Predicting future market structure by combining social and financial network information

Thársis T. P. Souza[a,*], Tomaso Aste[a]

[a]*Department of Computer Science, UCL, Gower Street, London, WC1E 6BT, UK*

Abstract

We demonstrate that future market correlation structure can be predicted with high out-of-sample accuracy by a multiplex network approach that combines information from social media and financial data. Market structure is measured by quantifying co-movement of asset prices returns while social structure is measured as the co-movement of social media opinion on the same assets. Prediction is obtained by a simple model that uses link persistence and link formation by triadic closure across both financial and social media layers. Results show that proposed model can predict future market structure with up to 40% out-of-sample performance improvement compared to a benchmark model that assumes a time-invariant financial correlation structure. Social media information leads to improved models for all settings tested, particularly in the long-term prediction of financial market structure. Surprisingly, financial market structure showed higher predictability than social opinion structure.

Keywords: Financial Networks; Network Link Prediction; Correlation Structure Prediction; Information Filtering Networks; Correlation-Based Networks; Social Media

1. Introduction

Financial markets can be regarded as a complex network in which nodes represent different financial assets and edges represent one or many types of relationships among those assets. Filtered correlation-based networks have been successfully used in the literature to study financial markets structure particularly from observational data derived from empirical financial time series [1, 2, 3, 4, 5]. The underlying principle is the use of correlations from empirical financial time series to construct a sparse network representing the most relevant connections. Analyses on filtered correlation-based networks for information extraction [6, 7, 3] are widely used to explain market interconnectedness

*Corresponding author
Email address: T.Souza@ucl.ac.uk (Thársis T. P. Souza)

from high-dimensional data. Applications range from asset allocation [8] to market stability assesments [9] and hierarchical structure analyses [2, 3, 4, 10, 11] and the identification of lead-lag relationships [12].

Most of the literature so far have been focusing on the analysis of financial time series. However, in recent years a large amount of information about financial markets have become available from exogenous sources such as social media. It is reasonable to conceive that changes in social media sentiment [13] and changes in asset prices might be related. Some previous studies have indeed shown the existence of relations including lead-lag relationships which in some cases indicate that social media can be used to predict changes in asset prices [14, 15, 16, 17, 18, 19]. When new information hits the markets investors may react rationally or irrationally [20, 21] expressing opinions on social media that can become market actions enabling opportunities to forecast future asset prices. However, it has also been highlighted that not all assets behave in the same way with some that are more influenced by social media sentiment and others that are, on the contrary, more influential on the social media sentiment [22]. Beside each single financial asset, the question that we address in this paper is whether the entire financial market structure is related to the structure constructed from social media sentiment and whether there exist lead-lag relationships that can be used for forecasting one structure in terms of the other.

In this work, we use dynamical Kendall correlations computed over rolling windows to investigate the temporal evolution of market structure represented by filtered correlation-based networks constructed from stock market prices and from Twitter-sentiment signals. We generate two networks: one from log-returns of stock prices and the other from twitter-sentiment. The two networks are treated as a multilayer problem with two layers of networks that share the same nodes but have different edge sets. We investigate whether financial market structure can be better predicted by combining past financial information with past social media sentiment information. The market structure forecasting problem is formulated as a link prediction problem where we estimate the probability of addition or removal of a link in the future from the information about the structure of the two financial and social networks from the past.

2. Methods

2.1. Financial and Social Networks

We selected $N = 100$ most capitalized companies that were part of the S&P500 index during the period 09/05/2012 to 08/25/2017. The list of company names is reported in the Appendix A.1. For each stock i the financial variable is defined as the daily stock's log-return $R_i(\tau) = \log Price(\tau) - \log Price(\tau - 1)$, where $Price(\tau)$ designates closing price at time τ. The social media variable is defined as the the social media opinion O_i over the stock i which is estimated as the total number of bullish daily tweets related to the stock i at time τ. Twitter sentiment data were provided by PsychSignal.com [23]. A Twitter message is defined to be related to a given stock if its ticker is mentioned. The dataset is

based on English language content and it is agnostic to the country source of the Twitter message.

Stock returns R_i and social media opinion scores O_i each amounts to time series of length equal to 1024 trading days. The series are divided time-wise into $M = 224$ windows $t = 1, 2, \ldots, M$ of width $T = 126$ trading days. A window step length parameter $\delta T = 5$ defines the displacement of the window, i.e., the number of trading days between two consecutive windows. The choice of window width T and window step δT is arbitrary and it is a trade-off between too dynamic and too smooth in the analysis taken. The smaller the window width and the larger the window steps the more dynamic the data are.

In order to characterize the synchronous time evolution of assets, we use the equal time Kendall's rank coefficients between assets i and j defined as

$$\rho_{i,j}(t) = \sum_{t'<\tau} sgn(V_i(t') - V_i(\tau))sgn(V_j(t') - V_j(\tau)), \tag{1}$$

where t' and τ are time indexes within the window t and $V_i \in \{R_i, O_i\}$.

Kendall's rank coefficients fulfill the condition $-1 \leq \rho_{i,j} \leq 1$ and form an $N \times N$ correlation matrix $C(t)$, which serves as the basis for the networks constructed in this paper. For the purpose of constructing the asset-based financial and social networks we define a distance between a pair of stocks. This distance is associated with the edge connecting the stocks and it reflects the level at which the stocks are correlated. We use a simple non-linear transformation $d_{i,j}(t) = \sqrt{2(1 - \rho_{i,j}(t))}$ to obtain distances with the property $2 \geq d_{i,j} \geq 0$, forming a $N \times N$ symmetric distance matrix $D(t)$.

We extract the $N(N-1)/2$ distinct distance elements from the upper triangular part of the distance matrix $D(t)$ which are then sorted in an ascending order and form an ordered sequence $d_1(t), d_2(t), \ldots, d_{N(N-1)/2}(t)$. Since we require the graph to be representative of the market, it is natural to build the network by including only the strongest connections in it. The number of edges to include is, of course, arbitrary. Here we include the edges in the bottom quartile, i.e., the 25% shortest edges in the graph (largest correlations), thus giving $E(t) = \{d_1(t), d_2(t), \ldots, d_{\lfloor N/4 \rfloor}(t)\}$. The presented mechanism for constructing networks defines them uniquely and, consequently, no additional hypothesis about graph topology is required.

Let us denote $E^F(t)$ and $E^S(t)$ as the set of edges constructed from the distance matrices derived from stock returns $R(t)$ and social media opinion $O(t)$, respectively. Two networks are considered as two layers of a duplex structure $\mathcal{G} = \{G^F, G^S\}$ where $G^F = (V, E^F)$ and $G^S = (V, E^S)$ with V the vertex set of stocks which is common to both layers.

2.2. Persistence

The state of an edge between vertices u and v in the financial layer at time t is represented with the corresponding adjacency matrix element $E^F_{u,v}(t)$: a binary variable with $E^F_{u,v}(t) = 1$ indicating the existence of the edge and $E^F_{u,v}(t) = 0$ its absence. Analogously the variable $E^S_{u,v}(t)$ accounts for the presence or absence

of edge (u, v) in the social (S) layer. The variable $E_{u,v}(t) = E^F_{u,v}(t) \vee E^S_{u,v}(t) = 1$ indicates instead the presence of at least one edge between u and v in the two layers; $E_{u,v}(t) = 0$ indicates that no edges are present between u and v in any layer.

2.3. Triadic Closure

Let $\mathcal{N}_{uv}$ be the set of nodes that are common neighbors to vertices u and v. We define the triadic closure $T^F_{u,v}(t)$ of an edge (u, v) at layer F and time t as the mean of the clustering coefficients of vertices in $\mathcal{N}_{uv}$:

$$T^F_{u,v}(t) = \frac{1}{|\mathcal{N}_{uv}|} \sum_{i \in \mathcal{N}_{uv}} C^F_i(t), \tag{2}$$

where term C^F_i is the clustering coefficient of node i which accounts for the fraction of triads in the neighbors of i that are closed in triangles

$$C^F_i = 2 \frac{\text{Number of triangles with a vertex on } i}{k_i(k_i - 1)} = \frac{\sum_{j,k \in \mathcal{N}_i} E^F_{j,k}}{k_i(k_i - 1)} \tag{3}$$

with k_i the degree of vertex i and $\mathcal{N}_i$ the neighborhood of i.

In the multiplex case, we keep the same definition but in this case triangles can be formed across several layers [24, 25]. For the multiplex case we shall use the symbol $T_{u,v}(t)$.

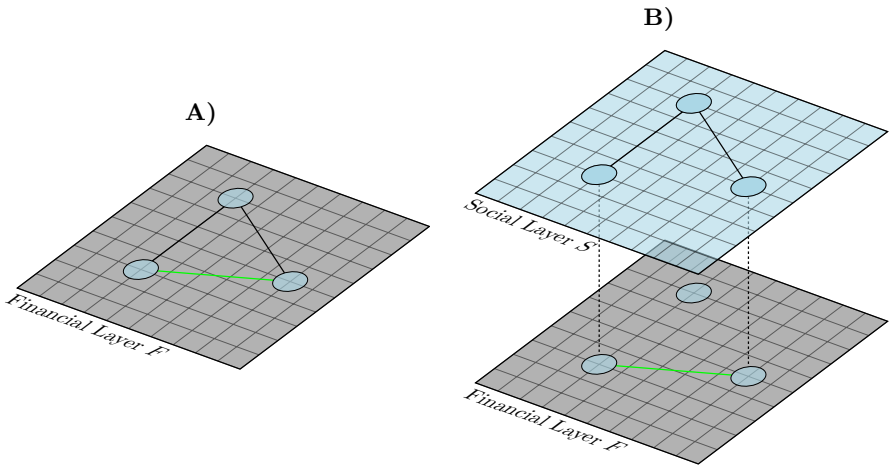

Figure 1: Triads on a single layered network (Panel A) and on a multiplex network (Panel B). The clustering coefficient of node i accounts for the fraction of triads in the neighborhood of i that are closed in triangles. The triadic closure of an edge (u, v) at layer F is a function of the clustering coefficients of the common neighbors of the vertices u and v. Triangles can be formed in a single layer or across layers.

2.4. Link Prediction

We aim to predict the probability that an edge is inserted or removed in the financial network, $G^F(t+h)$, at a future time $t+h$ by using the information about the past structures of the financial and social networks at previous times $t' \leq t$. For this purpose we consider two mechanisms:

1) the tendency of an edge present at a previous time to persist in the future (*edge persistence*);

2) the propensity to close triangles within a layer or across layers (*triadic closure*).

Persistence quantifies the tendency for an edge, present in the graph at time t to be present also in the graph at a later time $t+h$, whereas triadic closure quantifies the propensity to close a triangle by adding a given edge. The mechanism of growth by triadic closure is based on a principle of transitivity largely observed in real-world networks where there is a tendency to form triangles. Under that principle, two nodes tend to be connected if they share common neighbors with high transitivity, i.e., propensity to close triangles.

repetition

The probability to insert an edge in the future is then computed by means of a logistic regression of the edge persistence and the triadic closure coefficients. Regression coefficients are estimated by best fitting on a training set which is composed of rolling windows of 126 trading days initially ranging from 09/05/2012 and 09/10/2014. Prediction concerns the presence of edges in the financial network at $h = 1$ to $h = 20$ weeks ahead the end of the training set. The test set originally ranges from 09/17/2014 to 08/25/2017. The procedure is repeated by moving forward the training window in 1-week steps.

The probability $p_{u,v}(t+h)$ to observe vertices u, v connected by an edge at $t+h$ can be inferred in terms of the set of previous triadic closure coefficients, $T_{u,v}(t)$, and edge persistence scores $E_{u,v}(t)$. We first consider a restricted model that uses financial information only which is given by the following logistic model

$$\log \frac{p^F_{u,v}(t+h)}{1-p^F_{u,v}(t+h)} = \tilde{\beta}^h_0 + \tilde{\beta}^h T^F_{u,v}(t) + \tilde{\gamma}^h E^F_{u,v}(t), \tag{4}$$

where we perform a 1-step ahead prediction for $h \in (1, 2, \ldots, 19, 20)$ weeks.

In order to calibrate the parameters in Eq. 4, we consider a training window of $W = 126$ days which ends at time t. The log-likelihood function over the training window for the logistic model from Eq. 4 is given by [26]

$$\mathcal{L}^F(t) = \sum_{t'=t-W+1}^{t} \sum_{uv \in E^F(t'+h)} -\log\left(1 + e^{\tilde{\beta}^h_0 + \tilde{\beta}^h T^F_{u,v}(t') + \tilde{\gamma}^h E^F_{u,v}(t')}\right) + \sum_{t'=t-W+1}^{t} \sum_{uv \in E^F(t'+h)} \left(1 - E^F_{uv}(t'+h)\right)\left(\tilde{\beta}^h_0 + \tilde{\beta}^h T^F_{u,v}(t') + \tilde{\gamma}^h E^F_{u,v}(t')\right). \tag{5}$$

We differentiate the log-likelihood function given by Eq. 5 in order to find maximum log-likelihood estimates for the coefficients of Eq. 4.

In order to verify whether the multiplex information is relevant in predicting links in the financial network compared to past financial network alone, we consider a full regression model that takes the set of previous triadic closure coefficients and edge persistence from the financial layer $(T^F_{u,v}(t), E^F_{u,v}(t))$, social layer $(T^S_{u,v}(t), E^S_{u,v}(t))$ and the multiplex network $(T^F_{u,v}(t), E^F_{u,v}(t))$. The full model is

$$\begin{aligned}\log \frac{p_{u,v}(t+h)}{1-p_{u,v}(t+h)} = &\beta_0^h + \beta_1^h T^F_{u,v}(t) + \beta_2^h E^F_{u,v}(t) + \\ &\gamma_1^h T^S_{u,v}(t) + \gamma_2^h E^S_{u,v}(t) + \theta_1^h T_{u,v}(t) + \theta_2^h E_{u,v}(t).\end{aligned} \tag{6}$$

The log-likelihood function $\mathcal{L}(t)$ of the full model in Eq. 6 and the model fitting can be obtained analogously to the previous procedure performed for the restricted model from Eq. 4.

The likelihood ratio statistic

$$\lambda(t) = -2(\mathcal{L}_{max}(t) - \mathcal{L}^F_{max}(t)) \tag{7}$$

where $\mathcal{L}_{max}(t)$ and $\mathcal{L}^F_{max}(t)$ are the maxima of the log-likelihood functions for the full and restricted models, respectively. Under some not too restrictive assumptions [26], $\lambda(t)$ can be assumed to follow a χ^2 distribution with 4 degrees of freedom where a value of $\lambda > 18.47$ is assumed to be statistically significant at $p = 0.001$, In that case, there is evidence to accept the full model that considers social and financial information compared to the restricted model that considers financial information only.

The model performance is estimated by counting the true positive (edges predicted to be there and indeed present in the future network) and false positive (edges predicted to be there but not present in the future network) and measuring of AUC (area under the receiver operating characteristic curve) in the test set which originally ranges from 09/17/2014 to 08/25/2017. AUC ranges from 0.50 to 1.00, with higher values indicating that the model discriminates better between the two categories (edge-present, edge-absent).

3. Results

3.1. Market structure dynamics

We first investigate financial network persistence by comparing the financial network $G^F(t)$ at time t with a future financial network, $G^F(t+h)$, h steps ahead. In order to quantify the changes in the correlation network structure we use two measures: A) the fraction of new edges in $G^F(t+h)$ that were not present in $G^F(t)$; B) the Jaccard Distance, defined as

$$Jaccard(G^F(t'), G^F(t)) = \frac{\|G^F(t') \cap G^F(t)\|}{\|G^F(t') \cup G^F(t)\|}.$$

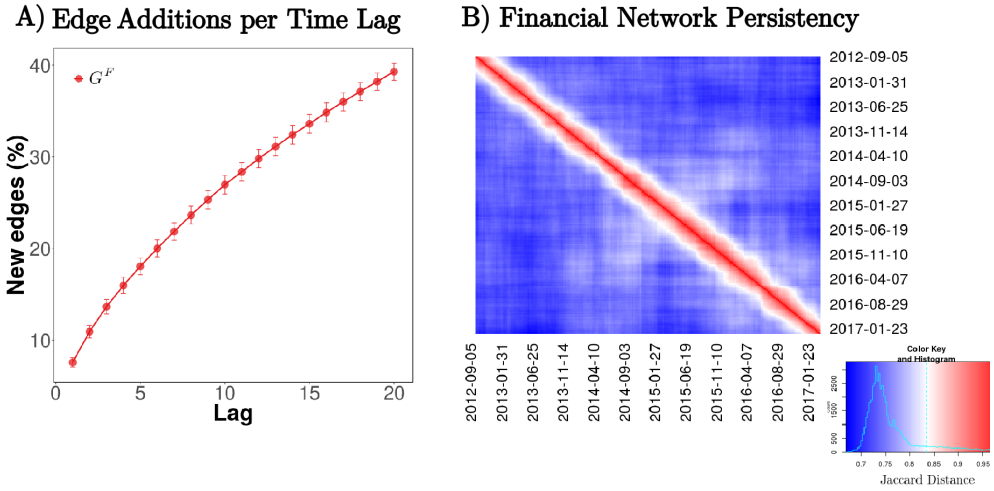

Figure 2: **Evidence that financial correlation structure changes considerably with time**. Panel A) shows the mean percentage of new edges in the financial network at time $t+h$ with respect to the edge set at time t ($1 \leq h \leq 20$ trading weeks). We observe that edges change considerably in the financial network with almost 40% of edges in financial networks changing after a period of $h = 20$ trading weeks. Panel B) shows the cross-similarity among financial networks measured as the Jaccard Distance between $G^F(t')$ and $G^F(t)$ with t and t' ranging from 09/05/2012 and 21/02/2017. We observe that edge changes (persistence) is quite stable overtime, i.e., the amount of edges that change is similar throughout the period. Network $G^F(t)$ are constructed at each time t from a correlation structure estimated from a sliding window of 126 trading days starting at time t. The windows move with time step of 1 trading week. Error bars in Panel A) indicate standard error.

Results are reported in Fig. 2, panels A) and B), respectively.

Fig. 2 panel A) shows the mean percentage of new edges in the financial network at time $t+h$ with respect to the edge set at time t ($1 \leq h \leq 20$ trading weeks). We observe that edges change considerably in the financial network with almost 40% of edges in financial networks changing after a period of $h = 20$ trading weeks. Fig. 2 panel B) shows the cross-similarity among financial networks measured as the Jaccard Distance between $G^F(t')$ and $G^F(t)$ with t and t' ranging from 09/05/2012 to 21/02/2017. We observe that edge changes (persistence) is quite stable overtime, i.e., the amount of edges that change is similar throughout the period. Hence, results indicate that the financial networks constructed are time-variant across the entire period studied with a stable rate of edge changes over time.

3.2. Prediction of Stock Market Structure

We use Eq. 6 to predict a the financial network, $G^F(t+h)$, at a future time $t+h$ by using the information about the past structures of the financial and social networks at previous times $t' \leq t$. Fig. 3 panel A) shows performance obtained in the prediction of out-of-sample edges for $h \in (1, 5, 10, 15, 20)$ trading steps

ahead. We observe that we achieve an overall high out-of-sample performance in financial network link prediction with performances in the range from 73% to 95% depending on time-lag and time-period with better prediction power for smaller number of steps ahead.

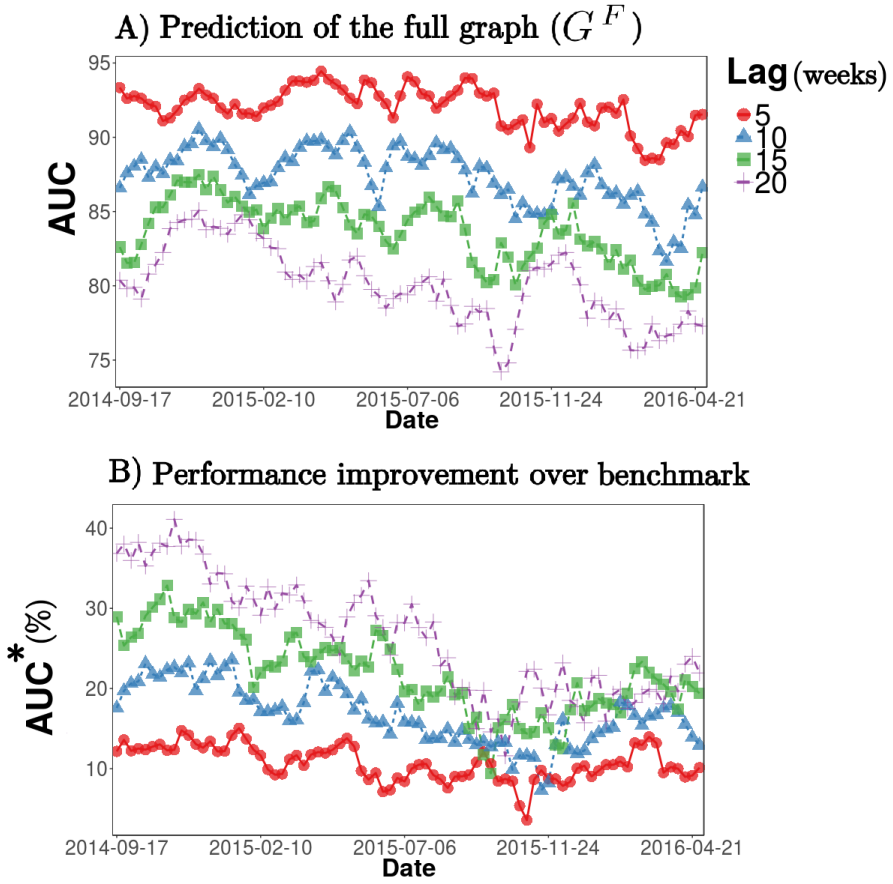

Figure 3: **Out-of-sample performance for financial network link prediction.** Plots show performance results (AUC), where each date (t) represents the performance obtained with a model trained with information up to time t and predicts edges in a network at time $t+h$. Panel A) shows performance obtained in the prediction of out-of-sample edges for $h \in (1, 5, 10, 15, 20)$ trading weeks. Panel B) shows performance improvement compared to a naive benchmark that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$.

We compare results against a benchmark model that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. Performance improvement against the benchmark is estimated as $AUC^* = (AUC - 0.5)/(\widehat{AUC} - 0.5) - 1$, where AUC represents the performance of the proposed model and $\widehat{AUC}$ is the performance of the benchmark. From Fig. 3 panel B) we observe

that the higher the number of steps ahead the higher the performance improvement over benchmark. Let us note that performance improvement over naive benchmark reached values as high as 40% for a long-term prediction with a lag of 20 trading weeks.

Fig. 4 reports an aggregate overview of the previous results for out-of-sample prediction in terms of the number of weeks ahead. We observe that the higher the lag the worst is the prediction performance (panel A), however the better is the improvement over the naive benchmark (panel B).

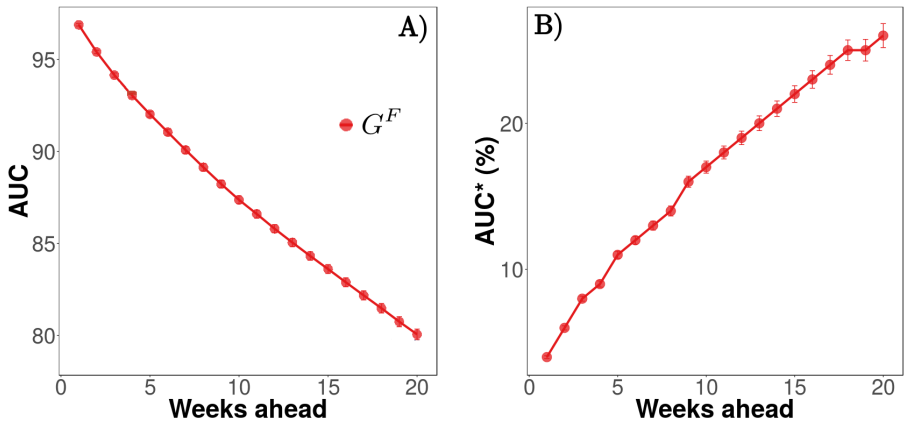

Figure 4: **Effect of time-lag in out-of-sample predictive performance.** Panel A) shows mean performance (AUC) in the prediction of out-of-sample edges of the full financial network G^F. Panel B) shows the performance improvement against a naive benchmark that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. Error bars indicate standard error.

In Appendix A.2, we report results obtained by using an expanding window instead of a rolling window as a training set. We observe that expanding the training set does not necessarily lead to better performance. In fact, the rolling window analysis yield better performance overall.

In order to verify whether the multiplex network is providing extra information with respect to the information from the financial network only, we re-computed the same out of sample edge prediction by using the financial network only which we compared against the full model that considers both information layers: financial and social. Comparison between the two models was performed by comparing their respective likelihoods. We have also segregated **disaggregated** the prediction of insertion of new edges E^+ and the prediction of edge deletions E^-. Results are reported in Table 1, where we show the likelihood values along with AUC performance obtained for each fit model.

We observe that models that include both financial and social information better fit the data compared to a model that considers financial data only, particularly for the case of prediction of insertion of new edges. The likelihood ratio increases with prediction lag indicating that full models (i.e. those that consider both financial and social networks) are particularly important in long-

Table 1: **Financial Link Prediction Performance Results.** High out-of-sample AUCs obtained indicate that the model has high performance balancing both false positives and false negatives predictions relative to true positive and negative values. Log-likelihood ratio (λ) increases with prediction lag indicating that social media features are particularly important for long-term prediction. AUC^* indicates performance improvement compared to a naive benchmark model that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. Results show that proposed model can achieve up to 27% improvement compared to the commonly-used assumption of stationarity in the correlation structure. Table show mean values obtained over the test period with corresponding standard deviation in parentheses. Models were trained in an rolling window with initial start and end dates of 09/05/2012 and 09/10/2014, respectively. Test period ranges from 09/17/2014 and 08/25/2017.

		E^+		E^-		G^F	
Lag	New edges (%)	AUC	λ	AUC	λ	AUC	AUC^* (%)
1	7.6 (0.51)	87 (0.33)	21 (0.76)	93 (0.11)	34 (1.2)	97 (0.064)	4 (0.091)
2	11 (0.69)	87 (0.37)	33 (1.2)	93 (0.1)	45 (1.5)	95 (0.092)	6 (0.14)
3	14 (0.79)	86 (0.39)	48 (1.5)	93 (0.11)	60 (1.6)	94 (0.11)	8 (0.17)
4	16 (0.88)	86 (0.39)	65 (2)	93 (0.11)	65 (1.9)	93 (0.13)	10 (0.21)
5	18 (0.92)	85 (0.41)	85 (2.6)	93 (0.11)	66 (1.9)	92 (0.15)	11 (0.24)
6	20 (0.93)	85 (0.41)	100 (3.2)	93 (0.1)	74 (2)	91 (0.16)	12 (0.27)
7	22 (0.95)	84 (0.42)	120 (3.5)	93 (0.1)	70 (2.2)	90 (0.18)	13 (0.3)
8	24 (0.98)	84 (0.43)	150 (4.3)	93 (0.1)	72 (1.9)	89 (0.19)	15 (0.33)
9	25 (0.99)	83 (0.44)	180 (5.7)	93 (0.1)	74 (2.2)	88 (0.21)	16 (0.37)
10	27 (1)	83 (0.43)	220 (6.3)	93 (0.096)	79 (1.9)	87 (0.21)	17 (0.4)
11	28 (1)	82 (0.43)	260 (7.2)	93 (0.094)	78 (2)	87 (0.22)	18 (0.43)
12	30 (1)	82 (0.42)	300 (7.9)	93 (0.09)	86 (2.4)	86 (0.22)	19 (0.45)
13	31 (0.99)	82 (0.43)	330 (7.9)	93 (0.09)	95 (2.1)	85 (0.22)	20 (0.49)
14	32 (1)	81 (0.43)	360 (9.2)	93 (0.084)	100 (2.4)	84 (0.23)	21 (0.51)
15	34 (1)	81 (0.43)	390 (9.9)	93 (0.083)	110 (2.3)	84 (0.24)	22 (0.55)
16	35 (1)	81 (0.43)	410 (10)	93 (0.08)	120 (3)	83 (0.24)	23 (0.58)
17	36 (0.99)	80 (0.43)	440 (11)	94 (0.079)	130 (2.6)	82 (0.25)	24 (0.62)
18	37 (0.97)	80 (0.44)	470 (12)	94 (0.076)	150 (3)	82 (0.25)	25 (0.67)
19	38 (0.94)	80 (0.46)	500 (12)	94 (0.072)	160 (3.6)	81 (0.27)	26 (0.71)
20	39 (0.95)	80 (0.48)	510 (12)	94 (0.068)	170 (3.7)	80 (0.28)	27 (0.79)

*A likelihood ratio of $\lambda > 18.47$ indicates statistical significance at $p = 0.001$.

term link prediction. Results confirm that the multiplex network is distinctively better than the single financial layer with all likelihood ratios with p-value $<$ 0.001 for all configurations tested.

3.3. Prediction of Social Media Structure

We have so far established that social opinion structure can provide statistically-significant information about future financial market structure. In this section, we investigate whether financial market structure can also significantly improve the prediction of future social opinion structure and whether this effect is larger or smaller than the previous.

The comparison between performance results is summarized in Fig. 5, where the prediction of social opinion structure G^S is plotted together with the results for the prediction of financial market structure G^F discussed previously. Surprisingly, results suggest that financial market structure has higher predictability than social opinion structure. We also observe that both the financial network and social opinion network predictions lead to an improvement compared to the

naive benchmark that considers time-invariance in social network structure. As previously observed, relative performance improvement increases with time lag. In this case, relative improvement is higher for the social media prediction than for the financial network as observed in Fig. 5 panel B).

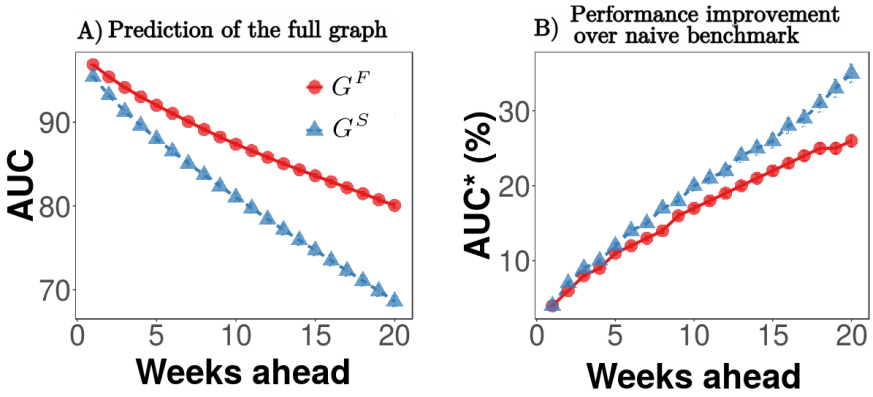

Figure 5: **Evidence that financial market structure has higher predictability than social media structure.** Panel A) shows mean performance (AUC) in the prediction of out-of-sample edges of the full financial network G^F and social network G^{SM}. Panel B) shows the performance improvement against a naive benchmark that assumes that correlation structure is time-invariant. Error bars indicate standard error.

One of the possible reasons why social opinion structure is less predictable compared to financial network structure is the higher structural variability of the former compared to the latter. Fig. 6 provides evidence that social media structure is less stable than financial market structure in terms of number of edge changes in time. More edges changed in the social opinion network than in the financial network for all lags tested. We observe that more than 50% of the edges in the social media structure changed over a time lag of 20 trading weeks compared to a change of about 40% in the financial network.

4. Discussion and Conclusions

We investigated whether financial market structure can be better predicted by combining past financial information with past social media sentiment information. We considered $N = 100$ most capitalized companies that were part of the S&P500 index in the period between May 2012 and August 2017. We generated two networks: A financial network constructed from log-returns of equity prices and a social network constructed from twitter-sentiment analytics. We constructed filtered correlation-based networks by keeping the strongest top quartile correlations only considering a rolling window of $T = 126$ trading days. The two networks were treated as a multiplex problem with two layers of networks that share the same nodes (stocks) but have different edge sets.

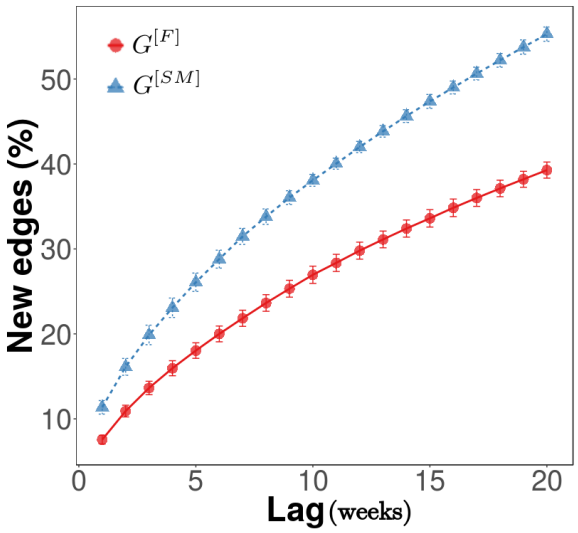

Figure 6: **Evidence that social media structure is less stable than financial market structure in terms of number of edge changes in time.** We observe that almost 40% of edges in Financial Networks changed after a period of 20 trading weeks while the social media structure changed more than 50% of its edges over the same time lag. A network at time t is constructed from a correlation structure estimated from a sliding window of 126 trading days starting at time t that moves with time step of 1 trading week. The financial network measures co-movement of stock returns while the social network measures co-movement of opinion over the same stocks. Error bars indicate standard error.

The financial market structure forecasting problem was formulated as a link prediction problem to estimate the probability of addition or removal of a financial link in the future from the information about the structure of the two financial and social networks in the past.

We proposed that financial network links were formed by a combination of two mechanisms: (i) triadic closure and (ii) edge persistence. The first mechanism assumes that two stocks have a propensity to be correlated if they share common neighbors. The edge persistence mechanism assumes that two connected stocks tend to remain connected in the future. A logistic model was trained over a set of data between 09/05/2012 and 09/10/2014 and then results were reported for the validation set over the following period from 09/17/2014 and 08/25/2017.

Results indicate that financial market structure is considerably time-variant which invalidates the commonly-used assumption of time-invariance in the stocks correlation structure. The proposed model showed high out-of-sample performance in financial network link prediction particularly in the case of long-term predictions achieving up to 40% performance improvement over a naive benchmark that assumed time-invariance in market structure. Likelihood ratio analysis demonstrated that models that considered both financial and social information better fit the data when compared to a restricted model that considers

financial information only. This provides evidence that supports the use of social information in the prediction of financial market structure.

Finally, findings indicated that social opinion structure is less stable than financial market structure. Surprisingly, the prediction of financial market structure using past social and financial information presented higher performance compared to the problem of predicting social opinion structure using past social and financial information.

Let us note that network link formation can occur due to mechanisms beyond the ones here studied. For instance, networks can grow as a result of a growth process that adds new nodes in the network, e.g., IPOs can generate growth in a financial network. Among other possible mechanisms, link formation can occur due to preferential attachment, a phenomenon widely observed in real networks where new nodes tend to link to the more connected nodes [27].

In sum, this study indicates that social opinion structure is relevant to predict future financial correlation structure. This result has important consequences because of the fundamental importance of financial correlation structure in Modern Portfolio Theory (MPT) [28], Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) [29]. Future work will focus on the investigation of further mechanisms of financial link formation and on applications in portfolio allocation strategies.

5. Acknowledgments

This work was supported by PsychSignal.com, which provided social media data. T.A. acknowledges support of the UK Economic and Social Research Council (ESRC) in funding the Systemic Risk Centre (ES/K002309/1). T.T.P.S. acknowledges financial support from CNPq - The Brazilian National Council for Scientific and Technological Development.

[1] M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, R. N. Mantegna, Statistically validated networks in bipartite complex systems, PLoS ONE 6 (3) (2011) 1–11. doi:10.1371/journal.pone.0017994.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0017994

[2] R. N. Mantegna, Hierarchical structure in financial markets, The European Physical Journal B - Condensed Matter and Complex Systems 11 (1) (1999) 193–197. doi:10.1007/s100510050929.
URL http://dx.doi.org/10.1007/s100510050929

[3] T. Aste, W. Shaw, T. Di Matteo, Correlation structure and dynamics in volatile markets, New Journal of Physics 12 (8) (2010) 085009.

[4] M. Tumminello, F. Lillo, R. N. Mantegna, Correlation, hierarchies, and networks in financial markets, Journal of Economic Behavior & Organization 75 (1) (2010) 40 – 58, transdisciplinary Perspectives on Economic Complexity. doi:http://dx.doi.org/10.1016/j.jebo.2010.01.004.
URL http://www.sciencedirect.com/science/article/pii/S0167268110000077

[5] M. Tumminello, T. Aste, T. Di Matteo, R. N. Mantegna, A tool for filtering information in complex systems, Proceedings of the National Academy of Sciences of the United States of America 102 (30) (2005) 10421–10426. arXiv:http://www.pnas.org/content/102/30/10421.full.pdf, doi:10.1073/pnas.0500298102.
URL http://www.pnas.org/content/102/30/10421.abstract

[6] T. Aste, W. Shaw, T. D. Matteo, Correlation structure and dynamics in volatile markets, New Journal of Physics 12 (8) (2010) 085009.
URL http://stacks.iop.org/1367-2630/12/i=8/a=085009

[7] W.-M. Song, T. Aste, T. Di Matteo, Analysis on filtered correlation graph for information extraction, Statistical Mechanics of Molecular Biophysics (2008) 88.

[8] F. Pozzi, T. Di Matteo, T. Aste, Spread of risk across financial markets: better to invest in the peripheries, Scientific reports 3.

[9] R. Morales, T. Di Matteo, R. Gramatica, T. Aste, Dynamical generalized hurst exponent as a tool to monitor unstable periods in financial time series, Physica A: Statistical Mechanics and its Applications 391 (11) (2012) 3180–3189.

[10] N. Musmeci, T. Aste, T. di Matteo, Clustering and hierarchy of financial markets data: advantages of the dbht., CoRR.

[11] W.-M. Song, T. Di Matteo, T. Aste, Hierarchical information clustering by means of topologically embedded graphs, PLoS One 7 (3) (2012) e31929.

[12] C. Curme, H. E. Stanley, I. Vodenska, Coupled network approach to predictability of financial market returns and news sentiments, International Journal of Theoretical and Applied Finance 18 (07) (2015) 1550043.

[13] O. Kolchyna, T. T. P. Souza, P. Treleaven, T. Aste, Twitter sentiment analysis: Lexicon method, machine learning method and their combination, in: G. Mitra, X. Yu (Eds.), Handbook of Sentiment Analysis in Finance, 2016, Ch. 5.

[14] J. Manfield, D. Lukacsko, T. T. P. Souza, Bull bear balance: A cluster analysis of socially informed financial volatility, in: 2017 Computing Conference, 2017, pp. 421–428. doi:10.1109/SAI.2017.8252134.

[15] T. T. P. Souza, O. Kolchyna, P. Treleaven, T. Aste, Twitter sentiment analysis applied to finance: A case study in the retail industry, in: G. Mitra, X. Yu (Eds.), Handbook of Sentiment Analysis in Finance, 2016, Ch. 23.

[16] I. Zheludev, R. Smith, T. Aste, When Can Social Media Lead Financial Markets?, Scientific Reports 4.

[17] P. C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, The Journal of Finance 62 (3) (2007) 1139–1168.

[18] M. Alanyali, H. S. Moat, T. Preis, Quantifying the relationship between financial news and the stock market, Sci. Rep. 3.

[19] H. Mao, S. Counts, J. Bollen, Quantifying the effects of online bullishness on international financial markets, European Central Bank Workshop on Using Big Data for Forecasting and Statistics, Frankfurt, Germany.

[20] W. F. M. D. Bondt, R. Thaler, Does the stock market overreact?, The Journal of Finance 40 (3) (1985) pp. 793–805.
URL http://www.jstor.org/stable/2327804

[21] A. Shleifer, Inefficient Markets: An Introduction to Behavioral Finance, Clarendon Lectures in Economics, OUP Oxford, 2000.

[22] T. T. P. Souza, T. Aste, A nonlinear impact: evidences of causal effects of social media on market prices, ArXiv e-printsarXiv:1601.04535.

[23] The psychsignal website (Oct. 2015).
URL https://www.psychsignal.com

[24] F. Battiston, V. Nicosia, V. Latora, The new challenges of multiplex networks: Measures and models, The European Physical Journal Special Topics 226 (3) (2017) 401–416. doi:10.1140/epjst/e2016-60274-8.
URL http://dx.doi.org/10.1140/epjst/e2016-60274-8

[25] E. Cozzo, M. Kivelä, M. D. Domenico, A. Solé-Ribalta, A. Arenas, S. Gómez, M. A. Porter, Y. Moreno, Structure of triadic relations in multiplex networks, New Journal of Physics 17 (7) (2015) 073029.
URL http://stacks.iop.org/1367-2630/17/i=7/a=073029

[26] J. Faraway, Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, CRC Press, 2006.

[27] A. Barabási, M. PÃ3sfai, Network Science, Cambridge University Press, 2016.
URL https://books.google.co.uk/books?id=ZVHesgEACAAJ

[28] D. Luenberger, Investment Science, Oxford University Press, 2014.
URL https://books.google.co.uk/books?id=YMSeDAEACAAJ

[29] J. Campbell, J. Campbell, A. Lo, A. MacKinlay, J. Champbell, A. LO, A. MacKinlay, P. Lo, O. Campbell, The Econometrics of Financial Markets, Princeton University Press, 1997.
URL https://books.google.co.uk/books?id=lkeKhnqUHx8C

Appendix

A.1. List of Selected Companies

AAPL, AMZN, NFLX, MSFT, GS, GOOGL, BAC, JPM, IBM, DIS, GILD, INTC, YHOO, WMT, GE, XOM, SBUX, CSCO, WFC, NVDA, PCLN, JNJ, MCD, NKE, BA, VZ, ES, PFE, KO, CVX, CAT, MU, MRK, CELG, EBAY, MS, CRM, FCX, QCOM, TGT, HD, CHK, BMY, AMGN, PG, HPQ, ORCL, FSLR, WFM, COST, BIIB, PEP, EA, AXP, WYNN, CMCSA, CL, AIG, DOW, NEM, MA, BBY, COP, LOW, TWX, ADBE, HAL, LLY, UNH, LUV, MMM, CVS, MO, FDX, DD, ED, KR, MON, UTX, ABT, SLB, YUM, MCO, AMAT, EXPE, AET, DE, GPS, UPS, VLO, CBS, HAS, COH, ALL, WDC, JWN, TXN, PM, UNP, EOG.

A.2. Prediction Results Using Expanding Window Training Set

In this section, we report results using models that were trained in an expanding window, instead of a rolling window, with initial start and end dates of 09/05/2012 and 09/10/2014, respectively. Test period ranges from 09/17/2014 and 08/25/2017.

Table AA: **Financial Link Prediction Performance Results using an Expanding Window Training Set.** High out-of-sample AUCs obtained indicate that the model has high performance balancing both false positives and false negatives predictions relative to true positive and negative values. Likelihood ratios ($\lambda > 18.47$) demonstrate that models considering social media network fit the data significantly better than the restricted model that considers financial features only at a significance level of $p = 0.001$. Likelihood ratio increases with prediction lag indicating that social media features are particularly important for long-term prediction. AUC^* indicates performance improvement compared to a naive benchmark model that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. Results show that proposed model can achieve up to 27% improvement compared to the commonly-used assumption of stationarity in the correlation structure. Table show mean values obtained over the test period with corresponding standard deviation in parentheses. Models were trained in an expanding window with initial start and end dates of 09/05/2012 and 09/10/2014, respectively. Test period ranges from 09/17/2014 and 08/25/2017.

	E^+		E^-		G^F	
Lag	AUC	λ	AUC	λ	AUC	AUC^* (%)
1	87 (0.33)	21 (0.76)	93 (0.11)	34 (1.2)	97 (0.064)	4 (0.091)
2	87 (0.37)	33 (1.2)	93 (0.1)	45 (1.5)	95 (0.092)	6 (0.14)
3	86 (0.39)	48 (1.5)	93 (0.11)	60 (1.6)	94 (0.11)	8 (0.17)
4	86 (0.39)	65 (2)	93 (0.11)	65 (1.9)	93 (0.13)	10 (0.21)
5	85 (0.41)	85 (2.6)	93 (0.11)	66 (1.9)	92 (0.15)	11 (0.24)
6	85 (0.41)	100 (3.2)	93 (0.1)	74 (2)	91 (0.16)	12 (0.27)
7	84 (0.42)	120 (3.5)	93 (0.1)	70 (2.2)	90 (0.18)	13 (0.3)
8	84 (0.43)	150 (4.3)	93 (0.1)	72 (1.9)	89 (0.19)	15 (0.33)
9	83 (0.44)	180 (5.7)	93 (0.1)	74 (2.2)	88 (0.21)	16 (0.37)
10	83 (0.43)	220 (6.3)	93 (0.096)	79 (1.9)	87 (0.21)	17 (0.4)
11	82 (0.43)	260 (7.2)	93 (0.094)	78 (2)	87 (0.22)	18 (0.43)
12	82 (0.42)	300 (7.9)	93 (0.09)	86 (2.4)	86 (0.22)	19 (0.45)
13	82 (0.43)	330 (7.9)	93 (0.09)	95 (2.1)	85 (0.22)	20 (0.49)
14	81 (0.43)	360 (9.2)	93 (0.084)	100 (2.4)	84 (0.23)	21 (0.51)
15	81 (0.43)	390 (9.9)	93 (0.083)	110 (2.3)	84 (0.24)	22 (0.55)
16	81 (0.43)	410 (10)	93 (0.08)	120 (3)	83 (0.24)	23 (0.58)
17	80 (0.43)	440 (11)	94 (0.079)	130 (2.6)	82 (0.25)	24 (0.62)
18	80 (0.44)	470 (12)	94 (0.076)	150 (3)	82 (0.25)	25 (0.67)
19	80 (0.46)	500 (12)	94 (0.072)	160 (3.6)	81 (0.27)	26 (0.71)
20	80 (0.48)	510 (12)	94 (0.068)	170 (3.7)	80 (0.28)	27 (0.79)

*A likelihood ratio of $\lambda > 18.47$ indicates statistical significance at $p = 0.001$.

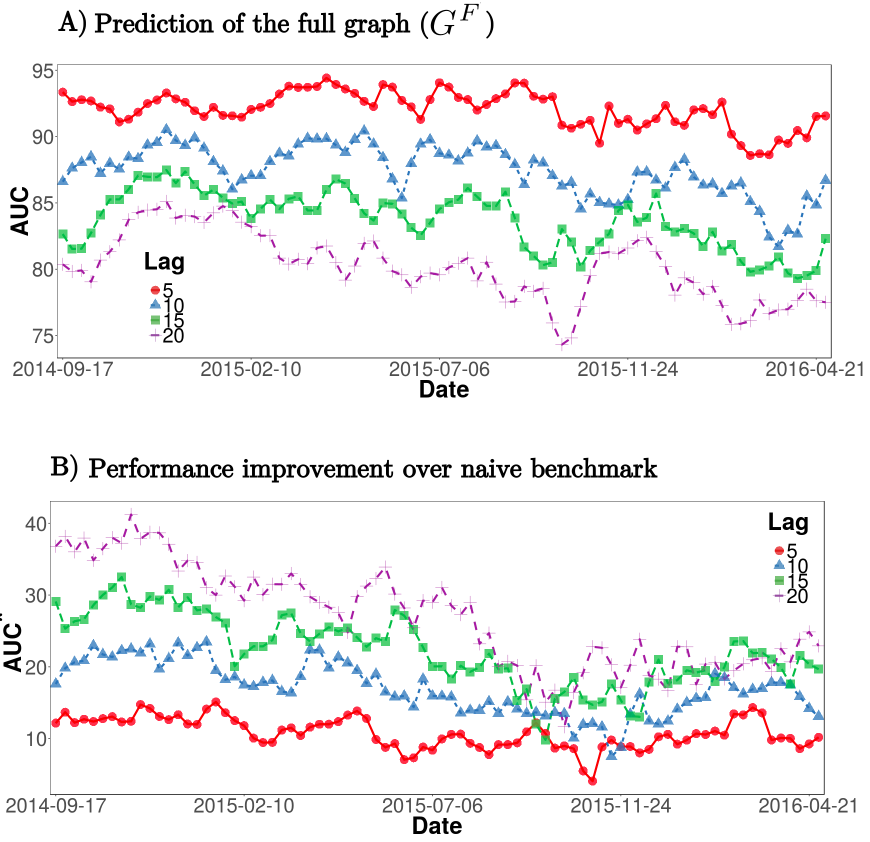

Figure 7: **Link prediction results using an expanding window training set - Evidence of high out-of-sample performance in financial network link prediction.** Models were trained in an expanding window with initial start and end dates 09/05/2012 and 09/10/2014, respectively. Test period ranges from 09/17/2014 and 08/25/2017. Plots show performance results (AUC), where each date (t) represents the performance obtained with a model trained with information up to time t and prediction of edges of a network at time $t + h$. Panel A) shows performance obtained in the prediction of out-of-sample edges for $h \in (1, 5, 10, 15, 20)$ trading weeks. Panel B) shows performance improvement compared to a naive benchmark that assumes that correlation structure is time-invariant, i.e., $G^F(t + h) = G^F(t)$.

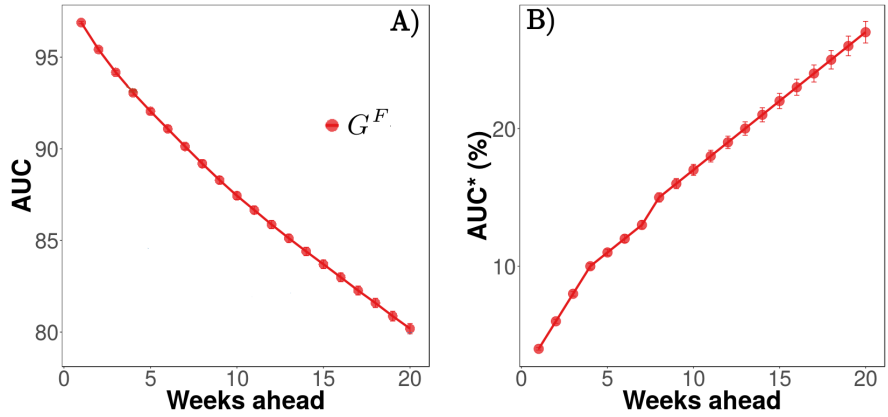

Figure 8: **Link prediction results using an expanding window training set - Effect of time-lag in out-of-sample predictive performance.** Plot shows mean performance (AUC) over the testing period in the prediction of out-of-sample new (E^+) and removed (E^-) edges as well as the prediction of the full financial network G^F. Error bars indicate standard error.

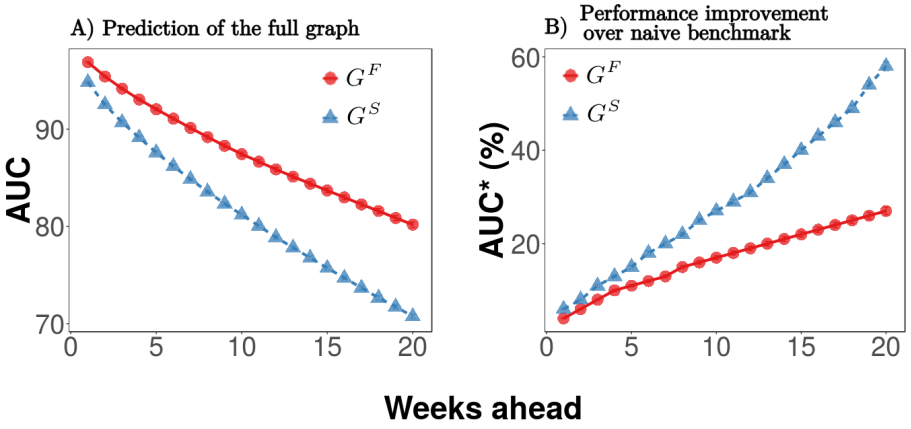

Figure 9: **Link prediction results using an expanding window training set - Evidence that financial market structure has higher predictability than social media structure.** Models were trained in an expanding window with initial start and end dates of 09/05/2012 and 09/10/2014, respectively. Test period ranges from 09/17/2014 and 08/25/2017. Panel A) shows mean performance (AUC) in the prediction of out-of-sample edges of the full financial network G^F and social network G^S. Panel B) shows the performance improvement against a naive benchmark that assumes that correlation structure is time-invariant. Error bars indicate standard error.