Complex networks reconstruction using renormalization theory

David Dobáš

FNSPE CTU, Prague

3.9.2024

Outline

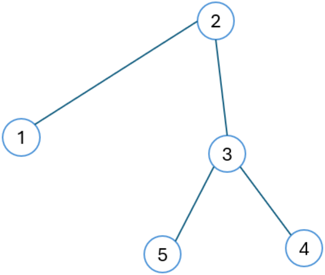
1. Graph theory essentials
   - Basic definitions
   - Graph properties
   - Complex networks
2. Random graph models
   - Exponential random graphs
   - Scale-invariant model
3. Network reconstruction
4. NR using SIM
   - Using the original SIM
   - Out-degree corrected SIM
   - Degree corrected SIM

Acknowledgements

This work is a result of cooperation with prof. Diego Garlaschelli and his PhD candidate Jingjing Wang during an exchange programme at the Leiden University in Netherlands.

Graph theory essentials

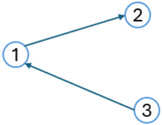
A **graph** (network) $\mathcal{G}$ is a set of **vertices** (nodes) V and a set of pairs of vertices E called **edges** (links).



$$V = \{1,2,3,4,5\}$$
$$E = \{\{1,2\},\{2,3\},\{3,4\},\{3,5\}\}$$

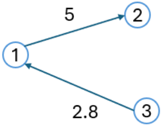
Graph theory essentials

Graph can be directed or undirected



$$E = \{(1,2),(3,1)\}$$

Graphs can be weighted, i.e. exists $W : E \to \mathbb{R}$



$$E = \{(1,2),(3,1)\}$$
$$W((1,2)) = 5$$
$$W((3,1)) = 2.8$$

Graph theory essentials

Graphs can be represented using adjacency matrix $\mathbb{A}$, whose entries are

$$a_{ij} = \begin{cases} 1 \text{ if there is an edge } i \rightarrow j \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

Weighted graphs can be represented by weighted adjacency matrix $\mathbb{W}$, such that

$$w_{ij} = \begin{cases} \text{weight of an edge } i \rightarrow j \text{ if the edge is present} \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

Graph properties

Degrees:

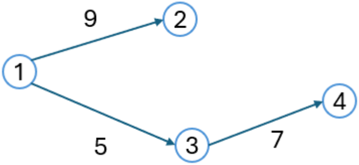
- Undirected: $k_i = \sum_{i=1}^{N} a_{ij}$
- Directed:

$$k_i^{in} = \sum_{i=1}^{N} a_{ji}$$
$$k_i^{out} = \sum_{i=1}^{N} a_{ij}$$

Strengths:

- Undirected: $s_i = \sum_{i=1}^{N} w_{ij}$
- Directed:

$$s_i^{in} = \sum_{i=1}^{N} w_{ji}$$
$$s_i^{out} = \sum_{i=1}^{N} w_{ij}$$



$$k_1^{out} = 2 \quad k_1^{in} = 0$$
$$k_3^{out} = 1 \quad k_3^{in} = 1$$

$$s_1^{out} = 14 \quad s_1^{in} = 0$$
$$s_3^{out} = 7 \quad s_3^{in} = 5$$

Graph properties

Average nearest neighbor degree (ANND):

- Undirected:
  $k_i^{nn} = \frac{\sum_{j \neq i} a_{ij} k_j}{k_i}$
- Directed:
  $k_i^{nn,out} = \frac{\sum_{j \neq i} a_{ij} k_j^{out}}{k_i^{out}}$
  $k_i^{nn,in} = \frac{\sum_{j \neq i} a_{ij} k_j^{in}}{k_i^{in}}$

Local clustering coefficient (undirected)

$$c_i = \frac{\sum_{j \neq i} \sum_{k \neq i,j} a_{ij} a_{jk} a_{ki}}{\sum_{j \neq i} \sum_{k \neq i,j} a_{ij} a_{ki}}$$

Complex networks

Complex networks

- Do not have simple or regular structure - their topology is complex
- Are observed in many real-world situations:
  - Citations network
  - Protein-protein interactions
  - World-wide web
  - World trade network
  - Many others...

Complex networks theory tries to find models, which can replicate properties of real-world networks.

Random graph models

In general, random graph models are such models, which to each graph G from given set $\mathcal{G}$ assign a probability $P(G)$,
s.t. $\sum_{G\in\mathcal{G}} P(G) = 1$

We may call these methods *ensemble models*, instead of generating single graph, we have a whole ensemble of graphs.

Exponential random graphs

- Approach analogical to statistical physics
- Prescribes to find such a probability, which maximizes the Shannon entropy

$$S = -\sum_{G \in \mathcal{G}} P(G) \ln P(G)$$

- Canonical ensemble is obtained by imposing soft constraints

$$c_i^* = \sum_{G \in \mathcal{G}} c_i(G) P(G)$$

Exponential random graphs

- Constrained maximization leads to

$$P(G|\vec{\theta}) = \frac{e^{-H(G,\vec{\theta})}}{Z(\vec{\theta})}$$

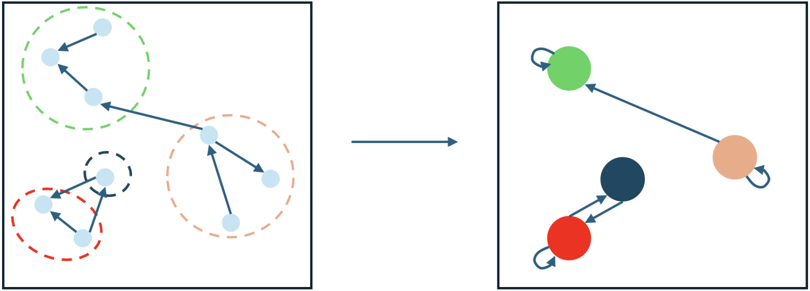
  where $H(G,\vec{\theta}) = \sum_{i=1}^{k} c_i(G)\cdot\theta_i$ is the graph Hamiltonian and $Z(\vec{\theta}) = \sum_{G\in\mathcal{G}} e^{-H(G,\vec{\theta})}$ is the partition function.

- For example, one can use degrees as constraints

$$H(G,\vec{\theta}) = \sum_{i\neq j} \theta_i k_i(G) \tag{3}$$

Scale-invariant model

The goal is to find a model, which is consistent across multiple scales



Scale-invariant model

- Assumes the following parameters
  - each node i has the so-called fitness x_i determining its tendency to form connections
  - z for overall density
- Also assumes, that upon coarse-graining, the fitnesses are additive, i.e. for a block-node I, its fitness $x_I = \sum_{i \in I} x_i$
- Lastly, it assumes that edges are independent random variables
- Having these assumptions, the aim is to find such a model, whose graph probability has the same form for all possible coarse-grainings and where only parameters change using some renormalization rule.

Scale-invariant model

It was shown[1] that these assumptions lead to a link probability

$$p_{ij} = 1 - \exp(-z x_i x_j)$$

For directed case, one can assume existence of outward fitnesses x_i and inward fitnesses y_i. Then one obtains a model

$$p_{ij} = 1 - \exp(-z x_i y_j)$$

The links are independent, therefore the random graph model is fully defined (probability of graph is product of link probabilities)

[1] Elena Garuccio, Margherita Lalli, and Diego Garlaschelli. "Multiscale network renormalization: Scale-invariance without geometry". In: *Physical Review Research* 5.4 (Oct. 2023), p. 043101. ISSN: 2643-1564.

Network reconstruction

- The goal is to plausibly reconstruct a given empirical network using only partial information about it
- For example, we want to find a reconstruction using only the knowledge of node out-strengths, in-strengths and number of links
- In such case, we want to find $O(N^2)$ entries of weighted adjacency matrix, using only the knowledge of $O(N)$ strengths.
- We may try to find one reconstructed network, however, the more robust approach is to generate an ensemble and then evaluating properties over the whole ensemble
- Example: interbank network

Interbank network

- Interbank network is a network of loans between banks
- It is a weighted network, where each link corresponds to a loan and its weight is the amount of the loan
- Banks do not disclose their individual loans, they only announce their total assets and liabilities, i.e. for bank i, we know A_i and L_i
- In terms of network properties, this corresponds to strengths $A_i = s^{in}_{0,i}$, $L_i = s^{out}_{0,i}$
- To study dynamical processes like stress-propagation, we need to know the full weighted adjacency matrix, therefore a reconstruction method is needed

Max-Ent algorithm

Prescribes to maximize

$$S = -\sum_{i,j=1}^{N} w_{ij} \ln w_{ij}$$

Leads to

$$w_{ij}^{ME} = \frac{s_{0,i}^{out} s_{0,j}^{in}}{\hat{W}} \qquad \forall i,j \tag{4}$$

where $\hat{W} = \sum_{i=1}^{N} s_{0,i}^{out} = \sum_{i=1}^{N} s_{0,i}^{in}$. However, if all strengths are nonzero, then all entries of the weighted adjacency matrix are nonzero - fully connected network

Two step algorithms

We can divide the reconstruction in two steps

1. First, we use some method to find the the topology of the reconstructed network (i.e. we find the adjacency matrix)
2. Only after that we assign weights

Several successful methods like that were proposed.[2]

[2] Tiziano Squartini et al. "Reconstruction methods for networks: The case of economic and financial systems". In: *Physics Reports* 757 (Oct. 2018), pp. 1–47. ISSN: 0370-1573.

Network reconstruction using the Scale-invariant model

Motivation: the Scale-invariant model might work better than other methods in situations, where different scales are present simultaneously.

For example, we know detailed information about banks in Czech republic, but only aggregate information about all banks in France, i.e. banks of France form only one node in our network.

Network reconstruction using the Scale-invariant model

We assume the knowledge of node out-strengths $x_i \equiv s_{0,i}^{out}$, in-strengths $y_i \equiv s_{0,i}^{in}$ and the total number of links L_0

We propose to use the following two step model

1. Sample edges using the connection probability given by the Scale-invariant model, i.e.

$$p_{ij}^{SIM} = 1 - \exp(-z x_i y_j)$$

2. Assign weights using the corrected gravity model

$$w_{ij} = \begin{cases} 0 & \text{if } a_{ij} = 0 \\ \frac{x_i y_j}{W p_{ij}^{SIM}} & \text{if } a_{ij} = 1 \end{cases} \tag{5}$$

where $W \equiv \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$

Network reconstruction using the Scale-invariant model

Before sampling edges, we need to find the value of z. We do that by fitting the expected number of links over an ensemble to the empirical value

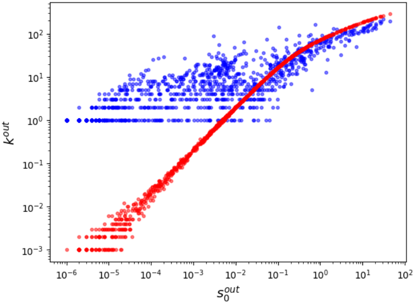
$$\langle L\rangle = \mathbb{E}(\sum_{i,j=1}^{N} a_{ij}) = \sum_{i,j=1}^{N} p_{ij} = \sum_{i,j=1}^{N} (1 - \exp(-z x_i y_j)) \overset{!}{=} L_0$$

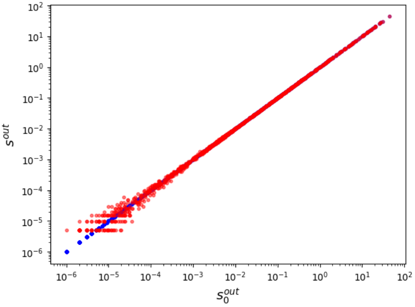
Network reconstruction using the Scale-invariant model

To evaluate our method, we might want to try reconstructing some real-world network. The primary aim is the interbank network, however, there is no dataset publicly available.

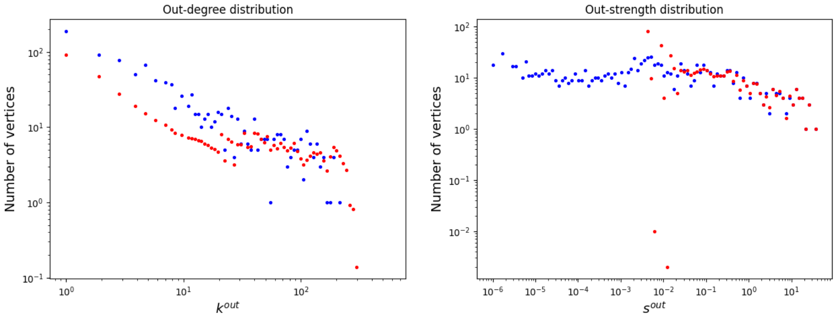
In this research project, we used the network of flights between 1021 US cities in the year 2015. Weights are numbers of passengers. The total number of links in our network is 19199.

Results

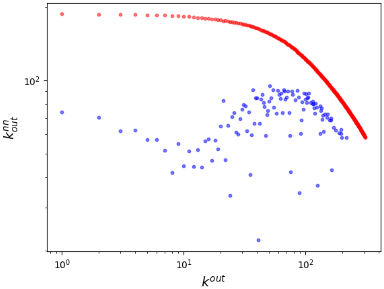

k^{out}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
s_0^{out}


s^{out}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
s_0^{out}

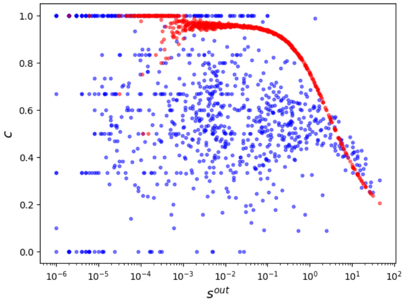
Results

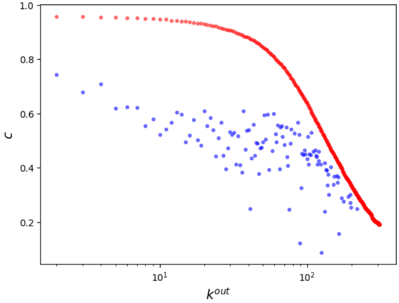

Out-degree distribution
Number of vertices
10^{2}
10^{1}
10^{0}
10^{-1}
10^{0}
10^{1}
10^{2}
k^{out}
Out-strength distribution
Number of vertices
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
s^{out}

Results



Results


c
1.0
0.8
0.6
0.4
0.2
0.0
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
s^{out}


c
1.0
0.8
0.6
0.4
0.2
10^{1}
10^{2}
k^{out}

Correction of degrees

We encountered a significant problem: too many isolated nodes

- Out of 1021 nodes, there were on average 502.745 out-isolated nodes and 503.166 in-isolated nodes in the reconstructed ensemble
- However, all nodes in the empirical network have nonzero strength, therefore are not isolated
- This can lead to poor results when evaluating stress-propagation on the network

Out-degree corrected SIM

Let us first demand that all out-degrees must be nonzero. How to modify the original model in the most "unbiased" way?
In such case, our space of allowed graphs is smaller, we denote it Γ. On this space, we can define

$$\begin{aligned} P_{oDSIM}\left(\{a_{ij}\}_{i,j=1}^{N}\right) &= P_{SIM}\left(\{a_{ij}\}_{i,j=1}^{N} \,\middle|\, k_i^{out} > 0\,\forall i\right) \\ &= \frac{P_{SIM}\left(\{a_{ij}\}_{i,j=1}^{N}\right)}{P_{SIM}\left(k_i^{out} > 0\,\forall i\right)} \qquad \forall\{a_{ij}\}_{i,j=1}^{N} \in \Gamma \end{aligned}$$

*Analytical derivation

Since rows of the adjacency matrix remain independent, the denominator can be analytically computed

$$P_{SIM}\left(k_i^{out} > 0\,\forall i\right) = \prod_i P_{SIM}\left(k_i^{out} > 0\right)$$

and

$$\begin{aligned} P_{SIM}(k_i^{out} > 0) &= 1 - P(k_i^{out} = 0) = 1 - \prod_j (1 - p_{ij}) = \\ &= 1 - \prod_j \exp(-z x_i y_j) = 1 - \exp(-z x_i \sum_j y_j) = \\ &= 1 - \exp(-z x_i W) \end{aligned}$$

Results

Although the number of out-isolated nodes decreased to zero, the number of in-isolated nodes stayed basically untouched

Degree corrected SIM

What if we demand, that both out-degrees and in-degrees shall be nonzero?

This once again defines a restricted space of graphs Γ, on which we can define the corretced model as

$$P_{DSIM}\left(\{a_{ij}\}_{i,j=1}^{N}\right) = \frac{P_{SIM}\left(\{a_{ij}\}_{i,j=1}^{N}\right)}{P_{SIM}\left(\left(k_i^{out} > 0\,\forall i\right) \cap \left(k_i^{in} > 0\,\forall i\right)\right)} \qquad \forall \{a_{ij}\}_{i,j=1}^{N} \in \Gamma$$

Degree corrected SIM

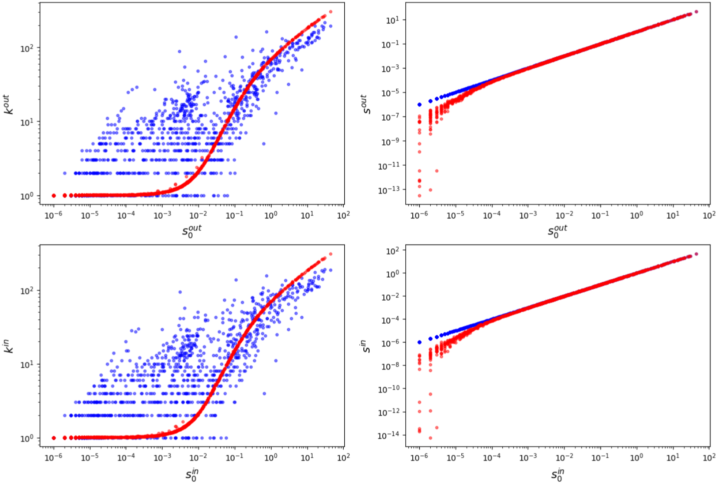
Now, we made all entries of the adjacency matrix dependent. Computation of the denominator is rather involved. And even if we compute it, we would need to sample the whole graph at once.

Instead, we propose to use the Metropolis-Hastings algorithm.

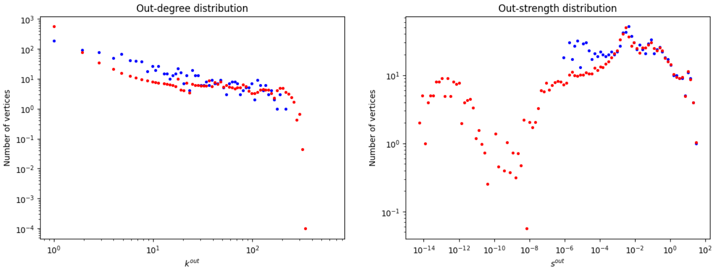
Metropolis-Hastings algorithm - key features

- Markov chain Monte Carlo algorithm for sampling from probability distributions
- Generates a sequence of samples, next sample is accepted or rejected based on ratio of probabilities
- That makes it possible to avoid the computation of the denominator
- At each step, we can flip one entry of the adjacency matrix, which makes the ratio of probabilities rather simple

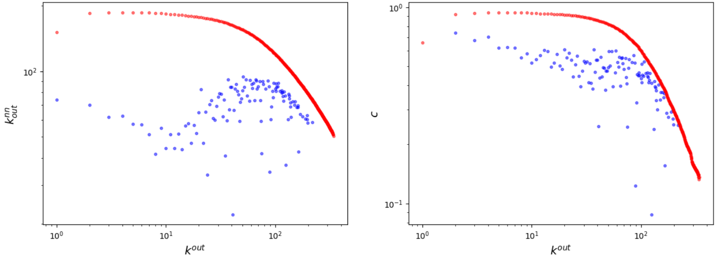
Results


k^{out}
10^2
10^1
10^0
10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^0 10^1 10^2
s_0^{out}
s^{out}
10^1
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^0 10^1 10^2
s_0^{out}
k^{in}
10^2
10^1
10^0
10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^0 10^1 10^2
s_0^{in}
s^{in}
10^2
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^0 10^1 10^2
s_0^{in}

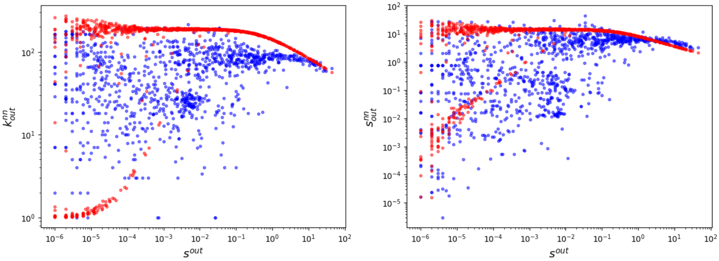
Results


Out-degree distribution
Number of vertices
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^0
10^1
10^2
k^{out}
Out-strength distribution
Number of vertices
10^1
10^0
10^{-1}
10^{-14}
10^{-12}
10^{-10}
10^{-8}
10^{-6}
10^{-4}
10^{-2}
10^0
10^2
s^{out}

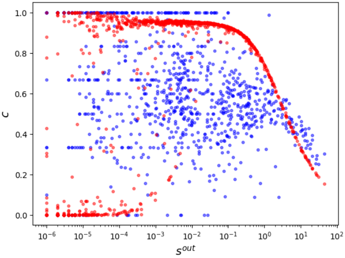
Results

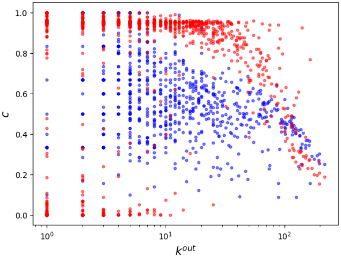


Results


k^{nn}_{out}
10^0
10^1
10^2
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
10^1
10^2
s^{out}
s^{nn}_{out}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
10^1
10^2
s^{out}

Results


1.0
0.8
0.6
0.4
0.2
0.0
c
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
s^{out}


1.0
0.8
0.6
0.4
0.2
0.0
c
10^{0}
10^{1}
10^{2}
k^{out}

Conclusion

Further work can be done

- Improving the sampling strategy
- Different models are applicable in different situations - test on the actual interbank network
- Comparison with already existing methods, showing possible advantage on multiscale data
- Study dynamical processes - stress propagation

Garuccio, Elena, Margherita Lalli, and Diego Garlaschelli. "Multiscale network renormalization: Scale-invariance without geometry". In: *Physical Review Research* 5.4 (Oct. 2023), p. 043101. ISSN: 2643-1564.

Squartini, Tiziano et al. "Reconstruction methods for networks: The case of economic and financial systems". In: *Physics Reports* 757 (Oct. 2018), pp. 1–47. ISSN: 0370-1573.