Predicting the future trend of popularity by network diffusion

An Zeng and Chi Ho Yeung'

Citation: Chaos **26**, 063102 (2016); doi: 10.1063/1.4953013
View online: http://dx.doi.org/10.1063/1.4953013
View Table of Contents: http://aip.scitation.org/toc/cha/26/6
Published by the American Institute of Physics

Predicting the future trend of popularity by network diffusion

An Zeng[1] and Chi Ho Yeung[2,a)]
[1]*School of Systems Science, Beijing Normal University, Beijing, People's Republic of China*
[2]*Department of Science and Environmental Studies, The Education University of Hong Kong, Taipo, Hong Kong*

(Received 13 July 2015; accepted 18 May 2016; published online 1 June 2016)

Conventional approaches to predict the future popularity of products are mainly based on extrapolation of their current popularity, which overlooks the hidden microscopic information under the macroscopic trend. Here, we study diffusion processes on consumer-product and citation networks to exploit the hidden microscopic information and connect consumers to their potential purchase, publications to their potential citers to obtain a prediction for future item popularity. By using the data obtained from the largest online retailers including Netflix and Amazon as well as the American Physical Society citation networks, we found that our method outperforms the accurate short-term extrapolation and identifies the potentially popular items long before they become prominent. *Published by AIP Publishing*. [http://dx.doi.org/10.1063/1.4953013]

Anticipating the sales of a product accurately is important for retailers as they can increase the inventory of a product accordingly, to avoid the potential loss due to the shortage of inventory and the unsatisfied demand of some customers. On the other hand, predicting the citation of a research paper at its early stage can help identify important research works to accelerate scientific progress. Although it may be trivial to predict the short-term popularity of an item by extrapolating its present popularity, it is difficult to predict its long-term trend which is dependent on the intrinsic value of the product. Conventional macroscopic information is not sufficient, and detailed microscopic information hidden under the macroscopic trend has to be used. In this paper, we exploit the microscopic structure of the consumer-product network and the citation network to identify product or research papers with long-term potential. We show that although our method is not as accurate as extrapolation to predict the immediate popularity, it outperforms extrapolation in the long run and identifies the potentially popular items long before they become prominent.

I. INTRODUCTION

Predicting the future based on the current state or historical data is a crucial task in many applications. While mathematical models and prediction techniques are sufficiently mature for some systems such as numerical weather forecast,[1,2] reliable prediction approaches are missing in many other applications such as earthquake prediction.[3] One may identify chaotic nature to be the major difficulty, yet the lack of understanding of the underlying principles may indeed be the real obstacle. For instance, applications such as predictions for financial crisis[4] or epidemic outbreaks[5] would be more feasible if one can identify the fundamental mechanism despite their high stochasticity.

Identifying the underlying mechanism in various systems has been a long lasting research topic, especially due to its potential to make important predictions. While these studies are usually interdisciplinary, physicists play a unique role to make macroscopic predictions based on microscopic interactions. For instance, while techniques such as time series analysis[6,7] are widely applied to predict stock price, physicists build simple models to mimic individuals interacting in financial markets;[8] the established mechanism is applied on virtual trading.[9] Similarly, models to predict epidemic spreading[10,11] are also studied. Such microscopic approaches often reveal phenomena overlooked by conventional studies. In contrast, conventional approach based on extrapolation ignores the underlying mechanism and may lead to less accurate prediction.

Among the various kinds of predictions, the growth of commodity sales and the citation to publications have been extensively studied.[12,13] In particular, the Bass diffusion model[12] predicts sales dynamics by considering macroscopic information, for instance, drawing relation between the number potential buyers to that of existing buyers. A similar rationale is also employed to explain the growth of node degree in networks.[14] Physicists then play a crucial role in understanding the fundamental consumer-product interactions as an origin of macroscopic phenomenon. For instance, models are introduced to explain the spread of news among readers in online social networks.[15,16] The competition between movies is modeled as interactions among viewers and reviewers on online recommendation systems.[17] On the other hand, citation networks are found to exhibit self-similar structure[18] and the increase in paper citation[19] was modeled by dynamical equations.[20] Future publication citations and journal impact factors are predicted in Refs. 21 and 22, but instead of the detailed network topology, only the number of obtained citations was considered in the predictions. Unlike previous studies which focus on the macroscopic dynamics, the physics approach incorporates individual interactions for making predictions.

a)Electronic mail: chyeung@ied.edu.hk

In this paper, we explore the microscopic consumer-product interaction to identify the potential popular commodities in the future. The group of products we are going to study are durable goods such as books and DVDs where replacement is usually not required. Specifically, we make use of the diffusion process to correlate each buyer with their potential purchase, which collectively provide a macroscopic trend for all products. Unlike the conventional studies of sales dynamics,[12,13] we make use of the microscopic relations between buyers and products to make predictions. We will apply our diffusion process on the real data obtained from *Movielens*, *Netflix*, *Delicious*, and *Amazon*, and show that in all the datasets our method outperforms the accurate short-term predictions obtained by linear extrapolation and identifies the potentially popular items long before they become prominent. These results indicate that microscopic information is crucial in identifying quality products. Finally, we apply the diffusion method to the American Physical Society (APS) citation network and found that our method works effectively in predicting the future trend of research paper popularity.

II. METHOD

Specifically, we analyze the data collected by online retailers where consumers are allowed to rate or review individual products on their web sites. Such relation can be projected on bipartite networks with two set of nodes representing users and items, respectively, and a link exists when a user have rated an item. These networks are growing since new reviews occur as time evolves. We thus denote $N(t)$ and $Z(t)$ to be the number of users and the number of items, respectively, at time t. The topology of the network is characterized by a time-evolving *adjacency matrix* $A(t)$, where the element $a_{i\alpha}(t) = 1$ if user i has rated item α at time t and $a_{i\alpha}(t) = 0$ otherwise (throughout the paper we will use Latin and Greek letters for indices to label users and items, respectively). The *degree* $k_\alpha(t)$ of an item α is defined as the number of users who have rated α at time t, i.e., $k_\alpha(t) = \sum_i a_{i\alpha}(t)$, which can be interpreted as the *cumulative popularity* of α. As the sales data are usually confidential, such review data would serve as a representative of product sales in our study.

Our task is to identify the potentially popular items in the future. Instead of the cumulative popularity, the number of new buyers would be a better indication for the concurrent popularity. For instance, the movie *Jurassic Park* was extremely popular in the 1990s, but in 2013 we do not expect it has the same number of new viewers compared to recent movies such as *Avatar*, since most viewers who are interested in *Jurassic Park* have already watched it, while there may exist interested viewers who have not watched *Avatar*. We thus quantify the *popularity* of item α between time t and $t + T_f$, where T_f is a period of time to the future after time t, to be $\Delta k_\alpha(t; t + T_f)$ instead of $k_\alpha(t + T_f)$, such that

$$\Delta k_\alpha(t; t') = k_\alpha(t') - k_\alpha(t). \tag{1}$$

Our goal is to identify the items with high $\Delta k(t; t + T_f)$ within a given period of T_f in the future based on the existing data at time t.

Finally, we define a *ranking operator* $\mathcal{R}[x_\alpha]$ which ranks a quantity x_α of item α among the corresponding quantity of all items. The item which comes first in the rank is given a rank value $\mathcal{R}[x_\alpha] = 1$, while the item which comes last in the rank is given a rank value $\mathcal{R}[x_\alpha] = Z$. For instance, $\mathcal{R}[k_\alpha(t)] = 1$ implies that item α has the highest cumulative popularity at time t. When m nodes have equal values of x_α, and there are n other nodes having x_α higher than that, the ranks of the first ones are assigned randomly from $n + 1$ to $n + m$, respectively. Therefore, if there are multiple nodes with the same x_α, the last value of the rank is still equal to Z.

A. Linear extrapolation—The Benchmark method

Before describing our prediction method, we describe the benchmark method called *Linear extrapolation* (LE) with which we compare our results. In our context, linear extrapolation assumes that the future increment of product sales or the citation counts of a paper is proportional to its existing popularity. Similar idea was introduced before to explain the growth of sales[12] and is indeed similar to the preferential attachment mechanism which explains network growth.[14] It is straightforward to use LE as an approach to predict sales. Since LE assumes rich-get-richer, we can use the present rank of product popularity directly as the prediction for their future rank. In this case, given the cumulative popularity of all items at time t, the predicted rank $\tilde{R}_\alpha^{(\mathrm{LE})}$ of the popularity of item α at time T_f is given by

$$\tilde{R}_\alpha^{(\mathrm{LE})}(t, T_f) = \mathcal{R}[k_\alpha(t)]. \tag{2}$$

One may argue that this simple approach is dominated by items with large cumulative degree, for instance *Jurassic Park* in the above example, while the future popularity is more dependent on recent demand. Thus, instead of the cumulative degree $k_\alpha(t)$, one can use the degree increment $\Delta k_\alpha(t - T_h; t)$, where T_h is a period of time to the past before time t, to predict the future rank

$$\tilde{R}_\alpha^{(\mathrm{LE})}(t, T_f, T_h) = \mathcal{R}[\Delta k_\alpha(t - T_h; t)]. \tag{3}$$

The rank predicted by Eq. (3) may thus reflect a more recent popularity of the product compared to that predicted by Eq. (2).

In principle, LE and the well-known preferential attachment only differ in the fact that the latter is probabilistic. In the context of network growth, various studies have shown that preferential attachment does not perfectly characterize the increase of node degrees. For instance, microscopic interactions involving quality and relevance[20] play a significant role. We thus expect that microscopic interactions are also crucial in identifying potentially popular items.

B. Microscopic diffusion method

Here we introduce our diffusion-based prediction method which makes use of the microscopic relations between buyers and products. The rationale is to examine the similarity of the group users who have collected the item; if the similarity among the group is high, we believe that this group of users

did not select the object by chance, but instead the product is of high quality in a particular interest group and has attracted a group of similar users. As a result, more similar users will collect the item which contributes to its increasing popularity. In other words, similarity between the users who collected the product is an indirect indication of the product quality in a specific interest group. It is different from the case that a product is collected by a group of random users who are not similar in their interests. We note that the diffusion processes have been applied to design the *recommendation algorithms* which can be regarded as a prediction problem[23,24] as well. In general, these algorithms utilize the history of users to infer their future preferences.

Specifically, we devise a self-avoiding mass diffusion (SAMD) method which explores the user-item bipartite networks to predict the future popularity of products. The SAMD method is given as follows: for each item α, we assign one unit of resource to each of the users who collected α and then redistribute the resource on the user-item network through a self-avoiding diffusion. Suppose predictions are made for a target item α at time t. Mathematically, we denote $m_i^{(\alpha)}(0) = a_{i\alpha}(t)$ to be the initial resources on the users before the diffusion. We then denote $m_i^{(\alpha)}(n)$ to be the resource on the user i at the n-th step of diffusion, and the diffusion is represented by $\vec{m}^{(\alpha)}(n+1) = \mathbf{W}(t) \cdot \vec{m}^{(\alpha)}(n)$, where $\vec{m}^{(\alpha)}(n)$ is a N-component vector with the i-th element corresponds to $m_i^{(\alpha)}(n)$, and $\mathbf{W}(t)$ is given by

$$W_{ij}(t) = \frac{1-\delta_{ij}}{k_i(t)} \sum_{\beta=1}^{Z} \frac{a_{i\beta}(t)a_{j\beta}(t)}{k_\beta(t)}, \tag{4}$$

which is a conventional diffusion matrix with $k_i(t) = \sum_{\gamma=1}^{Z} a_{i\gamma}(t)$ and $k_\beta(t) = \sum_{l=1}^{N} a_{l\beta}(t)$ denoting the cumulative degree of user j and item β, respectively, at time t. In the expression for $W_{ij}(t)$, the Kronecker delta $\delta_{ij} = 1$ if $i = j$, which makes sure the diffusion is self-avoiding. We note that each step (i.e., an increase of n by 1) includes a diffusion from the user side to the item side, and then a diffusion from the item side back to the user side. The first diffusion identifies all the items selected by a certain user i, and the second one identifies the users who have selected the same items as i. Therefore, the resource $m_j^{(\alpha)}(1)$ user j receives from user i after one step (i.e., the two diffusions) can be regarded as the similarity between user i and user j. To ensure mass conservation in the diffusion process, one needs to replace $k_\beta(t)$ in Eq. (4) by $k_\beta(t) - 1$. However, in this case, one has to exclude all the items with $k_\beta(t) = 1$ in the diffusion process since the denominator of Eq. (4) for these items would become zero. In order to keep the formulation, the process, and the implementation simple, we adopted the present form of Eq. (4) as a non-conservative diffusion. We found that the performance of SAMD is closely similar to that if $k_\beta(t)$ in Eq. (4) is replaced by $k_\beta(t) - 1$. After the whole diffusion process, we sum up the resources on all the users who have collected item α, given by

$$M^{(\alpha)}(t) = \sum_{i=1}^{N} a_{i\alpha}(t) m_i^{(\alpha)}(1), \tag{5}$$

which is denoted as the *diffusion score*. The higher the diffusion score, the more similar are the users who have already collected item α.

As we mentioned in Subsection II A, future trend should be inferred from the recent demand. To consider only the recent purchases, we modify the initial condition and $\mathbf{W}$ in Eq. (4) as follows: (i) the initial condition is modified as $m_i^{(\alpha)}(0) = a_{i\alpha}(t) - a_{i\alpha}(t - T_h)$ and (ii) in the diffusion from users to users we only consider connections between time $t - T_h$ and t, such that the modified $\mathbf{W}'(t, T_h)$ becomes

$$W'_{ij}(t, T_h) = \frac{1-\delta_{ij}}{\Delta k_i(t-T_h;t)} \sum_{\beta=1}^{Z} \frac{\Delta a_{i\beta}(t-T_h;t)\Delta a_{j\beta}(t-T_h;t)}{\Delta k_\beta(t-T_h;t)}, \tag{6}$$

where $\Delta a_{j\beta}(t-T_h;t) = a_{j\beta}(t) - a_{j\beta}(t-T_h)$, and $\Delta k_i(t-T_h;t)$ and $\Delta k_\beta(t-T_h;t)$ are given by Eq. (1); since $t \geq t-T_h$, $\Delta k_i(t-T_h;t) \geq 0$ and $\Delta k_\beta(t-T_h;t) \geq 0$. We remark that whenever the denominator of $W'_{ij}(t,T_h)$ becomes zero, i.e., $\Delta k_i(t-T_h;t) = 0$ or $\Delta k_\beta(t-T_h;t) = 0$, the numerator of $W'_{ij}(t,T_h)$ is also zero such that $W'_{ij}(t,T_h) = 0$ and $\mathbf{W}'(t,T_h)$ remains a diffusion matrix. Modification (i) identifies the recent buyers of item α. This is crucial if user i has recently shifted his or her interest, for instance, from comedies to sci-fi movies. Modification (ii) calculates user similarity based on recent activities between time $t-T_h$ and t only. A schematic diagram which shows the diffusion process is given by Fig. 1. The predicted ranking of popular items is given by

$$\tilde{R}_\alpha^{(\mathrm{SAMD})}(t, T_f, T_h) = \mathcal{R}[M'^{(\alpha)}(t, T_h)], \tag{7}$$

where

$$M'^{(\alpha)}(t, T_h) = \sum_{i}^{N} \Delta a(t - T_h; t) m_i'^{(\alpha)}(1), \tag{8}$$

and

$$\vec{m}'^{(\alpha)}(1) = \mathbf{W}'(t, T_h) \cdot \vec{m}'^{(\alpha)}(0). \tag{9}$$

We remark that the quantities M' and m' are computed by the matrix $\mathbf{W}'(t, T_h)$ which make use of the recent data between time $t - T_h$ and time t only. Similar to Eq. (5), the higher the diffusion score $M'^{(\alpha)}$, the more similar are the users who have already collected item α. Although it has been shown in Ref. 25 that the mass diffusion process leads to a similar macroscopic trend as LE, we will see in Sec. III that the two methods have significant differences microscopically, which contributes to the success of SAMD in identifying potentially popular items. In fact, the diffusion score can also be used as a measurement for node centrality. Similar applications have been discussed in Ref. 26.

C. Metrics

We then describe the metrics we employed to measure the prediction accuracy. There are many ways to quantify *prediction precision*, here we define it to be the fraction

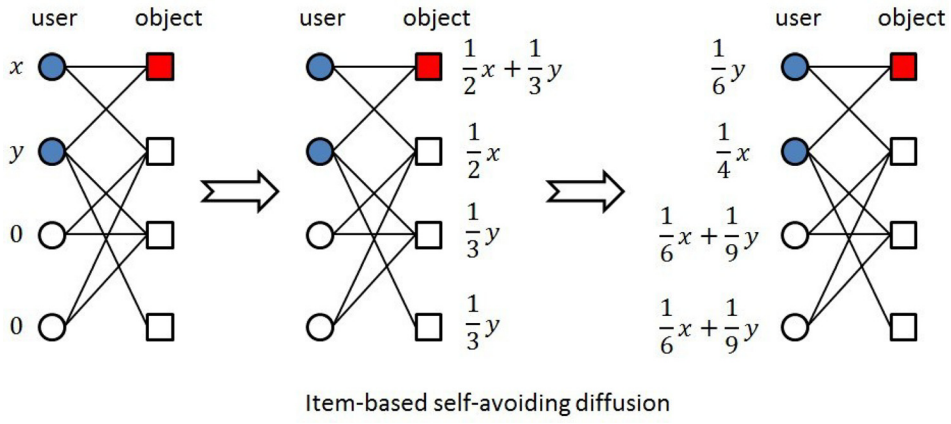

FIG. 1. A simple example to illustrate the diffusion process on bipartite network. The red square is the target item for popularity prediction. The blue circles are the users who have already selected the target item. After the diffusion, only the score of the blue circles are taken into account. In SAMD method, $x = y = 1$.

$Q_L(t, T_f)$ of the top-L predicted items to be found in the actual top-L list of popularity. In other words, $Q_L(t, T_f)$ is given by

$$Q_L(t, T_f) = \frac{n\left(\mathcal{R}\left[\Delta k_\alpha(t; t+T_f)\right] \le L \cap \tilde{R}_\alpha(t, T_f, T_h) \le L\right)}{n\left(\mathcal{R}\left[\Delta k_\alpha(t; t+T_f)\right] \le L\right)}, \tag{10}$$

where $n(S)$ corresponds to the number of elements in the set S, $\mathcal{R}[\Delta k_\alpha(t; t+T_f)]$ is the actual ranking of degree increase of α in the period $t \le t' \le t+T_f$, and $\tilde{R}_\alpha(t, T_f, T_h)$ is the predicted ranking either by Eq. (3) or (7).

Besides the precision, we also consider the Kendal's tau ranking correlation between the top-L items.[27] We consider all the items with $\mathcal{R}[\Delta k_\alpha(t; t+T_h)] \le L$ as the candidates (denoted as Γ_L) and calculate τ between their $\tilde{R}_\alpha(t, T_f, T_h)$ and $\mathcal{R}[\Delta k_\alpha(t; t+T_f)]$ where $\alpha \in \Gamma_L$.

In commercial web sites, online retailers usually only show top-10 or top-20 popular items on their first page. This indicates that top-10 and top-20 popular items are most interested by the online retailers as well as by the customers. We thus first fix this number to be a constant 20 for every dataset. In order to see how general our result is, we test as well top-50 and top-100 in each figure.

D. Data description

We will apply our prediction method on four datasets, obtained, respectively, from *Movielens*, *Netflix*, *Delicious*, and *Amazon*. The data from Movielens and Netflix are review rating given by users on movies, which can be mapped into viewer-movie bipartite networks with links exist when the rating is higher than a threshold value. Similarly, the data of bookmarks collected by users in Delicious and review written by users on books in the Amazon can be mapped into user-bookmark and reader-book bipartite networks, respectively. The time at which each individual review is written is known. As the original datasets are large, sampling is done by randomly picking N users with at least 20 movie ratings in Movielens or Netflix and all their collected items. The number of users, items, and links and the period where the data are collected are given in Table I.

III. RESULTS

A. Accuracy-oriented prediction

Before discussing results obtained by the SAMD method, we examined the performance of LE in predicting the top-L most popular item in the future. We observe similar results in four datasets, i.e., LE performs best by using the most recent history (small Δt_h) to predict the most near future (small Δt_f). The prediction accuracy decreases rapidly for long-term prediction. Moreover, the prediction is worse when historical information increases beyond $T_h = 90$ days. The optimal history length is roughly $T_h = 90$ days in all the four datasets.

In order to compare SAMD fairly with LE, we adopt the history length $T_h = 90$ days at which LE performs best. The dependence of Q_L and τ_L on Δt_f in these two methods is shown in Figs. 2 and 3, respectively. In Movielens, Netflix, and Delicious, Q_L obtained by SAMD decreases more slowly than that obtained by LE. These results indicate that LE focuses more on predicting the timely popular items, while SAMD discovers items with real high quality and these items will constantly attract attention even after a long time. The advantages of SAMD in these three datasets are more obvious in τ_L. As we can see from both Figs. 2 and 3, the performance of SAMD is consistent in all prediction length L considered. However, we also observe that SAMD cannot

TABLE I. Statistics of the datasets: The number of users, items, and links are, respectively, denoted by N, Z, and E.

Dataset	N	Z	E	Start date	End date
Movielens	5000	7533	7.3×10^5	1st January, 2002	1st January, 2005
Netflix	4968	16 331	1.2×10^7	1st January, 2000	31st December, 2005
Delicious	223 628	23 338	1.2×10^6	1st September, 2003	1st October, 2007
Amazon	99 621	383 553	1.0×10^6	31st May, 1996	15th September, 2005

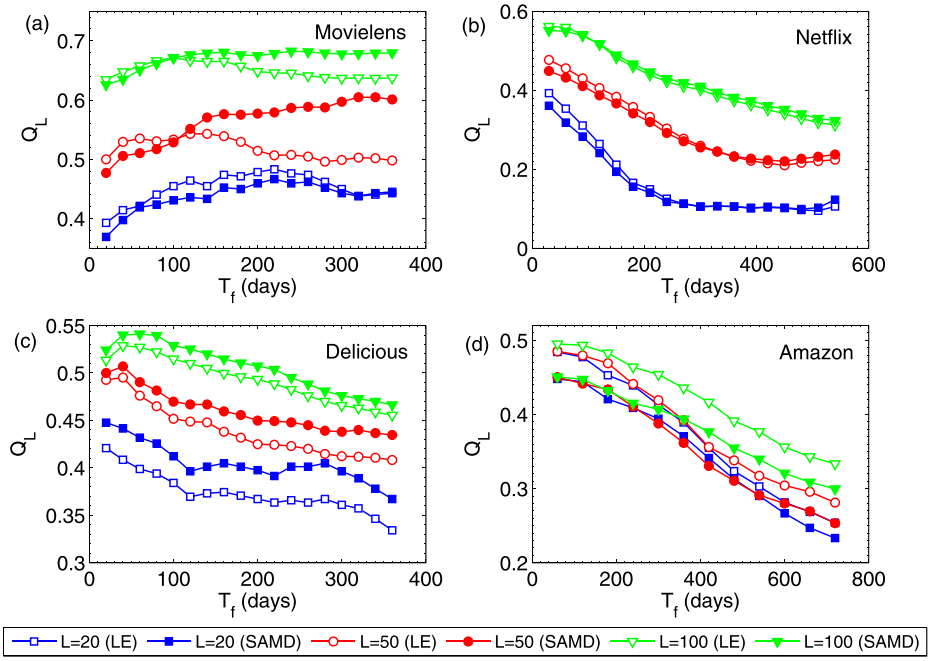

FIG. 2. The dependence of Q_L on T_f for the top-L item prediction by SAMD and LE in (a) Movielens, (b) Netflix, (c) Delicious, and (d) Amazon. $T_h = 90$ days with which the LE method roughly achieves its best performance. The results are averaged over different testing time t.

outperform LE in the Amazon dataset, which we will discuss in Subsection III B.

To further understand the benefit of SAMD over LE, we spot the top-L items which are successfully predicted by SAMD but not LE in the space of $\Delta k(t-T_h;t)$ against $\Delta k(t;t+T_f)$. We observe that SAMD is able to identify some future top-L items which were not popular in the past, i.e., items with relatively small $\Delta k(t-T_h;t)$ but large $\Delta k(t;t+T_f)$. However, we also remark that SAMD is not able to identify the top-L items with very small $\Delta k(t-T_h;t)$ such as between 1 and 20. Some of these niche items may become popular in the future, but they have small current degree because they just enter the market. Therefore, a method which can identify these potential items is necessary and of great importance.

B. Preferential diffusion

As shown above, SAMD cannot outperform LE in the Amazon dataset. To identify the potential items in the dataset, we slightly modify SAMD and introduce a preferential diffusion mechanism. The transition matrix $\mathbf{W}'(t,T_h)$ becomes

$$W'_{ij}(t,T_h) = \frac{1-\delta_{ij}}{\mathcal{M}} \sum_{\beta=1}^{Z} \frac{\Delta a_{i\beta}(t-T_h;t)\Delta a_{j\beta}(t-T_h;t)}{\left(\Delta k_\beta(t-T_h;t)\right)^{1-\theta}}, \quad (11)$$

where $\mathcal{M}$ is a normalization factor which equals to $\sum_{\beta=1}^{Z} \Delta a_{i\beta}(t-T_h;t)(\Delta k_\beta(t-T_h;t))^\theta$. Eq. (11) reduces to the ordinary SAMD when $\theta = 0$. We test the preferential SAMD method on two sparse datasets, namely, Delicious and

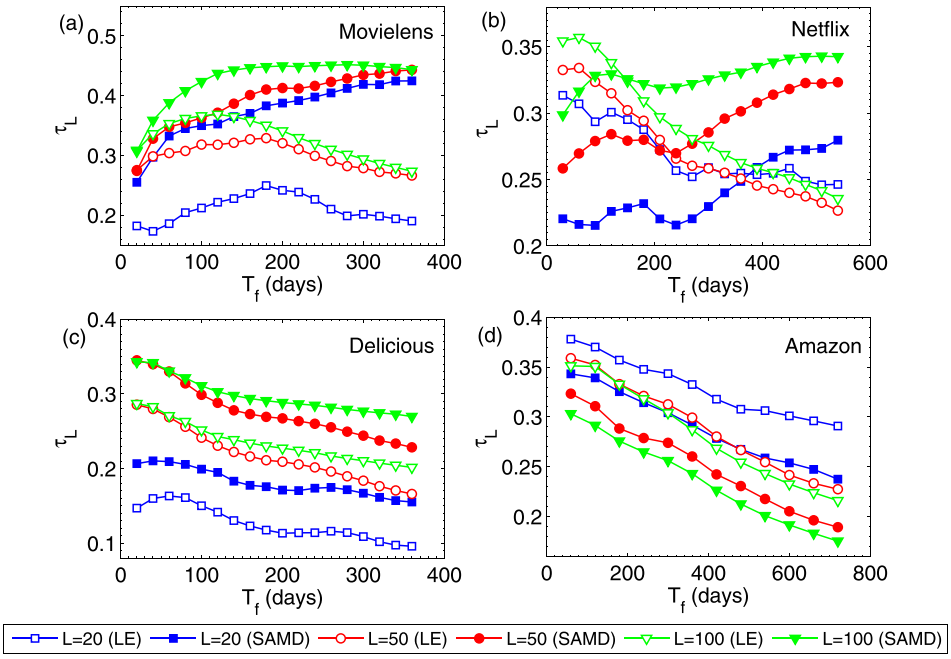

FIG. 3. The dependence of τ_L on T_f for the top-L item prediction by SAMD and LE in (a) Movielens, (b) Netflix, (c) Delicious, and (d) Amazon. $T_h = 90$ days at which LE approximately achieves its best performance. The results are averaged over different testing time t.

Amazon, and report the results in Fig. 4. Clearly, both Q_L and τ_L increase significantly with θ. This phenomenon is especially obvious in the Amazon dataset. Specifically, preferential SAMD outperforms LE when $\theta > 1$. We also tested the preferential SAMD method in the dense datasets, namely, Movielens and Netflix where interestingly the optimal θ is around 0.

To explain the effectiveness of preferential SAMD in Delicious and Amazon, we first note that a positive θ in Eq. (11) increases the contribution by the more popular items in the diffusion process. A positive optimal θ may imply that the information about user preference is strongly embedded in the popular items, and the less popular items carry little information. This happens especially in sparse networks where a large number and variety of items are present, such that the less popular items are only collected by a small number of users and may not provide any statistical relevant information.

C. Potential-oriented prediction

One of our ultimate goals is to identify potentially popular item long before they become prominent. In order to achieve this objective, we focus on the less popular items with $\Delta k_\alpha(t - T_h; t)$ from 1 to 50 in Movielens and Netflix, 1–30 in Delicious, and 1–25 Amazon (since Delicious and Amazon datasets are more sparse). We identify items with the same $\Delta k_\alpha(t - T_h; t)$. The LE method does not distinguish the future prospect of these items with the same current popularity $\Delta k_\alpha(t - T_h; t)$, while SAMD does. It is thus meaningful to apply SAMD to identify potentially popular items. For instance, a non-zero correlation between the diffusion score $M_\alpha(t, T_h)$ and future degree increase $\Delta k_\alpha(t; t + T_f)$ would help identify the items with the greatest potential among the items with the same $\Delta k_\alpha(t - T_h; t)$.

Here, we measure the Kendal's tau correlation between $M^{(\alpha)}(t, T_h)$ and $\Delta k_\alpha(t; t + T_f)$ and report τ of the items with different $\Delta k_\alpha(t - T_h; t)$ in Fig. 5. One can easily see that τ is positive in almost all $\Delta k_\alpha(t - T_h; t)$, which indicates that the items with high $M_\alpha(t, T_h)$ are more likely to have high popularity increment in the future. In other words, items with large diffusion score are those with great potential. Moreover, τ generally increases with $\Delta k_\alpha(t - T_h; t)$. This is because the amount of information for prediction increases with $\Delta k_\alpha(t - T_h; t)$, so that our method becomes more accurate. Note that τ of LE is always 0 in Fig. 5.

D. Application on the APS citation networks

Predicting highly cited papers long before they become prominent is useful in facilitating scientific development. Actually, our method can be easily extended to monopartite directed networks such as citation networks. We use the APS citation data as an example.[28] The APS data consist of 449 935 papers with 4 672 812 citations from 1897 to 2009. When we extend SAMD to citation networks, the diffusion start from the papers citing the target paper to their references, and go backward from these references to the papers which cited them.[18] The results are reported in Fig. 6 where the training period is set to be $T_h = 5$ years in each panel.

Figs. 6(a) and 6(b) show the performance of the preferential SAMD method in the APS datasets. Like the Amazon dataset, the prediction accuracy Q_L and τ_L can be improved by θ, with an optimal $\theta = 1$. In Figs. 6(c) and 6(d), we compare the performance of the SAMD method ($\theta = 1$) and LE. One can see that SAMD is comparable with LE in Q_L and better than LE in τ_L. In Fig. 6(e), which show the power of SAMD ($\theta = 1$) in detecting papers of great potential. One can see that the correlation is always positive, which indicates that SAMD can distinguish the potential papers among all the papers with the same recent increment of citation count $\Delta k_\alpha(t - T_h; t)$. As an example, we set $\Delta k_\alpha(t - T_h; t) = 25$ and plot the evolution of the mean citation count of the top-three ranked papers in Fig. 6(f), compared to the average of all the papers with $\Delta k_\alpha(t - T_h; t) = 25$. Clearly, the top-three potential papers identified by our method are cited

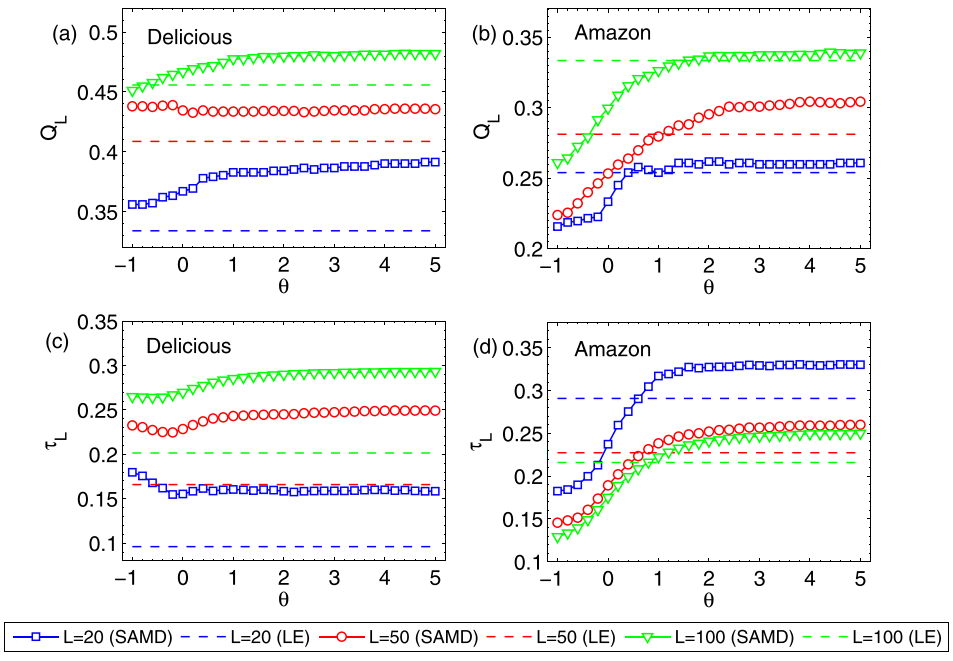

FIG. 4. The performance of the preferential SAMD method in (a) and (c) Delicious and (b) and (d) Amazon datasets. The history length $T_h = 90$ days in both datasets. The future length is $T_f = 365$ days in Delicious and $T_f = 730$ days in Amazon. The results are averaged over different testing times t.

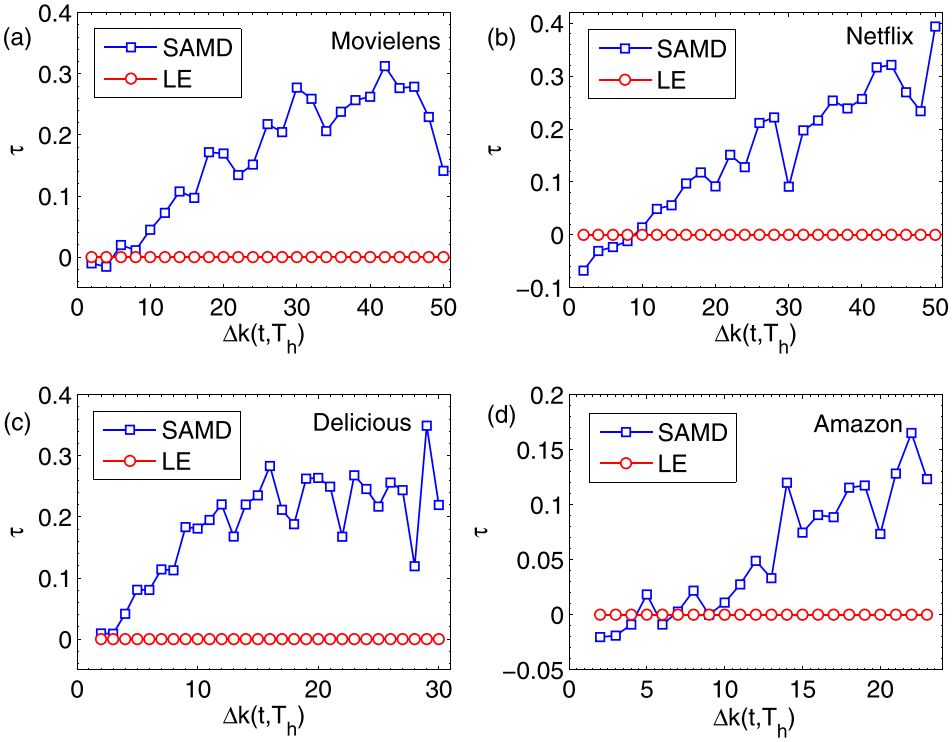

FIG. 5. The kendal's tau correlation between $M_\alpha(t,T_h)$ and $\Delta k_\alpha(t;t+T_f)$ among items with the same $\Delta k_\alpha(t-T_h;t)$ in (a) Movielens, (b) Netflix, (c) Delicious, and (d) Amazon datasets. The history length $T_h=90$ days in all datasets. The future length is $T_f=365$ days in Movielens, Delicious, and $T_f=540$ days in Netflix, and $T_f=730$ days in Amazon. In Delicious and Amazon dataset, $\theta=2$. The results are averaged over different choices of testing times t.

much more times than the average. We remark that the results in Figs. 6(e) and 6(f) show that SAMD can identify the papers with high potential, and are consistent with the findings in Subsection III A, III B, and III C.

In fact, prediction of individual paper citation is a hot research topic in recent years and a number of works have been devoted to derive such predictions.[21,22] However, most of these studies use only the evolution history of the number of citations of a paper (i.e., the time series of its citation history), but ignore the detailed topological information in the citation network (i.e., the relations between a paper and its cited papers). In our SAMD, we utilize these topological information to predict the future trend of the paper citations, which is different from the previous two studies which only utilize the more macroscopic information on the evolution history of the citation of papers. As shown in our paper, we find that this topological information is valuable for citation prediction. By using this information, SAMD outperforms

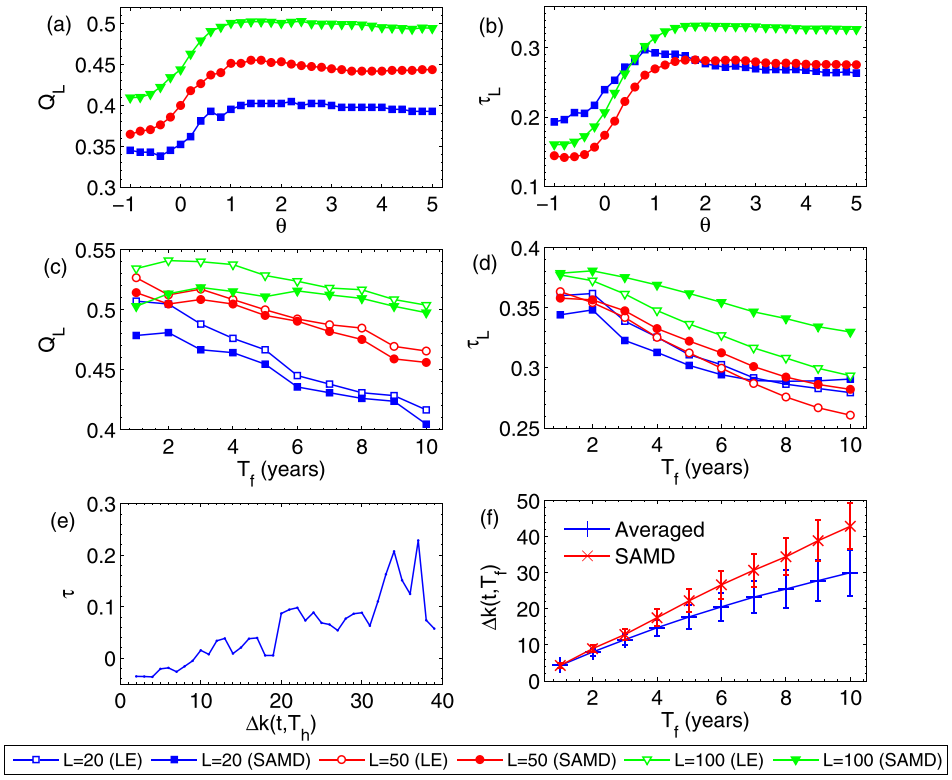

FIG. 6. (a) and (b) show the performance of the preferential SAMD method in the APS dataset. (c) and (d) compare the performance of the SAMD method ($\theta=1$) and that of the LE method in the APS dataset. (e) shows the ability of SAMD ($\theta=1$) in detecting the potential papers. (f) The evolution of the mean citation count of the top-three ranked papers among all the papers with $\Delta k_\alpha(t-T_h;t)=25$. $T_h=5$ years and $T_f=10$ years in (e). The results are averaged over different testing time t.

the linear extrapolation method. As such, our method can be considered as complementary to the existing methods.

IV. CONCLUSION

Predicting the popularity of a product in the near future is easy, since its present popularity usually gives a good prediction of its popularity in the near future. Nevertheless, this present trend becomes less relevant to the popularity in the far future and predicting their long-term popularity becomes difficult. In this case, one has to explore the constituent microscopic details underlying the macroscopic trend. By exploring the relation between buyers and products with a diffusion mechanism, we introduce a measure to rank the future popularity of items. Our results show that simple extrapolation performs better for short-term predictions, but our method gives a more accurate prediction of popular items in the long term. This implies that the microscopic information hidden under the macroscopic trend is essential in the identification of items with potential and the prediction of their long-term trend.

Identifying potentially popular products is useful in various respects. For instance, our method identifies the promising products among a group of products with similar present sales, which can be used to inform sellers to increase the stocks of those promising items; our method also identifies research papers which are likely to be cited frequently in the future, which can be used to identify pioneer research at an early stage to accelerate research and development. We believe that there are other potential applications of our findings, and the present manuscript opens up a new direction to identify methods and applications for predicting long-term popularity trend.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 11547188) and the Youth Scholars Program of Beijing Normal University (Grant No. 2014NT38). C.H.Y. acknowledges the Internal Research Grant No. RG71/2013-2014R and the Dean's Research Fund 04115 ECR-5 2015 of HKIEd, and the Research Grant Council of Hong Kong (Grant No. ECS 28300215).

[1]L. F. Richardson, *Weather Prediction by Numerical Process* (Cambridge University Press, Cambridge, UK, 1922).
[2]A. D. Ahrens, *Meterology Today: An Introduction to Weather, Climate and the Environment* (Brooks Cole Publishing, Belmont, CA, USA, 2007).
[3]K. Mogi, *Earthquake Prediction* (Academic Press, Tokyo, 1985).
[4]R. J. Shiller, *The Subprime Solution: How Today's Global Financial Crisis Happened and What to Do About It* (Princeton University Press, Princeton, NJ, USA, 2008).
[5]J. D. Murray, *Mathematical Biology* (Springer-Verlag, Berlin, 1993).
[6]J. D. Hamilton, *Time Series Analysis* (Princeton University Press, 1994).
[7]R. S. Tsay, *Analysis of Financial Time Series* (Wiley, 2002).
[8]D. Challet, M. Marsili, and Y.-C. Zhang, *Minority Games* (Oxford University Press, Oxford, UK, 2005).
[9]C. H. Yeung, K. Y. M. Wong, and Y.-C. Zhang, "Models of financial markets with extensive participation incentives," Phys. Rev. E. **77**, 026107 (2008).
[10]R. Pastor-Satorras and A. Vespignani, "Immunization of complex networks," Phys. Rev. E **65**, 036104 (2002).
[11]P. C. Pinto, P. Thiran, and M. Velerli, "Locating the source of diffusion in large-scale networks," Phys. Rev. Lett. **109**, 068702 (2012).
[12]F. M. Bass, "A new product growth for model consumer durables," Manag. Sci. **15**, 215 (1969).
[13]*New-Product Diffusion Models*, edited by V. Mahajan, E. Muller, and Y. Wind (Kluwer, Boston, MA, USA, 2000).
[14]A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," Science **286**, 509 (1999).
[15]M. Medo, Y.-C. Zhang, and T. Zhou, "Adaptive model for recommendation of news," Europhys. Lett. **88**, 38005 (2009).
[16]R. A. Banos, J. Borge-Holthoefer, and Y. Moreno, "The role of hidden influentials in the diffusion of online information cascades," EPJ Data Sci. **2**, 6 (2013).
[17]C. H. Yeung, G. Cimini, and C.-H. Jin, "Dynamics of movie competition and popularity spreading in recommender systems," Phys. Rev. E **83**, 016105 (2011).
[18]S. Gualdi, C. H. Yeung, and Y.-C. Zhang, "Tracing the evolution of physics on the backbone of citation networks," Phys. Rev. E **84**, 046104 (2011).
[19]R. Sinatra, P. Deville, M. Szell, D. Wang, and A.-L. Barabsi, "A century of physics," Nat. Phys. **11**, 791 (2015).
[20]M. Medo, G. Cimini, and S. Gualdi, "Temporal effects in the growth of networks," Phys. Rev. Lett. **107**, 238701 (2011).
[21]D. Wang, C. Song, and A.-L. Barabasi, "Quantifying long-term scientific impact," Science **342**, 127 (2013).
[22]H. Shen, D. Wang, C. Song, and A.-L. Barabasi, "Modeling and predicting popularity dynamics via reinforced Poisson processes," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014), p. 291.
[23]L. Lu, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems," Phys. Rep. **519**, 1–49 (2012).
[24]A. Fiasconaro, M. Tumminello, V. Nicosia, V. Latora, and R. N. Mantegna, "Hybrid recommendation methods in complex networks," Phys. Rev. E **92**, 012811 (2015).
[25]A. Zeng, C. H. Yeung, M.-S. Shang, and Y.-C. Zhang, "The reinforcing influence of recommendations on global diversification," Europhys. Lett. **97**, 18005 (2012).
[26]Y. Li, P. Pin, and C. Wu, "A network centrality method for the rating problem," Plos One **10**(4), e0120247 (2015).
[27]M. Kendall, "A new measure of rank correlation," Biometrika **30**, 81 (1938).
[28]See http://www.aps.org for APS citation network; retrieved 11th July 2013.