Visitor flow pattern of Expo 2010*

Fan Chao(樊 超)[a)b)] and Guo Jin-Li(郭进利)[a)†]

[a)] *Business School, University of Shanghai for Science and Technology, Shanghai* 200093, *China*
[b)] *College of Arts and Sciences, Shanxi Agricultural University, Taigu* 030801, *China*

(Received 31 October 2011; revised manuscript received 21 January 2012)

Expo 2010 Shanghai China was a successful, splendid, and unforgettable event, leaving us with valuable experiences. The visitor flow pattern of the Expo is investigated in this paper. The Hurst exponent, the mean value, and the standard deviation of visitor volume indicate that the visitor flow is fractal with long-term stability and correlation as well as obvious fluctuation in a short period. Then the time series of visitor volume is converted into a complex network by using the visibility algorithm. It can be inferred from the topological properties of the visibility graph that the network is scale-free, small-world, and hierarchically constructed, confirming that the time series are fractal and a close relationship exists among the visitor volumes on different days. Furthermore, it is inevitable that will be some extreme visitor volumes in the original visitor flow, and these extreme points may appear in a group to a great extent. All these properties are closely related to the feature of the complex network. Finally, the revised linear regression is performed to forecast the next-day visitor volume based on the previous 10-day data.

Keywords: fractal pattern, time series, visibility graph, complex network

PACS: 02.50.–r, 89.75.–k, 89.75.Hc, 89.65.–s **DOI:** 10.1088/1674-1056/21/7/070209

1. Introduction

From 1 May 2010 to 31 October 2010, Expo 2010 was successfully held in Shanghai, China. Given its theme "Better City, Better Life", Expo 2010 was expected to be a "successful, splendid and unforgettable" event and a platform for every exhibitor to explore sustainable models of urban development and better city life. During the 6 months of the exhibition, 246 nations and organizations participated in this event, which attracted over 73 million visitors. Both the exhibition scale and the total visitor volume hit a world record. How to manage so many visitors in such a small area and such a long period is a problem worth considering.

Besides the Expo, various kinds of fairs, expositions, sports meetings, and large conferences are held all over the world every year. The successful experiences from Expo 2010 may provide references for the visitor management of such large-scale activities. Unfortunately, we rarely find any work about this subject. To summarize the valuable experience of Expo 2010, we investigate the visitor flow pattern through statistics, time series, and complex network analyses in the present work.

Statistical and time series analyses are relatively mature techniques, which have been used widely in many fields. By contrast, the complex network theory[1−3] is a new branch in statistical physics to describe complex systems with networks, in which the nodes and the edges represent the entities and the relationships between them, respectively. It can be used to describe many networks in the real world, such as social networks,[4] transportation networks,[5] and human dynamics on networks.[6] Recently, several efforts have been made to bridge the time series and the complex networks.[7−10] Among all the methods, the so-called visibility graph algorithm proposed by Lacasa *et al.*[9] attracts much attention due to its simplicity and high efficiency, and a set of achievements have been obtained through it.[11−14] In our research, the Expo visitor flow pattern is studied mainly by using the network analysis, and the network is obtained through the visibility algorithm.

The rest of this paper is organized as follows. The data source and the overview are given in Section 2. In Section 3, the general statistical properties of the Expo daily visitor volume are studied to find the stability and fluctuation feature of the visitor flow. The time series of visitor volume are converted into a com-

*Project supported by the National Natural Science Foundation of China (Grant No. 70871082), the Shanghai Leading Academic Discipline Project, China (Grant No. S30504), and the Science and Technology Innovation Foundation of Shanxi Agricultural University, China (Grant No. 201208).

†Corresponding author. E-mail: phd5816@163.com

© 2012 Chinese Physical Society and IOP Publishing Ltd

plex network by using the visibility graph algorithm, and some topological parameters are calculated to investigate the correlation among data points of visitor volume in Section 4. The revised linear regression is performed to forecast the future visitor volume in Section 5. Finally, some conclusions and discussion are given in Section 6.

2. Data specification

The data of daily visitor volume used in this paper are collected from the official website of Expo 2010 (http://www.expo2010.cn). During the exhibition period of 184 days, there are approximately 73091000 visitors in total and 397234 daily visitors on average. Figure 1 exhibits the general fluctuation pattern of Expo 2010, with the horizontal axis representing the time (day), the left and the right vertical axes representing daily and accumulative visitor volumes, respectively.

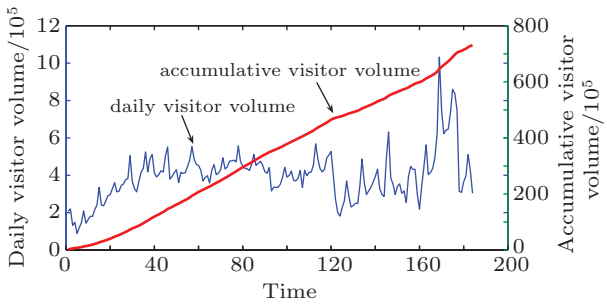


Fig. 1. (colour online) Daily and accumulative visitor volumes of Expo 2010.

3. General feature of Expo visitor flow

In this section, the visitor flow of Expo 2010 is investigated from two perspectives: statistical and time series analyses.

We calculate the mean value and the standard deviation of each month and the whole period, which are shown in Table 1. From the results, we can see the general fluctuation pattern of monthly visitor flow. Firstly, much fewer people went to the Expo in May, owing to the fact that people thought that they had plenty of time to visit, thus there was no need to hurry. In addition, the operation of the Expo was not in the best condition then, which made people hesitant to visit. Secondly, the visitor flow in July is the most stationary and the second greatest, which is the result of the steady flow of tour groups and students on their summer vocation. Thirdly, during the six months of the exhibition, October has the greatest visitor volume and standard deviation. The increase of visitor volume at the end of the exhibition is not surprising. Meanwhile, the obvious fluctuation in the visitor flow is the result of interweaving of small visitor volumes on the 14 designated days and the extremely great visitor volumes on several standard days, such as 1032700 on 16 October, 859900 on 22 October, and 837400 on 23 October.

Table 1. Mean values and standard deviations of Expo 2010 visitor volume.

Period	May.	Jun.	Jul.	Aug.	Sept.	Oct.	Total
Mean value/10^3	259.2	436.9	444.8	401.9	333.6	506.3	397.2
Standard deviation/10^3	96.0	55.1	40.3	73.4	89.2	194.7	132.0

On the other hand, we take the daily visitor volumes as the observation of a time series to study the visitor flow pattern of Expo 2010. More specifically, we use the Hurst exponent, namely the long-range dependence exponent (H for short), to measure the long-term memory effect of the time series, i.e., the autocorrelation of the time series, which has been extensively used in many fields, such as the stock market. The closer to 0.5 the value of H is, the more noise and fluctuation there will be in the time series. Meanwhile, when H deviates more from 0.5 and approaches 1, the time series will be more regular and persistent. Conversely, the time series are deemed to be anti-persistent when H declines to 0.

Here we use the method of rescaled range analysis[15,16] to obtain the Hurst exponent and the V-statistic (used to identify the length of non-periodical cycle) of the time series. As shown in Fig. 2(a), the result is that the Hurst exponent of the whole period daily visitor flow is 0.856 with $R^2 = 0.985$. It can be concluded that the visitor flow of the Expo does not obey the random walk but exhibits long-term stability and correlation, as the Hurst exponent exceeds 0.5. The deviations of the visitor flow in the future tend

to keep the same sign as the past.

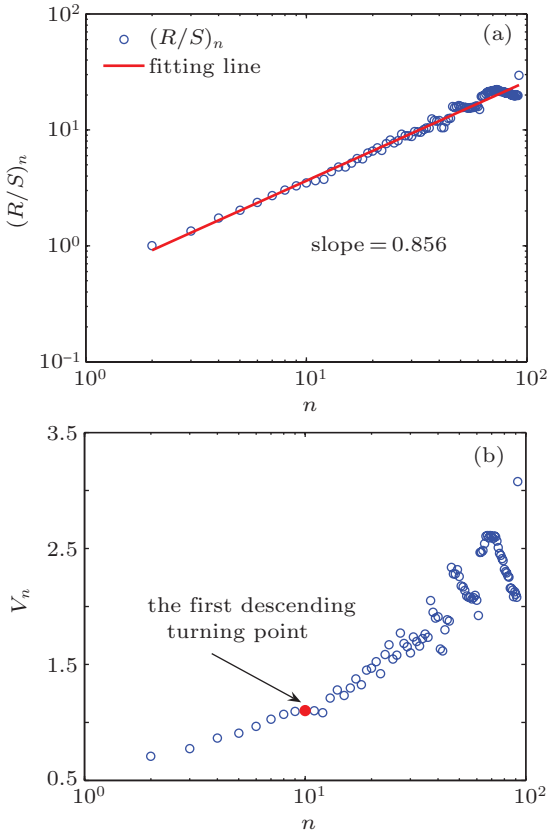


Fig. 2. (colour online) Plots of (a) $(R/S)_n$ and (b) V-statistic of visitor volume versus n. The slope of the curve in panel (a) indicates the value of the Hurst exponent, and the first crossover point in panel (b) corresponds to the cycle length of the time series.

Moreover, we can see from Fig. 2(b) that the V-statistic grows with time n in an accelerative trend, which means that the length of non-periodical cycle is far beyond the exhibition period of 184 days. Nevertheless, we may judge from Fig. 1 that there exists a short-term cycling period in the visitor flow, since the curve exhibits some kind of periodic variation. A closer inspection reveals that the curve of V_n versus n begins to shake after n has reached 10 in Fig. 2(b), and $n = 10$ is the first descending turning point, indicating that the cycling period is 10 days. This is confirmed by performing the linear regression to forecast the future trend of the visitor flow in Section 5, which finds that the prediction error reaches the minimum when the cycle is 10 days. For this reason, the results of rescaled range analysis are really helpful for forecasting the time series of the Expo visitor flow.

In short, the general law of the Expo visitor flow exhibits both fluctuation and stability features. The daily volumes are correlated with the long-term perspective, and they show obvious fluctuation in a short period of time.

4. Statistical properties of the visibility graph of Expo visitor volume

We use the algorithm introduced in Ref. [9]. A visibility graph is obtained by mapping a time series into a network according to the following visibility criterion: two arbitrary data (t_a, y_a) and (t_b, y_b) in the time series have visibility and consequently become two connected nodes in the associated graph, if any other data (t_c, y_c) with $t_a < t_c < t_b$ fulfills

$$y_c < y_a + (y_b - y_a)\frac{t_c - t_a}{t_b - t_a}.$$

Figure 3 shows a typical example of this algorithm. In Fig. 3(a), the data are displayed as vertical bars ordered according to the time sequence, where the heights indicate the values of observations, and the dashed lines express the visibilities between the data points. The converted network is shown in Fig. 3(b), where the nodes correspond to the data in the same order, and an edge connects two nodes if there is visibility between them. The visibility graph inherits several properties of the time series in its structure, i.e., a periodic series is converted into a regular graph, while a random series is converted into an exponential random graph, and a fractal series is converted into a scale-free

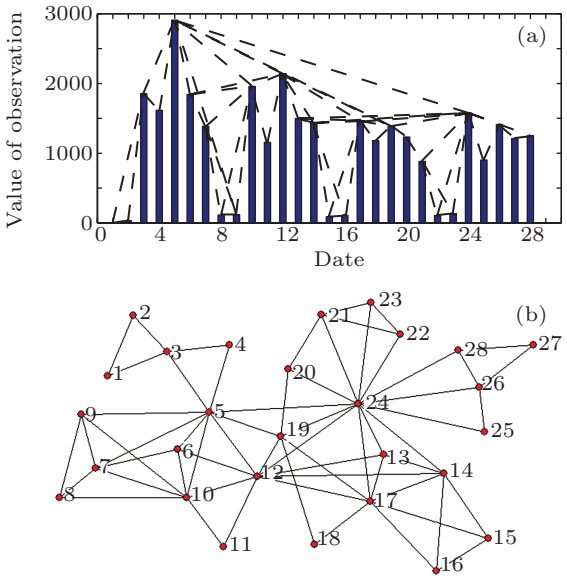


Fig. 3. (colour online) A typical example of visibility graph algorithm. In panel (a), the data are displayed as vertical bars ordered according to the time sequence, panel (b) shows the converted network.

network.

As shown in Fig. 4, the time series of the daily visitor volume of the Expo is converted into a visibility graph using the algorithm introduced above. In order to investigate the feature of the original time series of visitor flow, some characteristics of the visibility graph with 184 nodes and 676 edges are explored as follows.[2]

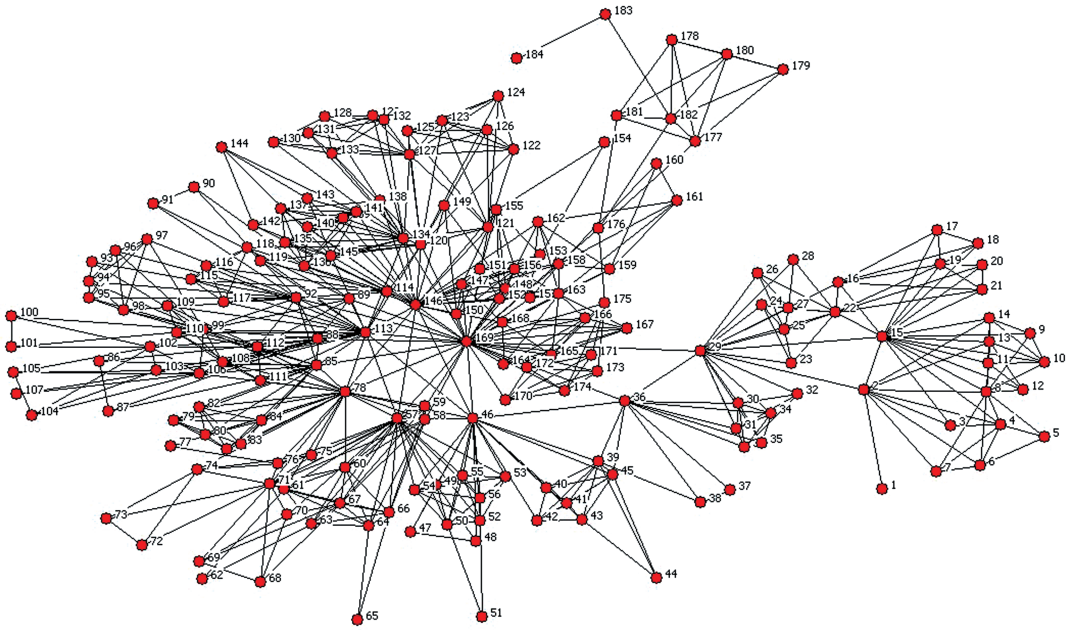


Fig. 4. (colour online) Network mapped from the time series of daily visitor volume of Expo 2010. The 184 nodes correspond to the days in the exhibition period, and the edges are established according to the visibility algorithm.

1) Average degree of network $\langle k\rangle$, which is the mean value of degrees of all the nodes in the network, is $\langle k\rangle = \sum_{i=1}^{N} k_i \Big/ N = 7.348$.

2) Average clustering coefficient $\langle C\rangle$, which is the mean value of clustering coefficients of all the nodes in the network, is $\langle C\rangle = \sum_{i=1}^{N} C_i \Big/ N = 0.789$.

3) Diameter of network D, which is the maximal distance between any pair of nodes, is $D = \max_{i,j} d_{ij} = 8$, where d_{ij} refers to the number of edges on the shortest path connecting nodes i and j.

4) Average path length L, which is the mean value of the distance between any pair of nodes, is $L = \dfrac{2}{N(N+1)} \sum_{i\geq j} d_{ij} = 3.447$.

5) Degree distribution $P(k)$ is the probability of a certain node to have degree k. In a scale-free network, the degree distribution obeys a right-skewed power law $P(k) \propto k^{-\gamma}$. For reducing the noise, we study the accumulative degree distribution, which obeys a power law $P(k) = 44.25k^{-2.37}$ with $R^2 = 0.991$. Therefore, the visibility graph is scale-free.

6) The small-world effect. We calculate the average path length step by step while increasing the total number of nodes N in the network. If L increases logarithmically with N increasing, namely, $L(N) \propto \ln N$, or slower, and the network keeps a large clustering coefficient at the same time, then the network is considered to have the feature of small world. As shown in Fig. 5(b), it can be observed that L increases with N more slowly than that in the logarithmical pattern, verifying that the visibility graph is a small-world network.

7) Hierarchical structure. The weighted average values of clustering coefficients of nodes with degree k are calculated as $\bar{C}(k) = \langle C|k\rangle = \sum_{i=1}^{n} C_i \Big/ n$, where n refers to the number of different clustering coefficients that a node with degree k has. The network is considered to be hierarchically constructed if $\bar{C}(k) \propto k^{-\alpha}$. In our case, $\bar{C}(k) = 5.597k^{-0.97}$.

8) The Pearson correlation coefficient r.[17] There are many hub nodes in a scale-free network, which have much larger degrees than the other nodes. Whether the interaction among the hubs of the network is attractive or repulsive can be determined from

the correlation between the degrees of different nodes. The degree correlation can be quantified by the Pearson correlation coefficient defined as

$$r = \frac{N^{-1}\sum_i k_1 k_2 - \left[N^{-1}\sum_i \frac{1}{2}(k_1+k_2)\right]^2}{N^{-1}\sum_i \frac{1}{2}(k_1^2+k_2^2) - \left[N^{-1}\sum_i \frac{1}{2}(k_1+k_2)\right]^2},$$

where k_1 and k_2 are the degrees of the nodes at the two ends of edge i. It can be deduced from the result of $r = 0.115$ that the network is positively correlated with the hub nodes being attractive to each other.

9) Nearest neighbors average connectivity.[18] The relation of degree between one node and its nearest neighbors can be measured by quantity $\langle K_{nn}\rangle = \sum_{k'} k' P(k'|k)$, where the conditional probability $P(k'|k)$ denotes the probability of a node with degree k connecting to a node with degree k'. Therefore, $\langle K_{nn}\rangle$ is used to investigate the relation between the degree of a certain node and the average degree of its nearest neighbors. From Fig. 5(d), we can see clearly that $\langle K_{nn}\rangle$ and k are positively correlated.

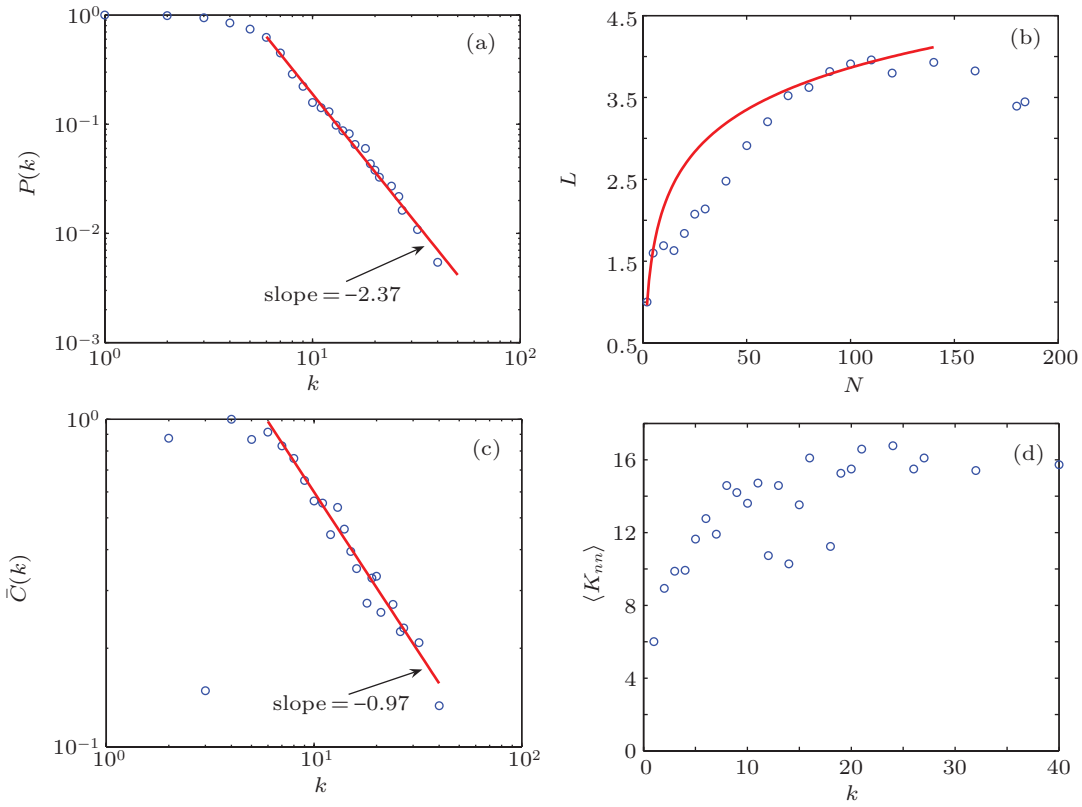


Fig. 5. (colour online) Topological features of the visitor flow network. Panel (a) shows that the accumulative degree distribution obeys a power law with exponent −2.37. Panel (b) shows that the average path length grows with the total number of nodes in the network more slowly than that in the logarithmical pattern, implying that the network is a small-world one. Panel (c) shows that the weighted average value of the clustering coefficients decreases with the degree obeying a power law with an exponent near to 1, suggesting that the network is hierarchically constructed. In panel (d), $\langle K_{nn}\rangle$ versus k shows the positive relation of degree between a node and its neighbors.

From the topological parameters calculated above, some explanations and conclusions can be given as follows.

Firstly, the visibility graph being a scale-free network verifies that the time series of the Expo visitor flow shows fractal characteristics. This result confirms the fact that the power-law degree distribution is related to fractality, which has been intensively discussed recently.[9,11−14,19−22] More accurately, the total degree of the top 29 nodes each with a degree no less than 10 accounts for up to 36.1% of the total degree of the whole network, and the average degree of the remaining 155 nodes is only 5.574, obviously showing the inhomogeneity of the degree distribution.

Secondly, the fact that the network has a large clustering coefficient and a small average path length growing slowly with the total number of nodes verifies the small-world phenomenon, which means that there

is a tight connection between the nodes even they are located far away from each other, since there are visibility lines between the corresponding data points in the time series. Consequently, it can be deduced that there is a certain relation between the quantities of visitor flow at different exhibition times. In other words, it is not random or unconnected between the past and the future in the time series of human behaviors.

Thirdly, it can be inferred from the result $\bar{C}(k) = 5.597k^{-0.97}$ that the visibility graph is hierarchically constructed, which means that if a node in the network has a larger degree, its neighbors do not tend to connect with each other. The nodes with larger degrees are the ones that have relatively greater observations in the time series than their directly connected and even unconnected neighbors. These extreme points correspond to the hub nodes in the scale-free network. For example, node 15 has a comparatively high degree $k = 18$ and a low clustering coefficient $C = 0.261$, and there are 335300 visitors on 15 May, much larger than the numbers of visitors of its neighbor nodes. (There are 163000, 180400, 180100, 215500, 240300 visitors on the 5 days before 15 May and 241500, 236400, 261900, 290600, 296400 visitors after 15 May respectively. Obviously, the daily visitor numbers are significantly smaller but more uniform before and after 15 May.) Correspondingly, node 15 has many neighbors in the network which are separated into two parts and connected with each other with a small probability. Thus, it can be concluded that it is inevitable that some extreme visitor volumes will be shown in the homogeneous flow on such a pattern.

Finally, the result $r = 0.115$ means that the network is assortative mixing, namely, the nodes with high degrees tend to link with the nodes also having high degrees. Moreover, the fact that $\langle K_{nn} \rangle$ and k are positively correlated implies that the larger degree a certain node has, the larger the average degree of its neighbors is. To make the statement clearer, the relation between the node degree in the network and the average visitor volume in the time series is discussed. As shown in Fig. 6, generally, the nodes with larger degrees correspond to the data points with more visitors. To sum up, the clustering phenomenon of the hub nodes in the network means that the extreme points in the time series appear in the form of group, in other words, a large visitor flow is always accompanied with other large visitor flows. Consequently, there are some consecutive periods of extreme visitor volumes in the whole exhibition, for instance, from 14 October to 24 October with an average daily visitors of 726255 and from 10 July to 28 July with an average daily visitors of 466942.

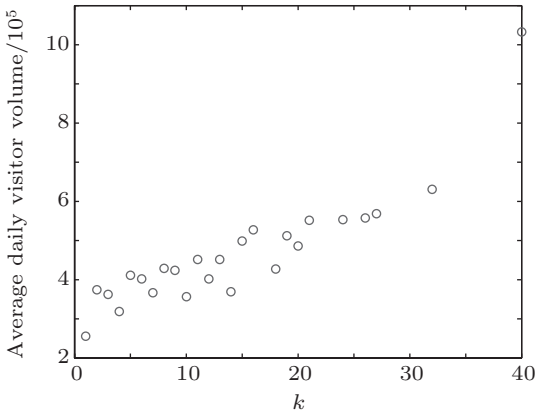


Fig. 6. Node degrees positively correlated with visitor volumes.

In conclusion, the visibility graphs converted from the time series are scale-free, have the small-world effect, and are hierarchically constructed. Thus the original time series are deemed to have a fractal feature, and there is a close relationship among the data points, especially those extreme points. Furthermore, the extreme visitor volumes are unavoidable and appear successively for several days.

5. Forecasting of Expo visitor volume

The time series analysis provides a method of using past observations collected at regular intervals to make a quantitative prediction about future events. Our study on the visitor flow pattern is also aimed at providing some help for forecasting the future visitor volume, which has a realistic significance. The fact that the visitor volume is influenced by many interweaving factors, such as weekday, holiday, weather, traffic, and designated days or standard days, makes accurate forecasting almost impossible. For simplicity, the classic time series forecasting method is used to predict future visitor volumes, and only the influence of the weekday is taken into account in our prediction below.

Firstly, the adjustment coefficients based on weekdays are calculated as

$$c_k = \bar{X}_k / \bar{X}, \quad k = 1, 2, \ldots, 7,$$

where $\bar{X}_k$ and $\bar{X}$ refer to the average visitor volumes of each day in a week and the whole first month. All the results are listed in Table 2.

Table 2. Total and daily average visitors in May and the adjustment coefficients.

	Mon.	Tues.	Wed.	Thur.	Fri.	Sat.	Sun.	Sum
Number of days	5	4	4	4	4	5	5	31
Total visitors	1173100	936700	913100	1009100	1098700	1618200	1285500	8034400
Average visitors	234620	234175	228275	252275	274675	323640	257100	259174
Coefficients	0.9053	0.9035	0.8808	0.9734	1.0598	1.2487	0.9920	

Then the linear regression is performed to predict the $(i+k)$-th visitor volume based on the data of the former k days. Here we take $k = 10$, since 10 days is the cycling period obtained in Section 3 and it can minimize the prediction error. In order to make the prediction more accurate, the results are revised by the corresponding adjustment coefficients calculated above. Figure 7 shows the prediction results. It is clear that our forecasting values are very close to the real observed data, especially when the daily values are accumulated during a period of 10 days. Comparing the results of adjusted values and unadjusted ones, we find that such a revision can bring down the deviation of the forecasting value from the observed data by nearly 7%.

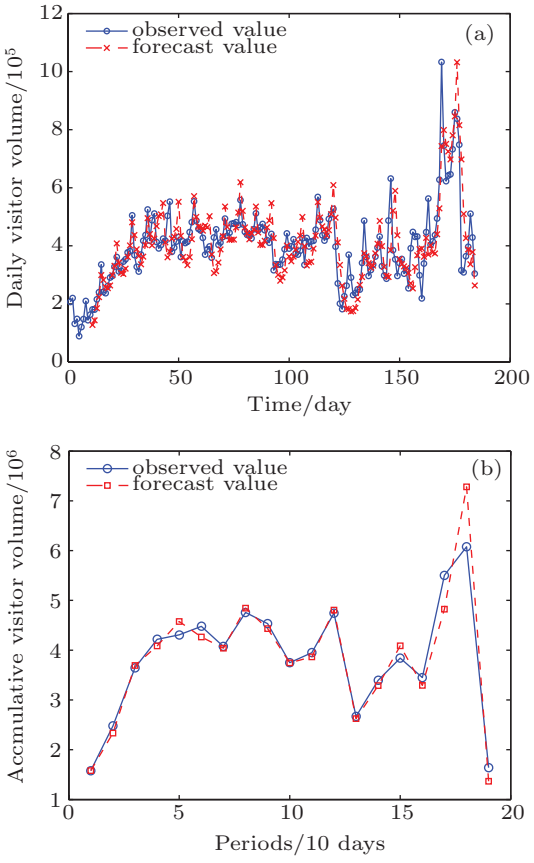


Fig. 7. (colour online) (a) Daily and (b) accumulative observation values (blue) and forecasting values (red) of the Expo visitor volume.

In summary, just as mentioned above, it is difficult to forecast future visitor volume accurately due to so many complicated factors, such as the unexpected but inevitable extreme visitor volume discussed in Section 5. When performing the prediction, we should not only use the well established methods, but also draw lessons from the historical experiences to make the prediction more flexible and accurate. For instance, we know from Expo 2008 that the visitor flow would increase significantly at the end of the exhibition period.

6. Discussion and conclusion

Expo 2010 in Shanghai, China, was a wonderful event, with a grand scale, long exhibition period, and a huge amount of visitors, leaving us with good memories and valuable experiences. In our research, the visitor flow is investigated from three different viewpoints.

First, the general feature of the daily visitor volume is discussed from the perspective of statistical and time series analyses. The results of mean value, standard deviation, Hurst exponent, and cycle length verify that the Expo visitor flow shows mixed properties of stability in the long term and fluctuation in the short term.

Second, the complex network converted from the time series using the visibility algorithm exhibits a scale-free property, the small-world effect, and a hierarchical structure, which confirms that the original time series is fractal and the data points are intensively connected with each other. Furthermore, the relations between the degree of one node and that of its neighbors, the degree and the clustering coefficient, as well as the degree and the visitor volume prove that the extreme visitor volumes are inevitable and appear in groups.

Third, the linear regression is performed to forecast the next-day visitor volume based on the previous 10-day data. In particular, the results are revised by using the adjustment coefficients according to the

week, which is verified to be an effective way to reduce the prediction error.

The method and the conclusion of our work may be helpful for managing and forecasting visitor flow, hiring staff and security guards, recruiting volunteers, designing pavilion structure, and so on, in large-scale exhibitions, spots meetings, or other human repetitious behaviors on a collective level. Moreover, our work may enrich the research of the correlation between the time series and complex networks. When the converted visibility graph is verified to be scale-free, how to make use of the topological structure and the dynamical mechanism of the complex network, as well as the fractal feature of the time series to perform the prediction, is worth further discussion.

References

[1] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[2] Albert R and Barabasi A L 2002 *Rev. Mod. Phys.* **74** 47
[3] Newman M E J 2003 *SIAM Rev.* **45** 167
[4] Si X M and Liu Y 2011 *Acta Phys. Sin.* **60** 078903 (in Chinese)
[5] Qian J H, Han D D and Ma Y G 2011 *Acta Phys. Sin.* **60** 098901 (in Chinese)
[6] Zhao F, Liu J H, Zha Y L and Zhou T 2011 *Acta Phys. Sin.* **60** 118902 (in Chinese)
[7] Zhang J and Small M 2006 *Phys. Rev. Lett.* **96** 238701
[8] Yang Y and Yang H J 2008 *Physica* A **387** 1381
[9] Lacasa L, Luque B, Ballesteros F, Luque J and Nuno J C 2008 *Proc. Natl. Acad. Sci. USA* **105** 4972
[10] Gao X Y, An H Z, Liu H H and Ding Y H 2011 *Acta Phys. Sin.* **60** 068902 (in Chinese)
[11] Lacasa L, Luque B, Luque J and Nuno J C 2009 *Europhys. Lett.* **86** 30001
[12] Luque B, Lacasa L, Ballesteros F and Luque J 2009 *Phys. Rev.* E **80** 046103
[13] Elsmer J B, Jagger T H and Fogarty E A 2009 *Geophys. Res. Lett.* **36** L16702
[14] Liu C, Zhou W X and Yuan W K 2010 *Physica* A **389** 2675
[15] Hurst H E 1951 *Transaction of American Society Civil Engineers* **116** 770
[16] Yoon S M and Kang S H 2008 *The Journal of the Korean Economy* **9** 403.
[17] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701
[18] Pastor-Sarorras R, Vazequez A and Vespignani A 2001 *Phys. Rev. Lett.* **87** 258701
[19] Song C M, Havlin S and Makse H A 2005 *Nature* **433** 392
[20] Goh K I, Salvi G, Kahng B and Kim D 2006 *Phys. Rev. Lett.* **96** 018701
[21] Kim J S, Goh K I, Kahng B and Kim D 2007 *New J. Phys.* **9** 177
[22] Gallos L K, Song C M and Makse H A 2007 *Physica* A **386** 686