Credit Scoring by Incorporating Dynamic Network Information

Yibei Li, Ximei Wang, Boualem Djehiche, Xiaoming Hu
Department of Mathematics, KTH Royal Institute of Technology, Sweden
{yibei, ximei, boualem, hu}@kth.se

Abstract

In this paper, the credit scoring problem is studied by incorporating network information, where the advantages of such incorporation are investigated in two scenarios. Firstly, a Bayesian optimal filter is proposed to provide a prediction for lenders assuming that published credit scores are estimated merely from structured individual data. Such prediction is used as a monitoring indicator for the risk warning in lenders’ future financial decisions. Secondly, we further propose a recursive Bayes estimator to improve the accuracy of credit scoring estimation by incorporating the dynamic interaction topology of clients as well. It is shown that under the proposed evolution framework, the designed estimator has a higher precision than any efficient estimator, and the mean square errors are strictly smaller than the Cramér–Rao lower bound for clients within a certain range of scores. Finally, simulation results for a specific case illustrate the effectiveness and feasibility of the proposed methods.

Keywords: credit scoring, network information, Bayesian filtering, average risk minimization

JEL Classification: G20, G32, G15
1. Introduction

Evaluating or estimating the credit scores according to clients’ financial background when one applies a loan is important for lenders such as banks or lending institutions (Beaver 1966, Thomas et al. 2002). It can not only decide whether the clients can get the loan or not but also closely influence the price for the loan such as the lending duration and lending rates (Emekter et al. 2015).

Most of the current credit scoring methods are based on statistical approaches which consider multi-dimension attribute information about the applicant of the loan. For instance, the logistic regression model, the ordered probit model, artificial neural network (ANN) algorithms, support vector machines (SVM) are widely used for predicting the probability of default for clients (Byanjankar et al. 2015, Xia et al. 2018, Ignatius et al. 2018, Wang et al. 2019b). While it is not always the “best” model in credit scoring problem (West 2000). According to a large numbers of studies for credit scoring modeling, there is no consistent conclusion on which method is the most accurate method to be used. Sometimes there are conflicts comparing the findings in different studies (Baesens et al. 2003, Miguéis et al. 2013).

Usually, around 10 to 30 attributes which generally consist of the key financial measures of clients such as their income or debt level, credit history and payments frequency are considered as inputs in the credit scoring models. Moreover, different countries in the world have different ways to handle the credit scoring estimation. For example, in the US, the well known credit scores calculated by the Fair Isaac Corporation (FICO) is composed by 35% payment history, 30% debt burden, 15% length of credit history time in file, 10% types of credit used and 10% recent searched for credit (Arya et al. 2013). In the UK, credit scores of the client are based on factors payment history, age of accounts, and credit utilization. Furthermore, getting on the electoral register to vote (or explaining why you’re not eligible to) can help improve one’s credit scorings (Hand and Henley 1997). In Japan, credit for consumer is always based on factors like length of employment and salary (Mohammadi and Zangeneh 2016).

However, in the past few years, the power of data, algorithms and technologies have led a dramatic change in credit scoring problem (Ntwiga and Weke 2016, Campbell-Verduyn et al. 2017). Network information is becoming popularly used in credit scoring since it is about the relations (De Benedictis et al. (2014)) which can reflect some economic activities. Furthermore, it is believed that network information among customers such as social networks, financial trading networks can be considered as an effective way for improving credit ratings for clients (Herrero-Lopez 2009, De Benedictis et al. 2014). Governments as well as more and more start-up companies or lending platforms are intended to rely on network based data to evaluate the creditworthiness for clients. For start-up companies, social or financial network information helps them to picture the customers in details (Rusli 2013, Freedman and Jin 2017). Network profiles of the clients such as the employment relationship history, number of friends, or financial transfer activities may influence or determine the credit scorings for each individual (Bolhuis 2015, Wang et al. 2019a). While there are several issues in the existing literatures. Firstly, most of the studies are data-drive decision making (Masyutin 2015). Freedman and Jin (2017) use the data from Prosper.com, which is the largest peer-to-peer consumer lending platform to examine that social networks facilitate online markets lending business. De Cnudde et al. (2015) use data from Lenddo, where they use social network data such as Facebook, Twitter, LinkedIn to provide unique insights about clients’ creditworthiness. Secondly, existing theoretical work about network information mainly relies on static analysis (Wozabal and Hochreiter 2012, Wei et al. 2015), which cannot ensure the accuracy of the credit updating in a dynamic model.

Motivated by the growing interest in using network information in practice, our study analyzes the importance of network information from a theoretical point of view. A dynamic interaction network among the clients is constructed according to homogeneous preference based on others credit assessments reported by the credit bureau (Zeng and Xie 2008). It is shown that such correlation reflected by the network formulation can be developed to further improve the scoring
precision. In particular, two scenarios are considered. Firstly, when the publishing of the scores is merely based on individual attributes, an optimal Bayesian filter is designed and the history network observations are used to make credit predictions for each client. Such prediction serves as a key monitoring indicator for the risk warning and management in lenders. Furthermore, a recursive Bayes estimator is proposed to improve the accuracy of the score publishing by incorporating the dynamic network topology as well. A one-step optimal estimate is given through average risk minimization at each period. It is shown that under the proposed evolution framework, the designed estimator has a higher precision than any other efficient estimator *, and the mean square errors are strictly smaller than the Cramér–Rao lower bound for clients within a certain range of scores. Finally, a special case is considered where the true credits are assumed to be uniformly distributed. Simulation results illustrate the effectiveness and feasibility of the proposed methods.

The rest of this paper is organized as follows. In Section 2, we give an overview of the mathematical modeling for credit scoring problem with network information. In Section 3, a Markov process is used for prediction and the corresponding recursive Bayesian filter is derived to estimate individual credit scores based on history observations. Section 4 proposes an online scoring framework to improve the score accuracy recursively. In section 5, we present a simulation study for the investigation of the proposed algorithm. Finally, conclusions and future work for credit scoring incorporating the network information are discussed in Section 6.

2. Mathematical model of credit scoring

2.1. Network modeling

In this paper, we consider the credit scoring problem for $N$ clients. Let $x_i \in \mathbb{R}$ denote the true credit of client $i$ ($i = 1, \ldots, N$), which evolves according to a linear model

$$x_i(t + 1) = a(t)x_i(t) + b(t)u_i(t) + w_i(t), \quad (1)$$

where $a(t) \in (0, 1]$ and $b(t) \in \mathbb{R}$ are given constants for any $t \geq 0$, $u_i(t)$ denotes the asset change of client $i$ at time $t$, and $w_i(t) \sim \mathcal{N}(0, Q_t)$ is the uncertainty.

At each time, each client $i$ establishes financial connections based on his own credit $x_i$ and others’ credit assessment $y_j$ reported by the lender. The network is formed based on homogeneous preference, which is inspired by Wei et al. (2015). Clients prefer to form connection with people with similar credit levels. Each pair of clients meet with a probability $\nu > 0$. Client $i$ forms a connection with client $j$ if and only if they have met and

$$m > |x_i - y_j|,$$

where $m$ is a random variable denoting the match threshold.

We model the financial network between agents by a time-varying graph denoted by $G_t = (\mathcal{V}, \mathcal{E}_t)$, where the vertex set $\mathcal{V} = \{1, 2, \ldots, N\}$ denotes the clients in the network and $\mathcal{E}_t \subset \mathcal{V} \times \mathcal{V}$ is the edge set at time $t$. We say that at time $t$ agent $j$ is a neighbor of agent $i$ if $(i, j) \in \mathcal{E}_t$, and the set of neighbors of agent $i$ is denoted by $\mathcal{N}_i(t) = \{j : (i, j) \in \mathcal{E}_t\}$. We use $n_i(t)$ to denote the number of neighbors of client $i$. In the remaining part of the paper we use $\{g_{ij}(t)\}_{i,j=1}^N$ to denote the elements of $\mathcal{E}_t$. And we define that

$$g_{ij}(t) = \begin{cases} 1 & \text{if } (i, j) \in \mathcal{E}_t, \\ 0 & \text{otherwise}. \end{cases}$$

*An efficient estimator is an unbiased estimator whose variance reaches the Cramér–Rao lower bound.
Hence client \( i \) forms a connection with client \( j \) with probability

\[
\Pr(g_{ij}(t) = 1) = \Pr(m > |x_i - y_j| \mid i \text{ and } j \text{ meet}) \\
= \Pr(m > |x_i - y_j|) \Pr(i \text{ and } j \text{ meet}).
\]  

We choose the distribution of \( m \) according to the following criteria:

(i) \( m \) is positive;
(ii) Clients are connected based on homophily preference, which means that \( \Pr(m > |x_i - y_j|) \) is larger with smaller credit difference \( |x_i - y_j| \).

Based on the above criteria, we choose \( m \) from a Rayleigh distribution with parameter \( k \). Later in this paper we will see that such choice of an exponential family not only satisfies the two criteria, but also makes it possible to derive some useful analytic expressions in Bayesian inference. Without loss of generality, we normalize the model by taking \( k = 1 \) and scale the credit scores to a positive interval, i.e. \( x_i(t) \in [0, M] \) for any \( t \geq 0 \). Then it holds that

\[
\Pr(g_{ij}(t) = 1) = \nu e^{-\frac{(x_i(t) - y_j(t))^2}{2}}.
\]

Note that in the most of the existing methodologies, \( y_j \) is derived merely based on structured personal data. Later in the paper, we will show that the network information can also be incorporated to achieve a higher estimation accuracy.

Unlike Wei et al. (2015), here we assume that at each time the lender only use a partial observation of the network. Assume that for each client \( i \), the lender has an observation for its financial network by \( g_i(t) = \{g_{ij}(t)\}_{j \in N_i(t)} \). Note that it is more reasonable to use information only from neighbors, since including the whole network as in Wei et al. (2015) is quite time-consuming and computationally heavy.

2.2. Problem formulation

Up until recently, the lender assesses a client’s creditworthiness only based on his own financial history and individual abilities such as salary and education level. As is shown in the previous part, the economic engagement between agents is closely related to their credits homogeneity. Therefore, such correlation between the financial network and individual credits can be used to improve the scoring accuracy. In this paper, the following two scenarios are considered:

(i) Risk prediction
   The true credits of clients evolve according to (1). At each period, the lender publishes an estimated credit score for each agent only based on his own properties. Meanwhile, a risk evaluator is observing that dynamic process, whose task is to provide a more precise prediction about the credit scores based on history observations of published scores and financial networks. Such prediction can then serve as a suggestion for the lender on future financing practices.

(ii) Recursive accuracy improvement based on dynamic interaction
   As for a further step, the credit estimation process and the evolution of the network are considered in an integrated manner. At each period, the lender publishes an optimal estimate for each score based on the current financial network as well as individual assets, which is then used to form a new network in the next period. The score estimation is coupled with the financial network in the sense that it is not only determined by the current network, but also will influence the formation of new connections next time. The estimation accuracy can then be improved iteratively through the dynamic interaction between the lender and clients.
3. Optimal Bayesian filtering

In this section, the risk prediction problem is studied. The credit scoring process is modeled as a Markov process, and the corresponding recursive Bayesian filter is derived to estimate individual credit scores based on history observations.

We assume that for client \( i \), based on his individual assets, the lender can only have a noisy observation of his credit score:

\[
y_i(t) = x_i(t) + v_i(t),
\]

where \( v_i(t) \sim \mathcal{N}(0, R_t) \) denotes the observation noise.

The filtering model can be given by

\[
\begin{cases}
x_i(t + 1) = a(t)x_i(t) + b(t)u_i(t) + w_i(t), \\
y_i(t) = x_i(t) + v_i(t), \\
g_{ij}(t) \sim \text{Ber}\left(e^{-\frac{(y_i(t) - y_j(t))^2}{2}}\right), \quad j \in \mathcal{N}_i(t),
\end{cases}
\]

where \( w_i(t) \sim \mathcal{N}(0, Q_t) \) and \( v_i(t) \sim \mathcal{N}(0, R_t) \) are independent Gaussian process with \( R_t > 0 \) and \( Q_t \geq 0 \). We denote \( z_i(t) = [y_i(t), g_{ij}(t)]_{j \in \mathcal{N}_i(t)} \in \mathbb{R}^{u_i(t) + 1} \) as the observation of client \( i \) by the risk evaluator at the time \( t \). Let \( Z_{i,t} = [z_i(0), ..., z_i(t)] \) denote the sequence of observation history.

**Theorem 3.1** Assume \( x_i(0) \sim \mathcal{N}(\bar{x}_0, P_{i,0}) \), and

\[
\mathbb{E}[w(t)w(t + \tau)] = 0, \mathbb{E}[v(t)w(t + \tau)] = 0, \quad \text{for any} \quad \tau \neq 0,
\]

\[
\mathbb{E}[w(t)x_0] = 0, \mathbb{E}[v(t)x_0] = 0, \quad \text{for any} \quad t \geq 0.
\]

Then \( x_i(t) | Z_{i,t-1} \) and \( x_i(t) | Z_{i,t} \) are Gaussian with

\[
x_i(t) | Z_{i,t-1} \sim \mathcal{N}(\hat{x}_i(t|t - 1), P_i(t|t - 1))
\]

\[
x_i(t) | Z_{i,t} \sim \mathcal{N}(\hat{x}_i(t|t), P_i(t|t)).
\]

where

\[
\begin{align*}
\hat{x}_i(t|t - 1) &= a(t)\hat{x}_i(t - 1|t - 1) + bu_i(t - 1), \\
P_i(t|t - 1) &= a(t)^2P_i(t - 1|t - 1) + Q_{i,t-1}, \\
\hat{x}_i(t|t) &= \hat{x}_i(t|t - 1) + K_{i,t}(y_i(t) - \hat{x}_i(t|t - 1)) + H_{i,t}\sum_{j \in \mathcal{N}_i(t)}(y_j(t) - \hat{x}_i(t|t - 1)), \\
P_i(t|t) &= (1 - K_{i,t} - n_i(t)H_{i,t})P_i(t|t - 1),
\end{align*}
\]

\( K_{i,t} \) and \( H_{i,t} \) are given by

\[
K_{i,t} = P_i(t|t - 1)/(R_t + P_i(t|t - 1) + n_i(t)R_tP_i(t|t - 1)),
\]

\[
H_{i,t} = P_i(t|t - 1)R_t/(R_t + P_i(t|t - 1) + n_i(t)R_tP_i(t|t - 1)).
\]

**Proof.** Due to the Markov property of process (4), by Bayesian rule it holds that

\[
p(x_i(t + 1)|Z_{i,t+1}) = \frac{p(z_i(t + 1)|x_i(t + 1))p(x_i(t + 1)|Z_{i,t})}{p(z_i(t + 1)|Z_{i,t})}
\]

(8)
Since $w(t)$ and $v(t)$ are Gaussian, we have

$$p(y_i(t)|x_i(t)) = p_v(y_i(t) - x_i(t)) = \mathcal{N}(x_i(t), R_i)$$  \hspace{1cm} (9a)

$$p(x_i(t+1)|x_i(t), Z_{i,t}) = p_{x_i}(x_i(t+1) - ax_i(t) - bu_i(t)) = \mathcal{N}(ax_i(t) + bu_i(t), Q_i).$$  \hspace{1cm} (9b)

The Bernoulli distribution of $g_i(t)$ is given by

$$p(g_i(t)|x_i(t), y_i(t)) = \prod_{j \in \mathcal{N}(i)} \nu e^{-\frac{(x_i(t) - y_{ij}(t))^2}{2}}.\hspace{1cm} (10)$$

Firstly, we prove (6) by induction. To begin with, we notice that

$$x_i(0)|Z_{i,0} \sim \mathcal{N}(\hat{x}_i(0|0), P_i(0|0)),$$

with $\hat{x}_i(0|0) = \bar{x}_0$ and $P_i(0|0) = P_i$. Assume that at time $t$, it holds that

$$x_i(t)|Z_{i,t} \sim \mathcal{N}(\hat{x}_i(t|t), P_i(t|t)) \hspace{1cm} (11)$$

for some $\hat{x}_i(t|t)$ and $P_i(t|t)$.

Next, we want to show that the above equation also holds at time $t+1$.

Notice that $p(x_i(t+1)|Z_{i,t})$ can be computed by

$$p(x_i(t+1)|Z_{i,t}) = \int p(x_i(t+1)|x_i(t), Z_{i,t})p(x_i(t)|Z_{i,t})dx_i(t).$$

Then by (9b) and (11), we obtain

$$p(x_i(t+1)|Z_{i,t}) = \mathcal{N}(\hat{x}_i(t+1|t), P_i(t+1|t)), \hspace{1cm} (12)$$

where

$$\hat{x}_i(t+1|t) = a(t)\hat{x}_i(t|t) + bu_i(t), \hspace{1cm} P_i(t+1|t) = a(t)^2P_i(t|t) + Q_i.$$

In addition,

$$p(y_i(t+1), g_i(t+1)|x_i(t+1)) = p(y_i(t+1)|x_i(t+1))p(g_i(t+1)|y_i(t+1), x_i(t+1)).$$

Then by (8), $p(x_i(t+1)|Z_{i,t+1})$ is given by

$$p(x_i(t+1)|Z_{i,t+1}) \propto e^{-\frac{(y_i(t+1)-x_i(t+1))^2}{2R_{i+1}}} \cdot e^{-\frac{(x_i(t+1)-\hat{x}_i(t+1|t))^2}{2P_{i+1}(t+1|t)}} \cdot \prod_{j \in \mathcal{N}(i+1)} e^{-\frac{(x_j(t+1)-y_{ij(t+1)})^2}{2}}.$$
Through some simple computation, it is obvious that \( x_i(t + 1)|Z_{i,t+1} \) is Gaussian, i.e.

\[
x_i(t + 1)|Z_{i,t+1} \sim \mathcal{N}(\hat{x}_i(t + 1|t + 1), P_i(t + 1|t + 1)),
\]

where \( x_i(t + 1|t + 1) \) and \( P_i(t + 1|t + 1) \) coincide with (7).

Hence we have proved (6), and (5) can be then derived as shown in (12).

Based on Theorem 3.1, a recursive Bayesian filtering algorithm is then designed to estimate the credit scores iteratively at time steps \( t = 1, \ldots, T \). Here the mean-squared error (MSE) is chosen as the criterion to derive the optimal filter. The filtering equations can then be derived based on (5) to (7), where the MMSE estimator is chosen as the conditional mean, i.e.

\[
MMSE(\hat{x}_i(t)) = \mathbb{E}[x_i(t) | Z_{i,t}] = \hat{x}_i(t|t).
\]

Similar to the Kalman filter, at each time step the proposed algorithm is executed in two phases: “predict” and “update”, which is given in Algorithm 1.

**Algorithm 1** Recursive Bayesian filtering algorithm

**Initialize:** \( \hat{x}_i(0|0) = \bar{x}_0, P_i(0|0) = P_i, 0 \)

1: for \( t = 0 : T \) do
2:     for \( i = 1 : N \) do
3:         **Predict:**
4:             \( \hat{x}_i(t + 1|t) = a(t)\hat{x}_i(t|t) + b(t)u_i(t) \)
5:             \( P_i(t + 1|t) = a(t)^2P_i(t|t) + Q_t \)
6:         **Update:**
7:             \( K_{i,t+1} = P_i(t + 1|t)/(R_t + P_i(t + 1|t) + n_i(t + 1)R_{t+1}P_i(t + 1|t)) \)
8:             \( H_{i,t+1} = P_i(t + 1|t)R_t/(R_t + P_i(t + 1|t) + n_i(t + 1)R_{t+1}P_i(t + 1|t)) \)
9:             \( \hat{x}_i(t + 1|t + 1) = \hat{x}_i(t + 1|t) + \sum_{j \in N_i(t + 1)} K_{i,t+1}(y_i(t + 1) - \hat{x}_i(t + 1|t)) \)
10: \( P_i(t + 1|t + 1) = (1 - K_{i,t+1} - n_i(t + 1)H_{i,t+1})P_i(t + 1|t) \)
11: end for
12: end for

It is obvious that the precision of the derived MMSE estimator is strictly higher than that of the observation by the lender, i.e.

\[
P_i(t|t) = \frac{R_tP_i(t|t - 1) - R_tP_i(t|t - 1)R_t}{n_i(t)P_i(t|t - 1)R_t + R_t + P_i(t|t - 1)} < R_t.
\]

Hence, it is reasonable to adopt the proposed filter for the risk evaluator to provide a useful credit prediction for the lender. At each time, such a prediction can then be used as a reference for the lender to make long-term financial decisions. For example, the lender may consider lowering the loan of a company if its credit score is predicted to decrease. Furthermore, since we have shown that taking the network information into account is able to improve the scoring accuracy, it is natural to speculate that the lender can also incorporate the network information in its own assessment, which will be studied in the next section.
4. Recursive scoring based on dynamic interaction

4.1. Dynamic scoring framework

In this part, an online scoring framework is used to recursively improve the score accuracy based on the dynamic interaction between the lender and the clients.

During each period, the lender publishes the current score prediction, based on which the clients then form a new network with homogeneous preference. Then at the end of the period, the lender updates a new estimate for current individual credit scores based on the observation of the network, which will be used to make new predictions at the beginning of the next period.

Here we use \( \bar{x}_i(t) \) and \( \hat{x}_i(t) \) to denote the prediction and corrected estimation for the credits of agent \( i \) at period \( t \). In each period, the interaction process mentioned above is divided into the following four steps, as shown in Fig. (1).

(i) **True credits update:**
   The true credits of the clients evolve according to the system model:
   \[
   x_i(t) = a(t-1)x_i(t-1) + bu_i(t-1) + w_i(t-1) \quad (13)
   \]

(ii) **Publishing by the lender:**
   The lender publishes a new prediction of credit scores based on individual financial contributions:
   \[
   \bar{x}_i(t) = a(t-1)\hat{x}_i(t-1) + bu_i(t-1) \quad (14)
   \]

(iii) **Network formation:**
   Each agent \( i \) forms a new network \( g_i(t) \) based on its own credit \( x_i(t) \) and others’ scores \( \bar{x}_j(t) \) from the lender.

(iv) **Score correction:**
   The lender updates a new credit estimation \( \hat{x}_i(t) \) for each agent based on the observation of the current network \( g_i(t) \).

![Figure 1.: Interaction process at period \( t \).](image)

Note that at the beginning, the lender gives the initial estimate \( \hat{x}_i(0) \) only based on individual financial history of clients. In accordance with Section 3, here we also assume that \( \hat{x}_i(0) \) is unbiased and Gaussian.

**Assumption 1** For each agent \( i \), the initial credit estimate \( \hat{x}_i(0) \) is Gaussian with \( \hat{x}_i(0) \sim N(x_i(0), P_i(0)) \).
4.2. One-step optimal estimator

In this part, the formula for the score correction step is investigated. During each period, a one-step optimal estimator \( \hat{x}_i(t) \) is updated based on the observation of the current network \( g_i(t) \).

It is well-known that efficient estimators are uniformly optimal in the class of unbiased estimators, which realize the Cramér–Rao lower bound (CRLB) (Lehmann and Casella 2006). Therefore, in order to further improve the estimation precision (the reciprocal of the variance), biased estimators can be considered to realize lower variance. In this part, the Bayes estimate is derived through average risk minimization, which is able to realize a lower MSE than efficient estimators for clients in the middle class.

In order to derive the optimal estimate, we define an average risk measure

\[
 r(\hat{x}_i, \alpha) = \int_{-\infty}^{+\infty} E[L(x_i, \hat{x}_i)] \alpha(x_i) \, dx_i, \tag{15}
\]

where \( L(x_i, \hat{x}_i) \) is some risk function and \( \alpha(x_i) \) is a positive weighting function indicating how important it is to have a low risk for different values of \( x_i \).

Without loss of generality, \( \alpha(x_i) \) can be normalized by

\[
 \int_{-\infty}^{+\infty} \alpha(x_i) \, dx_i = 1.
\]

Then the optimal estimator with weighting function \( \alpha(x_i) \) is obtained by minimizing the average risk in (15), i.e.

\[
 \hat{x}_{i, \alpha}(g_i) = \arg\min_{\hat{x}_i} r(\hat{x}_i, \alpha), \tag{16}
\]

which is also known as the Bayes estimator.

It is easy to see that when we consider the quadratic cost \( L(x_i, \hat{x}_i) = ||x_i - \hat{x}_i||^2 \), (16) can be actually regarded as the MSE estimator of \( x_i \) given observation \( G_i \). It is also quite inspiring to consider the Bayesian interpretation of (16). By considering \( x_i \) as the outcome of a random variable \( X_i \) whose prior distribution is given by \( \alpha(x_i) \), the solution to (16) turns out to be the posterior mean, i.e.

\[
 \hat{x}_{i, \alpha}(g_i) = \mathbb{E}[x_i | G_i = g_i]. \tag{17}
\]

Under the framework of recursive scoring based on dynamic interaction, at each period, the “prior knowledge” \( \alpha(x_i) \) is chosen as the pdf of the predicted score \( \bar{x}_i \).

**Theorem 4.1** Under Assumption 1, the posterior \( x_i(t) | G_i(t) = g_i(t) \) in (17) is Gaussian at each time step \( t \geq 0 \), i.e. \( x_i(t) | g_i(t) \sim N(\bar{x}_i(t), \bar{P}_i(t)) \), with

\[
 \bar{x}_i(t) = \bar{x}_i(t) + \frac{\bar{P}_i(t)}{1 + \bar{P}_i(t)n_i(t)} \sum_{j \in \mathcal{N}_i(t)} (\bar{x}_j(t) - \bar{x}_i(t)),
\]

\[
 \bar{P}_i(t) = \frac{\bar{P}_i(t)}{1 + \bar{P}_i(t)n_i(t)}, \tag{18}
\]

9
\[
\vec{x}_i(t) = a(t-1)\vec{x}_i(t-1) + bu_i(t-1), \\
\vec{P}_i(t) = a^2\vec{P}_i(t-1) + Q_{t-1}.
\]

**Proof.** Here we prove the \( x_i(t) \mid g_i(t) \) is Gaussian by induction.

To begin with, we have \( \hat{x}_i(0) \sim N(x_i(0), \hat{P}_i(0)) \).

We assume that at time \( t-1 \), it holds that:
\[
x_i(t-1) \mid g_i(t-1) \sim N(\hat{x}_i(t-1), \hat{P}_i(t-1)),
\]
for some \( \hat{x}_i(t-1) \) and \( \hat{P}_i(t-1) \).

Then we want to show that the above also holds at time \( t \). By (13) and (14), it is easy to see that the prior distribution of \( x_i(t) \) is Gaussian with \( N(\bar{x}_i(t), \bar{P}_i(t)) \), where \( \bar{x}_i(t) \) and \( \bar{P}_i(t) \) are given by (19) and (20) respectively. In the sequel, “(t)” is omitted for the sake of brevity.

By Bayesian rule, the probability density function (pdf) of \( x_i \mid g_i \) is given by
\[
p(x_i \mid g_i) = \frac{\alpha(x_i)p(g_i \mid x_i)}{p(g_i)}.
\]

Since we only use neighbors’ information in \( g_i \), the Bernoulli distribution of \( g_i \) is given by
\[
p(g_i \mid x_i) = \prod_{j \in \mathcal{N}_i(t)} \nu e^{-\frac{(x_i - \bar{x}_j)^2}{2}}.
\]

Hence, \( p(x_i \mid g_i) \) can be computed by
\[
p(x_i \mid g_i) \propto e^{-\frac{(x_i - \bar{x}_j)^2}{2\bar{P}_i(t)}} \cdot \prod_{j \in \mathcal{N}_i(t)} e^{-\frac{(x_i - \bar{x}_j)^2}{2}}.
\]

Therefore, \( x_i(t) \mid g_i(t) \) must be Gaussian with
\[
x_i(t) \mid g_i(t) \sim N(\hat{x}_i(t), \hat{P}_i(t)),
\]
where \( \hat{x}_i(t) \) and \( \hat{P}_i(t) \) are given in (18), which can be obtained by comparing the coefficients.

In conclusion, at each time \( t \), given the observation of the financial network, the credit estimation of each agent \( i \) is updated by an optimal estimator
\[
\hat{x}_{i,\alpha}(g_i) = E[x_i(t) \mid g_i(t)] = \hat{x}_i(t),
\]
which is given by (18) and is used in the credit correction step in Fig. 1.

4.3. **Performance analysis**

4.3.1. **Estimation error and precision.**

The MSE is a widely-used criterion to analyze the performance of estimators, which can be split
$MSE[\hat{x}_i] = E[\|\hat{x}_i - x_i\|^2]$

$= E[\|\hat{x}_i - E[\hat{x}_i]\|^2 + E[\hat{x}_i] - x_i\|^2$

$= Var(\hat{x}_i) + Bias(\hat{x}_i)\]$

(22)

where $Var(\hat{x}_i)$ and $Bias(\hat{x}_i)$ denote the variance and bias of the estimator respectively.

It is well-known that efficient estimators are uniformly optimal in the class of unbiased estimators, which realize the Cramér–Rao lower bound (CRLB). Therefore, biased estimators are considered to further improve the estimation precision (the reciprocal of the variance). We will then show that the proposed Bayes estimator is able to realize a strictly smaller variance than CRLB.

**Lemma 4.2** At each period $t$ when $x_i(t)$ is required to be estimated from $g_i(t)$, the Bayes estimator given in (18) has a higher estimation precision than all the unbiased estimators, which means $\hat{P}_i(t)$ is strictly smaller than $CRLB(x_i(t))$.

**Proof.** First, the Fisher Information Matrix (FIM) of $x_i$ can be computed from the log-likelihood function by:

$I_F(x_i(t)) = -E\left[\frac{\partial^2 \log(p(g_i(t); x_i(t)))}{\partial x_i(t)^2}\right] = n_i(t)$.

Then by (18), it holds that

$\hat{P}_i(t)^{-1} = n_i(t) + \bar{P}_i(t)^{-1} > I_F(x_i(t))$,

and thus

$\hat{P}_i(t) < I_F(x_i(t))^{-1} = CRLB(x_i(t))$.

**Lemma 4.3 (Bounds of prediction value)** Assuming $|\bar{x}_i(0)| \leq M_0$, where $M_0 = \max_{i=1,...,N} |\bar{x}_i(0)|$. The prediction value of the credit scorings given by equation (19) is bounded if $0 < a(t) < 1$, for any $t \geq 0$.

**Proof.** According to equation (19), we have

$\hat{x}_i(t) = \frac{1}{a(t)} (\bar{x}_i(t + 1) - b(t)u_i(t))$.

Thus

$\frac{1}{a(t)} (\bar{x}_i(t + 1) - b(t)u_i(t)) = \bar{x}_i(t) + \frac{\bar{P}_i(t)}{1 + \bar{P}_i(t)n_i(t)} \sum_{j \in N_i} (\bar{x}_j(t) - \bar{x}_i(t))$, 

11
then
\[
\tilde{x}_{i}(t+1) = a(t)[\tilde{x}_{i}(t) + \frac{\hat{P}_1(t)}{1 + \hat{P}_1(t)n_i(t)} \sum_{j \in N_{i,t}} (\tilde{x}_j(t) - \tilde{x}_{i}(t))] + b(t)u_i(t)
\]
\[
= a(t)[\frac{1}{1 + \hat{P}_1(t)n_i(t)} \tilde{x}_{i}(t) + \frac{\hat{P}_1(t)}{1 + \hat{P}_1(t)n_i(t)} \sum_{j \in N_{i,t}} \tilde{x}_j(t)] + b(t)u_i(t).
\]
Considering the first step, we have that
\[
|x_i(1)| = |a(0)||\frac{1}{1 + \hat{P}_{1,0}n_i,0} \tilde{x}_{i}(0) + \frac{\hat{P}_{1,0}}{1 + \hat{P}_{1,0}n_i(t)} \sum_{j \in N_{i,t}} \tilde{x}_j(0)| + b(0)u_i(0)|
\]
\[
= |a(0)||\frac{1}{1 + \hat{P}_{1,0}n_i,0} \tilde{x}_{i}(0) + \frac{\hat{P}_{1,0}}{1 + \hat{P}_{1,0}n_i(t)} \sum_{j \in N_{i,t}} \tilde{x}_j(0)| + b(0)u_i(0)|
\]
\[
\leq |a(0)M_0 + b(0)u_i(0)|
\]
\[
\leq |a(0)M_0| + |b(0)u_i(0)|.
\]
Iteratively, it holds,
\[
|x_i(t)| \leq M_0 \prod_{k=0}^{t-1} a(k) + \sum_{k=0}^{t-1} \left[ \prod_{l=k+1}^{t-1} a(l) \right] |b(k)u(k)|.
\]
Since \(0 < a(t) < 1\), \(\tilde{x}_i(t)\) can be bounded by a constant \(M\) for any \(t \geq 0\), i.e. \(|\tilde{x}_i(t)| \leq M\).

**Assumption 2** \(Q_t\) is bounded with an lower bound and an upper bound \(Q_l, Q_u\), respectively, i.e., \(Q_l \leq Q_t \leq Q_u\) for \(t > 0\).

**Theorem 4.4 (Bounds of estimation variance)** The estimation precision \(\hat{P}_1(t)\) is bounded by a lower and upper bound, respectively, i.e., \(P_t(t) \leq \hat{P}_1(t) \leq P_u(t) < CRLB(x_i(t))\).

**Proof.** First, we prove \(\hat{P}_1(t)\) has a lower bound. According to equation (18),
\[
\hat{P}_1(t) = a(t-1)^2 \hat{P}_1(t-1) + Q_{t-1} \geq Q_l,
\]
thus \(\frac{1}{\hat{P}_1(t)} \leq \frac{1}{Q_l}\). From Eq. (18), we have
\[
\frac{1}{\hat{P}_1(t)} - \frac{1}{P_i(t)} = \frac{1}{P_i(t)} + n_i(t) \leq \frac{1}{Q_l} + N,
\]
where \(N\) is the number of agents, denoting \(P_l = \left(\frac{1}{Q_l} + N\right)^{-1}\), we have \(\hat{P}_1(t) \geq P_l\).

Secondly, we prove that \(\hat{P}_1(t)\) has an upper boundary. According to Eq. (18), (20), and Assumption 2, we have
\[
\hat{P}_1(t) = \frac{a(t-1)^2 \hat{P}_1(t-1) + Q_{t-1}}{1 + (a(t-1)^2 \hat{P}_1(t-1) + Q_{t-1})n_i(t)}
\]
hence

\[
\frac{1}{\hat{P}_i(t)} = \frac{1}{\hat{P}_i(t-1)} \frac{1}{(a(t-1)^2 + \hat{P}_i^{-1}(t-1)Q_{t-1})} + n_i(t) \\
\geq \frac{1}{\hat{P}_i(t-1)} \frac{1}{a(t-1)^2 + (Q_{t-1}^{-1} + N)Q_u} + n_i(t) \\
\geq m_0^t \frac{1}{\hat{P}_i(0)} + \sum_{k=0}^{t} m_0^k n_i, \]

where \( m_0 = \frac{1}{\bar{a}^2 + (Q_i^{-1} + N)Q_u} \) and \( \bar{a} = \max_{0 \leq k \leq t} a(k) \), then we have

\[
\hat{P}_i(t) \leq \left( m_0^t \frac{1}{\hat{P}_i(0)} + \sum_{k=0}^{t} m_0^k n_i(t-k) \right)^{-1},
\]

which means that \( \hat{P}_i(t) \) is bounded by \( P_u = \left( m_0^t \frac{1}{\hat{P}_i(0)} + \sum_{k=0}^{t} m_0^k n_i(t-k) \right)^{-1} \), which is strictly smaller than \( CRLB(x_i(t)) \).

4.3.2. A special case.

When only unbiased estimators are considered, efficient estimators are optimal with MMSE equal to CRLB. However, when all estimators are taken into account, there does not exist an estimator that is uniformly optimal for all values of \( x_i \in [0, M] \). One reason is that the bias is always dependent on the specific value of the parameter to be estimated. Therefore, the Bayes estimator is designed to realize a better estimation performance for a subset of clients. In this part, such intuition is illustrated by a special case, where the proposed estimator is able to realize a lower MSE than all efficient estimators for clients in the middle class.

**Assumption 3** Consider a short period where the true credits are assumed to be constant, i.e. \( x_i = x_i(0) \), for any \( t \geq 0 \). Assume \( \{x_i\}_{i=1}^N \) is uniformly distributed on \([0, M]\) and \( M > 6 \).

**Theorem 4.5** For the middle-class clients with \( x_i(1) \in [3, M-3] \), the estimator in (18) is unbiased when \( N \to +\infty \) (which also indicates consistency of the estimator) and the corresponding MSE is lower than that of the efficient estimator, i.e. \( MSE(\hat{x}_i(t)) \leq CRLB(x_i(t)) \).

**Proof.** To begin with, we consider the first step

\[
\bar{x}_i(1) = a(0)\bar{x}_i(0) + b(0)u_i(0), \\
\hat{x}_i(1) = \bar{x}_i(1) + \frac{\hat{P}_i(1)}{1 + \hat{P}_i(1)n_i(1)} \sum_{j \in \mathbb{N}, i} (\bar{x}_j(1) - \bar{x}_i(1))
\]

for \( x_i(1) \in [3, M-3] \).

Under **Assumption 1**, it is easy to see that \( \bar{x}_j(1) \) is unbiased with \( \bar{x}_j(1) \sim \mathcal{N}(x_j(1), \hat{P}_j(1)) \) for any \( j = 1, \ldots, N \), which can also be rewritten as:

\[
\bar{x}_j(1) = x_j(1) + e_j(1), \quad \text{with } x_j(1) \sim \mathcal{U}[0, M] \text{ and } e_j(1) \sim \mathcal{N}(0, \hat{P}_j(1)).
\]
Then the pdf of $\bar{x}_j(1)$ can be given by the convolution of two density functions as

$$p_{\bar{X}}(\bar{x}_j(1)) = \int_{-\infty}^{+\infty} p_X(x_j(1)) p_E(\bar{x}_j(1) - x_j(1)) dx_j(1)$$

$$= \frac{1}{2M} [erf\left(\frac{M - \bar{x}_j(1)}{\sqrt{2P_j(1)}}\right) - erf\left(\frac{-\bar{x}_j(1)}{\sqrt{2P_j(1)}}\right)],$$

where $erf$ denotes the error function.

Recall that $N_i(1)$ is generated based on (3). It is easy to see that $E[n_i(1)] \to +\infty$ as $N \to +\infty$. Then by the law of large number, for any given $x_i(1) \in [3, M - 3]$, it holds that

$$\lim_{N \to +\infty} E[\hat{x}_i(1)] - x_i(1) = E[\bar{x}_j(1) - \bar{x}_i(1)]$$

$$= \int_{-\infty}^{+\infty} \nu e^{-\frac{(x_j(1) - x_i(1))^2}{2}} (\bar{x}_j(1) - x_i(1)) p_{\bar{X}}(\bar{x}_j(1)) d\bar{x}_j(1)$$

$$= \int_{x_i(1) - 3}^{x_i(1) + 3} \nu e^{-\frac{(x_j(1) - x_i(1))^2}{2}} (\bar{x}_j(1) - x_i(1)) p_{\bar{X}}(\bar{x}_j(1)) d\bar{x}_j(1) = 0$$

where the last line results from the three-sigma rule and the fact that $p_{\bar{X}}(\bar{x}_j(1))$ is identical on $[3, M - 3]$ since $\bar{P}_j(1)$ is far smaller than $M$.

In this situation, we have that $\hat{P}_i(t + 1) = \frac{\hat{P}_i(t)}{1 + \hat{P}_i(t)n_i(t)}$ is decreasing with $t$. Then the recursion can be executed in the same manner as shown above, thus leading to an unbiased estimator for any $t \geq 0$, i.e.

$$\lim_{N \to +\infty} E[\hat{x}_i(t)] - x_i(t), \text{ for any } t \geq 0,$$

for any $x_i \in [3, M - 3]$.

Hence it is straightforward that

$$MSE(\hat{x}_i(t)) = \hat{p}_i(t) < CRLB(\hat{x}_i(t)), \text{ for any } t \geq 0,$$

which is also converging to zero with probability almost 1.

Remark 1 As for the clients with credit scores on the boundary, it is obvious that the proposed estimator is biased. For example, the estimation for a low-income client will be lifted since most of his neighbors have a higher score. Recall that the observation that merely relying on individual assets is unbiased with a high variance. Then the Bayes estimator realizes a lower variance at the expense of bias error, which can be considered as a trade-off between individual attributes and network information. Such trade-off between bias and variance can be considered by the lender by adjusting the weighting function $\alpha(x_i)$ in (15). Or alternatively, for those clients with extremely low or high income, the lender can also choose to make assessments only based on individual financial attributes and consider the network information as a reference for risk prediction as shown in Section 3.

5. Numerical simulations

In this section we present a numerical study considering a network with the 30 clients, i.e. $N = 30$. Here we consider the special case as given in Section 4.3.2, where the parameters in Eq. (13) are
setting chosen as $a = 1$, $b = 0$, $Q_t = 0$, respectively. The scoring process is executed on time interval $t = 1, \cdots, 60^*$. The scenario of online scoring is considered. Initially, the lender can only obtain a noisy credit estimation based on limited information of individual assets. At each time, each client forms a new homophily-based network according to others’ credit reports published by the lender, which is then used by the lender to make a one-step optimal predictor for the next period. Under such framework of dynamic interaction, the links among clients are reconstructed at each period with the updated publishing of the score by the lender. Fig. 2 presents the network structure at time step $t = 60$. It is a directed network, where clients with similar credit scores are connected with a probability of $Pr(g_{ij}(t) = 1) = e^{-\frac{(x_i(t) - x_j(t))^2}{2}}$.

Figure 2.: Network among 30 clients at time step $t = 60$

In this simulation, the recursive Bayes estimator based on dynamic interaction in Section 4 is used for lenders to update the credit scores for each client. Estimation results and the error covariance are shown in Fig. 3. The real state (red dots in the figure) stands for the precise credit states for each agent, which are presented in an ascendant order. The estimation values at time step $t = 1$, $t = 30$, $t = 60$ are presented for each agent, respectively. We can see that for the middle class, i.e., the clients with credit scores around $4 \sim 12$, the estimation converges to their true value, while the low–income class and high–income class have a positive and negative bias respectively. Such conclusion is consistent with the theoretic results in Section 4.3.2. The covariance of the estimation errors on the right of Fig. 3 shows that $\hat{P}_i(t)$ is decreasing to a lower boundary $P_i$. Furthermore, since there is no system noise $w(t)$, $\hat{P}_i(t)$ will converge to 0 with probability almost 1.

Figure 3.: Credit scoring estimation (left) and error covariance (right)

*Each time step stands for a period with a specific number of days.
6. Conclusions and future work

It is common practice to use the structured financial data such as loan characteristics (purpose of the loan and its duration), clients characteristics (age, gender, education) and credit history (repayment of previous loans) for estimating the credit score. In this paper, we have formulated a theoretical framework to study the credit scoring problem by incorporating network information based on the homogeneous preference. Two scenarios are considered, respectively. Firstly, we proposed a Bayesian optimal filter to predict clients’ credit scores if the publishing of the credit scores are estimated merely from structured individual data. Such prediction is used as a monitoring indicator for the risk warning in lenders’ future financial decisions. Secondly, we developed a recursive Bayes estimator to improve the accuracy of score estimation by incorporating network topologies as well. It was shown that under the proposed evolution framework, the designed estimator has a higher precision than any efficient estimator, and the mean square errors are strictly smaller than the Cramér–Rao lower bound for clients within a certain range of scores. For further investigation, simulation results for a case where true credits are uniformly distributed illustrate the effectiveness and feasibility of the proposed methods.

Our findings suggest that network information analysis plays a quite important role for estimating the credit scorings from a point of theoretical view. A straightforward problem afterwards is to study the connection rules for the network formation among the clients in order to improve oneself credit scoring.
References


Rusli, E., Bad credit? start tweeting. startups are rethinking how to measure creditworthiness beyond FICO. *Wall Street Journal*, 2013.


