Mapping stochastic processes onto complex networks

A. H. Shirazi,[1] G. Reza Jafari,[2,3] J. Davoudi,[4] J. Peinke,[5] M. Reza Rahimi Tabar,[5,6] and Muhammad Sahimi[7]

[1] *Students' Scientific Research Center, Tehran University of Medical Sciences, Tehran, Iran*

[2] *Department of Physics, Shahid Beheshti University, G.C., Evin, Tehran 19839, Iran*

[3] *IFISC, Instituto de Fisica Interdisciplinar y Sistemas Complejos (UIB-CSIC), Campus UIB, 07122 Palma de Mallorca Spain*

[4] *Department of Atmospheric Physics, University of Toronto, Toronto, Canada*

[5] *Carl von Ossietzky University, Institute of Physics, D-26111 Oldenburg, Germany*

[6] *Department of Physics, Sharif University of Technology, Tehran 11155-9161, Iran*

[7] *Mork Family Department of Chemical Engineering & Materials Science, University of Southern California, Los Angeles, California 90089-1211, USA*

Abstract - We introduce a method by which stochastic processes are mapped onto complex networks. As examples, we construct the networks for such time series as those for free-jet and low-temperature helium turbulence, the German stock market index (the DAX), and the white noise. The networks are further studied by contrasting their geometrical properties, such as the mean-length, diameter, clustering, average number of connection per node. By comparing the network properties of the investigated original time series with those for the shuffled and surrogate series, we are able to quantify the effect of the long-range correlations and the fatness of the probability distribution functions of the series on the constructed networks. Most importantly, we demonstrate that the time series can be *reconstructed* with high precisions by a simple random walk on their corresponding networks.

There has been much recent interest in the study of complex networks and their applications in a variety of fields, ranging from computer science and communications, to sociology and epidemiology [1, 2, 3]. Graph theory provided the mathematical foundations for the study of complex networks, which began in the 18th century with Euler's solution to the famous *Bridges of Konigsberg* problem. The work of Erdös and Renyi [4, 5] in the 1950s on the theory of random graphs has also been very influential. A network - or a graph - is simply a collection of nodes or vertices connected by links or edges. The links may be directed or undirected, and weighted or unweighted. Much recent research has shown that many, and perhaps most, natural or even man-made phenomena may be usefully and fruitfully described in terms of networks and their properties. The brain, for example, is a huge network of neurons linked by synapses [6].

In parallel with the recent work on complex networks, there has also been much interest in, and work on, analyzing nonlinear dynamical and stochastic processes which fluctuate widely and contain correlations that may be very extended. Such processes occur in many natural and man-made phenomena, ranging from various indicators of economic activities, such as the stock market, to velocity fluctuations in turbulent flows, heartbeat dynamics, and many other phenomena [7, 8].

It is of interest and practical importance to develop frameworks that may connect the two different representations of complex systems, namely, networks and widely fluctuating stochastic processes. There have already been some attempts to map some random processes onto equivalent network representations. For example, spatial networks that represent complex dynamics in high-dimensional state spaces [9, 10, 11], as well as networks of pseudo-periodic time series [12], have been constructed.

In this paper we propose a general method by which a given stochastic process is mapped onto a complex network with distinct geometrical properties. The relation between the statistical properties of the stochastic time series, such as the intermittency and correlation length, and their stochastic behavior, as well as the properties of their equivalent networks are then studied. We show that a network equivalent of a stochastic time series enables us to address a central question in the field of analysis of the data for complex systems: Given a fluctuating, sequentially-measured set of experimental data, how is it possible to *reconstruct* the time series, representing the data, with high precision? The reconstruction is essential to making predictions for the *future*. We show that for such processes the corresponding time series can be reconstructed with high precision by a random walk on the network equivalent. As examples, we construct complex networks for several distinct time series, such as

those for the free-jet turbulence, financial markets (the German stock market index, the DAX), and the white noise.

As is well-known, a given process with a degree of randomness or stochasticity may have a finite, or even an infinite, Markov-Einstein (ME) time (or length) scale. The ME time (length) scale is the minimum time interval (length scale) over which the data can be considered as a Markov process [8]. To ascertain such a property, the data are first examined to see whether they follow a Markov chain and, if so, their ME time (length) scale T is estimated [12, 13, 14, 15, 16]. We note that the correlation and the ME scales are two independent scales in a time or spatial series [17]. For example, for stochastic series that describe the properties of turbulent flows, which typically contain long-range correlations, the ME length scale is the Taylor length scale [18]. The time series that we study in this paper have finite ME time scales T.

The determination of the ME scale T of a stochastic series is done based on the least-squares method and the likelihood statistical analysis [17, 19], by computing numerically the three-point probability density function (PDF) of the series, and comparing the results with what is expected for a Markov process. Then, the least-squares method is used to minimize the difference between the two, in order to delineate the Markov properties of the series, and in particular the ME T. The details of the method has been described in our previous publications [17, 19] and, therefore, need not be repeated here.

Now let us consider a discrete stationary process $x(t)$, with unit ME time scale $T = 1$ (in units of the data lag). If M is the number of bins needed to precisely represent, or evaluate, the PDF of $x(t)$, the data are partitioned into M bins, with each bin having an equal number of data points. Each bin is represented by a node in the equivalent complex network of the series, which would have M nodes. When the PDF of the data is not Gaussian, we use the surrogate technique to transfer it to a Gaussian function, while preserving the correlations in the data. Increasing the number of nodes (bins) also increases the resolution of the network, but also the statistical errors. The statistical errors increase due to the reduction of the number of data points in each bin. Thus, there is an optimal number of nodes (bins) that has the highest resolution, such that the statistical errors do not eliminate any meaningful information. The optimal number of the bins is estimated by minimizing the distance (difference) of the PDF of the surrogated time series from (with) the Gaussian PDF. This is done by the chi-square test.

Nodes i and j are linked if as the time increases the value of $x(t)$ in bin i changes to that in bin

j in one time step. A weight w_{ij} is attributed to a link ij, which is the number of times that a given value of $x(t)$ changes from its value in bin i to bin j in one time step, and is normalized at each node. The transition matrix $\mathbf{W} = [w_{ij}]$ is, in general, not symmetric. Thus, the network is both *directed* and *weighted*. For the time series with a finite ME time scale $T > 1$, one can construct the transition matrix with entries given by $p[x(t)|x(t - T)]$, and attribute to each node a set of data arrays with length T (see below). In order to clarify the process of generating the network equivalents of the series, we show in Fig. 1 the schematic network made from a simple series.

We constructed the equivalent complex networks for three series, namely, those for the free-jet turbulence (FJT), the German financial market (the DAX), and the white noise (WN) time series. The FJT data were for a fluid velocity, measured at various times at a fixed point in the flow system. We also computed the ME time scales for the three series. Figure 2 presents the three network equivalents of the three series. Each network was constructed using 3×10^6 normalized data points in each series, and contains about 10^3 nodes, with the distance of each node from the network's center being proportional to its number, meaning that the nodes with higher number of links are located at larger distances from the network's center. The intensity of the colors associated with the links is proportional to their weight w_{ij}.

In Table I the networks' characteristics, such as the mean link length, clustering coefficient, diameter, average and maximum number of outgoing and incoming links (the nodes' degree), $\langle k_{\text{out}}^2 \rangle / \langle k_{\text{out}} \rangle$, and $\langle k_{\text{in}}^2 \rangle / \langle k_{\text{in}} \rangle$ are presented, where k is the nodes' degree or connectivity. The mean length for a directed weighted network was evaluated with Dijkstra's algorithm [20], which accounts for the number of nodes in a probable path between every two edges. The clustering coefficient was calculated by counting the normalized number of connected triplet of vertices [20], i.e.,

$$C_i = \frac{A_i}{B_i} = \frac{\sum_{j,k} w_{ij} w_{ki} w_{jk}}{\sum_{j,k} w_{ij} w_{ki}} \ . \tag{1}$$

The clustering coefficient has a clear meaning in stochastic processes with Markov properties. A_i is the probability of returning to state i after three ME time scales, i.e., $A_{=} w_{ii}(3T)$, where we have used the stationarity of the time series and the Chapman–Kolmogorov (CK) equation for the Markov series. Note that w_{ij} is a conditional PDF that, for a process with the ME times scale $T = 1$, satisfies the CK equation, whereas for $T > 1$ it satisfies the CK equation with, $P(i, t + t_M|j, t) = w_{ij}$. The CK equation states that $w_{ij}(2T) = \sum_k w_{ik} w_{kj}$. The coefficient B_i is also simplified because of the nature of w_{ij} as the conditional probabilities. Indeed, $\sum_j w_{ij}(T) = 1$, due to the normalization, and

$\sum_k w_{ki}(T) = p[x(t + T) = i]$ is the probability of being in a state i in one step of the size of the ME time scale. Note also that the latter sum is an unconditional probability measure of being in state i.

In terms of the network structure and properties, the DAX time series exhibits the interesting feature that, its mean length and diameter are approximately equal to the corresponding values for the WN network. However, the clustering coefficient of the DAX network is one order of magnitude larger than that of the WN and is, in fact, comparable with that of the FJT. To investigate the effect of intermittency and rare events in the corresponding networks of the time series, we repeat the procedure for the FJT for a very large Reynolds number (see Table I). Due to the very high fluctuations of the velocity field at the high Reynolds number, the clustering coefficient decreases. But, at the same time, due to the increase in the rate of "rare" events with increasing Reynolds number, the network's diameter increases too.

To determine the effect on the network structure of long-range correlations and the fatness of the tails of the time series' PDF, we shuffled and surrogated the time series. In principle, long-range correlations and the fatness of the PDF can lead to the intermittent or multifractality of a time series. If the long-range correlations are the origin of the multifractality, then the corresponding shuffled time series should exhibit monofractal scaling, since the long-range correlations are destroyed by the shuffling. If, however, the multifractality is also due, in part, to the fat tails of the PDF, the shuffled series will exhibit weaker multifractality than the original time series, as the multifractality due to the fatness of the PDF of the time series is not affected by the shuffling.

To investigate the possibility of multifractality due to the broadness (fat tails) of the PDF, we surrogate the time series. In this method the phases of the coefficients of the discrete Fourier transform of the time series are replaced with a set of pseudo-independent quantities, uniformly distributed in $(-\pi, \pi)$. The correlations in the surrogate series do not change, but the probability function changes to the Gaussian distribution. Our analysis indicates that for the shuffled time series the clustering coefficients, diameter, and the mean length of the equivalent network take on values that are essentially equal to those for the WN series, while the network for the surrogate of the time series increases slightly in diameter and the mean length, but its clustering coefficients remains unchanged. This implies that the clustering coefficients are almost independent of the structure of the time series' PDFs.

We find for the time series that we study that the degree distributions of the outgoing and incoming links, $N(k_+)$ and $N(k_-)$, have the property that, $N(k_+) \simeq N(k_-)$. Moreover, the network for the turbulence time series $x(t)$ has, on average, about 90-120 nodes, whereas the corresponding numbers

5

for the DAX's and WN's networks are about 900 and 950 (on the order of the bin numbers). Thus, the FJT time series $x(t)$ does not contain the possibility of "jumps" from one bin to another arbitrary bin, whereas such possibilities do exist for the DAX and WN series; see Fig. 3. Due to the different numbers of the accessible nodes for the turbulence and the other time series, we have plotted the $N(k)$ for the DAX and WN in the inset of Fig. 3. Moreover, the plots of $N(k)$ for the turbulence series with the Reynolds numbers Re=36100 and 757000 indicate that, at the higher Reynolds number the peak of the distribution shifts to higher values of k, but the distribution still does contain smaller values of k as well.

We also find that the PDF of the weight differences $w_{ij} - w_{ji}$ of the networks' links has a positive skewness for the FJT and DAX series, but vanishing skewness for the WN series, implying that in the former case their third moments $S_3 \geq 0$. The typical values for $S_3 = \langle (w_{ij} - w_{ji})^3 \rangle / \langle (w_{ij} - w_{ji})^2 \rangle^{3/2}$ are 0.047, 0.001, and 0.000 for the FJT at Reynolds number Re=36100, the DAX, and the WN series, respectively. Thus, the FJT and DAX time series do not have symmetric adjacent matrices. The symmetry is broken even more strongly in the FJT. Moreover, we find that, $S_3 = 0.0185$ for the time series for turbulence with low-temperature helium (as the fluid) at Re=757000, hence indicating that the adjacent matrix is more symmetric at high Reynolds numbers Re.

An important aspect of this work is that, the networks enable us to *reconstruct* the time series, such that they will be similar, in the *statistical* sense, to the original time series. The dynamics of conditional PDF $P(x, t | x_0, t_0)$ is governed by the Kramers-Moyal (KM) equation [19]:

$$\partial P(x, t | x_0, t_0) / \partial t = \sum_{l=1}^{\infty} (-\partial/\partial x)^l [D^{(l)}(x) P(x, t | x_0, t_0)].$$

The KM coefficients $D^{(l)}(x)$ may be written in terms of the weights w_{ij} as, $D^{(l)}(x_j) = \int dx_i \, (x_i - x_j)^l w_{ij}$. The n-point joint PDF can be written in terms of $P(x, t | x_0, t_0)$ and $P(x, t)$. This means that knowledge of the weights w_{ij} yields all the correlation functions of the process $x(t)$. This fact enables us to reconstruct the time series with very high precision.

In practice, to reconstruct a given time series, we first construct its network equivalent and the weights w_{ij} and, then, perform a random walk on the network, with the transition probabilities of the walk being the weights w_{ij}. The random walk then generates a time series (of the visited nodes) with the same statistical properties as those of the original series. The generalization of the reconstruction procedure to time series with a ME time scales $T > 1$ is also straightforward. One constructs the transition matrix with its entries given by, $p(x_t | x_{t-T})$, and attributes to each node a set of data

arrays with length $T > 1$. Therefore, to reconstruct the time series with $T > 1$, a stochastic walker moves from one node to another, such that at each node i we set the array, $x_i(t) = (x_1, x_2, \cdots, x_T)$. Moreover, we may use the same method for constructing the equivalent networks for *coupled* or *multidimensional* time series. For example, for the coupled time series $x(t)$ and $y(t)$, we assign the vector $(x_1, \cdots, x_{T_1}, y_1, \cdots, y_{T_2})$ to each node, where the time series $x(t)$ and $y(t)$ have the ME time scales T_1 and T_2, respectively.

As an example, consider a time series with with $T \simeq 20$, shown in Fig. 4. There, we have plotted the data for the FJT with a Reynolds number of 36100, and reconstructed the time series, from the top to the bottom, respectively. The reconstructed time series preserves all the statistical properties of the original series over *every* scale. To make a quantitative check on this claim, we also present in Fig. 4 the PDFs of the turbulence time series increments, i.e., $P(x(t+\tau) - x(r); \tau)$ for $\tau = 1, \ 10, \ 100$ and 800, respectively [21]. As shown in Fig. 4, the PDFs of the reconstructed time series have the same properties as the original time series. This implies that the reconstruction also preserves the cascade nature of the turbulence series in scale.

Figure 4 also presents the scaling behavior of the structure function of the velocity increments, i.e., $S_q = \langle |x(t+\tau) - x(t)|^q \rangle$, for $q = 2, \ 3$, and 4. The results indicate again that they have the same scaling nature and properties. Thus, the agreement between the statistical properties of the original and reconstructed time series is excellent, at least with respect to the general structure functions.

Finally, we note that knowledge of the weights w_{ij} enables us to determine the level-crossing frequency of the time series [22, 23, 24]. The level crossing is characterized by the quantity ν_α^+, which is the average frequency of positive-slope crossing of a level α in a time series. The frequency ν_α^+ of the crossings is deduced from the underlying weight matrix w_{ij}. For discrete time series, the frequency is given by,

$$\nu_\alpha^+ = \int_\infty^\alpha \int_\alpha^\infty w_{ij} P(x_j) \ dx_i dx_j. \tag{2}$$

In summary, we introduced a method by which stochastic processes are mapped onto equivalent complex networks. We described the physical interpretation of the networks' geometrical properties, such as their mean length, diameter, clustering, average number of connection per node, and their stochastic interpretations. As example, we determined the network characteristics of the free-jet and low-temperature helium turbulence, the German stock market index (the DAX), and the white noise. For a given process with a finite Markov-Einstein time (or length) scale, the corresponding network enables us to *reconstruct* the time series with high precision, by performing a random walk in the

network.

As a distinct application, generation and regeneration of large surfaces would be possible by sampling a real surface with high resolution (with the same resolution as nanoscope imaging, e.g., the atomic force microscope images). This would then be applicable to computer simulation of surfaces and interfacial processes, such as diffusion of materials between rough surfaces, the effect of surface roughness on the friction, and so on [14]. It was shown that conditional probability distribution function $P(x, t|x_0, t_0)$, satisfies the Kramers-Moyal equation, with the coefficients of the equation given by the weights of the corresponding adjacent matrix. We also established the relation between the level-crossing frequency of the time series in terms of the weights w_{ij} attributed to each link of the network. The method described here is applicable to a wide variety of stochastic processes and, unlike many of the previous methods, does not require the data to have any scaling feature.

References

[1] ALBERT R. and BARABASI A. L., *Rev. Mod. Phys.*, **74** (2002) 47.

[2] DOROGOVTSEV S. N. and MENDES J. F. F., *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[3] COHEN R. and HAVLIN S., *Complex Networks: Stability, Structure and Function* (Cambridge University Press, Cambridge, 2008).

[4] ERDÖS P. and REYNI A., *Publ. Math.* (Debrecen), **6** (1959) 290.

[5] ERDÖS P. and REYNI A., Publications of the Mathematical Inst. of the Hungarian Acad. of Sciences, **5** (1960) 17.

[6] PALLA G., BARABASI A. L., and VICSEK T., *Nature*, **446** (2007) 664.

[7] MANTEGNA R. and STANLEY H. E., *An Introduction to Econophysics: Correlations and Complexities in Finance* (Cambridge University Press, New York, 2000).

[8] FRIEDRICH R., PEINKE J., and RAHIMI TABAR M. R., in *Encyclopedia of Complexity and System Science*, edited by R. Meyers (Springer, Berlin, 2009).

[9] SCALA A., AMARAL L. A. N., and BARTHELEMY M., *Europhys. Lett.*, **55** (2001) 594.

[10] RAO F. and CAFLISH A., *J. Mol. Bio.*, **342** (2004) 299.

[11] SHREIM A., GRASSBERGER P., NADLER W., SAMUELSSON B., SOCOLAR J. E. S., and PACZUSKI M., *Phys. Rev. Lett.*, **98** (2007) 198701; LACASA L., et al. Proc. Natl Acad. Sci. U.S.A., **105**, 13 (2008) 4972.

[12] ZHANG J. and SMALL M., *Phys. Rev. Lett.*, **96** (2006) 238701.

[13] FRIEDRICH R. and PEINKE J., *Phys. Rev. Lett.*, **78** (1997) 863.

[14] JAFARI G. R., FAZLEI S. M., GHASEMI F., VAEZ ALLAEI S. M., RAHIMI TABAR M. R., IRAJI ZAD A., and KAVEI G., *Phys. Rev. Lett.*, **91** (2003) 226101.

[15] JAFARI G. R., MAHDAVI S. M., IRAJI ZAD A. and KAGHAZCHI P., *Surface and Interface Analysis,* **37** *(2005) 641; GHASEMI F., SAHIMI M., PEINKE J., FRIEDRICH R., JAFARI G. R., and RAHIMI TABAR M. R., Phys. Rev. E,* **75** *(2007) 060102(R); JAFARI G. R., BAHRAMINASB A. and NOROUZZDEH P., International Journal of modern physics C,* **18**, *(2007) 1223.*

[16] FAZALI S. M., SHIRAZI A. H. and G. R. JAFARI, New J. Phys. **10** *(2008) 083020; KIMIAGAR S., JAFARI G. R. and RAHIMI TABAR M. R., J. Stat. Mech. (2008) P02010.*

[17] GHASEMI F., BAHRAMINASAB A., MOVAHED M. S., RAHVAR S., SREENIVASAN K. R., and RAHIMI TABAR M. R., J. Stat. Mech. (2006) P11008.

[18] LÜCK St., RENNER Ch., PEINKE J., and FRIEDRICH R., Phys. Lett. A, **359** *(2006) 335.*

[19] RISEKEN H., The Fokker-Planck Equation (Springer, Berlin, 1984).

[20] NEWMAN M. E. J., SIAM Rev., **45** *(2003) 167.*

[21] FRISCH U., Turbulence, The Legacy of A. N. Kolmogorov (Cambridge University Press, Cambridge, 1995).

[22] RICE S. O., Bell System Tech. J., **23** *(1944) 282; ibid.,* **24** *(1945) 46.*

[23] SHAHBAZI F., SOBHANIAN S., RAHIMI TABAR M. R., KHORRAM S., FOROOTAN G. R., and ZAHED H., J. Phys. A **36** (2003) 2517; JAFARI G. R., MOVAHED M. S., FAZELI S. M., RAHIMI TABAR M. R. and MASOUDI S. F. J. Stat. Mech., P06008 (2006).

[24] BAHRAMINASAB A., SADEGH MOVAHED M., NASIRI S. D., MASOUDI A. A., and SAHIMI M., J. Stat. Phys., **124** (2006) 1471; VAHABI M., JAFARI G. R., Physica A **385** (2007) 583.

Network's Properties of different time series

Networks	Mean length	Clustering	Diameter	$k_{\mathrm{Max,out}}/k_{\mathrm{Max,in}}$	$\langle(k_{\mathrm{out}}^2\rangle/\langle k_{\mathrm{out}}\rangle)/(\langle k_{\mathrm{in}}^2\rangle/$
White Noise	1.0498	0.001	2	970/973	0.0067/0.0065
Stock Market (DAX)	1.1044	0.013	2	964/960	0.0922/0.0935
Turbulence, Re=36100	3.820	0.038	15	131/128	0.3359/0.3335
Turbulence, Re=757000	3.776	0.023	23	178/181	0.3458/0.3469
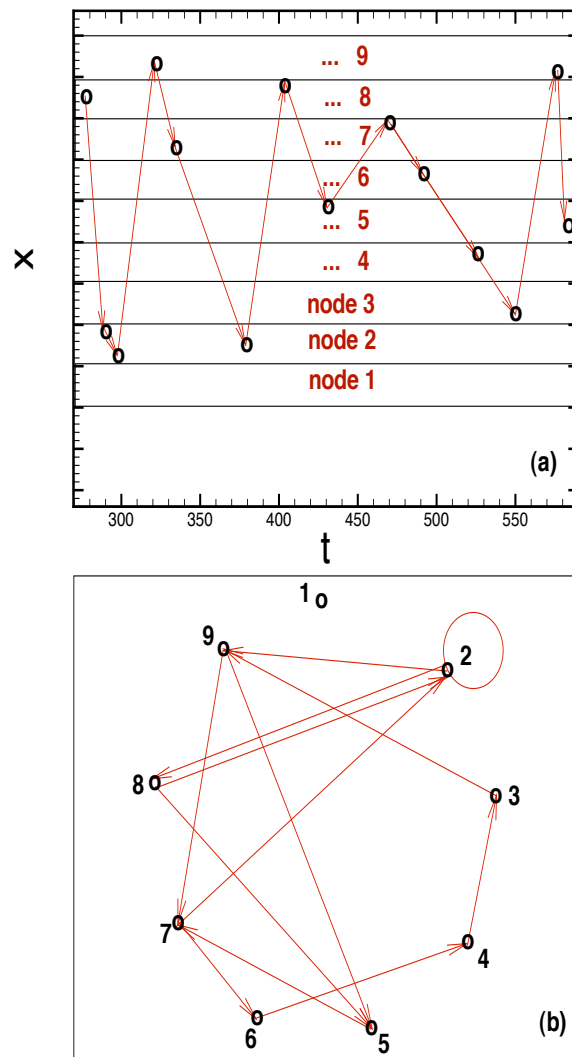
Figure 1: A schematic network (b) made from a signal(a).
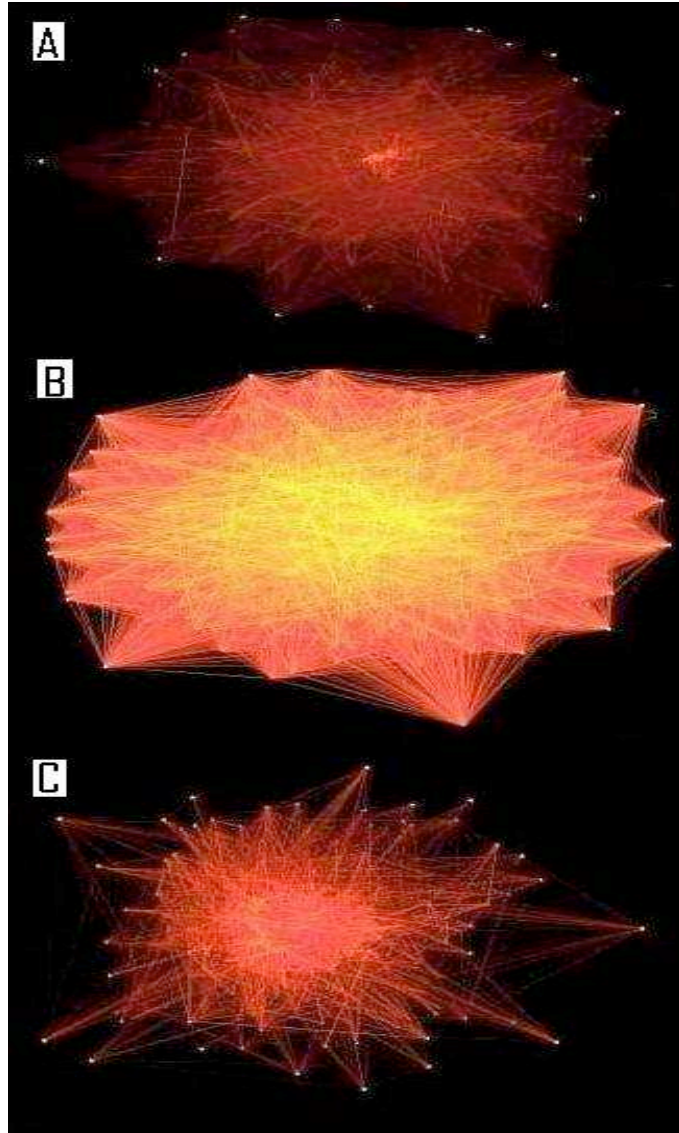
Figure 2: Networks for the time series of, (A) turbulence; (B) the white noise, and (C) log-returns of the DAX. The networks have 1000 nodes and are constructed by 3×10^6 normalized data points.
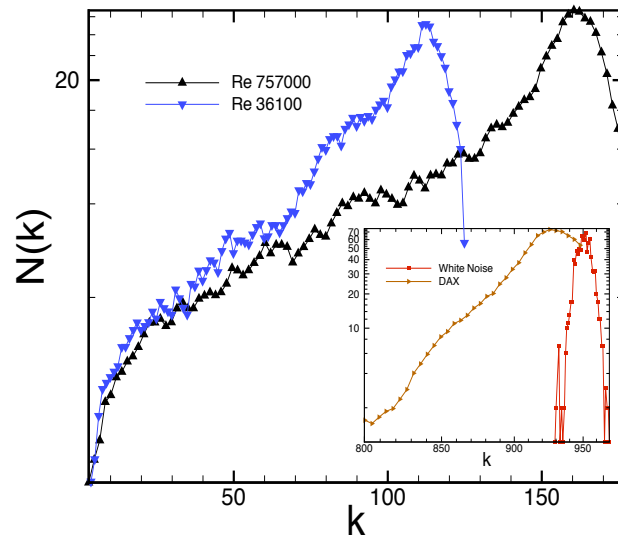
Figure 3: The number distribution of vertices with a degree equal to k for the white noise, turbulence time series with Reynolds numbers of 36100 and 757000, and the log-returns of the DAX.
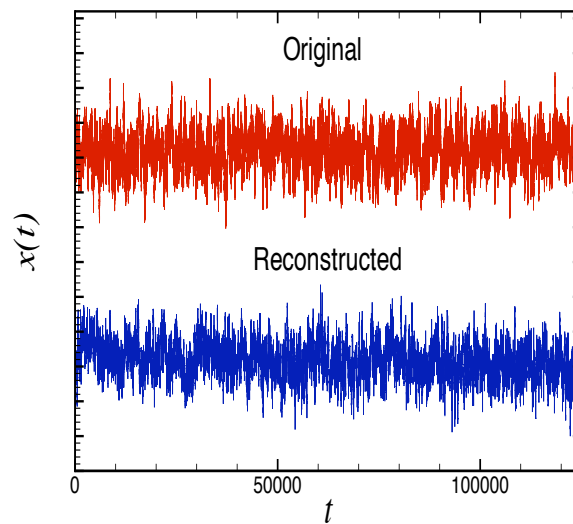

Figure 4: Samples of the original turbulence time series (upper) as the input, and the reconstructed data (down) as the output.
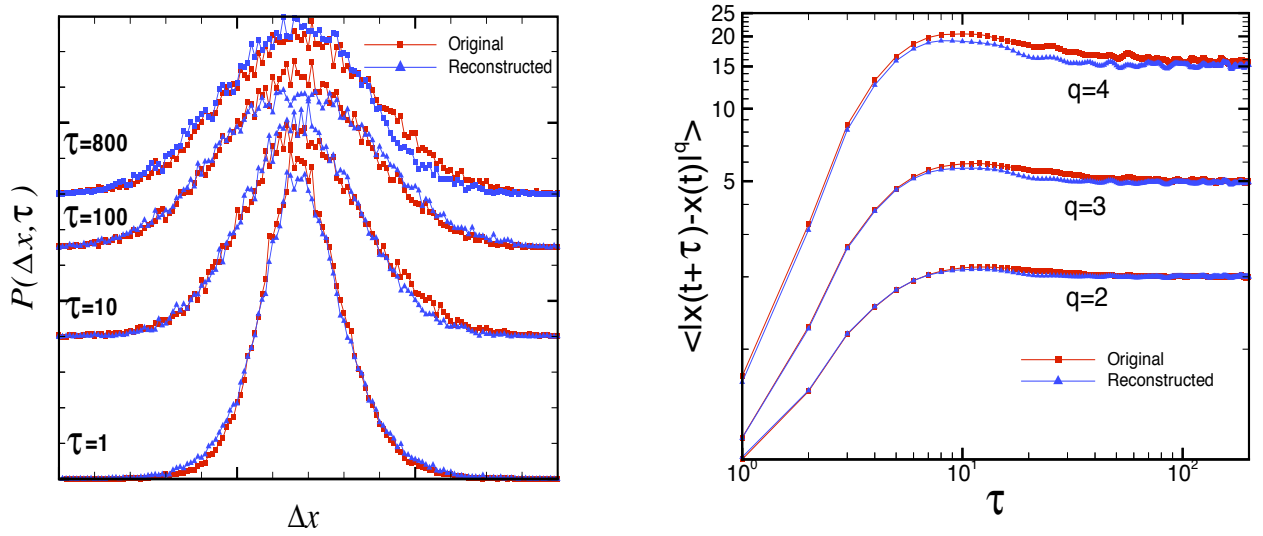
14

Figure 5: Comparison of the directly-evaluated PDF $P(x(t+\tau) - x(r); \tau))$ for $\tau = 1,\ 10,\ 100$, and 800. Bottom: Comparison of the moments, $S_q = \langle |x(t+\tau) - x(t)|^q \rangle$ ($q = 2,\ 3$, and 4) with those of the original (square) and reconstructed (triangle) time series.