Cluster and network analysis for business models in the Mexican banking system

Dr. Serafín Martínez Jaramillo
Content*

- Introduction
- Business models
- Clustering methods and random forests
- Financial networks and stochastic block models
- Results

*The views expressed in this presentation are exclusively the responsibility of the author and do not necessarily reflect those of CEMLA or Banco de México.
Introduction
General overview

- The last crisis show the need from the regulators to understand the role that individual banks play in the financial system.

- In this work, two possible forms were used to identify the role played by a bank in the system: Cluster and Network analysis.

- The first step is to identify clusters which can be associated with business models, through high dimensionality reduction techniques (Factorial and Reduced K-Means) and then we use random forests to identify the most important variables.

- The second aim is to study the role played by the banks in the banking system network, this was done by the application of Stochastic Block Models.
New approaches to study the banking system

- Banking systems have become more complex in recent times, due to aspects as the evolution of Banks, the arrival of new players and the emergence/transformation of markets, and these changes could have left possible blind spots not covered yet.

- Also, the high level of interconnectedness in the banking system is related to the fact that exists many channels of interaction (lending, buying others banks’ securities, engaging in foreign exchange transactions, etc.) and markets (repurchase agreements, derivatives, among others).

- A new set of techniques that study the banking system have recently gained acceptance among central bankers under the wide umbrella of Machine Learning.
  - Some of their functions are: pattern recognition, object classification and function approximation.
Business models
Business Models for financial institutions

- A business model is translated into the activities that such institution performs in order to obtain funding, provide liquidity to the system and maximize its profits; it also involves the intensity, their geographical reach and the target of such activities.

- If we have in the system institutions which are Too-Big-To_fail and Too-Interconnected-To-Fail, and also these institutions have similar business models and connectivity patterns, then the banking system is perceived as being fragile. For that reason, this study is important.

- As an example, in Lucas et al. (2017), Farn and Vouldis (2017) and León (2017), the authors identify different business models in the first two cases, for the european banking system, and in the last case, for the Colombian banking system, by using clustering methods.
Data

- The data used for this study was low granularity balance sheet information from a regulatory report known as “Catálogo Mínimo”. The time span of the data for this study goes from 2007 to 2017.

- After eliminating variables which have value 0 for all the institutions, we ended up with 400 variables. In a similar way to Farn and Vouldis (2017) an important pre-processing procedure was performed, which consisted in the elimination of highly correlated variables.

- The clustering methods were performed in three different data sets in order to explore the optimal level of granularity. It’s worth to mention the use of an agnostic approach for the identification of groups, with the objective to allowing the data to speak for itself.
Clustering methods and random forests
For the first stage of this study, unsupervised techniques were used (meaning that our training data does not include an output variable of interest), in particular, clustering methods.

The idea behind the application of clustering techniques is to find groups of observations with similar characteristics; but this depends critically on the definition of similarity that is used. For this work euclidean distance were the metric used.

The clustering algorithms used here are designed to work with fath data, which means that we have more variables than objects, being for this case, more variables than banks. The algorithms are both variants from K-Means: Reduced and Factorial K-Means, and looks to reduce the dimensionality of the variables.
Factorial and Reduced K-Means

- K-Means clustering methods aim to separate all the observations into K different groups, maximizing the separation between groups while minimizing the distance of observations within the cluster, in relation to the cluster’s centroid.
  - An important issue for this approach, compared with hierarchical clustering methods, is that the number of clusters must be specified beforehand.

- The algorithm only guarantees convergence to a local minimum, results can change depending on the initial grouping; for this reason it is recommendable to execute the algorithm several times with different initial groupings.

- In addition to obtain groups (based on K-Means), algorithms selected for this study, also reduce the dimensionality of the input data, which is useful for fat data problems. Both methods assume that the groups centroids are located in a sub-space of the space generated by all the variables.
Decision Trees and Random Forests

- Decision Trees are supervised learning techniques and it aims to separate the space of independent variables in exclusive regions, having that the same prediction is assigned for all the observations within a region with the next criteria:
  - Regression case: The prediction will be the average of the response variable for all the observations in a given region.
  - Classification case: The prediction will be the most common class of the response variable for all the observations in a given region.

- Given that the partitions of the dependant variables are not linear, decision trees could become very complex or present overfitting. In order to overcome these issues, random forests were developed. This technique creates bootstrap samples and train a decision tree on each sample, averaging the prediction of every tree trained, reducing the variance of the overall model.

- It is possible to obtain the relevance of the independent variables for the prediction of the dependent variable.
Networks and stochastic block models
Financial Networks

- Network models in finance are now well established. These models are used to measure systemic risk, solvency and funding contagion and also in stress testing. More recently, network models are also being used to understand and model interactions in many different market and activities.

- We can clearly see that the banking exposures network have had important structural changes.

(a) Beginning of the period of study  
(b) End of the period of study
Data

- The data for this study comes from a database at Banco de México, which is used for contagion and systemic risk studies.

- The dataset contains daily current exposures among Banks from 2005 onwards and from 2011 the daily current exposures among financial intermediaries like brokerage houses, investment funds, pension funds and foreign financial intermediaries.

- The current exposures in such a dataset used in this paper come from the following bank activities:
  - Interbank call money loans
  - Cross holding of securities
  - Derivatives trading
  - Foreign exchange transactions
Stochastic Block Models

- Once covered the identification of groups in relation to its business models, another important aspect is to uncover groups of banks which behave similarly in relation to its interbank relationships. Stochastic Block Models is the technique selected in this work to uncover this structure.

- Many of the community detection methods are subject to overfitting and could also be affected by random variations in the connectivity of a system. To overcome these problems new methods use probability theory that translates the concept structural equivalence into a probabilistic framework.

- All the methods which use probability to identify communities in a network use as assumption the existence of $n$ nodes, for which is possible to observe the variable $y_{ij}$ called relationship from $i$ to $j$. 
The most successful approach to model the existent relationships in a network has been to assume that such relationships can be modeled conditioning its probability of existence to a vector $x$ of node attributes, in this case, the $x_i$ entry $i \in \{1; 2; \ldots; n\}$ represents the community to which node $i$ belongs to. In this, it is possible to model the relationship:

$$y = (y_{ij})_{1 \leq i \neq j \leq n}$$

conditional to the vector $x = (x_1, x_2, \ldots, x_n)$ where $i$ belongs to the set $C = \{1, 2, 3, \ldots, c\}$, that contains all the possible groups to which the node can belong to. This is equivalent to think that the probability that two nodes $i$ and $j$ form a link in the network depends on the community that they belong to.
Stochastic Block Models

- In this work is not possible to take an a priori approach, which requires knowledge on the groups’ structure, that we currently don’t have, therefore is needed an posteriori approach to find the groups’ structure by using the relationships $y$, known from the adjacency matrices.

- The model is described as follows:
  - $N$, denotes the set of all the node pairs $(i, j), i, j \in \{1, 2, \ldots, n\}$.
  - Each node $i$ belongs to a class $k, k \in \{1, 2, \ldots, c\}$.
  - We only observe the relationship matrix $Y$, and we want to estimate the groups vector $X$. 
Stochastic Block Models

First, the joint probability of the group of each node is defined, these are the result of the random variables $X_i$. In this way we have:

$$P(X_i = x_i, X_2 = x_2, \ldots, X_n = x_n) = \theta_i^{m_i} \cdots \theta_c^{m_c}$$

Where $m_i$ represents the number of nodes which belong to class $i$ and $\theta_i$ represents the probability that a node belongs to class $i$.

One of the strong assumptions is that the existence of a connection between two nodes depends on the class they belong to, in this form, given the vector $X = x$ we have that:

$$P(Y_{ij} = 1 \mid X = x) = \eta(x_i, x_j)$$

Where $\eta(k, h)$ is the probability of connection that depends on the grouping of $i$ and $j$ and also, the sum of all $\eta(k, h)$ for all $k, h \in C$, it’s equal to 1.
Stochastic Block Models

Moreover, the probability $\eta(k, h)$ can be split in $\eta(k, k)$ and $\eta(k, h)$, that represents the probability of connection between two nodes from the same group and between nodes of different groups respectively. The taking into account the previous partition it is possible to write:

$$
P(y \mid x, \theta, \eta) = \left( \prod_{1 \leq k < h \leq e} (\eta(k, h))^{e(k, h)} \right) \times \left( \prod_{k=1}^{c} (\eta(k, k))^{e(k, k)} \right)
$$

Where $e(k, h)$ is the number of existing relations between blocks $k$ and $h$ that is computed as:

$$
e(k, h) = \sum_{(i, j) \in N} I\{y_{ij} = 1\} I\{x_i = k\} I\{x_j = h\}.
$$
Stochastic Block Models

The model that is needed is given by the joint distribution of the relationships and the groups structure, i.e. the joint distribution of \((Y, X)\), which is given by:

\[
P(y, x \mid \theta, \eta) = \theta_1^{m_1} \cdots \theta_c^{m_c},
\]

\[
\times \left( \prod_{1 \leq k < h \leq c} (\eta(k, h))^{e(k, h)} \right),
\]

\[
\times \left( \prod_{k=1}^{c} (\eta(k, k))^{e(k, k)} \right).
\]

The three components of the previous equation represent, respectively: distribution of the nodes groups, the connections between nodes of different classes and the connections between nodes of the same group.
Stochastic Block Models

Given that only the structure of $Y$ is known, the previous probability can be re-written as:

$$P(y \mid \theta, \eta) = \sum_{x \in C^n} P(y, x \mid \theta, \eta)$$

This is, we sum over all possible combination of classes for the $n$ nodes in the network. In order to apply the previous equation it is necessary to estimate the parameters and using the posterior distribution:

$$f(\theta, \eta \mid y) = \sum_{x} f(\theta, \eta, x \mid y)$$

With these parameters it is possible to estimate the classes vector $x$ using the posterior predictive density

$$P(x \mid y) = \int f(\theta, \eta, x \mid y) d\eta d\theta$$
Stochastic Block Models

Finally, Gibbs sampling is used to obtain the conditional probability

\[ f(\theta, \eta, x \mid y). \]

From which, it is possible to make inference for the two previous equations. In this work the network analysis package *Graph-Tool* (Peixoto, 2014) were used. In Peixoto (2017), the author details the methodology used for the Python package.
Results
Results (Clustering methods)

Results will be presented for different sets; having level N1 as the more general configuration, followed by the level N2, composed by more granular data.

From the image below, it is possible to see the fitting to the data with the highest level of aggregation (lowest granularity), N1. The figure presents two panels, panel a) presents the groups including all the variables at level N1; panel b) presents the groups without including the highly correlated variables. From this figure it is possible to see that the fitting is complex as the subspace is tri-dimensional, some of the groups are not well differentiated.
Results (Clustering methods)

In this figure, a more detailed data set were used. Here, it is possible to see that there are more groups but better separated and if the highly correlated variables are eliminated, the complexity is reduced to two dimensions.
Results (Clustering methods)

We have the highest level of granularity for the next figure. For this case, again the groups are well separated and the elimination of correlated variables reduces complexity.
Results (Clustering methods)

Figure below shows the change in the number of banks and on the composition of the groups for the information set N1 eliminating the highly correlated variables. The classification provided by the Factorial K-Means algorithm does not seem to make sense as it basically produces 1 large group and two very small groups most of the time.
Results (Clustering methods)

The incorporation of more information does not produce interesting results as it produces more groups with many changes across time. It is important to say that a large number of groups is not bad in itself, but for such a small banking system in terms of the number of banks doesn't seem to be right.
Results (Clustering methods)

This figure shows the results for the Reduced K-Means algorithm for the highest information aggregation level. The results show that two dimensions are enough for the identification of the clusters, the number of groups is smaller and the dispersion of the institutions is more realistic.
The second information level shows similar results to the N1 information set. The only difference is the existence of a new group in the case including all the variables and in the case where the highly correlated variables are eliminated.
Results (Clustering methods)

In order to see the evolution of the groups identified by the algorithm we will resort again to the alluvial diagram. Figure below shows the alluvial diagram for the N3 information set. In this figure it is shown that there are important changes in the structure and composition of the groups in December 2007 and March 2008. The number of groups and the banks that belong to such groups is quite stable.
Results (Random Forests)

Once we know to which group each institution belongs, then it is necessary to understand what is causing institutions to belong to a group. For this we resorted to random forest in order to determine the importance of the balance sheet accounts for the groups.
Results (Random Forests)

For the highest aggregation level, N1, figure on the left shows that the most important variables for the classification are the contributed capital and capital gained; after them, current credit portfolio, traditional deposits and securities investments have big jumps. Taking into account dates with important jumps, the jump for the current credit portfolio and securities investment during the second half of 2009 is worth emphasizing. This change in the business model might be related to the Mexican GDP fall around those days.
Results (Random Forests)

Given that the accounts taken in the last figure are very general, the information provided by them is still small. As a consequence of the above result we decided to perform the same exercise for different levels of data granularity.

Table 2: Name of the most important variables from the RKM N2 SinCol adjustment

<table>
<thead>
<tr>
<th>ID</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>118300000000</td>
<td>Contributed Capital - Increase Due To Capital Stock Actualization (4)</td>
</tr>
<tr>
<td>129200000000</td>
<td>Capital Gain - Increase Due To Capital Reserves Actualization (4)</td>
</tr>
<tr>
<td>129400000000</td>
<td>Capital Gain - Increase Due To Previous Years Results Actualization (4)</td>
</tr>
<tr>
<td>129100000000</td>
<td>Capital Gain - Capital Reserves (4)</td>
</tr>
<tr>
<td>410100000000</td>
<td>Contributed Capital - Capital Stock (4)</td>
</tr>
<tr>
<td>211100000000</td>
<td>Traditional Deposits - Term Deposits (2)</td>
</tr>
<tr>
<td>223000000000</td>
<td>Collateral Sold Or Given As Guarantee (2)</td>
</tr>
<tr>
<td>129100000000</td>
<td>Investment In Securities - Titles To Negotiate (1)</td>
</tr>
<tr>
<td>170100000000</td>
<td>Permanent Investments - Subsidiaries (1)</td>
</tr>
<tr>
<td>236300000000</td>
<td>Interbank Loans And From Other Institutions - Long Term (1)</td>
</tr>
</tbody>
</table>

Table 3: Name of the most important variables from the RKM N3 SinCol adjustment

<table>
<thead>
<tr>
<th>ID</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>420900000000</td>
<td>Capital Gain - Previous Years Results - Results Due To Accounting Changes (4)</td>
</tr>
<tr>
<td>420910000000</td>
<td>Capital Gain - Previous Years Results - Results Pending To Apply (4)</td>
</tr>
<tr>
<td>231100000000</td>
<td>Traditional Deposits - Term Deposits - From General Public (2)</td>
</tr>
<tr>
<td>139100000000</td>
<td>Current Loan Portfolio - Commercial Credits - Without Restrictions (1)</td>
</tr>
<tr>
<td>293100000000</td>
<td>Deferred Credits - Commissions Charged On Initial Loan Grant (2)</td>
</tr>
<tr>
<td>221400030000</td>
<td>Securities Trading - Derivatives - For Negotiation Purposes - Options - Valuation (2)</td>
</tr>
<tr>
<td>221400030000</td>
<td>Securities Trading - Derivatives - For Negotiation Purposes - Options (2)</td>
</tr>
<tr>
<td>221400030000</td>
<td>Securities Trading - Derivatives - For Negotiation Purposes (2)</td>
</tr>
<tr>
<td>221400030000</td>
<td>Other Assets - Advanced Payments (1)</td>
</tr>
<tr>
<td>238200000000</td>
<td>Interbank Loans - Short Term - Loans From Development Banks (2)</td>
</tr>
</tbody>
</table>
Results (Random Forests)

It is worth notice that the majority of selected series for the last two figures are sub-levels of the accounts showed in the first level of aggregation. One important aspect is that there is consistency between the results of the determination of the variables importance at the different levels of aggregation. If this was not the case the interpretations as well as the justification would be much more difficult to achieve. Once again, the capital accounts are present and are also the most important ones. However, there are not sub-accounts of the credit portfolio for the first figure of the last slide. The third level of granularity (which correspond to the inferior figure of the last slide) shows the that the sub accounts of the important accounts at the previous level are also important. Moreover, a sub sub account of the current credit portfolio appears to be important: unrestricted commercial credit. Also relevant is the presence of securities investment but on the liability side of the balance sheet. Interbank liabilities accounts are important for all the granularity levels.
Results (Stochastic Block Models)

In order to visualize the stability of the groups obtained by the SBM we use the same alluvial diagram as the ones used for the illustration of the clusters obtained using balance sheet information. Figure shows that changes in the groups are more frequent than for the clustering methods. However, the number of groups is more stable in this approach.
Results (Stochastic Block Models)

In the figure below it is possible to see the relevance of the centrality metrics in the formation of groups. It is not surprising to find that the degree metrics are the most relevant followed by closeness.
Results (Stochastic Block Models)

Now, it is possible to observe why the degree metrics are the most relevant for the classification as the distributions are well differentiated for each group being cluster 1 the one with the most connected nodes and cluster 3 the one with the less connected nodes. Moreover, cluster 1 is the one with the highest closeness to the rest of the network.
Results (Stochastic Block Models)

From the metrics shown in the figure, DebtRank distinguishes itself from the rest given that despite that the mean is not very different among groups, group 1 presents considerable more extreme values than the other groups. This implies that the most systemic banks belong to group 1. Therefore in addition to the existence of similar business models, these banks also share similar connectivity patterns.
Conclusions

- The banks' business models change accordingly with the economic environment.
- The RKM clustering method is sensible to the information set and to extreme values from the input variables.
- For the Mexican case, capital accounts seem to be relevant for the groups created by the method for the whole period of study.
- The credit portfolio and other financial assets (securities and derivatives) are very important in some points in time; for example, during the GFC,
- The use of data with higher degree of granularity can provide better insight on the results.
- Comparing the results obtained by the two different methods it is possible to see that banks in Mexico form groups in a non-trivial way. Therefore, it is necessary to study the banking and the financial system in more than only one dimension.