A Bayesian Investment Model for Online P2P Lending*

Xubo Wang, Defu Zhang, Xiangxiang Zeng**, and Xiaoying Wu

Department of Computer Science, Xiamen University, Xiamen, China
xwang@stu.xmu.edu.cn, {dfzhang,xzeng}@xmu.edu.cn,
wuxiaoying0720@126.com

Abstract. P2P online lending is an emerging economic lending model. In this marketplace, borrowers submit requests for loans, and lenders make bids on them. It has put forward new challenges to investors about how to make effective investment decisions. Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies. In the paper, we calculate the mutual information of every two variables to measure their mutual dependence and build a Bayesian network model to select loans that would pay back with high confidence. We perform abundant experiments on the data from the world's largest P2P lending platform Prosper.com. Experimental results reveal that Bayesian network model can significantly help investors make better investment decisions than other investment models.

Keywords: P2P lending, Classification, Bayesian network, Tree Augmented Naïve Bayesian.

1 Introduction

P2P lending , also called online social lending, allows direct lending and borrowing between individuals on an Internet-based platform, without the participation of traditional financial intermediaries such as banks (Wang, 2009). In this way, it provides convenient online services for reallocating small funds in credit transaction. There are more than 30 online P2P lending markets in more than 10 countries in the world, such as Zopa in UK and Prosper in the US. In recent years, advances in P2P lending marketplaces have provided new research opportunities with the availability of massive amounts of P2P transaction data. In this study, we focus on Prosper (http://www.prosper.com), the largest online P2P lending market in US, which has helped its 1.26 million members receive over $314 million loans. In this marketplace, borrowers submit requests for loans (called listing), and then lenders make bids on them. Prosper handles the aggregation and disbursement of funds to borrowers and then services the loans, collecting and distributing payments and interest back to the loan investors.

* The work was supported National Natural Science Foundation of China (61202011 and 61272385).

** Corresponding author.

J. Su et al. (Eds.): ICoC 2013, CCIS 401, pp. 21–30, 2013.

P2P lending, as a novel economic model, has been studied extensively in recent years, and is mainly focused on borrowers' social networks and personal information, loan attributes, lenders' decisions and so on. As for social networks, Freedman & Jin (2008) have investigated whether they solve the information asymmetry problem in peer-to-peer lending. They found that loans with friend endorsements and friend bids have fewer missed payments, but the estimated return of group loans is lower than those of non-group loans due to lender's learning and the elimination of group leader rewards. Lin et al. (2009) distinguished between structural and relational aspects of networks, and found the relational aspects are consistently significant predictors of the funding probability, interest rates, and ex-post default. Collier and Hampshire (2010) built a theoretical framework for the evaluation and design of community reputation systems. Sergio (2009) also built a model-based clustering method to measures the influence of social interactions in the risk evaluation of a money request.

To help the lenders make better decision, Luo et al. (2011) proposed a data driven investment decision-making framework, which exploits the investor composition of each investment for enhancing decisions making in P2P lending. They revealed that following some investors who have good investment performance in the past will make more correct investment decisions. Katherine & Sergio (2009) examined the behavior of lenders and find that, while there exists high variance in risk-taking between individuals, many transactions represent sub-optimal decisions on the part of lenders. Klafft (2008) showed that following some simple investment rules improves profitability of a portfolio. Kumar (2007) empirically proves that lenders mostly behave rationally and charge appropriate risk premiums for antecedents of loan default. Iyer (2009) also find that lenders are able to use available information to infer a third of the variation in creditworthiness that can be captured by a borrower's personal information. Puro et al. (2010) developed a borrower decision aid system, which helps the borrowers quantify their strategic options, such as starting interest rate, and the amount of loan to request. Wu & Xu (2011) proposed a decision support system based on intelligent agents in P2P Lending to help borrowers getting loan more efficiently, by providing borrowers with individual risk assessment, eligible lender search, lending combination and loan recommendation.

On Prosper, loan transactions between borrowers and lenders are conducted in an information-rich environment. When posting a listing, borrowers also submit their personal portfolios, such as Amount-Requested, Credit-Score, Homeowner, Category (or purpose), debt information and so on. All these information have influence on investors' decision. Li & Qiu (2011) displayed that borrower' decisions, e.g., loan amount, interest rate will determine whether successfully fund loan or not. Herzenstein et al (2008) also explored the determinants of funding successfully, found that borrowers' financial strength and efforts after they post a listing are major factors. The role of financial intermediaries on the P2P online market was analyzed by Berger & Gleisner (2009), which 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. Pope & Sydnor (2008) analyzed discrimination in Prosper, found that loan listings with blacks in the attached picture are 25 to 35 percent less likely to receive funding than those of whites with similar credit profiles. Badunenko et al. (2010) observed that female borrowers pay on average higher interest rates than males at the largest German P2P lending platform, due to female borrowers deliberately offer higher interest rates in anticipation that they would be otherwise discriminated.

The above researches mainly focus on one or part of information of loans. In this paper, we try to investigate all the loan information in a uniform framework. Specifically, we develop a Bayesian network model with all the information in table listing, including the amount of loan to request, interest rate, category of loan, borrowers' credit score, homeowner, dept-to-income-rate, month-loan-payment. Using a large sample of paid or default loan data of Prosper from 2008 to 2011, we construct a Tree Augmented Naïve (TAN) Bayesian network model. Then we experimentally tested this model, using the data in 2012, and compared them to logistic regression, and Luo's method (Luo et al, 2011). Experimental results reveal that TAN Bayesian network can significantly help investors make better investment decisions than other models.

The rest of this paper is organized as follows: The base knowledge of TAN Bayesian network model is provided in Section 2. In Section 3, a Bayesian network model for P2P lending is built and compared to other investment models. Finally, we conclude the work in Section 4.

2 Bayesian Networks

Tree Augmented Naïve (TAN) Bayesian network algorithm (Chow & Lui, 1968) is used mainly for classification. It efficiently creates a simple Bayesian network model, allowing for each predictor to depend on another predictor in addition to the target variable. Its main advantages are its classification accuracy and favorable performance compared with general Bayesian network models. As for the paper, the target variable loan status will be simplified as 1=paid or 0=default two classes, then a listing with portfolios can forecast to classified as 0 or 1 by the Bayesian network model constructed by the past loans.

2.1 TAN Classifier Learning Procedure

Let X = (X1, X2, … , Xn) represent a categorical predictor vector and Y represent the target category, The learning procedure is summarized in Fig. 1 and illustrated in more detail below.

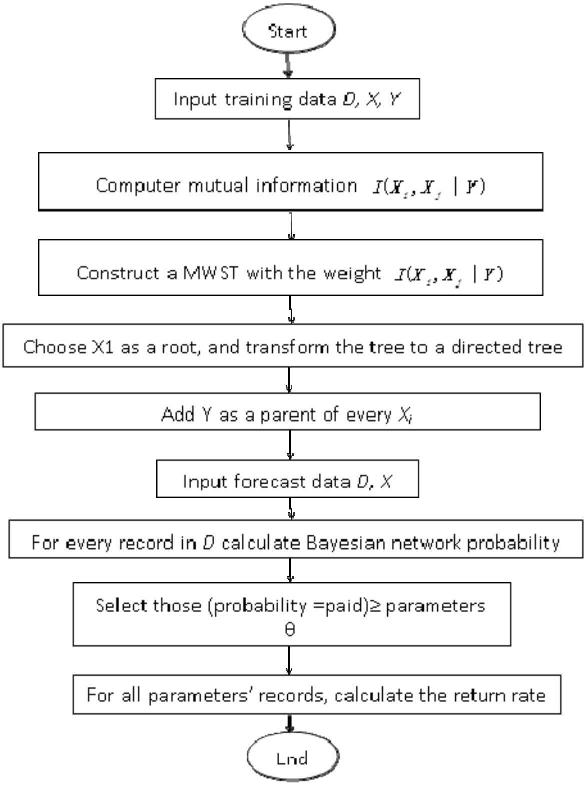
Start
Input training data D, X, Y
Computer mutual information $I(X_i, X_j \mid Y)$
Construct a MWST with the weight $I(X_i, X_j \mid Y)$
Choose X1 as a root, and transform the tree to a directed tree
Add Y as a parent of every X_i
Input forecast data D, X
For every record in D calculate Bayesian network probability
Select those (probability =paid)≥ parameters θ
For all parameters' records, calculate the return rate
Lnd

Fig. 1. Bayesian Network Algorithm

1. Take the training data *D* as input.
2. Compute the conditional mutual information[21] by

$$I(X_i, X_j \mid Y) = \sum_{x_i, x_j, y_k} P(x_i, x_j, y_k) \times \log\left(\frac{p(x_i, x_j \mid y_k)}{p(x_i \mid y_k) p(x_j \mid y_k)}\right) \tag{1}$$

In probability theory and information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two random variables. Learning a tree-like network structure over D by using the structure learning algorithm outlined below.

3. Using Prim's algorithm (Prim, 1957) to construct a maximum weighted spanning tree with the weight of an edge connecting X_i to X_j by $I(X_i, X_j \mid Y)$.
4. Transform the resulting undirected tree to directed one by choosing X_1 as a root node and setting the direction of all edges to be outward from it.
5. Add *Y* as a parent of every X_i where $1 \le i \le n$.

2.2 Probability Calculating

The Bayesian network classifier is a simple classification method, which classifies a case by determining the probability of it belonging to the i-th target category Yi. As investors, we main concern is the loans that will pay back with high probability. These probabilities are calculated as

$$
\begin{aligned}
&P\left(BStatus = 1 \mid X_1 = x_1^j, X_2 = x_2^j, \ldots\ldots, X_n = x_n^j\right) \\
&= \frac{P\left(BStatus = 1, X_1 = x_1^j, X_2 = x_2^j, \ldots\ldots, X_n = x_n^j\right)}{P\left(X_1 = x_1^j, X_2 = x_2^j, \ldots\ldots, X_n = x_n^j\right)} \\
&= \frac{P(BStatus = 1)\prod_{k=1}^{n} P\left(X_k = x_k^j \mid parrent(X_k)\right)}{P\left(X_1 = x_1^j, X_2 = x_2^j, \ldots\ldots, X_n = x_n^j\right)} \qquad (2)
\end{aligned}
$$

3 Experiment Results and Comparison

3.1 Dataset

The dataset used in our experiments is from Prosper.com. Prosper includes six relational data tables, which are Members, Groups, Credit Profile, Listings, Loans and Bids data tables. The Listing table is the most important for our modeling. A Listing is created by a Borrower 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 then after the Listing period ends it will become a Loan.

In our experiments we use seven attributes from the Listing table, which are described in details below.

AmountRequested The amount that the member requested to borrow.

BorrowerRate The rate is computed as the LenderRate + GroupLeaderRewardRate (if applicable) + BankDraftFeeAnnualRate (if applicable).

CreditScore The credit score of the borrower at the time the listing was created

Category The Category is one of the following numerical values : 0 Not available, 1 Debt consolidation, 2 Home improvement, 3 Business, 4 Personal loan, 5 Student use, 6 Auto, 7 Other.

DebtToIncomeRatio The debt to income ratio of the borrower at the time the listing was created.

IsBorrowerHomeowner This attributes specifies whether or not the member is a verified Homeowner.

BidCount The total number of Bids.

3.2 Data Preprocessing

Bayesian nodes deal with discrete data, however, only category (0~7) and IsBorrowerHomeowner (0=false, 1=true) are discrete, the others are continuous values.

Therefore, we digitize these data by width-fixed method. For an attribute, assume that the maximum value of the attribute be $V_{\max}$, and the minimum value of the attribute be $V_{\min}$, we set the separation width to be $d = (V_{\max} - V_{\min}) / 5$, then the attribute is digitized to be 0, 1, 2, 3, 4, 5, when the value belongs to {$V_{\min}$, $V_{\min}$+*d*}, {$V_{\min}$+*d*, $V_{\min}$+2*d*}, 2{$V_{\min}$+2*d*, $V_{\min}$+3*d*}, 3{$V_{\min}$+3*d*, $V_{\min}$+4*d*}, and {$V_{\min}$+4*d*,$V_{\max}$}, respectively.

3.3 Forecast of Return Probability

The data to construct a Bayesian network is selected from the duration from 2008 to 2010. The network aims at predicting the return rate from Jan 1st, 2011 to April 30th, 2011. As we just concern about the people who would pay back as the model classified. The accuracy is calculated as follows:

$$R = \frac{f_{11}}{f_{11} + f_{01}} \tag{3}$$

Where f_{11} is the number of Status=1 and B-Status=1, f_{01} is the number of really Status=0 but B-Status=1. When the Bayesian network model is built, we can calculate the Bayesian probability with the information input. The Bayesian network algorithm is described as Figure3.

Specifically, we select the data from 2008.1 to 2010.1 to build model and use the data of 2011.1 to check the model. Next, add the 2010.2 data to the learning data while the check data is 2011.2, and by this analogy. With the TAN Bayesian method and model, we calculate the Bayesian probability of all the check data. Then the return rate of different probabilities can be calculated, that is, select those B-Status=1 loan that its Bayesian probability is higher than parameterθand compare with the really Status. The result is shown by Table1 and Figure 2.

Table 1. Return Rate of Different Bayesian Network Probability

θ	Real Return Rate	0.5	0.6	0.7	0.8	0.9
2011.1	0.74	0.78	0.81	0.79	0.90	1.00
2011.2	0.84	0.85	0.86	0.88	0.95	1.00
2011.3	0.79	0.81	0.82	0.84	0.89	1.00
2011.4	0.75	0.75	0.76	0.78	0.91	0.95

In P2P lending, our investment decision model ranks loans, from the best to the worst, according to the probability by Bayesian network. Investors can choose the top ones as the candidate set. We find empirical evidence to show the effectiveness of our model and the influence of different parameters. From Figure 3, the paid loans' distributions of Bayesian network probability is markedly higher than the default loans.

In Figure 4, we compare the rate of return by our model against three baselines, the average rate of return by investing on all loans, logistic regression model, investor composition method by Luo et al.(2011).We can find that, the higher γ we choose, the higher return rate of the candidates chosen by our model has than others, whereγis the top probability loans to invest on.

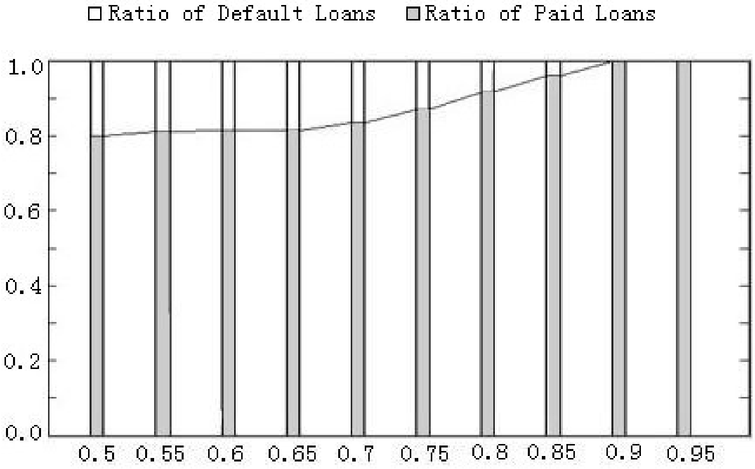

Fig. 2. Ratio of Loans Status by Bayesian Network Probability

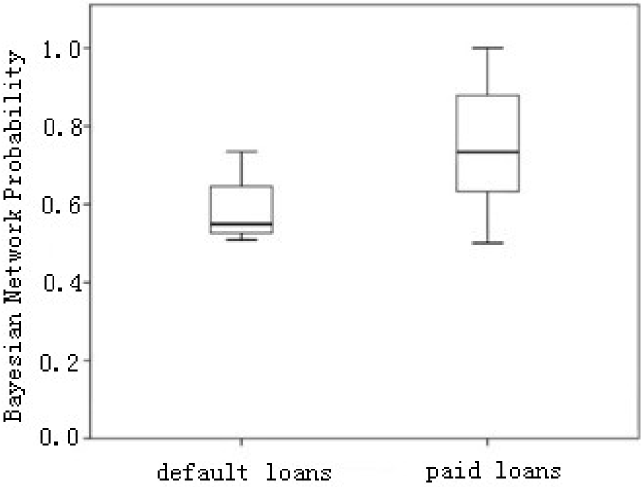

Fig. 3. Distributions of Bayesian Network Probability by Loans Status

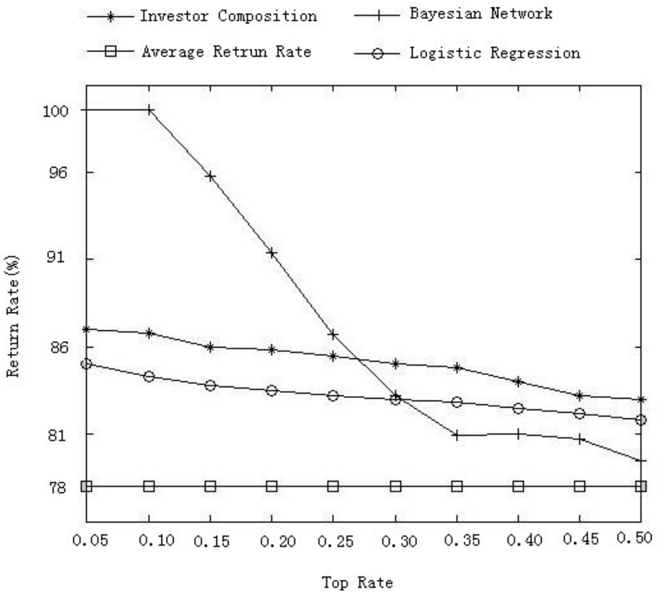

Fig. 4. A Comparision of Return Rate

3.4 Different Time Span of Training Data as an Indicator

Bayesian probability is learning from experiment data and expert knowledge. Whether the larger the learning data is the better? We choose the data from 2011.1 to 2011.4 as check data, and the training data is N months data that one year before the check data. That is, we first choose 2011.1 as the check data, then the three months training data is from 2009.11 to 2010.1, the four months training data is from 2009.10 to 2010.1, and so on. After doing all the experiments, we put the all three months' check data together, and compare with the really data, the result is shown by Table2. From the table, the different time span of training data doesn't make any difference. However, the table shows that different Bayesian probabilities may make significant different.

Table 2. Return Rate of Different Time Span of Training Data

Training data	Return Rate	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
3months	0.78	0.79	0.81	0.82	0.84	0.84	0.87	0.91	0.92	1.00	1.00
4months	0.78	0.82	0.82	0.82	0.84	0.84	0.86	0.87	0.89	1.00	1.00
5months	0.78	0.80	0.82	0.83	0.83	0.83	0.85	0.89	0.91	1.00	1.00
6months	0.78	0.80	0.81	0.82	0.82	0.84	0.86	0.88	0.90	1.00	1.00
7months	0.78	0.79	0.81	0.82	0.82	0.83	0.87	0.88	0.91	1.00	1.00
8months	0.78	0.79	0.81	0.83	0.84	0.87	0.87	0.90	0.91	1.00	1.00
9months	0.78	0.79	0.80	0.81	0.83	0.84	0.86	0.91	0.92	1.00	1.00

3.5 The Newest Bayesian Network Model

With those experiments, Bayesian network model is proved as an effective model for P2P lending loan. We use all the information and data in table loan since 2008 to build a new model. The model is shown in Fig. 5.

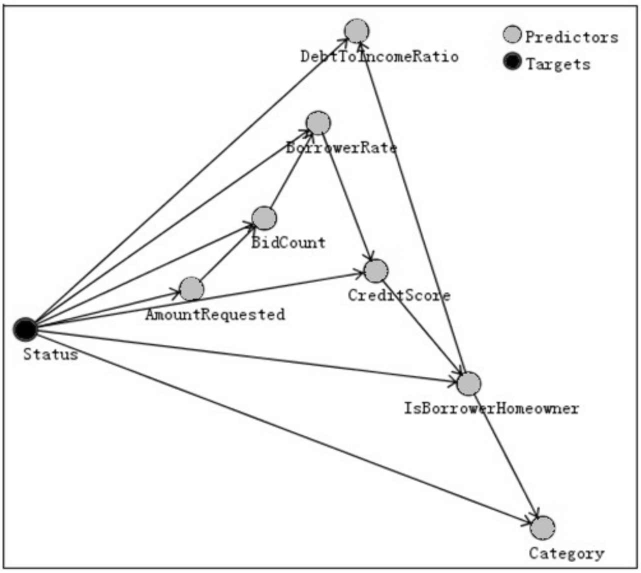

Fig. 5. A Bayesian Network Model

4 Conclusions

In this paper, we build Bayesian network model with all the borrower information and loan information in the table listing. First, we calculate the mutual information of every two variables and create a maximum weighted spanning tree (MWST) with them. When the weight matrix is created, the MWST algorithm gives an undirected tree that can be oriented with the choice of a root. A Bayesian network model is built when we add Status as the parent of every node. Then we check the model by the data a year later, if the Bayesian probability is higher, the rate of return is higher too. Experimental results reveal that Bayesian network model can improve investment performances.

References

1. Wang, H., Greiner, M., Aronson, J.: People-to-People Lending: The Emerging E-Commerce Transformation of a Financial Market. Value Creation in E-Business Management 36, 182–195 (2009)
2. Prosper Marketplace, http://www.prosper.com

3. Freedman, S., Jin, G.: Do social networks solve information problems for peer-to-peer lending? Evidence from prosper.com, NET Institute Working Papers No. 08-43, Indiana University (2008)
4. Lin, M., Nagpurnan, R.P., Viswanathan, S.: Judging Borrowers By The Company They Keep: Social Networks and Adverse Selection in Online Peer-to-Peer Lending, Western Finance Association 2009 Annual Meeting Paper, University of Maryland, College Park, MD (2009)
5. Collier, B., Hampshire, R.: Sending mixed signals: multilevel reputation effects in peer-to-peer lending markets. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, New York, pp. 197–206 (2010)
6. Sergio, H.: Social interactions in P2P lending. In: Proceeding SNA-KDD 2009 Proceedings of the 3rd Workshop on Social Network Mining and Analysis, New York, Article No. 3 (2009)
7. Luo, C., Xiong, H., Zhou, W., et al.: Enhancing investment decisions in P2P lending: an investor composition perspective. In: KDD 2011(Knowledge Discovery in Databases), San Diego, California, USA, pp. 292–300 (2011)
8. Krumme, K., Sergio, H.: Do Lenders Make Optimal Decisions in a Peer-to-Peer Network? In: Proceedings of the IEEE/WIC/ACM Int'l Joint Conf. on Web Intelligence and Intelligent Agent Technology, Milan, Italy, pp. 124–127 (2009)
9. Klafft, M.: Online Peer-to-Peer Lending: A Lenders' Perspective. In: Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, Las Vegas, pp. 371–375 (2008)
10. Kumar, S.: Bank of one: empirical analysis of peer-to-peer financial marketplaces. In: Americas Conference on Information Systems, Association for Information System Electronic Library, Keystone, Colorado (2007)
11. Iyer, R., Khwaja, A., Luttmer, E.: Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? Working Paper, National Bureau of Economic Research (2009)
12. Puro, L., Teich, J., Wallenius, H., Wallenius, J.: Borrower decision aid for people-to-people lending. Decision Support System 49, 52–60 (2010)
13. Wu, J., Xu, Y.: A Decision Support System for Borrower's Loan in P2P Lending. Journal of Computers 6(6) (June 2011)
14. Li, S., Qiu, J.: Do Borrowers Make Homogeneous Decisions in Online P2P Lending Market? An Empirical Study of PPDai in China. In: Service Systems and Service Management (ICSSSM), Tianjin, China, pp. 1–6 (2011)
15. Herzenstein, M., Andres, R., Dholakia, U., Lyandres, E.: The Democratization of Personal Consumer Loans? Determinants of Success in Online Peer-To-Peer Lending Communities. Working Paper, University of Delaware (2008)
16. Berger, S., Gleisner, F.: Emergence of financial intermediaries in electronic markets: The case of online p2p lending. BuR Business Research Journal 2(1), 39–65 (2009)
17. Pope, D., Justin, S.: What's in a picture? Evidence of discrimination from Prosper.com. Working Paper, University of Pennsylvania (2008)
18. Badunenko, N.B., Schäfer, D.: Are women more credit-constrained than men?– Evidence from a rising credit market. JEL Classification: G21, J16 (2010)
19. Chow, C., Lui, C.: Approximating discrete probability distributions with dependence trees. IEEE Trans. on Info. Theory 14, 462–467 (1968)
20. Prim, R.: Shortest connection networks and some generalisations. Bell System Technical Journal 36, 1389–1401 (1957)
21. Friedman, N.: Learning belief networks in the presence of missing values and hidden variables. In: Proceedings of the 14th International Conference on Machine Learning, pp. 125–133. Morgan Kaufmann (1997)