Information Processing and Management 60 (2023) 103190

Contents lists available at ScienceDirect

Information Processing and Management

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

Utilizing statistical physics and machine learning to discover collective behavior on temporal social networks

Yi-Xiu Kong [a,*], Rui-Jie Wu [b], Yi-Cheng Zhang [c,d,*], Gui-Yuan Shi [b,*]

[a] School of Science, Beijing University of Posts and Telecommunications, Beijing, China
[b] International Academic Center of Complex Systems, Beijing Normal University, Zhuhai, China
[c] School of Computer Science, Xijing University, Xi'an, China
[d] Department of Physics, University of Fribourg, Fribourg, Switzerland

ARTICLE INFO

Keywords:
Social collective behavior
Evolutionary machine learning
Social networks
Time-correlation-functions

ABSTRACT

Computational social science has become a branch of social science that uses computationally intensive ways to investigate and model social phenomena. Exploitation on mathematics, physics, and computer sciences, and analytic approaches like Social Network Analysis (SNA), Machine Learning (ML), etc, develops and tests the theories of complex social phenomena. In the emerging environment of social media, the new characteristics of social collective behavior and its extensive phenomena have become the hot spot of common concern across many disciplines. In this paper, we propose a general quantitative framework to discover the social collective behavior in temporal social networks. The general framework incorporates the Time-Correlation Function (T.C.F.) in statistical physics and evolutionary approach in Machine Learning, and provides the quantitative evidence of the existence of social collective behavior. Results show collective behaviors are observed and there exists a tiny fraction of users whose behavior are constantly replicated by public, disregard of the behavior itself. Our method is assumption-independent and has the potential to be applied to various temporal systems.

1. Introduction

Understanding human behavior in social systems lies in the center in the social sciences. Despite the complexity embedded in our society, social scientists have advanced to observe and understand our society in both theory and application from many different perspectives (Bak-Coleman et al., 2021; Blumer, 1971; Mukkamala & Beck, 2018). In recent years, the development of data science and computer science has brought new possibilities to the quantitative analysis of social science problems. The study of social sciences involves the understanding of social agents, the interaction among these agents, and also the result of those interactions on the social mixture (Macy & Willer, 2002). Though the topics and methodologies in scientific discipline take issue from those in scientific or engineering science, many of the approaches employed in modern social simulation originated from fields like physics and AI (Epstein & Axtell, 1996; Gilbert & Troitzsch, 2005). A number of the approaches that originated from this field are also applied into the social sciences, like measures of network spatial relation from the fields of Social Network Analysis and network science. However, studying social sciences faces more challenges than natural sciences like math, physics etc., due to both the intrinsic complexity of social systems. As said by John von Neumann (Alt, 1972), we quote here, 'If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.'

* Corresponding authors.
E-mail addresses: yixiu.kong@bupt.edu.cn (Y.-X. Kong), yi-cheng.zhang@unifr.ch (Y.-C. Zhang), sgy@bnu.edu.cn (G.-Y. Shi).

https://doi.org/10.1016/j.ipm.2022.103190
Received 30 June 2022; Received in revised form 20 October 2022; Accepted 20 November 2022
Available online 1 December 2022
0306-4573/© 2022 Published by Elsevier Ltd.

One of the most complex and intriguing topics in social sciences is the collective behavior. Collective behavior (Park & Burgess, 2019), includes panics, fads, trends, crazes, public, etc (Kim, 2018; Lu et al., 2020; Victor, 1998), as well as more structured phenomena such as reform and revolutionary social movements (Goldstone & Ritter, 2019). As one of the most commonly seen collective behaviors, herding behavior is widely seen in many scenarios of both animal herd and human society (Cote & Sanders, 1997; Dugatkin, 2013; Trueman, 1994; Werner & Dyer, 1993), it can be generalized as a group of interacting individuals behaving according to the behavior of others rather than their own judgment. Many studies of herding behavior were made from different perspectives like psychology (Drury, 2020), sociology (Millward & Takhar, 2019), or financial market (Dyer, Johansson, Helbing, Couzin, & Krause, 2009), as we will show in the following. These results indicate that humans act not only based on personal judgment but also influenced by the environment. But we still lack a general quantitative method which can be applied to various contexts of social systems to capture the herding behavior among interacting humans. A general quantitative framework to detect the herding behavior and other collective behaviors in many time-dependent systems is in dire need.

2. Related works

Collective behavior as an emergent topic in social science studies recently attracts attention of scientists from diverse backgrounds. In 1995, Vicsek presented the famous Vicsek model (Csahók & Vicsek, 1995), in which the statistical characteristics of many self-driven particle motions can be reproduced. Vicsek's idea is very simple. In order to avoid "collision", the self-driving particles are determined by the average direction of their neighbors' motion (plus some noise and other factors), so they pass through the neighbors in the model. The pointing angle is averaged to obtain the moving direction of the particles, and then the conformation of the system is updated according to the moving direction. The Vicsek model offers a brand new methodology to quantitatively study the general collective behavior.

Goldstone and Janssen (2005) develop a computational agent-based models to provide quantitative and empirically verifiable accounts of how individual decisions lead to the emergence of group-level organizations. Wittenbaum and Park (2001) discover that groups prefer to discuss shared information that all members know instead of unshared information that a single member knows, unveiling the potential collective mechanism. Raafat, Chater, and Frith (2009) researched herding from cognitive neuroscience background, they proposed that the mechanism of information spreading among the individuals and the patterns of connections of individuals can be combined to understand the herding phenomenon.

There are many more systems in which collective behavior has been observed: a few examples in non-living systems including nematic fluids (Kudrolli, 2010); boats (Ibele, Mallouk, & Sen, 2009; Narayan, Ramaswamy, & Menon, 2007); simple robots (Suematsu, Nakata, Awazu, & Nishimori, 2010); macromolecules (Butt et al., 2010; Schaller, Weber, Semmrich, Frey, & Bausch, 2010) as well as humans (Helbing, Schweitzer, Keltsch, & Molnár, 1997). Importantly, Dyer et al. (2008) tested these predictions on human groups in which the experimental subjects were naive and they did not use verbal communication or any other active signaling. The experiments indeed supported the predictions. Other experiments investigated the relationship between the spatial position of informed individuals and the speed and accuracy of the group motion (Dyer et al., 2009).

Fowler's famous paper. (Fowler & Christakis, 2008) in political sciences use social networks analysis to study the influence amongst people, and find that the relationship between people's happiness extends up to three degrees of separation. Their results show that the influence of people can spread by the social network with others. Lederrey and West (2018) study the herding behavior of human on online beer rating sites. They use the data of same item on two separate rating sites, focusing on items that received a high first rating on one site, and a low first rating on the other, and thus construct a natural control group to test the herding effect. Bouri, Gupta, and Roubaud (2019) examine the existence of herding behavior in the financial market. By using a logistic regression with rolling-window analysis, they find that herding tends to occur as uncertainty increases in the market.

There exist more studies on herding behavior in human crowds. Despite their success on either discover the herding according to specific patterns, or to use the quantitative methods to reveal the transmission-based herding behavior (Raafat et al., 2009), the existing research still lacks a quantitative research method that does not rely on specific assumptions or dynamic processes. In this paper, inspired by the above research, we adopt the Time correlation Function method in statistical physics to study herding behavior. Based on the simple meaning of herding behavior and combined with the machine learning process, we develop a new statistical algorithm, which is independent of specific application scenarios and dynamic models, and is a universal research method. It will put forward a new idea for the future research of herding behavior.

3. Data and methods

Now we start with a brief introduction of the data from online systems. We obtain 4 temporal social online datasets from open access, which are Amazon DVD purchase data, Delicious bookmark data, Weibo repost (a social media like twitter or Facebook) data and a twitter retweets data. A typical record in the data denotes a user i collect an item (a DVD, bookmark or a message) α at time t. Note that in our dataset there is no repeated actions which means a user can choose an item only once at most. We analyze the data through commonly used Social Network Analysis (SNA) method. With SNA, we project the data to a bipartite network $G = (N, L)$, which consists of two set N and L. N is the set of nodes (or vertices), which in our case is the set of users and items. L is the set of links (or edges). The interactions of the network can be represented by an adjacent matrix $A_{i\alpha}$ with elements $a_{i\alpha} = 1$ if there is a link between user i and item α, and 0 otherwise (throughout this paper we use Latin and Greek letters for user related indices and item related indices respectively). Thus, for any given time t, we can obtain the adjacent matrix $A_{i\alpha}(t)$ which contains all the information of the input data at time t. We denote the degree of item and user nodes by k_α and k_i, respectively, which are number of users who collected item α and the number of items user i has collected. And, we denote $\Delta k_\alpha(t, t+\tau) = k_\alpha(t+\tau) - k_\alpha(t)$ the degree increase of item α during the period $t \rightarrow t+\tau$.

3.1. Datasets and the statistics

1. Amazon movie purchase data

The dataset were obtained from open access (McAuley & Leskovec, 2013). While the data span 5546 days (August 1997–October-2012), we only use the data from days 2000 to 5000 because the rest of the data show comparably low activity of users. After this operation, there are 960,374 links whereas 497,309 users and 88,859 items have at least one link. The fundamental statistic of the data can be found in Fig. 1.

2. Delicious bookmark data

The data obtained by downloading publicly-available data from the social bookmarking website delicious.com in May 2008. Due to processing speed constraints, we randomly sampled 50% of all users available in the source data and included all their bookmarks. To solve the possible ambiguity of various web addresses pointing to the same web page, reduce the number of items and thus increase the data density, bookmarks are represented only by their base www-address without the initial protocol specification, possible leading "www." and trailing slash (e.g., http://www.edition.cnn.com/US/ is modified to edition.cnn.com). Time stamps are counted in hours from 01/09/2003 and run from 0 to 36,027. For the same user activity reasons as in Amazon, we only use the data from hour 15,000 to 35,000. There are 107,810 users, 2,435,912 items and 9,322,949 links in the resulting data. The fundamental statistic of the data can be found in Fig. 1.

3. Weibo reposts data.

The weibo data was obtained through open access (Fu, 2013). The data collects the reposts of users on Weibo.com (A Chinese social media which is a mixture of Facebook or Twitter) for one week, time spans from 2nd Feb 2012 to 9th Feb 2012. Each record of data is a temporal information about a user repost a microblog at a certain time. The time resolution is second, but in our case, we regroup the time information into 168 h. After the pre-treatment, the data consists of 1,141,319 users, 2,041,854 messages, there are 4,790,108 records during the 168 h in total. The fundamental statistic of the data can be found in Fig. 1.

4. Twitter Higgs Boson data.

The Higgs dataset has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the hard to catch Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are collected. The data used in our study have two sub-datasets: the first one is the retweets data, the data records the user of the tweet, the source user who is retweeted, as well as the time of each of the tweet. The second dataset is the user-following data, in which the following relations of users are recorded.

3.2. Methods

In this section we will briefly introduce our method called the Evolutionary User Score (EUS), including Time-Correlation Function(T.C.F.) serving as the basis of this method. TCF is an intuitive and effective way of describing the dynamics of temporal systems and is one of most common tools of studying time-dependent dynamics. It has also been widely applied in many fields like statistical physics, signal processing, informatics and is generally applicable to any time-dependent process. All these applications share the same fundamental idea, and differ only in the specific scenarios they are applied to. In general, a T.C.F. $C_{AB}(t,\tau)$ has the following form:

$$C_{AB}(t,\tau) = \langle A(t) \cdot \ B(t+\tau)\rangle \tag{1}$$

where τ is a parameter that decides the time delay between two temporal variables $A(t)$ and $B(t)$. The bracket$\langle\rangle$ here means ensemble average depending on the specific system that is considered. It measures the statistic correlation between two variables $A(t)$ and $B(t)$ (with a time difference τ). In physics, $A(t)$ is regarded as an perturbation induced to the system, and $B(t+\tau)$ represents the system response to the perturbation, with the T.C.F we can study the properties of the system revealed from its statistic correlations.

According to the concept of herding phenomenon, that a group of interacting individuals behaving according to the behavior of others, we realize the underlying potential of T.C.F. to study this herding behavior. In the following we apply the idea of T.C.F. and derive a quantitative measurement of social herding behavior in online systems. Similarly, as in physics we treated an injected particle as a perturbation to the system and measure the response of the system, we now regard each action of users, in our case is that user i collecting an item α at time t, as an perturbation to our system. Depending on the influence of the user, the response can vary. Thus we measure the system response to the action, i.e. the number of users who have replicated the user's choice in a given time, which in our case is the degree growth of item α in the near future period $t \to t+\tau$.

Based on the above introduction, we derive our perturbation term $A(t) = a_{i,\alpha}(t)$, representing i collects α at time t, and system response $B(t,\tau) = \ \Delta k_\alpha(t,t+\tau)$. Note that $\Delta k_\alpha(t,t+\tau)$ is the total number of users who collect item α during time $t \to t+\tau$. By giving the above definition, the T.C.F. will be representing the statistic correlation between the action that user i collects an item α at time t and how many users replicate the action during the period $t \to t+\tau$, In other words, the T.C.F. we defined is a measure of collective herding behavior in the system towards user i, as shown in the following equation:

$$UserScore_i(t,\tau) = \ \langle a_{i,\alpha}(t) \cdot log[\Delta k_\alpha(t,t+\tau)]\rangle_\alpha \tag{2}$$

Here the bracket means the T.C.F averages over all the items user i has collected at time t. This treatment helps to normalize the difference of the measurement at different time t. Fig. 2 gives an illustration of the measured online behavior. As shown in the figure the user b has much less influence to the public than user a.

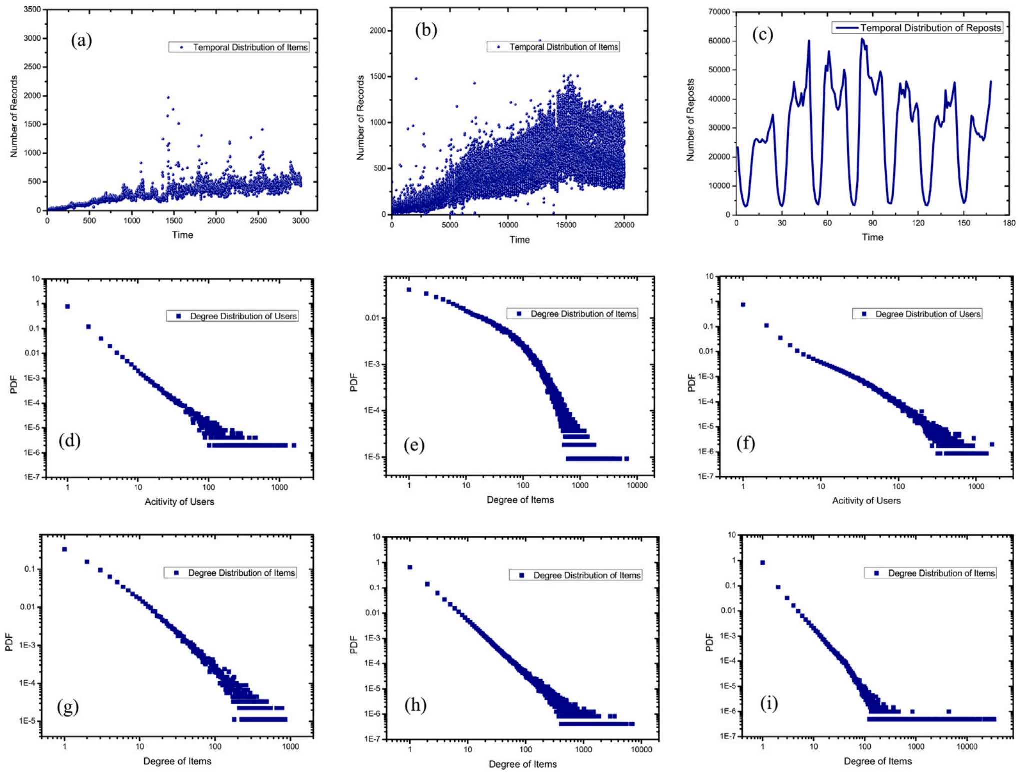

Fig. 1. Fundamental statistics of datasets. **(a)**, **(b)**, and **(c)** show the temporal distribution of items in Amazon, Delicious, and Weibo data, respectively. **(d)**, **(e)**, and **(f)** show the degree distribution of users in Amazon, Delicious, and Weibo data, respectively. **(g)**, **(h)**, and **(i)** show the degree distribution of items in Amazon, Delicious, and Weibo data, respectively.

Algorithm 1 EUS Computation.

Input: dataset, t_0, t_f, t_d
Output: $EUS(t_0)$

Initialization: Matrix $UserScore_i(t)$ based on Equation (2)
Compute $EUS(t=1)$ based on
Sort $EUS(t=1)$
Classify Group High, Group Medium, Group Low according to the relative ranking of $EUS_i(t)$ by top 1%, 1 - 10% and rest.
For $t < t_0$
Update $EUS_i(t)$ based on Equation (3)
Update Group High, Group Medium, Group Low based on the relative ranking.
$t \rightarrow t+1$
Return $EUS(t_0)$

Next, we use the concept of evolutionary approach in ML to process the UserScore at each time step. Similar to evolution and natural selection process, we use the following evolutionary learning approach to automatically classify the users in to three groups depending on the ranking of their influence to the public as measured by UserScore. The selection of groups evolve through the temporal evolution of UserScore of all users in the system.

In detail, first the UserScore at each time step is calculated separately. If a user has a UserScore at time t, that user is awarded a 'bonus' score of $1/\Omega_t$ which equals to how many time larger than average UserScore. Note the computation of UserScore includes a normalization term introducing the competing interaction among all users. Next, for a given time t_0, as shown by the formula (3), we calculate the historic UserScore considering all $t < t_0$. Finally we obtain the evolutionary score all $t < t_0$, called Evolutionary User Score (EUS). In particular, in order to suppress the system's first-mover advantage of old users, we apply an exponential decay of the weight. The decay rate can be conditioned by a preset parameter.

$$EUS_i(t,\tau) = \sum_{t'<t} UserScore_i\left(t',\tau\right) \cdot \ exp[(t'-t)/t_d]/\Omega_t, \tag{3}$$

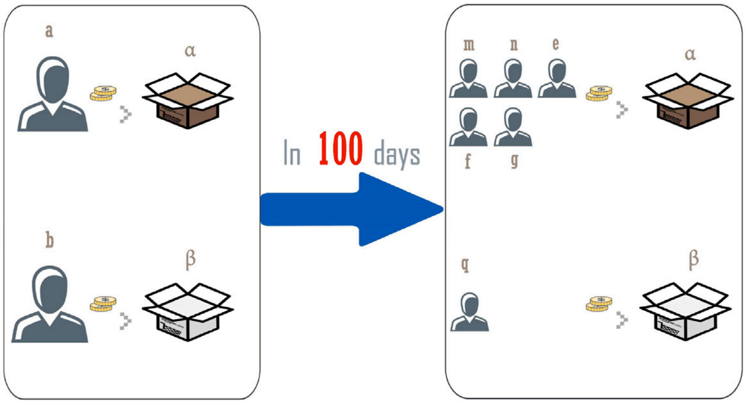

Fig. 2. Illustration of the online herding behavior which will be measured by correlation functions. As shown in the figure, a user i collects an item α at time t, after t_f, a number of users copy the choice of user i.

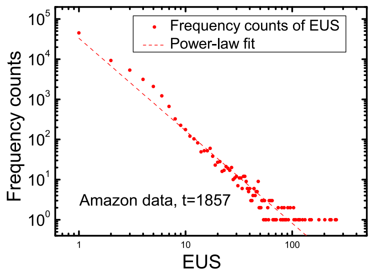

Fig. 3. The distribution of EUS at a randomly chosen time in Amazon, t = 1857.

here $\Omega_t = \langle UserScore_i^t \rangle$ is the mean value of UserScore over all users at time $t.t_d$ is the parameter that measures how fast is the decay rate of the historical UserScore of all users.

Note that the UserScore is a long-tailed power-law distribution, as is shown in Fig. 3., the vast majority of UserScore are less than 1. At each time step, the EUS of all users will be ranked and divided into three groups, Group High(Top 1%), Group Medium (Top 1%–10%) and Group Low (the rest). We are most interested in the top 1% of users with highest UserScores, and the algorithm will process every user in the system.

Because the user's behavior can be regarded a random perturbation to one's exsiting EUS score. In order to obtain the two groups of users with high UserScore, we need to change and select the groups according to the evolution of each user, that is, at each step of evolution, optimize the following supervised learning function:

$$Group_{High}(t) = argmax(\sum_{i=1}^{\epsilon * N} EUS_i(t)) \tag{4}$$

$$Group_{Medium}(t) = argmax(\sum_{i=1}^{\theta * N} EUS_i(t)), i \notin Group_{High}(t), \tag{5}$$

where N is the total number of users, $\epsilon = 0.01$ is the parameter that regulates the portion of users in Group High, while $\theta = 0.1$ represents the portion of users in Group Medium, except those in Group High.

Only few of users can have high scores and be enrolled into Group High. Thus, for a user in Group High at time t, under the null hypothesis, the probability that a user is still in the high user group in the next step is only 1%, and the probability of exceeding the average is also very small. Thus, this evolutionary algorithm will give rigorous rewards and punishments to each user's performance at each time t. The ones with high performance will be promoted, and the ones with low performance will be downgraded.

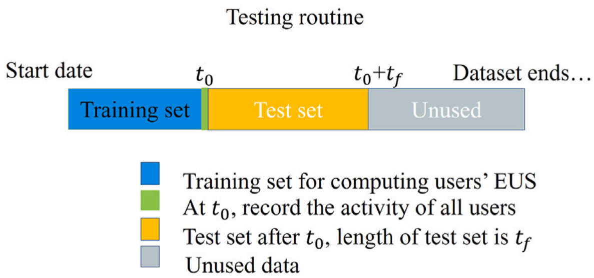

Fig. 4. Illustration of the test schema. At any given time t_0, we use the historic data ($t < t_0$) to calculate the EUS. We then test the future degree increase of the items picked by different users to see whether these items perform differently in the future.

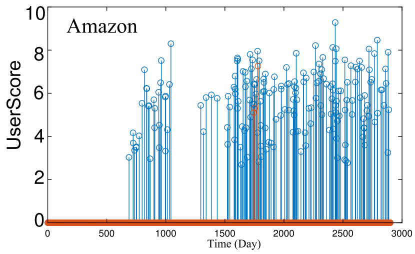

Fig. 5. The temporal pattern of correlation functions of two typical users. The blue one denotes a user from Group High, the red one denotes a user from Group Medium. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Results

4.1. Test routine

With the EUS, we can investigate how this EUS measures the users' influence that can affect the subsequent evolution of the system. At any given time t_0, we use the historic data ($t < t_0$) to calculate the EUS and classify the three user groups. We then test the future degree increase of the items picked by the three groups to see whether these items perform differently in the future time window $t_0 \rightarrow t_0 + t_f$, as shown in Fig. 4. t_f is the parameter to determine the length of future time window.

(1) High influential group (Group High), including top 1% of users with the highest capability score EUS_i^t.

(2) Medium influential group (Group Medium), including 10% of users with the highest aggregate prediction capability (excluding users in Group High).

(3) Low influential capability group (Group Low), including the remaining 90% of users. Next, we will check whether the items selected by the three groups of users (we name them Item High, Item Medium and Item Low respectively) will have different evolution trajectories of their degree increase in the coming period $t_0 \rightarrow t_0 + \tau$. Note that here the items can be chosen by all three groups simultaneously so the overlap of three groups of items are permitted. The activities of two typical users in Amazon data are shown in Fig. 5.

4.2. Future degree increase of items

Fig. 6(a), (c), and (e) show the results of average future degree increase of items picked by the three groups in Amazon, Delicious, and Weibo datasets, respectively. We can see that in all datasets, the items selected by the three groups have significant different degree growth in the future (p-value $\ll 0.05$).

We can find that the three groups of items have great differences in their long-term degree growth. The degree growth of items picked by each group decreased in turn referring to High, Medium, Low group. Among them, the Item High group average increment is significantly higher than the other three groups. Error bars indicate the standard deviations of the statistics. Since the number of people in Group High is only 1% of the total population, the number of Group Medium is 9% of the total population, and the rest are Group Low users, so the numbers of items selected by these three groups of users are largely different. Specifically, the total number of Group High's daily selection on the Amazon data set is about 30 items with a variation around 5 during the all time steps, Group Medium selects about 40 items, and Group Low selects almost all items sold on the same day, about 300. Since the items selected by each group are allowed to intersect, and the Group Low group contains almost all items, it can be said that the items selected by Group High and Group Medium are a subset of the items selected by Group Low. Thus we can conclude that the significant increase in the average growth of the two groups compared to Group Low is related to the exclusion of those items that will not be popular in the selection of these two groups of users, in other words, The group user's choice includes only items with

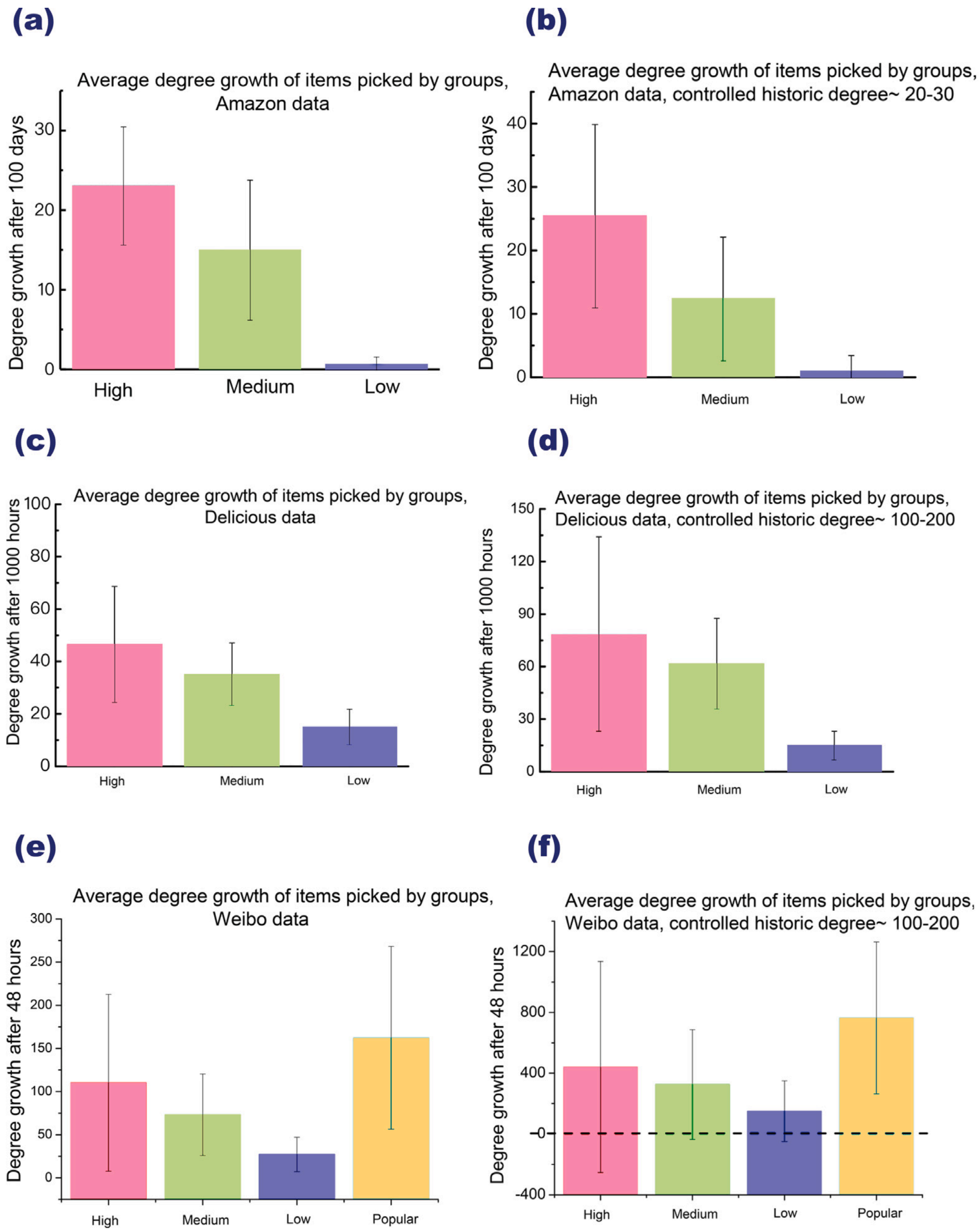

Fig. 6. The future degree increase of items picked by 3 groups of users. The results are averaged over 100 realizations. **(a)** Amazon data. $t_f = t_d = 100$ days. Error bars indicate the standard deviations of the respective statistics. **(b)** Amazon data, control group that historic degree of items are filtered to be between 20 to 30. $t_f = t_d = 100$ days. **(d)** Delicious data. $t_f = t_d = 1000$ h. **(b)** Delicious data, control group that historic degree of items are filtered to be between 100 to 200. $t_f = t_d = 1000$ h. **(e)** Weibo data. $t_f = t_d = 48$ h. **(f)** Weibo data, control group that historic degree of items are filtered to be between 100 to 200. $t_f = t_d = 48$ h. The dashed line indicates the zero value. Note that in **(e)** and **(f)** we show the results of additional column named Popular, which means the most popular historic items in Weibo data, the number of the Popular items are the same with items picked by Group High.

large future increments. It can be seen from the above analysis that the EUS computation can identify the users that can choose the items that have systematic potential in the future degree increments.

What mechanism causes this difference? This may be the issue we are most concerned about. The most well known mechanism about the growth of degree is the Preferential attachment. This mechanism was proposed by Barabási and Albert (1999). This mechanism points out that in many social and natural systems, such as wealth distribution, urban population, oil reserves in oil fields, etc., all conform to the power-law. A preferential attachment process is any of a class of processes in which some quantity, is defined among a broad range of magnitudes. a number of individuals or items according to how much they already have. Here we want to know if the observed difference in the degree growth of items comes from the preferential attachment: that is, the future degree growth of some users is large, just because they chose items that were popular in the past. So we have the following hypothesis, and we intend to test whether the hypothesis are statistically true.

Hypothesis 1. The difference in the future average increment of the items selected by different groups of users is derived from the degree of history of the item. This difference comes from the degree of the item itself rather than the difference in the user group. There is no significant statistical difference in the degree of increase in the similarity of the item.

To test this hypothesis, we set an experiment with controlled historic degrees of items, and using the same testing routine to observe whether these items have different growth once they are selected by different groups of users. The results on Amazon, Delicious, and Weibo data are shown in Fig. 6(b), (d), and (f), respectively. As can be seen from the figure, on all three different data sets, the items selected by the three groups of users will have a larger average degree growth in the future. In Amazon and Delicious data, the items selected by the high group, even in the case of similar historical degrees, still outperform the items selected by the other two groups with a strong statistical significance. Therefore, H1 cannot be established on these two datasets. That is to say, the degree increase of the items are not resulted from the Preferential attachment mechanism. The herding behavior are validated on these two datasets.

Interestingly, we find that in the Weibo dataset, the items picked by Group High have a higher average degree increment than the other groups. However, comparing the data of Group High and Group Medium, we found that H1 could not be rejected from the t-test perspective (p-value of t-test is 0.14), and the p-value of Wilcoxin test is 0.65.

The reason of such phenomenon comes from the statistical distribution of the testing results. Due to some unknown reasons in the Weibo data, the degree growth results for 100 replicates have a power-law-like frequency distribution, and have very large variation, as shown in Fig. 6(f), which is completely inconsistent with the behavior of other data sets. Power-law distributions can have infinite variance, in which cases the large sample guarantees from the Central Limit Theorem (CLT)will not apply. As we know, the commonly used t-test and Wilcoxin test can apply only if the data satisfy CLT, i.e. are of finite variance, but the Weibo data set shows an obvious non-convergence tendency of variance. Therefore, such conventional statistical tests like t-test and Wilcoxin test are not applicable to our results. Unfortunately, at present there is no well-accepted statistical method to judge this situation. Based on the current result of average degree increase of items, especially considering the growth of most popular historic items, which is much larger than all three item groups, we will agree that the weak herding behavior are likely to exist in Weibo and the most dominant factor of degree growth can be related to the specific mechanism that Weibo system promotes hot topics.

4.3. Stability and relaxation of the EUS

As shown in the above analysis, EUS can measure the existence of herding behavior in temporal systems with diverse dynamics. From the perspective of real-world application, such metric can be used to predict the future trends of the public to a certain extent. If we capture the behavior of High Group users, we can know that the user crowd will have a tendency to follow their behavior, and thus resulting a trend in the near future. But before using EUS for predicting, an important issue is how stable is the EUS metric over time. If the EUS indicator decays too quick over time, then the EUS High Group does not guarantee to have a strong influence to the system in the long term future. Or if the EUS of users change too fast with respect to time, then it is difficult to obtain a relative stable Group High in terms of constant observing their behavior.

To investigate this, we first calculate the EUS of all users at any given time t_0 and then compute the Pearson Correlation of EUS values on different time steps on Amazon and Delicious dataset. The result is given in Fig. 7(a) and (b). The result shows that the EUS score of users are relatively stable over time. This is not surprising because the EUS is the weighted cumulative value. Then we test the relaxation of degree growth of items selected by user groups in the long term. In Fig. 7(c) and (d), we show the degree growth time-relaxation of items picked by different groups, in Amazon and Delicious data, respectively. It can be seen that the relaxation is slow over time for all of the groups. And items from group high have consistently larger degree growth in the far future.

4.4. Predictive power of the EUS

In the following we test whether EUS has a predictive power of the collective trends . We perform the test on Twitter higgs boson dataset to observe that whether a user with higher EUS will have more retweets.

As the EUS is based on herding behavior and can be a used as a metric for measuring the user's influence towards the public, here we compare EUS with two well-accepted metrics for assessing the influence of users, which are the degree of users and the PageRank of users. In Twitter higgs boson dataset, the static social network data allows us to compute the degree and PageRank of direct user–user interaction network, while our EUS metric is computed through a dynamic process according to the behaviors of the users. We compute three metrics of all users and observe that how the metrics are related to the number of retweets of their tweets in a given time. The result is shown in Fig. 8. As we can see in the figure, on average, the group of users with the highest degree level will have about 2 retweets in 48 h; the group of users with the highest PageRank level will have about 5 retweets in 48 h; while the group of users with the highest EUS level will have about 10 retweets in 48 h. The result shows our EUS metric is a much better predictor to predict the future trend in real-world systems.

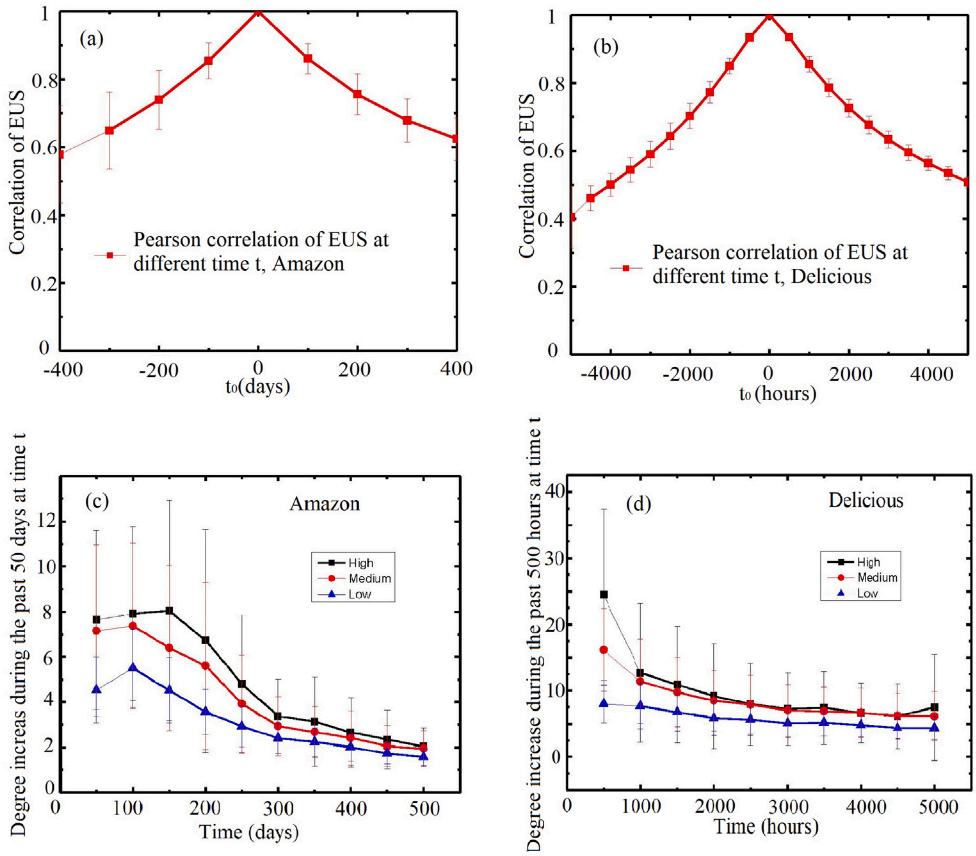

Fig. 7. **(a)** and **(b)** The linear correlation coefficients of EUS values at different time t on Amazon and Delicious data. The result is averaged over 100 realizations. **(c)** and **(d)** The time relaxation pattern of items picked by different groups of users on Amazon and Delicious datasets.

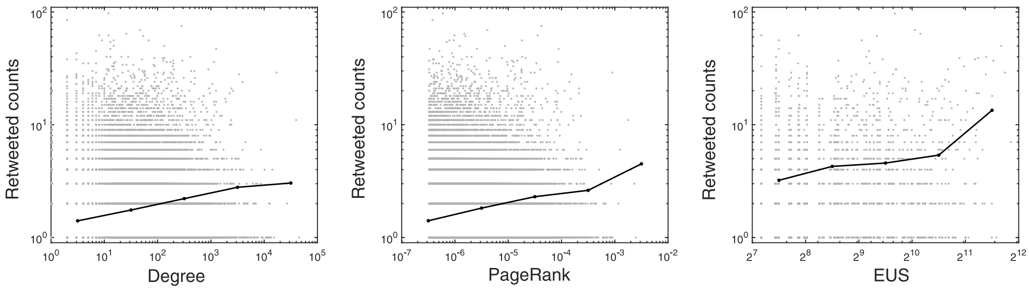

Fig. 8. The predictive power of metrics in Twitter higgs boson dataset. Each dot represent a user. The solid lines indicate the interval mean of retweeted counts with 48 h after time $t = 70$. The data are divided into five intervals according to the horizontal axis. Note that the figures are in log–log scale. **(a)** Retweeted counts versus degree of users. **(b)** Retweeted counts versus PageRank of users. **(c)** Retweeted counts versus EUS of users.

5. Conclusion and discussion

Now, we can conclude that the EUS score calculated according to the correlation function can measure the herding behavior in temporal online systems and predict the future growth of items to a certain extent. We should note that we did not use any feature information about the items or users in our calculations, we only calculate the statistical correlation between the users' behavior. Because the degree growth of an item over a period of time reflects the behavior of all users as a whole. As shown in our analysis, there exists a small group of users who can impact the whole system so that the public choices are biased to replicate their behavior. We can calculate EUS score to predict the collective behavior of all users in the future. From this perspective, EUS can also measure the users' collective behavior and predict the future trends.

If we consider the definition of herding behavior, we will find that the T.C.F defined here can be interpreted as the collective herding behavior of a group towards a specific user. Although the number of the high EUS user group are few, it is indisputable that these users can affect the behavior of all users in the future to some extent. So we can conclude that EUS measures the impact of a user on other users in the system.

As naturally being a measure of correlation, EUS does not provide any causal relation. We suppose there are two possible mechanisms that results such phenomenon. The first conjecture is that high influence users affect other users through social networks or direct user–user interactions, such as on twitter and Weibo datasets. But unfortunately we lack such social network data that describe the detailed influence sequences corresponding to the temporal data to test this hypothesis. The second conjecture is that for systems without apparent social networks, such as Amazon, the users' influence can spread through recommendation system. However, this require additional massive research to test and is divergent to the topic of this paper. Overall, EUS can serve as a metric for measuring the influence of users, but the reason why users with high EUS score have a higher influence to the system still awaits further investigation.

In conclusion, combined with T.C.F. and evolutionary approach in ML, we devise a simple metric called the EUS that can measure the herding behavior in online social networks. We analyze three temporal online datasets which are Amazon DVD purchase data, delicious bookmark data, and Weibo (a Chinese social media like twitter or Facebook) re-posts data with this new model. We discover herding behaviors in these datasets using our method and find that there exists a tiny fraction of users whose behaviors are consistently replicated by all users in the system, disregard of what their choices are. We then validate our results by generating a control group in which the preferential attachment mechanism is ruled out. Furthermore, we apply our method on twitter dataset which records the retweets data about Higgs boson, and the results show our EUS have a significantly better predictive power on the future trends than two commonly used metric that measures influence of users, the degree and PageRank.

CRediT authorship contribution statement

Yi-Xiu Kong: Designed the research, Designed the figures, Performed numerical experiments, Writing – original draft. **Rui-Jie Wu:** Performed theoretical analysis, Writing – original draft. **Yi-Cheng Zhang:** Designed the figures, Performed numerical experiments, Writing – original draft. **Gui-Yuan Shi:** Performed theoretical analysis, Designed the figures, Performed numerical experiments, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We acknowledge Ms Rui ZHOU for the assistance on the preparation of figures. This project has received funding from the National Science Foundation of China (NSFC) under grant agreement No. 71731002, the Fundamental Research Funds for the Central Universities, China from BUPT, under grant No. 505022019 and No. 500422415, and Swiss National Science foundation, under grant No. 200021_213158.

References

Alt, F. L. (1972). Archaelogy of computers: Reminiscences, 1945–1947. *Communications of the ACM*, *15*(7), 693–694.
Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., et al. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, *118*(27), Article e2025764118.
Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.
Blumer, H. (1971). Social problems as collective behavior. *Social Problems*, *18*(3), 298–306.
Bouri, E., Gupta, R., & Roubaud, D. (2019). Herding behaviour in cryptocurrencies. *Finance Research Letters*, *29*, 216–221.
Butt, T., Mufti, T., Humayun, A., Rosenthal, P. B., Khan, S., Khan, S., et al. (2010). Myosin motors drive long range alignment of actin filaments. *Journal of Biological Chemistry*, *285*(7), 4964–4974.
Cote, J., & Sanders, D. (1997). Herding behavior: Explanations and implications.
Csahók, Z., & Vicsek, T. (1995). Lattice-gas model for collective biological motion. *Physical Review E*, *52*(5), 5297.
Drury, J. (2020). Recent developments in the psychology of crowds and collective behaviour. *Current Opinion in Psychology*, *35*, 12–16.
Dugatkin, L. A. (2013). *Principles of animal behavior: Third international student edition*. WW Norton & Company.
Dyer, J. R., Ioannou, C. C., Morrell, L. J., Croft, D. P., Couzin, I. D., Waters, D. A., et al. (2008). Consensus decision making in human crowds. *Animal Behaviour*, *75*(2), 461–470.
Dyer, J. R., Johansson, A., Helbing, D., Couzin, I. D., & Krause, J. (2009). Leadership, consensus decision making and collective behaviour in humans. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*(1518), 781–789.
Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom Up*. Brookings Institution Press.
Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study. *Bmj*, *337*.

Fu, W. (2013). Assessing student physical therapists' clinical reasoning competency.
Gilbert, N., & Troitzsch, K. (2005). *Simulation for the social scientist*. McGraw-Hill Education (UK).
Goldstone, R. L., & Janssen, M. A. (2005). Computational models of collective behavior. *Trends in Cognitive Sciences*, *9*(9), 424–430.
Goldstone, J., & Ritter, D. (2019). Revolution and social movements. In *The Wiley blackwell companion to social movements* (pp. 682–697). Hoboken: John Wiley & Sons Ltd, Wiley Online Library.
Helbing, D., Schweitzer, F., Keltsch, J., & Molnár, P. (1997). Active walker model for the formation of human and animal trail systems. *Physical Review E*, *56*(3), 2527.
Ibele, M., Mallouk, T. E., & Sen, A. (2009). Schooling behavior of light-powered autonomous micromotors in water. *Angewandte Chemie*, *121*(18), 3358–3362.
Kim, J. (2018). Why do people take to the streets? Understanding the multidimensional motivations of protesting publics. *Public Relations Review*, *44*(4), 501–513.
Kudrolli, A. (2010). Concentration dependent diffusion of self-propelled rods. *Physical Review Letters*, *104*(8), Article 088001.
Lederrey, G., & West, R. (2018). When sheep shop: Measuring herding effects in product ratings with natural experiments. In *Proceedings of the 2018 world wide web conference* (pp. 793–802).
Lu, Y., Yu, L., Zhang, T., Zang, C., Cui, P., Song, C., et al. (2020). Exploring the collective human behavior in cascading systems: A comprehensive framework. *Knowledge and Information Systems*, *62*(12), 4599–4623.
Macy, M. W., & Willer, R. (2002). From factors to factors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*(1), 143–166.
McAuley, J. J., & Leskovec, J. (2013). From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on world wide web* (pp. 897–908).
Millward, P., & Takhar, S. (2019). Social movements, collective action and activism. *Sociology*, *53*(3), NP1–NP12.
Mukkamala, A., & Beck, R. (2018). The role of social media for collective behavior development in response to natural disasters.
Narayan, V., Ramaswamy, S., & Menon, N. (2007). Long-lived giant number fluctuations in a swarming granular nematic. *Science*, *317*(5834), 105–108.
Park, R. E., & Burgess, E. W. (2019). *Introduction to the science of sociology*. Good Press.
Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*(10), 420–428.
Schaller, V., Weber, C., Semmrich, C., Frey, E., & Bausch, A. R. (2010). Polar patterns of driven filaments. *Nature*, *467*(7311), 73.
Suematsu, N. J., Nakata, S., Awazu, A., & Nishimori, H. (2010). Collective behavior of inanimate boats. *Physical Review E*, *81*(5), Article 056210.
Trueman, B. (1994). Analyst forecasts and herding behavior. *The Review of Financial Studies*, *7*(1), 97–124.
Victor, J. S. (1998). Moral panics and the social construction of deviant behavior: A theory and application to the case of ritual child abuse. *Sociological Perspectives*, *41*(3), 541–565.
Werner, G. M., & Dyer, M. G. (1993). Evolution of herding behavior in artificial animals. In *From animals to animats 2: Proceedings of the second international conference on simulation of adaptive behavior, vol. 2* (p. 393). MIT Press.
Wittenbaum, G. M., & Park, E. S. (2001). The collective preference for shared information. *Current Directions in Psychological Science*, *10*(2), 70–73.