Journal of Complex Networks (2021) **00**, 1–17
doi: 10.1093/comnet/cnab022

Markets as ecological networks: inferring interactions and identifying communities

CLIVE EMARY†
School of Mathematics, Statistics and Physics, Newcastle University, Newcastle-upon-Tyne NE1 7RU, UK

AND

HUGO FORT
Institute of Physics, Faculty of Science, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay
†Corresponding author. Email: clive.emary@newcastle.ac.uk

Edited by: Petter Holme

[Received on 16 February 2021; editorial decision on 10 June 2021; accepted on 15 June 2021]

Financial markets are paradigmatic examples of complex systems and have been compared to ecological networks in which different species (firms) interact and co-evolve. A central object governing species dynamics in ecology is the community matrix, whose elements are closely related to pairwise interspecific interaction coefficients. Using this ecological analogy we propose a method, based on the Maximum Entropy (MaxEnt) principle, that allows us to infer candidates for an economic community matrix from time series data of market values. To assess the usefulness of this picture, we construct community matrices for a set of companies belonging to the Fortune 500 list and perform a community analysis on the resultant networks. This analysis shows these networks to strongly reflect the known industry groupings of the firms. We conclude therefore that our community matrices capture non-trivial information about the interaction of firms, not immediately apparent from the covariance of market values. We anticipate our approach being useful in elucidating further aspects of market structure, as well as forming the basis of forecasting market dynamics.

Keywords: MaxEnt, business ecosystem, ecological networks, community detection, modularity.

1. Introduction

Financial markets are paradigmatic examples of complex systems, comprising a large number of interacting agents (traders and algorithms) whose decisions dictate their ongoing evolution. Correspondingly, markets have been regarded as being similar to ecosystems in which species interact and co-evolve [1–5], and population dynamics has been proposed as a model of the dynamics of companies regarded as species in the 'business ecosystem' [3, 4]. This ecological paradigm has been applied to different industrial sectors, such as the newspaper industry [6], the airline and oil-drilling industries [7], internet [8] and software firms [9].

Ecosystems comprise a biotic component (communities of living organisms) and an abiotic one (the non-living chemical and physical parts of the environment that affect living organisms), interacting as a single system [10]. A central piece of the biotic component is the ecological interactions between organisms. Such biotic interactions, which can be regarded as operating either between individual organisms or entire species, are often difficult to define and measure [11, 12]. The most direct procedure to estimate the

pairwise interaction coefficients, either in natural or artificial biological assemblies of species, is through pairwise competition trials that compare the species yields in biculture relative to monoculture [13–15]. These experiments, common in community ecology and agriculture science, are not feasible in systems like markets since we cannot isolate single firms from the rest in order to study their evolution under controlled conditions. It becomes necessary, therefore, to resort to indirect methods to infer the nature and strength of interactions between firms. In this paper we approach this problem using the principle of maximum entropy (MaxEnt).

The MaxEnt principle, introduced by Jaynes [16, 17], is a general method to make the least-biased inferences compatible with available data. Jaynes' MaxEnt formulation can be viewed as a method of making predictions from limited data by assuming maximal ignorance about unknown degrees of freedom. MaxEnt has been successfully used to infer interactions from datasets in a wide variety of biological systems: from tropical forests [18–20] to networks of neurons [21, 22], and from gene expression in yeast [23] to flocks of birds [24].

Here, we use MaxEnt to infer the effective interaction coefficients for a set of companies in the Fortune 500 list. In total, Fortune 500 companies represented in 2019 two-thirds of the U.S. GDP with $13.7 trillion in revenues, $1.1 trillion in profits, and $22.6 trillion in market value [25]. The set we study includes the 38 companies with revenues among the largest in 2018 that were in this list for five consecutive years in a row, 2014–2018.

In theoretical ecological terms, the interaction coefficients that we derive for these firms can be interpreted as belonging to a *community matrix* [26], an object that has played [27, 28] and continues to play [29–32], a key role in understanding the collective properties of assemblies of interacting species. Our MaxEnt approach produces several related candidates for the community matrix for the firms we consider. We translate these community matrices into adjacency matrices [33, 34] by mapping interaction strengths onto edge weights. This then establishes a precise analogy between the relationships between firms in a market and an ecological network in which the nodes represent species (firms) and the edges, pairwise interactions.

To explore how well the resultant networks reflect salient information about the relationships between firms, we subject the networks to a community analysis [34], the aim of which is to identify groups of firms (communities) that interact more strongly with one another than they do with the rest of the network. Again, such an analysis has its counterpart in ecological theory [35, 36], where the connection between modularity and ecosystem stability has been a theme of particular interest [31, 37, 38]. We find strong evidence for a modular structure in network of firms. Moreover, we compare the communities obtained from our network analysis with an independent classification of firms into groups, namely the 24 GICS industry groups [39]. This comparison shows that the network modules derived from MaxEnt interaction matrices strongly reflect the industry groupings. By considering overlap with the GICS industry group as a measure of fitness, we are able to determine which of the possible MaxEnt community matrices is the best choice for describing the clustering of interactions within the network.

Thus, we present a methodology for directly extracting from empirical time-series market data a network representation of the interactions of firms within a market. We speculate that this methodology will be useful in the future both for the analysis of market structures, as well as forming the basis of an approach towards the dynamics of market values.

2. MaxEnt and the community matrix

The maximum entropy (MaxEnt) principle is a general method to make the least-biased inferences compatible with available data [16, 17]. The lack of knowledge we generally have of a real system can

be modelled by a probability distribution for the different possible observable states of that system. This probability distribution will typically not be known, and in fact there are many possible choices for it that are compatible with observations. Of all such distributions, the recipe of MaxEnt is to choose the probability distribution that maximizes the information entropy subject to the constraints of the available data. We consider a system of N random variables, $\{v_i \mid 1 \leq i \leq N\}$, arranged into the vector $\mathbf{v}$. Then, if we assume that the only information we possess about these variables is their mean $\overline{\mathbf{v}}$ and their covariance matrix Σ, the MaxEnt principle posits the following joint probability distribution

$$P(\mathbf{v}) = [\mathrm{Det}\, 2\pi\, \Sigma]^{-1/2} \exp\left[-\frac{1}{2}\,(\mathbf{v}-\overline{\mathbf{v}})\cdot\Sigma^{-1}\cdot(\mathbf{v}-\overline{\mathbf{v}})\right]. \tag{1}$$

This distribution has previously been interpreted in terms of the equilibrium statistical mechanics of a generalised Ising model [18] in which $J = -\Sigma^{-1}$ is the matrix of interaction strengths between 'spins' of length v_i. In this work, however, we consider an interpretation of Eq. (1) in terms of ecosystem interactions where, as we now show, the matrix J can be related to a community matrix.

Let $\mathbf{\Delta} = \mathbf{v}(t) - \overline{\mathbf{v}}$ be the vector of deviations from the mean, and let $\mathbf{F}(t)$ be a random force vector with zero mean and temporal correlations described by

$$\langle \mathbf{F}(t)\mathbf{F}(t')\rangle = 2B\,\delta\left(t-t'\right), \tag{2}$$

were B is a symmetric matrix. In this light, the MaxEnt distribution of Eq. (1) can then be interpreted as the stationary probability distribution of the Langevin equation [40]

$$\frac{d}{dt}\mathbf{\Delta} = C\cdot\mathbf{\Delta} + \mathbf{F}(t), \tag{3}$$

provided that matrix C satisfies

$$C\cdot\Sigma + \Sigma\cdot C^{T} = -2B, \tag{4}$$

with T denoting the matrix transpose. In a population-dynamics setting, Eq. (3) describes behaviour of populations $\mathbf{v}$ near stationary point $\overline{\mathbf{v}}$ with $\mathbf{F}(t)$ describing external (typically abiotic) forcing [41] and with matrix C the community matrix, describing interactions between the species[1].

Further progress can be made by splitting C into two components: $C = K + \Omega$, the first of which gives the symmetric part of $C\cdot\Sigma$, and the second, the antisymmetric part. From Eq. (4), we see that the symmetric part obeys

$$K = -B\,\Sigma^{-1}. \tag{5}$$

This equation shows that although Σ^{-1} does not determine K uniquely by itself, with the addition of a noise model (B) it does. In principle, the noise affecting different species could be correlated, but here, we consider all correlations to come from interactions, such that the noise for each firm is

[1] Technically, the community matrix is defined as the Jacobian of a set of nonlinear equations of motion, such as a generalised Lotka–Volterra model, evaluated at a particular stationary point. Its elements describe the influence of species on one another for small changes in populations near their equilibrium value.

independent. We will consider two particular cases. Firstly, we posit that the noise is of constant strength for each firm with amplitude scaled to one, that is, $B_{ij} = \delta_{ij}$ with δ_{ij} the Kronecker delta. This gives us the community-matrix contribution

$$K = J = -\Sigma^{-1}, \tag{6}$$

which is the same interaction matrix implied by the Ising-model interpretation. Ignoring for a second interactions between firms (i.e. just looking at the diagonal elements of K), in this model the relative size of the fluctuations in v_i is determined by the size of the 'restoring rate' $|K_{ii}|$, since the noise driving each market value is the same. In our second model, we assume that the restoring rate $|K_{ii}|$ is constant for all firms, and that it is the differences in noise amplitude that drives the difference in the size of fluctuations. Thus, we set $B_{ij} = |J_{ii}|^{-1}\delta_{ij}$ in Eq. (5) such that the community-matrix contribution reads

$$K = I \qquad \text{with} \qquad I_{ij} = J_{ij}/|J_{ii}|. \tag{7}$$

This has diagonal elements $I_{ii} = -1;\ \forall i$. Thought of in term of Ising-model interaction strengths, matrix element I_{ij} is the strength of the interaction between firms i and j relative to the self-interaction of i. Note that whereas matrix J is symmetric, matrix I is not.

The MaxEnt procedure with fixed mean and covariance tells us nothing about the antisymmetric part Ω of the product $C \cdot \Sigma$. For want of additional information, we therefore assume this contribution is zero and base our subsequent analysis exclusively on the part K—a point we will return to in the discussions.

3. Market data

The NYSE market has 2800 listed firms and because we do not have time series data for all 2800 firms and working with 2800×2800 matrices would increase the difficulty of the analysis considerably, we consider a subset of these firms, roughly corresponding to the "largest" firms in our data set. This is similar to the situation in ecological networks, which typically only include the most relevant species, i.e. the most abundant ones or those species that are *a priory* expected to exhibit stronger effects over other species [18]. In detail, then, the firms we consider in our analysis were selected according to the following criteria:

(1) They are all amongst companies with the largest revenues in the Fortune 500 list as of March 29, 2018 [42], coinciding with day 1000 of the time series we have.

(2) They simultaneously were in the list for 5 years in a row, from 2014 to 2018[2].

(3) The market values of these firms were available in the 2014–2018 Fortune 500 lists[3].

This results in a list of 39 firms. However, inspecting the market value times series, we found that those corresponding to the two home mortgage companies created by the U.S. Congress, Fannie Mae and Freddie Mac, are almost exactly linearly dependent. This poses a problem for inferring the effective interaction matrix (see below) since this requires inversion of the covariance matrix Σ which becomes

[2] Thus, for example firms like Dell Technologies, ranked 35 in the 2018 Fortune 500 list [42], was not included because from 2014–2016 it was not in the Fortune 500 list.

[3] For example, there is no market value listed for State Farm Insurance in 2018 [42].

singular with the above linear dependence. To overcome this difficulty we treat these two firms as a single firm, FNMA+FMCC, by summing their market values. In this way, we obtain the set of 38 firms in Table 1. For each firm we also list its ticker symbol and its Industry Group based on Global Industry Classification Standard (GICS) [39, 43]. Taken together they represent always at least 25% of the total NYSE market value along this period [44], and thus this set constitutes a significant and important sample of the market.

For each of the firms in our set we consider a 1000-day time series of market values from 4 October 2014 to 29 March 2018. Let us define $v_i(t)$ to be the market value of firm $1 \leq i \leq 38$ in day $1 \leq t \leq 1000$ of this time series. In our subsequent analysis, we consider a coarse-graining of the data over different time scales. To this end, we divide the 1000-day data into time slices, each of length T and labelled by $1 \leq n \leq n_{\max} = \lfloor 1000/T \rfloor$. Within time slice n the mean market value of firm i is

$$\overline{v}_i^{(n,T)} = T^{-1} \sum_{t=(n-1)T+1}^{nT} v_i(t), \tag{8}$$

and market-value covariance matrix has elements

$$\Sigma_{ij}^{(n,T)} = T^{-1} \sum_{t=(n-1)T+1}^{nT} \left[v_i(t) - \overline{v}_i^{(n)} \right] \left[v_j(t) - \overline{v}_j^{(n)} \right]. \tag{9}$$

4. Community and adjacency matrix construction

In this section, we use matrices $\Sigma^{(n,T)}$ to construct community matrices and the interaction networks they imply. Our first step is to combine the matrices from different time slices. For the covariance and J matrices, we simply take the mean of the individual matrices over the time slices

$$\Sigma^{\text{tot}} = \overline{\Sigma} = \frac{1}{n_{\max}} \sum_{n=1}^{n_{\max}} \Sigma^{(n,T)}; \qquad J^{\text{tot}} = \overline{J} = \frac{-1}{n_{\max}} \sum_{n=1}^{n_{\max}} \left[\Sigma^{(n,T)} \right]^{-1}. \tag{10}$$

We could similarly define a total I matrix as the average over the individual I-matrices defined as in Eq. (7) for each time slice. Empirically, however, we find that for the following analysis a preferable procedure is to construct a total I matrix by obtaining mean values of on- and off-diagonal elements of J separately, and then combining. Thus we define the matrix I^{tot} elementwise as

$$I_{ij}^{\text{tot}} = \frac{J_{ij}^{\text{tot}}}{|J_{ii}^{\text{tot}}|}. \tag{11}$$

We then use the matrices Σ^{tot}, J^{tot} and I^{tot} as the bases of adjacency matrices of the interaction networks. We specifically want to maintain information about interaction strengths in these networks and so the networks will be weighted. Interpreting them directly as adjacent matrices brings up a number of issues, particularly in the context of community detection. The first is that the matrices are signed. Although community analysis can be carried out on signed networks [45], it is not clear that the sign here is relevant for defining communities. The second issue is the diagonal elements. Again, whilst self loops can in principle be included in community analysis, the significance of them here is unclear. Thus,

TABLE 1 *Firms included in our analysis ordered by their mean market value in the 1000-day period from 4 October 2014 to 29 March 2018. Also listed for each firm is its ticker symbol and its industry group.*

#	Firm	Ticker	Industry group	Mean market value [$M]
1	Apple	AAPL	Tech hardware	675216
2	Alphabet	GOOGL	Media	519446
3	Microsoft	MSFT	Software Svcs	453842
4	Berkshire Hathaway	BRK.B	Insurance	377161
5	Exxon Mobil	XOM	Energy	361658
6	Amazon	AMZN	CD Retail	325291
7	Johnson & Johnson	JNJ	Bio Pharma	312114
8	Wells Fargo	WFC	Banks	269755
9	JP Morgan	JPM	Banks	265640
10	General electrics	GE	Capital goods	251764
11	Walmart	WMT	CS Retail	237163
12	P&G	PG	CS Products	222644
13	AT&T	T	Telecoms	215103
14	Chevron	CVX	Energy	202676
15	Verizon	VZ	Telecoms	200803
16	Bank of America	BAC	Banks	197126
17	Home Depot	HD	CD Retail	160504
18	Citigroup	C	Banks	159949
19	Comcast	CMCSA	Media	158147
20	IBM	IBM	Software Svcs	154334
21	UnitedHealth Group	UNH	HC Svcs	134340
22	Boeing	BA	Capital goods	108381
23	CVS Caremark	CVS	HC Svcs	95810
24	Walgreens Boots Alliance	WBA	CS Retail	82163
25	Costco	COST	CS Retail	66937
26	Lowe's	LOW	CD Retail	64926
27	Ford Motors	F	Auto	53559
28	General motors	GM	Auto	53379
29	Phillips 66	PSX	Energy	43770
30	Target	TGT	CD Retail	41185
31	Anthen	ANTM	HC Svcs	40344
32	McKesson	MCK	HC Svcs	40181
33	Valero Energy	VLO	Energy	30120
34	Kroger	KR	CS Retail	29837
35	Marathon Petroleum	MPC	Energy	25858
36	Cardinal Health	CAH	HC Svcs	25099
37	AmerisourceBergen	ABC	HC Svcs	19339
38	Fannie Mae+Freddy Mac	FNMA+FMCC	Banks	4583

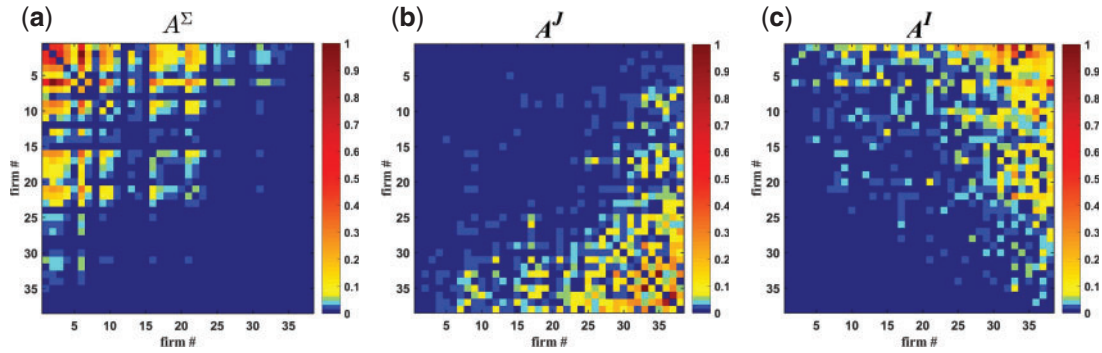

FIG. 1. Plots of the adjacency matrices A^{Σ}, A^{J} and A^{I} for $T = 1000$ with firms ordered by mean market value as in Table 1. Each matrix has been normalized to have a maximum element of one.

in defining adjacency matrices, we take the absolute value of the corresponding interaction matrix and drop the diagonal elements, and thus the adjacency matrices we consider are

$$A^{\Sigma}_{ij} = |\Sigma^{\text{tot}}_{ij}| \left(1 - \delta_{ij}\right); \qquad A^{J}_{ij} = |J^{\text{tot}}_{ij}| \left(1 - \delta_{ij}\right); \qquad A^{I}_{ij} = |I^{\text{tot}}_{ij}| \left(1 - \delta_{ij}\right). \tag{12}$$

A plot of the matrices A^{Σ}, A^{J} and A^{I} is given in Fig. 1. With firms ranked by mean market value, this representation shows that both A^{Σ} and A^{J} are quite nested [46], albeit with the direction of the nesting occurring in opposite directions. The matrix A^{I} also has a concentration of large matrix elements, but this occurs in the upper right quadrant.

The adjacency matrix A^{I} presents an additional problem for community detection as it is non-symmmetric and thus generates a directed network. A number of approaches have been proposed to address community detection in directed networks [47], with the simplest being to consider standard (undirected) community detection algorithms applied to an appropriately symmetrized version of the original matrix [48]. The most obvious symmetrization is to take the arithmetic mean of the adjacency matrix and its transpose: $\frac{1}{2}\left[A^{I} + (A^{I})^{T}\right]$. However, in this work we employ that the geometric mean symmetrization $A^{I}_{\text{sym}} = \sqrt{I \odot I^{T}}$, where $\odot$ denotes the Hadamard product, or in terms of matrix elements

$$\left[A^{I}_{\text{sym}}\right]_{ij} = \sqrt{I_{ij} I_{ji}} = |J^{\text{tot}}_{ij}| / \sqrt{(J^{\text{tot}}_{ii} J^{\text{tot}}_{jj})}. \tag{13}$$

The significant difference between the two symmetrizations is that, in the arithmetic case, significant values can be obtained when only one of I_{ij} or I_{ji} is large, whereas in the geometric case, this requires both I_{ij} and I_{ji} to be significant. Thus, two-way links are picked out as stronger that those in a single direction. In the following, we will only give results for the geometric symmetrization A^{I}_{sym} and discuss briefly those from the arithmetic symmetrization at the end.

4.1 *Thresholding*

From Fig. 1, it is clear that these adjacency matrices contain a large proportion of very small values, and it might be wondered whether these represent genuine interactions or are simply a product of data imperfections or some random process not significant to the properties under consideration. To investigate how this potential "noise" effects our results, we introduce a threshold τ below which values in the

adjacency matrix are set to zero. Specifically, for a weighted adjacency matrix with elements $A_{ij} \geq 0$ of which $A_{ij}^{\max}$ is the maximum value, we define a thresholded version as

$$A_{ij}(\tau) = A_{ij}\,\theta[A_{ij} - \tau A_{ij}^{\max}], \tag{14}$$

where $0 \leq \tau \leq 1$ is the (fractional) threshold and $\theta[x]$ is the unit-step function. Increasing the threshold causes the initially fully connected networks to first become more sparse and eventually break up into disconnected pieces.

Due to the different distributions of matrix element magnitudes, comparing different matrices with the same value of the threshold is not a like comparison. We thus define the "weight removed" in the thresholding procedure as

$$w(A, \tau) = 1 - \frac{\sum_{ij} A_{ij}(\tau)}{\sum_{ij} A_{ij}}, \tag{15}$$

such that the impact of thresholding two different matrices to the same value of weight w is roughly comparable. In the following, we thus plot our results as a function of this weight.

4.2 *Null model*

To interpret community analysis results, it is important to compare with a null model [49]. The null model we consider here results from randomizing the temporal order of original market value data for each firm separately. This preserves mean and variance for the individual firms but randomizes their correlations. Adjacency matrices are constructed exactly as above, and we present results here obtained from 200 such randomizations.

5. Community structure

The modularity $Q(g, C)$ is a measure of the extent to which network g is connected along the lines of community structure C [50]. We will not repeat its definition here but note that it is applicable to both weighted and unweighted networks [51]. Modularity is bounded $|Q| \leq 1$, with a value of $Q \to 1$ indicating that g possesses exactly structure C. A value of $Q = 0$ indicates agreement no better than random, and a value $Q < 0$ indicates a tendency for the graph to clustered in a fashion 'opposite' to the way defined by C (i.e. more links between the communities of C than within them). We will look at the modularity of the network defined above in two ways. First we consider the modularity with respect to the communities C_{ind} defined by the industry groups of Table 1. This value of modularity we denote as $Q_{\text{ind}} = Q(g, C_{\text{ind}})$. Second, we consider the modularity for a particular network maximized over all community structures $Q_{\max} = Q(g, C_{\max}) = \max_C Q(g, C)$, where $C_{\max}$ is the maximizing community structure. Given the small size of the network, it is possible to find exactly the community structure that maximizes the modularity. This procedure is slow and, given that we will sweep over threshold and consider 200 random instances in the null model, to consistently use this exact optimization is impractical. We therefore consider instead the 'greedy' algorithm of Ref. [52] to obtain approximations to $Q_{\max}$ and $C_{\max}$. We ran both the greedy and full optimization algorithms [4] for a test case of matrix A^I_{sym} with $T = 100$. In this case, the 'greedy communities' were found to have a modularity $Q_{\max}$ to within 4% of

[4] All calculations were performed using the IGraph/M v0.4 package [53] on Wolfram Mathematica v12.0.1.

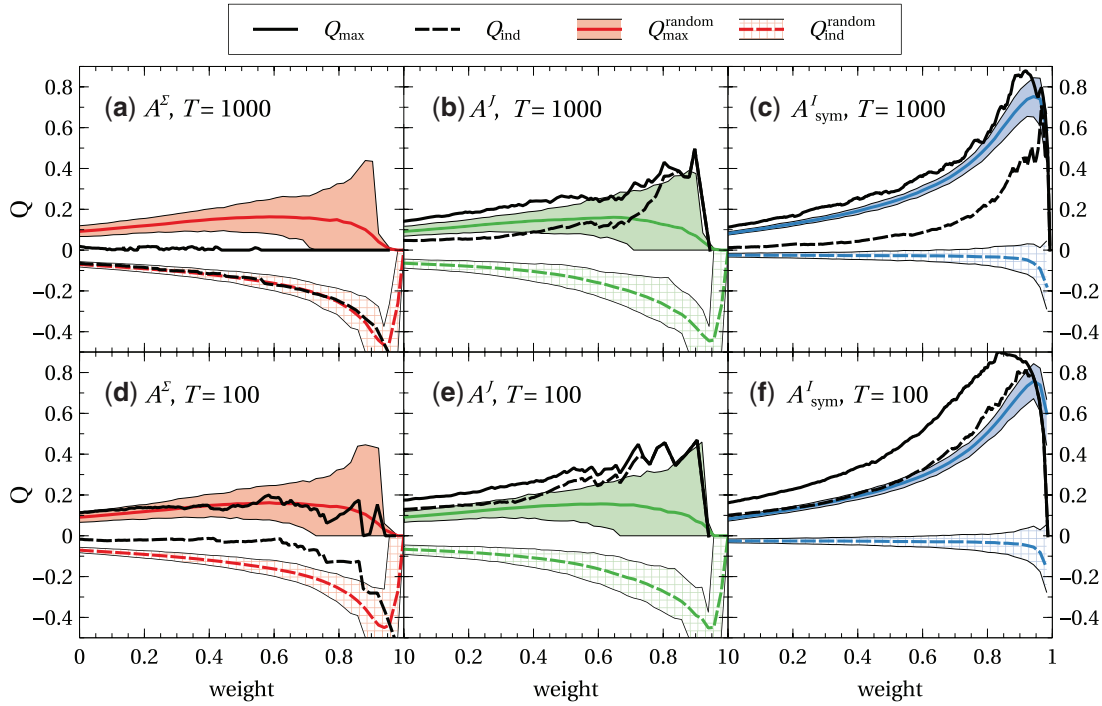

FIG. 2. Network modularities for the adjacency matrices A^{Σ} (a, d), A^{J} (b,e), and A^{I}_{sym} (c, f) as a function of weight remaining after thresholding. The results in the top row were obtained with $T = 1000$, whilst those on the bottom were obtained with $T = 100$. The continuous black lines show the optimal modularity Q_{max} and the dashed black lines show the industry-sector modularity Q_{ind}. The coloured lines show the same quantities for the randomized null model together with the 10:90 quantiles of same. The extent to which Q_{max} and Q_{ind} exceed their randomizations is indicative of the presence of well-defined modules and overlap with industrial groups, respectively.

the exact maximum across the whole τ range. For $\tau > 0.25$ ($w > 0.75$), when the networks become sufficiently sparse, the optimal communities found by both algorithms were exactly the same. From this we conclude that the greedy algorithm gives a close enough approximation to the optimal communities for our purposes here, and we use this algorithm to generate the results that follow.

Figure 2 shows the modularities Q_{max} and Q_{ind} as a function of weight removed for the three matrices A^{Σ}, A^{J} and A^{I}_{sym} after thresholding. Results are given for time-slices of $T = 1000$ (top row) and $T = 100$ (bottom) and are compared with those obtained from the null-model randomizations.[5] We begin by considering the results for network derived from the covariance matrix. With $T = 1000$, the maximum modularity Q_{max} remains close to zero across the full range of thresholding (Fig. 2a), a result which, compared with the randomizations, is anomalously low. This can be explained by noting that the full A^{Σ} matrix (at $\tau = 0$) is far more nested that is typical for the randomized ensemble and, at high connectances, there is an inverse relation between modularity and nestedness [35] because the dominant firms (i.e. those with the largest mean market value) connect strongly to most other firms in the network. Moreover, as

[5] In obtaining the mean and 10:90 quantiles, the sample used was taken from results with the same weight to within ± 0.005 of the target value.

the threshold increases, the high nestedness means that it is single nodes that become disconnected from the network (rather than more sizeable subgraphs), and these individual nodes contribute nothing to the modularity. For $T = 100$, the $Q_{\max}$ modularity of the covariance network is increased slightly with a maximum value $Q_{\max} \approx 0.2$ (Fig. 2d). This is consistent with the randomization results, and thus not indicative of any particular structure in the network. Concerning the industry-group modularity Q_{ind} in the covariance case, the $T = 1000$ results match with randomizations, whereas for $T = 100$, Q_{ind} is slightly higher than expected from the randomizations, but still negative. Overall, then, the covariance adjacency matrix (both values of T) shows little trace of any particular community structure, and certainly none relating to the industry groups.

Turning now to the results for the A^J matrix, we first comment that the randomizations have similar modularity properties to those for the A^Σ matrix. The properties of the actual A^J matrices are significantly different, however. For both $T = 100$ and $T = 1000$, the maximum modularity $Q_{\max}$ lies at or above the upper end of expectations from randomizations, and this is most pronounced in the $T = 100$ case (panel 2e). The industry-group modularity Q_{ind} is now positive across the range of threshold (compared with the negative values from the randomizations) and for $T = 100$ lies around the upper end of the randomization range for the maximum $Q_{\max}$.

This trend towards increasing modularity (both $Q_{\max}$ and Q_{ind}) is continued by the symmetric-interaction matrix A^I_{sym}. The strong increase in $Q_{\max}$ as a function of the threshold is not in itself especially significant, as this is also demonstrated by the randomizations. However, for $T = 1000$, the actual $Q_{\max}$ lies slightly above expectation from randomizations (panel 2c), and for $T = 100$, it lies significantly above the random results (panel 2f). As for A^J, the industry group modularity $Q_{\max}$ is positive across the range, and in the $T = 100$ case, it is large and lies around or above the upper end of expectation for the optimal modularity $Q_{\max}$ found from the randomization. These results mean that the matrices A^J and A^I both show a more modular structure that would otherwise be anticipated for the null model. Moreover, Q_{ind} being high for both these networks, especially in the $T = 100$ case, hints that the industry-group communities, whilst not the optimal partitioning of these networks, are playing a significant role in structuring them.

We can investigate this latter point quantitatively by comparing the community structures C_{ind} and $C_{\max}$ using the 'adjusted Rand index' (ARI) [54]. This assumes a value of 1 when the communities are identical, and 0 when they are only as closely related as chance would suggest. The ARI can take negative values for anti-correlation. Again, we interpret these results through comparison with the randomization null model [49]. Figure 3 shows the ARI comparing $C_{\max}$ and C_{ind} as a function of the weight for the same adjacency matrices as in Fig. 2. The first general trend is that the ARI for the $T = 100$ case (bottom row) is higher and more distinct from the randomizations than is the $T = 1000$ case (top row). Moreover, for $T = 100$ results as we go from A^Σ to A^J to A^I_{sym}, the ARI values increase markedly. In particular, the results for A^I_{sym} with $T = 100$ stand out with ARI $\gtrsim 0.35$ for most of the threshold range, well in excess of the randomization, and with a peak value of 0.61 (Fig. 3f). This peak value occurs for large thresholds (large weight) where the network is heavily disconnected. At the other end of the spectrum, the ARI for the A^Σ matrix with $T = 1000$ shows that the optimal communities bear no significant relation to the industry groups (Fig. 3a).

From these results, we therefore expect the A^I_{sym} network for $T = 100$ to strongly reflect the structure of the industry groups, and this is apparent when we visualize the network. The main panel of Fig. 4 shows the network for A^I_{sym} with $T = 100$ and no thresholding applied. The nodes are colour coded according to industry group, and grouped together into the optimal communities $C_{\max}$. There are six modules in the optimal structure. The only module to match perfectly with a community in C_{ind} is the telecoms community consisting of T and VZ. Nevertheless, there a number of modules with clear connections to

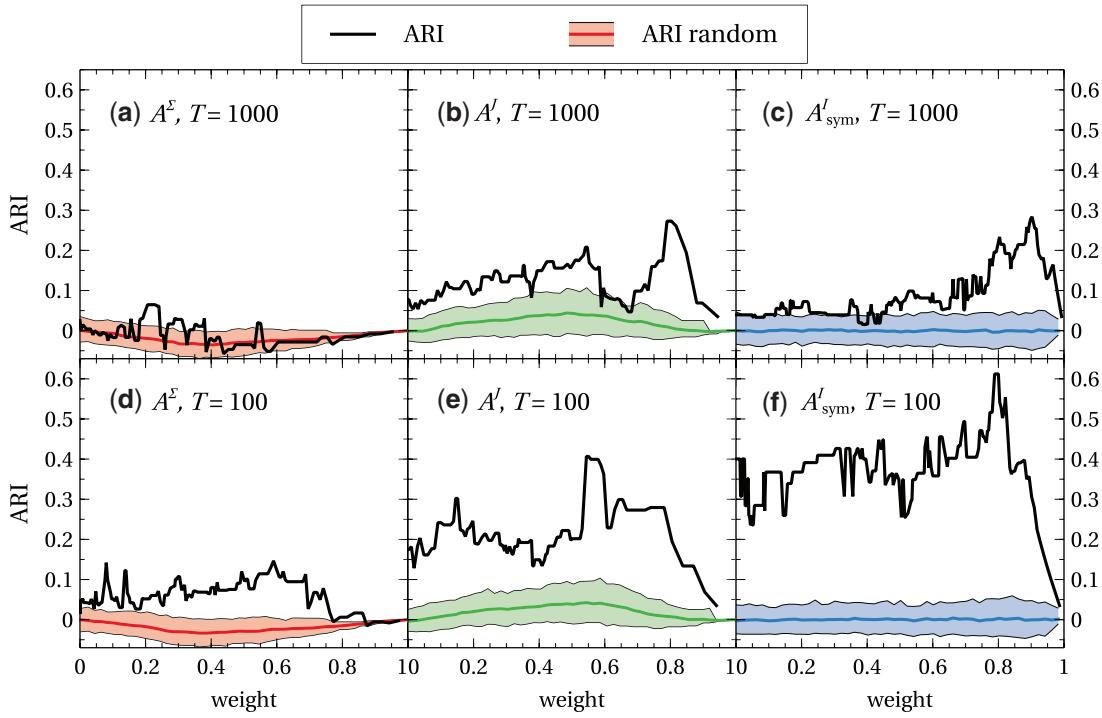

FIG. 3. Plot of the ARI measure of similarity between maximum-modularity communities $C_{\max}$ and the industry groups C_{ind} as a function of weight removed by thresholding. Layout of the panels is as in Fig. 2. Continuous black lines show the results for the adjacency matrix in question; coloured lines show results for the randomized null model including 10:90 quantiles. The degree to the ARI exceeds the randomization results indicates how the communities found by community detection reflect the externally determined industry groups.

industry groups. All five energy firms are bunched together, along with the two auto companies F and GM. All five banks are bunched together, along with the other finance-sector firm BRK.B, IBM, and the capital goods firms GE and BA. One module contains most of the healthcare firms plus the only biotech-pharms firm in the network, plus the seemingly unrelated firms LOW, HD and CMCSA. Then there are two modules which, although the firms within them belong to different industry groups, are clearly closely related. There is a 'tech community' consisting or AMZN, GOOGL, MSFT and AAPL, and one consisting of COST, KR, WMT and WBA (CS retail) together with PG (CS products) and CVS and TGT.

Figure 4b shows the network diagram for A^{I}_{sym} with $T = 100$ again but this time with a threshold (corresponding to a weight removed of 0.79 obtained with a threshold $\tau = 0.271$) chosen to give the maximum ARI value for this network in Fig. 3. With this high a threshold, the network is highly disconnected, and only strongly-connected modules persist. These include a healthcare-services module (5 firms, all HC Svcs minus CVS), an energy module (5 firms), a finance module (4 banks pls BRK.B, minus FNMA+FMCC which is perhaps an exception), a 6-firm module that groups together four CS retail firms COST-KR-WBA-WMT with CVS and TGT and tech module consisting of AMZN, GOOGL and MSFT. There are also several pairs. Within industry groups we have T-VZ, HD-LOW and F-GM. We

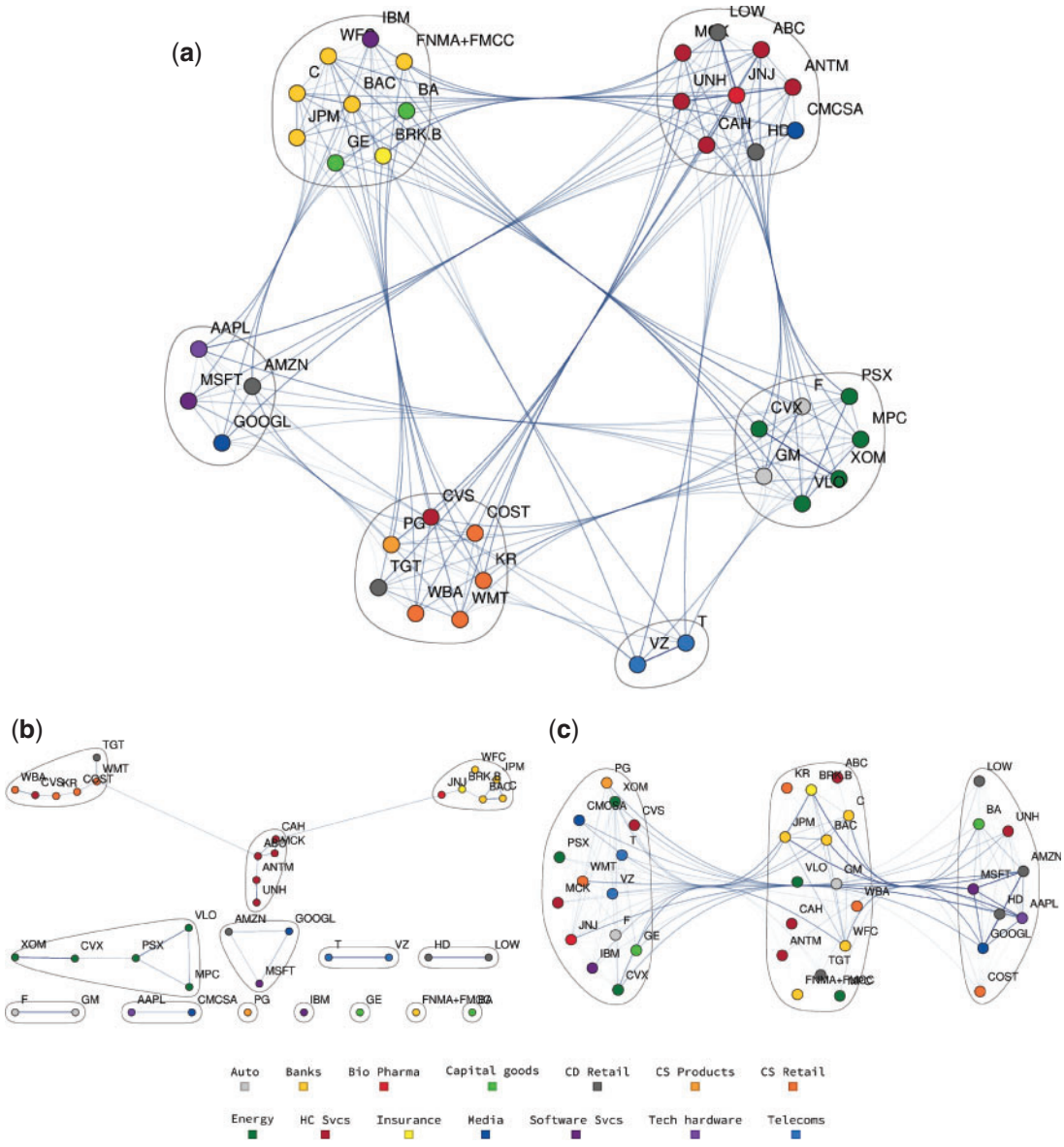

FIG. 4. Network visualizations with nodes (firms) colour-coded by industry group, and divided into modules according to the optimal community detection. The strength of line connecting nodes corresponds to the weight of the edge. The three networks shown are for the adjacency matrices: (a) A^I_{sym} matrix without thresholding; (b) A^I_{sym} matrix with threshold $\tau = 0.271$ corresponding to a removed weight of $w = 0.79$, which gives the maximum ARI value for this matrix; and (c) A^Σ without thresholding. All networks derived for $T = 100$.

also have the AAPL–CMCSA pair, a relationship which seems plausible given Apple's role as a media provider.

In strong contrast to these networks, Fig. 4c shows the network based on covariance matrix A^{Σ} with $T = 100$ and zero threshold. Here, community detection finds three large modules, each of which contains firms from various industry groups. This division into modules is not particularly robust with respect to thresholding, as when the threshold is changed, a number of firms switch modules. At high levels of thresholding, we obtain mostly disconnected single nodes. The only robust non-trivial modules that appear are the single pair CVX–XOM and one including the all financial firms except FNMA+FMCC. All of which reinforces what we found by studying the modularity and the ARI index, namely that the traces of the industry groups in the covariance network are extremely faint.

Finally, we address in more detail the impact of the time-slice length. Figure 5 shows the ARI for A^{I}_{sym} constructed using a range of values of T from 50 to 1000. Whilst there is considerable variation with threshold, it is clear from this figure that overall the ARI shows a non-linear dependence on T. In particular, the results for $T = 50$ and for $T \geq 250$ are generally lower than those obtained for $T = 100$ and $T = 200$. Although the results differ in detail, the overall character of the $T = 100$ and $T = 200$ results is similar. From this, we see that a choice around $T = 100$ to $T = 200$ gives the greatest similarity with the industry groups.

6. Discussion

In this article, we have presented a method that uses the principle of MaxEnt along with time-series data of market values to infer community matrices and networks that describe the interactions between firms in a fashion similar to how theoretical ecology pictures the interaction of species in an ecosystem.

We then considered the question of whether the networks of interacting firms so-derived exhibit a significant community structure. If we were to base our answer to this question directly on the covariance matrix Σ itself, we would conclude that the answer is no, as the corresponding adjacency matrix shows a modularity no greater than we would expect by chance and no particular trace of the GICS industry groups. A very different story emerges, however, when we look at the interaction networks derived from the MaxEnt community matrices. Here, we see modularities above what chance would suggest and, more importantly, we see a strong overlap between the optimal communities of networks and the externally-determined GICS industry groups. And, although the match between industry groups and the optimal communities of the A^{I}_{sym} network is not perfect, some of the optimal communities have a logical coherence of their own, for example, the tech stocks grouped together. For some purposes these communities might conceivably be preferable to GICS groups. A related question that this community analysis enables us to answer is which of the community matrices derived from MaxEnt is the best. If we take overlap with the GICS classification as the metric, then it is clear that, although the 'Ising' matrix J shows some positive features, it is the matrix I, with adjacency matrix symmetrized as discussed, that shows the strongest signatures of the industry group structure. This analysis therefore clearly picks out matrix I as the best candidate for interaction matrix, at least in the present context.

Our second main observation in this regard is that we obtain far greater overlap with the industry groups when we split the time-sequence data into 100-day chunks and then average, rather than considering the 1000-day series as a whole. We take this as a sign that the interactions between firms change over the course of the 1000 days, and that building the community matrices over this complete time period tends to average out some of these interactions. Dividing the data up into 100-day time-slices and then constructing community matrices preserves more of this interaction information. This implies that the Langevin interpretation of Eq. (1) should be thought of as holding quasi-statically, with matrix J and stationary vector $\overline{\mathbf{v}}$ evolving on a scale longer than 100 days.

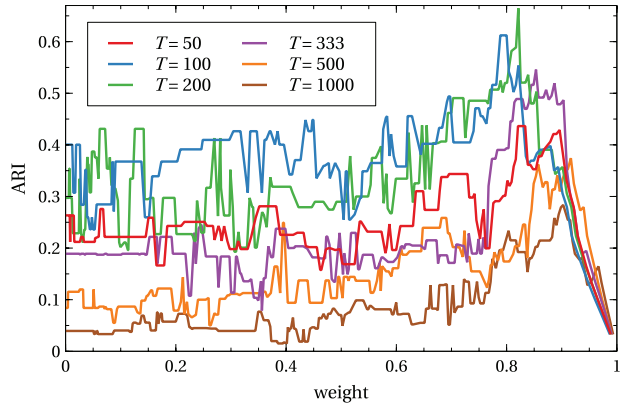

FIG. 5. ARI for the A^I_{sym} matrix as a function of removed weight obtained for different time-slice lengths $T = 50$, 100, 200, 333, 500 and 1000. Overall, the highest values are obtained for $T = 100$ and $T = 200$.

The results presented here represent a selection from a number of possibilities for constructing adjacency matrices that we considered. When set against a metric of reflecting the GICS industrial groups, these alternative approaches were found to be inferior to, or at least no better than the approaches reported above. For example, in building the adjacency matrices, we looked at taking the positive elements of the matrix rather than the absolute value, and we looked at exchanging the order of taking the mean with taking the absolute value or positive part. Concerning the matrix I^{tot} we considered constructing the individual I matrices for each time slice and then averaging. We also considered the arithmetic symmetrization $A^I + (A^I)^T$, which gave modularity results roughly comparable with those obtained from the A^J matrix, as well as the 'bibliometric symmetrization' $A^I(A^I)^T + (A^I)^T A^I$ [47], which gave no evidence of community structure. Finally, we note that we also used the classification in terms of 11 GICS industry sectors, rather than industry groups, and obtained very similar results. This can be appreciated in Fig. 4a, where firms in the same industry sector but different industry group are grouped in the same optimal community, for example, financial and health care sectors.

Inspection of the community matrices [55] here shows that for every pair of species i, j matrix elements K_{ij} and K_{ji} are always of the same sign. In ecological terms, these interactions represent therefore either competition $(--)$ or mutualism $(++)$ [56]. For the purposes here, this difference was not important, as we used the absolute value of interaction strengths as basis for the adjacency matrix weights. However, for further understanding the relationship between firms, this information would seem highly significant. The third major class of ecological interaction is that of antagonism, where diagonally opposite elements in the community matrix have opposite sign $(+-)$. In ecological terms, this might represent a predator–prey or a parasite–host interaction, and here it would indicate that one firm profits at another's expense. These interactions are absent in the community matrices here, as their appearance requires that the antisymmetric component Ω be finite, and indeed that it dominates the symmetric component for the antagonistically-interacting firms. However, due to its asymmetry, matrix I can model situations approaching either commensalism $(+0)$ or amensalism $(0-)$, and actually we found many instances in which $|I_{ij}|$ and $|I_{ji}|$ are very different in size. Providing an estimate for Ω is outside the current analysis, and will be the subject of further investigations. At any event, the fact that we obtain sensible community detection without an explicit consideration Ω implies that the effect of this component on the overall network structure is perhaps small.

Based on its ability to reproduce non-trivial community structure of interacting firms, we hypothesis that the MaxEnt network approach described here may prove to be useful in elucidating further aspects of the structure and behaviour of economic systems. As Foster [57] explains, a network formulation is crucial to understand market dynamics, driven by positive feedbacks [58], from a complex systems perspective upon the economy. The approach outlined here provides an empirical way of putting this perspective on a quantitative footing. A future application is the detailed modelling of the dynamics of the community of firms with an eye to the forecasting of future stock prices.

Acknowledgements

The authors are grateful for support from a Royal Society Challenge-led Grant (Grant number: CHL\R1\180156) in facilitating the initial stage of this collaboration.

References

1. Alderson, W. (1965) *Dynamic Marketing Behavior: A Functionalist Theory of Marketing*. Irwin, IL.
2. Lo, A. W. (2004) The adaptive markets hypothesis. *J. Portfolio Manag.*, **30**, 15–29.
3. Moore, J. F. (1993) *The Death of Competition: Leadership and Strategy in the Age of Business Ecosystems*. New York: Harper Paperbacks.
4. Moore, J. F. (1993) Predators and prey: a new ecology of competition. *Harvard Bus. Rev.*, **71**, 75–86.
5. Niederhoffer, V. (1998) *The Education of a Speculator*. New York: Wiley.
6. Carroll, G. R. & Delacroix, J. (1982) Organizational mortality in the newspaper industries of Argentina and Ireland: an ecological approach. *Admin. Sci. Q.*, **27**, 169–198.
7. Mascarenhas, B. & Sambharya, R. B. (1996) The pattern of density dependence in two global industries. *Manag. Int. Rev.*, **36**, 331–354.
8. Javalgi, R., Cutler, B. & Todd, P. (2004) An application of an ecological model to explain the growth of strategies of internet firms: the cases of eBay and Amazon. *Eur. Manag. J.*, **22**, 464–470.
9. Lakka, S., Michalakelis, C., Varoutas, D. & Martakos, D. (2013) Competitive dynamics in the operating systems market: modeling and policy implications. *Technol. Forecast. Soc. Change*, **80**, 88–105.
10. Chapin III, F. S., Matson, P. A. & Vitousek, P. (2011) *Principles of Terrestrial Ecosystem Ecology*, 2nd edn. New York: Springer.
11. Harrison, S. & Cornell, H. (2008) Toward a better understanding of the regional causes of local community richness. *Ecol. Lett.*, **11**, 969–979.
12. Michael Begon, Colin R. Townsend, J. L. H. (2005) *Ecology: From Individuals to Ecosystems,* 4th edn. New York: Wiley.
13. Carrara, F., Giometto, A., Seymour, M., Rinaldo, A. & Altermatt, F. (2015) Inferring species interactions in ecological communities: a comparison of methods at different levels of complexity. *Methods Ecol. Evol.*, **6**, 895–906.
14. Fort, H. (2018) On predicting species yields in multispecies communities: quantifying the accuracy of the linear Lotka–Volterra generalized model. *Ecol. Model.*, **387**, 154–162.
15. Halty, V., Valds, M., Tejera, M., Picasso, V. & Fort, H. (2017) Modeling plant interspecific interactions from experiments with perennial crop mixtures to predict optimal combinations. *Ecol. Appl.*, **27**, 2277–2289.
16. Jaynes, E. T. (1957) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630.
17. Jaynes, E. T. (1957) Information theory and statistical mechanics. II. *Phys. Rev.*, **108**, 171–190.
18. Fort, H. (2020) *Ecological Modelling and Ecophysics: Agricultural and Environmental Applications*. Bristol, UK: IOP ebooks, IOP.
19. Harte, J. (2011) *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*. Oxford: Oxford University Press.

20. Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. (2009) Inferring species interactions in tropical forests. *Proc. Natl. Acad. Sci. USA*, **106**, 13854–13859.
21. Bialek, W. & Ranganathan, R. (2007) Rediscovering the power of pairwise interactions.
22. Schneidman, E., Still, S., Berry, M. J. & Bialek, W. (2003) Network Information and Connected Correlations. *Phys. Rev. Lett.*, **91**, 238701.
23. Lezon, T. R., Banavar, J. R., Cieplak, M., Maritan, A. & Fedoroff, N. V. (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci. USA*, **103**, 19033–19038.
24. Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M. & Walczak, A. M. (2012) Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. USA*, **109**, 4786–4791.
25. Fortune (2019) https://fortune.com/fortune500/2019/.
26. Novak, M., Yeakel, J. D., Noble, A. E., Doak, D. F., Emmerson, M., Estes, J. A., Jacob, U., Tinker, M. T. & Wootton, J. T. (2016) Characterizing species interactions to understand press perturbations: what is the community matrix? *Annu. Rev. Ecol. Evol. Syst.*, **47**, 409–432.
27. May, R. M. (1972) Will a large complex system be stable? *Nature*, **238**, 413.
28. May, R. M. (2001) *Stability and Complexity in Model Ecosystems*. Princeton University Press.
29. Allesina, S. & Tang, S. (2012) Stability criteria for complex ecosystems. *Nature*, **483**, 205.
30. Allesina, S. & Tang, S. (2015) The stability–complexity relationship at age 40: a random matrix perspective. *Popul. Ecol.*, **57**, 63–75.
31. Landi, P., Minoarivelo, H. O., Brännström, O., Hui, C. & Dieckmann, U. (2018) Complexity and stability of ecological networks: a review of the theory. *Popul. Ecol.*, **60**, 319–345.
32. Stone, L. (2018) The feasibility and stability of large complex biological networks: a random matrix approach. *Sci. Rep.*, **8**, 8246.
33. Biggs, N. (1994) *Algebraic Graph Theory*, 2nd edn. Cambridge University Press.
34. Newman, M. (2018) *Networks*, 2nd edn. Oxford: OUP.
35. Fortuna, M. A., Stouffer, D. B., Olesen, J. M., Jordano, P., Mouillot, D., Krasnov, B. R., Poulin, R. & Bascompte, J. (2010) Nestedness versus modularity in ecological networks: two sides of the same coin? *J. Anim. Ecol*, **79**, 811–817.
36. Olesen, J. M., Bascompte, J., Dupont, Y. L. & Jordano, P. (2007) The modularity of pollination networks. *Proc. Natl. Acad. Sci. USA*, **104**, 19891–19896.
37. Grilli, J., Rogers, T. & Allesina, S. (2016) Modularity and stability in ecological communities. *Nat. Communications*, **7**, 12031.
38. Stouffer, D. B. & Bascompte, J. (2011) Compartmentalization increases food-web persistence. *Proc. Natl. Acad. Sci. USA*, **108**, 3648–3652.
39. MSCI. https://www.msci.com/gics (accessed on 31 July 2020).
40. Zwanzig, R. (2001) *Nonequilibrium Statistical Mechanics*. OUP.
41. Nisbet, R. M. & Gurney, W. S. C. (1982) *Modelling Fluctuating Populations*. New York: Wiley.
42. Fortune (2018) https://fortune.com/fortune500/2018/.
43. Bloomberg. https://www.bloomberg.com (accessed on 31 July 2020).
44. NYSE. https://www.nyse.com/market-cap (accessed on 10 July 2020).
45. Gómez, S., Jensen, P. & Arenas, A. (2009) Analysis of community structure in networks of correlated data. *Phys. Rev. E*, **80**, 016114.
46. Mariani, M. S., Ren, Z.-M., Bascompte, J. & Tessone, C. J. (2019) Nestedness in complex networks: observation, emergence, and implications. *Phys. Rep.*, **813**, 1–90.
47. Malliaros, F. D. & Vazirgiannis, M. (2013) Clustering and community detection in directed networks: a survey. *Phys. Rep.*, **533**, 95–142.
48. Satuluri, V. & Parthasarathy, S. (2011) Symmetrizations for Clustering Directed Graphs. *Proceedings of the 14th International Conference on Extending Database Technology*, EDBT/ICDT 11. New York, NY, USA: Association for Computing Machinery, p. 343–354.
49. Gates, A. J. & Ahn, Y.-Y. (2017) The impact of random models on clustering similarity. *bioRxiv*.

50. NEWMAN, M. E. J. & GIRVAN, M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 026113.
51. NEWMAN, M. E. J. (2004) Analysis of weighted networks. *Phys. Rev. E*, **70**, 056131.
52. CLAUSET, A., NEWMAN, M. E. J. & MOORE, C. (2004) Finding community structure in very large networks. *Phys. Rev. E*, **70**, 066111.
53. HORVÁT, S. (2020) IGraph/M.
54. HUBERT, L. & ARABIE, P. (1985) Comparing partitions. *J. Class.*, **2**, 193–218.
55. MAY, R. M. (1973) Qualitative stability in model ecosystems. *Ecology*, **54**, 638.
56. INGS, T. C., MONTOYA, J. M., BASCOMPTE, J., BLÜTHGEN, N., BROWN, L., DORMANN, C. F., EDWARDS, F., FIGUEROA, D., JACOB, U., JONES, J. I., LAURIDSEN, R. B., LEDGER, M. E., LEWIS, H. M., OLESEN, J. M., VAN VEEN, F. F., WARREN, P. H. & WOODWARD, G. (2009) Review: Ecological networks beyond food webs. *J. Anim. Ecol.*, **78** 253–269.
57. FOSTER, J. (2005) From simplistic to complex systems in economics. *Cambridge J. Econ.*, **29**, 873–892.
58. SHAPIRO, C. & VARIAN, H. (1999) *Information Rules: A Strategic Guide to the Network Economy*. Boston, MA: Harvard Business School Press.