Published in "Physics Procedia 3(5): 1867-1876 , 2010"
which should be cited to refer to this work.

Weighted Bipartite network and Personalized Recommendation

Xin Pan[a], Guishi Deng[a], Jian-Guo Liu[b,c,d]

[a]*Institute of Systems Science, Dalian University of Technology, Dalian 116024, P. R. China*
[b]*Research Centre of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, P. R. China*
[c]*Business School, University of Shanghai for Science and Technology, Shanghai 200093, P. R. China*
[d]*Department of Physics, University of Fribourg, Fribourg CH-1700, Switzerland*

Abstract

In this paper, the degree distributions of a bipartite network, namely Movielens, are investigated. The statistical analysis shows that the distribution of the degree product, $k_u k_o$, has an exponential from, where k_u and k_o denote the user and object degrees respectively. By introducing the edge weight effect on the recommendation performance, an improved recommendation algorithm based on mass diffusion (MD) process is presented. We argue that the edges weight of the user-object bipartite network should be taken into account to measure the object similarity. By taking into account the user and object degree correlations, the weighted bipartite network is constructed. The numerical results of the MD algorithms on the weighted network indicate that both of the accuracy and diversity could be increased at the optimal case. More importantly, we find that, at the optimal case, the edge weight distribution would change from the exponential form to the poisson form. This work may shed some light on how to improve the recommendation algorithm performance by considering the statistical properties.

Keywords: Personalized recommendation, network-based algorithm, mass diffusion, degree effects

1. Introduction

The last few years have witnessed tremendous activity devoted to the understanding of complex networks [1, 2, 3, 4, 5, 6, 7]. In recent years, the bipartite networks have catch lots of attentions. In the bipartite networks, the nodes are divided into two sets X and Y, and only the connection between two nodes in different sets is allowed. Many systems are naturally modeled as bipartite networks [8]: The scientific collaboration network consists of researchers and papers [9], the human sexual network consists of men and women [10], etc. With the advent of the Internet, the exponential growth of the World-Wide-Web [11] and routers confront people with an information overload. Consequently, how to efficiently help people obtain information that

they truly need is a challenging task nowadays [12]. Actually, most of these systems could be demonstrated by a bipartite network called the "opinion network" [13, 14], where each node in the user-set is connected with its collected objects in the object set. For example, personalized recommender systems have been used to recommend books and CDs at Amazon.com, movies at Netflix.com, and news at Versifi Technologies (formerly AdaptiveInfo.com) [15], and so on. A landmark for information filtering is the use of search engine, by which users could find the relevant web pages with the help of properly chosen keywords. However, the search engine has some disadvantages. For example, it does not take into account personalization and returns the same results for people with far different habits. Being an effective tool to address this problem, the recommender system has caught increasing attentions from researchers to engineers. Motivated by its significance in economy and society, the design of an efficient recommendation algorithm becomes a joint focus from engineering science to marketing practice. Various kinds of recommendation algorithms have been proposed, including the correlation-based methods [16, 17, 18, 19, 20, 21, 22], content-based methods [23, 24], iteratively self-consistent refinement [25], bipartite-network-based methods [26, 27, 28, 29], and so on (see the review article [15, 30] and the references therein).

Recently, Zhang *et al.* [26, 28] have successfully applied the classical physical processes, such as the heat conduction and mass diffusion, to deal with the personalized recommendation problem. The original algorithms require a kind of steady states, and to arrive at these states is time consuming. Zhou *et al.* [27, 29] thus proposed the simplified versions where only one step of heat conduction and/or random walk is taken into account. These simplified algorithms are considerably more accurate than the standard collaborative filtering and much faster with competitive accuracy compared with the matrix decomposition techniques [25]. In the above methods, all of the objects and users with far different degrees have been treated equally, in other words, the degree correlations between objects and users are neglected. For example, suppose a user with small-degree has collected a small-degree object, the edge connecting them represents a very special taste of the user, while the information contained in the edges connecting an active user and a popular object is less meaningful. Therefore, we argue that the user similarity index could be improved by considering the degree correlation of the user-object bipartite network.

This paper is organized as follows. Firstly, we investigate the degree distributions of the Movielens data. Secondly, the recommendation algorithm based on the mass diffusion process is presented. Thirdly, the improved algorithm is introduced. Finally, some conclusions and discussions are demonstrated.

2. Degree Distributions of MovieLens dataset

Suppose there are m objects and n users in a recommender system. Denote the object set as $O = \{o_1, o_2, \cdots, o_m\}$ and the user set as $U = \{u_1, u_2, \cdots, u_n\}$, a recommender system can be fully described by an adjacent matrix $A = \{a_{ij}\} \in R^{m,n}$, where $a_{ij} = 1$ if o_i is collected by u_j, and $a_{ij} = 0$ otherwise. A benchmark dataset, namely MovieLens[1], were used in this paper. The MovieLens data is a randomly-selected subset of the huge data, which consists of 1682 movies (objects) and 943 users. The user could rate collected objects from one to five. We argue that the ratings one user given to the objets could reflect their likelihood, the rating five indicates that he/her likes this object, while the rating one contains the dislike information. Since the weighted

[1] http://www.grouplens.org

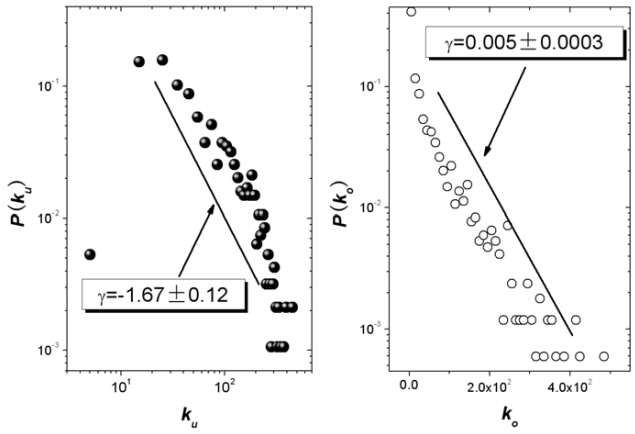

Figure 1: The degree distributions of user k_u and object k_o, from which one can see that the user degree distribution approximately has power-law form, where the exponent of the user degree distribution is $\gamma = -1.67 \pm 0.12$, while the one of the object has exponential form and the exponent is -0.015 ± 0.0003.

network is constructed to improve the user similarity definition, only the ratings which could reflect the users' interests or habits are considered, therefore, a movie is set to be collected by a user only if the giving rating is larger than 2. The original data contains 10^5 ratings, 82.52% of which are ≥ 3, that is, the user-object (user-movie) bipartite network after the coarse gaining contains 82,520 edges.

2.0.1. Degree distribution

The degree k_i of a node i is the number of edges incident with the node, and is defined in terms of the adjacency matrix A. The degree k_u of user u could be given as follows.

$$k_u = \sum_{i=1}^{m} a_{iu}. \tag{1}$$

The degree k_o of object o is computed as

$$k_o = \sum_{j=1}^{n} a_{oj}. \tag{2}$$

The most basic topological characterization of a bipartite network can be obtained in terms of the degree distribution $P(k)$, defined as the probability that a node chosen uniformly at random has degree k or, equivalently, as the fraction of nodes in the graph having degree k. The distributions of k_u and k_o are demonstrated in Fig.1, from which one can find that the distribution of user degree k_u approximately has the scale-free property while the one of object degree k_o approximately has

the exponential form. It should to be noticed that the above two distribution pictures are not clear. We investigate the distribution of the degree product of the user and object $k_u k_o$, which is used to measure the edge weight and implemented to the algorithm based on mass diffusion process. The exponential distribution is demonstrated in Fig.2.

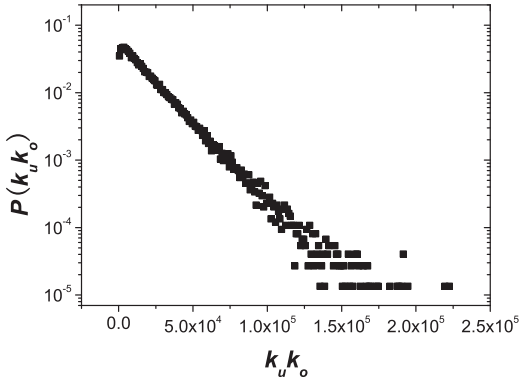

Figure 2: The distribution of the degree $k_u k_o$, which has an exponential form.

3. Improved recommendation algorithm based mass diffusion process

We assume a certain amount of resource (i.e. recommendation power) is associated with each object, and the weight w_{ij} represents the proportion of the resource o_j would like to distribute to o_i . For example, in the book-selling system, the weight w_{ij} contributes to the strength of recommending the book o_i to a customer provided he has already bought the book o_j. The weight w_{ij} can be determined following a network-based diffusion process [31] where each object distributes its initial recommendation power to all the users who have collected it, and then each user sends back what he has received to all the objects he has collected.

For a given user u_i, we assign some resource (i.e., recommendation power) on those objects already been collected by u_i. In the simplest case, the initial resource vector $\mathbf{f}$ can be set as

$$f_j = a_{ji}. \tag{3}$$

That is to say, if the object o_j has been collected by u_i, then its initial resource is unit, otherwise it is zero. After the resource-allocation process, the final resource vector is

$$\widehat{\mathbf{f}} = \mathbf{W}\mathbf{f}. \tag{4}$$

Accordingly, all u_i's uncollected objects o_j ($1 \leq j \leq n$, $a_{ji} = 0$) are sorted in the descending order of $\widehat{f}_j$, and those objects with highest values of final resource are recommended.

The *Mass diffusion*(MD) is akin to a random walk process on the bipartite user-object network [27]. Assigning objects on the network an initial level of resource denoted by the vector $\mathbf{f}$, we then redistribute it via the transformation $\widehat{\mathbf{f}} = \mathbf{W}\mathbf{f}$, where

$$w_{o_\alpha o_\beta} = \frac{1}{k_{o_\beta}} \sum_{i=1}^{n} \frac{a_{\alpha i} a_{\beta i}}{k_{u_i}} \tag{5}$$

is a column-normalized $m \times m$ probability matrix representing the diffusion process. Accordingly, the contribution of the edge connecting u_i and o_l should be $(k_{u_i} k_{o_l})^\lambda$, where λ is a tunable parameter. The edge weight is given in the following way

$$w_{ij} = (k_{u_i} k_{o_l})^\lambda. \tag{6}$$

Based on the weighted bipartite network, the object-object similarity of MD could be given as

$$w_{o_\alpha o_\beta} = \frac{1}{k_{o_\beta}} \sum_{i=1}^{n} \frac{a_{\alpha i} w_{\alpha i} a_{\beta i} w_{\beta i}}{k_{u_i}}. \tag{7}$$

Recommendations for a given user u_i are obtained by setting the initial resource vector f_i in accordance with the objects the user has already collected, that is, by setting $f_\alpha^i = a_{\alpha i}$. The resulting recommendation list of uncollected objects is then sorted according to $\widehat{f_i}$ in descending order.

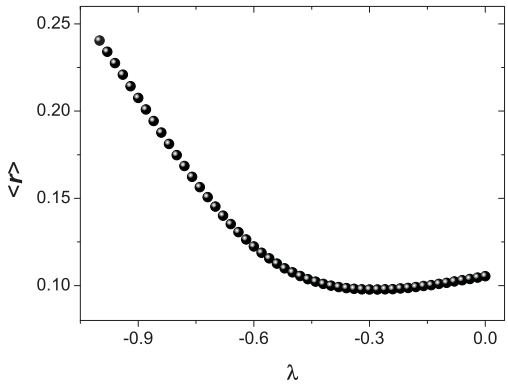

Figure 3: The average ranking score $\langle r \rangle$ vs. λ for the MD algorithm on the weighted bipartite network. The optimal λ_{opt} of MD algorithm, corresponding to the minimal $\langle r \rangle = 0.0977$, is $\lambda_{\mathrm{opt}} = -0.3$. All the data points are averaged over ten independent runs with different data-set divisions.

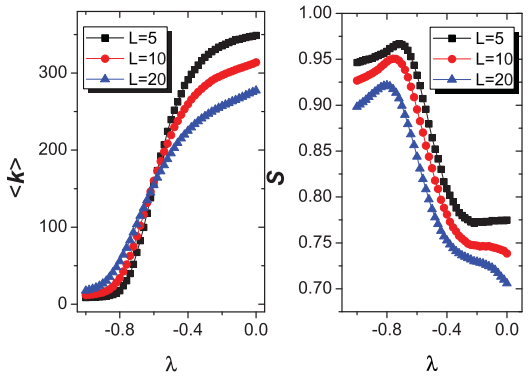

Figure 4: The popularity $\langle k \rangle$ and diversity S vs. λ for the MD algorithm on the weighted bipartite network. All the data points are averaged over ten independent runs with different data-set divisions.

4. Algorithmic performance metrics

4.1. Average ranking score

The average ranking score is adopted to measure the accuracy, which is defined as follows. For an arbitrary user u_i, if the entry u_i-o_j is in the probe set (according to the training set, o_j is an uncollected object for u_i), we measure the position of o_j in the ordered list. For example, if there are $L_i = 10$ uncollected objects for u_i, and o_j is the 3rd from the top, we say the position of o_j is 3/10, denoted by $r_{ij} = 0.3$. Since the probe entries are actually collected by users, a good algorithm is expected to give high recommendations to them, leading to small r_{ij}. Therefore, the mean value of the position r_{ij}, $\langle r \rangle$, averaged over all the entries in the probe, can be used to evaluate the algorithmic accuracy: the smaller the average ranking score, the higher the algorithmic accuracy, and vice verse.

4.2. Popularity

The average degree of all recommended objects, $\langle k \rangle$, and the mean value of Hamming distance, S, are taken into account to measure the algorithmic popularity and diversity [29]. The smaller average degree, corresponding to the less popular objects, are preferred since those lower-degree objects are hard to be found by users themselves.

4.3. Diversity

The personalized recommendation algorithm should present different recommendations to different users according to their tastes and habits. The diversity can be quantified by the average Hamming distance, $S = \langle H_{ij} \rangle$, where $H_{ij} = 1 - Q_{ij}(L)/L$, L is the length of recommendation list and $Q_{ij}(L)$ is the overlapped number of objects in u_i and u_j's recommendation lists. The largest $S = 1$ indicates the recommendations to all of the users are totally different, in other words, the

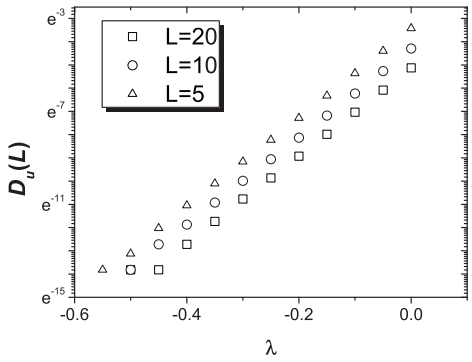

Figure 5: The diversity $D_u(L)$ of the user list vs. λ when $L = 5, 10$ and 20, from which one can see that the D_u would decrease when λ decreases.

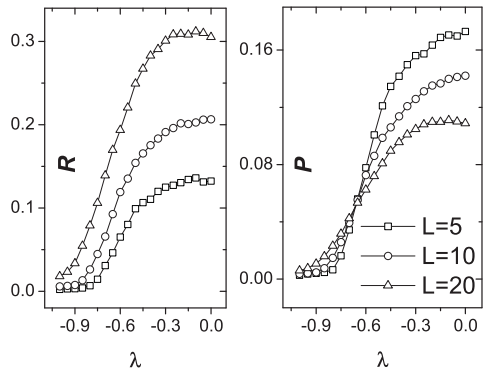

Figure 6: The precision P and recall R vs. λ to different recommendation list length L. One can find that both of P and R decrease when λ decreases.

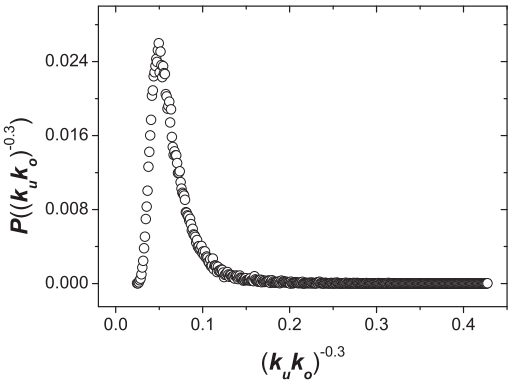

Figure 7: The distribution of the degree times $(k_u k_o)^{\lambda}$ at the optimal case $\lambda = -0.3$, which has a poisson form.

system has highest diversity. While the smallest $S = 0$ means all of recommendations are exactly same. In addition, the diversity $D_u(L)$ of the user lists is also investigated. In this paper, $D_u(L)$ is defined by using the similarity between the objects in the top-L recommended ones for each user, which could be given in the following way

$$D_u(L) = \frac{1}{n}\sum_{i=1}^{n}\frac{2}{(L-1)L}\sum_{j\in\Gamma_i(L)} w_{ij}, \tag{8}$$

where $\Gamma_i(L)$ is the object set which user i has not collected and ranked in the top-L position.

4.4. Precision and Recall

Recall is defined as the ratio of number of recommended objects appeared in the probe to the total number of data entries in the probe, which could be defined as

$$R = \frac{1}{N}\sum_{i}\frac{h_i}{N_{ip}}, \tag{9}$$

where h_i is the number of recommend objects appeared in user i's list and N_{ip} is the number of unselected objects of user i in the probe set. The larger recall corresponds to the better performance.

Precision is defined as the ratio of number of recommended objects appeared in the probe to the recommendation list length L. A larger precision corresponds to a better performance. Precision is also called the hitting rate in the literature.

5. Numerical results

Applying the improved algorithm on the Movielens data, the accuracy, popularity, diversity, precision and recall are investigated respectively. Figure 3 reports the algorithmic accuracy as

a function of λ. The curve has a clear minimum around $\lambda = -0.30$, which indicates that to depress the influence of the users or objects with large degrees could enhance the accuracy. Compared with the routine case ($\lambda = 0$), the average ranking score can be reduced by 7.3% at the optimal case. Figure 4 reports the average degree of all recommended movies as a function of λ. When $\lambda < 0$, $\langle k \rangle$ is positively correlated with λ, thus to depress the influence of edges connecting active users and popular objects gives more opportunity to the unpopular objects, which is consistent with our expectation. Figure 4 also exhibits the correlation between S and λ, indicating that to depress the influence of the edges connecting active users and popular objects makes the recommendations more personalized. In addition, the diversity of the user list, denoted by $D_u(L)$, is demonstrated in figure 5. One can find that, when L equals to 5, 10 and 20, $D_u(L)$ is exponentially correlated with λ. Increasing the parameter λ could enhance the diversity of each user list. Figure 6 shows that precision and recall of different L to the parameter λ, from which one can see that both of the P and R decrease as λ decrease from zero. In addition, the product distribution of the user and object degrees is investigated in Fig.7. One can see from Fig.7 that, different from the exponential distribution form of the product when $\lambda = 0$, the product distribution approximately has a poisson form at the optimal case.

6. Conclusions and discussions

In this paper, the edge weight of the bipartite network is investigated and one definition of edge weight is given according to the nodes connected by the edge, namely the user and object degrees. The edge weight is then introduced to the recommendation algorithm based on the mass diffusion process. By introducing a tunable parameter, the correlation between the parameter and the algorithmic performance in investigated. The numerical results shows that the accuracy, measured by the average ranking score, could be improved from 0.1054 to 0.0977 at the optimal case $\lambda = -0.3$, and both of the system diversity and the diversity of each user's list could be improved slightly. We also find that, at the optimal case, the edge weight distribution be changed from the exponential form to the poisson one. Although it's hard to find the correlation between the edge weight distribution and the algorithmic performance, the weighted bipartite network could be implemented in other structure-based algorithm to improve the accuracy and diversity. This work may shed some on the construction of weighted bipartite network and information filtering.

From this work, we could get the following conclusions and discussions. Firstly, the edge weight of the bipartite network affects the recommendation algorithm performance. In the improved algorithm, although the accuracy is only improved 7.3%, when $L = 5$, the diversity could be improved from 0.77 to 0.97, which is a great improvement. Secondly, the further work should focus on the correlation between the statistical properties and the algorithmic performance. Since there are less of metrics to describe the bipartite network, the basic measurements should be proposed and introduced to present the adaptive algorithm according to the network statistical properties.

Acknowledgement

We acknowledge *GroupLens* Research Group for providing us the data set. This work is partially supported by National Natural Science Foundation of China (Grant Nos. 10905052,

70901010, 10635040 and 60744003), the Swiss National Science Foundation (project 205120-113842), the Specialized Research Fund for the Doctoral Program of Higher Education of China.(Grant No. 20060358065), and Shanghai Leading Discipline Project (No. S30501).

Reference

[1] L. A. N. Amaral, A. Scala, M. Barthélémy, H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. 97, 11149 (2000).
[2] S. H. Strogatz, Nature 410, 268 (2001).
[3] R. Albert, A.-L. Barabási, Rev. Mod. Phys. 74, 47 (2002).
[4] S. N. Dorogovtsev, J. F. F. Mendes, Adv. Phys. 51, 1079 (2002).
[5] M. E. J. Newman, SIAM Rev. 45, 167 (2003).
[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Huang, Phys. Rep. 424, 175 (2006).
[7] L. da F. Costa, F. A. Rodrigues, G. Travieso, P. R. V. Boas, Adv. Phys. 56, 167 (2007).
[8] P. Holme, F. Liljeros, C. R. Edling, B. J. Kim, Phys. Rev. E 68, 056107 (2003).
[9] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Aberg, Nature 411, 907 (2001).
[10] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabási, Nature 407, 651 (2000).
[11] G.-Q. Zhang, G.-Q. Zhang, Q.-F. Yang, S.-Q. Cheng, T. Zhou, New Journal of Physics 10, 12307 (2008).
[12] P. Resnick, H. R. Varian, Commun. ACM 40, 56 (1997).
[13] S. Maslov, Y.-C. Zhang, Phys. Rev. Lett. 87, 248701 (2001).
[14] M. Blattner, Y.-C. Zhang, S. Maslov, Physica A 373, 753 (2007).
[15] G. Adomavicius, A. Tuzhilin, IEEE Trans. Know. & Data Eng. 17, 734 (2005).
[16] J. L. Herlocker, J. A. Konstan, K. Terveen, J. Riedl, ACM Trans. Inform. Syst. 22, 5 (2004).
[17] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, J. Riedl, Commun. ACM 40, 77 (1997).
[18] J.-G. Liu, B.-H. Wang, Q. Guo, Int. J. Mod. Phys. C 20, 285 (2009).
[19] J.-G. Liu, T. Zhou, B.-H. Wang, Y.-C. Zhang, Physica A 389, 881 (2010).
[20] R.-R. Liu, C.-X. Jia, T. Zhou, D. Sun, B.-H. Wang, Physica A 388, 462 (2009).
[21] D. Sun, T. Zhou, J.-G. Liu, R. -R. Liu, C. -X. Jia, B. -H. Wang, Phys. Rev. E 80, 017101 (2009).
[22] J.-G. Liu, T. Zhou, Z.-G. Xuan, H.-A. Che, B.-H. Wang, Y.-C. Zhang, arXiv:0907.1228.
[23] M. Balabanović, Y. Shoham, Commun. ACM 40, 66 (1997).
[24] M. J. Pazzani, Artif. Intell. Rev. 13, 393 (1999).
[25] J. Ren, T. Zhou, Y.-C. Zhang, EPL 88, 38005, (2009).
[26] Y.-C. Zhang, M. Medo, J. Ren, T. Zhou, T. Li, F. Yang, Europhys. Lett. 80, 68003 (2008).
[27] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Phys. Rev. E 76, 046115 (2007).
[28] Y.-C. Zhang, M. Blattner, Y.-K. Yu, Phys. Rev. Lett. 99, 154301 (2007).
[29] T. Zhou, L.-L. Jiang, R.-Q. Su, Y.-C. Zhang, Europhys. Lett. 81, 58004 (2008).
[30] J.-G Liu, M.Z.-Q. Chen, J. Chen, F. Deng, H.-T. Zhang, Z. Zhang, T. Zhou T. Int. J. Inf. and Sys. Sci. 5(2), 230 (2009).
[31] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin, Phys. Rev. E 75, 021102 (2007).