A network exploratory analysis of firm ownership at subgraph level, in the case of France

A Hazan, A Vialfont

A network exploratory analysis of firm ownership at subgraph level, in the case of France

A. Hazan*, A. Vialfont**

* LISSI/IUT-SF/UPEC - 36/37 rue Georges Charpak, Lieusaint, 77567.
**ERUDITE/FSEG/UPEC, 61 av. du Général de Gaulle 94010 Créteil Cedex.

Abstract

An exploratory analysis of French ownership network is proposed, based on ORBIS dataset. We compare groups of firms that form subnetworks, with approximately half of firms bearing attributes, and focus on their organization. In particular, we aim at characterizing the relation between groups organization and anticompetitive behaviors observed in the French competition authority decisions at the 4 digit sector level.

1 Introduction

The firm ownership network was shown to behave as a complex system, with heterogenous anticipating agents in a time-dependent context. This network can be modelled in a simplified way as a graph G.

Quantifying the notion of corporate control and studying large samples of firms allowed to show that, at the large scale and global level, control is highly concentrated [GB09, VGB11], and firm groups are structured in a highly hierarchical way [HV22], in comparison to other economic networks models.

Our objective is to establish whether this hierarchical structure presents some particularities in terms of topological characteristics and economic concentration ratios when we consider anticompetitive practices observed in the French competition case law.

2 Datasets

In order to analyze the French ownership network, we rely on the ORBIS database from which we have extracted more than 3 millions entities with and without legal personality. Among these entities, we find 1.3 million of links (or edges), denoted by s if the downstream entity is a subsidiary (has legal personality), e for a secondary establishment (no legal personality), and p for a private equity downstream relation (that is a limited partnership with no active role in management).[1]

As basic statistics, we find :

1. a giant component made of 513 k entities. The existence of a giant weakly connected component is coherent with random network theory since $N_e > N_n$, see [BP16]. The second largest connected component has size 1,5 k.
2. We have a total of 210k connected components.
3. The distribution of the size s of non-giant components can be fitted either by a lognormal or an exponential law, because it doesn't span a wide interval of magnitudes ($s \in [10^1; 10^3]$).
4. The large number of firms in non-giant connected component is an important observation, overlooked by mesoscale models (e.g. bow-tie, or core-periphery).

In order to address our objective, we also consider a database constructed out of the Competition authority decisions that provides us with the complete set of anti-competitive practices detected in France since 2000.[2] This database allow us to identify fine grained sectors affected by anti-competitive behaviors.

[1] The construction of this database is known to be complex and we follow [KOSVS+15] to be as representative as possible. In particular, we also plan to compare the generated networks to those associated to the LIFI database used in [HV22].

[2] The database can be found on the Github of the French Competition Auhtority: https://github.com/AutoriteDeLaConcurrence/decisions-adlc.

3 Methods

As explained in 2, the input data is composed first of node records that associate descriptors to a set of firms.

$$\{(n^0, f_0^0, \ldots, f_{F_n}^0), \ldots, (n^N, f_0^N, \ldots, f_{F_n}^N)\}$$

where $\{(n^i\}_{i\in[1,N]}$ are the indexes of the nodes that describe firms, $(f_0^i, \ldots, f_{F_n}^i)$ the F_n node features that describe node i. Node features can be either numeric of categorical.
Furthermore, the dataset includes a set of attributed edges, given by:

$$\{(n_{src}^0, n_{tar}^0, g_0^0, \ldots, g_{F_e}^0), \ldots (n_{src}^E, n_{tar}^E, g_0^E, \ldots, g_{F_e}^E)\}$$

where $((n_{src}^i, n_{tar}^i)$ are the source and target node indexes in edge i. Further, $g_0^i, \ldots g_{F_e}^i$ are a set of F_e edge features corresponding to edge i, that describe the type of relationship between the source and target node. Edge features can be either numeric of categorical. For example, the kind of ownership relationship between two entities can be "subsidiary", "private equity", etc...

Graphs considered in our work are all directed unless explicitly stated. Further, graphs are created from edge files, in such a way that isolated firms are discarded from graph analysis (around 1,8 million of entities in the present state of the database).

Subgraphs are investigated in this work because they reflect group of firms that are meaingful from an economic point of view. Thus, they need to be characterized. Three types of subgraph features are used:

- subgraph topological features. To each subgraph G a set of classical topological features are associated: number of nodes, number of edges, acyclicity, diameter, statistical descriptors of in-degree and out-degree (min, max, etc...)
- subgraph aggregate node features. Subgraph G is considered as a set of nodes described by features, (edge information is discarded), and aggregate descriptors such as the mean are computed for each feature type. For example, since not all firms in the edge file have associated firm data in the node file, we compute the rate of missing data concerning firms in individual subgraphs.
- subgraph aggregate edge features. Subgraph G is considered as a set of edges described by features, (node features are discarded), and aggregate descriptors such as the edge feature count are computed. For example, since edges bear the feature "ownership types" with modalities "s"/"e"/"p", we obtain three aggregate descriptors for a given subgraph: number of "s" edges, number of "p" edges, ...

Considering subgraphs seems natural because of the distribution of the sizes of the non-giant connected components (see 2). Still there is a giant component containing more than 500k firms, that needs to be cut into subgraphs in a principled way, taking into account the directed and very hierarchical structure of the ownership network. This "node clustering" problem has been largely addressed in many fields such as computer science, network science, but most works concern undirected networks. In the case of directed edges, much less algorithms are available. We first consider nonparametric stochastic block model (SBM) inference as a benchmark, because of the availability of theoretical works [Pei14b] as well as open source efficient and scalable implementations [Pei14a] providing minimum-description length (MDL) model selection, that is choosing the right number of clusters. Moreover we compare this benchmark to recently published spectral node clustering algorithms which remarkable performance was recognized in the network science community [LS20].

Dimension reduction and visualization is carried-on thanks to the manifold learning algorithm UMAP [MHM18], and compared to non-negative matrix factorization (NMF) [Gil21] that is better suited to explain why data points belong to a given cluster.

4 Preliminary results and perspectives

At this stage we are able to provide basic statistics at subgraph level concerning 10k weakly connected non-giant components, that represent 170k firms. More work is needed in particular to inject the type of anticompetitive practices inside the ownership network and to perform node clustering in the weakly connected giant component.

Still the following findings can be highlighted:

- topological features: the average node number is 17 and the average edge number is 20, both with a large associated variance, suggesting an important variability, which hints at considering binning subgraphs by size, and comparing bins. 99% of the top 10k non-giant subgraphs are directed acyclic graphs (this means that there is no directed loop path in the subgraph). The maximum indegree ($> 2k$) is much larger than the maximum indegree (42). The 75% quantile of the directed diameter is 2. Those indicators depict a typical subgraph structure that is very hierarchical, with a top-level root-node having many children, that are often leaf nodes, not linked to each other. This confirms the observations we made in a recent paper with a different dataset [HV22] and another hierarchy quantifier.

- aggregate node features. the median missing data rate is 58%, which is slightly above the mean expected rate of 47% evoked in sec.2. This may indicate an heterogeneity in missing node data among the sample, which calls for an explanation. The "group size" field has an average value of 18, which is coherent wit the topological average node number found above. However the maximum "group size" is $4,1k$, which is way above the size of the second-largest component ($1.5k$) mentionned in sec. 2.

- aggregate edge features. We observe that "e" edges are predominant, which is coherent with full-sample statistics.

Dimension reduction of topological subgraph features has been done with UMAP, and nice clusters appear. However an explanation for the appearance of clusters is yet to find, which is a known drawback of such manifold learning algorithm.

More work is thus needed, and we propose to leverage NMF to do so because of its higher explainability. Since some features are categorical, they should be separated from numeric features and treated separately, which might lead to a large dimensional problem similar to topic modelling.

In terms of perspectives, a comparison with deep unsupervised algorithms that operate directly on node-attributed subgraphs (rather than at node-level) would be interesting, but the literature is scarce.

Next, further examinations are required to reach a satisfactory node clustering of the giant connected component, prior to performing subgraph-level analysis as proposed above. First, a proper visualization is needed, getting rid of the many leaf nodes that obfuscate the giant component projection. An SBM-based node clustering (including model selection, thanks to graph-tool) was done, but fails (all nodes end up in the same cluster). An explanation for this is required. The spectral directed clustering algorithms have been tested on a synthetic tested and perform well, still we lack a adequate model selection procedure to choose the number of clusters. This is an open problem, hardly addressed in the literature.

Lastly, clustering of all nodes in the database, at node-level will be performed, comparing UMAP, NMF and deep unsupervised methods.

References

[BP16] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, Cambridge, United Kingdom, 2016. OCLC: ocn910772793.

[GB09] J. B. Glattfelder and S. Battiston. Backbone of complex networks of corporations: The flow of control. *Physical Review E*, 80(3):036104, September 2009.

[Gil21] Nicolas Gillis. *Nonnegative matrix factorization*. Number 2 in Data science. Society for Industrial and Applied Mathematics, Philadelphia, 2021.

[HV22] Aurélien Hazan and Arnold Vialfont. Réseau de liaisons financières entre entreprises françaises. In *FRENCH REGIONAL CONFERENCE ON COMPLEX SYSTEMS – FRCCS 2022*, Paris, France, June 2022.

[KOSVS+15] Sebnem Kalemli-Ozcan, Bent Sorensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yesiltas. How to construct nationally representative firm level data from the orbis global database: New facts and aggregate implications. Working Paper 21558, National Bureau of Economic Research, September 2015.

[LS20] Steinar Laenen and He Sun. Higher-order spectral clustering of directed graphs. *Advances in neural information processing systems*, 33:941–951, 2020.

[MHM18] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. Publisher: arXiv Version Number: 3.

[Pei14a] Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014.

[Pei14b] Tiago P. Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X*, 4(1):011047, March 2014.

[VGB11] Stefania Vitali, James B. Glattfelder, and Stefano Battiston. The Network of Global Corporate Control. *PLoS ONE*, 6(10):e25995, October 2011.