚类 1 显示了一个金融情况良好（高利息、良好的信用、低债务），但是任何不属于可信组织的部分通常不会导致贷款。

	1	2	3	4	5	6
P	%13	%5	%15	%30	%25	%11
Max Borrower Ratio	0.25	0.16	0.21	0.21	0.14	0.28
Final Borrower Ratio	0.25	0.16	0.18	0.19	0.11	0.27
Credit Grade (7-0)	4.92	2.75	4.17	4.13	1.98	5.85
Debt To Income Ratio	0.35	0.47	0.56	0.37	0.30	0.26
Is Home Owner? (1-0)	0.41	0.65	0.48	0.47	0.66	0.26
# Endorsements	0.91	0.46	3.62	0.47	0.36	0.72
# 1 <sup>st</sup> degree Friends	1.01	0.58	4.53	0.53	0.51	0.73
Group Rating (5-0)	1.56	1.46	3.87	3.90	3.94	3.87
Percent Funded	0.27	0.25	0.94	0.95	0.97	0.94
	✗	✗	✓	✓	✓	✓

图 5：“积极的 Listings”的聚类

最突出的聚类中心——聚类 4，表示所有的金融和社会特征是对贷款人的吸引力的一般情况。在这个聚类中，借款人拥有很高的组织评分、低债务、良好的信用评分和合理的高利率。

在聚类 3 和 6 的例子中，我们发现在其他一些特征强的情况下可以帮助克服弱点特征。例如，在聚类 3 中，高负债收入比率被很多推荐人和朋友因素克服；此外，在聚类 4 中没有抵押品可以通过提供高利率和良好的信用信息来克服。

通过聚类 5 我们得出一个最重要的结论，也是 P2P 贷款的基础：成为高度信任的社区一员可以促进用户获得贷款，这在银行借贷中是不可被接受的。在这种情况下，低利率和低信用等级的用户由那些愿意冒着风险将他们的钱借给他们信任的人的社区成员接受。

#### 4.1.3 “消极的 Listings”的聚类

相反的情况下，bids 的数量少（小于或等于平均值）的情况提供了一个完全不同的结果，如图 6 所示。提供高利率和拥有良好的信用等级并不能保证获得贷款。此外，一种常见的特征是那些获得低于平均值 bids 数量的 Listing 往往隶属于一个低评分组织。

	1	2	3
P	%35	%60	%5
Max Borrower Ratio	0.17	0.18	0.17
Final Borrower Ratio	0.17	0.18	0.17
Credit Grade (7-0)	5.16	6.39	6.07
Debt To Income Ratio	0.33	0.30	4.59
Is Home Owner? (1-0)	0.37	0.30	0.21
# Endorsements	1.04	0.09	0.31
# 1 <sup>st</sup> degree Friends	1.20	0.14	0.40
Group Rating (5-0)	1.10	1.00	1.00
Percent Funded	%6	%1	%1
	✗	✗	✗

图6: “消极的Listings”的聚类

#### 4.1.4 “可信组织的Listings”的聚类

组织评分分布是一个偏向极限评分的分布形式: 89.3%的 Listings 属于评分为 0 和 1 的, 6.7%属于评分 5 的, 4%属于其他评分的。对此, 评分大于 1 的组织被考虑为可信组织, 而评分为 0 或 1 的被认为是开放组织。






	1	2	3	4	5
	%7	%38	%31	%17	%6
Max Borrower Ratio	0.21	0.13	0.26	0.19	0.22
Final Borrower Ratio	0.21	0.13	0.26	0.16	0.22
Credit Grade (7-0)	4.82	5.44	6.35	3.92	5.7
Debt To Income Ratio	0.25	0.51	0.34	0.55	3.24
Is Home Owner? (1-0)	0.35	0.28	0.23	0.40	0.20
#Endorsements	1.31	0.72	0.95	1.53	2.97
# 1 <sup>st</sup> degree Friends	1.06	0.69	0.82	1.64	3.68
#Bids	31	2	3	156	2
# Group Members	2.24	1	1	4	1
Percent Funded	%41	%2	%3	%99	%2
					

图 7：“可信组织的 Listings”的聚类

图 7 显示了对信任组织成员的 listings 经过聚类算法得到了五个聚类中心。这些聚类中心中借贷成功率达到 17%（聚类 4）。这些结果很明显可以看出，属于一个可信组织的成员并不能保证获得贷款，还需要提供一个合理的报价。P2P 借贷的社会特征在请求的弱点吸收中发挥关键作用，如在这种情况下的借款、用户的推荐、朋友和可信组织都很可能会帮助 listings 获得贷款。

聚类 1 代表了以下部分贷款情况：这些 listings 被认为是有吸引力；然而，缺乏推荐和朋友等社会支持阻止他们获得资金。

聚类 5 代表了极端的例子。在这种情况下，提供的情况是有吸引力的，用户也是高度认可，同时也有朋友的支持。然而，即使是组织的成员也认为债务数量太大，对他们来说这样风险太大。因此，聚类 5 的 listings 不能获得贷款。

聚类 2 和 3 是最大的两个聚类，代表了存在其他一些原因没有获得贷款的情况。

#### 4.1.5 “开放组织的 Listings”的聚类

在组织评分小于或等于 1 的 P2P 借贷数据子集中发现了四个聚类中心，如图 8 中



所描述的。通过比较高评分组织的成功 Listings 的聚类结果和低评分组织的成功 Listings 的聚类结果，我们发现，借贷成功的概率从 17% 减少到 9%。这意味着由于组织部分的影响获得借贷的机会减少了一半。

低评分组织通常是尽量开放来使得新成员加入，因此这些都是用户最密集的组织。如今，一些情况比如社区行为或是成员间的信任减弱（这些组织没有产生选择现象），或是可信任机构提供的信贷请求描述并不能获得贷款。贷款人对这些情况产生的 Listings 表现出了更保守的态度。

	1	2	3	4
	%35	%9	%21	%35
Max Borrower Ratio	0.25	0.20	0.15	0.12
Final Borrower Ratio	0.25	0.18	0.15	0.12
Credit Grade (7-0)	6.23	3.76	5.58	5.82
Debt To Income Ratio	0.55	0.33	0.53	0.57
Is Home Owner? (1-0)	0.29	0.50	0.36	0.35
#Endorsements	0.41	0.62	0.86	0.13
#1 <sup>st</sup> degree Friends	0.48	0.76	1.00	0.19
#Bids	3.92	149	3.66	1.52
#Group Members	1.04	3.9	1.03	1.00
Percent Funded	%4	%96	%3	%1
	✗	✓	✗	✗

图 8：“开放组织的 Listings”的聚类

成功案例表明在这种情况下只有低债务配合一些抵押品（例如拥有住房）才能获得贷款。结果表明加入一个高评分组织对于吸引更多的 bids 是很重要的。

## 5. 结论

对 *Prosper.com* 的 P2P 贷款数据库的分布研究发现社会特征会影响获得融资贷款请求的概率。受信任的群体不仅可以更容易获得完全资助的贷款请求，而且同时可以使借款人更好的确立合理的初次贷款的利率。没有吸引力的特征,比如高债务或低信用评级可以通过增加财务吸引力提供（例如高利息），或增加社会特征如加入高评分的

组织，得到代言或朋友的支持。

这项研究发现了这样一个事实：有吸引力的贷款请求由那些隶属于低评分的开放组织的借款人发出则不容易获得贷款，贷款人会表现出更加怀疑的态度，因此对他们产生更加保守的行为。而那些成功情况主要取决于隶属于这些开放组织的人数和他们的金融情况。

不幸的是，Prosper 提供的社交网络应用功能并没有得到合理利用。很少有用户利用他们的社交网络，试图从中受益。今天，大多数 listing 都是在不信任的气氛中进行的，主要是金融特征决定了它们的成功。然而，这次研究发现并证明那些促进社交网络发展的人更能从小额信贷中受益。

我们推测，Prosper 的社交网络没有更加流行的原因是相关的用户不具有足够的时间去同时维护多个社交网络或认识到 P2P 借贷中社交网络的重要性。此外，Prosper 的社交网络的特定目的只是为了吸引了那些想借出或者借入的用户进行多次借贷，这是一个由一个人可以生成的非常严格的过滤的整个社交网络。集成 P2P 借贷进入现有的社交网络如 Facebook, Orkut 或 MySpace 可能促进在 P2P 贷款社会特征的潜力，并且可能更能突出本文提出的结论。

## 6. 未来的工作

最近的经济活动在世界范围内对 P2P 借贷是否是有益的金融服务，对经济困难的家庭能否产生积极的影响产生了质疑。同样，传统的金融机构开始对 P2P 借贷公司的发展产生了兴趣。这个开创性的工作分析实际数据，揭示了在增加信贷的成功率中社会交往的影响。但仍然有几个问题尚未得到解答，所以提出了以下几点需要进一步的研究的方面：

- 贷款性能分析：来自组织成员的直观的、同等的压力将催促借款人按时还款。因此，组织中的友好关系对借款人的影响是需要被衡量的。

- 隐性的社交行为：由直接朋友、间接朋友和组织成员产生的相对数量的 bids 是隐性的社交特征，这也需要使用特征选择技术来衡量计算其相关性。

- 预测 listing 的成功率：给定一个成员的金融特征、社会特征和 listing 特点，预测这个潜在的信贷请求是否会获得最终的贷款。

- 预测贷款表现：给定一个借款人目前的金融特征、社会特征和贷款特征，预测贷款是否会被及时还清。根据[8]中所述，组织成员之间的相似性和社会影响可以作为预

测未来行为的依据。因此，社区的动态情况可以潜在地辅助预测和衡量交易风险。

- 混杂现象：环境（上下文），例如用户可能拥有类似的工作，在地理上相互接近，或是用户追求相同的目标，是否会影响 P2P 借贷过程中的社交情况，这是非常值得研究的。Anagnostopoulos 等人对这一现象进行了建模 [13]。

- 无标度网络：用户的 Bids 引起了其他用户的注意，并最终形成一个导致整个信贷请求推进的动量效应。这就是所谓的优先连接。这种效应与小世界网络中的一种名叫无标度网络的原则[19]相似——新顶点优先连接到网络中连接的顶点。P2P 借贷的社交网络与小世界网络的相似之处仍需要在未来的工作中被研究。

## 7. 鸣谢

我们需要感谢给我们帮助的 Ray Garcia, Prof. Rosalind Picard 和 the MAS 662J students，他们在研究中给了我们很多建设性的建议。

我们也需要感谢 Prosper.com 开放了他们的数据库，给我们研究。

这项研究是由 the EJ/GV Researcher Formation Fellowship BFI.08.80 支持的。

## 外文文献原稿

# Social Interactions in P2P Lending

Sergio Herrero-Lopez  
MIT  
77 Massachusetts Avenue  
Cambridge, MA 02139-4307  
+1-617-417-6055  
sherrero@mit.edu

## ABSTRACT

Access to capital in the form of credit through money lending requires that the lender to be able to measure the risk of repayment for a given return. In ancient times money lending needed to occur between known parties or required collateral to secure the loan. In the modern era of banking institutions provide loans to individuals who meet a qualification test. Grameen Bank in Bangladesh has demonstrated that small poor communities benefited from the “microcredit” financial innovation, which allowed a priori non-bankable entrepreneurs to engage in self-employment projects. Online P2P (Peer to Peer) lending is considered an evolution of the microcredit concept, and reflects the application of its principles into internet communities. Internet ventures like Prosper.com, Zopa or Lendingclub.com, provide the means for lenders and borrowers to meet, interact and define relationships as part of social groups. This paper measures the influence of social interactions in the risk evaluation of a money request; with special focus on the impact of one-to-one and one-to-many relationships. The results showed that fostering social features increases the chances of getting a loan fully funded, when financial features are not enough to construct a differentiating successful credit request. For this task, a model-based clustering method was applied on actual P2P Lending data provided by Prosper.com.

## Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – *Data Mining*.

J.4 [Social and Behavioral Sciences]: Sociology.

## General Terms

Economics, Human Factors.

## Keywords

Social Networks, Clustering, Gaussian mixture.

## 1. INTRODUCTION

P2P Lending on *Prosper.com* is based on an online reverse auction where individuals request either to borrow money, taking the *Borrower* role, or buy loans, taking the *Lenders* role. Borrowers set the amount of money they need and the maximum interest rate they would be willing to pay by posting a *Listing*; and lenders bid on their loans partial amounts and set the minimum interest rate they want to receive. P2P Lending users can adopt the borrower role, lender role or both. The main difference with traditional banking industry is that the lender not only has borrower’s financial information available, but they can also evaluate the risk of an operation based on the social characteristics of the individual.

The management of the auction is carried out by the P2P Lending Company. It collects and displays listings or bids made by users, and provides a social networking engine that enables users establishing relationships with other users and joining or creating groups across the internet.

The goal of this project is not only to reveal hidden patterns in the behavior of P2P Lending users, but also to confirm or refute a priori obvious behaviors that are often assumed. This work is focused on the application of several data clustering techniques in order to discover and analyze the underlying distribution patterns of the P2P Lending data.

The organization of this paper is as follows: Section II summarizes previous work on P2P Lending. Section III introduces the structure of the Prosper dataset, and describes the data pre-processing that had to be applied prior to proceeding to the analysis of the information. This section also presents the common social features of P2P Lending sites. The application of data clustering techniques on P2P Lending data is described in Section IV. The conclusions of this study are explained in Section V. Finally, Section VI proposes future work plan to conduct to further an understanding of this emerging financial service.

## 2. RELATED WORK

The economic value of social capital in P2P lending has not been extensively explored in academic research. Most analysis so far has been done in an economic context.

Knack & Keefer [1], present evidence that social capital does matter for measurable economic performance.

Hulme & Wright [2] provide an extensive and in-depth treatment of the social lending subject from multiple perspectives.

Whether social networks solve the information asymmetry problem in peer-to-peer lending is studied in [3], revealing that estimated returns of group loans are lower than those of non-group loans due to lender learning and the elimination of group leader rewards.

The role of financial intermediaries on the P2P online market is analyzed by Berger & Gleisner [4], and it demonstrates that the recommendation of a borrower significantly enhances credit conditions, and the intermediary’s bid on a credit listing has a crucial impact on the resulting interest rate.

Dominants of success in securing a loan on Prosper are explored by Herzenstein & Andrews [5], where authors find that borrowers’ financial strength and efforts after they post a listing are major factors in determining whether a solicitation will be successful when compared to demographic factors.

The work presented by Sydnor, Pope [6] shows evidence of discrimination by examining how lenders respond to signaling characteristics such as race, age, and gender.

J.Ryan, K.Reuk and C.Wang [7], analyzed the regression model and correlation between variables using real-world data. The purpose of this study was to weight the relative relevance of each of the financial and social features independently, and determine their influence in the success on the conversion of a listing to a loan. Their study showed that financial features are determinant.

To our best knowledge, no research has been done from the perspective of pattern recognition and knowledge discovery on P2P Lending data. This is the reason why we consider that the analysis of a real P2P Lending dataset may reveal meaningful patterns that can significantly impact the trajectory of this emerging financial service. This work is focused on finding existing concentrations in real data, and directly associating these concentrations with the characteristics of the social network on which P2P Lending business model is premised.

### 3. P2P LENDING DATA

Prosper Marketplace's database is available to researchers to facilitate the understanding of its business model. Even though different P2P Lending social networks may utilize different naming conventions, the underlying details are conceptually similar. Consequently, any conclusions derived from analyzing Prosper's database may be generalized for the P2P Lending industry.

The Prosper relational model is composed by nine tables: *Bid*, *Category*, *Credit Profile*, *Listing*, *Loan*, *Group*, *Loan Performance*, *Marketplace* and *Member* as shown in Figure 1. For this project, Member, Group, Listing and Bid tables were joined in order to construct a flat representation that could be processed. In order to facilitate the understanding of the different elements a brief description of each of these tables is given:

<b>Bid</b> <ul style="list-style-type: none"> <li>• Amount</li> <li>• Minimum Rate</li> <li>• Listing Status</li> <li>• ...</li> </ul>	<b>Category</b> <ul style="list-style-type: none"> <li>• Key</li> <li>• Name</li> <li>• Hierarchy</li> <li>• ...</li> </ul>	<b>Group</b> <ul style="list-style-type: none"> <li>• Member Key</li> <li>• Group Rating</li> <li>• City</li> <li>• ...</li> </ul>
<b>Listing</b> <ul style="list-style-type: none"> <li>• Amount Funded</li> <li>• Amount Remaining</li> <li>• Bid Count</li> <li>• ...</li> </ul>	<b>Loan</b> <ul style="list-style-type: none"> <li>• Credit Grace</li> <li>• Borrow Rate</li> <li>• Debt to Income Ratio</li> <li>• ...</li> </ul>	<b>Member</b> <ul style="list-style-type: none"> <li>• Key</li> <li>• Friend Member Keys</li> <li>• Group Key</li> <li>• ...</li> </ul>
<b>Marketplace</b> <ul style="list-style-type: none"> <li>• Groups Count to Date</li> <li>• Interest Rates Table</li> <li>• Loans Closed Count</li> <li>• ...</li> </ul>	<b>Credit Profile</b> <ul style="list-style-type: none"> <li>• Amount Delinquent</li> <li>• Bankcard Utilization</li> <li>• Credit Grade</li> <li>• ...</li> </ul>	<b>Loan Performance</b> <ul style="list-style-type: none"> <li>• Cycle Number</li> <li>• DPD (Date Past Due)</li> <li>• Net Defaults</li> <li>• ...</li> </ul>

Figure 1: P2P Lending Relational Model

- A Member is a registered user of the P2P Lending site. Members may have one or multiple roles that determine which actions the Member is allowed to perform on the site. A collection of Members who share a common interest or affiliation join into a Group.
- A Group is a community of Members that have a common interest. The Group is managed by a leader, who is in charge of admitting/rejecting applicants, and distributing rewards

among Group Members. Each Group is rated by Prosper according to the loan performance of its Members. In order to avoid misunderstanding, it is necessary to clarify that the reputation of a Group is assessed by the P2P Lending site based merely on marketplace activity and not on the credit profiles of the members that are part of it. Therefore, members with poor Credit Scores can also be part of communities with the highest reputation in the marketplace, as long as they are accepted. Hence, it is wrong to assume that Groups with high reputation are formed by members with good credit profiles. Once admitted, individual members help positively or negatively the Group as with their performance. Groups are not rated until they have a representative history of activity, and after that they are rated on a one to five scale.

- A Bid is created when a Lender wishes to lend money to a Borrower in response to a Listing the Borrower created to solicit Bids. Bids are created by specifying an amount and a minimum rate the lender wishes to receive. In order to become a Loan, the Bids need to win the auction.
- Borrowers create Listings to solicit bids by describing themselves and the reason they are looking to borrow money. If the Listing receives enough bids by Lenders to reach the amount requested before the Listing period ends, it will become a loan.

### 3.1 Social Components in P2P Lending

This study is focused on understanding social interactions in P2P Lending transactions. Like other online networking engines, P2P Lending marketplaces follow the fundamental properties of social networks, and behave according to two phenomena: *social influence* [9] and *selection* [10, 11]. First, it is understood by *social influence* when a person diffuses its ideas to another person through interaction. Second, *selection* phenomenon is a reason why people tend to form communities and have relationships with persons who are similar to them. These two phenomena are present in peer-to-peer lending through the following common components of:

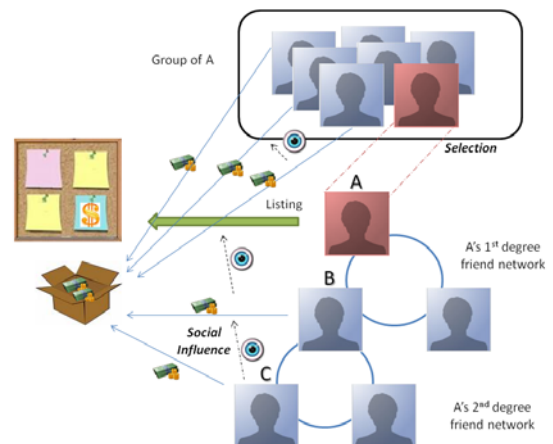


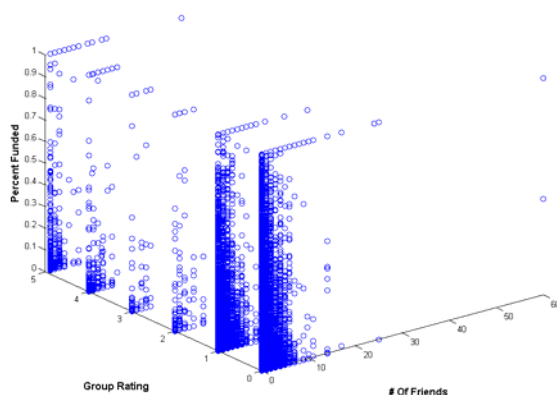
Figure 2: Social Influence and Selection

- "Friend": It represents a one-to-one link from a member to other borrowers or lenders. This relationship between members is usually based on family, friendship or

previous transaction history in the P2P Lending context. It is made public and intends to motivate lenders within borrower's second degree social network to bid based on indirect trust. For example in Figure 2, Person A posted a listing and gets bids from its friend Person B. By *social influence*, Person C, who is friend of Person B, but not necessarily friend of Person A, is going to bid on Person A's listing as well.

- “Group”: Members are allowed to form communities. Group members help each other and the Group Rating depends on their performance. Groups are formed by the *selection* phenomenon, where individuals tend to trust those that share similarities with them. The trust among group members not only facilitates creating successful listings, but also generates peer pressure on colleague borrowers to force them to have an appropriate loan performance. Groups are managed by group Leaders who bring borrowers to the P2P Lending site, maintain the presence of the group in the site, and collect or share group rewards. Borrowers who are members of a group often get better interest rates because Lenders tend to have more confidence in Borrowers that belong to trusted groups. Group formation in social networks was studied in [12].
- “Endorsement”: Members are allowed to give public feedback on previous transactions with other members. This public feedback might alter the impression of friends or group colleagues through *social influence*.

*Social influence* and *selection* phenomena in the scope of P2P Lending are illustrated in Figure 2.



**Figure 3: Social Network Utilization**

Figure 3 shows the percent that a listing managed to collect classified by the Group Rating of the group that the borrower belongs to and the number of friends that the borrower has. The Group Rating is a quantitative evaluation generated by Prosper that measures the loan performance of the members in the group. Rating 0 is the worst, usually assigned to open groups where anybody can join and are highly populated, whereas rating 5 is assigned to exclusive groups with no open access and very few members that know each other very well. The largest concentration is located in the low values of percent funded of

groups ratings 0 and 1 and few friends. Simple statistics indicate that the social networking features are not often used, since 88.35% of the members have no friends and 77.19% belong to groups with rating 0.

### 3.2 Data Preprocessing

The Prosper dataset contains all the transaction and member data since its inception in November 2005. This is a considerable volume of information that encloses approximately (by December 2008) 6 Million bids, 900,000 members, 4,000 groups and 350,000 listings. In order to facilitate the analysis of the data, the dataset was filtered to contain all the listings created in calendar year 2007, the bids created for these listings, the subset of members that created these listings and bids, and finally, the groups these members are affiliated with. The filtering resulted in considerable reduction of the dataset: 1.5 Million bids, 74,000 members, 2,000 groups and 93,000 listings. The reduction of the dataset based on the listing creation date allowed having a snapshot of the entire dataset without losing information on the activity and interactions between members.

In order to analyze the P2P Lending data, it was necessary to remove unnecessary features and encode some of the remaining into numerical values. Features related with time, geo-location and personal descriptions were ignored for this study, while financial features like the credit grade, which is graded from AA to E, were transformed to numerical scales. Similarly, Group Rating (0-5 stars) and home ownership (No, Yes) indicators were encoded into numbers. It is considered that descriptive features may have relative relevance for lenders when evaluating the risk of a listing. Nevertheless, the quantitative evaluation of a descriptive feature is not considered in this work.

There is previous research of Prosper's dataset that carried out feature selection and analyzed the regression model and correlation between variables [7]. In this paper, different feature selection algorithms were applied to confirm or refute those results and reveal new findings. Next, a succinct description of the features analyzed in subsequent sections is provided:

- Max Borrower Ratio: Maximum interest the borrower is willing to pay if the listing becomes a loan.
- Final Borrower Ratio: The final interest that the borrower will pay if the listing becomes a loan.
- Credit Grade: Borrower's credit grade. It goes from 7 (best) to 1(worst).
- Debt To Income Ratio: Comparison of borrower's debt to its income.
- Homeowner: Indicates whether the borrower owns its home. True/False.
- # Endorsements: Number of positive endorsements received by the borrower requesting the loan.
- # 1st Degree Friends: Number of direct friends of the borrower requesting the loan.
- # Group Members: Number of users in the group that the borrower is affiliated with.



- **Group Rating:** Quantitative evaluation of the performance of the group given by Prosper based on the transaction history. 5 (best) to 0 (worst).
- **Total # of Bids:** Number of bids a listing receives during the auction.
- **Percent funded:** Fraction of the total amount that the borrower managed to collect during the lifetime of the listing. Listings with 90% funded or above are considered “successful listings”, while the rest are “failed listings”.

## 4. P2P LENDING DATA CLUSTERING

In statistical pattern recognition, clustering may be approached using model-based methods, such as Gaussian Mixture Models, or heuristic methods, like K-Means. The advantage of the utilization of model-based methods is that the selection of the number of clusters or the validation of a given model can be done formally.

In this case, an arbitrarily complex probability density function (pdf) needs to be modeled. It is known that mixture models can represent complex class-conditional pdfs, consequently, mixture models are considered an appropriate fit for this problem.

Conventional approaches try to fit finite mixture models to data using the *expectation-maximization* algorithm (EM) [15],[16],[17], and stop when converging to a maximum likelihood (ML) estimate. Unfortunately, due to the greedy nature of the EM algorithm, initialization can dramatically impact the result, if the algorithm converges in the boundary of the feature space.

Usually, selecting the number of mixture components can be problematic: the selection of too many components will certainly over fit the data, while too few components will not represent the distribution appropriately.

In order to avoid these two problems, the clustering technique presented by M. Figueredo and A. Jain in [14] was used for this study. The mixtures selected were Gaussians for convenience. This technique uses a variant to the EM algorithm and has some remarkable advantages:

- It is less sensitive to the initialization than standard EM and by avoiding the boundary of the feature space it does not converge to a local maximum.
- The optimum number of components is selected based on a tradeoff between the fit of the model to the data (log likelihood) and the complexity of the mixture (number of Gaussians).

The results shown in subsection 4.1 correspond to the mean vectors obtained after the execution of the Gaussian Mixture Models (GMM) technique. The covariance matrices are omitted for simplicity but are available upon request.

In order to compare the results of this clustering technique with

heuristics based methods, K-Means was applied to the same datasets for a number of K equal to the number of mixtures provided by GMM for every case. K-Means was initially not considered due to its incapacity to select the number of clusters formally. The obtained results were consistent with the GMM results and the same conclusions could be obtained from them. Due to the similarity to GMM, the cluster configurations obtained for K-Means are omitted from this publication.

### 4.1 Gaussian Mixture Models

The application of Gaussian Mixture Models follows the approach described by in [14]. In order to facilitate the interpretation of the clusters generated by this algorithm, the calculation constrained the Gaussians to have diagonal but different covariance matrices.

The initialization method follows a high entropy method called *random starting* [17]. This technique will make the algorithm show a self-annealing behavior [18]. The performance of this initialization method has been reported in [17].

Each column in the tables of Figures 4-8 represents a Gaussian. The probability of the cluster is indicated on the top, along with the values of the mean vector for each of the Gaussian models. The individual values of the mean vectors are color-coded to represent its grade of attractiveness for lenders. Green denotes attractive, while red represents repelling for lenders to risk their money. The degradation of the colors is based on the maximum/minimum values for each row in the table, and consequently the corresponding dimension in the mean vector. The outcome of the listing is indicated at the bottom by symbols  $\checkmark$ , for successful credit request, and X for failed credit requests. For this publication, it is assumed that a percent funded greater than 90% should be considered successful.

Initially, the clustering algorithm is going to be applied to the entire 2007 dataset aiming to differentiate two groups: “successful listings” and “failed listings”. Once that this groups are found, the data is sub-partitioned based on gained insights and re-clustering is applied in order to obtain the characteristics that best define each of these groups.

P	%10	%44	%46
Max Borrower Ratio	0.20	0.24	0.12
Final Borrower Ratio	0.17	0.24	0.12
Credit Grade (7-0)	3.77	6.16	5.67
Debt To Income Ratio	0.34	0.55	0.55
Is Home Owner? (1-0)	0.49	0.30	0.35
# Endorsements	0.66	0.53	0.31
# 1st degree Friends	0.79	0.61	0.39
Group Rating (5-0)	3.89	1.04	1.01
# Bids	149.54	3.76	2.30
Percent Funded	%95	%3	%2



Figure 4: Clustering entire 2007 data

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. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

The 3rd SNA-KDD Workshop '09 (SNA-KDD'09), June 28, 2009, Paris, France. Copyright 2009 ACM 978-1-59593-848-0...\$5.00.



#### 4.1.1 Clustering on the entire 2007 dataset:

The clustering algorithm found three clusters as shown in Figure 4, two of which represent the “failed listings” and enclose 90% of the occurrences; and the other the “successful listings”, which represents 10% of the listings posted. Simple inspection of the values of the means for each cluster indicates that being affiliated to a group with high rating and having a low debt to income ratio are determinant factors to attract more bids and eventually get the listing completely funded.

Besides, it can be seen that being a home owner, having endorsements and friends are in the background. It is interesting to see that some loan requests, even if they offer attractive interests and carried out by people with high credit grades (see Cluster 2), do not attract the attention of the lender community.

From the clustering results, it is contemplated that Group Rating and the # of bids are determinant variables for the definition of the clusters. Therefore, these two variables are considered the best to be used to divide the entire dataset into subsets. The number of bids divides “Active Listings” from “Passive Listings”, while Group Rating separates “Trusted Group Listings” from “Open Group Listings”. Next, the clustering algorithm is going to be applied on each of these subsets so as to find their defining characteristics.

#### 4.1.2 Clustering on “Active Listings”

The average number of bids for all the listings generated in 2007 is 17. Therefore, the entire dataset is filtered to keep only those listings that received more bids than the average, and consequently created significant activity in the P2P Lending context.

	1	2	3	4	5	6
P	%13	%5	%15	%30	%25	%11
Max Borrower Ratio	0.25	0.16	0.21	0.21	0.14	0.28
Final Borrower Ratio	0.25	0.16	0.18	0.19	0.11	0.27
Credit Grade (7-0)	4.92	2.75	4.17	4.13	1.98	5.85
Debt To Income Ratio	0.35	0.47	0.56	0.37	0.30	0.26
Is Home Owner? (1-0)	0.41	0.65	0.48	0.47	0.66	0.26
# Endorsements	0.91	0.46	3.62	0.47	0.36	0.72
# 1 <sup>st</sup> degree Friends	1.01	0.58	4.53	0.53	0.51	0.73
Group Rating (5-0)	1.56	1.46	3.87	3.90	3.94	3.87
Percent Funded	0.27	0.25	0.94	0.95	0.97	0.94
	✗	✗	✓	✓	✓	✓

Figure 5: Clustering “Active listings”

The clustering algorithm found six clusters that are described in Figure 5. Approximately 82% of the listings that received more than 17 bids were successful. The Group Rating is shown to be a

determinant factor for the success and is the prominent difference between successful Clusters (3,4,5,6) and failed Clusters (1,2). Specifically, Cluster 1 shows a financially good offer (high interest, good credit, low debt), but out of any of the Trusted Groups does not usually lead to getting the loan.

The most prominent concentration, Cluster 4, represents the general case when all the features are attractive for lenders. In this cluster, high Group Rating, low debt, good credit score and good high interest rates are offered.

Clusters 3 and 6 are examples in which being strong in some features can help overcome other weak characteristics. For example, a high debt to income ratio is concealed by having several endorsements and first degree friends in Cluster 3; furthermore, not having collateral can be overcome by offering a high interest and a good credit profile as in Cluster 4.

Cluster 5 leads to one of the most important conclusions, which are also the basis of P2P Lending: Being member of a highly trusted community facilitates getting a loan for members that a priori should be classified as non-bankable. In this case, low interest and low credit grades are assimilated by the members of the community, who are willing to risk their money for people they trust.

#### 4.1.3 Clustering on “Passive Listings”

The opposite situation, the case in which the number of bids is less or equal to the average provides a completely different result, as shown in Figure . Offering a high interest rate and a good credit grade do not guarantee getting the loan fully funded. Besides, being affiliated to a low rating group is a common characteristic of those listings that receive fewer bids than the average.

	1	2	3
P	%35	%60	%5
Max Borrower Ratio	0.17	0.18	0.17
Final Borrower Ratio	0.17	0.18	0.17
Credit Grade (7-0)	5.16	6.39	6.07
Debt To Income Ratio	0.33	0.30	4.59
Is Home Owner? (1-0)	0.37	0.30	0.21
# Endorsements	1.04	0.09	0.31
# 1 <sup>st</sup> degree Friends	1.20	0.14	0.40
Group Rating (5-0)	1.10	1.00	1.00
Percent Funded	%6	%1	%1
	✗	✗	✗

Figure 6: Clustering "Passive Listings"

#### 4.1.4 Clustering on "Trusted Group Listings"

The Group Rating distribution is biased to the extreme rating values: 89.3% of the listings belong to groups with rating 0 and 1, 6.7% to group with rating 5 and 4% to the rest. For this analysis, Trusted Groups are considered those with Group Rating greater than 1, while Open Groups are those with rating 0 or 1.

Figure 7 shows the five concentrations obtained after running the clustering algorithm on the listings posted by members that are affiliated to Trusted Groups. The success rate of this concentration is 17% and is represented by Cluster 4. It is clear from these results that being member of a Trusted Group does not guarantee to get a loan, it is still necessary to present a reasonable offer. The social characteristics of P2P Lending will play a key role to assimilate weaknesses of the request, like the debt in this case, and members with endorsements, friends and populated groups will be likely to get the listings funded.

Cluster 1 represents the partially funded case. These listings are considered attractive; however, the lack of social support in form of endorsements and friends prevented them getting the funding.

Cluster 5 represents the extreme case. In this case, the offer is attractive and the members are highly endorsed and supported by friends. Nevertheless, the debt is too large to be assumed even for members of the group and it is too risky even for them. Therefore, listings in Cluster 5 do not get the loan.

Clusters 2 and 3 have the highest concentrations and represent the cases in which for several reasons the listings do not get funded.

	1	2	3	4	5
	%7	%38	%31	%17	%6
Max Borrower Ratio	0.21	0.13	0.26	0.19	0.22
Final Borrower Ratio	0.21	0.13	0.26	0.16	0.22
Credit Grade (7-0)	4.82	5.44	6.35	3.92	5.7
Debt To Income Ratio	0.25	0.51	0.34	0.55	3.24
Is Home Owner? (1-0)	0.35	0.28	0.23	0.40	0.20
#Endorsements	1.31	0.72	0.95	1.53	2.97
#1 <sup>st</sup> degree Friends	1.06	0.69	0.82	1.64	3.68
#Bids	31	2	3	156	2
#Group Members	2.24	1	1	4	1
Percent Funded	%41	%2	%3	%99	%2



Figure 7: Clustering "Trusted Group Listings"

#### 4.1.5 Clustering on "Open Group Listings"

Four concentrations were found in the subset of the P2P Lending data with Group Rating less or equal to 1, as described in Figure 8. By comparing the successful cluster in the high Group Rating results to the successful cluster of the low Group Rating results, it is necessary to indicate that the probability has been reduced from 17% to 9%. This means that the chances of getting a listing funded are reduced to half due to the influence of the group component.

Low rating groups are usually open to new members to join and consequently these are highly populated. Components like the community behavior and the trust between members are attenuated (these groups were not created by the *selection* phenomenon), and credit request profiles that were funded in trusted scenarios are not accepted this time. Lenders show to be more conservative towards listings generated in this scenario.

Success stories show that only low debts are accepted in this situation, along with some collateral (such as having a house). The results show that is also important to be affiliated to highly populated groups to attract more bids.

	1	2	3	4
	%35	%9	%21	%35
Max Borrower Ratio	0.25	0.20	0.15	0.12
Final Borrower Ratio	0.25	0.18	0.15	0.12
Credit Grade (7-0)	6.23	3.76	5.58	5.82
Debt To Income Ratio	0.55	0.33	0.53	0.57
Is Home Owner? (1-0)	0.29	0.50	0.36	0.35
#Endorsements	0.41	0.62	0.86	0.13
#1 <sup>st</sup> degree Friends	0.48	0.76	1.00	0.19
#Bids	3.92	149	3.66	1.52
#Group Members	1.04	3.9	1.03	1.00
Percent Funded	%4	%96	%3	%1



Figure 8: Clustering "Open Group Listings"

## 5. CONCLUSIONS

The study of the distribution of the *Prosper.com* P2P Lending database confirms that social features influence the probability of getting funding for a loan request. Affiliation with Trusted Groups not only doubles the probability of getting a loan request fully funded, but also, establishes the scenario for borrowers with a priori non-bankable profile to get a loan with reasonable rates. Unattractive features such as high debt or low credit scores can be

overcome by either financially attractive offers (high interest), or social capital such as affiliation with highly rated groups, endorsements or friend support.

This work uncovered the fact that attractive loan requests made by people affiliated with Open Groups do not easily get a loan, and lenders show to be more skeptical and hence have a more conservative behavior towards them. The success of these strongly depends on the number of people affiliated to these Open Groups and their financial profile.

Unfortunately, the social networking application provided by Prosper is not properly exploited. Few members utilize their social network and try to benefit from it. Today, the majority of the listings are performed in a distrust atmosphere where predominantly financial features dictate their success. However, this studies finding demonstrate that those that do foster the social network do get benefit from it and replicate the microcredit financial principles.

We speculate that the cause of the Prosper social network not being more populated may be associated users do not having enough time to maintain several social networks simultaneously or realizing its important in the lending determination. Furthermore, the specific purpose of Prosper's social network only attracts to those users that intend to lend or borrow possibly multiple times, which is a very strict filter of the entire social network that a single person can generate. The integration of the P2P Lending in an existing social network such as Facebook, Orkut or MySpace may boost the potential of the social interactions in P2P Lending and probably emphasize the results that were presented in this paper.

## 6. FUTURE WORK

Recent economic events worldwide have raised the question whether P2P Lending may be a beneficial financial service and might positively impact families in economic trouble. Similarly, traditional financial institutions are following with interest evolution of P2P Lending companies. This pioneering work analyzes real data and reveals the influence of social interactions in increasing the success of a credit request. Several questions remain unanswered and are proposed for further research:

- **Loan Performance Analysis:** Intuitively, peer pressure from group members will push borrowers to pay on time. Therefore, the impact of a group affiliation in the performance of a borrower needs to be measured.
- **Hidden Social Interactions:** The relative amount of bids originated by first degree friends, second degree friends and group members, are hidden social features that need to be calculated and their relevance measured using feature selection techniques.
- **Listing Success Predictability:** Given a member's current financial profile, social profile and listing characteristics, predict whether a potential credit request would eventually be funded.
- **Loan Performance Predictability:** Given a borrower's current financial profile, social profile and loan characteristics, predict whether the loan would be timely paid. According to [8] similarities between group members and *social influence*

can serve as predictors of future behavior. Therefore, the dynamics of the community could potentially help predict and measure the risk of a transaction.

- **Confounding phenomenon:** It is necessary to study, whether the environment (in the sense of context), where users might have similar jobs, be geographically close to each other, or pursue the same goal, influences social interactions in P2P Lending. Anagnostopoulos et al. modeled this phenomenon in [13].
- **Scale-Free network:** Bids from Members call the attention of other Members and eventually generate a momentum effect that leads to fund the entire credit request. This is known as preferential attachment. This effect is comparable to the principles of scale-free networks [19], a type of small-world networks, in which new vertices connect preferentially to the more highly connected vertices in the network. The resemblance of P2P Lending social networks with small-world networks needs to be studied in future work.

## 7. ACKNOWLEDGMENTS

We want to give our thanks to Ray Garcia, Prof. Rosalind Picard and the MAS 662J students for giving constructive feedback in this work.

We also appreciate *Prosper.com* opening their database for research.

This work was supported by the EJ/GV Researcher Formation Fellowship BFI.08.80

## 8. REFERENCES

- [1] S. Knack, P. Keefer. Does Social Capital Have An Economic Payoff? A Cross-Country Investigation, Quarterly Journal of Economics, 1997.
- [2] M. Hulme., C. Wright. Internet Based Social Lending: Past, Present and Future. Social Futures Observatory. October 2006.
- [3] S. Freedman, G.Z. Jin. Do Social Networks Solve Information Problems for Peer-to-Peer Lending? Evidence from Prosper. Com - papers.ssrn.com.
- [4] S. Berger, F. Gleisner. Emergence of Financial Intermediaries on Electronic Markets: The Case of Online P2P Lending - papers.ssrn.com
- [5] M. Herzenstein, R. Andrews, U .Dholakia, E. Lyandres. The Democratization of Personal Consumer Loans? Determinants of Success in Online Peer-to-Peer Lending Communities - papers.ssrn.com
- [6] J.R. Sydnor, D.G. Pope. What's in a Picture? Evidence of Discrimination from Prosper. com - papers.ssrn.com
- [7] J.Ryan, K.Reuk, C.Wang, "To Fund Or Not To Fund: Determinants Of Loan Fundability in the Prosper.com Marketplace.", Stanford Graduate School of Business.
- [8] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Las Vegas, Nevada, USA,

- August 24 - 27, 2008). KDD '08. ACM, New York, NY, 160-168.
- [9] N. E. Friedkin. A Structural Theory of Social Influence. Cambridge University Press, 1998.
  - [10] P. Lazarsfeld and R. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Bergen, T. Abel, and C. Page, editors, *Freedom and Control in Modern Society*. Van Nostrand, 1954.
  - [11] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 2001.
  - [12] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA, August 20 - 23, 2006). KDD '06. ACM, New York, NY, 44-54.
  - [13] Anagnostopoulos, A., Kumar, R., and Mahdian, M. 2008. Influence and correlation in social networks. In *Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA, August 24 - 27, 2008). KDD '08. ACM, New York, NY, 7-15
  - [14] Figueiredo, M. A. and Jain, A. K. 2002. Unsupervised Learning of Finite Mixture Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 3 (Mar. 2002), 381-396.
  - [15] Dempster, A., Laird, N., and Rubin, D. Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm. *J. Royal Statistical Soc. B*, vol. 39, pp. 1-38, 1977.
  - [16] McLachlan, G., Krishnan, T. *The EM Algorithm and Extensions*. New York: John Wiley & Sons, 1997.
  - [17] McLachlan, G., Peel, D. *Finite Mixture Models*. New York: John Wiley & Sons, 2000.
  - [18] Rangarajan, A. "Self Annealing: Unifying Deterministic Annealing and Relaxation Labeling," *Energy Minimization Methods in Computer Vision and Pattern Recognition*, M. Pellilo and E. Hancock, eds., pp. 229-244, Springer Verlag, 1997.
  - [19] Amaral, L.A.N., Scala, A., Barthelemy, M., Stanley, H.E., Classes of small-world networks, *Proceedings of the National Academy of Sciences of the United States of America*. PNAS, October 10, 2000.