Visual analytics for networked-guarantee loans risk management

Zhibin Niu[*] Tianjin University
Dawei Cheng[†] Shanghai Jiao Tong University
Liqing Zhang[‡] Shanghai Jiao Tong University
Jiawan Zhang[§] Tianjin University

Abstract

Groups of enterprises can guarantee each other and form complex networks in order to try to obtain loans from banks. Monitoring the financial status of a network, and preventing or reducing systematic risk in case of a crisis, is an area of great concern for the regulatory commission and for the banks. We set the ultimate goal of developing a visual analytic approach and tool for risk dissolving and decision-making. We have consolidated four main analysis tasks conducted by financial experts: i) Multi-faceted Default Risk Visualization, whereby a hybrid representation is devised to predict the default risk and an interface developed to visualize key indicators; ii) Risk Guarantee Patterns Discovery. We follow the Shneiderman mantra guidance for designing interactive visualization applications, whereby an interactive risk guarantee community detection and a motif detection based risk guarantee pattern discovery approach are described; iii) Network Evolution and Retrospective, whereby animation is used to help users to understand the guarantee dynamic; iv) Risk Communication Analysis. The temporal diffusion path analysis can be useful for the government and banks to monitor the spread of the default status. It also provides insight for taking precautionary measures to prevent and dissolve systematic financial risk. We implement the system with case studies using real-world bank loan data. Two financial experts are consulted to endorse the developed tool. To the best of our knowledge, this is the first visual analytics tool developed to explore networked-guarantee loan risks in a systematic manner.

Index Terms: H.5.2 [User Interfaces]: User Interfaces—Graphical user interfaces (GUI); H.5.m [Information Interfaces and Presentation]: Miscellaneous

1 Introduction

Networked-guarantee loans (also known as guarantee circles) are an economic phenomenon unique to Asia countries, especially China, and they are attracting increasing attention from the banks and the government. In order to obtain loans from banks, groups of small and medium enterprises back each other to enhance their financial security. When more and more enterprises are involved, they form complex directed-network structures [25]. Figure 1 shows a guarantee network consisting of more than 600 enterprises. The existing mechanism in the financial industry for loan decision-making falls behind the demand for loans from businesses. Most of the criteria are designed for *independent major* players, while, in practice, the s-mall and medium enterprises may provide inaccurate or manipulated data or induce intertwined risk factors [17]. Thousands of guarantee networks of different complexities have coexisted for a long period and have evolved over time. This requires an adaptive strategy in order to prevent, identify, and dismantle systematic crises.

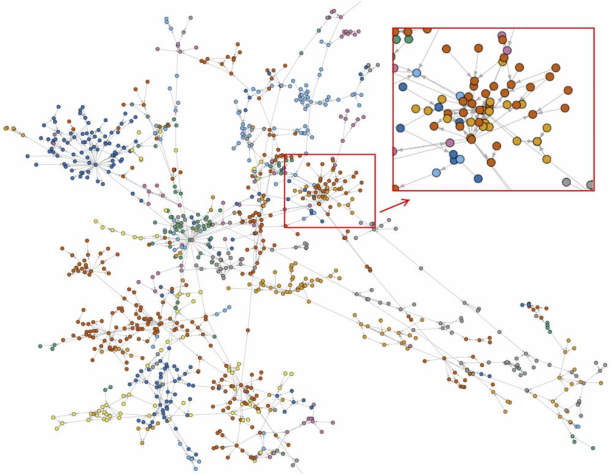


Figure 1: A real-world loan guarantee network formed from bank records, with each node representing an enterprise.

Highlighted by the complex background of the growth period, the structural adjustment of the pain period, and the early stage of the stimulus period, structural and deep-level contradictions have emerged in the economic development system. Many kinds of risk factors have emerged throughout the guarantee network that might accelerate the transmission and amplification of risk, and the guarantee network may be alienated from the "mutual aid group" as a "breach of contract". An appropriate guarantee union may reduce the default risk, but significant contagious damage throughout the networked enterprises may still occur in practice [24]. The guaranteed loan is a debt obligation promise; if one corporation gets trapped in risks, it may spread the contagion to other corporations in the network. When defaults diffuse across the network, a systemic financial crisis may occur. The contagion to risk loan guarantee, especially malicious guarantee, is still relatively limited. Monitoring the financial status is so difficult that it is usually only after a capital chain rupture that the regulators can study a case in depth. With the economic slowdown, the need for credit risk management is more urgent than ever before.

We propose a visual analytics approach for networked-guarantee loan risk management. The main contributions are:

1. We identify and provide practical solution to the problem of credit risk management for networked-guarantee loans, which is driven by finance industry demands, and we believe this is an important research problem to the data mining and visual analytics community;
2. We implement intuitive visual analytic tools for i) Multi-faceted Default Risk Visualization; ii) Risk Guarantee Patterns Discovery; iii) Network Evolution and Retrospective; and iv) Risk Communication Analysis. We perform empirical studies and verified the ecacy.
3. We conduct interviews with two domain experts and have our approach endorsed. We highlight three risk patterns that are difficult to discern without using a visual analytic approach.

The rest of the paper is organized as following: Section 2 describes works involving different aspects related to our problem; Section 3 details the four visual analytic tasks and our approaches;

[*] e-mail: zniu@tju.edu.cn
[†] e-mail: dawei.cheng@sjtu.edu.cn
[‡] e-mail: lq-zhang@sjtu.edu.cn
[§] e-mail: jwzhang@tju.edu.cn

2165-8773/18/$31.00 ©2018 IEEE
DOI 10.1109/PacificVis.2018.00028

Section 4 describes the data and the case study; and we report the user study results in Section 5. Conclusions and future works are described in Section 6.

2 Related Work

We introduce several relevant works on network analytics in the financial domain and works on financial security visualization.

Credit risk evaluation Since the seminal "Partial Credit" model [23], numerous data-driven approaches have been introduced for credit scoring [5]. Jan Vanthienen and others interpreted and visualized the learned knowledge embedded in neural networks based credit scoring approach [4]. Andrew W. Lo and others propose consumer credit risk prediction models based on consumer behavior (debt-to-income ratio and consumer banking transactions), linear regression model, and time-windowed data set. They claim a 85% default prediction accuracy and can save cost between 6% and 25% [18]. In this paper, we adopt a similar idea and propose a hybrid representation to predict the enterprise default rate.

Financial network analytics The relationship between network structure and financial system risk has been studied carefully and several insights have been drawn: Network structure has little impact on system welfare, but plays an important role in determining systematic risk and welfare in the short term debt [1]. After the 2008 global financial crisis, network theory attracted more attention: The crisis brought about by Lehman Brothers spread to connected corporations in a similar infectious way as the epidemic of Severe Acute Respiratory Syndrome (SARS) in 2002 – both were small damages that hit a networked system and caused serious events [8]. The journal of Nature Physics published a special edition on how to understand some fundamental economic issues using network theory. For example, the dynamic network produced by bank overnight fund loans may act as an alert of a crisis [11]. Contrary to the conventional stereotype that large institutions are "too big to fail", the truth is that the position of an institution in a network is equally, and sometimes more, important than its size [6]. The more central the vertex is to the graph, the more influential it is to the whole economic network when default occurs [11]. Although considerable efforts have been made to understand fundamental problems in financial systems [7], there is little work on system risk analysis in the networked-guarantee loans, except for preliminary work [26], where a positive correlation between the K-shell decomposition value of the network and default rates was reported. Readers are referred to [28,36] for more references on graph related applications.

Visualization in financial systems Visualization and visual analytics have been introduced to the financial sector, including transactions monitoring, price fluctuations, and complex decision-making [14]. Animation is used to visually analyze large amounts of time-dependent data [2,3]. The 3D treemap is introduced to monitor real-time stock market performance and to identify a particular stock that has produced unusual trading patterns [16]. The interactive exploratory tool is designed to help the casual decision-maker to quickly choose between various financial portfolios [30]. Coordinated specific keywords visualization within wire transactions are used to detect suspicious behaviors [12]. The Self-Organizing Map (SOM), a neural-network-based visualization tool, is often used in financial risk visualization analysis for monitoring the occurrence of sovereign defaults in less developed countries [32], for the visual analysis of the evolution of currency crises by comparing clusters of crises between decades [31], and for discovering imbalances in financial networks [33]. Readers are referred to [21] for more references on financial visualization. The visual analytic approach is also employed to analyze contagion in networks and in the simulation of contagion effects [34]. Motifs are employed to analyze and visualize the network [19,22,35]. We are inspired by the various technologies and designed a visual interface for networked-guarantee loan risk management.

3 Risk Management and Visualization

We consult with financial experts and set the ultimate goal of developing an interactive tool for the government and banks to monitor default spread status and provide insight for taking precautionary measures to prevent and dissolve systematic financial risk. Based on the goal, we consolidate four analysis tasks. The tasks include:

Task1: Multi-faceted Default Risk Visualization. The current loan credit rating system is based on the pure financial status of the individual borrower. The credit assessor can usually access the first layer of the guarantee chain, thus cannot trustfully evaluate the risks. It is necessary to carry out a systematic analysis of the enterprise to avoid inadequate risk assessment.

Task2: Risk Guarantee Patterns Discovery. Fraud guarantee patterns may lead to default and diffusion. Identifying new high default patterns helps banking experts to single out and tackle the principal default problem. Visual analytics tools should be developed to thoroughly analyze the network.

Task3: Network Evolution and Retrospective. Understanding the network dynamic helps financial experts to understand how firms are connected together temporally. It requires visualizing the evolution of the guarantee network based on historical data.

Task4: Risk Communication Analysis. Before a crisis occurs, forecasting the default diffusion path and monitoring the default spread status will help the government and banks to take precautionary measures, conduct research, and take effective measures to prevent and dissolve risks.

Fig. 2 gives the workflow. In the data preprocessing stage, guarantee networks are constructed from the bank records. Then, the spatiotemporal information is utilized during the visual analytics stage. In task 1, forecasted default risk and network related measurements are visualized to help to locate hotpot efficiently. In task 2, an interactive interface is designed to help the experts to explore and discover possible malicious loan frauds. In task 3, the evolution of the network provides insights of the past enterprises' activity and task 4 provides the possible default spread path in the future. In the risks dissolving stage, with the insights obtained from previous stage will help to divide the guarantee network so that no regional or systematic financial risks occur. We next describe the detailed algorithms, strategy, and interactions.

3.1 Default Risk Prediction and Visualization

The loan records reveal that guarantee network and default rates are both increasing, and the network structures show a strong correlation with the defaults. We construct feature vectors consisting of hybrid information and employ the supervised learning approach to train the prediction model. In what follows, we discuss the hybrid features used in our model.

In order to build a highly representative feature that can reliably reflect the statistical relationships between the customers information and their repayment ability, we clean the data and construct the features as: (1) Basic Profile, the essential company registration information, which reflects the character, capital, collateral, capability, condition, and stability [26]. We use business nature, registered capital, enterprise scale, employee numbers and other information to make up the corporations basic profile. Most banks require the company to update this basic information when the enterprise makes a loan application; we choose to use the latest information as the basic profile features of the loan. (2) Credit Behavior, historical behavior, e.g. credit history, default records, default amount, total loan amount and loan count, total loan frequency (if any), total default rates. This is calculated using all the loan records before the active loan contract. (3) Active Loan, the loan contract in its execution period. It contains active loan amount, active loan number, type of capital return and interest return, etc. (4) Network Structure, network features such as centralities are extracted. Note that, as discussed above, the basic profile may not be completely trustworthy, as the businesses may

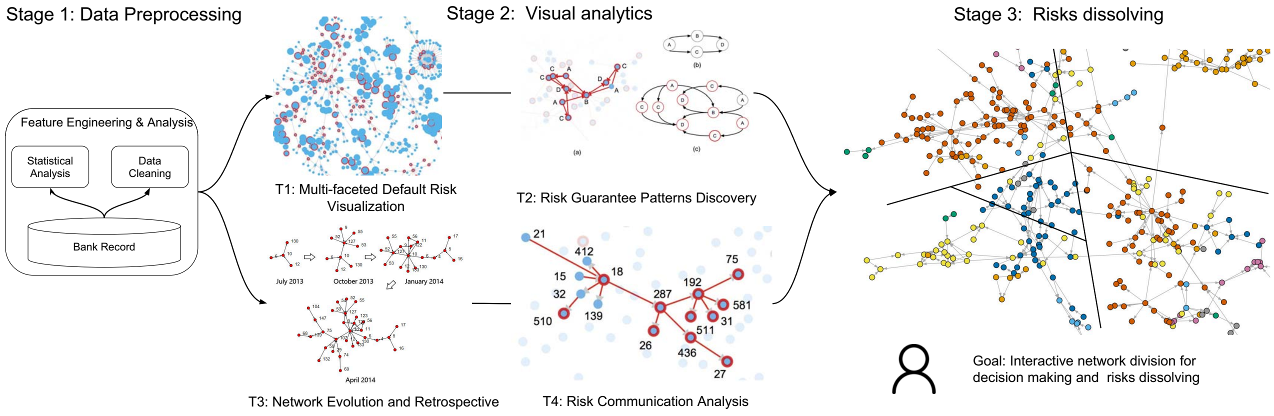

Stage 1: Data Preprocessing
Feature Engineering & Analysis
Statistical Analysis
Data Cleaning
Bank Record
Stage 2: Visual analytics
T1: Multi-faceted Default Risk Visualization
T2: Risk Guarantee Patterns Discovery
(a)
(b)
(c)
July 2013
October 2013
January 2014
April 2014
T3: Network Evolution and Retrospective
21
412
75
15
18
192
32
287
581
510
139
511
31
26
436
27
T4: Risk Communication Analysis
Stage 3: Risks dissolving
Goal: Interactive network division for decision making and risks dissolving

Figure 2: Overview of the system and tasks.

provide out-of-date or even false information to the bank. However, the guarantee network uses trustworthy information, as the bank can build the data from its own records.

The prediction of default for a customers loan guarantee can be modeled as a supervised learning problem. We choose to use logistic regression based on a gradient boosting tree to predict the risk for the reason that it is reportedly successful in many data science problems. Also, note that our task is to visualize the risk for different enterprises. We do not compare the prediction performance of various regression methods in this paper; these will be demonstrated in our future work.

In the XGboost, the tree ensemble model using K additive function to prediction output can be represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i) \tag{1}$$

In Eq. 1, f_k is the k_{th} decision tree, X_i is the training feature and $\hat{y}_i$ is prediction results. Finding parameters of the tree model is turned into minimizing the objective function problem and it can be trained in an additive manner [13].

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \ where \ \Omega(f) = \gamma T + \frac{1}{2}\lambda \, ||\omega||^2 \tag{2}$$

where $\sum_i l(\hat{y}_i, y_i)$ is a training loss function that measures the difference between the prediction and the target; $\Omega(f)$ is a smoothing regularization term to avoid over-fitting.

We design and implement a visual interface enabled to view the network with various multiple measurements. Fig. 3 gives the interface, by which users can adjust the node size by the predicted default risk (proportional to the diameters of the sphere) and by the following network centrality measurements: Hub score and Authority score, K-Shell decomposition score, PageRank, Eigenvector centrality scores, Betweenness centrality, and Closeness centrality. Fig. 4 gives a part-visualization of a real guarantee network. In the graph, all defaulted enterprises are highlighted by red circles. Node size is proportional to predicted risk (a), K-shell value (b), and authority score (c). Through the interface, users can also observe the rolling prediction risk of an enterprise over a month and highlight it on the whole network by choosing it on the heatmap.

3.2 Risk Guarantee Patterns Discovery

Empirical studies by bank risk control specialists suggest risk guarantee patterns, including mutual guarantee and revolving guarantee (see Fig. 7). Such interactions are currently legal in China but in the banking industry, specialists in the bank risk control department only have SQL query capability to detect relatively simple guarantee patterns. Understanding more complicated risk guarantee patterns is difficult due to the tools limitation. An arbitrary guarantee pattern, which has a high default rate, can lie underneath the complex network structures. Thus, it is impossible exhaustively to compare all network patterns to determine whether it is in high default. In this work, we develop a visual analytics tool to help the experts explore, discover and further understand what has happened. We follow Ben Shneidermans mantra of information visualization, and the approach includes two steps: first, high default group detection; then risk guarantee pattern discovery.

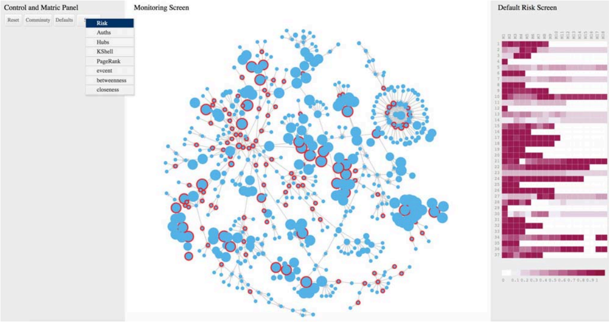


Figure 3: The interface for Visual Analytics for Enterprise Default Risk. We use a heatmap to code the rolling prediction risks over a month.

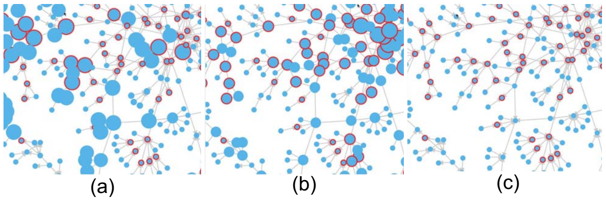


Figure 4: Visualization of the network with (a) the rolling prediction risk, (b) K-shell value, and (c) authority score.

High default group detection. Recognizing high default groups narrows down the search scope of the risk guarantee relationship. Based on the conjecture that defaults tend to occur in clusters, we divide the whole network into several distinct sets by community detection. Theoretically, community structure in the graph is defined as the node sets that interact with each other internally more frequently than with those outside it. Identifying such substructures provides insight into understanding the structure of complex networks (both the functions and the topology affect each other).

We use a force-directed graph with colored communities and revised treemap interface to visualize the community detection results. The community label and default rates are displayed on the flat colored blocks. The treemap chart is used for navigation here; thus, the sum of the area does not necessarily need to be one. The larger blocks reveal the high default communities saliently.

Fig. 5 (a) shows the results on a typical independent subgraph that we constructed from bank loan records. The communities are marked using a separate color background and the average default rates are labeled. There are 36 communities, of which defaults occur in 27, with an average 38% to 8.6% default rate, all other 9

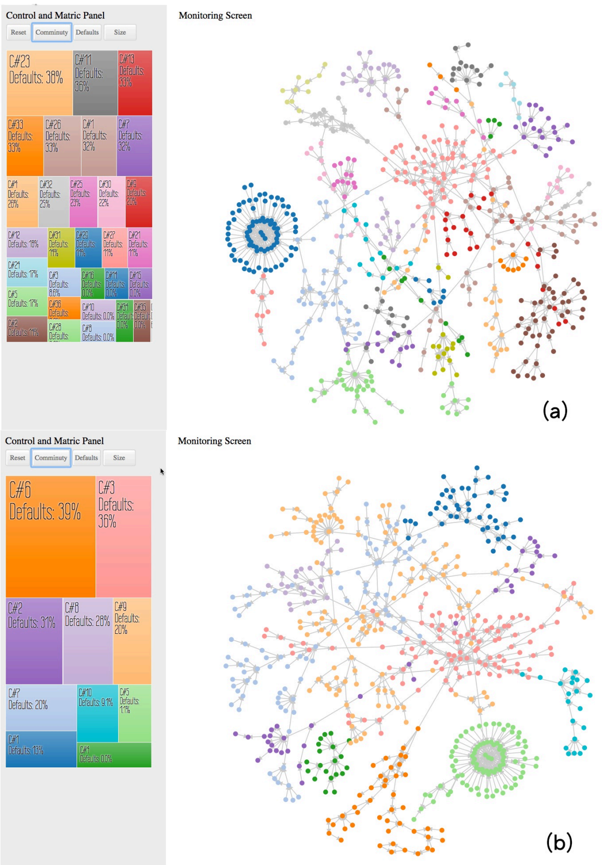


Figure 5: Defaults occur in clusters and we interactively edit the clusters. (a) 30 communities generated by a random walks algorithm; (b) 10 communities after interactive editing. The ratios for defaulting firms are labeled separately on the left-hand side treemaps.

communities have no defaults. We adopt the random walk algorithm [29]. A similar phenomenon is observed on random walks, edge betweenness, and spineless community.

However, the evaluation of community detection is still an open question [20]. As the community detection algorithm only considers the link information and neglects the node attribute information, the partition may not be consonant to the actual conditions. The basic rule for community detection is to minimize the number of links between communities, and this uses pure network structure information. In financial practice, each node in the network comes with rich information, such as enterprise sectors, changes in deposits, assets, loan amount, etc. It would be unreliable to discard such attributes when dividing the network. By interaction, we enable the users to edit the communities into coherent ones by referring to the relevant financial metric. We allow users to interactively perform the following manipulation actions.

Interactive community editing. We enable users to explore the financial information and interactively edit the communities by merging strongly associated communities, to reassign the community labels for the structural hole spanners (a key role in the information diffusion) [10], or to split a community into several distinct smaller groups. The generated subgraphs are noted as groups of interest (GOI), in which the high-risk guarantee pattern is often hidden.

Reassign. The reassign operation allows the user to change the community labels of the structure hole spanner. The structure hole spanner is the bridge node that connects different communities in a network. Fig. 6 is reproduced from [15]; it illustrates a network with three communities and six structural hole spanners. Empirical study suggests that individuals would benefit from filling the "holes" (an alternate name for the structural hole spanners) between people or groups that are otherwise unconnected [9]. We observed high default in structure hole spanners with their neighboring internal nodes. We enable the users to investigate the financial metrics and reassign the community labels of the structure hole spanners. With the interface, he/she first double-clicks the "tile" on the treemap highlighting all the connected communities. Single-clicking the structure hole spanner node can reassign it to the opposite community.

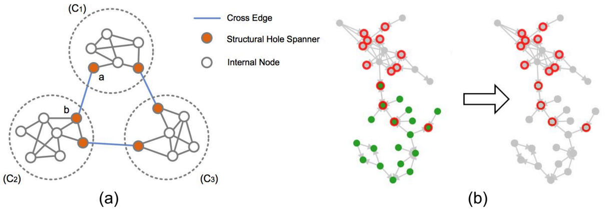


Figure 6: (a) Structural hole spanner illustration example, reproduced from [15]; the structural hole spanners are editable for merging or reassigning actions. (b) Example of merging two communities on the structural hole spanner.

Merge. Naive community detection divides a graph based purely on links in the graph, it may generate many communities where some of them share a common sector category or similar network structures. Merging the communities referring to a specific financial metric can produce medium-sized and more tractable subgraphs. With the interface, he/she first double-clicks one "tile" on the treemap highlighting all the connected communities and then double-clicking the structure hole spanner node to merge the two communities.

Split. When the default is unevenly distributed, we need to split the community and cut off the stable parts to reduce the next motif related computation complexity. With the interface, he/she first double-clicks one "tile" on the treemap highlighting the connected communities and then double-click the edge making the two opposite parts of the subgraph be split into two communities.

The interface also has a financial radar view to encode the key financial infromation. The key indices include: *Defaults*, historic default behavior; *LA/RC* the ratio of loan amount to registered capital. It would be more insightful to use the ratio of loan amount to enterprise net assets; however, the latter information is not always available, so we use registered capital instead. *Deposit loss* the rapid decrease of deposit and shorting of money may imply business out of the situation. *Sector* the enterprise sector related to the macroeconomic conditions and is an important clue when editing communities. *GA/RC* the ratio of guarantee amounts to registered capital. The ratio of guarantee amount to enterprise net assets is a crucial factor for the stability of the financial system. Also because of lacking information transparency, we use registered capital instead. *Credit rating* is the review rating of bank experts; this is also a key clue when editing communities.

Risk Guarantee Pattern Discovery and Visualization. The guarantee patterns that are prone to default may exist underneath the GOIs. A complex guarantee network is always connected by several smaller subgraphs bridged by the structural hole spanners. The subgraphs inside the communities may reveal certain risk patterns; even a fraud pattern. The motifs are the most basic building blocks for a graph and they may reflect functional properties. In this work, we obtain a set of motifs by first detecting motifs from the GOI. The motifs are ranked by their default rates (Eq. (3)). High default rate motifs are noted as a pattern of interest (POI); these may need to be investigated by banking experts as a priority.

$$priority = \left(\frac{\sum default\_node\_number(m))}{\sum node\_number(m)}\right) \tag{3}$$

where m is a motif. All motifs are possible risk guarantee patterns.

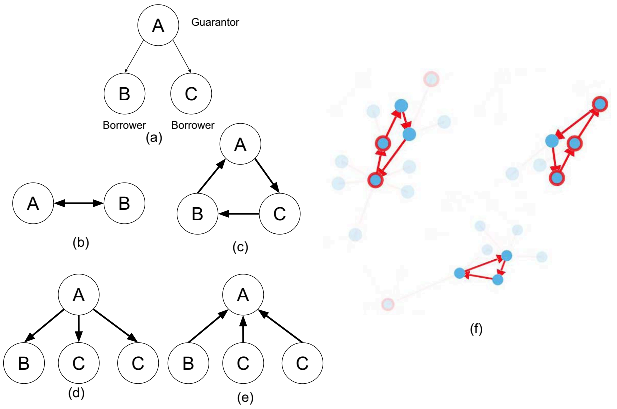


Figure 7: (a) guarantee network, where enterprise A (guarantor) guarantees B and C (borrowers) to get loans from the bank (lender). The (b–e) graphs are classic loan guarantee patterns, specifically: (b) mutual guarantee, (c) revolving guarantee, (d) star shape guarantee, (e) joint liability guarantee. (f) revolving guarantees detected from a real-world loan guarantee network.

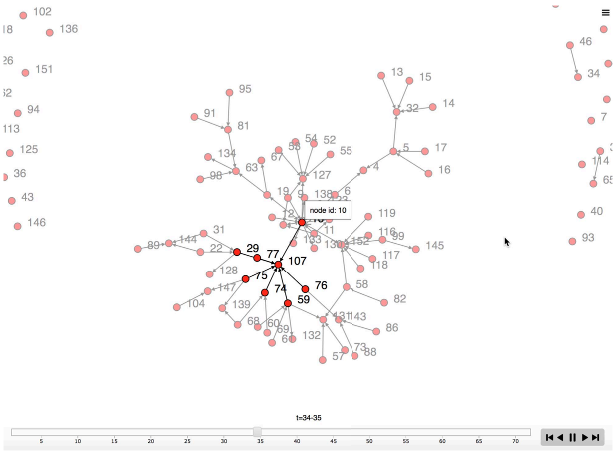


Figure 8: Visual analytics interface for evolving loan guarantees. The numbers in the graph are node ID

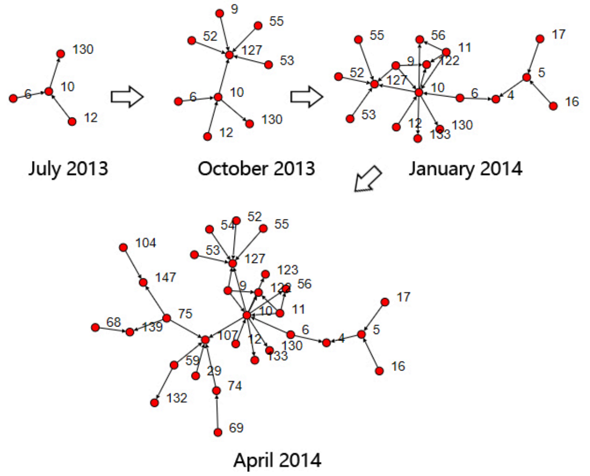


Figure 9: The guarantee network keeps evolving from July 2013 to April 2014. The numbers in the graph are node ID

However, it is still computationally challenging to obtain all POIs using the approach above for the following reasons. Firstly, motif structures increase rapidly with an increase in node number; for example, a four-node motif gives rise to more than 3000 possibilities. It is therefore impossible to enumerate all the motif structures. Secondly, motif matching is exhaustively searched from the query graph into the large network, and it presents a subgraph isomorphism problem. It still takes too much time for motifs with more nodes to be matched on the network. With the interface, we enable interactive motif editing. Users can refer to the financial radar view of adjacent nodes and add new nodes to the motifs to generate a more complex POI without exhaustively compute all possibilities.

3.3 Network Evolution and Retrospective

Network evolution over time is observed from the guarantee network. The topology of the network keeps changing: some nodes are connected to the network or removed from it; some communities are connected together through the guarantee of the structural hole spanner. Like many other real networks, competitive decision-making is taking place in the guarantee network: When a firm lacks the security to obtain a loan from a bank, it may resort to a guarantee corporation or third-party firms. To some extent, the new guarantors may improve the overall rationality of the system, but may also induce an unstable factor as the network becomes even more complex. Understanding the network dynamic helps financial experts to understand how the firms are connected together temporally.

Animation is employed to visualize the evolution of a guarantee network. With the interface, users can drag the time bar to backtrack how the network has evolved over time. They can hover the cursor over a node to view the companys financial information. This will help the financial experts to understand what has happened historically. Fig. 9 gives an example of a real network evolved from July 2013 to April 2014. By combining enterprises financial information of different time, financial experts would be able to make the analysis.

3.4 Risk Communication Analysis

As a new phenomenon, the understanding of the systematic risk of the networked-guarantee loan is still insufficient. Sophisticated guarantee relationships tend to cause credit granted by multiple lenders and excessive credit. In the loan guarantee, a guarantor takes on the debt obligation if the borrower defaults; therefore, if the guarantee cannot be paid back to the bank, it may resort to its guarantors. In this case, the default may propagate throughout the network, like a virus. The default contagion increases both the possibility of the occurrence of risks and the transmission of risks. Especially in a period of economic downturn, some enterprises will face operational difficulties and the financial crisis will have a domino effect: the default phenomenon may spread rapidly in the network, and this could make a large number of enterprises fall into an unfavorable situation. The government and the banks always wish to monitor the default spread status and understand the complexity of the current issue of risks so that they can take precautionary measures, conduct research, and dissolve the risks to ensure that no regional or systematic financial risk occurs.

Based on relevant knowledge and experience, we develop a visual analytics tool to aid the default path discovery by visualization. A principle of the default diffusion can be described, as the vulnerable nodes are the guarantors. Fig. 10 gives a diffusion path illustration. (a) is a guaranteed network with eight nodes, where node E provides a guarantee to five adjacent nodes and C, D provide a guarantee to B and then to A; (b) is the possible diffusion path: the default of node A may lead to B, C, D, and even E defaulting. It is noted that nodes G, F, and H are not connected with node E, and therefore the default of E will not affect the repayment status of G, F, or H.

In practice, there may be multiple possible propagation paths, as each node can serve as a guarantor or get guaranteed. It is difficult to outline the main propagation path from the entire graph. We make the following assumption: the node on multiple propagation paths is the key to prevent large-scale default diffusion and thus should be highlighted. We compute all the propagation paths, count

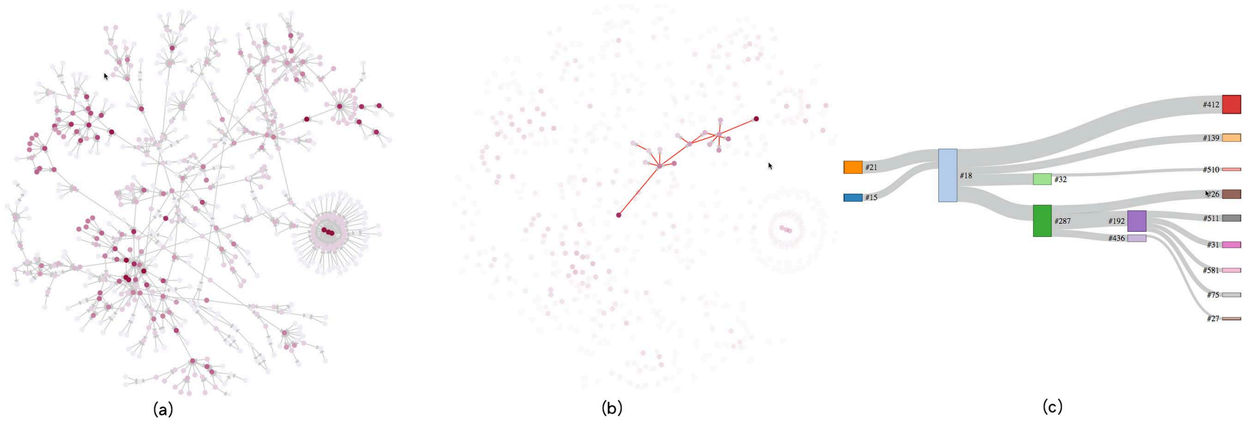
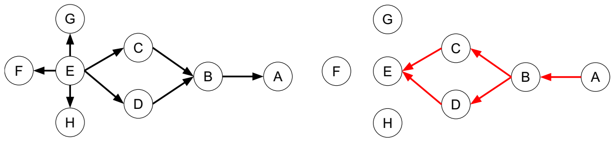


Figure 11: One real diffusion path and the corresponding Sankey diagram.



(a) Loan guarantee network

(b) Possible default diffusion path

Figure 10: Default path for a real network. The characters represent different enterprises.

occurrences, and highlight the node on the network. We use color to illustrate the *propagation risk importance*.

We design the visual analytics tool, which enables financial experts to take into account several factors on the judgment of defaults. These factors include the financial information on the corporation and the guarantee contract amount information. The former information is plainly listed when the user hovers the mouse pointer over the node, while a Sankey diagram is used to represent the guarantee flow. The widths of the Sankey diagram bands are directly proportional to the guarantee amount.

Fig. 11 (a) gives results on a real guarantee network, when we choose one nodefor example, node 32. The whole potential propagation path is highlighted in (b), while (c) is the corresponding Sankey diagram. It can be seen that upstream companies usually provide more in guarantees than they receive. For example, node 18 provides much more guarantee than it receives. The imbalance between the guarantee amount and the collateral amount provides a clue for credit line assessment. The real situation is even more complex. The default may be diffused like a viral infection, and the virus must identify and bind to its receptor (guarantor). As mentioned earlier, each enterprise has more than 3000 financial entries, and it is therefore difficult to quantify each enterprises ability to resist infection. We enable users to look up multiple financial statuses and cut off the propagation path. We also note that the propagation model provides more insights to end users, and we plan to perform an in-depth study of the topic and provide a simulation interface in the future.

4 Case Study

We first introduce the loan process, data exploration and then describe the experiments. As Fig. 12 shows, there is often more than one guarantor per loan transaction, and there may be several loan transactions for a single guarantor in a period. Once the loan is approved, the business can usually obtain the full size of the loan immediately, and starts to repay the bank regularly by an installment plan until the end of the loan contract. The banks need to collect as much fine-grained information as possible concerning the repayment ability of the enterprise. The information falls into four categories: transaction information; customer information; asset information such as mortgage status; and history of loan approval records, etc.

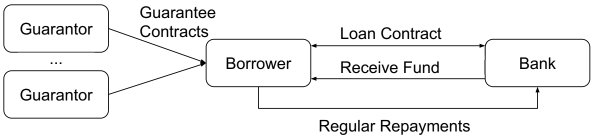


Figure 12: Loan guarantee process. The borrower wishing to get a loan from a bank first needs to sign loan guarantee contracts with guarantors before signing a loan contract. After the company receives its loan from the bank, it repays the loan by installments.

We collect loan records spanning ten years from our cooperated commercial bank and construct the guaranteed network. The names of the customers in the records are encrypted and replaced by an ID.

4.1 Multi-faceted Default Risk Visualization

We propose a multi-faceted default risk visualization interface (see Fig. 3) and it includes forecasting default risk, centrality measurements (Authority score, hub score, K-shell, PageRank, Event, Betweeness, and, Closeness). We next explain them separately.

Default risk prediction. As illustrated in Section 3.1, a hybrid representation and gradient boosting tree based approach is employed to predict the default risk. In the following experiments, we define Node-wise (NW) feature as the vector composed of basic profile, credit behavior, active loan information; define Network (N) feature as only network structure features; define Community Behavior (CB) feature as loan history behavior associated with graph community; define Hybrid (H) feature consists of both node-wise feature, network feature, and community behavior feature.

Besides, we choose to employ a three-month sliding window setting for training, observation, prediction, and evaluation. The reasons are two-folds: (1) Prediction shall be adapted to a dynamic setting with a regularly updated forecasting results. In fact, using sliding window is a typical way for rolling prediction as commonly adopted in event prediction practices. (2) The business often runs on a quarterly basis, which can also be observed from the record that the default happens intensively at each end of quarter. Thus from a business demand perspective, it would be helpful to know the borrowers who may be default on a quarterly basis. As Fig. 13 shows, in the training stage, for all customers who obtain bank loans from 2013 Q1 (first quarter of 2013), the features are extracted in that period; the repayment status in 2013 Q2 are the labels to train the model. In the testing stage, we use the trained model to predict the customers who obtain loans between 2013 Q2 and use the real repayment status from 2013 Q3 to evaluate the performance when reaching the end of September 2013.

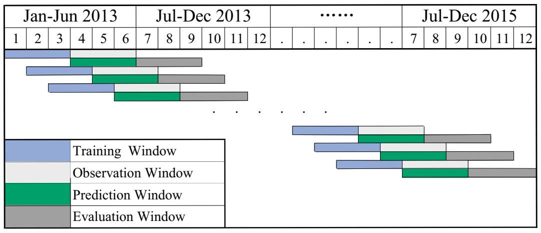


Figure 13: Illustration for the rolling sliding windows protocol. Features are extracted in the training window, and the corresponding outcome default label is collected in the observation window. Then the features and default outcome are used to train the model. The trained model is used by collecting the input features during the prediction window and verifying its performance when we reach the end of the evaluation window.

Table 1: AUC of forecasting models

Period	NW	NW,CB	NW, N	H
2013 Q3	0.910	0.924	0.917	0.925
2013 Q4	0.905	0.926	0.920	0.931
2014 Q1	0.901	0.929	0.923	0.930
2014 Q2	0.907	0.931	0.928	0.933
2014 Q3	0.908	0.935	0.933	0.937
2014 Q4	0.910	0.933	0.939	0.941
2015 Q1	0.908	0.937	0.946	0.946
2015 Q2	0.902	0.938	0.942	0.945
2015 Q3	0.911	0.935	0.946	0.952
2015 Q4	0.907	0.935	0.954	0.959

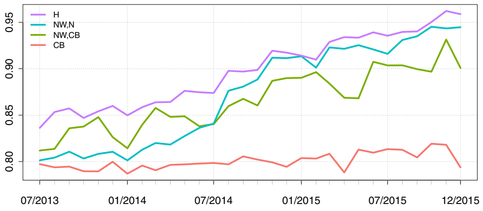


Figure 14: Recall of forecasting models using different feature representation over time. Refer to Section 4.1 for the abbreviations.

We perform risk predictions using the proposed hybrid representation via an ablation test. The AUC (Area under Cure) of the models with different sliding windows are listed in Table 1. As expected, the models using the hybrid feature always outperform other models with naive node-wise feature. It is worth noting that before 2014 Q4, the node-wise and community behavior feature (NW,CB) performs better than node-wise and network (NW,N) feature yet the latter outperforms since 2014 Q4. The recall curves in Figure 14 also reveal such a phenomenon, which perhaps is attributed to the increase of guarantee network complexity over time.

We also compare the prediction importance of node-wise, network, community behavior and our hybrid feature representation. By counting the times each feature is split to a branch of a decision tree in XGBoost regression, we can obtain relative importance of the features. As Figure 15 shows, node-wise feature, community behavior and network feature take opposite trends over time. Initially, node-wise and community behavior features share similar weights and four times more than network features; With the network structure more and more complex, the network feature importance are increased and even account for nearly one-third importance at 2015Q4. This is consistent with the statistics observation that as the guarantee relationships becomes more complex over time, the network centrality related features become more important. Moreover, since node-wise feature only assumes customers are independent, it has weak discriminations when the enterprise are involved in a complex network.

Centrality measurements. We now report some observations derived from the data. Centrality indicators are helpful to identify the relative importance of nodes in the network. Fig. 16 gives the histogram of several of the most complex subgraphs on how the defaults are distributed with different centrality indicator values. It is noted that defaults occur more frequently on nodes with large authority values and small hub values. This is consistent with intuition – if an enterprise works as a hub and backs a large number of other corporations, it can be supposed that it is relatively stable and operates in good condition. Conversely, if an enterprise works as an authority and accepts guarantees from many other corporations, this is an indication that it lacks funding security and is at a higher risk of trouble. The statistics signal to the lender that it should watch the status of the "authority" high nodes in the guarantee network. Although the underlying assumption of PageRank is quite like the authority score, we did not observe a similar correlation between the values and the default rates (see Fig. 16).

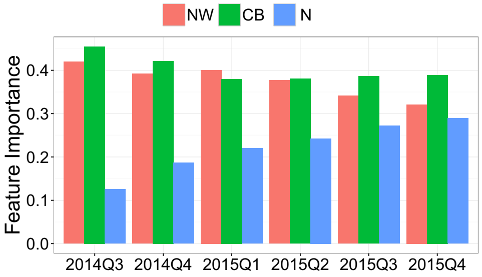


Figure 15: Feature importance score from 2014Q3 to 2015Q4. Refer to Section 4.1 for the abbreviations.

Community ID	1	2	3	4	5	32	33	34	35
Firms	44	42	35	19	29	4	3	4	4
Defaults	14	6	3	5	5	1	1	0	0
Ratio for default firms	32%	14%	9%	26%	17%	25%	33%	0	0
Ratio for default amount	68%	37%	4%	92%	83%	72%	100%	0	0
Structural hole spanner	7	3	5	2	2	1	1	1	1
Neighbour communities	5	3	4	3	2	1	1	1	1
Total loan amount	1071	518	1503	292	1282	18	48	57	105
Total default amount	733	190	62	270	1065	13	48	0	0

Table 2: Statistics for communities generated by the random walk community detection algorithm [29].

However, it is difficult to reliably quantify the correlation of graph centrality indicators with enterprise defaults, in this case, interactive analytics tools provides the possibility to fuse the financial expert domain knowledge with the data-driven indicators. In the multi-faceted risk visualization interface (see Fig. 3) , different risks are highlighted by various diameter spheres and the users are able to explore enterprises from different point of views. This will help they make a better decision in the following analysis tasks.

4.2 High Default Group Detection

High default group detection can reduce analysis scope and thus further help risk pattern discovery and it usually includes automatically community detection and interactive community editing. The experiments is performed on a independent guarantee network with 116 nodes. It is first automatically divided it into 36 communities. The statistics are given in Table 2.

We edit the community following basic guidelines: (1) consider default status, loan amount, and other financial statistics comprehensively; (2) small communities can be either merged with large neighboring communities or pruned. For example, communities 35 and 34 both have four nodes and these firms never default. There is a low possibility that they will become high default groups in the future. Conversely, community 23 has eight nodes, three of which have a default history. They could be merged with the neighboring communities. (3) Structural hole spanner nodes should be given special attention. Usually, defaults happen on the structural hole spanners, so the adjacent communities can be merged. Finally, we obtain ten communities, seven of which have relatively high default rates as Table 3 and as Fig. 17. The seven medium-sized groups of subgraphs can be efficiently processed for further tasks. It is noted that the merge and reassign operations are based on user expertise and the user may choose various criteria, the final treemap

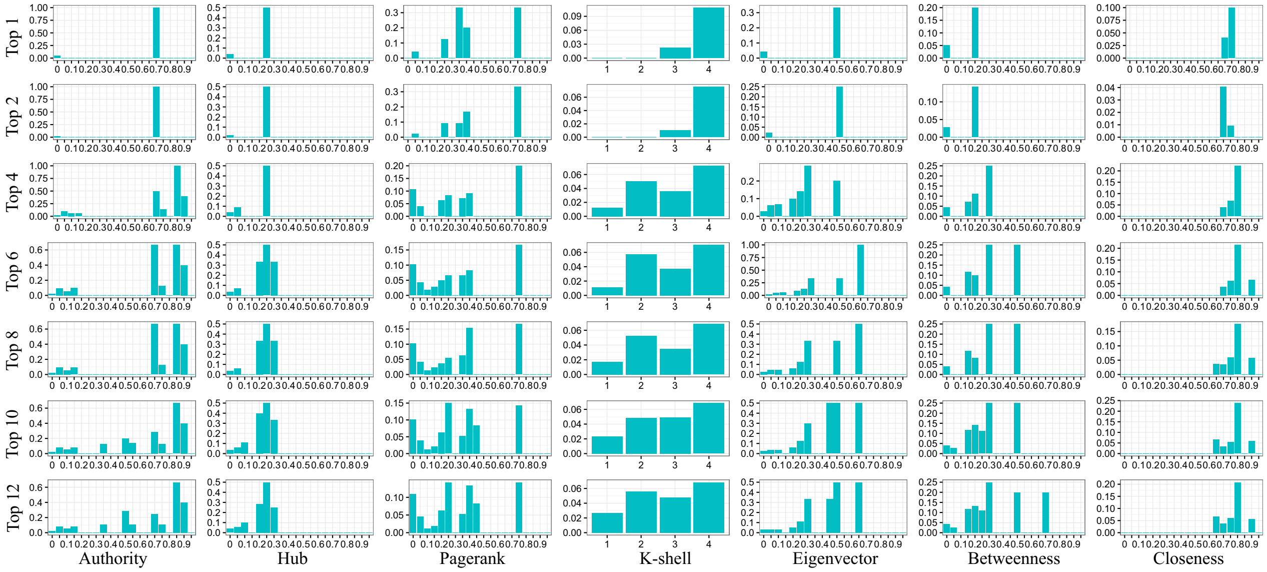


Figure 16: Overdue rates for different graph metric values. From left to right, each column is for a kind of graph metric, namely Authority score, Hub score, PageRank value, K-shell value, Eigenvector centrality, Betweenness centrality, Closeness centrality; From the top down, each row is the most complex independent subgraph.

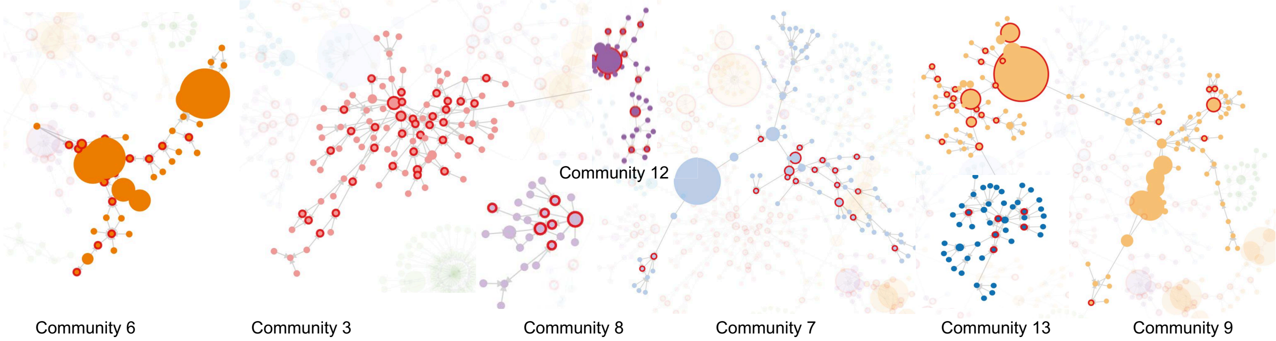


Figure 17: High default groups after interactive editing.

Community ID	13	12	3	6	7	8	9
Firms	46	36	103	44	88	25	128
Defaults	6	11	37	17	18	7	25
Ratio for default firms	13%	31%	36%	39%	20%	28%	20%
Ratio for default amount	31%	97%	85%	41%	40%	78%	51%
Structural hole spanner	4	1	17	3	7	1	5
Neighbour communities	2	1	6	1	2	1	2
Total loan amount	623	826	1695	2080	2273	512	4045
Total default amount	191	804	1441	863	918	398	2083

Table 3: Statistics for communities after interactive editing.

can demonstrate different combinations and default rates.

4.3 Risk Guarantee Patterns Discovery

With the high default groups, we are able to focus and explore risk patterns more efficiently. This includes (1) automatic motif detection from high default groups. Specifically, we employ the gtrieScanner (http://www.dcc.fc.up.pt/gtries/) approach. (2) Matching the motifs with the entire network and calculating the ratio for default firms. (3) Ranking the motifs in descending default order, and they are high default patterns. (4) The user interactively edits the high default patterns by adding more nodes, and the system will automatically match the new subgraph with the entire network and produce the ratio for default firms.

Matching all those motifs on the whole network would be time-consuming. Theoretically, there are 199 and 9364 possible combinations for 4- and 5-vertex motifs for a directed network, respectively. We start from the 4-vertex-motifs and by interactively editing risk motifs, the user can explore more complex patterns efficiently. In the case study, we choose to analyze community 3, which consists of 103 enterprises; 36% of them default the 85% loans from the bank, as Table 3 shows. Fig. 18 gives the twenty 4-vertex-motifs automatic algorithm detected from community 3, and Table 4 shows the statistical information.

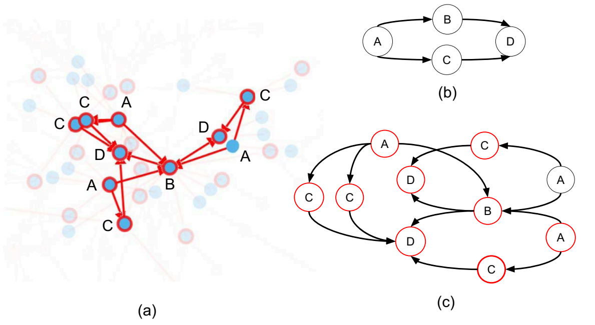


Figure 19: (a) Pattern 15 highlighted on the loan guarantee network. (b) Pattern 15 model. (c) Alternative way to understand pattern 15.

Although there are nearly 200 kinds of 4-vertex node motif shapes, only 20 exist in the high default group. We thus perform analysis only on the 20 motifs rather than on every shape. Most of them have rather complex structures; however, some of them are known to banking experts – for example, motif 6 is a joint liability loan. Some others can be understood by a combination of smaller guarantee

Motif ID	19	15	20	16	17	8	3	7	10	14	4	12	5	18	13	11	2	1	6	9
Motifs	1	4	1	4	4	74	169	92	23	6	164	17	151	1	13	22	312	437	95	24
Firms	4	10	4	28	18	165	238	179	125	24	202	101	304	25	106	138	410	522	478	176
default firms	4	9	3	18	11	79	110	69	48	9	70	35	89	7	28	32	95	111	79	26
Ratio for default firm	100	90	75	64	61	48	46	39	38	38	35	35	29	28	26	23	23	21	17	15
Ratio for default amount	100	100	100	55	75	56	53	71	45	47	59	37	58	24	31	49	64	49	46	44
Total loan amount	36	78	64	955	218	3259	5442	3583	3602	263	4872	3157	6975	1364	3134	4930	8919	11963	10546	3433
Total default amount	36	78	64	522	163	1829	2897	2547	1607	123	2871	1166	4072	331	970	2405	5686	5822	4836	1507

Table 4: Statistical information for the high default motifs.

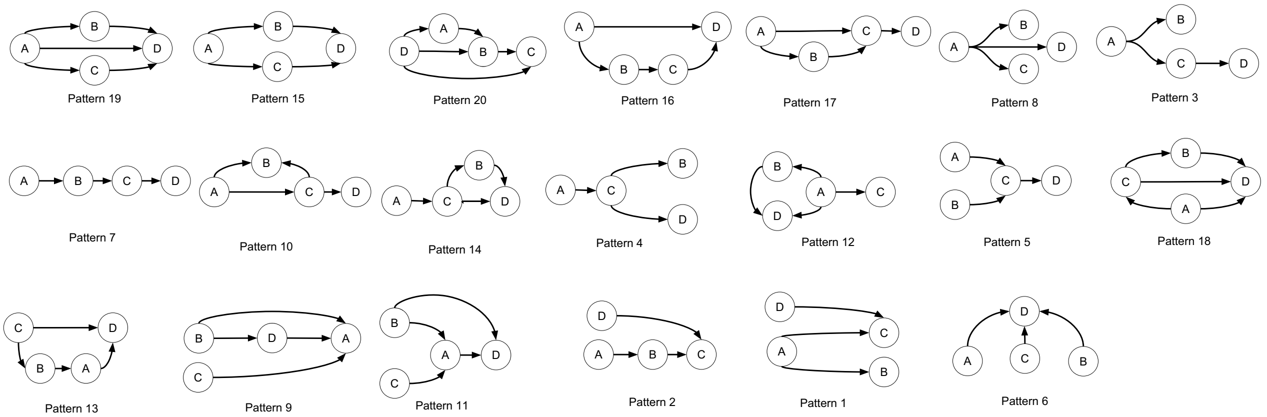


Figure 18: All the patterns (4-vertex-motif structures) detected from community 3. Among them, patterns 15, 16, and 17 show single-input, single-output, and feed-forward structures.

patterns. For example, motif 5 is a combination of a joint liability guarantee loan with a single guarantee. Three of the motifs, 15, 16, and 17, attracted our attention for a number of reasons: (1) high default rates for the patterns (ranging from 61% to 90% in the ratio for default firm and 55% to 100% in the ratio for default amount); (2) a relatively small number of instances (4 or 5) are detected from the whole network; (3) the top five risk motifs show *single input, single output, feed forward structures*. Fig. 19 gives all the instances of pattern 15 that are detected from the entire network. Some of the nodes coincide together. These three patterns are interesting; for example, where pattern 15 occurs five times in a group, the bank lost *all* the money lent to the enterprises with such guarantee structures (see Table 4). There is a high possibility that a fraud loan guarantee may happen several times, and the local bank failed to recognize the fraud pattern. A similar analysis implies that patterns 16 and 17 may also be risk patterns.

5 User Study

We conduct interviews with two banking loan experts. Expert-A comes from the financial regulator. He has more than five years of experience in guarantee network research and has published several important investigation reports and books on the topic. He is also the expert who together with us to consolidate the four research tasks. Expert-B comes from our cooperated bank. He has ten years of loan approval experience and is able to access the complete data set. Both interviewees are attracted by and immediately understand the force-directed graph based view, however, they have difficult to further explore more functions. So, we give them both a 15 minutes training, introducing the main tasks, motivation and the operations of the interface. The interviewees could ask questions and operate the interface to warm up. Then, the interviewees are required to run the tool in 30 minutes and write down their feedback.

Expert-A is familiar with all the four tasks. In the Task 1, he agrees the indicators are useful but suggest the interface should be reorganized with buttons, as the current drop-down menu is a bit difficult to choose. In the Task 2, he is rather interested in the community editing. He said that when he and his colleagues try to resolve the financial risks in guarantee networks, a major operation is to split the loan guarantee network into smaller ones in order to avoid the default diffusion. The editing function of the tool provides users with a powerful weapon to achieve their target. He agrees on illegally conveyed benefits might exist under the suggested risk patterns. In the Task 3, he likes the animation but also suggested the dynamic information needs further investigation. In the Task 4, he suggested that the diffusion path and the corresponding Sankey diagram are useful, but better diffusion model should be developed. Finally, he suggested, the four sub-tools should be reorganized and integrated into one view so that that they can maximize the potential for the ultimate risk isolation operations.

Expert-B expressed that with the tool he is able to grasped the intricate connections between enterprises clearly when assessing a loan. In the Task 1, He likes the force-directed graph based monitoring view but expressed concern about visual clutter issue. He mentioned that in practice, the guaranteed network size could be as large as thousands nodes (although it is very rare) and in such case, it would be difficult to visualize them in the naive force-directed graph. In the Task 2, he thinks the treemap gives an intuitive understanding of the guarantee groups; he likes the financial radar view which we did not expect. He is very interested in the discovered risk patterns. He noted the ID of the nodes and investigated in-depth what had happened. Two weeks later, he sends us feedback that in pattern 15 (see Fig. 19), all default enteritises were sued in court one after another a year ago. With the given names, we confirm that the enterprises are mostly in printing or related industry. However, because the businesses are very small (three to five employees on average), and the information is not transparent, we are not able to dig more. In the Task 3, he expressed that the animation is rather intuitive. It helps to understand how the network was generated but lacks a strong connection with other tools. In the Task 4, he understands the diffusion path and meaning of the Sankey diagram. Because of the tool are currently could only analysis preloaded data, he suggests we further develop the interface and tests the tool on more data set.

Discussion. The above case studies and domain expert interviews confirm the effectiveness of the system in networked-guarantee loan risk management. We also notice there are some limitations. (1) Visual clutters. The case studies are performed on an independent subgraph with more than 600 nodes, the experts are able to zoom in to see the details on ordinary laptops without difficulty. In practice, extreme complex independent subgraphs are very rare and obey the power law [27]. Our statics shows 85.1% are graphs with fewer than 50 vertexes while about 6.6% are graphs composed of more than 300 vertexes in a real dataset. So, the current system can be applied to the majority networked-guarantee loan risk management tasks. However, analyzing the large guarantee networks is also important, we believe classic graph simplification algorithms (for example, community-based clustering) may help to reduce the visual clutter and improve the visualize performance. (2) Visual interface optimization. The current system has separate sub-tool views for different tasks and the operation are relatively complex even for domain expert. Since we have conducted case study with domain experts, next, we will introduce it to the visual analytics experts and perform pair analytics. With the feedback, we will optimize the visual analytics work flow and the interface around risk isolation–the ultimate goal. (3) Default diffusion prediction. All the vulnerable nodes are highlighted in the current system and it will be inevitably introduce misjudgment. Future work will include computational modeling of default diffusion.

6 Conclusion

We present a visual analytics approach for networked-guarantee loan risk management. To our best knowledge, this is the first work using visual analytics approaches to address the guarantee loan default issue. It can help the government and banks to monitor default spread status and can provide insight for taking precautionary measures to prevent and dissolve systematic financial risk.

Acknowledgments

We would like to thank the anonymous reviewers for their useful feedbacks. This research was sponsored by the Open Project Program of the State Key Lab of CAD&CG (Grant No.A1824), Zhejiang University and NVIDIA Corporation GPU Grant.

References

[1] F. Allen, A. Babus, and E. Carletti. Financial connections and systemic risk. Technical report, National Bureau of Economic Research, 2010.

[2] D. Archambault, H. Purchase, and B. Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552, 2011.

[3] D. Archambault and H. C. Purchase. Can animation support the visualisation of dynamic graphs? *Information Sciences*, 330:495–509, 2016.

[4] B. Baesens, R. Setiono, C. Mues, and J. Vanthienen. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science*, 49(3):312–329, 2003.

[5] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

[6] S. Battiston, M. Puliga, R. Kaushik, P. Tasca, and G. Caldarelli. Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific reports*, 2:srep00541, 2012.

[7] D. Bisias, M. Flood, A. W. Lo, and S. Valavanis. A survey of systemic risk analytics. *Annu. Rev. Financ. Econ.*, 4(1):255–296, 2012.

[8] S. Bougheas and A. Kirman. Complex financial networks and systemic risk: A review. In *Complexity and Geographical Economics*, pp. 115–139. Springer, 2015.

[9] R. S. Burt. Structural holes and good ideas. *American journal of sociology*, 110(2):349–399, 2004.

[10] R. S. Burt. Secondhand brokerage: Evidence on the importance of local structure for managers, bankers, and analysts. *Academy of Management Journal*, 50(1):119–148, 2007.

[11] M. Catanzaro and M. Buchanan. Network opportunity. *Nature Physics*, 9:121–123, 2013.

[12] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pp. 155–162. IEEE, 2007.

[13] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM, 2016.

[14] M. Dumas, M. J. McGuffin, and V. L. Lemieux. Financevis. net: A visual survey of financial data visualizations. In *Poster Abstracts of IEEE Conference on Visualization*, vol. 2, 2014.

[15] L. He, C.-T. Lu, J. Ma, J. Cao, L. Shen, and P. S. Yu. Joint community and structural hole spanner detection via harmonic modularity. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 875–884. ACM, 2016.

[16] M. L. Huang, J. Liang, and Q. V. Nguyen. A visualization approach for frauds detection in financial market. In *Information Visualisation, 2009 13th International Conference*, pp. 197–202. IEEE, 2009.

[17] M. Jian and M. Xu. Determinants of the guarantee circles: The case of chinese listed firms. *Pacific-Basin Finance Journal*, 20(1):78–100, 2012.

[18] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[19] C. Klukas, F. Schreiber, and H. Schwöbbermeyer. Coordinated perspectives and enhanced force-directed layout for the analysis of network motifs. In *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60*, pp. 39–48. Australian Computer Society, Inc., 2006.

[20] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[21] F. Lindskog et al. *Modelling dependence with copulas and applications to risk management*. PhD thesis, Master Thesis, ETH Zürich, 2000.

[22] E. Maguire, P. Rocca-Serra, S.-A. Sansone, J. Davies, and M. Chen. Visual compression of workflow visualizations with automated detection of macro motifs. *IEEE transactions on visualization and computer graphics*, 19(12):2576–2585, 2013.

[23] G. N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

[24] D. Mcmahon. Loan guarantee chains in china prove flimsy. *The Wall Street Journal*, 27, 2014.

[25] X. Meng, Y. Tong, X. Liu, Y. Chen, and S. Tan. Netrating: Credit risk evaluation for loan guarantee chain in china. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pp. 99–108. Springer, 2017.

[26] X. L. X. Meng. Credit risk evaluation for loan guarantee chain in china. 2015.

[27] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[28] Z. Niu, R. R. Martin, F. C. Langbein, and M. A. Sabin. Rapidly finding cad features using database optimization. *Computer-Aided Design*, 69(C):35–50, 2015.

[29] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[30] S. Rudolph, A. Savikhin, and D. S. Ebert. Finvis: Applied visual analytics for personal financial planning. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 195–202. IEEE, 2009.

[31] P. Sarlin. Clustering the changing nature of currency crises in emerging markets: an exploration with self-organising maps. *International Journal of Computational Economics and Econometrics*, 2(1):24–46, 2011.

[32] P. Sarlin. Sovereign debt monitor: A visual self-organizing maps approach. In *Computational Intelligence for Financial Engineering and Economics (CIFEr), 2011 IEEE Symposium on*, pp. 1–8. IEEE, 2011.

[33] P. Sarlin. Chance discovery with self-organizing maps: Discovering imbalances in financial networks. *Advances in Chance Discovery*, pp. 49–61, 2013.

[34] T. von Landesberger, S. Diel, S. Bremm, and D. W. Fellner. Visual analysis of contagion in networks. *Information Visualization*, 14(2):93–110, 2015.

[35] T. von Landesberger, M. Görner, R. Rehner, and T. Schreck. A system for interactive visual analysis of large graphs using motifs in graph editing and aggregation. In *VMV*, vol. 9, pp. 331–340, 2009.

[36] J. Yan, M. Cho, H. Zha, X. Yang, and S. M. Chu. Multi-graph matching via affinity optimization with graduated consistency regularization. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 38(6):1228–1242, 2015.