Fan *et al. Applied Network Science* (2019) 4:69
https://doi.org/10.1007/s41109-019-0171-y

Applied Network Science

RESEARCH Open Access

Topology of products similarity network for market forecasting

Jingfang Fan[1,2*], Keren Cohen[3], Louis M. Shekhtman[2], Sibo Liu[4], Jun Meng[1], Yoram Louzoun[3*] and Shlomo Havlin[2*]

*Correspondence:
j.fang.fan@gmail.com;
louzouy@math.biu.ac.il;
havlin@ophir.ph.biu.ac.il
[1]Potsdam Institute for Climate Impact Research, Potsdam 14412, Germany
[2]Department of Physics, Bar Ilan University, Ramat Gan 52900, Israel
Full list of author information is available at the end of the article

Abstract

The detection and prediction of risk in financial markets is one of the main challenges of economic forecasting, and draws much attention from the scientific community. An even more challenging task is the prediction of the future relative gain of companies. We here develop a novel combination of product text analysis, network theory and topological based machine learning to study the future performance of companies in financial markets. Our network links are based on the similarity of firms' products and constructed using the Securities Exchange Commission (SEC) filings of US listed firms. We find that several topological features of this network can serve as good precursors of risks or future gain of companies. We then apply machine learning to network attributes vectors for each node to predict successful and failing firms. The resulting accuracies are much better than current state of the art techniques. The framework presented here not only facilitates the prediction of financial markets but also provides insight and demonstrates the power of combining network theory and topology based machine learning.

Keywords: Topology, Network, Economic, Machine learning

Introduction

Network science has been used to predict many natural and technological phenomena, including among many others, the evolution of a scientist's impact (Sinatra et al. 2016), forecast disease epidemics (Eubank et al. 2004; Colizza et al. 2006; Brockmann and Helbing 2013), predict the spatial development of urban areas (Li et al. 2017) and forecast climate extreme events (Ludescher et al. 2014; Boers et al. 2014; Meng et al. 2017; Meng et al. 2018). Networks have been demonstrated to be useful tools in the study of many real world systems, such as, physics, biology, and social systems (Newman 2010; Cohen and Havlin 2010; Zhao et al. 2013). Recently, a network approach has been applied to describe the instability in financial systems (Bardoscia et al. 2017), and to study the relationship between the structure of the financial network and the likelihood of systemic failures due to contagion of risk (Miura et al. 2012; Acemoglu et al. 2015).

Over the last few years, networks have been combined with machine learning approaches to predict the classes of nodes or of full networks. The main approach to node classification was the homophily assumption that neighboring nodes have similar classes. This has been often used to project some input over the Laplacian eigenvectors, and the usage of the projection for classification (Bruna et al. 2013). This basic approach

was then used by most following studies, where either the graph itself was used (in such a case, the eigenvectors themselves are used) or an input to the graph was used. In such a case a convolution with these eigenvectors was used (Masci et al. 2015; Monti et al. 2017). A Multi-Dimensional-Scaling (MDS) projection of the points in the graphs was also used for a similar goal (Belkin and Niyogi 2002; Levy et al. 2015). Other successful approaches using the same relation between neighborhood and class similarity, include DeepWalk (Perozzi et al. 2014), where each node has an ID. A truncated random walk is performed on nodes. It then uses these ordered sers of vertices as an input to skip-gram to compute a projection of each word into R^N maximizing the order probability. Planetoid (Yang et al. 2016) also uses random walks combined with negative sampling. Duvenaud et al. used a translation of subgraphs to hash functions for a similar task in the context of molecule classifications (Duvenaud et al. 2015). Recently, Kipfs and collaborators propose a simplification to spectral based convolutions (Kipf and Welling 2016a; Schlichtkrull et al. 2018; Kipf and Welling 2016b) and instead use a two-layer convolutional neural network.

An important weakness of most of these approaches is the usage of the graph only as a similarity measure and ignoring more complex features of topology of graphs, focusing on the above-mentioned assumption that proximity in the graph implies similarity in labels. Different works have used variants of this idea, each using smoothness and graph distance differently (e.g. (Belkin and Niyogi 2004; Sindhwani et al. 2005)). This smoothness can be obtained by encouraging cross–edge label smoothness. For example, a quadratic penalty with fixed labels for seed nodes can be used (Zhou et al. 2004; Zhu et al. 2003). Multiple parallel methods using diffusion have also been proposed (Rosenfeld and Globerson 2017). Grover and Leskovec developed an algorithmic framework, node2vec, to detect continuous feature representations for nodes in network (Grover and Leskovec 2016).

However, recently a complementary approach suggested that, in contrast with images that are typically overlaid on a 2D lattice, graphs have a complex topology. This topology is highly informative of the properties of nodes and edges (Rosen and Louzoun 2015; Naaman et al. 2018; Benami et al. 2019) and can thus be used to classify their classes. We here propose that such topology–based methods can be used to predict companies future performance.

In the present study, we use text-based analysis of SEC Form 10-K product descriptions to construct the network of product similarity between firms. The firms are regarded as network nodes, and the level of similarity between the product descriptions of different firms represents the network links (strength). The 10-K product descriptions are obtained from the Securities Exchange Commission (SEC) Edgar website (https://www.sec.gov/). We then analyze the topological structure of the network, and determined how measures such as, clustering coefficient correlate with membership in the Standard & Poor's 500 (S&P 500). We also analysed the K-shell structure of the network (Carmi et al. 2007; Kitsak et al. 2010) and find that it reveals that firms in more outer shells have higher risk to collapse (or merge). Furthermore, we combine the network structure and machine learning methods to predict both the successful and collapsed firms. We find that the forecasting rates by using our combined method are significantly higher than random guessing and other methods (Breiman 2001).

Data

Product similarity data

In this study, we use text-based analysis of 10-K product descriptions to obtain the product similarity between firms, representing the links. For any two firms i and j, we have a product similarity, which is a real value in the interval [0,1] describing the similarity of words used by firms i and j in their 10-K forms. To compute the "product similarity" between two firms using the basic cosine similarity method (Hoberg and Phillips 2010; 2016), We first build the database for each year by taking the list of unique words used in all product descriptions in that year. We then take the text from each firm's product description and construct a binary N-vector summarizing its word usage. A given element of this N-vector is 1 if the given dictionary word is used in firm i's product description. For each firm i, we denote this binary N-vector as $\vec{P}_i$. We define the normalized vector $\vec{V}_i$ as,

$$\vec{V}_i = \frac{\vec{P}_i}{\sqrt{\vec{P}_i \cdot \vec{P}_i}}. \tag{1}$$

To measure the similarity of products of firms i and j, we compute the dot product of their normalized vectors, which is then the basic cosine similarity:

$$w_{i,j} = \frac{\vec{V}_i \cdot \vec{V}_j}{\mid V_i \mid\mid V_j \mid}. \tag{2}$$

In this study, we use 18 years (from 1996 to 2013) of data. For a more detailed description see Ref. (Hoberg and Phillips 2010). We think that the similarity measure employed is better than the standard approach of determining industry membership by the traditional industry classification such as SIC codes. This is since, traditional SIC codes are assigned to a specific firm and remain unchanged over time. The assignment is not accurate as firms can have multiple product segments that belong to different sectors. Firm can also adjust their product lines according to the market conditions; a pattern cannot be captured by a static traditional industry code. In contrast, the text-based method used in this study has several appealing features. First, based on their mandatory disclosure, the method measures the similarity of the products of two specific firms, while traditional industry code only provides static zero-one membership classifications (Hoberg and Phillips 2010). Built on this text-based method, we can therefore explore the properties of the product networks. Second, the method reflects a dynamic industry structure allowing us to examine the intertemporal features of the networks.

External financial data

Apart from the product similarity data, we also compile some external information to construct financial variables. We download the daily closing price of the S&P 500 index from the Center for Research in Security Prices (CRSP) (http://www.crsp.com/). This time-series can be used to measure the overall market dynamics. Several firm-specific variables are also taken to control the range of observed firm traits. Specifically, firm size is assumed to be the total market capitalization following as suggested in Ref. (Fama and French 1992); the book-to-market ratio is the book value of equity over the market value of equity and measures a firm's growth opportunities; leverage is a measure of capital structure defined by the ratio of long-term debt and other debt (current liabilities) to

the sum of long-term debt, debt in current liabilities, and stockholders' equity; profitability is defined as income before extraordinary items over lagged total assets; prior year return is the return of the stock in the past year; investment is the year-over-year percent growth in total assets following (Cooper et al. 2008); liquidity is a measure of firm liquidity found using reference (Amihud 2002); and the Altman Z-score is a measure of default risk according to reference (Altman 1968). The data can be downloaded from the Compustat database (http://www.compustat.com).

Methods

Product similarity network

For each year, we construct a weighted undirected network based on the product similarity data where each firm or company is a node and the links have a given strength, w, representing the level of similarity between pairs of nodes. The links are thresholded such that only links with a weight greater than some significant value are kept in the network. Here, we only consider the similarity values that are above 10^{-4}. This threshold is calculated based on the coarseness of the three-Digit Standard Industrial Classification (SIC). The level of coarseness thus matches that of three digit SIC codes, as both classifications result in the same number of firm pairs being deemed related (Hoberg and Phillips 2010). We present a specific product similarity network for the year 2012 in Fig. 1.

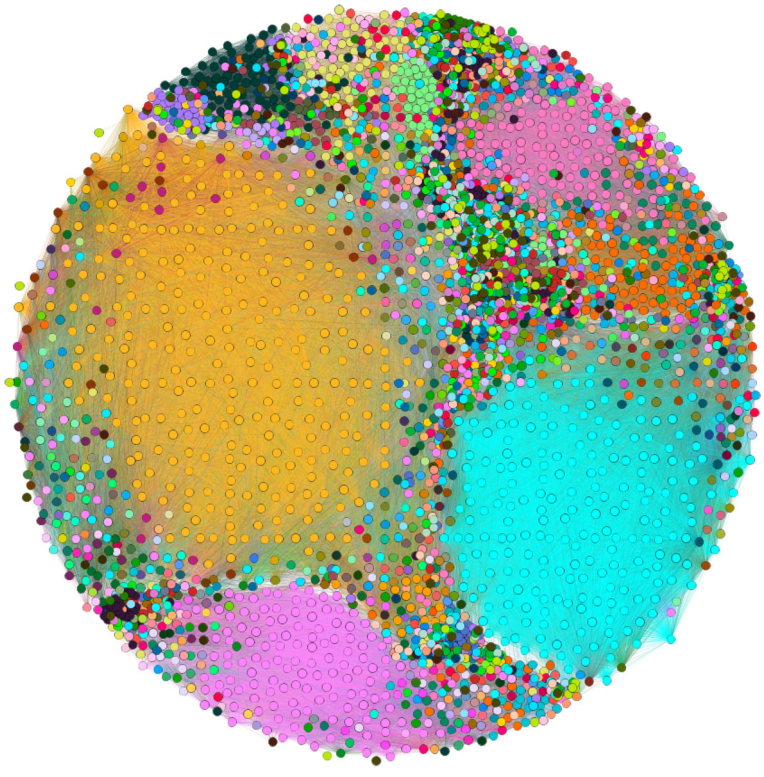

Fig. 1 A typical snapshot of the product descriptions similarity network for the year 2012. The system size is $N = 3925$, average degree is $< k >= 94.78$. Different colors represent different shells (Carmi et al. 2007)

Network topology measures for machine learning

Multiple network topological measures have been used to predict the future gain of companies, as well as the probability of their collapse. The general approach is based on methods described in Rosen et al. (2016) and Naaman et al. (2018). In brief, a vector representing a set of topological features has been computed for each node, and this vector was then used in a machine learning framework as decribed below. The following features were used for the topological network attribute vector (NAV):

- Degree (number of neighbors).
- Betweenness Centrality (Everett and Borgatti 1999). Betweenness is a centrality measure of a node (or a link). It is defined as the number of shortest paths between all vertex pairs that pass through the node.
- Closeness Centrality (Sabidussi 1966). Closeness is a centrality measure of a node (or a link) defined as the average length of the shortest path between the vertex and all other vertices in the graph.
- Distance distribution moments. We compute the distribution of distances from each node to all other nodes using a Dijkstra's algorithm (Dijkstra 1959), and then take the first and second moments of this distribution.
- Flow (Rosen and Louzoun 2014). We define the flow a node as the ratio between the directed and undirected distances between this node and all other nodes.
- Network motifs (Milo et al. 2002). Network motifs are patterns of small connected sub-graphs. We use an extension of the Itzchack algorithm (Itzhack et al. 2007) to calculate motifs. For each node, we compute the frequency of each motif in which this node participates. Note that the above algorithm is an extension of the concept to undirected graphs.
- K-core (Batagelj and Zaveršnik 2011). The K-core of a network is a maximal subgraph that contains only vertices with degree k or more. Equivalently, it is the subgraph of G formed by recursively deleting all nodes of degree less than k.

The selection of features is based on their computationally availability, as well as from experience of other studies on informative features. Basically the features used can be divided into three group: A) Local features which basically count the frequency of subgraphs (e.g., degree, small scale motif, clustering coefficient.) B) Measures of hiearchy, which basically measure the position of a node in a network if the network would have been ordered according to some hierarchical mechanism. C) Measures of centrality (e.g., K-Cores). Each network feature reflects different types of correspondence between product similarities of different firms. These network features reflect competition in production and industry demand for certain products. However, the causal relationship between our predictive ability and the micro-level descriptions of the network features remains to be determined in future work.

Machine learning

We computed for each node an NAV as described above. All nodes within a given year were then split into a training set composed of 70% of nodes and a test set composed of the remaining 30%. Each node was also associated with two binary valued tags representing whether the company collapsed in the following year and if it was within the top 5% of returns in the following year. The values of the tags were then learned using

either a Random Forest (RF) (Ho 1995) classifier and a neural network with two internal layers. The Random Forest was trained with 200 trees and a balanced loss matrix, where the error cost was inversely proportional to the training fraction of the appropriate tag (0 or 1). The trees were limited to 10 samples per leave. All other parameters were the default parameters of the Matlab Treebagger. The neural network had a rectified linear unit (ReLU) activation functions and a cross entropy loss function. The solver used was an ADAM optimizer (Kingma and Ba 2014). L2 regularization has also been applied. All other parameters were as in the default of the Keras python library: https://keras.io.

Results

We first consider the probability density function (PDF) of links' strength, $p(w)$, for each network (year). The results for six specific years are shown in Fig. 2. We find that $p(w)$ is robust and approximately follows an exponential distribution. The values of product similarity can reflect the product market synergies and competition of mergers and acquisitions (Hoberg and Phillips 2010), i.e., higher w may mean the two firms are highly competitive or in cooperative relations.

Next, we consider two basic parameters to reveal the structure of the network: the weighted degree and clustering coefficient (Watts and Strogatz 1998). The weighted degree for node i is defined as,

$$s_i = \sum_{j=1}^{N} a_{ij} w_{ij}, \tag{3}$$

where a_{ij} is the adjacency matrix and s_i quantifies the strength of node i in terms of the total weight of its connections. In the case of the product similarity networks, it reflects the importance or impact of a firm i in the network. We find, see Fig. 3, that the strength $s(k)$ of nodes with degree k increases with k as,

$$s \sim k^{\beta}. \tag{4}$$

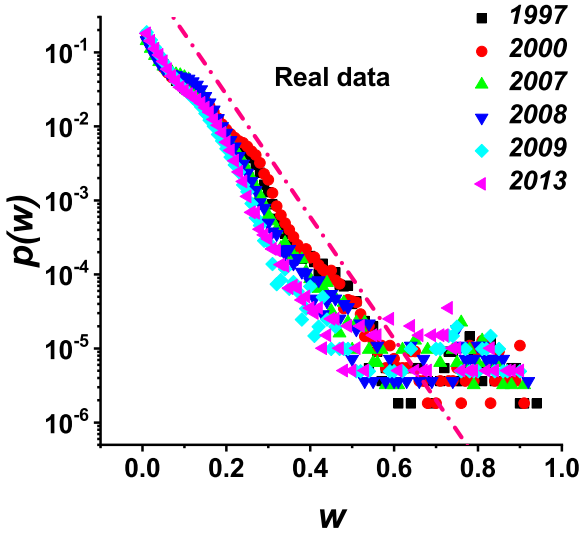

Fig. 2 The probability distribution functions of links weights for real networks of different years, the dashed line is a fitting line using an exponential distribution

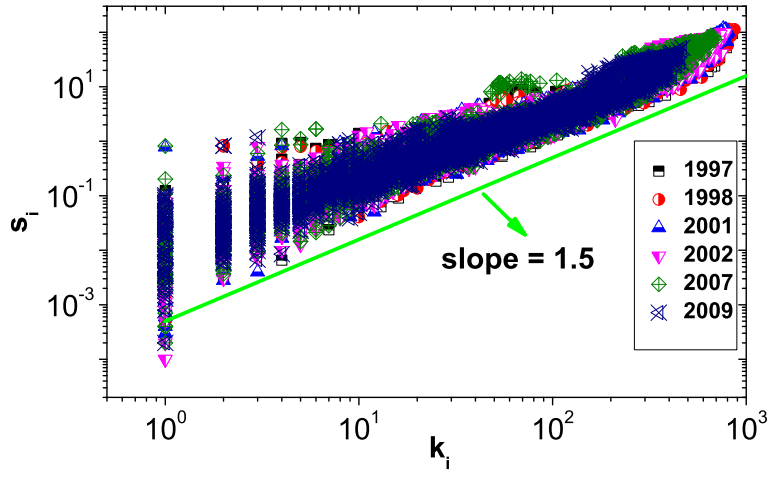

Fig. 3 Strength s as function of the degree k of nodes. The line shows that the real data follow a power law behavior with an exponent $\beta \approx 1.5$

We find that the power-law fit for the real data gives an exponent $\beta \approx 1.5$. This value implies that the strength of nodes grows faster than their degree, i.e., the weight of edges belonging to highly connected nodes tends to have a higher value. We notice that the universal power–law relationship is also observed in other real networks, e.g., the world-wide airport network, even with the same value of β (Barrat et al. 2004).

For weighted networks, the clustering coefficient for node i is defined as the geometric average of the subgraph edge weights (Saramäki et al. 2007),

$$c_i = \frac{1}{k_i(k_i-1)} \sum_{j,k} (\hat{w}_{i,j}\hat{w}_{i,k}\hat{w}_{j,k})^{1/3}, \tag{5}$$

where the edge weight is normalized by the maximum weight in the network, $\hat{w}_{i,j} = w_{i,j}/max(w)$. The average clustering coefficient is $C = \frac{1}{N}\sum_{i=1}^{N} c_i$, representing the presence of triplets in the network. Figure 4a shows the dynamical evolution of C with time (years),

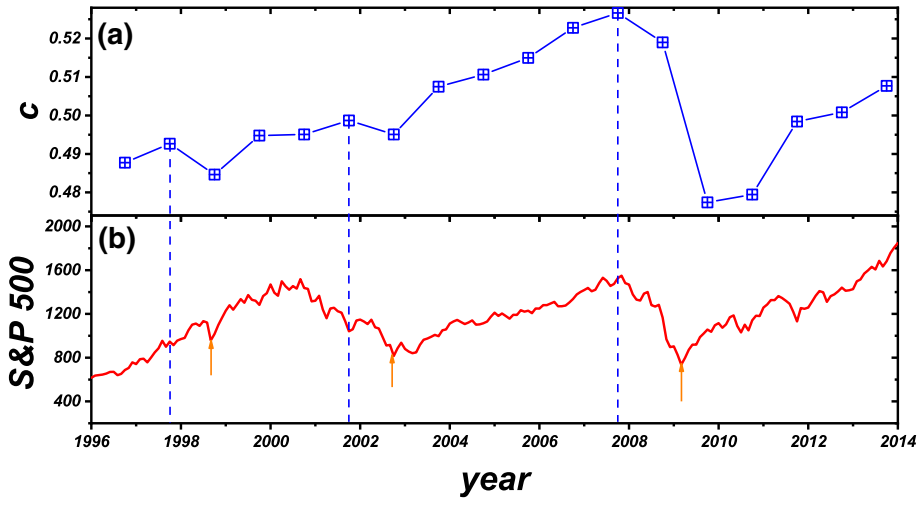

Fig. 4 **a** Evolution of the clustering coefficient C with years. **b** The evolution of the S&P 500 Index

Fig. 4b shows the S&P 500 Index. We find that the behavior of C (1 year ahead) and the S&P 500 Index is highly correlated (see Fig. 5). In particular, we can see 3 local maxima C (labeled by dashed blue lines) and 3 local minima S&P 500 (labeled by red lines with arrows), which stands for three financial crisis: the 1997-1998 Asian Financial Crisis, the 2002 dotcom bubble and the 2008 global financial crisis. The maxima values are always one year before the minima. This suggests that our network index C might be able to help in forecasting the following year's stock market returns.

Given the relation between the evolved topology of the network, we tested whether the network contains enough information to predict at the single node level whether the company represented by this node will collapse or make exceptional gains in the following year. We define the systemic risk as the disappeared firms (through bankruptcy, privatization or mergers) in the next year and the systemic return as the firms with the highest stock-market return ratio in the next year (we define firms as top firms if their stock returns are ranked as top 5% of the firm sample). Note that we define firm collapse as firms that are not listed companies any more if they bankrupt, get merged or choose privatization, among which bankruptcy and mergers are the most common events. The dynamics of product market networks changes firm's competition environment by affecting their revenue and market share leading to bankruptcy or merger. The intensified competition in the product network might also incentivize privatization.

To further show that network topology matters, we next analyze the K-shell structure of the network and find that the K-shell method is quite useful for predicting well above random, the disappeared firms. We present the receiver operating characteristic curve (ROC) using the K-shell method for predicting the disappeared firms of the 2012 network in Fig. 6. The area under the ROC curve (AUC) is significantly higher than the random case for all years (shown in Fig. 7). In particular, firms in the lower shells have higher market risk ratio. A possible reason is that the nodes in lower shells (less than 50) in the network are very fragile with higher risk. We present the scatter plot of risk ratio, ρ (the ratio of disappeared firms at each shell), as a function of K-shell, see Fig. 8 for three specific years, 1996, 2003 and 2012. Our results reveal that firms in more outer shells have higher risk to collapse.

Next a combined topological approach was tested to predict the same tags (collapsed or top) using a large set of topological measures and a Random Forest classifier. The combined features approach (Rosen and Louzoun 2016) significantly outperformed the

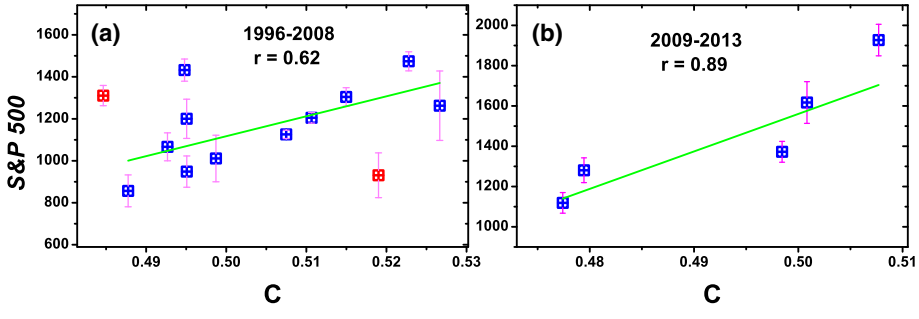

Fig. 5 The network Index C versus the annual average S&P 500 Index of next year. **a** The years from 1996 to 2008, i.e before financial crisis. **b** The years from 2009 to 2013, i.e after financial crisis. The straight lines show the best linear regression with correlation r

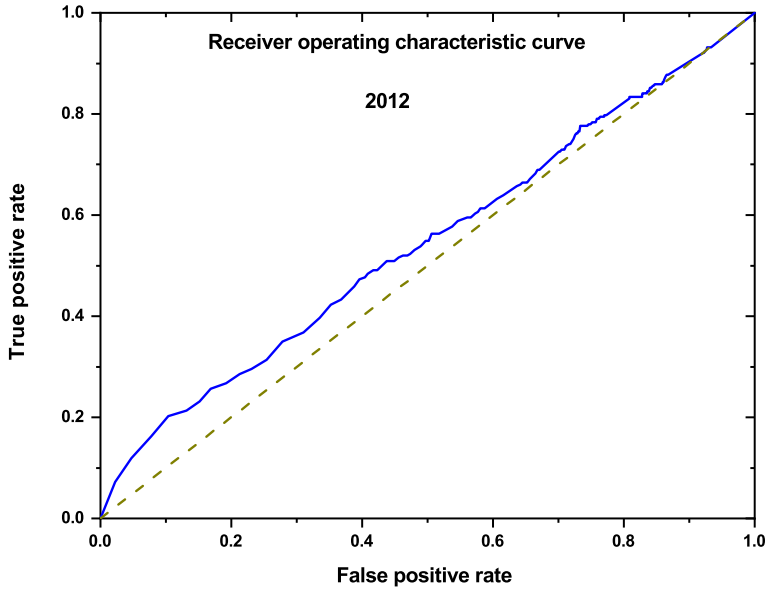

Fig. 6 The receiver operating characteristic curve (ROC) using the k-shell method for predicting the disappeared firms in the 2012 network

K-Core based classification (Fig. 9). For both collapsed and top companies, we find that the topological information of our network provides significantly better predictions than the random case (the AUC is shown in Fig. 9 upper and middle plots) for all 17 years. We also compared the AUC of machine learning using the network topology to the logistic regression based on the standard financial measures (non-network information): firm size, book-to-market ratio, leverage, profitability, prior year return, investment, liquidity, and Altman Z-score. The results are shown in Fig. 9 upper and middle plots.

Moreover when topological features are combined with the external features, the precision accuracy of RF method is significantly higher in almost all years and in both categories (see gain AUC in Fig. 9 bottom plot). This strongly suggests that the network topological information plays a prominent role in the prediction of future collapsed and top companies, it can be complementary to the traditional logistic regression methods. We note that during the years preceding the 2008 economic crisis, the network topology was extremely better than external data at predicting collapses. This is since most of the firms are influenced by the crisis, and some of them have been collapsed. As described above, our network features (e.g., k-shell) are quite sensitive to measure and predict the risk of firms. So, when topological features are combined with the external features, the precision accuracy of RF method is significantly higher for the collapsed category. In this study, we used the disappeared firms in the next year to measure the systemic risk. It contains bankruptcy, mergers or privatization. However, these are very different financial events, with potentially opposite meanings. Bankruptcy is always bad news, but a merger can be good news for stockholders. By separating the disappeared firms into bankruptcy and merger, we perform the same machine learning analysis, and we find that the internal information is important for bankruptcy; both topology and internal information have equal weight for mergers. But in both cases, the combination outperforms the internal information itself (see Fig. 10).

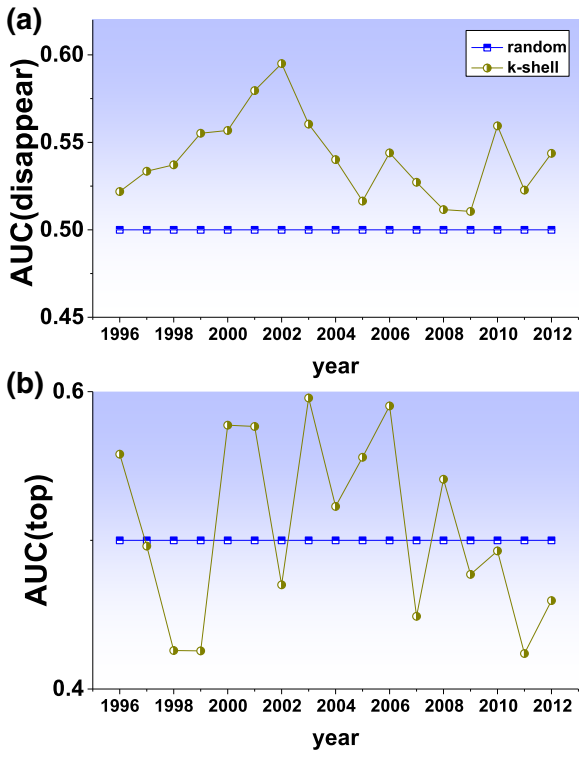

Fig. 7 The area under the ROC curve for (**a**) disappeared firms and (**b**) top return firms using forecasts from the *K*-shell methods

As discussed above, the top 5% and disappeared firms in year x are obtained from the next year $x + 1$. Which means that we need to know the future returns/disappearance of the training set to predict for the remaining 30% of the firms. We perform the same analysis using the entire data of the previous year as training and the entire data of the next year as test, i.e., we learn from year x to the outcome of year $x + 1$. We then test the effect of features of year $x + 1$ on the outcome of year $x + 2$ using the learning performed the year before. Our results are shown in Fig. 11, we find that the weights do not change drastically from year to year. The results for the bankruptcy are even better, but for the merger are slightly worse. The difference is probably because mergers are more volatile and the information from previous years is not as useful.

SUMMARY

In summary, we developed a combined network and machine learning algorithm to predict both collapsed and top companies in the financial market. Our network is based on the 10-K product similarities. We find that several key topological measures inherent in our networks can serve as good precursors for machine learning approach. A Random Forest approach is applied, but any other classifier can be used. The forecasting accuracies using our method are higher than using well-known logistic regression techniques. Moreover, when combining the external features and network topological measures, we find that the accuracy of the machine learning are significantly higher in almost all years and in both categories (collapsed and top gain). The proposed method and analysis can

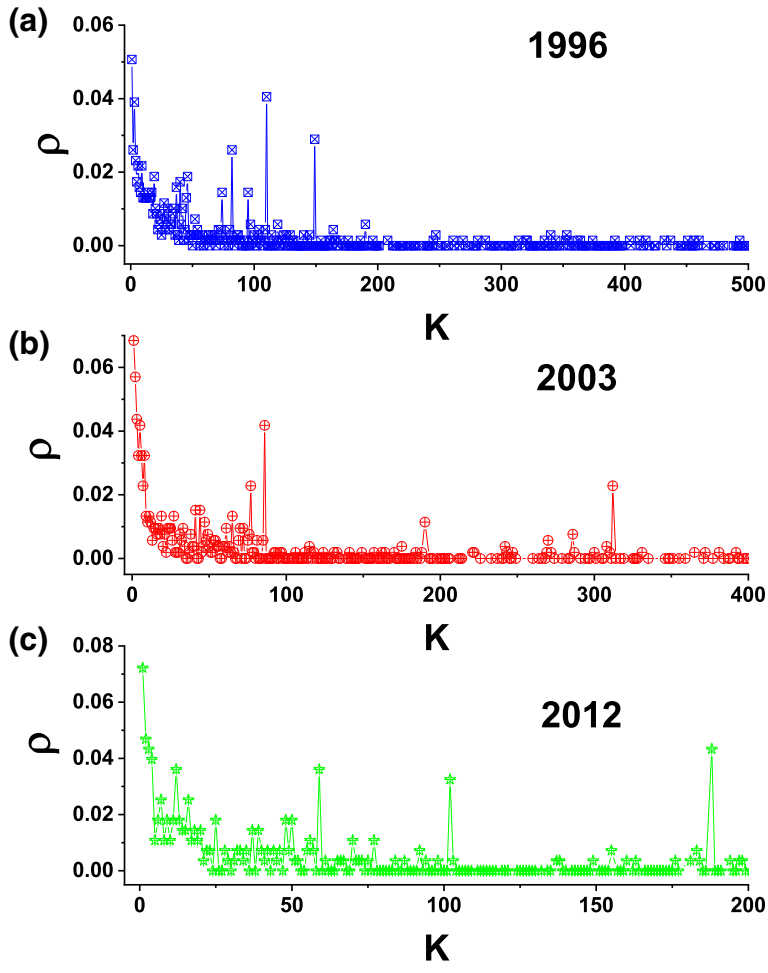

Fig. 8 The risk ratio, ρ, and number of nodes N in K-shell as a function of K for (**a**, **b**) 1996, (**c**, **d**) 2003 and (**e**, **f**) 2012 years

provide a new perspective for prediction of individual companies in the financial product markets and can potentially be used as a template to study other financial systems. Beyond its direct application in risk assessment for companies, the current work has two broad implications:

- First, the importance of product similarity networks. While such networks have been previously described (Hoberg and Phillips 2010), their implications for forecasting the future of companies is highly novel. The position of a company in the product space is highly correlated with both its risks and its growth potential. This relation is often not causal and may simply reflect common correlations between the product space topology and the company "situation". However, while the topology is easily defined, the situation of a company is currently described only by internal information on the company. Such internal information does not usually capture the effect of the market and of competition on its future. Indeed, combining the network product similarities and standard methods of ranking, improves the predictive power of both. Optimally positioning a company in this network may be the key for its future growth. While the current analysis was a pure prediction of the future of a

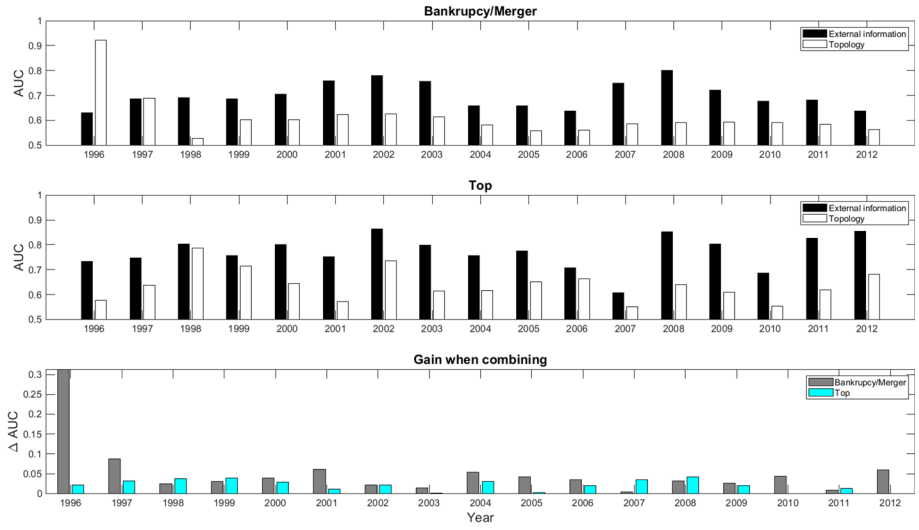

Fig. 9 (Upper and Middle) Comparison of AUC for logistic regression on external features and Random Forest Classifier on topological features for both collapsed and top companies. (Bottom) The gain in AUC for Random Forest on combined topological and external features for the collapsed and top companies

company, based on its position, we now plan to predict the opposite, which is the best position to ensure growth, and prevent collapse. Such a prediction would require to disentangle the causal effect of position in the network from its correlation with other unknown variables.

- Second, the importance of topology of networks. The success of a company is not diffusive, and similar companies often have different routes. While companies that share the same market may enjoy or suffer in common from fluctuation in the market size, they also compete. As such, beyond label diffusion, the detailed topology

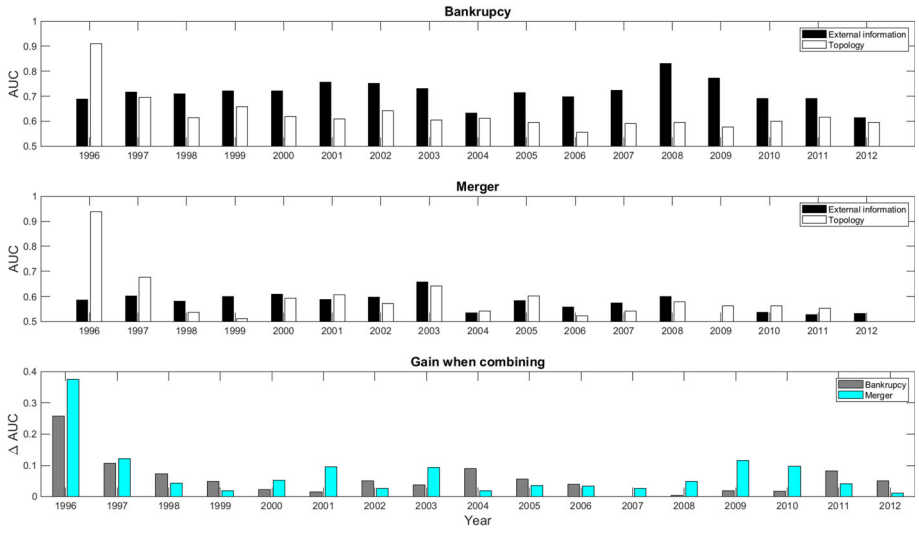

Fig. 10 (Upper and Middle) Comparison of AUC for logistic regression on external features and Random Forest Classifier on topological features for both bankruptcy and merger companies, respectively. (Bottom) The gain in AUC for Random Forest on combined topological and external features for the bankruptcy and merger companies

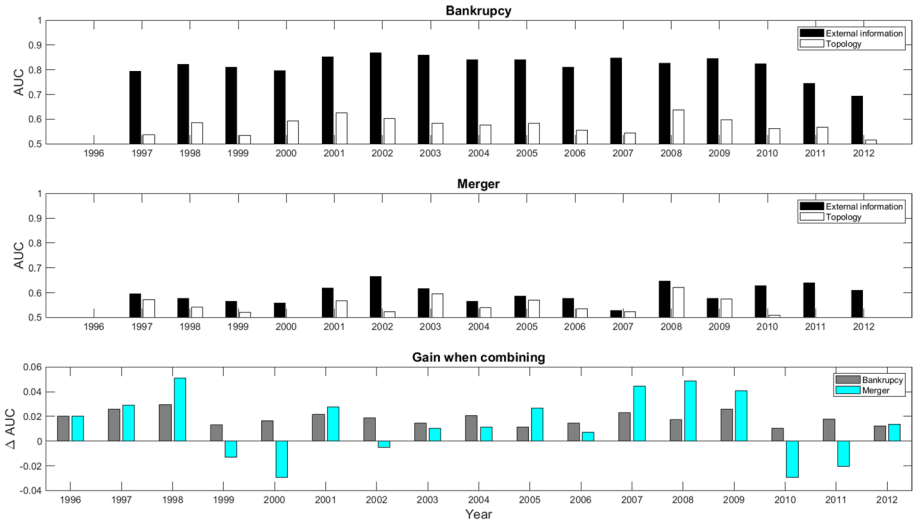

Fig. 11 The same as Fig. 10, but for predicting the next year

of the companies is important. Specifically their centrality in the network, as measured for example by their K-Core may be crucial.

We plan to expand this work to explicitly incorporate the dynamics of the network. The current analysis was static and based on separate classifiers for each year. However, the dynamics of the network may also be crucial. This will be tested using a recurrent neural network formalism.

Abbreviations

AUC: Area Under the ROC Curve; CRSP: Center for Research in Security Prices; MDS: Multi-Dimensional-Scaling; NAV: Network Attribute Vector; PDF: Probability Density Function; ReLU: Rectified Linear Unit; RF: Random Forest; ROC: Receiver Operating characteristic Curve; SEC: Securities Exchange Commission; S&P 500: Standard & Poor's 500

Acknowledgements

We thank Wolfgang Lucht or helpful discussions. PIK is a Member of the Leibniz Association.

Authors' contributions

JF, YL, and SH designed research; JF, KC, and YL performed research; JF, KC, LS, SL, JM, YL and SH analyzed data; and JF, KC, LS, SL, JM, YL and SH wrote the paper. All authors read and approved the final manuscript.

Funding

We acknowledge the Italy-Israel project OPERA, the Israel-Italian collaborative project NECST, the Israel Science Foundation, ONR, Japan Science Foundation, BSF-NSF, and DTRA (Grant no. HDTRA-1-10-1-0014), ARO, EPICC for financial support.

Availability of data and materials

All data and scripts for analysing the data are available from the corresponding authors upon reasonable request.

Competing interests

The authors declare no competing interests.

Author details

[1]Potsdam Institute for Climate Impact Research, Potsdam 14412, Germany. [2]Department of Physics, Bar Ilan University, Ramat Gan 52900, Israel. [3]Department of Mathematics, Bar Ilan University, Ramat Gan 52900, Israel. [4]Department of Economics, Lingnan University, Tuen Mun, Hong Kong.

Received: 25 January 2019 Accepted: 22 July 2019
Published online: 28 August 2019

References

Acemoglu D, Asuman O, Alireza T-S (2015) Systemic Risk and Stability in Financial Networks. Am Econ Rev 105(2):564–608. https://doi.org/10.1257/aer.20130456. https://www.aeaweb.org/articles?id=10.1257/aer.20130456

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Finance 23(4):589–609

Amihud Y (2002) Illiquidity and stock returns: cross-section and time-series effects. J Financ Mark 5(1):31–56

Bardoscia M, Stefano B, Fabio C, Caldarelli G (2017) Pathways towards instability in financial networks. Nat Commun 8:14416. https://doi.org/10.1038/ncomms14416. https://www.nature.com/articles/ncomms14416

Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci U S A 101(11):3747–3752. https://doi.org/10.1073/pnas.0400087101. http://www.pnas.org/content/101/11/3747

Batagelj V, Zaveršnik M (2011) Fast algorithms for determining (generalized) core groups in social networks. ADAC 5(2):129–145. https://doi.org/10.1007/s11634-010-0079-y. https://doi.org/10.1007/s11634-010-0079-y

Belkin M, Niyogi P (2002) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in neural information processing systems. MIT Press, Cambridge. pp 585–591

Belkin M, Niyogi P (2004) Semi-supervised learning on riemannian manifolds. Mach Learn 56(1-3):209–239

Benami I, Cohen K, Nagar O, Louzoun Y (2019) Topological based classification of paper domains using graph convolutional networks. arXiv:1904.07787 [cs, stat]. arXiv: 1904.07787. http://arxiv.org/abs/1904.07787

Boers N, Bookhagen B, Barbosa HMJ, Marwan N, Kurths J, Marengo JA (2014) Prediction of extreme floods in the eastern Central Andes based on a complex networks approach. Nat Commun 5:5199. https://doi.org/10.1038/ncomms6199. http://www.nature.com/doifinder/10.1038/ncomms6199

Breiman L (2001) Random Forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Brockmann D, Helbing D (2013) The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. Science 342(6164):1337–1342. https://doi.org/10.1126/science.1245200. http://science.sciencemag.org/content/342/6164/1337

Bruna J, Zaremba Wojciech, Szlam A, LeCun Y (2013) Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203

Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of Internet topology using k-shell decomposition. Proc Natl Acad Sci 104(27):11150–11154. https://doi.org/10.1073/pnas.0701175104. http://www.pnas.org/content/104/27/11150

Cohen R, Havlin S (2010) Complex networks: structure, robustness and function. Cambridge university press, Cambridge

Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci U S A 103(7):2015–2020

Cooper MJ, Gulen H, Schill MJ (2008) Asset growth and the cross-section of stock returns. J Finance 63(4):1609–1651

Dijkstra EW (1959) A Note on Two Problems in Connexion with Graphs. Numer Math 1(1):269–271. https://doi.org/10.1007/BF01386390. http://dx.doi.org/10.1007/BF01386390

Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. In: Advances in neural information processing systems. MIT Press, Cambridge. pp 2224–2232

Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429(6988):180. https://doi.org/10.1038/nature02541. https://www.nature.com/articles/nature02541

Everett MG, Borgatti SP (1999) The centrality of groups and classes. J Math Sociol 23(3):181–201. https://doi.org/10.1080/0022250X.1999.9990219

Fama EF, French KR (1992) The cross-section of expected stock returns. J Finance 47(2):427–465

Grover A, Leskovec J (2016) Node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. ACM, New York. pp 855–864. https://doi.org/10.1145/2939672.2939754. event-place: San Francisco, California, USA. ISBN 978-1-4503-4232-2. http://doi.acm.org/10.1145/2939672.2939754

Ho TK (1995) Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1. pp 278–2821. https://doi.org/10.1109/ICDAR.1995.598994

Hoberg G, Phillips G (2010) Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. Rev Financ Stud 23(10):3773–3811. https://doi.org/10.1093/rfs/hhq053. http://rfs.oxfordjournals.org/content/23/10/3773

Hoberg G, Phillips G (2016) Text-based network industries and endogenous product differentiation. J Polit Econ 124(5):1423–1465. https://doi.org/10.1086/688176

Itzhack R, Mogilevski Y, Louzoun Y (2007) An optimal algorithm for counting network motifs. Phys A Stat Mech Appl 381:482–490. https://doi.org/10.1016/j.physa.2007.02.102. http://www.sciencedirect.com/science/article/pii/S0378437107002257

Kingma DP, Ba J (2014) Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs]. arXiv: 1412.6980. http://arxiv.org/abs/1412.6980

Kipf TN, Welling M (2016a) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907

Kipf TN, Welling M (2016b) Variational graph auto-encoders. arXiv preprint arXiv:1611.07308

Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888–893. https://doi.org/10.1038/nphys1746. http://www.nature.com/nphys/journal/v6/n11/full/nphys1746.html

Levy O, Goldberg Y, Dagan I (2015) Improving distributional similarity with lessons learned from word embeddings. Trans Assoc Comput Linguist 3:211–225

Li R, Dong L, Zhang J, Wang X, Wang W-X, Di Z, Stanley HE (2017) Simple spatial scaling rules behind complex cities. Nat Commun 8(1):1841. https://doi.org/10.1038/s41467-017-01882-w. https://www.nature.com/articles/s41467-017-01882-w

Ludescher J, Gozolchiani A, Bogachev MI, Bunde A, Havlin S, Schellnhuber HJ (2014) Very early warning of next el nino. Proc Natl Acad Sci 111(6):2064–2066. https://doi.org/10.1073/pnas.1323058111. http://www.pnas.org/content/111/6/2064

Masci J, Boscaini D, Bronstein M, Vandergheynst P (2015) Shapenet: Convolutional neural networks on non-euclidean manifolds. Technical report

Meng J, Fan J, Ashkenazy Y, Havlin S (2017) Percolation framework to describe El Nino conditions. Chaos Interdisc J Nonlinear Sci 27(3):035807. https://doi.org/10.1063/1.4975766. http://aip.scitation.org/doi/abs/10.1063/1.4975766

Meng J, Fan J, Ashkenazy Y, Bunde A, Havlin S (2018) Forecasting the magnitude and onset of El Niño based on climate network. New J Phys 20(4):043036. https://doi.org/10.1088/1367-2630/aabb25. http://stacks.iop.org/1367-2630/20/i=4/a=043036

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298(5594):824–827. https://doi.org/10.1126/science.298.5594.824. http://science.sciencemag.org/content/298/5594/824

Miura W, Takayasu H, Takayasu M (2012) Effect of Coagulation of Nodes in an Evolving Complex Network. Phys Rev Lett 108(16):168701. https://doi.org/10.1103/PhysRevLett.108.168701. https://link.aps.org/doi/10.1103/PhysRevLett.108.168701

Monti F, Boscaini D, Masci J, Rodola E, Svoboda J, Bronstein MM (2017) Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proc. cvpr, Vol. 1. IEEE, New York. p 3

Naaman R, Cohen K, Louzoun Y (2018) Edge sign prediction based on a combination of network structural topology and sign propagation. J Complex Netw 7(1):54–66. https://doi.org/10.1093/comnet/cny012. https://academic.oup.com/comnet/advance-article/doi/10.1093/comnet/cny012/4999727

Newman M (2010) Networks: an introduction. Oxford university press, Oxford

Perozzi B, Al-Rfou R, Skiena S (2014) Deepwalk: Online learning of social representations. In: Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining. ACM, New York. pp 701–710

Rosen Y, Louzoun Y (2014) Directionality of real world networks as predicted by path length in directed and undirected graphs. Phys A Stat Mech Appl 401:118–129. https://doi.org/10.1016/j.physa.2014.01.005. http://www.sciencedirect.com/science/article/pii/S0378437114000090

Rosen Y, Louzoun Y (2015) Topological similarity as a proxy to content similarity. J Complex Netw 4(1):38–60

Rosen Y, Louzoun Y (2016) Topological similarity as a proxy to content similarity. J Complex Netw 4(1):38–60. https://doi.org/10.1093/comnet/cnv012. https://academic.oup.com/comnet/article/4/1/38/2366087

Rosenfeld N, Globerson A (2017) Semi-supervised learning with competitive infection models. arXiv preprint arXiv:1703.06426

Sabidussi G (1966) The centrality index of a graph. Psychometrika 31(4):581–603. https://doi.org/10.1007/BF02289527. https://doi.org/10.1007/BF02289527

Saramäki J, Kivelä M, Onnela J-P, Kaski K, Kertész J (2007) Generalizations of the clustering coefficient to weighted complex networks. Phys Rev E 75(2):027105. https://doi.org/10.1103/PhysRevE.75.027105. https://link.aps.org/doi/10.1103/PhysRevE.75.027105

Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. In: European semantic web conference. Springer, Berlin. pp 593–607

Sinatra R, Wang D, Deville P, Song C, Barabási A-L (2016) Quantifying the evolution of individual scientific impact. Science 354(6312):5239. https://doi.org/10.1126/science.aaf5239. http://science.sciencemag.org/content/354/6312/aaf5239

Sindhwani V, Partha N, Belkin M (2005) Beyond the point cloud: from transductive to semi-supervised learning. In: Proceedings of the 22nd international conference on machine learning. ACM, New York. pp 824–831

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442. https://doi.org/10.1038/30918. http://www.nature.com/nature/journal/v393/n6684/full/393440a0.html

Yang Z, Cohen WW, Salakhutdinov R (2016) Revisiting semi-supervised learning with graph embeddings. arXiv preprint arXiv:1603.08861

Zhao J-H, Zhou H-J, Liu Y-Y (2013) Inducing effect on the percolation transition in complex networks. Nat Commun 4:2412. https://doi.org/10.1038/ncomms3412. http://www.nature.com/ncomms/2013/130909/ncomms3412/full/ncomms3412.html

Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B (2004) Learning with local and global consistency. In: Advances in neural information processing systems. MIT Press, Cambridge. pp 321–328

Zhu X, Ghahramani Z, Lafferty JD (2003) Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th international conference on machine learning (icml-03). ACM, New York. pp 912–919

Publisher's Note