

RESEARCH

Open Access



Structure and influence in a global capital–ownership network

Sammy Khalife* , Jesse Read and Michalis Vazirgiannis

*Correspondence:
khalife@lix.polytechnique.fr
LIX, CNRS, Ecole
Polytechnique, Institut
Polytechnique de Paris,
Palaiseau, France

Abstract

Relationships between legal entities can be represented as a large weighted directed graph. In this work, we model the global capital ownership network across a hundred of millions of such entities with the goal of establishing a methodology for extracting and analysing meaningful patterns of capitalistic influence from the graph structure. To do so, we adapt and employ metrics from graph analytics and algorithms from the area of influence maximization. We characterize the relationships extracted and show that our analysis aligns with information from macro-economic studies; for example it recovers the presence of known tax heavens, which appear in dense subgraphs of countries. We also identify and quantify cases where capital is principally owned by others, corresponding to global influence. Beyond confirming known patterns and justifying our novel application of influence maximization methodology in this area, the outcome also offers new insight and metrics in this domain, by highlighting the existence of strong communities of capitalistic property. We leverage influence maximization methods as a means to evaluate the impact of entities in these contexts. Finally we formulate the results of our study into recommendations for future analyses of this kind.

Keywords: Complex networks, Legal entities, Capitalistic graphs, Centrality measures, Graph degeneracy, Influence maximization

Introduction

A *legal entity* is a juridic term that designates an individual, company, or organization that has legal rights and obligations. In the standard terminology, legal entities are usually divided into individuals or corporations (e.g companies). The economic entity principle, stating that financial transactions must be assigned to a specific business, is considered as one of the fundamental principles of accounting: each entity must have separate accounting records, except for its subsidiaries. This principle implies that the nature and clarity of the information describing legal entities is important for compliance standards. Several types of interactions can exist between these entities, for instance payments or capitalistic property, defining different network structures. Such networks have already been studied, namely the interbank market (Boss et al. 2004) and interbank payment flows (Soramäki et al. 2007). A study on systemic risk in the interbank network, where payment interactions are treated as a complex network was

described in Lenzu and Tedeschi (2012). Recent work attempted to describe the actual topologies observed in the financial system (Inaoka et al. 2004).

In this work, we propose a methodology built on recent developments in network science, to analyze a worldwide capitalistic ownership graph, and provide results on industrial data (described in Dijk (2018)). After an initial analysis using centrality measures, we propose an influence maximization (IM) algorithm in order to target specific nodes that have high importance in the ownership graph. This work significantly extends an earlier and preliminary analysis in Khalife et al. (2019). Specifically, this article contributes:

- a development and application of influence maximization explanation to the study of capitalistic graphs
- an extensive empirical evaluation comparing recent methodologies and techniques from network science to this area
- extraction and analysis of patterns relating to capitalistic ownership
- an in-depth discussion of the results, and
- a number of recommendations for future work.

Ownership network and influence maximization

Network model

We denote a capitalistic graph $G = (V, E)$. A vertex $i \in V$ represents a legal entity, an oriented edge e_{ij} with weight w_{ij} represents capitalistic property: an edge from i to j means that i owns capital of j in proportion w_{ij} . \mathcal{N}_i^+ and \mathcal{N}_i^- are respectively the out-neighbors and in-neighbors of vertex i (i.e., set of nodes connected to and from the vertex). $|I|$ is the cardinality of a finite set I . For mathematical convenience, we impose the following constraints:

$$\begin{aligned} \forall e \in E, w_e &\in [0, 1] \\ \forall i \in V, \sum_{j \in \mathcal{N}_i^-} w_{ji} &\leq 1 \end{aligned} \quad (1)$$

In Sect. 3.1, we describe how to ensure Eq. (1). Following standard juridic terms, there are two types of entities in the graph:

- Natural person: individual human being
- Legal person: incorporated organizations including corporations, government agencies; or non-governmental organizations. A legal person is composed of natural persons, but has a distinct juridic identity.

We suppose there does not exist inner links between the subgraph of natural people, even though natural person can have influence over others (we suppose this link is not explicit so it is not considered as an edge in the graph). We are interested in studying the influence of legal entities, so the distinction between natural entities and organizations raises an important discussion for the evaluation of such influence.

In view of the instance at our disposal, and in particular the fact that the data is poor with regard to personal attributes, we decided to consider the first situation, that is to say to keep only the links between organizations. Furthermore, unlike market payment

graphs which are very dynamic, the capitalistic graph evolves across a relatively long time scale and significant changes do not occur frequently (except rare events, such as the onset of an economic crisis). Therefore, we consider the capitalistic graph constant for our analysis and do not analyze its dynamic aspect, although this may be considered in future work. The data is provided by Bureau Van Dijk (2018), and lists all the physical and legal entities. Each of these entities has corresponding metadata: Name, location (country and continent), and description. They also have a weighted capital property over a list of other entities. The original data format is in a non-relational form; software development was necessary to parse and load data into a graph. For reasons of confidentiality, the data has been anonymized. Supplementary description is in Sect. 3.

Previous work

Location-sector analysis of the Orbis network has been proposed to identify and locate important entities (Nakamoto et al. 2019) and, in a 2015 snapshot, to analyze the location and sector of conduit firms likely to be used for treaty shopping (Nakamoto et al. 2019). In Heemskerk and Takes (2016) take a data-drive approach of community detection through modularity maximisation. In this work, we consider a different methodology in order to measure the capacity of some entities to influence economically the network. Differently, we here model an entity and its subsidiaries as separate, making the assumption that the influence of an entity is also spread through its subsidiaries in the diffusion. The macroeconomic indicators aforementioned will then be revealed using adequate centrality measures and influence maximization algorithms. Finally, in comparison with those references, we use a more recent version of the network (Orbis 2018). We defer a more in-depth discussion of the wider financial-analysis literature to Sect. 4.

Motivation and justification of the methodology

The notion of influence we consider concerns the capacity of an entity to influence economically and socially on the other entities of the network, and will be defined at different scales in our study. Firstly, at the level of legal entities, through the possession of capital, and then, using aggregation of attributes detailed in Sect. 2.7, at country and sector levels.

The motivations to analyse the knowledge graph under the prism of centrality measures and entity influence is two-fold. First, the detection of influential communities of entities has a natural interpretation in economic and financial terms. Second, the existence of low time complexity algorithms available for this purpose. Indeed, the considered network being very large, any non (quasi-) linear time complexity algorithm would result in computational time issue. Besides, several possibilities exist in order to model the influence of nodes, for instance with the influence interdiction problem (Omer and Mucherino 2020). However to the best of our knowledge, other formulations do not have linear time algorithm to obtain good approximate solutions efficiently. On the contrary, centrality measures (k -cores) and some influence maximization algorithms, which are relatively straightforward methods, have linear time complexity have proven to be very efficient in the detection of influential communities. In the remainder of this section we present the main components of our analysis in order to describe the topology of the graph, and to measure influence of entities, specifically:

- Degree distribution and connected components
- Centrality measures and K-cores
- Rooted influence graph
- Aggregation by attribute
- Influence maximization

Later, in Sect. 3, we provide the results of the analysis, with respect to location and sector.

Degree distribution and connected components

In the following A is the adjacency matrix of the network of the graph $G = (V, E)$.

$$A_{ij} = \begin{cases} w_{ij} & \text{if } e = (i, j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The inner-degree and outer-degree of a vertex i are respectively the sum of coefficients in column i and row i . The degree distribution is a fundamental and important measure in complex networks (especially Internet Siganos et al. 2003), because it allows to classify graphs according to the type of distribution (e.g power law graphs). In the following, we will consider degrees of the non-weighted graph, so that the in and out degrees of vertex i are respectively the number of in and out neighbors (i.e number of non-zero coefficients on the column and row i of A). A component $W \subseteq V$ of G is said connected if for each i and j in W , there exists a path from i to j .

Degeneracy and centrality measures

A graph is said to be k -degenerate if and only if every subgraph has a vertex of degree at most k . The degeneracy of a graph is defined as the smallest value of k for which it is k -degenerate. As such, it is a measure of sparsity. In order to enrich this notion, let us define the notion of k -core for an undirected graph. A k -core of an undirected graph is defined as the maximal subgraph $C_k \subset G$ such that each node has degree at least k in this subgraph:

$$C_k = \{i \in V \mid |C_k \cap \mathcal{N}_i| \geq k\} \quad (3)$$

If C_k is not empty, then G has degeneracy at least k since there exists a subgraph that contains nodes of degree k . Therefore, the degeneracy of G is the largest k for which C_k is not empty. The core number of a vertex is the maximum k for which he belongs to C_k . The notion of k -core extends to directed graphs, with in and out cores (i.e., \mathcal{N}_i becomes \mathcal{N}_i^- and \mathcal{N}_i^+ respectively). There exist linear-time algorithms to compute k -cores (Batagelj and Zaversnik 2003), and memory efficient methods (Cheng et al. 2011). As presented in Giatsidis et al. (2013), it is possible to extend the definition for directed graphs, considering 2-dimensions version (\mathcal{D} -cores), but they are not considered in this work. In the scope of this study, the computation of k -cores will yield dense communities of entities having important intra-ownership when k is close to the degeneracy of the graph.

Rooted influence graph

For each entity, we define a subgraph, the rooted influence graph (RIG), similarly to the rooted citation graph (RCG) in collaboration analysis (Giatsidis et al. 2019). The RIG of a vertex i is the subgraph of G induced by the set of vertices that contain i and all the vertices which can be reached by a directed path. That is, $j \in \text{RIG}$ if and only if there is a directed path from vertex i to vertex j . The resulting directed acyclic graph (DAG) contains all the entities that are directly or implicitly influenced by the entity i .

Based on this definition, it is natural to consider the following quantities in the RIG to measure influence of an entity:

- (a) out-degree
- (b) average degree of the nodes in the RIG
- (c) core influence (i.e core number of the considered entity of the undirected RIG)

In Sect. 3, we use the core influence measure since we are interested in the influence of communities rather than individuals. Indeed, coreness in the RIG of an entity is a measurement of how dense is its neighborhood.

Aggregation by attributes

The ownership graph also contain entity attributes (location (country or region) and a description of the activity (sector), cf Sect. 3.1 for a precise description). Here, we present a method to analyze the graph of entities by attributes. Let \mathcal{A} be the set of values for a given attribute of the entities. For $(a, b) \in \mathcal{A}^2$, let G_a (resp. G_b) be the set of entities having attribute a (resp. b). We define a new graph $G_{\mathcal{A}} = (\mathcal{A}, E_{\mathcal{A}})$ between attribute values in the following way:

$$w_{ab} = \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{i \in G_a \cap \mathcal{N}^-(j)} w_{ij} \tag{4}$$

In other words, $G_{\mathcal{A}}$ provides a kind of “meta-graph” based on the pairwise relationship between the values of \mathcal{A} . For example, a graph where $\mathcal{A} = \{a, b\}$ defines two countries, will be a graph of two nodes, inheriting the connectivity in the form of an aggregation w_{ab} on up to two directed edges $E_{\mathcal{A}}$. This definition of Eq. (4) insures the following mathematical conveniences, as also in Eq. (1):

$$\forall (a, b) \in \mathcal{A}^2$$

$$0 \leq w_{ab} \tag{5}$$

$$w_{ab} \leq \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{i \in G_a \cap \mathcal{N}^-(j)} w_{ij} \tag{6}$$

$$w_{ab} \leq 1$$

and

$$\begin{aligned}
 \sum_{a \in \mathcal{A}} w_{ab} &= \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{a \in \mathcal{A}} \sum_{i \in G_a \cap \mathcal{N}^-(j)} w_{ij} \\
 &= \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{i \in \mathcal{N}^-(j)} w_{ij} \\
 &\leq \frac{1}{|G_b|} \sum_{j \in G_b} 1 \\
 \sum_{a \in \mathcal{A}} w_{ab} &\leq 1
 \end{aligned}
 \tag{7}$$

Moreover, the choice of Eq. (4) is a good candidate for the influence of a attribute a over an attribute b (e.g influence of Germany over France), since it corresponds to a quantity representing the percentage or total capital owned by a over b . Therefore, The meta-graph $G_{\mathcal{A}}$ allows us to analyze interactions between attributes (i.e countries or sectors). However the limitation of this analysis lies in the assumption that entities have the same capital: we will come back to this limitation in Sect. 3.

Diffusion models and influence maximization

Influence maximization (IM) in a network is the problem of maximizing influence with regards to seed nodes using a diffusion model. It has been extensively studied recently due to its potential commercial value. An example of application of influence maximization is viral marketing (Domingos and Richardson 2001), where an organization wants to spread the adoption of a product from selected adopters. Influence maximization is also the corner stone in other important applications such as network monitoring, rumor control, and social recommendation. The first question is how to model the information diffusion process in a network, which affects the influence spread given a seedset.

In the previous section, we have defined the nature of the complex network we consider. In the following, the object of the spreading in this network is considered of factor of economic ownership which has different implications on economic and social influence. There exist several models for diffusion (Li et al. 2018). The first category of these models are called *progressive*. Activated nodes cannot be deactivated in later steps. Most IM algorithms consider the progressive models, and we will limit the discussion our study to those. In the next two paragraphs, we present two standard progressive models.

incoming neighbors are active. Each vertex v is associated a threshold τ_v , and gets activated if and only if the sum of incoming neighbors (incoming weights if the graph is weighted) is higher than τ_v .

The Independent Cascade (IC) model Independent Cascade (IC) is a classic and well-studied diffusion model (Goldenberg et al. 2001). A vertex v is activated by each of its incoming neighbors independently by introducing an influence probability to each edge. Given a seed set S at time step 0, the IC diffusion unfolds in discrete steps. Each active user u in step t will activate each of its outgoing neighbor v that is inactive in step $t + 1$ with probability p_{uv} . The activation process can be considered as flipping a coin with head probability p_{uv} : if the result is head, then v is activated; otherwise, v stays inactive. u has one chance to activate its outgoing neighbors. Then, u stays active and stops the activation. The diffusion instance terminates when no more nodes can be activated. The

influence maximization (IM) problem is theoretically complex in general. More precisely, it has been proven that there is no polynomial algorithm solving IM unless $P = NP$ under most of the diffusion models (Li et al. 2018). This theoretical result implies that it is very challenging to retrieve a solution close to the optimal seed set, and to scale to large graphs simultaneously. A naive algorithm exploring all the possible values for the seed sets of size k in a graph with n vertices yields a $\binom{n}{k}$ time-complexity (supposing the diffusion is in $O(1)$).

Influence function The definition of an influence function depends on the fundamental assumption that the diffusion model D terminates given any set of nodes S as input. In these conditions, $\sigma_{D,G}$ represents the function counting the number of vertices affected by the spread at the end of the process using the graph G . It operates on the set of subsets of V having k vertices. Given a diffusion model D , a graph G and a seed set size k , the IM problem corresponds to the following optimization problem:

$$S_{D,G}^* = \arg \max_{|S|=k} \sigma_{D,G}(S) \tag{8}$$

D, G are usually are implicit, so that $\sigma_{D,G}$ usually becomes σ . There exists a literature dedicated to IM (for a survey, see for instance Li et al. (2018)). This work is not intended as a survey on the topic, so we report some results and methods that we considered in order to measure entity influence in the capitalistic graph.

Greedy framework The greedy algorithm consists in adding iteratively nodes to S (starting from \emptyset), selecting a vertex u if u provides the maximum marginal gain to the influence function σ_D . The algorithm terminates when $|S| = k$. The theoretical guarantee of the greedy framework depends on the properties of σ (so implicitly on the diffusion model D and G). In the diffusion model we are considering (LT and IC), σ_D is a non-negative monotone and sub-modular function (for a proof, cf. for instance Mossel and Roch (2010)), but this property does not depend on the graph G considered. In these conditions, it is possible to estimate the approximation ratio of the greedy framework. If S^1 is the set of vertices returned by Algorithm 1, then Nemhauser et al. (1978):

$$\sigma_{D,G}(S^1) \geq (1 - (1 - \frac{1}{k})^k) \sigma_{D,G}(S^*) \tag{9}$$

Algorithm 1 Greedy method for IM

Require: $k \in \mathbb{N}$
 1: $S \leftarrow \emptyset$
 2: **for** $j = 1 \rightarrow k$ **do**
 3: $u^* = \arg \max_{u \in V \setminus S} (\sigma(S \cup \{u\}) - \sigma(S))$
 4: $S \leftarrow S \cup \{u^*\}$
 5: **end for**
 6: **return** S

Sketch-based methods There is a plethora of methods based on Monte-Carlo simulations in order to obtain approximate solutions of the IM problem. Sketch-based approaches are a range of methods that try to conceal theoretical efficiency (not only

practical efficiency) and preserving a constant approximation ratio. The main disadvantage of these methods is that they do not generalize to all diffusion models. However, they are compatible with LT and IC (using a generalization through the triggering model Kempe et al. (2003)). Specifically, Borgs et al. (2014) avoid the limitation of Greedy and propose a drastically different method for influence maximization under the IC model, called Reverse Influence Sampling (RIS). We first introduce two concepts:

Reverse reachable (RR) set Let v be a vertex in G , and g be a sub-graph of G obtained by removing each edge e in G with $1 - w_e$ probability ($w_e \in [0, 1]$). The reverse reachable (RR) set for v in g is the set of nodes in g that can reach v . (That is, for each node u in the RR set, there is a directed path from u to v in g).

Random (RR) set Let G be the distribution of g induced by the randomness in edge removals from G . A random RR set is an RR set generated on an instance of g randomly sampled from G , for a node selected uniformly at random from g .

Sketch-based methods consider θ sketches of the graph: $\{G_1, \dots, G_\theta\}$, and $\sigma_{D,G}(S)$ is computed as the average number of users reached by S on these sketches: $\hat{\sigma}_{D,G}(S) = \frac{1}{\theta} \sum_{i=1}^{\theta} \sigma_{D,G_i}(S)$. Borgs et al. (2014) showed that it is not necessary to estimate the influence using sketches on the entire graph.

Algorithm 2 Reverse reachable (RR) sketch

Require: $\theta \in \mathbb{N}$ and a RR generation method

```

1:  $S \leftarrow \emptyset$ 
2: Generate  $R_1, \dots, R_\theta$  reverse
3:  $R \leftarrow \{R_1, \dots, R_\theta\}$ 
4: for  $i = 1 \rightarrow k$  do
5:   Pick vertex  $v_i$  that covers the most RR sets in  $R$ 
6:    $S = S \cup \{v_i\}$ 
7:   Remove RR sets from  $R$  that are covered by  $v_i$ 
8: end for
9: return  $S$ 

```

The average infected seed set over the sketches using LT and IC can be proven to converge in probability to the infected set of the all graph (Kempe et al. 2003). Moreover, the quality of the solutions returned by some of the sketch-based methods (e.g IMM Tang et al. (2015)) have been proven to hold an approximation ratio of $(1 - \frac{1}{e} - \epsilon)$ with high probability (Tang et al. 2014).

Results

Description and preprocessing

Orbis is a database composed of about one hundred million entities, developed by a specialized group based in Bureau Van Dijk's Brussels office, aggregating several sources of data. A detailed description is given in Dijk (2018). The location of an entity is usually a country (or else a region such as Hawaiï (U.S), La Réunion (Fr.)) that defines where the legal entity has its headquarters. Therefore, a subsidiary and a parent company can have different locations. See Beddi and Mayrhofer (2010) for in-depth insights into the role of location in headquarters-subsidiaries relationships. The sector of a legal entity is a description of its activity. The database had some inconsistencies and missing values. In some cases entities (and edges) appear multiple times. We merged identical edges into a

single edge by taking the maximum weight. About 30% of weights are missing. According to the documentation of Orbis, these links are indirect, i.e they represent ownership through other entities and, since we wish to measure direct ownership, we simply removed these edges for our analysis.

Degree distribution and connected components

We used the igraph software implementation (Csardi and Nepusz 2006) to manage the graph structure emanating from this database. After pre-processing, the graph of non isolated entities (i.e. entities with at least one edge) is composed of 39, 398, 321 nodes and 80, 874, 728 edges. Following our discussion in the first section, we consider the subgraph of juridical entities (organizations) for our entity influence analysis. The subgraph of non-isolated organizations is composed of 6, 516, 332 nodes and 6, 670, 813 edges, with 1, 429, 853 components; additional numbers are in Table 1. An important metric in graph mining is the *density*, representing the average connectivity (i.e. abundance of edges) of the graph. The density of a directed graph is a real number in [0, 1], maximized for cliques and minimized for a graph of isolated nodes. For the graph at hand the density is:

$$D = \frac{|E|}{|V|(|V| - 1)} = 1.571e - 07$$

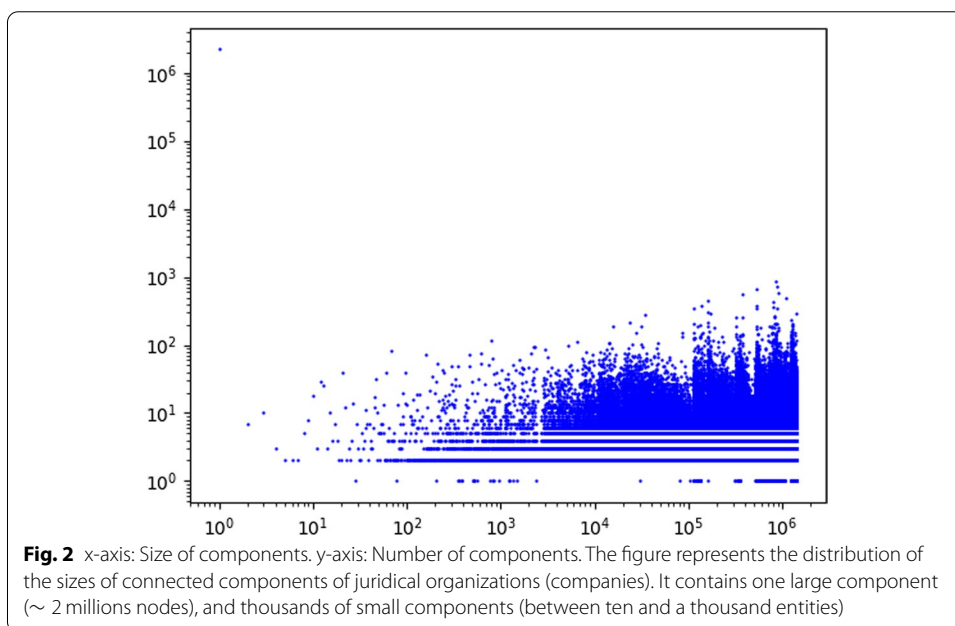
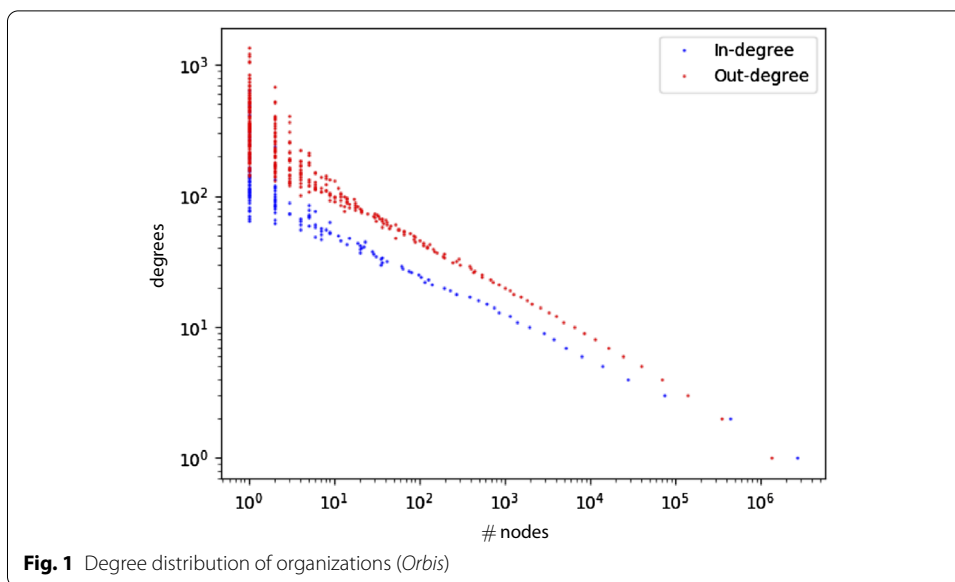
which means that the graph is very sparse. The degree distribution is in Fig. 1. We can see that the degrees vary importantly in the graph. More specifically there are few entities with up to more that a million capitalistic relations whereas the vast majority have less that one hundred ones.

The distributions of entities by country is displayed in Table 2. We see that the distribution is not uniform and is not correlated with the sizes of each country. There are relatively few entities in the United States (US) compared to some European or South American countries. This is because the data source has little knowledge of US entities. One possibility for refining the study in relation to this country would be to supplement *Orbis* with data from other sources (outside the scope of this work).

The subgraph of organizations contains a very large component (i.e. a set of entities where each pair of nodes is connected via a path) of 1,442,704 nodes and 1,816,874 edges. The other components are very small, smaller than one thousand nodes. The component-size distribution is in Fig. 2, which shows that there are a lot of small components (some hundred of nodes), and lots of isolated nodes. Two drawings of smaller components (less than a thousand nodes), are depicted in Fig. 3. We do not include a

Table 1 Capitalistic ownership graph instance (Orbis). (1) All edges, (2) Edges with unknown weights removed

Subgraph	Nodes		Edges	
	(1)	(2)	(1)	(2)
Organizations	81576517	81576517	6818574	4242843
Non isolated organizations	6518718	4789294	6818574	4242843
Total (individuals and organizations)	105426819	105426819	81111480	49777255



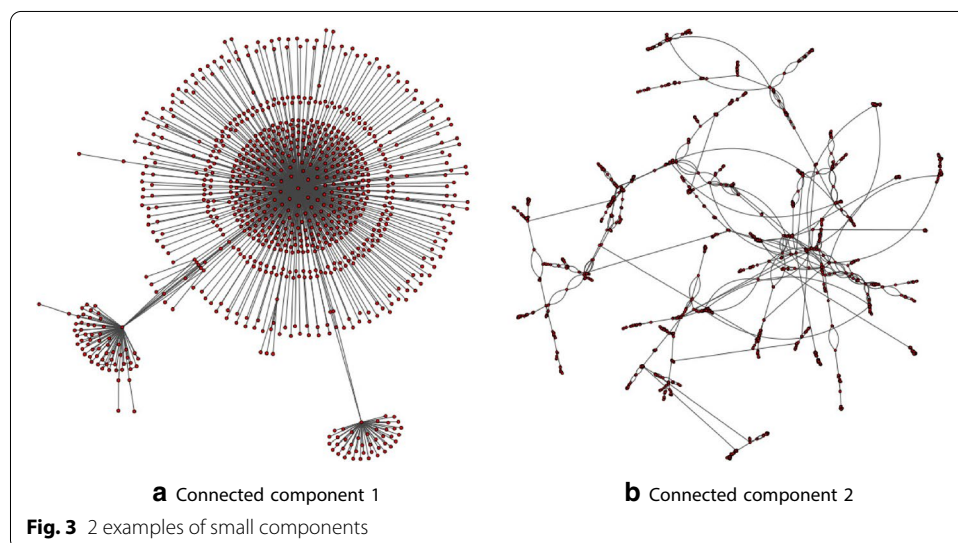
fine-grained classification of the structure of these small components. However, in the next subsections we provide the analytics on the largest component of juridical entities of 1, 442, 704 nodes following Sect. 2.

Degeneracy and centrality measures

In Sect. 2.5, we define the notion of k -degeneracy and k -cores, named C_k . The maximum k -core (defined as the k -core such that C_k is not empty and k is maximal) of the graph allows to find an approximation of the densest part of the graph in linear time (Batagelj and Zaversnik 2003). On the one hand, in the original graph of entities restricted to organizations, the large component of 1,442,704 million nodes has a degeneracy value

Table 2 Top 20 most frequent countries in total

Country	Number (total)
Brazil	19550646
China	9865149
Italy	4985420
United Kingdom	4030892
France	3740322
Russian Federation	3258544
Germany	2631038
Australia	2554195
Netherlands	2498228
Colombia	1924520
Czech Republic	1803264
Poland	1607002
Sweden	1594381
Japan	1489379
Spain	1163873
India	1139334
Mexico	965197
Bulgaria	898522
Taiwan	863927
Romania	853540



equal to 18 composed of 45 entities. C_{18} is a very dense community of entities, where each of them is being owned (resp. owns) the capital of (resp. by) at least 18 other entities in total. Recall that the other components are much smaller (i.e. few hundreds of entities) and sparse, so that degeneracy in these components is not very informative for data mining.

On the other hand, the distribution of countries and sectors in $\{C_k \mid k \geq 9\}$ (992 entities) are displayed in Tables 3 and 4. These tables show that countries and sectors are not uniform but shared between a limited amount of countries (Singapore, Australia,

Table 3 20 most frequent countries in the 10 top-k cores (992 entities)

Countries	Number
Singapore	440
Australia	331
Germany	256
Ukraine	125
France	69
Malaysia	52
New Zealand	52
Bermuda(GB)	33
India	32
Chile	26
Italy	19
Thailand	15
Hong Kong	10
United Kingdom	6
Cayman Islands(GB)	5
China	2
Ireland	2
Netherlands	2
Israel	1
Japan	1
Luxembourg	1
British virgin islands	1

Germany, Ukraine, and France), and sectors (Activities of holding companies, Buying and selling of own real estate, Other monetary intermediation, Rental and operating of own or leased real estate, Activities of head offices). These countries and sectors have a more intense interaction for capitalistic property.

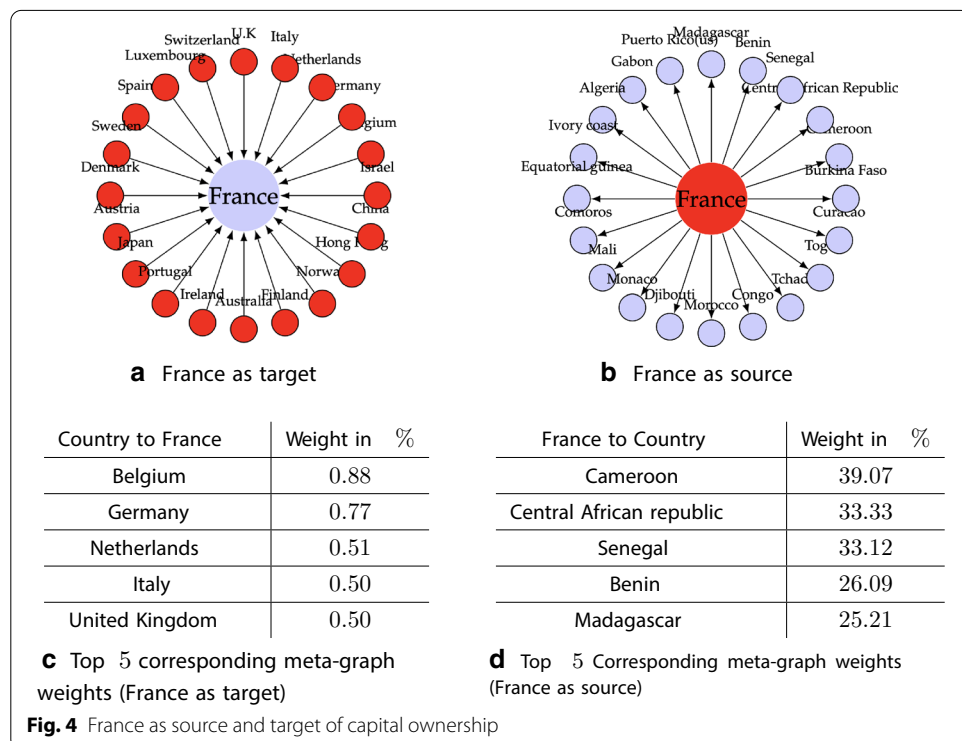
Aggregation by attribute

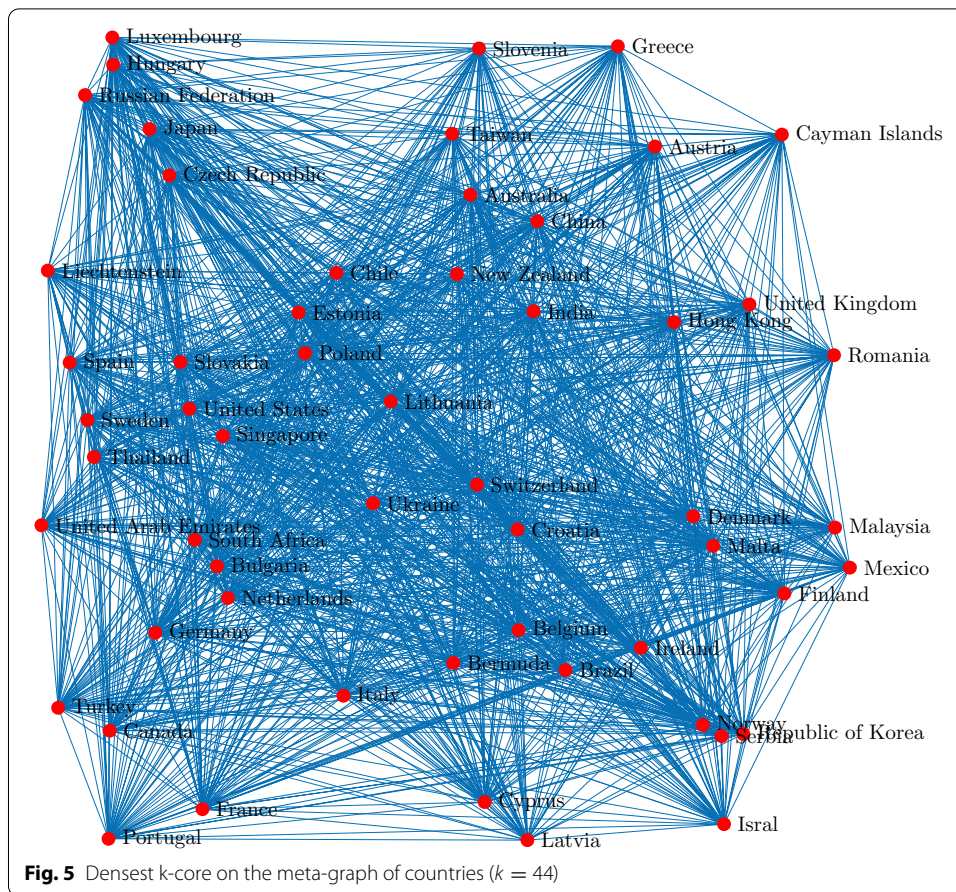
In order to provide an example of location analysis of such network, we consider the case of France. We computed the aggregated sites whose capital is most possessed by French entities (top 20 countries ranked by edge weight). Conversely, we computed the top 20 aggregate locations that own capital of French entities. We reported results in Fig. 4. The entities most held by the French entities are those located in their former colonies (ie Algeria, Togo, Congo, Côte d'Ivoire, Benin, Chad, Gabon, Senegal, Cameroon, Comoros, Madagascar, ...). Conversely, the French entities are mostly owned by those located in economically strong countries, such as Germany, China, the United States, Japan, the United Kingdom, the United Arab Emirates, as well as countries close to France geographically or sharing close historical relations (Belgium, Italy, Morocco, ...). We also reported top-5 neighbors sorted by weight in descending order (Fig. 4d, c). Top in weights are much lower than the top out weights, this means that French entities have a tendency to own capital of entities in other countries rather than to have capital owned at the international level.

Figure 5 depicts the densest k-core ($k = 44$) on the meta-graph of countries, and represents the most connected subgraph of countries of capitalistic ownership. We

Table 4 Most frequent sectors in top k-cores (992 entities)

Sector	Number
Activities of holding companies	314
Buying and selling of own real estate	84
Other monetary intermediation	64
Rental and operating of own or leased real estate	53
Activities of head offices	46
Other activities auxiliary to financial services, except insurance and pension funding	42
Trusts, funds and similar financial entities	37
Unknown	36
Activities of other membership organisations n.e.c.	36
Publishing of directories and mailing lists	35
Other financial service activities, except insurance and pension funding n.e.c.	32
Sea and coastal freight water transport	25
Real estate agencies	16
Business and other management consultancy activities	16
Precious metals production	15
Other business support service activities n.e.c.	14
Other credit granting	13
Mining of coal and lignite	12
Publishing of journals and periodicals	11
Mining of other non-ferrous metal ores	11
Security and commodity contracts brokerage	11
Fund management activities	10
Other professional, scientific and technical activities n.e.c.	10
Retail sale in non-specialised stores with food, beverages or tobacco predominating	10

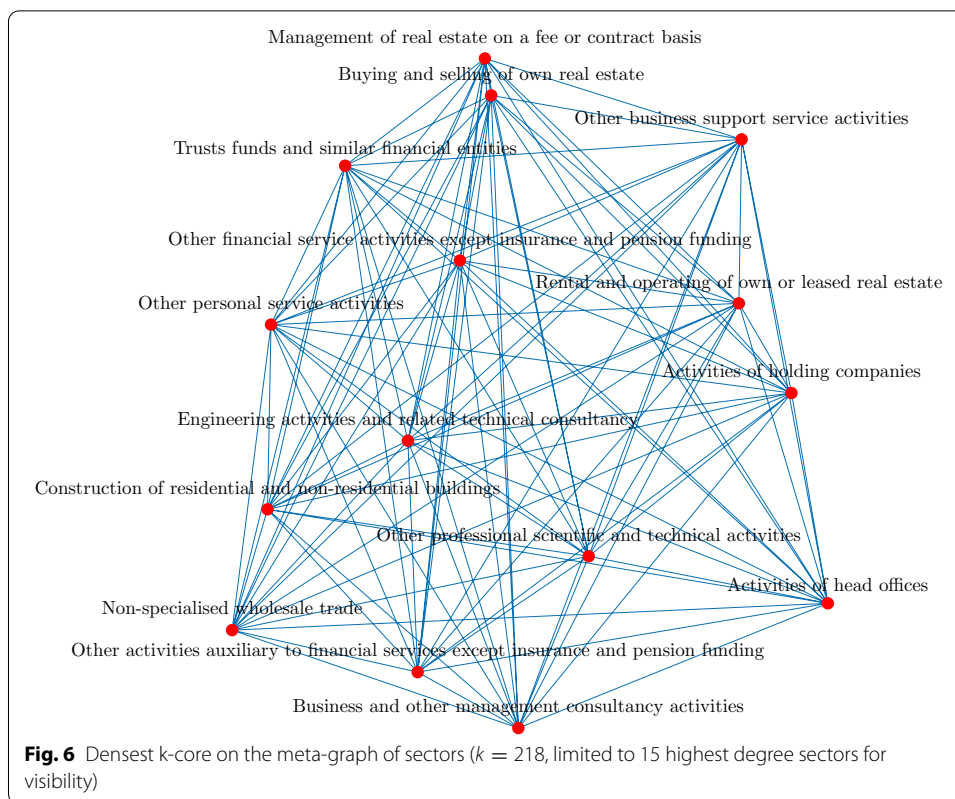




observe an important proportion of countries in Europe (France, Germany, United Kingdom, Italy, Spain, Portugal, ...). Other economically strong countries such as United States, China, United Arab Emirates and Republic of Korea are also represented. These two categories have a dense collaboration in terms of capital ownership. Two important aspects about Fig. 5:

- Some “tax havens” are present (Cayman Islands, Malta, Cyprus, ...) despite their relative low number of entities.
- The United States of America are in the densest k -core of the meta-graph of countries, even though under-represented in the dataset (less than 850000 of entities, cf. Table 3).

Figure 6 depicts the densest k -core ($k = 218$) on the meta-graph of sectors (limited to 15 sectors with highest degree for visibility). It reveals a dense collaboration between several economic sectors (Financial activities, real estate, engineering and technical consultancy, ...).



Influence analysis

Recall that, in the context of this paper, influence is *the capacity to have an effect on the development or decisions of an entity*. In this subsection we present two different types of methods to measure it. Then, we compare the results obtained using the same diffusion model. We used two coreness-based methods to measure influence indirectly. First, sorting nodes based on their coreness number in the graph. Second, following Sect. 2.6, we compute the coreness of each node in its RIG, and sort them by decreasing value. Then, we keep the top- k nodes and consider the distribution of attributes within these. The results with $k = 10000$ are in Table 5 (left side) for location analysis and in Table 6 for sector analysis.

The problem and method used are described in Sect. 2.8. We use the influence maximization (IM) paradigm in order to measure entity influence in the capitalistic graph. Here, we ran the simulations using influence maximization with martingales, with $\epsilon = 0.1$ and a seed set of 10,000 nodes. The output is a seed set of 10,000 nodes which is an approximation of the IM solution whose quality will be discussed in Sect. 4. We display the distributions within this seed set in the right part of Table 5 for location analysis, and Table 7 for sector analysis.

In order to estimate the quality of their respective solutions, we compared the number of infected nodes using our diffusion model (independent cascades), with entities obtained in the top k -cores of the graph. We define the ratio τ as the ratio of the seed set size and the number of vertices of the graph. For practical software reasons (NDlib library Rossetti et al. 2018), we considered larger seed sets ($\tau \in \{5\%, 10\%, 15\%, 20\%\}$) of

Table 5 Top 20 influential countries using coreness on the RIG (left) and influence maximization on the IM on the total network of entities (right)

Country	Number	Country	Number
Italy	1085	Germany	1322
Germany	995	China	1161
Ukraine	880	Australia	1010
France	866	France	642
India	600	Italy	628
Japan	555	United Kingdom	530
Australia	425	Austria	381
Spain	396	Norway	337
United Kingdom	334	Spain	328
Russian Federation	328	Ukraine	296
Norway	298	Japan	255
Portugal	277	Netherlands	220
Austria	258	Belgium	206
China	190	Russian Federation	169
Taiwan	190	Sweden	163
Belgium	187	Singapore	137
Netherlands	178	New Zealand	115
Thailand	163	Portugal	115
Singapore	151	Poland	108
Malaysia	124	India	98

Table 6 Top 20 influential sectors using coreness using RIG

Sector	Number of entities
Activities of holding companies	1079
Other monetary intermediation	659
Activities of head offices	636
Rental and operating of own or leased real estate	344
Other financial service activities, except insurance and pension funding n.e.c.	312
Other activities auxiliary to financial services, except insurance and pension funding	249
Business and other management consultancy activities	244
Unknown	243
Fund management activities	193
Other business support service activities n.e.c.	185
Trusts, funds and similar financial entities	175
Non-specialized wholesale trade	168
Construction of residential and non-residential buildings	137
Development of building projects	132
Buying and selling of own real estate	114
Retail sale in non-specialized stores with food, beverages or tobacco predominating	107
Security and commodity contracts brokerage	100
Production of electricity	97
Administration of financial markets	85
Insurance, reinsurance and pension funding, except compulsory social security	84
Life insurance	81
Non-life insurance	77

Table 7 Top 20 influential sectors using IMM

Sector	Number of entities
Unknown	1181
Activities of holding companies	520
Rental and operating of own or leased real estate	354
Other business support service activities n.e.c.	298
Activities of head offices	291
Buying and selling of own real estate	214
Business and other management consultancy activities	212
Other activities auxiliary to financial services, except insurance and pension funding	180
Development of building projects	155
Production of electricity	153
Construction of residential and non-residential buildings	133
Real estate agencies	132
Trusts, funds and similar financial entities	130
Engineering activities and related technical consultancy	119
Non-specialised wholesale trade	107
Management of real estate on a fee or contract basis	102
Computer programming activities	100
Other financial service activities, except insurance and pension funding n.e.c.	95
Other professional, scientific and technical activities n.e.c.	89
Other transportation support activities	85

Table 8 Comparison with the IC model for several seed set sizes (50 simulations)

Method	$\sigma(S) \pm std$			
	$\tau = 5\%$	$\tau = 10\%$	$\tau = 15\%$	$\tau = 20\%$
Top- k coreness	285250.8 \pm 1304.8	351887.4 \pm 928.8	418440.6 \pm 755.8	474402.2 \pm 454.3
IMM (Tang et al. 2015)	161485.4 \pm 1211.2	298874.9 \pm 774.3	419304.6 \pm 1055.5	521251.7 \pm 1033.4
Top- k RIG coreness	610138.7 \pm 549.8	701871.5 \pm 706.7	799244.1 \pm 661.1	903330.4 \pm 369.1

the considered graph where $\tau = 5\%$ corresponds to 75,136 nodes). The results are shown in Table 8.

Discussion and conclusion

Our analysis based on centrality measures and influence maximization is consistent with economic and historical facts such as the existence of tax heavens, and private capital ownership over its former colonies (we specifically studied the case of France). This therefore allowed us to reveal important locations and sectors communities in the economy, and potentially target most influential entities. Our method allows us to validate these intuitions and to provide indicators of the influence of countries via private capitalistic possession.

Initial seeds using k -cores and yield a better influence score for several seed set sizes ranging from 5 to 20% using the independent cascade diffusion model. Results suggest that several means should be considered (depending on the size of the seed set) to measure influence in large networks with similar degree distribution. Our

main result via k -core analysis on the capital–ownership data of organizations reveals a central component. And within, our analysis leveraging the different approaches listed in the previous section helped define communities of countries or sectors that are influencing the economy of capital ownership.

As with any study of this kind, there are possible limitations. For example, we may highlight the limitation of the approximation of different sized entities as having the same capital, since we aggregated the weights without taking in account the capital value. That is to say, the size of capital of entities is not explicitly modeled (entities are assumed to have the same capital). However, although a direct deterministic link cannot be established, there are strong links (Brailsford et al. 2002), in particular a positive relationship between ownership and capital structure. Therefore from a probabilistic point of view, one can argue that our analysis is based on structure with the volume of capital ‘marginalized out’. And this appears to be the case in so far as that our results and conclusions corroborate those of the literature. This, of course, is not to say that a follow up analysis (including capital holdings directly) would yield a finer-grained analysis.

The influence measurement following methods of RIG and IMM (Sects. 2.6 and 2.8, respectively) revealed significant difference in the distributions of attributes. For all seed set sizes, we have obtained better results using RIG coreness than with IMM. Influence based on coreness on the initial graph yields better solutions for $\tau \in \{5\%, 10\%, 15\%$ seed set sizes, and IMM performs a bit better than top- k coreness for $\tau = 20\%$. This is an interesting discovery. In spite of good theoretical guarantees, the IMM algorithm of influence maximization suffers on some graph instances and seed set sizes found in practice (such as this one). This is also in accordance with experiments of Giatsidis et al. (2019) suggesting that coreness can yield excellent influence measurement in our context of interest of this work. We thereby suggest that for large and sparse networks such as the one we considered, coreness-based influence measurement should be preferred.

Our results are inline with the literature. For example, Garcia-Bernardo et al. (2017) also use a data-driven approach to identify *sinks*, which included the Netherlands, the United Kingdom, Ireland, Singapore, and Switzerland—all identified in the top-44 core countries in our analysis, and nearly all (except Switzerland) in the top-20 core. In this context, sinks attract and retain foreign capital. An example was given in Garcia-Bernardo et al. (2017) where in three years (2007–2009) Google moved for tax reasons most of its profits generated outside the United States (US\$12.5 billion) through corporate entities in the Netherlands. On the other hand, other countries are attractive as intermediate destinations of international investments and transfers. For example our analysis uncovered the Cayman Islands. This is covered in discussion in the literature and, e.g., Roberts (1995); Fichtner (2016), specifically highlight and cover this case. Authors in Vitali et al. (2011) found that transnational corporations form central structure where a large portion of financial control flows to a small tightly-knit core of financial institutions; this also corroborates our results. The centralization we see has been explained as a partial results of tightening monetary policy (Braccaccio et al. 2019). It appears—and this is supported by, e.g., Vitali et al. (2011) who refer to an economic ‘super-entity’, that not only an effect of policy but which could have major future policy implications. Therefore it is important that ongoing research confirms these structures.

It may appear as surprising that U.S. entities did not have a relatively more important role, though we emphasise that they simply do not appear in this data. Although we have limited information on the U.S., other research, e.g., Heemskerk and Takes (2016)—who used community detection through modularity maximisation—discusses strong transatlantic connections between Europe and North America as part of a dense network of shared directors and control, as opposed to a relatively more separate Asian cluster still relatively subservient to the US-Europe center. In other words, traditional core countries still form dominant function. On a more detailed level, Glattfelder and Battiston (2009) show that although Anglo-Saxon countries have highly concentrated control in the hands of a few shareholders, whereas in European countries this concentration is relatively less significant. When we looked at France in particular, we see diversity among neighboring countries with outward flow, and inward flow (i.e., greater control) over former colonies. This provides further insight (and validation) of the functionality of Orbis instance Dijk (2018).

We notice that IMM improves over RIG. The influence measurement following methods of RIG and IMM (Sects. 2.6 and 2.8 respectively) revealed significant difference in the distributions of attributes. For all seed set sizes, we have obtained better results using RIG coreness than with IMM. Influence based on coreness on the initial graph yields better solutions for $\tau \in \{5\%, 10\%, 15\%$ seed set sizes, and IMM performs a bit better than top- k coreness for $\tau = 20\%$. This is an interesting discovery. In spite of good theoretical guarantees, the IMM algorithm of influence maximization suffers on some graph instances and seed set sizes found in practice (such as this one). This is also in accordance with experiments of Giatsidis et al. (2019) suggesting that coreness can yield excellent influence measurement in our context of interest of this work. We suggest that for large and relatively sparse network, influence measurement with coreness should be preferred.

We outlined results that should be treated with caution, due to the sparsity of data, particularly respective of some countries, for example—in this dataset—the United States. Nevertheless, the overall results provide important insight and information on the influence of the entities and are for the most part consistent with the current global economic situation. Of course, further enrichment of the capital ownership dataset with additional entities and capital information will likely provide an even more precise quantitative measure of influence. We could also consider studying the evolution of this graph over time (monthly or yearly), but this is out of scope of the current study and we leave that line of research for future work.

Abbreviations

IC: Independent Cascade; IM: Influence Maximization; IMM: Influence Maximization with Martingales; LT: Linear Threshold; RCG: Rooted Citation Graph; RIG: Rooted Influence Graph; RIS: Reverse Influence Sampling; RR: Reverse reachable.

Acknowledgements

We thank BNP Paribas CIB (Paris, Analytics Consulting) for sharing the Ownership network dataset assembled by Bureau Van Dijk.

Authors' contributions

SK: Conceptualization, methodology, writing, software, investigation, visualization. JR: Methodology, writing. MV: Discussion and supervision. All authors read and approved the final manuscript.

Funding

No funding to be declared.

Availability of data and materials

The dataset used for this analysis (Dijk 2018) was assembled by Bureau Van Duyk, and is subject to strict sharing rules (data available under purchase, no open source available).

Competing interests

The authors declare that they have no competing interests.

Received: 12 July 2020 Accepted: 3 February 2021

Published online: 18 February 2021

References

- Batagelj V, Zaversnik M (2003) An $o(m)$ algorithm for cores decomposition of networks. arXiv preprint cs/0310049
- Beddi H, Mayrhofer U (2010) The role of location in headquarters-subsidiaries relationships: an analysis of French multinationals in emerging markets. In: 36th annual EIBA (European International Business Academy) conference, p 26
- Borgs C, Brautbar M, Chayes J, Lucier B (2014) Maximizing social influence in nearly optimal time. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms. SIAM, pp 946–957
- Boss M, Elsinger H, Summer M, Thurner S (2004) Network topology of the interbank market. *Quant Finance* 4(6):677–684
- Brailsford TJ, Oliver BR, Pua SL (2002) On the relation between ownership structure and capital structure. *Account Finance* 42(1):1–26
- Brancaccio E, Giammetti R, Lopreite M, Puliga M (2019) Monetary policy, crisis and capital centralization in corporate ownership and control networks: a b-var analysis. *Struct Change Econ Dyn* 51:55–66. <https://doi.org/10.1016/j.strueco.2019.08.005>
- Cheng J, Ke Y, Chu S, Özsü MT (2011) Efficient core decomposition in massive networks. In: 2011 IEEE 27th international conference on data engineering. IEEE, pp 51–62
- Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal Complex Systems* 1695
- Dijk BV (2018) Source: Orbis
- Domingos P, Richardson M (2001) Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 57–66
- Fichtner J (2016) The anatomy of the Cayman islands offshore financial center: Anglo-America, Japan, and the role of hedge funds. *Rev Int Polit Econ*. <https://doi.org/10.1080/09692290.2016.1243143>
- García-Bernardo J, Fichtner J, Takes FW, Heemskerck EM (2017) Uncovering offshore financial centers: conduits and sinks in the global corporate ownership network. *Sci Rep* 7(1):6246. <https://doi.org/10.1038/s41598-017-06322-9>
- Giatsidis C, Thilikos DM, Vazirgiannis M (2013) D-cores: measuring collaboration of directed graphs based on degeneracy. *Knowl Inf Syst* 35(2):311–343
- Giatsidis C, Nikolentzos G, Zhang C, Tang J, Vazirgiannis M (2019) Rooted citation graphs density metrics for research papers influence evaluation. *J Inform* 13(2):757–768
- Glattfelder JB, Battiston S (2009) Backbone of complex networks of corporations: the flow of control. *Phys Rev E* 80(3):036104
- Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. *Mark Lett* 12(3):211–223
- Heemskerck EM, Takes FW (2016) The corporate elite community structure of global capitalism. *New Polit Econ* 21(1):90–118. <https://doi.org/10.1080/13563467.2015.1041483>
- Inaoka H, Ninomiya T, Taniguchi K, Shimizu T, Takayasu H et al (2004) Fractal network derived from banking transaction-an analysis of network structures formed by financial institutions. *Bank Jpn Work Pap* 4
- Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 137–146
- Khalife S, Read J, Vazirgiannis M (2019) Empirical analysis of a global capital-ownership network. In: International conference on complex networks and their applications. Springer, pp 656–667
- Lenzu S, Tedeschi G (2012) Systemic risk on different interbank network topologies. *Phys A Stat Mech Appl* 391(18):4331–4341
- Li Y, Fan J, Wang Y, Tan K-L (2018) Influence maximization on social graphs: a survey. *IEEE Trans Knowl Data Eng* 30(10):1852–1872
- Mossel E, Roch S (2010) Submodularity of influence in social networks: from local to global. *SIAM J Comput* 39(6):2176–2188
- Nakamoto T, Chakraborty A, Ikeda Y (2019) Identification of key companies for international profit shifting in the global ownership network. *Appl Netw Sci* 4(1):58
- Nakamoto T, Rouhban O, Ikeda Y (2019) Location-sector analysis of international profit shifting on a multilayer ownership-tax network. *Evol Inst Econ Rev* 17(1):219–41
- Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions-i. *Math Program* 14(1):265–294
- Omer J, Mucherino A (2020) Referenced vertex ordering problem: theory, applications and solution methods. HAL open archives
- Roberts SM (1995) Small place, big money: the cayman islands and the international financial system. *Econ Geogr* 71(3):237–256
- Rossetti G, Milli L, Rinzivillo S, Sirbu A, Pedreschi D, Giannotti F (2018) Ndlb: a python library to model and analyze diffusion processes over complex networks. *Int J Data Sci Anal* 5(1):61–79
- Siganos G, Faloutsos M, Faloutsos P, Faloutsos C (2003) Power laws and the as-level internet topology. *IEEE/ACM Trans Netw* 11(4):514–524

- Soramäki K, Bech ML, Arnold J, Glass RJ, Beyeler WE (2007) The topology of interbank payment flows. *Phys A Stat Mech Appl* 379(1):317–333
- Tang Y, Xiao X, Shi Y (2014) Influence maximization: Near-optimal time complexity meets practical efficiency. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data. ACM, pp 75–86
- Tang Y, Shi Y, Xiao X (2015) Influence maximization in near-linear time: a martingale approach. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data. ACM, pp 1539–1554
- Vitali S, Glattfelder JB, Battiston S (2011) The network of global corporate control. *PLoS ONE* 6(10):25995

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
