Physica A 447 (2016) 493–501

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# Finding modules and hierarchy in weighted financial network using transfer entropy

Soon-Hyung Yook *, Huiseung Chae, Jinho Kim, Yup Kim *

*Department of Physics and Research Institute for Basic Sciences, Kyung Hee University, Seoul 130-701, Republic of Korea*

## HIGHLIGHTS

- We construct the information transfer network based on the transfer entropy.
- We analyze the modular structure with various time resolutions.
- We compare the results with modular structure obtained from the cross correlations.
- We show that the transfer entropy provides a better modular structure with higher value of modularity.

## ARTICLE INFO

*Article history:*
Received 3 December 2014
Received in revised form 24 October 2015
Available online 28 December 2015

*Keywords:*
Complex networks
Financial networks

## ABSTRACT

We study the modular structure of financial network based on the transfer entropy (TE). From the comparison with the obtained modular structure using the cross-correlation (CC), we find that TE and CC both provide well organized modular structure and the hierarchical relationship between each industrial group when the time scale of the measurement is less than one month. However, when the time scale of the measurement becomes larger than one month, we find that the modular structure from CC cannot correctly reflect the known industrial classification and their hierarchy. In addition the measured maximum modularity, $Q_{max}$, for TE is always larger than that for CC, which indicates that TE is a better weight measure than CC for the system with asymmetric relationship.



## 1. Introduction

Recent development of network science has been provided very useful and comprehensive framework to investigate the interwoven connectivity patterns observed in a wide range of scientific disciplines from physics to biology and economics [1]. In many real networks such as social networks [2], brain networks [3], protein-interaction network [4], each node belongs to a module or community. The module is a group of nodes which form a tightly knit group with high density of within-group edges and a lower density of between-group edges [5]. Such modules or communities are mesoscale building blocks of complex networks, because they usually correspond to the fundamental functional blocks in a network. Therefore, classifying modules in a network has been a fundamental problem to understand the origin of the specific topological, functional, and dynamical properties of a network.

Most studies on the modular structure of a given network have been focused on the finding of an efficient algorithm from a given topological information. Examples include the modularity maximization [5], clique percolation [6], and spectral analysis of the non-backtracking matrix [7]. Due to the inherent complexity, developing a more efficient model algorithm is still an open problem in network science. Besides finding the efficient algorithm, uncovering the relationship between

* Corresponding authors.
*E-mail addresses:* syook@khu.ac.kr (S.-H. Yook), ykim@khu.ac.kr (Y. Kim).

http://dx.doi.org/10.1016/j.physa.2015.12.018
0378-4371/

the given modules is also an important quest to understand the organization of complex systems. Especially, the hierarchy between the modules in a network is one of the important and pervasive features of the organization of natural and artificial systems out of equilibrium [4,8–10]. Thus, finding the hierarchical relationship between modules potentially provides significant insight into the central aspects governing the physical properties of networks and their functionality.

There is an additional difficulty in finding modules and their relationship in many real networks. Many real networks are well described by the weighted networks in which each link is associated with a weight [11]. Examples of weighted networks include scientist collaboration network and airport network [12]. Even though the topological definition of the modularity for the weighted networks can be easily extended from that for the unweighted network [13], finding a good measure for weight of each link is not a trivial problem. Therefore, in order to understand the dynamical and topological properties of such weighted networks, it is very important to find more informative weight measure for network analysis of various systems.

One widely used measure for the weight is the cross correlation (CC), which is usually assumed to be symmetric [14–21]. For example, in financial system, Mantegna introduced a method to find a hierarchical arrangement of the stocks based on the CC of asset returns [14]. By defining an appropriate metric, they constructed the minimum spanning tree (MST) from the fully connected weighted graph and identified the clusters of companies. More recently, the study on the time dependent properties of CC distribution and the dynamic asset tree showed quantitative differences between the crash and the normal periods [21].

However, in many real complex systems, the relationship between each unit is not necessarily symmetric. One of important factors for such asymmetry is the causality. The causality in complex system was usually measured by the lagged CC [22], Granger causality [23], and the time-delayed mutual information [24]. The lagged CC is intuitive and simple measure for the asymmetric interaction between each unit in complex systems. By using the lagged CC, Kullmann et al. constructed a weighted directed network and quantitatively showed that there is some pulling effect between companies in financial system [22]. The causality network between global market indices based on the Granger causality was also studied [25]. Time-delayed mutual information provides more general and intuitive measure for the dependence between random variables. But it was recently shown that the mutual information does not explicitly distinguish the actually exchanged information due to a common history or input signal [26]. As an alternative measure of the information transfer, the transfer entropy (TE) was introduced to exclude such undesired influences [26]. In financial systems, such as global market indices, the causality measured by TE between the market indices is well represented by the weighted directed edges [27].

In this paper, to investigate how useful TE is as a weight measure for financial system, we consider the information transfer network (ITN), in which TE is used as the weight measure, and analyze the modular structure. The modular structures of ITN are compared with those of correlation network (CN) which uses the cross correlation to determine weight between companies. From the comparison, we find that the modules of both ITN and CN are consistent with the well known industrial classification [28] when the time scale of the measurement is small. However, if the time scale becomes larger, then the modules in CN significantly deviate from the known industrial classification. In addition, the measured maximum modularity, $Q_{\max}$, of ITN is always larger than that of CN, which indicates that TE is a better weight measure than CC for the systems in which the asymmetric relationship between each unit becomes important.

## 2. Data set and definition of states

In order to study modular structure of the financial network and their hierarchical relationship, we use the Standard & Poor's (S&P) 100 data traded from 03/01/1962 to 03/12/2010 [29]. From the obtained time series of S&P 100 index, we first define the state, $i_t$, of company $I$ at day $t$ to calculate TE. As the simplest choice of $i_t$ for a company $I$ we consider the binary state, i.e. $i_t = 1\,(0)$ if $Y_I(t+\Delta t) \geq Y_I(t)$ ($Y_I(t+\Delta t) < Y_I(t)$), where $Y_I(t)$ denotes the stock price of company $I$ at time $t$. Thus $i_t$ simply represents the increase (decrease) of price if $i_t = 1\,(i_t = 0)$.

## 3. Transfer entropy and cross correlation

Let $i_t$ ($j_t$) be the state of company $I$ ($J$) at time $t$. TE which represents the information flow from $J$ to $I$ is defined as [26]

$$T_{J\to I} = \sum p(i_{t+1}, i_t^{(k)}, j_t^{(\ell)}) \log_2 \frac{p(i_{t+1}|i_t^{(k)}, j_t^{(\ell)})}{p(i_{t+1}|i_t^{(k)})}. \tag{1}$$

Here we use the shorthand notation $i_t^{(k)} = (i_t, \ldots, i_{t-k+1})$. The sum in Eq. (1) represents the sum over all available realization of state $(i_{t+1}, i_t^{(k)}, j_t^{(\ell)})$ in a time series. $p(i_{t+1}, i_t^{(k)}, j_t^{(\ell)})$ is the joint probability that the combination of $i_{t+1}$, $i_t^{(k)}$ and $j_t^{(\ell)}$ has a particular value, and $p(i_{t+1}|i_t^{(k)}, j_t^{(\ell)})$ is the conditional probability that $i_{t+1}$ has a particular value when the values of the previous samples $i_t^{(k)}$ and $j_t^{(\ell)}$ are given. $k$ and $\ell$ in Eq. (1) are set as $k = \ell = 1$ [26]. In ITN the weight from a company $J$ to $I$ is assigned as $w_{JI} = T_{J\to I}$.

For a comparison we use CC as a weight between nodes to construct CN. CC between node $I$ and $J$, $G_{IJ}$, is defined as [15–20]

$$G_{IJ} = \frac{\langle R_I R_J\rangle - \langle R_I\rangle \langle R_J\rangle}{\sqrt{\sigma_I^2 \sigma_J^2}}. \tag{2}$$

Here $R_I \equiv \ln Y_I(t+\Delta t) - \ln Y_I(t)$ is the return and $\sigma_I^2 \equiv \left\langle R_I^2 - \langle R_I \rangle^2 \right\rangle$. The CN is obtained by assigning the weight from $J$ to $I$ as $w_{JI} = G_{JI}$. Since $G_{IJ}$ in Eq. (2) is symmetric, $w_{IJ} = w_{JI} = G_{IJ}$.

## 4. Modularity

To find the modular structure we use the modularity maximization method introduced in Ref. [5]. The modularity for a weighted directed network of size $N$ is defined as [30],

$$Q = \frac{1}{2M} \sum_{I,J} \left[ w_{IJ} + w_{JI} - \frac{w_I^{out} w_J^{in}}{M} - \frac{w_J^{out} w_I^{in}}{M} \right] \delta_{c_I, c_J}. \tag{3}$$

Here $c_I$ represents the community including the company $I$. $w_{IJ}$ is the weight from a company $I$ to $J$ and $w_I^{out} = \sum_{J=1}^{N} w_{IJ}$ ($w_I^{in} = \sum_{J=1}^{N} w_{JI}$) is the total weight from (to) the node $I$. $M = \sum_{I=1}^{N} \sum_{J=1}^{N} w_{IJ}$ is the total weight of networks. To find the modular structure, we use the agglomerative hierarchical clustering method [5]. The algorithm is as what follows. Start with a state in which each vertex is the sole member of one of $N$ communities. We repeatedly join communities together in pairs. The pairs are chosen at each step to maximize the increase in $Q$. The progress of the algorithm can be represented as a dendrogram, a tree showing the order of the joins. Different levels of cuts through this dendrogram give different levels of modules or communities. We can select the best cut by looking for the maximal value of $Q$.

## 5. Results

In Fig. 1(a), the dendrogram obtained from ITN with $\Delta t$ = 1-week is displayed. At each level of dendrogram, we calculate $Q$ and find that the maximum of $Q$ for $\Delta t$ = 1-week is $Q_{\max} = 0.1194$. When $Q = Q_{\max}$ (represented by black lines) we obtain four distinctive modules. Each module at $Q = Q_{\max}$ is composed of technology related companies, financial/goods distribution system/service related companies, material companies, and utility/health care/consumer goods related companies, respectively. These four distinctive modules can be divided into the small groups as shown in Fig. 1(a). Each small group agrees well with the known classification of the industries and shows that there exists a clear hierarchy between them. For example, the companies in the "material" group have strong tendency to be connected together at the early stage of the algorithm (see Fig. 1(b)). The second lowest level of group is composed of the companies in the "utilities" as shown in Fig. 1(c). The "utilities" group is combined with a small "technology" group. This small technology group is composed of two communication companies. The third lowest level of group is mainly composed of the "financial" companies as shown in Fig. 1(d). Most of the "technology" companies such as IBM and Intel belong to the fourth lowest level of group as shown in Fig. 1(e). At the higher level of hierarchy the "financial" group is merged with the companies of "consumer goods", "industrial goods", and "service" groups. Especially, the companies in the consumer goods and the industrial goods in this module are related to the goods distribution system such as Costco, Fedex, and UPS. As we repeat the algorithm, the groups of "consumer goods", "health care" and "utility" are merged into a single module at $Q = Q_{\max}$. The different level of modules clearly shows the hierarchy between industrial groups. Specifically, the companies which are related to the raw material industry have strong tendency to form a module at the lower level of cut. For the intermediate level, the companies in the financial/goods distribution system and the technology groups make their own modules. On the other hand, the companies in the consumer packaged goods such as health care and consumer goods make a single module at the higher level of cut.

In Fig. 2 the dendrogram obtained from CN with $\Delta t$ = 1-week is displayed. When CC is used for the weight on each directed edge, three distinctive modules are obtained at $Q_{\max} = 0.057$. In CN the companies in the material group are first connected together as in the case of ITN in Fig. 2(b). As we repeat the algorithm, the companies in the utility group are combined to make a small group (Fig. 2(c)). This utility group is connected with the material group. Then the financial and technology groups are formed as shown in Fig. 2(d) and (e). At the higher level of cut, the companies in the health care and the consumer goods make a single module. Even though the technology group and the financial group are not separated in CN at $Q = Q_{\max}$, CN with $\Delta t$ = 1-week shows a similar hierarchy and modular structure with ITN. Note that the obtained value of $Q_{\max}$ for CN is lower than that for ITN. This indicates that the companies in the same module for CN are more loosely connected than those for ITN.

Fig. 3 shows the dendrogram obtained from ITN with $\Delta t$ = 1-month. As shown in Fig. 3 we find that there are three distinctive modules at $Q = Q_{\max}$. In this case, the technology group and material group belong to the same module at $Q = Q_{\max}$. Fig. 3(b)–(e) show the small groups obtained from the five lowest level of cuts. Like the ITN with $\Delta t$ = 1-week, the companies in the material group are connected together at the lowest level of cut (Fig. 3(b)). Then the utility group and the financial group are formed, respectively (Fig. 3(c) and (d)). As we repeat the algorithm, the technology group is formed and merged with material group (Fig. 3(e)). Thus the modular structure and their hierarchy of ITN with $\Delta t$ = 1-month are almost the same as ITN with $\Delta t$ = 1-week.

In Fig. 4 we display the dendrogram for CN with $\Delta t$ = 1-month. At $Q = Q_{\max}$ we find two distinctive modules. However, we can find only three relatively well ordered small groups, material group (Fig. 4(b)), technology group (Fig. 4(c)), and financial group (Fig. 4(d)). The companies in the other groups are widely scattered and mixed together without any specific

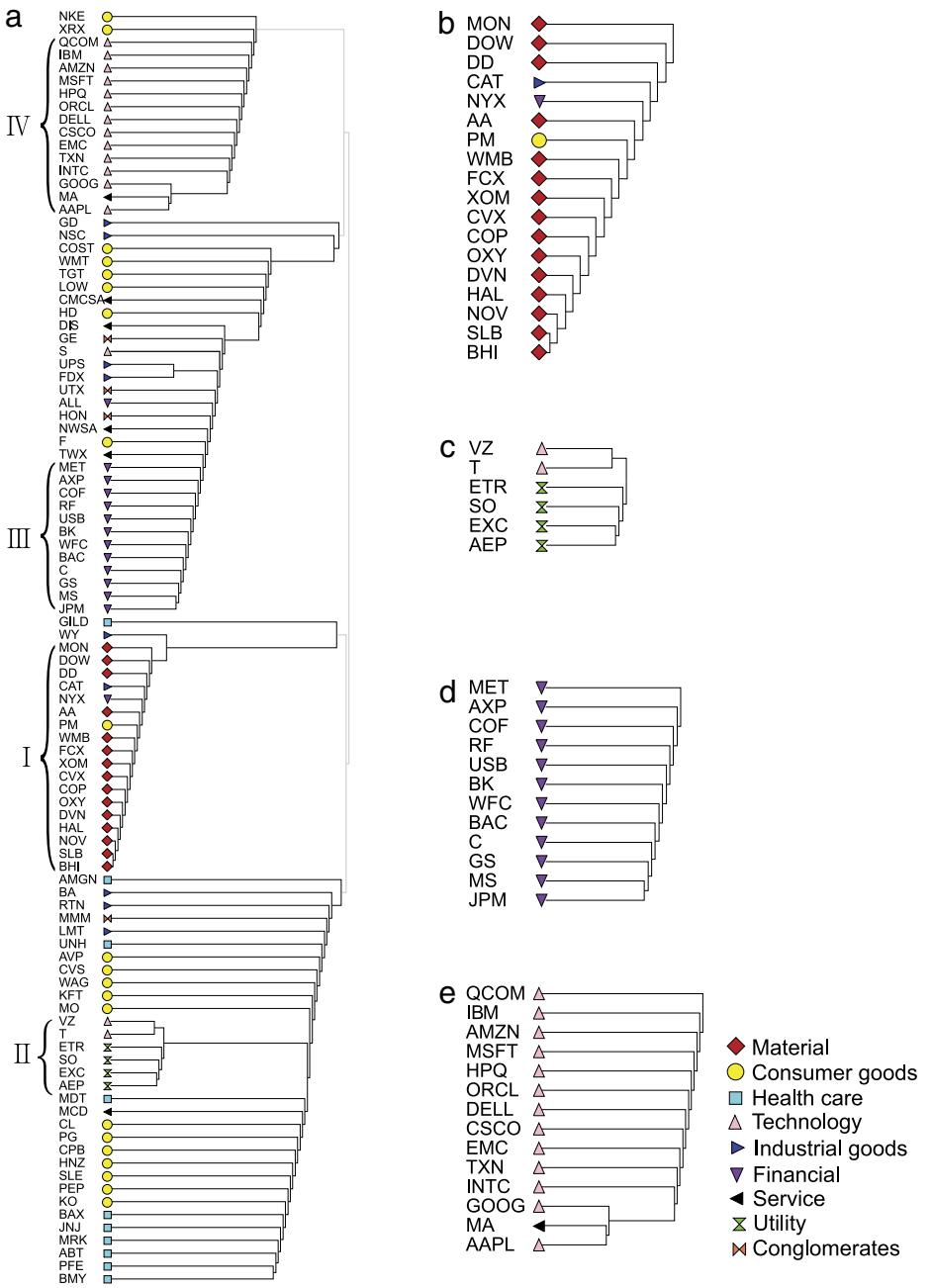


**Fig. 1.** (Color online) (a) Dendrogram obtained from ITN with $\Delta t = 1$-week. The modules and hierarchy represented by the black lines correspond to the case of $Q = Q_{\max}$. The gray lines at the top of the dendrogram denote the higher level of the hierarchy with $Q < Q_{\max}$. Enlarged plot of (b) regime I—material group, (c) regime II—utilities and technology groups, (d) regime III—financial group, and (e) regime IV—technology group.

order (see Fig. 4(a)). This indicates that when $\Delta t$ becomes large, the modules at higher level of cut cannot be correctly detected if CC is used as a weight measure. This inaccuracy in finding module with CC is originated from the absence of long-term correlation in financial market [31].

When $\Delta t = 3$-months, we still find three different modules in ITN at $Q = Q_{\max}$ with relatively well ordered subgroups as shown in Fig. 5(a). However, we cannot find any meaningful modular structure in CN with $\Delta = 3$-months, even though the companies in the utility, the material and the health care groups have tendency to be connected together as in Fig. 5(b).

For a quantitative analysis of the modular structures obtained from ITN and CN, we compare the value of $Q_{\max}$ for each network. In Fig. 6 we display the measured $Q_{\max}$ for various values of $\Delta t$. The data in Fig. 6 clearly shows that $Q_{\max}$ obtained from ITN is almost twice as large as that from CN for all $\Delta t$. From the definition of $Q$ in Eq. (3), high value of $Q$ corresponds to

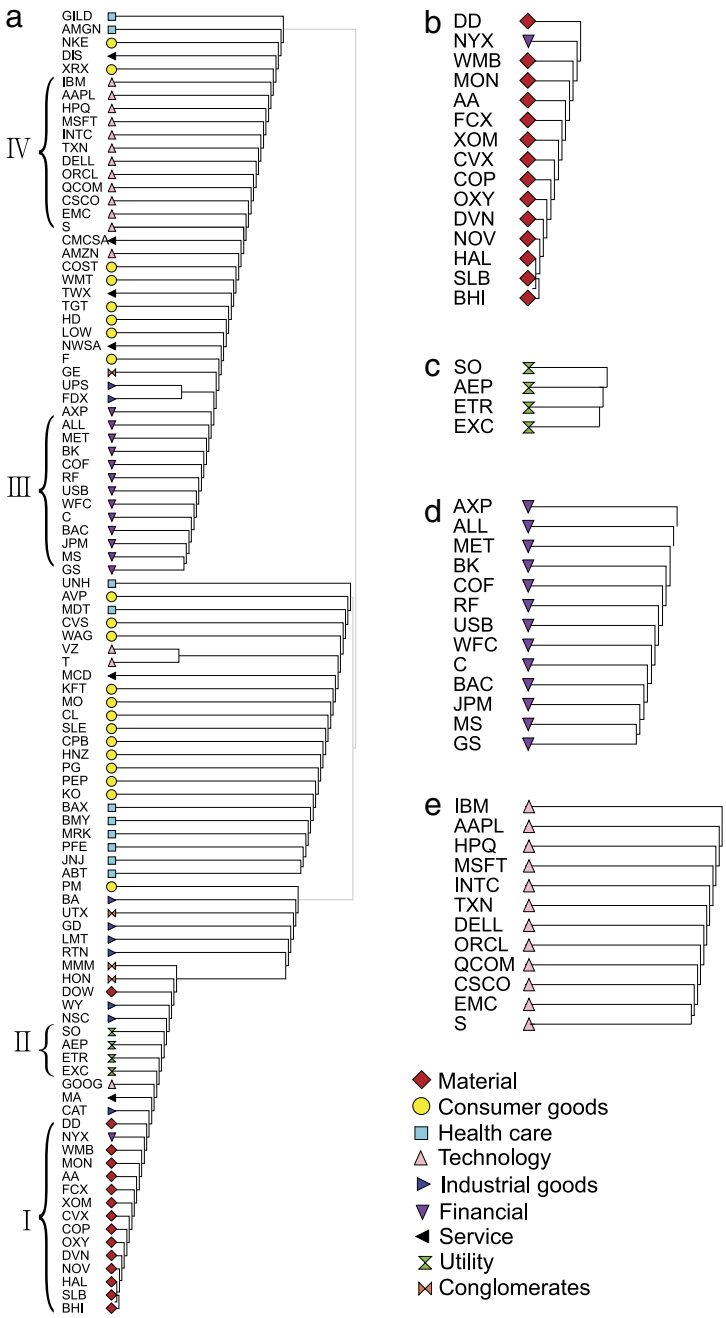


**Fig. 2.** (Color online) (a) Dendrogram obtained from CN with $\Delta t$ = 1-week. The modules and hierarchy represented by the black lines correspond to the case of $Q = Q_{max}$. The gray lines at the top of the dendrogram denote the higher level of hierarchy with $Q < Q_{max}$. Enlarged plot of (b) regime I—material group, (c) regime II—utility group, (d) regime III—financial group, and (e) regime IV—technology group.

the good division of networks into modules. Thus the result clearly shows that TE is a better weight measure than CC for the systems with the asymmetric relationship such as causality. In order to see how important the consideration of asymmetric relationship is, we also construct the symmetric rank correlation network (RCN) using the Kendall's $\tau_B$ [32]. Kendall's $\tau_B$ measures a symmetric rank correlation. As shown in Fig. 6, even though $Q_{max}$ from RCN for $\Delta t$ = 1-day is comparable with $Q_{max}$ from ITN, $Q_{max}$ from RCN becomes almost the same with that from CN when $\Delta t$ becomes larger than 1-day.

We also test how the obtained modules in each network are well divided through measuring the overlapping communities (or modules). To find the overlapping community in a weighted directed network, we use the algorithm suggested by Chen et al. [33]. In this algorithm the obtained module at $Q = Q_{max}$ for each network is given as the

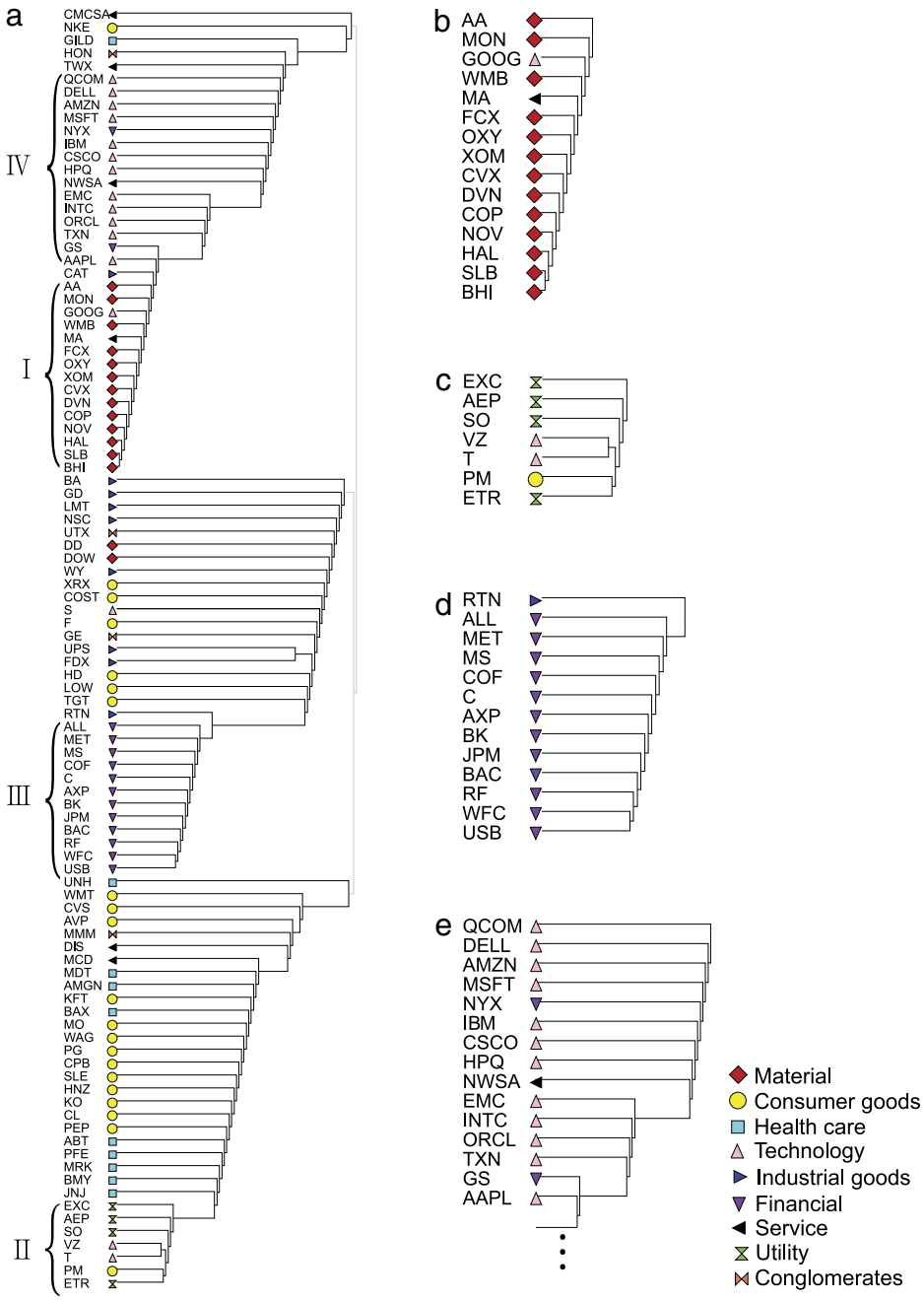


**Fig. 3.** (Color online) (a) Dendrogram obtained from ITN with $\Delta t = 1$-month. The modules and the hierarchy represented by the black lines correspond to the cut at $Q = Q_{\max}$. The gray lines at the top of the dendrogram denote the higher level of hierarchy with $Q < Q_{\max}$. Enlarged plot of (b) regime I—material group, (c) regime II—technology group, (d) regime III—financial group, and (e) regime IV—technology group.

initial module. The overlapping communities can be obtained by the following community expanding procedure. For each company $I$ we define the total weight $w_I$ as $w_I = \sum_{J=1}^{N}(w_{IJ} + w_{JI})$. For a company $I$ in a module $c$, the belonging degree $B(I, c)$ is defined as $B(I, c) = \sum_{J\in c}(w_{IJ} + w_{JI})/w_I$. Starting from the initial modules, (i) find all neighbors $N_c$ of the initial module $c$. (ii) For a given threshold of belonging degree $B^*$, if a company $I \in N_c$ satisfies the condition $B(I, c) > B^*$, then add $I$ into the module $c$ directly. (iii) Repeat (i)–(ii) until $B(I, c) \leq B^*$ for all $I \in N_c$. The procedures (i)–(iii) are also applied to all modules (see Ref. [33] for details). At the end of the community expanding procedure, some companies can belong to several modules, which are defined as the overlapping community. For various values of $B^*$ we find that the number of overlapping community in ITN is much smaller than that in CN for all $\Delta t$. For example, 56 companies belong to the

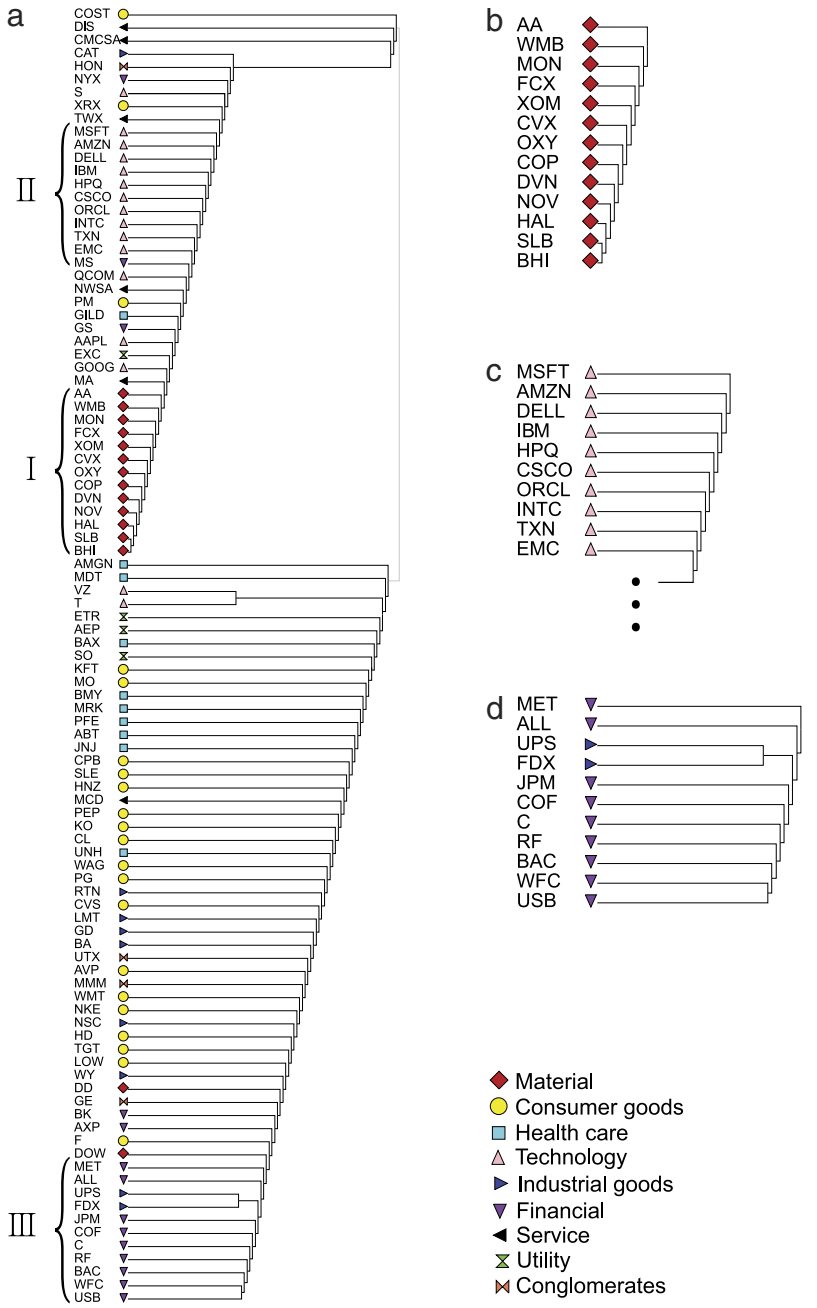


**Fig. 4.** (Color online) (a) Dendrogram obtained from CN with $\Delta t = 1$-month. The modules and the hierarchy represented by the black lines correspond to the cut at $Q = Q_{\max}$. The gray lines at the top of the dendrogram denote the higher level of hierarchy with $Q < Q_{\max}$. Enlarged plot of (b) regime I—material group, (c) regime II—technology group, (d) regime III—financial group.

overlapping community in CN while there is no overlapping community in ITN at $B^* = 0.4744$ with $\Delta t = 1$-week. This result also clearly shows that the modular structure in CN is much fuzzier than that of ITN.

## 6. Summary

In summary, we study the modular structure of financial market and the hierarchical relationship between them. We use two different physical quantities to assign the weight on each edge, TE and CC. Using the agglomerative hierarchical clustering algorithm, we find that the modular structures of ITN and CN are similar if $\Delta t < 1$-month. However, for large values of $\Delta t$ ($>1$-month) ITN still preserves the modular structure and their hierarchical relationship, while CN does not

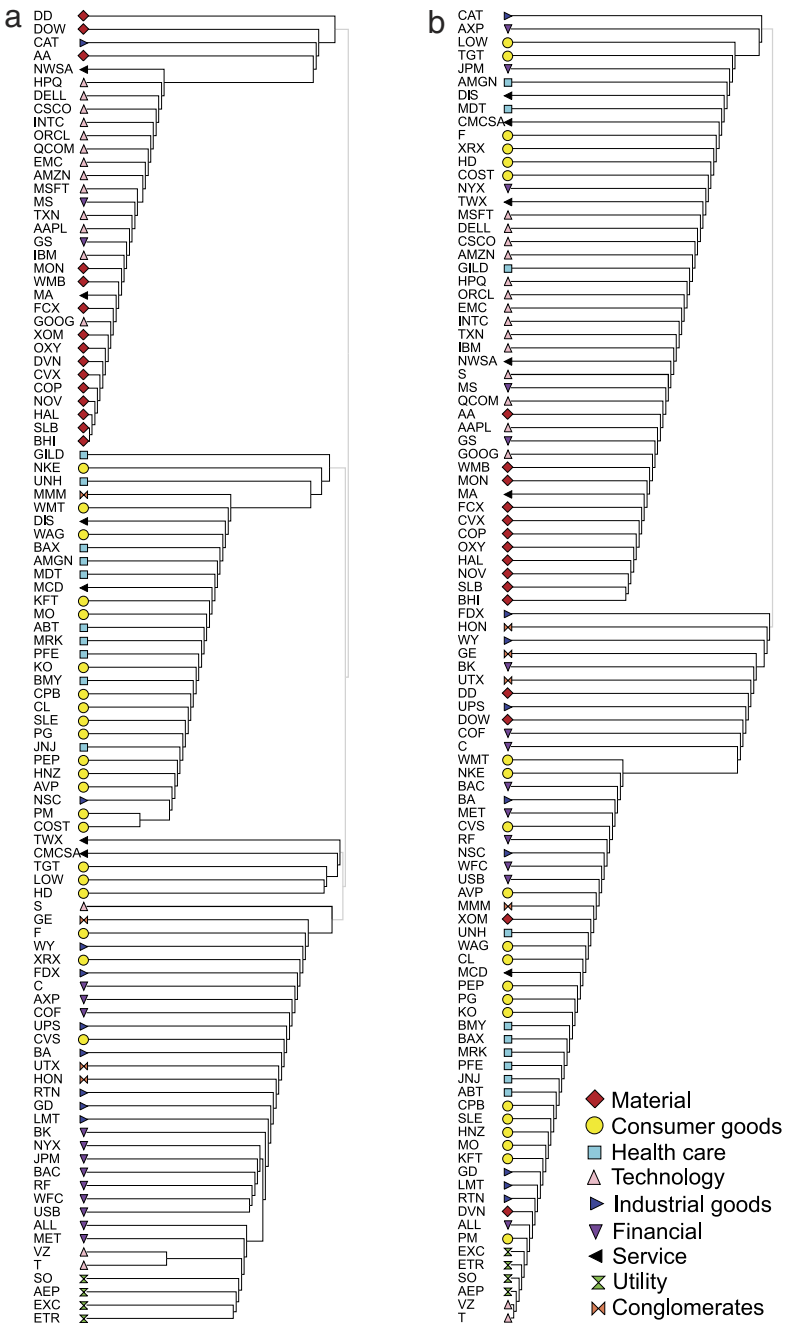


**Fig. 5.** (Color online) Dendrogram obtained from (a) ITN and (b) CN with $\Delta t = 3$-months. The modules and the hierarchy represented by the black lines correspond to the cut at $Q = Q_{\max}$. The gray lines at the top of the dendrograms denote the higher level of hierarchy with $Q < Q_{\max}$.

correctly reflect the known industrial classification and their hierarchy. In addition, we also find that the value of $Q_{\max}$ for ITN is always larger than that for CN, which implies that TE is a better weight measure than CC when the relationship between nodes is not symmetric.

## Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2013R1A1A2057791 and NRF-2012R1A1A2007430).

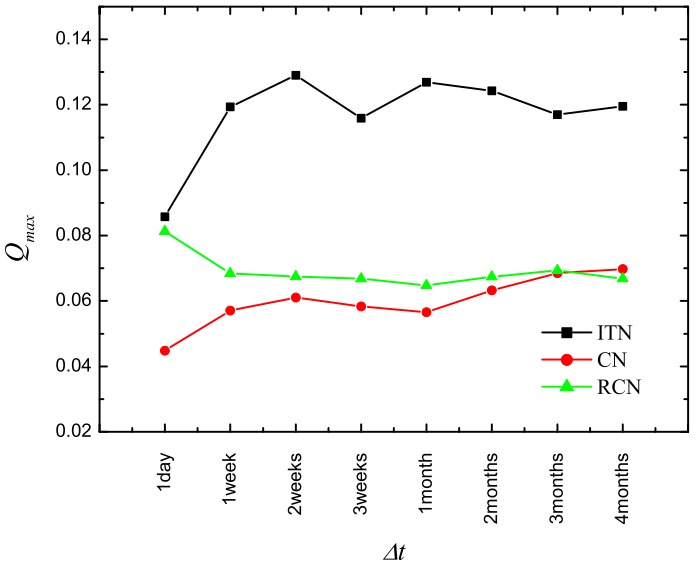


**Fig. 6.** Comparison of $Q_{\max}$ obtained from ITN, CN, and RCN.

## References

[1] M.E.J. Newman, Networks: An Introduction, Oxford University Press, New York, 2010.
[2] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. 99 (2002) 7821.
[3] M. Chavez, M. Valencia, V. Navarro, V. Latora, J. Martinerie, Phys. Rev. Lett. 104 (2010) 118701.
[4] S.-H. Yook, Z. Oltvai, A.-L. Barabási, Proteomics 4 (2003) 928.
[5] M.E.J. Newman, Phys. Rev. E 69 (2004) 066133.
[6] I. Derényi, G. Palla, T. Vicsek, Phys. Rev. Lett. 94 (2005) 160202.
[7] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, P. Zhang, Proc. Natl. Acad. Sci. 110 (2013) 20935.
[8] M.H.R. Amaral, C.H. Loch, D. Wilkinson, B.A. Huberman, Nature 379 (1996) 831.
[9] E. Ravasz, A.L. Somera, D.M. Mongru, Z.N. Oltvai, A.-L. Barabási, Science 297 (2002) 1551.
[10] C. Song, S. Havlin, H.A. Makse, Nat. Phys. 2 (2006) 275.
[11] S.-H. Yook, H. Jeong, A.-L. Barabási, Y. Tu, Phys. Rev. Lett. 86 (2001) 5835.
[12] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, Proc. Natl. Acad. Sci. 101 (2004) 374.
[13] M.E.J. Newman, Phys. Rev. E 70 (2004) 056131.
[14] R.N. Mantegna, Eur. Phys. J. B 11 (1999) 193.
[15] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, Physica A 287 (2000) 374.
[16] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, Phys. Rev. Lett. 83 (1999) 1467.
[17] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, H.E. Stanley, Phys. Rev. E 65 (2002) 066126.
[18] A. Sensoy, S. Yuksel, M. Erturk, Physica A 392 (2013) 5027.
[19] S. Kumar, N. Deo, Phys. Rev. E 86 (2012) 026101.
[20] A. Nobi, S.E. Maeng, G.G. Ha, J.W. Lee, J. Korean Phys. Soc. 62 (2013) 569.
[21] J.-P. Onnela, A. Chakraborti, K. Kasik, J. Kertész, A. Kanto, Phys. Rev. E 68 (2003) 056110.
[22] L. Kullmann, J. Kertész, K. Kaski, Phys. Rev. E 66 (2002) 026125.
[23] C.W.J. Granger, Econometrica 37 (1969) 424.
[24] C.E. Shannon, W. Weaver, The Mathematical Theory of Information, University of Illinois Press, Urbana, 1949.
[25] T. Výrost, S. Lyócsa, E. Baumöhl, Physica A 427 (2015) 262.
[26] T. Schreiber, Phys. Rev. Lett. 85 (2000) 461.
[27] Y. Kim, J. Kim, S.-H. Yook, Physica A 430 (2015) 39.
[28] http://www.stockmaven.com.
[29] http://finance.yahoo.com.
[30] Y. Kim, S.W. Son, H. Jeong, Phys. Rev. E 81 (2010) 016103.
[31] R.N. Mantegna, H.E. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, Cambridge, 2000.
[32] A. Agresti, Analysis of Ordinal Categorical Data, John Wiley & Sons, New York, 2010.
[33] D. Chen, M. Shang, Z. Lv, Y. Fu, Physica A 389 (2010) 4177.