Supporting information

Fig. S1. Distribution differences Normalized distributions of Funding, Indegree, Outdegree and Betweenness. All network centrality distributions are significantly different from the funding one: Indegree ($p \sim 10^{-16}$), Outdegree ($p \sim 10^{-6}$), Betweenness ($p \sim 10^{-4}$).

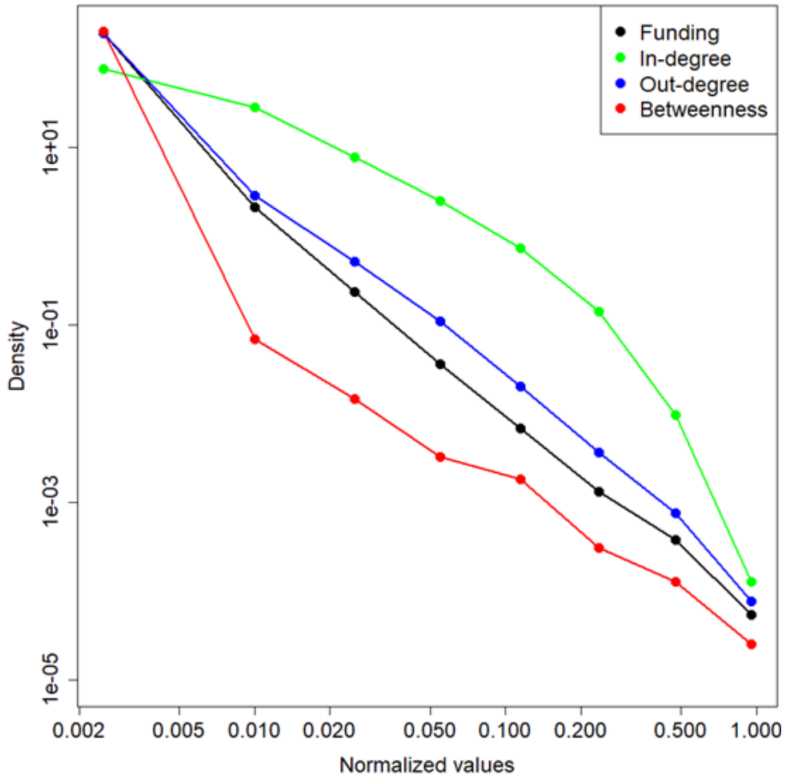


Fig. S2-a. Comparison between Indegree and Funding distributions for Nationality.

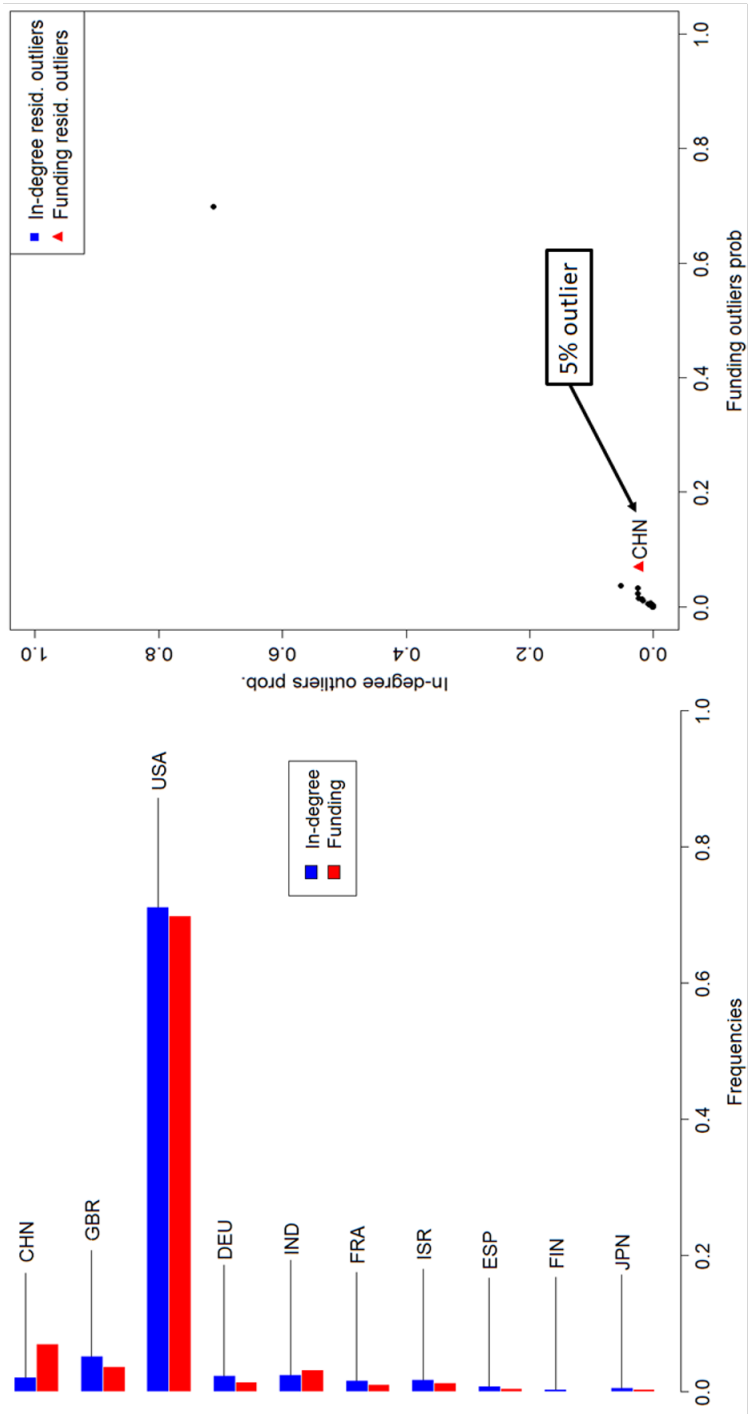


Fig. S2-b. Comparison between Indegree and Funding distributions for Investor Type.

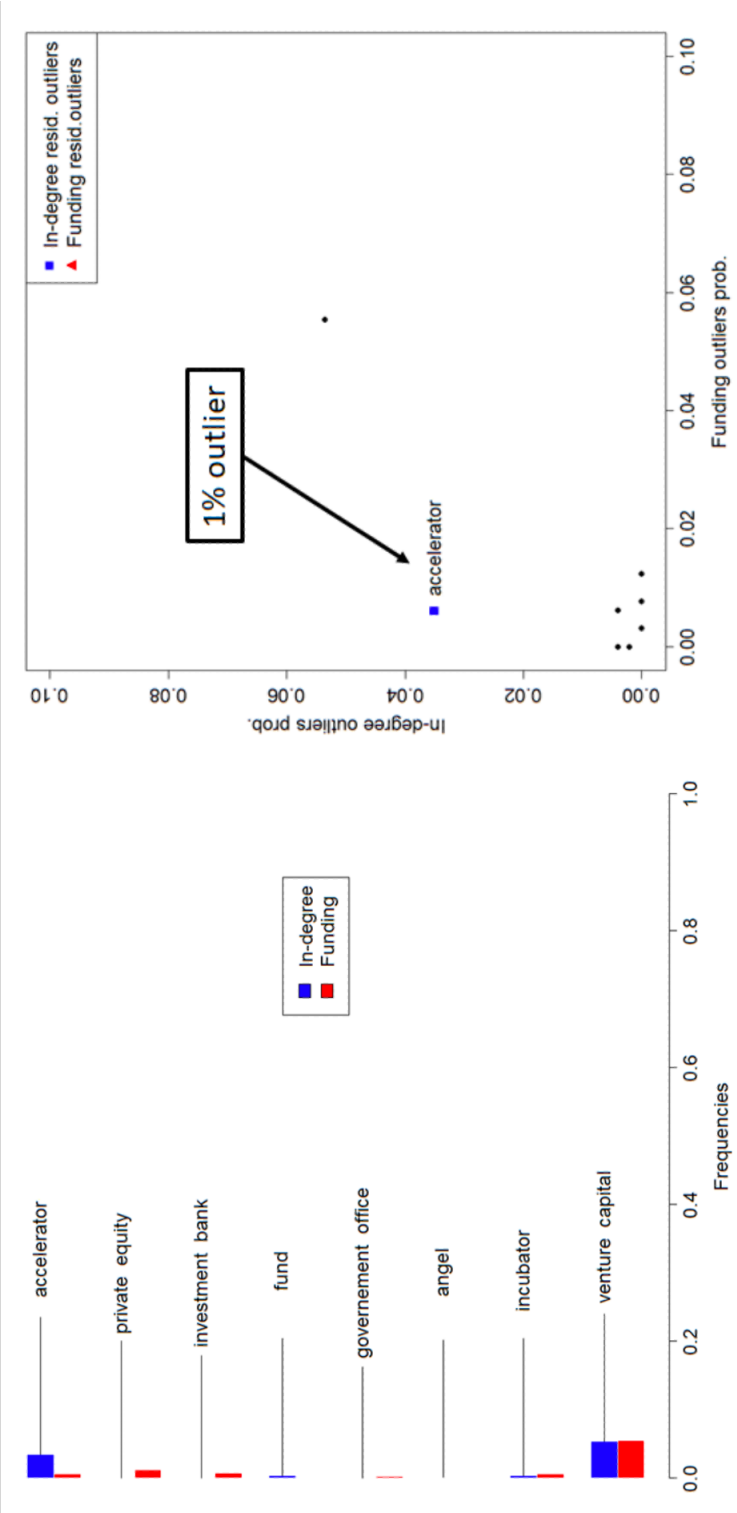


Fig. S2-c. Comparison between Indegree and Funding distributions for Economic Category.

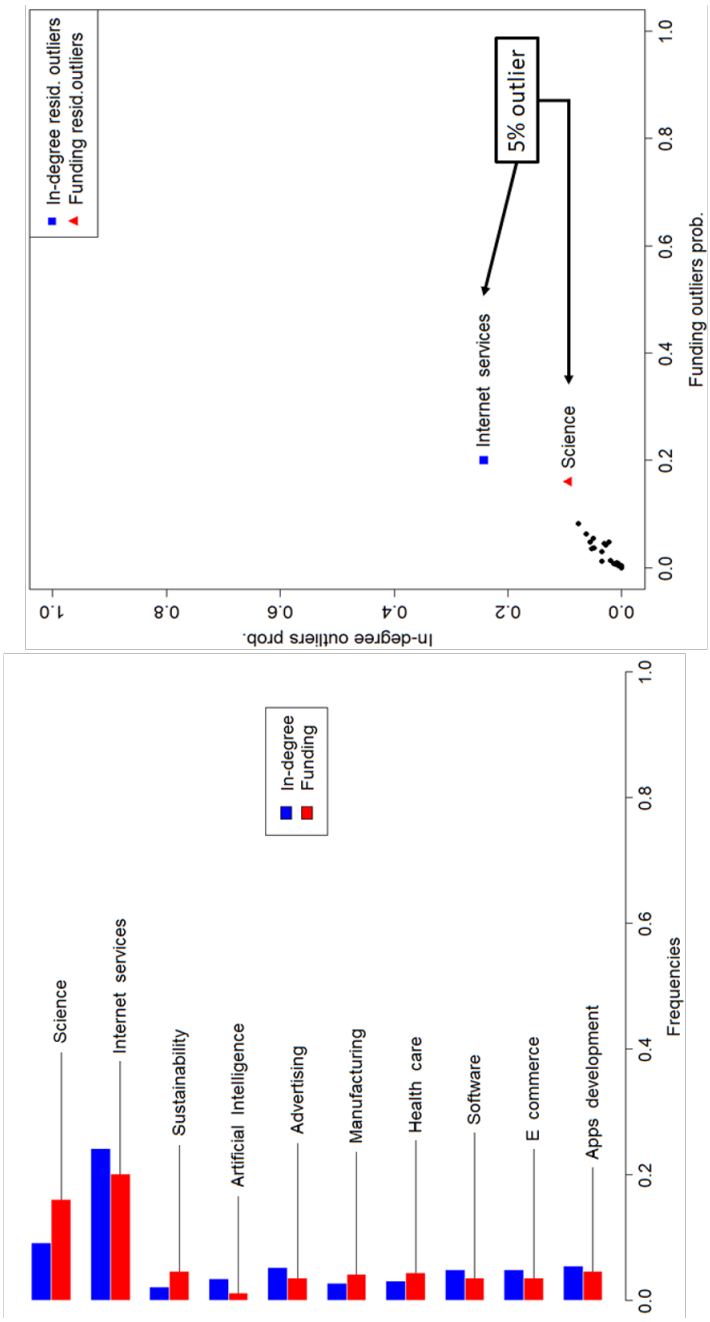


Fig. S3-a. Comparison between Outdegree and Funding distributions for Nationality.

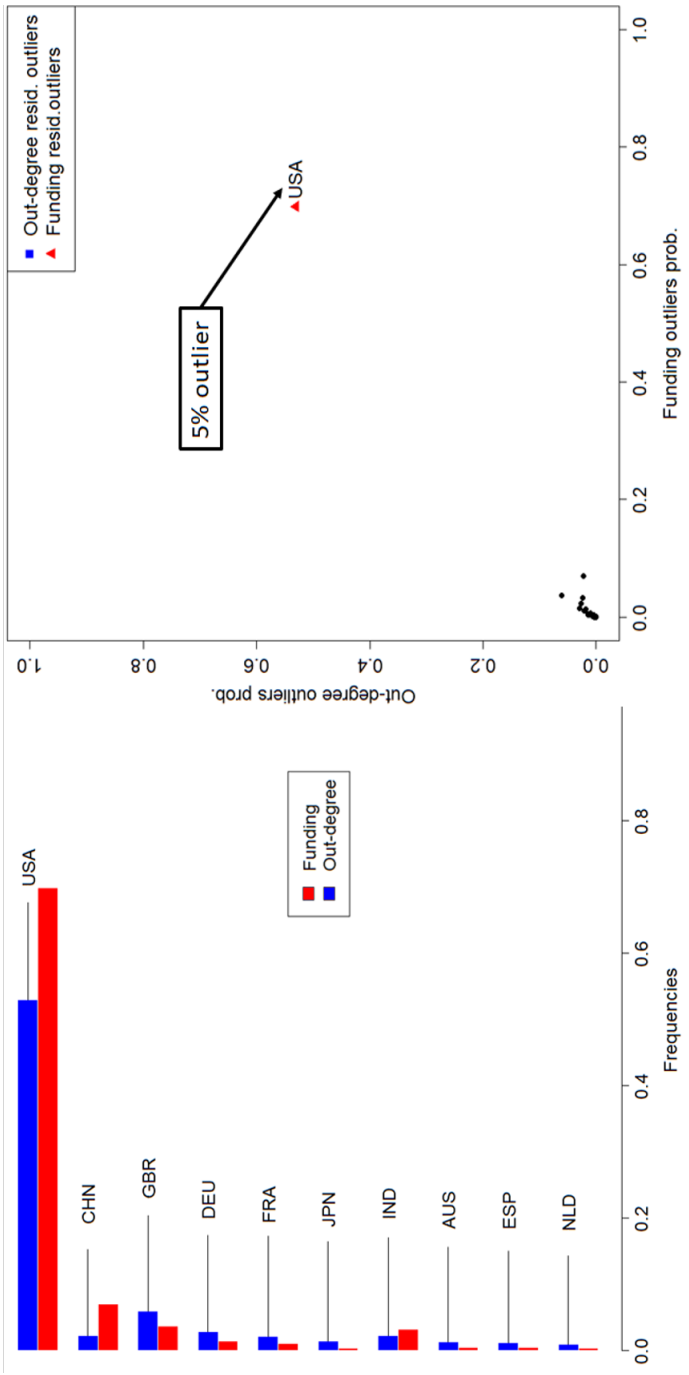


Fig. S3-b. Comparison between Outdegree and Funding distributions for Investor Type.

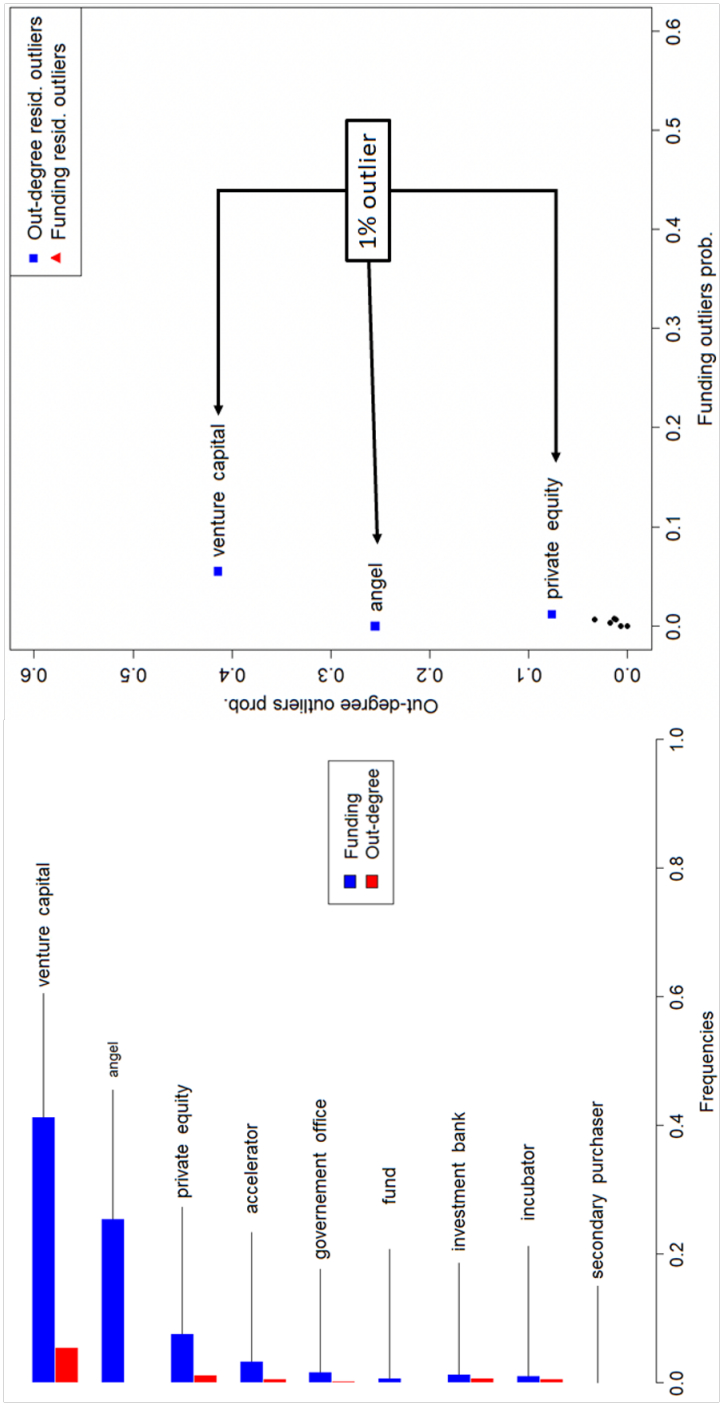


Fig. S3-c. Comparison between Outdegree and Funding distributions for Economic Category.

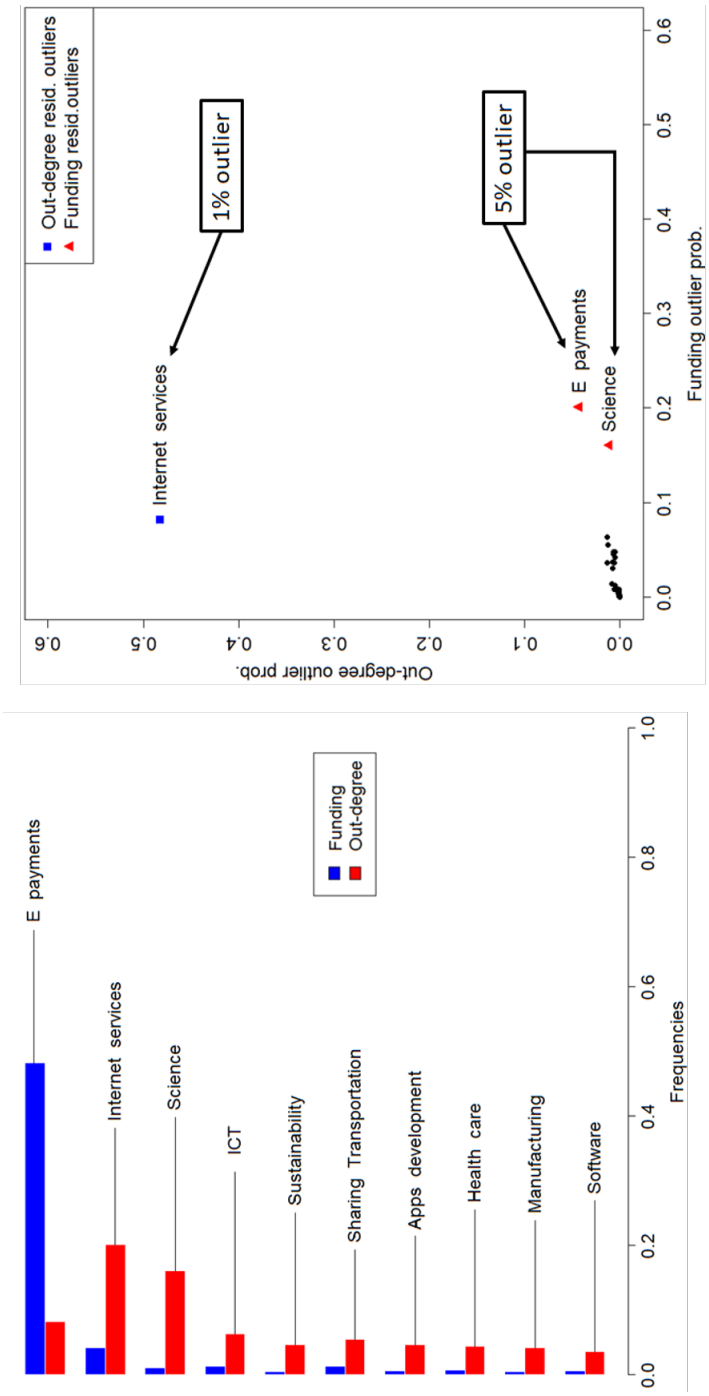


Fig. S4-a. Comparison between Betweenness and Funding distributions for Nationality.

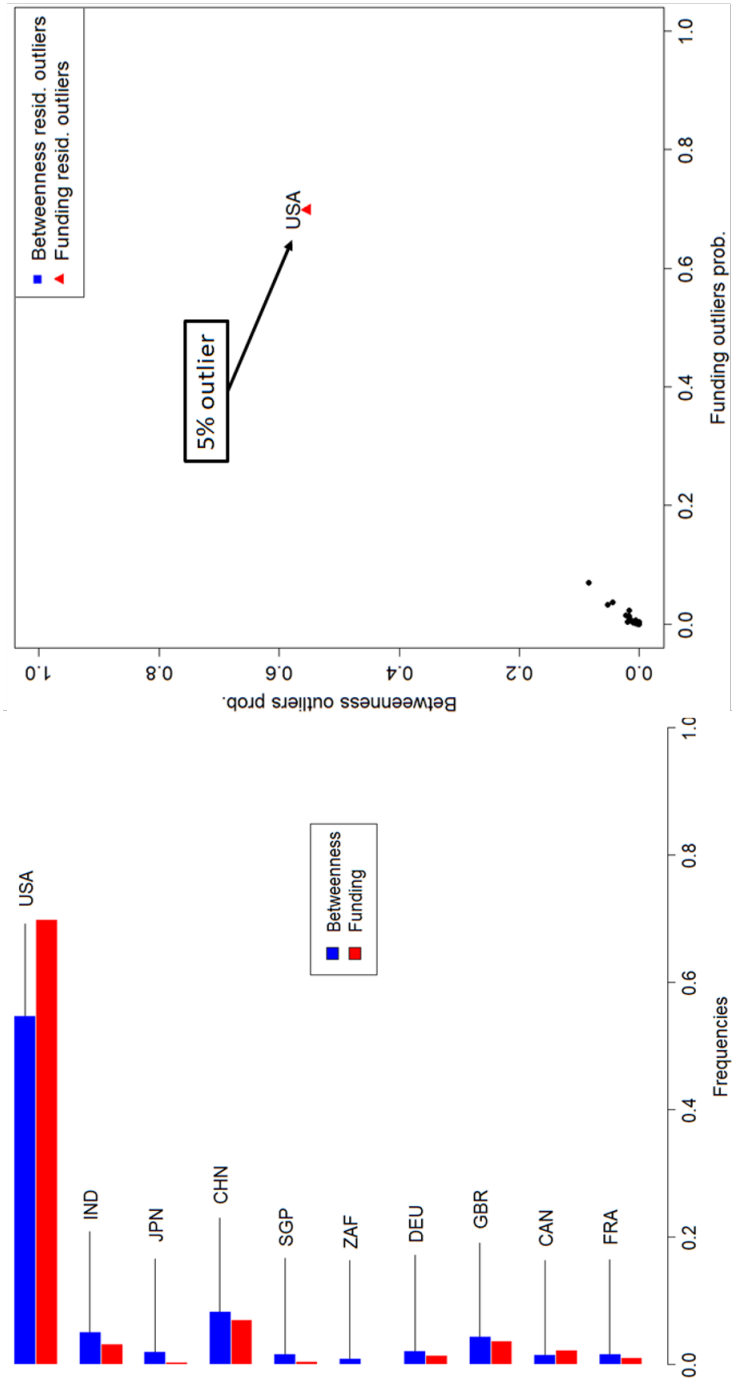
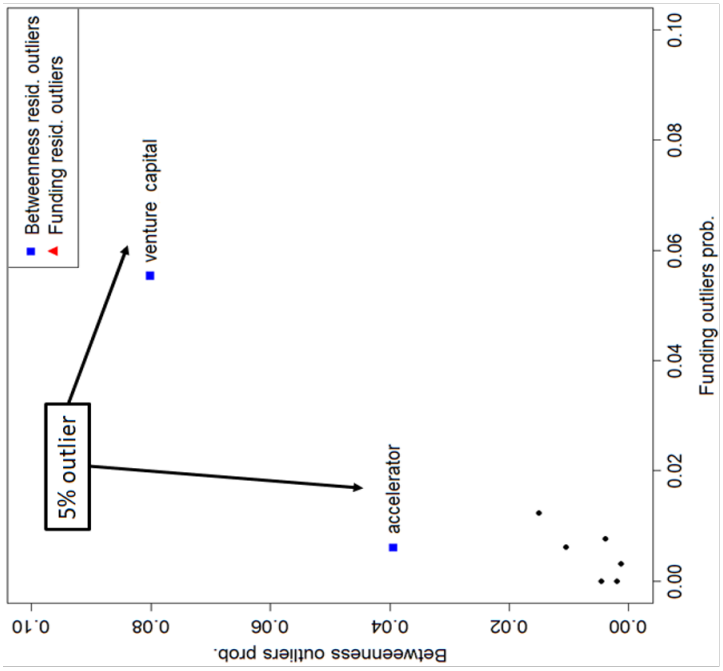


Fig. S4-b. Comparison between Betweenness and Funding distributions for Investor Type.



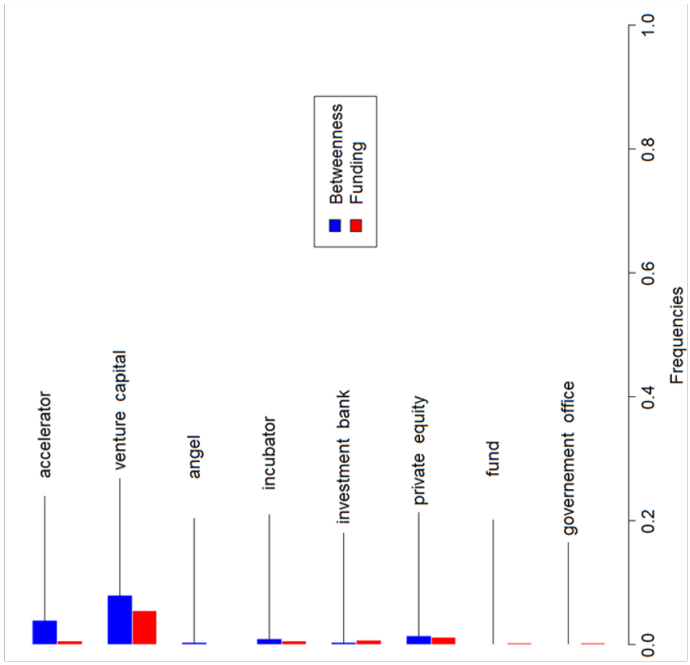


Fig. S4-c. Comparison between Betweenness and Funding distributions for Economic Category.

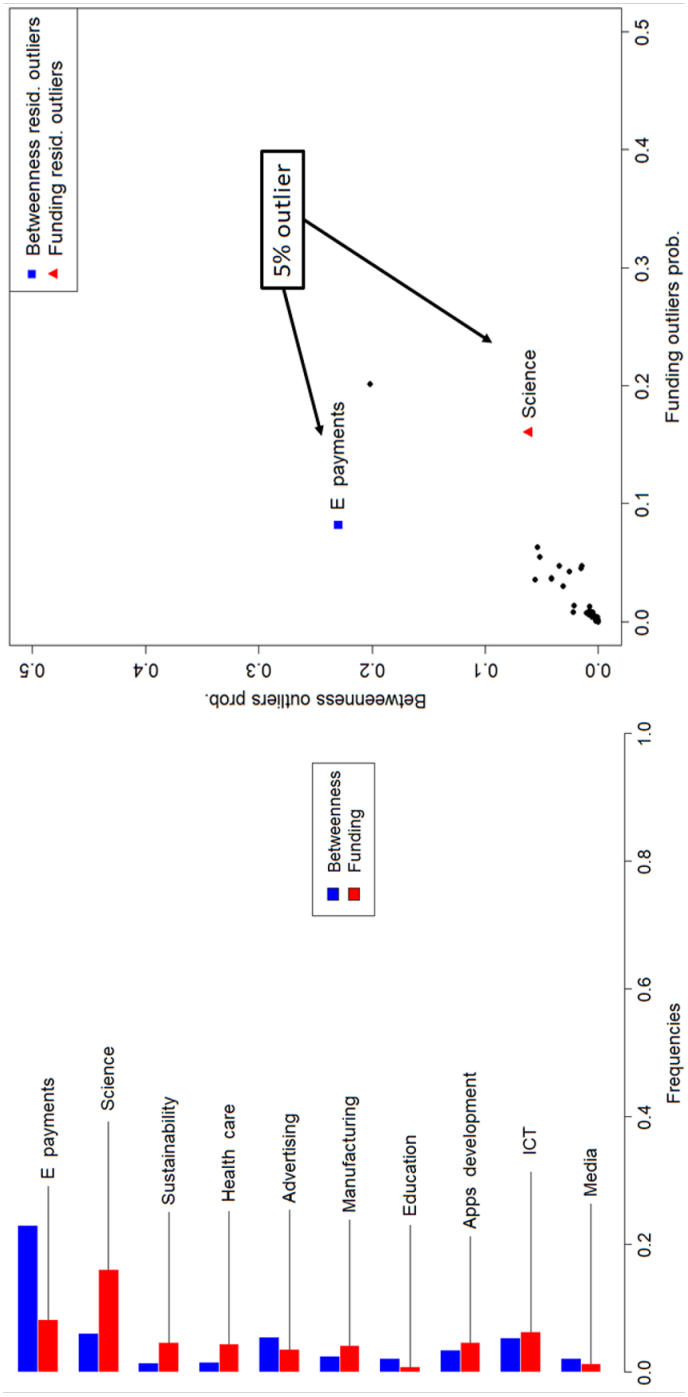
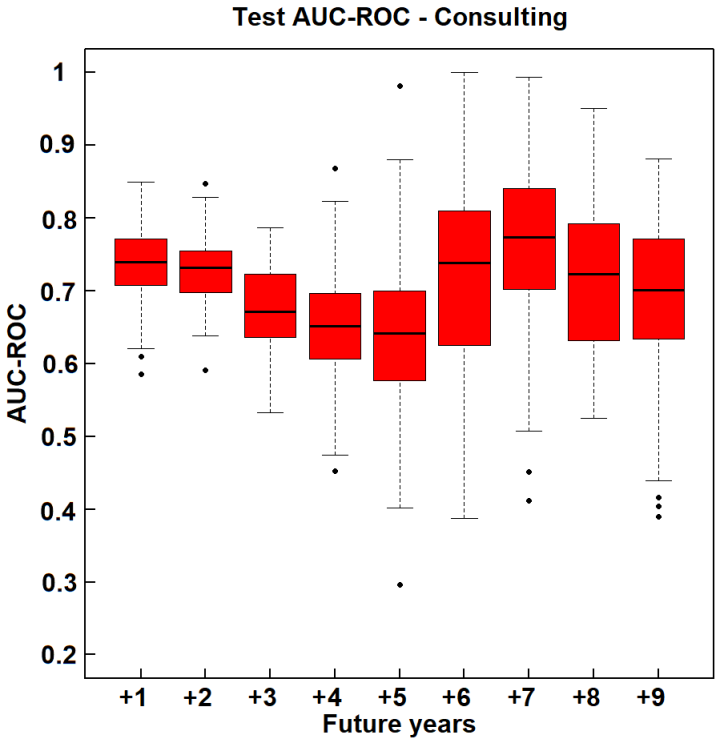


Fig. S5-a. AUC-ROC for the coarse Economic Categories, part 1.



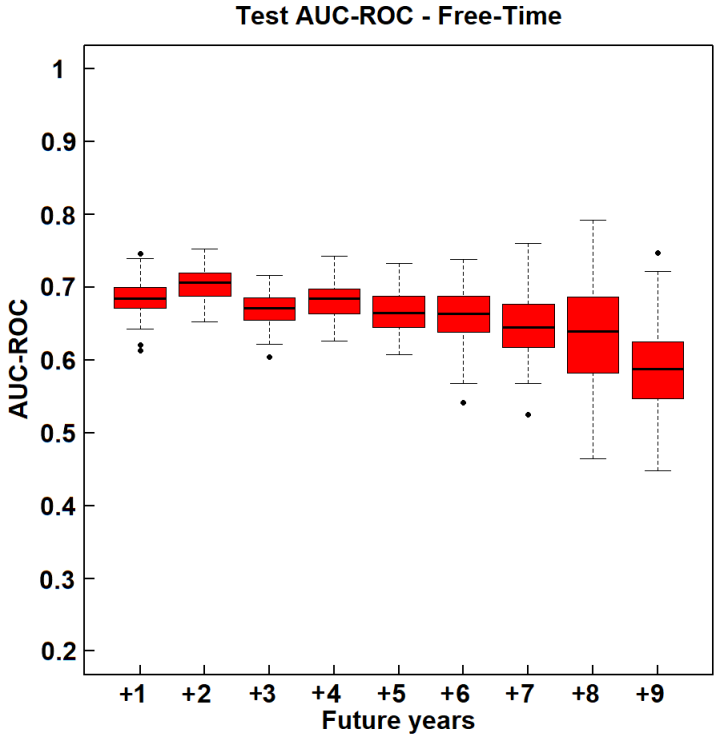
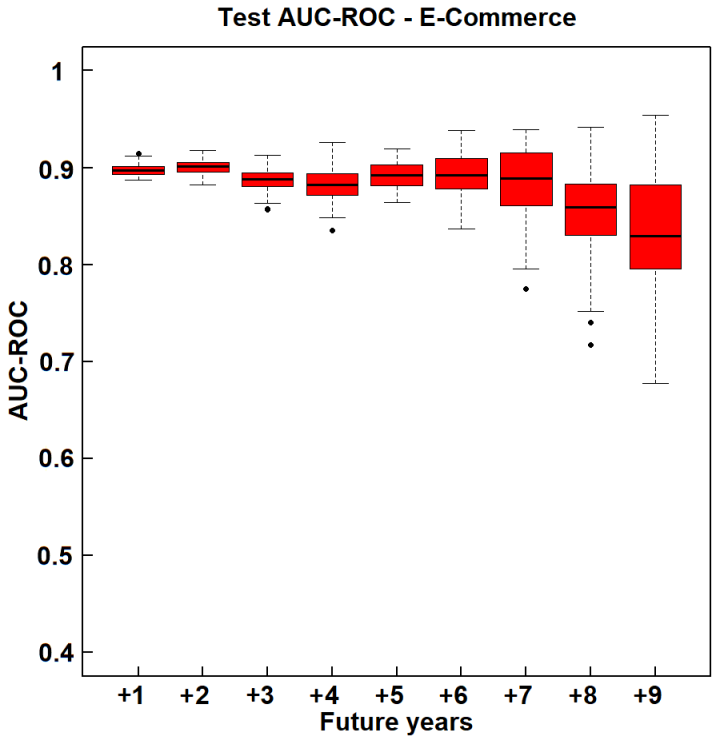


Fig. S5-b. AUC-ROC for the coarse Economic Categories, part 2.



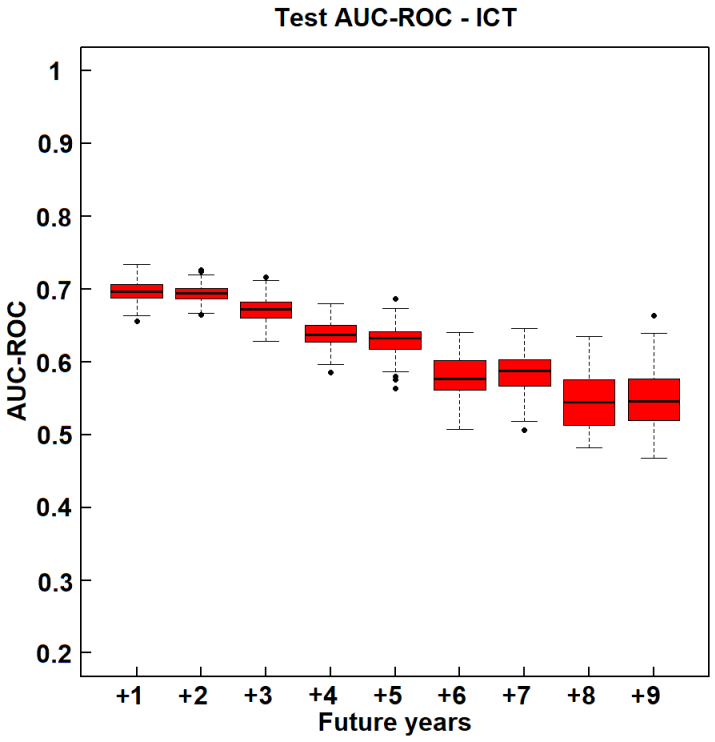
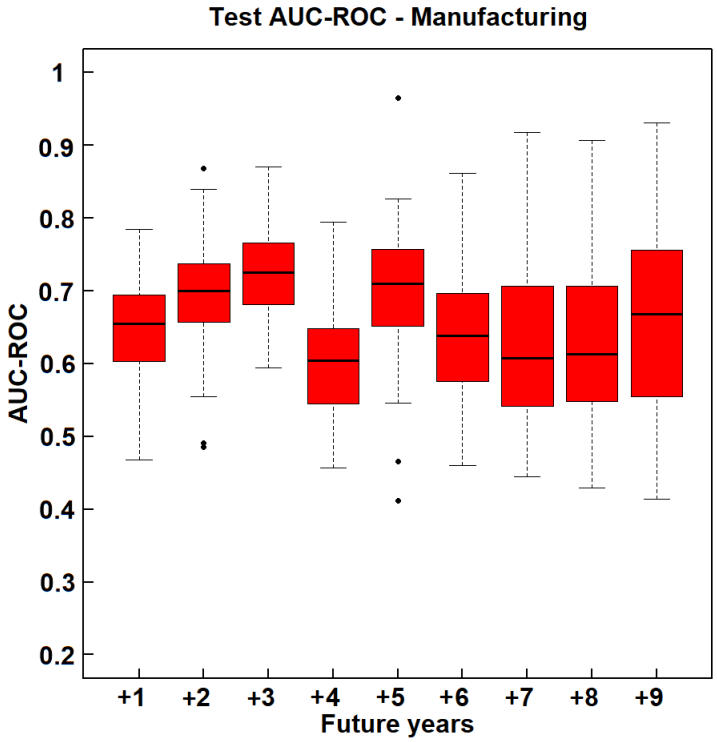


Fig. S5-c. AUC-ROC for the coarse Economic Categories, part 3.



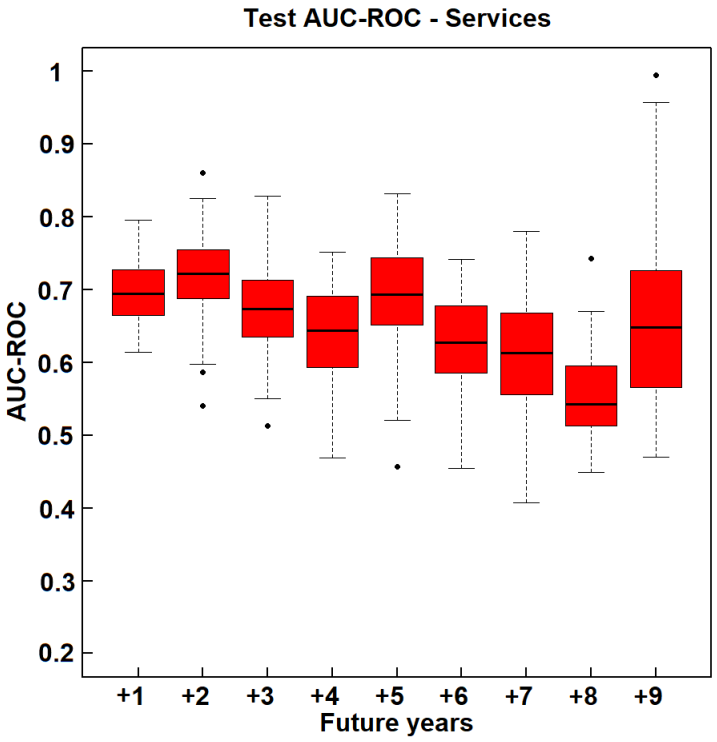


Fig. S5-d. AUC-ROC for the coarse Economic Categories, part 4.

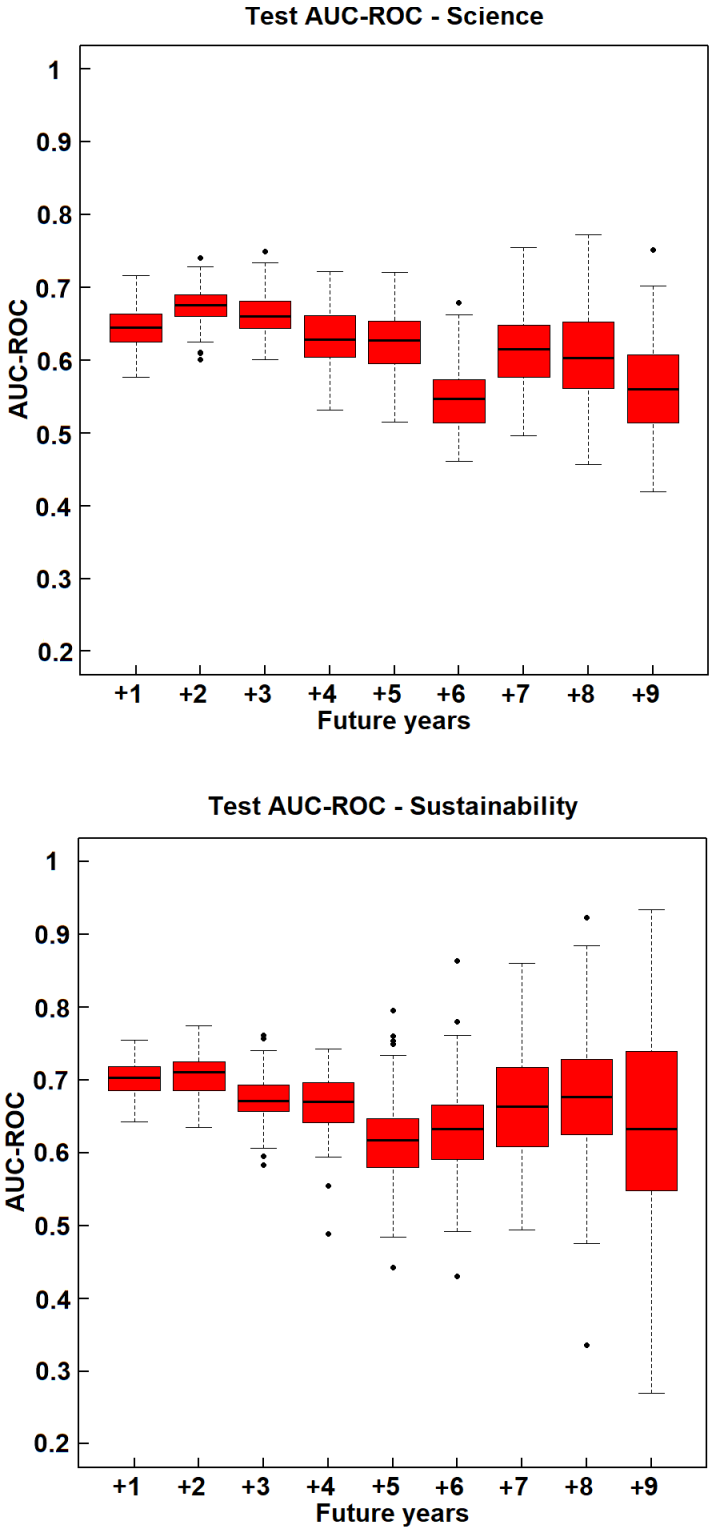
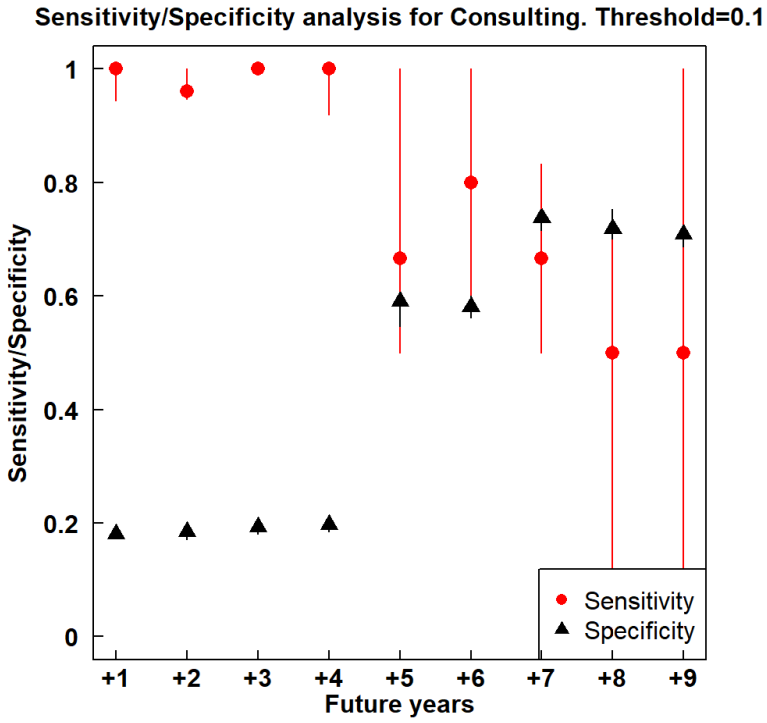


Fig. S6-a. Sensitivity and Specificity for the coarse Economic Categories, part 1.



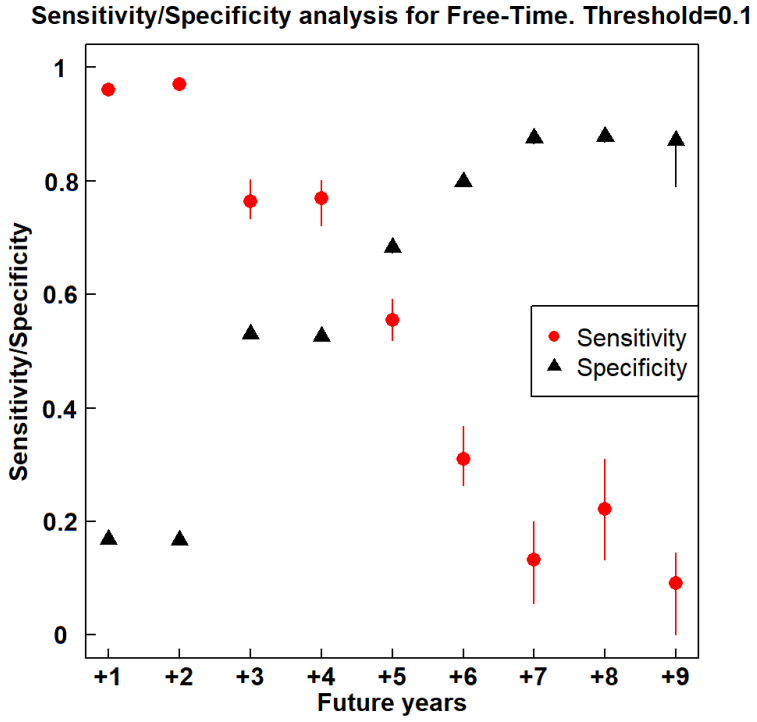


Fig. S6-b. Sensitivity and Specificity for the coarse Economic Categories, part 2.

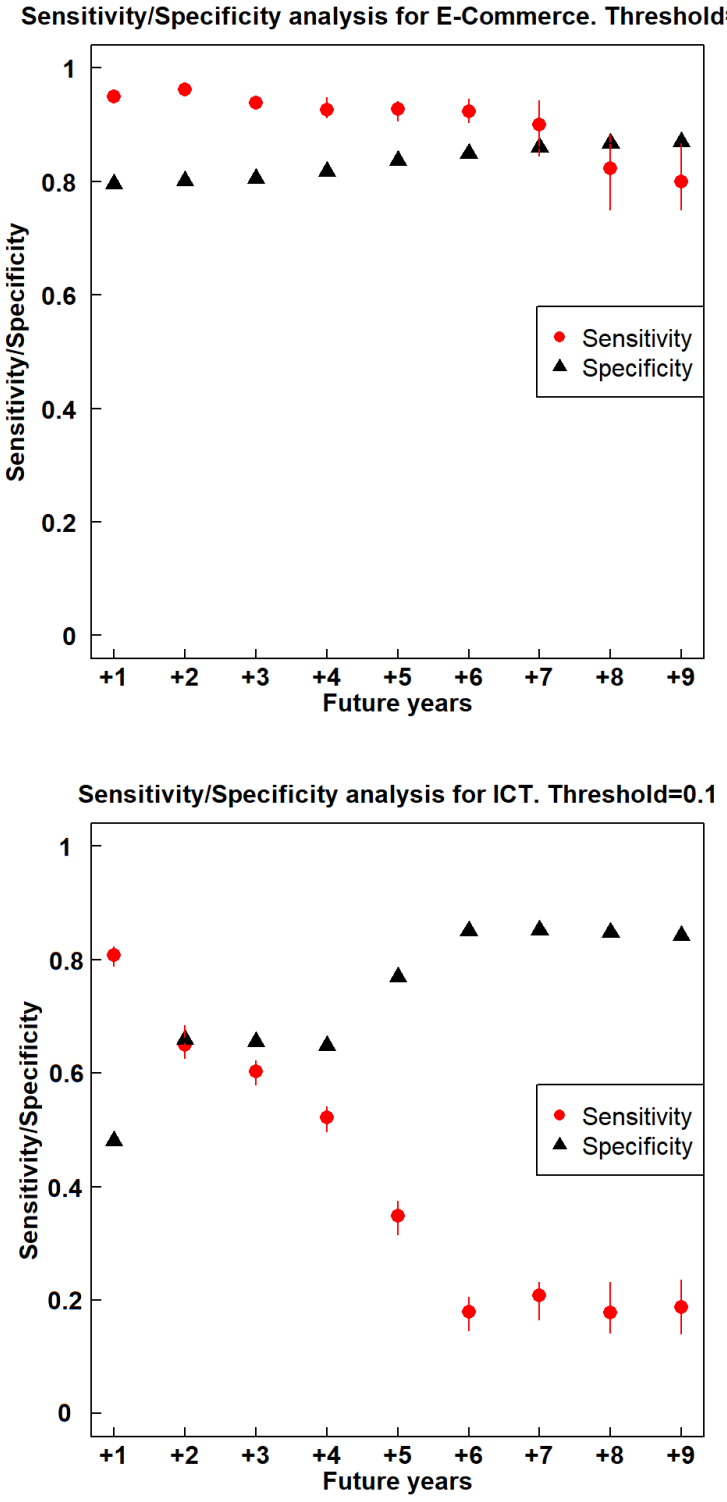
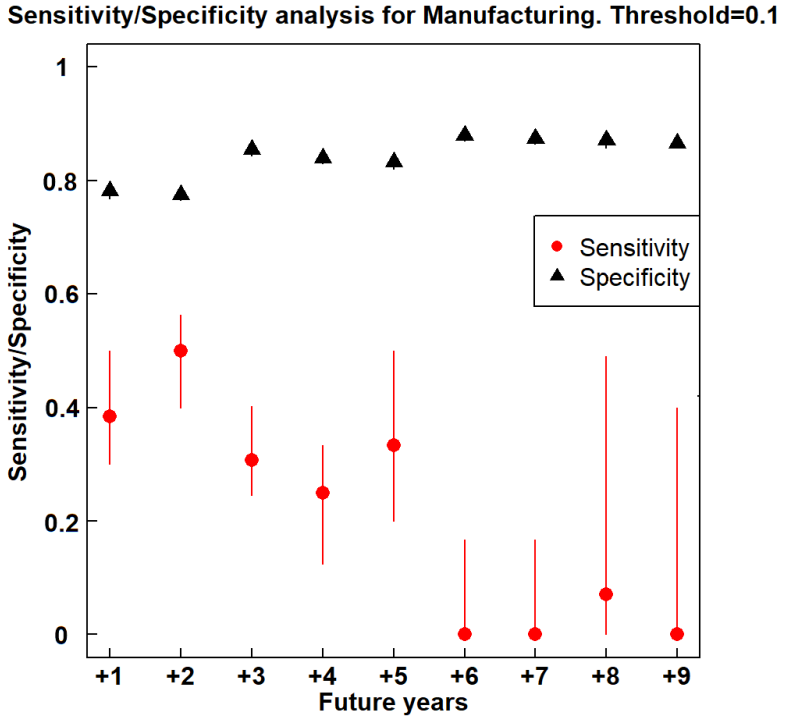


Fig. S6-c. Sensitivity and Specificity for the coarse Economic Categories, part 3.



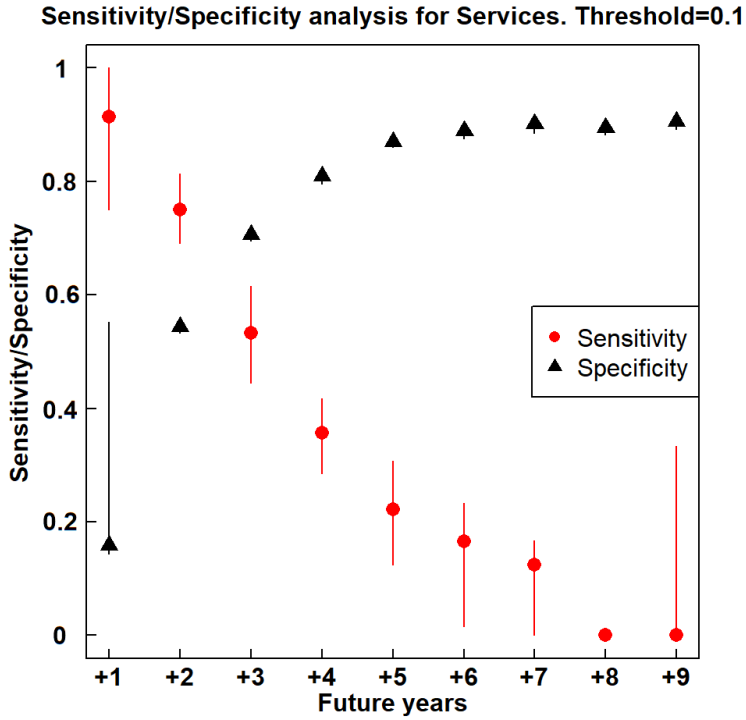
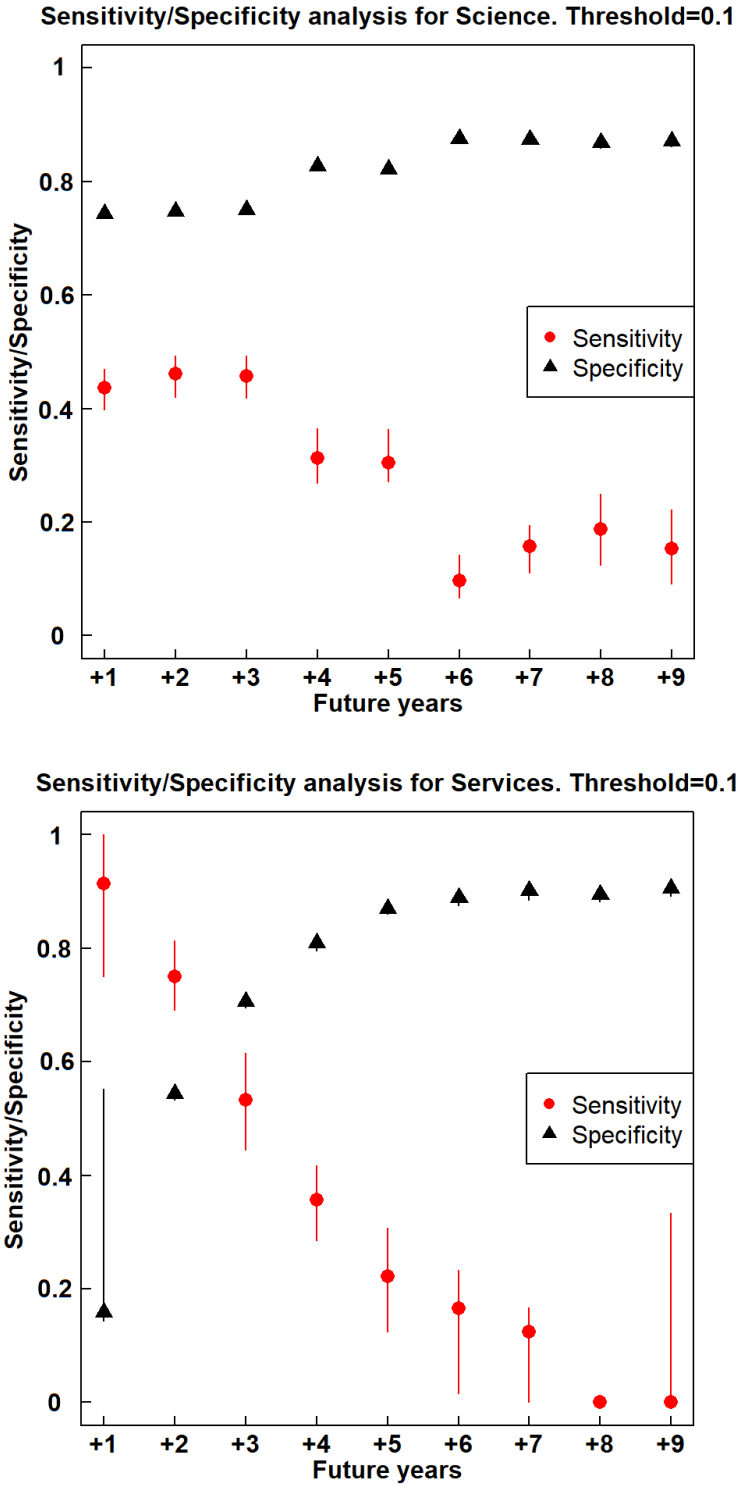


Fig. S6-d. Sensitivity and Specificity for the coarse Economic Categories, part 4.



Appendix S1. Statistical analyses.

By definition, an outlier is an observation exceeding a distance of 1.5 $IQR(X)$ from the first and third quartiles of its distribution X, where X = funding, in-degree, out-degree and betweenness. Funds and centrality measures have long-tail distributions with large skewness and this feature may entail an overestimation of outliers. To tackle this issue, we bootstrapped the experimental distributions ten thousands of times, estimated each time the left and right outlier thresholds and finally averaged these results. Accordingly, only observations exceeding these robust averaged left and right thresholds were identified as outliers.

All statistical tests performed in this work are non-parametric Kolmogorov-Smirnov tests. Calculated p-values were corrected according to multiple hypothesis testing with Bonferroni correction. We are aware this choice can be too conservative in many applications, however in this study we aimed at presenting a tool to quantitatively evaluate importance and the strategic nature of some elements, more than existing differences; although the interested reader can evaluate all existing differences and perspectives enlightened by our approach. We preferred limiting our discussions to those significant according to the most stringent criterion provided by Bonferroni.

Finally, to evaluate the effect of centrality measures on funding outliers, we considered centrality measures for each year, from 2000 to 2017 and assigned the label 1 to those firms resulting funding outliers in the future $1, 2, ..., 10$ years. Thus we obtained 10 distinct datasets d_i; for each one we performed 100 $10-$fold cross-validation analyses to determine the model accuracy: we randomly split a dataset d_i in 10 parts, one part was used for validation, the remaining ones for training.

The findings presented in this work exploit the informative content provided by aggregate funds collected by each firm until 2017. However, another possibility would consist in considering funds collected year by year. We chose to consider aggregate funding to maximize the information provided by Crunchbase data. Accordingly, we were able to take into account:

- the information deriving from the overall temporal series of collected funds and exploiting it to obtain an accurate model of success;
- the economic interplay established over time and lasting bonds which therefore shape the network structure;

Considering funds collected over a long temporal range makes the aggregate network less sensitive to statistical fluctuations, which on the contrary would be significant when considering only one year of funds. Aggregating funds and therefore connections weakens the weight of each year with respect of the whole time series; the longer the series, the weaker the importance of each year. Therefore, aggregating information can be useful to explore global trends and strengthen the model robustness.

To avoid any bias when estimating prediction accuracy, we excluded that information from the model by considering only the collected funds dating back to a year earlier, tow years earlier and so on.

Table S1. Organization of Crunchbase data.

File Name	Short Description
1. Acquisitions	Data about acquisitions
2. Category_groups	A list of all economic categories within the data
3. degrees	Educational qualification of tracked people
4. event_appearances	events and participating people
5. events	A list of all recorded events
6. funding_rounds	Description of funding rounds
7. funds	The file includes all present investment funds
8. investment_partners	partnerships established in funding rounds
9. investments	Information about leader investors in funding rounds
10. investors	A description of all Crunchbase investors
11. ipos	Firms at initial public offering stage
12. jobs	Job career of tracked people
13. org_parents	The list of subsidiaries and controller companies
14. organization_descriptions	Description of firm activities
15. organizations	A detailed description of all Crunchbase firms
16. people	A list of all people in Crunchbase
17. people_description	A description of tracked people

Table S2. Top ten ranking.

Ranking	Nationality	Economic category	Investor type
1.	USA (53.6%)	Internet services (19.3%)	Angel (60.4%)
2.	UK (7.6%)	e-Payments (14.4%)	Venture Capital (27.8%)
3.	IND (4.2%)	Software (6.1%)	Private equity (6.2%)
4.	CAN (3.0%)	Science (5.8%)	Accelerator (1.9%)
5.	CHI (2.9%)	ICT (5.6%)	Government Office (1.1%)
6.	DEU (2.8%)	e-Commerce (5.0%)	Incubator (1%)
7.	FRA (2.3%)	Sharing transportation (4.4%)	Investment bank (0.9%)
8.	ISR (1.7%)	Apps development (4.3%)	Fund (0.5%)
9.	AUS (1.5%)	Healthcare (4.1%)	Secondary purchaser (0.03%)
10.	ESP (1.3%)	Advertising (3.9%)	Startup competition (0.003%)

Table S3. Top fifty ranking.

Rank	Funding	Indegree	Outdegree	Betweenness
1.	Verzion Comm.	Uber	500 Startups	Y Combinator
2.	Tsinghua Unigr. Int.	Atrium LTS	Y Combinator	FundersClub
3.	Didi Chuxing	Flexport	Sequoia Capital	Techstars
4.	Tesla	DocuSign	New Enterprise Associates	StartX
5.	China Unicom	Pinterest	Intel Capital	Alibaba
6.	Uber	SeatGeek	Accel Partners	Alchemist Accel.
7.	Rosneft	Opendoor	NYSERDA	Groupon
8.	WeWork	CardioDx	Kl. Perk. Cauf. & Byers	Salesforce
9.	AT&T Wir. Mob. Gr.	Lyft	SOSV	Google
10.	Alibaba	Prosper	Wayra	Crowdcube
11.	Meituan-Dianping	Fab	Draper Fisher Jurvetson (DFJ)	Seedcamp
12.	Flipkart	Mattermark	SV Angel	Betaworks
13.	Clearwire	Active Network	Start-Up Chile	Startupbootcamp
14.	Hilton Worldwide	TransMedics	Bessemer Venture Partners	Seedrs
15.	Apple	Tesla	Techstars	WR Hambrecht
16.	SH Pudong Dev. Bank	Practice Fusion	Right Side Cap. Manag.	Baidu
17.	Sberbank	Domo	Technology Development Fund	DST Global
18.	COFCO	Neuronetics	Greylock Partners	AngelList
19.	Jumpstart Ltd	PTC Therapeutics	First Round Capital	500 Startups
20.	Charter Comm.	Airbnb	Goldman Sachs	AOL
21.	Ping An	EndoGastric Sol.	Index Ventures	Tencent Hld.
22.	Suning	ecomom	Lightspeed Venture Partners	Digital Curr. Gr.
23.	Ant Financial	Artsy	Battery Ventures	Slack
24.	Airbnb	Bluesmart	Plug and Play	Amplify.LA
25.	Gas Natural	Scopely	High-Tech Gruenderfonds	Yahoo
26.	Nvidia	Namely	Crowdcube	Visionplus
27.	Evonik Industries	Pivot3	Brand Capital	Rock Health
28.	First Data Corp.	Sun Basket	Venrock	OurCrowd-GCai
29.	Grab	Keen IO	Andreessen Horowitz	Didi Chuxing
30.	Ele.me	Memebox Corp.	Benchmark	JFDI.Asia
31.	AccorHotels	Klout	General Catalyst	CircleUp
32.	Xerox	Boxed	Khosla Ven.	Amazon
33.	Allegro	Spotify	Norwest Ven. Ptrs - NVP	Anthemis Group
34.	Toys —R—	Meru Networks	GV	Cisco
35.	Toutiao	Doppler Labs	Redpoint	Kickstarter
36.	Ola	Casper	Menlo Ventures	Uber
37.	Reliance Jio Inf. Ltd.	Kamcord	Canaan Partners	Xiaomi
38.	B2M Solutions	Proterra	Atlas Venture	Lighter Capital
39.	Magic Leap	Actelis Networks	Northstar Ventures	SeedInvest
40.	Roche	Luxe	Matrix Partners	Entrepreneur First
41.	Lazada Group	Calient Tech.	Pol. Partners	LetsVent.
42.	Snap Inc.	Slack	U.S. Venture Ptrs (USVP)	Garage Tech. Ven.
43.	Lyft	GENBAND	Seedrs	PayPal
44.	Safaricom	Black Duck Sw.	Silicon Valley Bank	Snapdeal
45.	Delivery Hero	ColorChip	Foundation Capital	Silver Lake Ptrs
46.	Spotify	SpotHero	Mayfield Fund	Wefunder
47.	Univ. Studios Jp.	Path	Kima Ventures	Imagine K12
48.	Infor	Optimizely	IDG Capital Partners	Rocket Internet
49.	One97 Comm.	Beepi	Startupbootcamp	HIGHLINEvc
50.	Xiaomi	LeadGenius	CRV	One97 Comm.

Appendix S2. Firm-level variables in the Crunchbase dataset

Although the focus of our model is to investigate and quantify the relationship between network centralities and the ability of startups to obtain funding in funding rounds, it is natural to wonder to what extent specific firm-level variables affect the classifier's performance. To answer this question, in Section **??** we developed a variant of the original model, obtained by combining both network centralities and categorical variables that characterize the startups available in the Crunchbase dataset. In particular, the firm-level variables that we considered in this task are as follows:

- Investor type. 34 619 of the 121 950 subjects in Crunchbase are actually investors, belonging to one of the following 10 categories: angel (20 930), venture capital (9 619), private equity (2 153), accelerator (667), government office (397), incubator (331), investment bank (322), fund (188), secondary purchaser (11) and startup competition (1). However, the value of this variable is not available in around 71.6% of the cases, indicating that the corresponding elements are not investors.

- Coarse economic category. The economic categories, available in Crunchbase for 81 322 economic subject out of 121 950 (66.7%), are numerous and too specific for some tasks: in a classification problem, some categories are underrepresented and cannot be efficiently employed. For this reason, we introduce 8 coarse economic categories, in which the original refined categories are grouped: Public services, Media and free time, ICT and Computer Science, Science, E-services, Consulting, Engineering and manufacturing, Green economy. A detailed review of the correspondence between the fine (38 categories) and coarse (8 categories) subdivisions is reported in Tab. S4 Table in Supplementary Information.

- Country development groups. Information on nationality is available for 104 628 economic subjects out of the 121 950 surveyed in Crunchbase (85.8%). Although these companies are based in a large number of countries (160), the geographical distribution of Crunchbase elements is extremely inhomogeneous, as we already observed in Section **??**. For this reason, using nationality of Crunchbase subjects as a categorical feature to predict funding would be unfeasible. However, since geographical information could be relevant in determining the ability of a startup in collecting funds, we applied the same logic as for the economic categories, grouping the 160 countries in 4 development groups. Such a partition is obtained in Ref. [**?**] by applying community detection to a complex network of United Nations Member States (UNMS), constructed from World Bank development indicators[1]. The vast majority (76.5%) of economic subjects in Crunchbase belongs to the most developed group, while the remainder is divided among the upper-middle development group (3.1%), the lower-middle development group (4.8%), the least developed one (0.6%) and subjects which are not assigned to any development group because their nationality is either missing in Crunchbase (14.2%) or not included in the UNMS set (0.9%).

- Employee count. Crunchbase also contains information on the size of 59 655 companies (48.9% of the cohort), quantified by a categorical variable related to the number of employees, that can fit into 9 different ranges: 1–10 employees (20 373 firms), 11–50 (21 560), 51–100 (6974), 101–250 (2396), 251–500 (2010), 501–1000 (2362), 1001–5000 (1405), 5001–10 000 (992), over 10 000 (1583).

Before training the classifier on the combined set of centrality metrics and firm-level features, the categorical variables were preprocessed to properly manage missing entries. Since missing entries in the Investor type value correspond to non-investor startups, we

[1] https://databank.worldbank.org

interpreted it as a category of its own. The same treatment has been done for the missing entries of the Coarse economic category, Country development group and Employee count variables: this strategy was adopted, instead of, e.g., replacing missing entries with the mean value of the available ones, because values are expressed as ranges or categories and not as numerical values.

S4 Table. Coarse economic category grouping.

Refined categories	Cardinality	Coarse categories	Total cardinality
Administrative services	563		
Community services	167	Public services	4124
Government office	86		
Healthcare	3308		
Advertising	3139		
Events	9		
Food and beverage	1407		
Media	2205	Media and free time	10 907
Sports and travels	943		
Travel	981		
Video production, editing	2223		
Apps development	3537		
Content editing	1		
Gaming	273		
Hardware	740		
ICT	4522	ICT and Computer Science	30 381
Internet services	16 079		
Platforms	4		
Privacy services	246		
Software	4952		
Telecommunications	27		
Artificial Intelligence	1851		
Data Analytics	297	Science	8697
Education	1808		
Science	4741		
E-commerce	4072	E-services	15 790
E-payments	11 718		
Consumer services	176		
Energy consulting services	170		
Professional services	1027	Consulting	2951
Real estate	1025		
Sales services	553		
Design	177		
Manufacturing	2430	Engineering and manufacturing	2631
Navigation	24		
Natural resources	309		
Sharing transportation	3553	Green economy	5841
Sustainability	1979		