密度優先探索に基づくコミュニティ抽出と入札データ分析への応用

Community Extraction Based on Density-first Search and Its Application to Bid Data

狩山和亮 [1*]　Marco Cuturi[1]　山本章博 [1]
Kazuaki Kariyama　Marco Cuturi　Akihiro Yamamoto

久保山哲二 [2]　福元健太郎 [2]
Tetsuji Kuboyama　Kentaro Fukumoto

[1] 京都大学大学院情報学研究科　[2] 学習院大学
Graduate School of Informatics, Kyoto University　Gakushuin University

Abstract: In this research, we propose a new biclustering method for extracting communities from binary matrices which represent a binary relation. A binary relation can be represented as a bipartite graph or a binary matrix. Many effective clustering methods for extracting communities from graphs and matrices have been proposed. In this paper, the objective data is a bid data which represent a participation record of companies in bids. A community in bid data means a set of companies which often participated in multiple bids. We aim at applying the community extraction to finding bid rigging groups. In order to achieve the goal, we propose a biclustering method based on the density of bipartite graphs and the characteristic extraction by the nonnegative matrix factorization.

1 Introduction

In recent years, in the field of data mining, many researchers have investigated various methods for extracting communities from a relational data, which represent binary relations. In this research, a relational data means a data representing links among several entities. Extracting communities means extracting groups whose members share some characteristics without concrete attributes. Methods developed for such extraction of communities can be applied to coauthor networks among scientists [11] and protein interaction prediction [12].

A binary relation can be represented as a bipartite graph. In terms of graphs, a community is a set of nodes with dense links. Extracting communities means extracting such communities from a given bipartite graph. A binary relation can be also represented as a matrix. Extracting communities from matrices is called biclustering. Biclustering is clustering of the rows and columns of matrices simultaneously.

In this research, the objective data of clustering is a data which represent a participation record of companies in bids for public works. Our goal is to propose a new biclustering method applicable to extracting communities from a relational data which represent a participation record of companies in bids for public works. We call the relational data *bid data.* We represent a relation between *companies* and public works (*items*) as a bipartite graph or a matrix. A relation between a company and an item represents that the company participated in the bid for the item. In bid rigging, a community is a set of companies which often simultaneously participate in multiple bids. For extracting communities, we apply the graph-based community extraction or the matrix-based community extraction. We aim at applying the community extraction to finding bid rigging groups.

In recent years, many researchers have proposed methods for analyzing the structure of bid rigging based on real bid data [1, 2, 8]. However, as far as

[*] 連絡先： 京都大学大学院情報学研究科
〒 606-8501 京都府京都市左京区吉田本町
E-mail: kariyama@iip.ist.i.kyoto-u.ac.jp

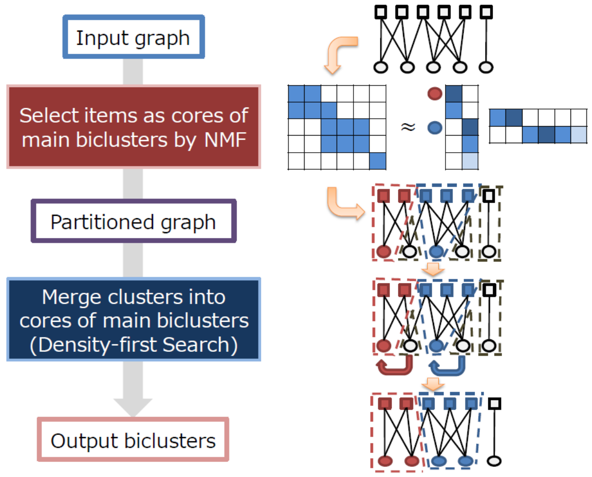

Input graph
Select items as cores of main biclusters by NMF
Partitioned graph
Merge clusters into cores of main biclusters (Density-first Search)
Output biclusters
≈

Fig 1: The overview of our method

the author knows, no methods for extracting companies suspected to join a bid rigging group by using a participation record of bids have been proposed.

In order to develop a new biclustering method applicable to real bid data, we make two assumptions of bids. First, since communities do not always participate in real bids, we assume that bid data can include items in which no community participate. We need not assign all items to extracted main biclusters. Second, we assume that if a community participates in a bid, few companies out of the community participate in the same bid.

Based on the assumptions, we propose a biclustering method for extracting communities from bid data. We focus on the density of bipartite subgraphs and feature extraction by nonnegative matrix factorization (NMF) [10]. We call the biclustering method based on the density of bipartite subgraphs *density-first search*. We show the overview of our method in Fig 1.

2 Preliminary

In this research, we treat binary relations between two sets X and Y of objects. Then a binary relation R is mathematically defined as a subset of $X \times Y$. The relation R is represented as a function $R : X \times Y \to \{0, 1\}$ where $R(x, y) = 1$ if and only if $(x, y) \in R$. Below we assume that $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$.

Table 1: Correspondence of Concepts

Bidding	Binary Relation	Graph	Matrix
Item	Object X	Teams X	Rows X
Company	Object Y	Actors Y	Columns Y
Participation	{0,1}	Existence of an edge	{0,1}

2.1 Biclusters in Bipartite Graphs

In this paper, a bipartite graph which represents a relation is called *a relational bipartite graph*. A bipartite graph G is defined as a triple (X, Y, E) where X and Y are disjoint nonempty sets of nodes and $E \subseteq X \times Y$. We call every element in X *a team*, and every element in Y *an actor*. If a team $x \in X$ and an actor $y \in Y$ are linked by an edge, we say the actor y belongs to the team x. In bids for public works, when we represent a participation record of companies in items as a bipartite graph, a team corresponds to an item, an actor correspond to a company, and a link between an item x and a company y represents that the company y participate in the bid for the item x. A bicluser C of a bipartite graph G is a bipartite subgraph (X', Y', E') such that $X' \neq \emptyset, Y' \neq \emptyset, X' \subset X, Y' \subset Y$, and $E' = \{(x, y) \mid (x, y) \in E, x \in X', y \in Y'\}$. In the paper, we intend to extract dense clusters, and so with the word "bicluster", we sometimes mean a dense cluster.

2.2 Biclusters in Matrices

In this paper, a binary matrix which represents a relation is called *a relational matrix*. Biclustering [5] is clustering of the rows and columns of matrices simultaneously. We consider a $m \times n$ relational matrix $\mathbf{V}$. The element $v_{i,j}$ of $\mathbf{V}$ is equal to 1 if and only if $(x_i, y_j) \in R$, otherwise 0. Let $I \subseteq X$ and $J \subseteq Y$ be the subsets of the rows and columns. The submatrix of $\mathbf{V}$ with a set of rows I and a set of columns J is denoted by $\mathbf{V}_{IJ}$. A bicluster $\mathbf{V}_{IJ}$ is a submatrix of the form $\mathbf{V}_{IJ}$ where rows in I are correlated across all columns in J, and vice versa.

2.3 Communities in Bid Data

In this research, *a bid data* is a data which represent a relation between items and companies in bids. If a company y_j participated in the bid for an item x_i, the relation $R(x_i, y_j) = 1$. We define a community in bids as a set of companies which often participate in multiple bids simultaneously. We summarize which concept in bid, relation, graph, or matrix corresponds to a concept in the other in Table 1.

3 Proposed Method

We propose a bottom-up approach to search and merge clusters of nodes in a relational bipartite graph greedily based on the density of bipartite subgraph. We call the method *density-first search*.

3.1 Preprocessing

At the beginning, we select teams as cores of large bicluster with dense links based on NMF. We use integers as identifiers of biclusters, and an identifier is assigned to every team as well as every actor. The identifier assigned to a team x and an actor y is denoted by $ID(x)$ and $ID(y)$, respectively. We assume that a bicluster consists of teams and actors with the same identifier. Let $G = (X, Y, E)$ be given where $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$.

We call a large bicluster with dense links *a main bicluster*. We select k teams $X_c = \{x_{i_1}, \ldots, x_{i_k}\}$ as cores of k main biclusters based on NMF. We let $ID(x_{i_l}) = l$ and $ID(y_j)$ be the identifier of some x which belongs to $\{x | x \in X_c, (x, y_j) \in E\}$. In order to make the selection deterministic, we let $ID(y_j) = \min_{x \in X_j} ID(x)$ where $X_j = \{x \mid x \in X_c, (x, y_j) \in E\}$. We let $C_1 = (\{x_{i_1}\}, Y_1, E_1), \ldots, C_k = (\{x_{i_k}\}, Y_k, E_k)$, where $Y_l = \{y_j \mid (x_{i_l}, y_j) \in E, y_j \notin \bigcup_{l'<l} Y_{l'}\}$ and $E_l = \{(x_{i_l}, y) \mid y \in Y_l, y \notin \bigcup_{l'<l} Y_{l'}\}$ for $l = 1, \ldots, k$. We assign identifiers to the other nodes. For $x_i \notin X_c$ and $y_j \notin Y_1 \cup \ldots \cup Y_k$, we let $ID(x_i) = -i$ and $ID(y_j)$ be the identifier of some x_i which belongs to $\{x_i \mid x_i \notin X_c, (x_i, y_j) \in E\}$. In order to make the selection deterministic, we let $ID(y_j) = \max_{x \in X_j} ID(x)$ where $X_j = \{x \mid x \notin X_c, (x, y_j) \in E\}$.

3.1.1 NMF

In the density-first search, we merge biclusters into cores of main biclusters. If we select teams in only one main bicluster as cores, we cannot extract the other main biclusters. In order to avoid the situation, it is desirable that actors which belong to each selected team are different respectively, that is, row vectors of the relational matrix corresponding selecting teams are as uncorrelated as possible. It is because we apply NMF to selecting such teams.

NMF [10] is one of the low rank approximation methods focusing on the analysis of latent characteristics of data matrices. It factorizes a nonnegative matrix $\mathbf{V}$ into two nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$. More precisely, the algorithm factorizes a data matrix $\mathbf{V} = \{v_{i,j}\}_{m \times n}$ into a basic matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k] = \{w_{i,l}\}_{m \times k}$ and a coefficient matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_k]^{\mathrm{T}} = \{h_{l,j}\}_{k \times n}$ so that the following approximate equation is established:

$$\mathbf{V} \approx \mathbf{WH}.$$

For NMF, the coefficient vectors $\mathbf{h}_1, \ldots, \mathbf{h}_k$ tend to be independent due to nonnegative constraints. In this research, we use the result of NMF for a relational matrix to the selection of teams. Regarding the ith column vector of the basic matrix $\mathbf{W}$ as vectors which represent weights of teams in a main bicluster, we select a team with the maximal value in the ith column vector for a main bicluster i.

NMF minimizes the difference of $\mathbf{V}$ and $\mathbf{WH}$. More precisely it is formulated as the problem solving

$$\mathbf{W}, \mathbf{H} = \underset{\mathbf{W},\mathbf{H} \geq 0}{\operatorname{argmin}} ||\mathbf{V} - \mathbf{WH}||^2, \tag{1}$$

where the formulation of NMF defining the objective function as the Frobenius norm of the error matrix. For solving this problem, multiplicative update rules [10] derived from an auxiliary function method are widely used. Since the algorithm initializes $\mathbf{W}$ and $\mathbf{H}$ at random, the solutions of the algorithms are affected by the initialized matrices. In order to solve this problem, we apply the initialization based on the singular value decomposition [4] .

3.2 Density-first Search

Selecting the teams based on NMF, we apply the density-first search to the relational bipartite graph. The relational bipartite graph is partitioned into k cores of main biclusters and $(m - k)$ other biclusters. We define the density D_l of the bicluster (X_l, Y_l, E_l) as

$$D_l = \frac{|E_l|}{|X_l| \cdot |Y_l|}.$$

Our proposed method is a bottom-up approach that we merge biclusters into cores of main biclusters such that the density of the obtained biclusters keeps high.

We define *a neighbor team* of a bicluster i as a team x which belongs to $\{x \mid \exists y \in Y_i, (x, y) \in E\}$. We define *the candidate clusters* of the bicluster i as clusters which include a neighbor team of the bicluster i. For $i = 1, \ldots, k$, we merge one of the candidate cluster of C_i into C_i such that the density of the obtained bicluster is highest. However, if the density is lower than a given threshold σ, we merge no clusters into C_i. We iterate the merging procedure. When no clusters can be merged into any biclusters, we output the identifiers assigned to all teams and actors.

4 Experiment

We tried an experiment for comparing the performance of our method with existing methods for extracting communities by using random bipartite graphs with explicit cluster structure and real bid data.

4.1 Evaluation with Random Bipartite Graph

Applying the proposed method in [7], we generated random bipartite graphs with explicit cluster structure. The procedure to generate random bipartite graphs consisting teams and actors is as follows:

1. Assign S actors to each cluster i for $i = 1, \ldots, N_C$ and cluster IDs $1, \ldots, N_C$ to all actors.
2. Generate a team a. Set the number of actors m_a belonging to a and the cluster ID of a. Select an ID from all the cluster IDs at random.
3. To each team, apply the followings:
    - With probability q, link edges between Team a and m_a actors according to the following two rules:
        - With probability p, link an edge between Team a and a random actor belonging to the cluster with the same cluster ID as the team.
        - With probability $1 - p$, link an edge between Team a and a random actor in all actors.
    - With probability $1 - q$, link edges between Team a and random m_a actors.
    - Generate N_t teams by iterating 2 and 3.

The *team homogeneity* p means the probability of selecting clusters with the same cluster ID, and The *cluster property* q means the probability of generating teams with cluster structure. If $p = 1$, all actors belong to a team are included in the same cluster. If $q = 1$ and p is high, all teams have implicit cluster structure. Corresponding to the concepts of bidding, the ratio of companies out of a community increases as p decreases. Moreover, the less q becomes, the more items in which no community participate are included in the participation record.

For random bipartite graphs with explicit cluster structure, We evaluated the performance of clustering based on the Louvain method (Louvain) and our methods (NMF.DFS) by using the normalized mutual information (NMI) [6]. Generating the graphs, we changed the two parameters: The team homogeneity p and the cluster property q. We let the other parameters be unchanged: $N_C = 10$, $S = 10$, $N_t = 100$, and $m_a = 8$. We generated 10 graphs for each p and q. In our method, the threshold of the density of merged biclusters in the density-first search $\sigma = 0.4$ and the number of cores $k = 10$. The number of iterations in the NMF algorithm was 100. We used the algorithms of the Louvain method[1] implemented by the author of the paper [3] .

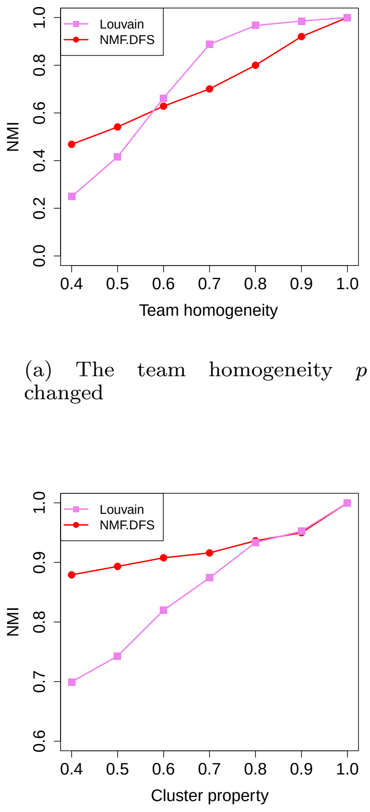
We investigated how the team homogeneity p and the cluster property q affect the NMI of the algorithms. When we changed p, we set $q = 1$. When we changed q, we set $p = 1$. Figure 2 shows the average of the NMI. For $p \leq 0.5$ or $q \leq 0.7$, NMF.DFS showed the better NMI than the Louvain method. As the result, we conclude that the NMI of the density-first search is better than the Louvain method for the bipartite graphs which include many data unrelated to communities.

4.2 Evaluation with Real Bid Data

We compared the performance of our method with existing methods for extracting communities by using real bid data. We applied each method to real bid data. We judged how likely bid rigging occurred in bids by the average *bid acceptance ratio.* The bid acceptance ratio of a bid is calculated as the ratio of the successful bid price to the upper limit of the bid price. The National Liaison Council of Ombudsman Groups[2] regards that if the bid acceptance ratio of a

[1] http://perso.uclouvain.be/vincent.blondel/research/louvain.html

[2] http://www.ombudsman.jp/dangou/



(a) The team homogeneity p changed

(b) The cluster property q changed

Fig 2: The NMI

bid for a public work is no less than 95%, it is strongly suspicious that a bid rigging group participates in the bid, and that if it is 90% to 95 %, a bid rigging group are supposed to participate in the bid.

As real bid data, we target the bid data for public works held in Chiba City in fiscal 2005 to fiscal 2009. Table 2 shows the number of items for which bids were held in each fiscal year, and the total number of companies which participated in the bids.

We evaluate the performance of the Louvain Method [3], the IRM [9], and NMF.DFS for extracting communities from the bid data. We define *an item cluster* I_t as a set of items in the bicluster t extracted by a biclustering method. Let $J_t = \bigcup_{i \in I_t} J_i$, where $J_i = \{j \mid (i,j) \in E\}$. If $|J_t| \geq 50$, we say that I_t is large. If the number of items in I_t is no more than one, we say that I_t is small. We defined the density of the item cluster I_t as follows:

$$\frac{\sum_{i \in I_t} c_i}{|I_t| \cdot |J_t|},$$

Table 2: The bid data in Chiba City

Fiscal year	#Item	#Company
2005	634	815
2006	528	669
2007	545	562
2008	363	452
2009	197	264

where c_i is the number of companies which participated in the bid for the item $i \in I_t$.

When the density of the item cluster I_t was very low, it would be hard to regard the set of companies J_t as a community. Assuming that the density of large item cluster is low, we investigated the average density of large item clusters. Similarly, it would be hard to guess that the participating companies in a small item cluster are members of a community. We counted the number of item clusters whose average bid acceptance ratio $\geq$ 95% and $\geq$ 90%, not including large and small item clusters. We investigated the following:

- # of item clusters (#Cluster)
- # of large item clusters (#Large)
- The average density of large item clusters (Average density)
- # of small item clusters (#Small)
- # of item clusters whose average bid acceptance ratio $\geq$ 95% (BAR $\geq$ 95%)
- # of item clusters whose average bid acceptance ratio $\geq$ 90% (BAR $\geq$ 90%)

We used the algorithms of the IRM[3] implemented by the author of the paper [9]. In our method, we used only main biclusters for the experiment. We set the threshold of the density of merged clusters in the density-first search $\sigma = 0.4$ and the number of cores $k = 20$.

We show the average of each value in fiscal 2005 to fiscal 2009 in Table 3. The Louvain method and the IRM extracted more large item clusters with low density than NMF.DFS. It would be because the Louvain method and the IRM cannot identify submatrices without cluster structure, and they extract submatrices as a very large cluster. Removing large and

[3] http://www.psy.cmu.edu/~ckemp/code/irm.html

Table 3: Evaluation with real bid data

Fiscal year	#Cluster	#Large	Average density	#Small
Louvain	31.6	6.4	0.105	15.8
IRM	11.8	6.8	0.109	0
NMF.DFS	20	0.4	0.405	0

(a) The detail of the partitioned sets of items

Method	BAR $\geq$ 95%	BAR $\geq$ 90%
Louvain	2.2	7.8
IRM	0.8	2.8
NMF.DFS	5.4	13.8

(b) The result

small item sets, NMF.DFS extracted more item clusters whose average bid acceptance ratio is no less than 95% and 90% than the other methods. We conclude that local search like the density-first search is effective in extracting communities from the bid data. Moreover, we can extract sets of items with higher density by changing σ based on NMF.DFS.

5 Conclusion

We have proposed a biclustering method for extracting communities from matrices which represent a participation record of companies in bids for public works based on the density of bipartite subgraphs and the characteristic extraction by NMF. We have presented that the density-first search is effective in extracting reasonable communities from bid data.

Though the bid acceptance ratio of bids in which bid rigging groups participate is liable to be high, it is not an evidence of bid rigging. This research contributes to the empirical analysis of bid rigging based on the approach of extracting communities. However, we assume that we need more background knowledges in order to make the approach practicable.

References

[1] Masaki Aoyagi. Bid rotation and collusion in repeated auctions. *Journal of Economic Theory*, 112(1):79–105, 2003.

[2] Patrick Bajari and Lixin Ye. Deciding between competition and collusion. *Review of Economics and Statistics*, 85(4):971–989, 2003.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[5] Yizong Cheng and George M Church. Biclustering of expression data. In *Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology*, pages 93–103, 2000.

[6] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[7] Roger Guimerà, Marta Sales-Pardo, and Luís A Nunes Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3):036102, 2007.

[8] Rieko Ishii. Favor exchange in collusion: Empirical study of repeated procurement auctions in japan. *International Journal of Industrial Organization*, 27(2):137–144, 2009.

[9] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

[10] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

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

[12] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.