Structure-oriented prediction in complex networks

Zhuo-Ming Ren [a,c], An Zeng [b,c,*], Yi-Cheng Zhang [c,a]

[a] *Alibaba Research Center for Complexity Sciences, Alibaba Business School, Hangzhou Normal University, Hangzhou 311121, China*
[b] *School of Systems Science, Beijing Normal University, Beijing 100875, China*
[c] *Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland*

Keywords:
Complex networks
Prediction
Network structure
Network dynamics

Complex systems are extremely hard to predict due to its highly nonlinear interactions and rich emergent properties. Thanks to the rapid development of network science, our understanding of the structure of real complex systems and the dynamics on them has been remarkably deepened, which meanwhile largely stimulates the growth of effective prediction approaches on these systems. In this article, we aim to review different network-related prediction problems, summarize and classify relevant prediction methods, analyze their advantages and disadvantages, and point out the forefront as well as critical challenges of the field.

Contents

[*] Corresponding author at: School of Systems Science, Beijing Normal University, Beijing 100875, China.
E-mail address: anzeng@bnu.edu.cn (A. Zeng).

1. Introduction

The ultimate goal of understanding a real system is to predict it. Prediction is thus always located in the core of scientific research. One can easily find the applications in various real systems such as predicting the most popular movies in online commercial systems; predicting the stock prices in financial systems; predicting the collective human dynamics in social systems; predicting the traffic congestions in transportation systems, just to name a few. However, prediction in these complex systems is particularly difficult due to their rich emergent properties and chaotic behaviors. Complex networks have been proved in the literature to be an effective tool for modeling and analyzing complex systems as they capture the intricate structure of interactions between components that lead to the complex phenomena [1]. In this context, there is a recent wave of investigating prediction problems in complex networks, aiming at not only forecasting the structural growth of the network itself but also revealing the future evolution of the dynamics taking place on the network [2]. For instance, a number of issues such as link prediction, trend prediction, human mobility prediction, information spreading prediction have been investigated.

The rapid development of the prediction research in complex networks is promoted by the accelerating availability of empirical data containing temporal information. Example includes the purchase records of online users, citations between scientific papers, instant human locations from GPS, and so on. However, the highly complex mechanisms behind these data bring in new challenge for prediction. Traditional simple methods such as polynomial regression and linear extrapolation are no longer effective. To extract accurate predictors from the empirical data, theories and methods based on statistical physics play a critical role. For instance, the mechanistic models combining multiple factors accurately predict the future citations of papers and authors [3,4]; the anisotropic rescaling of the distributions of human travel at different scales results in a statistically self-consistent microscopic model that accurately predicts individual human mobility [5]; Classic percolation theory is employed to predict the spreading of virus or information in social networks [6,7]. Numerous such methods were developed in recent literature. Thereupon, the review will cover a wide range of topics including predictions of network structures; predictions at microscopic and macroscopic levels; theoretical analysis of predictability and practical application of the prediction tools. In this review, besides survey the simple summarization of the prediction tools, we aim to classify methods according to their relations; quantitatively compare the effectiveness of reviewed algorithms; and discuss the underlying mechanisms that cause their advantage and disadvantage. The review will be useful as a guide for the practical use of the recently developed prediction tools. Additionally, it will be meaningful for future theoretical research as it will point out the forefront as well as the standing challenges of this field.

2. Framework of prediction in complex networks

2.1. Primary issues

When speaking about prediction, people usually refer to forecasting the future based on the historical data. However, prediction is actually much broader concept. It also includes unveiling and quantifying the correlation between variables of a system, such that one variable is called a "predictor" of another. In addition, if certain components of a system are hidden or missing, one can also "predict" them by pointing out which part they are. Generally speaking, prediction in science is a study focusing on uncertain events.

Real systems are highly complex in structure and constantly evolving, prediction thus is a particular challenging task. Prediction has wide applications in a variety of real systems. A typical example is the online commercial system where an accurate prediction of the most popular products will enable the online retailers better manage their inventory. Similarly, identifying the potential young researchers allow administrator decides who to hire and to whom the fundings should be given to. The benefit of accurate prediction in stock market is even more straightforward. Due to wide applications, large effort has been devoted to design prediction tools for various real systems. Numerous classic methods based on regression and machine learning have been developed by computer scientists and researchers from specific fields such as economics and biology.

The past twenty years have witnessed the rapid development of network science. Many data from real systems are naturally described by complex networks. Examples include the social networks, citation networks, airline networks, international trading networks and so on. The network tools not only lead to more effective understanding of the structure of the real system, but also inspire numerous prediction methods at different levels. Compared with traditional methods, the related works on prediction in complex networks have their own features which can be summarized into the following six aspects.

- **Mechanistic models.** Though the regression and machine learning methods are widely adopted, their prediction accuracy to some specific problems are not satisfactory as the prediction up-limit is constrained by the mis-chosen formulate. A mechanistic model, on the other hand, is constructed with the driving mechanisms of the system inferred from the historical data. Such models usually can significantly improve the prediction accuracy of the system's future evolution.
- **Reliability.** A well-performed prediction method should have both high accuracy and reliability. The data quality of real systems are not always guaranteed. In some case, the data can be overly sparse such that the method should be able to solve the cold-start problem. Moreover, the prediction has to be conducted in noisy environment where the observed data contains spurious information.
- **Bringing macroscopic and microscopic.** In complex systems, collective behaviors are emerged from the local interactions. Therefore, the macroscopic-level prediction can be achieved by aggregating the prediction at microscopic level in a non-trivial way. Such approach usually has high ability in identifying potential components.
- **Predictability.** Much effort in the literature has been devoted to improve the prediction accuracy. However, the up-limit of the accuracy in numerous real systems are bounded by the predictability. Therefore, knowing the theoretical predictability of a system is crucial for designing prediction methods.
- **Feedback.** The prediction in some cases may have feedback effect on the evolution of the system. This is because the system may react to the information from prediction. Typical examples are the recommender systems where the recommendation may guide users' selection by predicting what they may be interested in, and the stock market where people may purchase the stock whose price is predicted to go up.

2.2. Basic procedures

Prediction is a highly data-driven science. Data from real systems with time information are commonly used to examine the effectiveness of the prediction methods. Usually, the data from real systems are divided into a training set E^T and a probe set E^P according to time. E^T is regarded as the known information and the prediction algorithms run on it. E^P, on the other hand, is treated as unknown information and used to measure the prediction accuracy after the prediction is made. To ensure the prediction methods have sufficient data to extract valuable information for prediction, the size of E^T is usually rather large. However, the size of E^T in some works are deliberately decreased to simulate the cold-start problem. The size of E^P is usually small. Altering the size of E^P has another interpretation. A small E^P is corresponding to predicting the shorter future, while a large E^P is corresponding to long term prediction. The typical size ratios for E^T and E^P are 90% and 10%, respectively. In some cases where the time information of the data is not available, the real data are sometimes divided into E^T and E^P randomly.

In prediction, a method with more parameters may seem to have higher accuracy than a method with fewer parameters. However, it is very important to examine whether the seemingly high accuracy is truly due to the advantage of the method or just a result of over-fitting. A prediction method that has been over-fitted is very sensitive to minor fluctuations in the training data. An easy procedure to examine over-fitting is through a three-fold data division. Instead of dividing the data into a training set E^T and a probe set E^P, the real data are divided into three subsets, i.e. a training set E^T, a learning set E^L and a probe set E^P. Both E^T and E^L are treated as known information, E^L is used to estimate the optimal parameters which are finally used to predict E^L. The accuracy in predicting E^L is the final performance of the method.

To investigate the feedback effect of a prediction method, the predicted results (denoted by a set E^R) can be added to E^T to simulate the situation that the system adopts the prediction or approaches to the prediction in its future evolution. $E^T \cup E^R$ is then compared with the real case $E^T \cup E^P$ to reveal the influence of prediction on the evolution of the system. This approach is especially important for prediction in online systems. For instance, how the recommendation algorithms and search engines affect the popularity of online items. In this case, one can even make a more realistic assumption that only a fraction of the prediction results are adopted by the system and investigate how this faction influence the future evolution of the system. In addition, this approach is also used to examine the network reconstruction performance when the prediction methods are applied to recover missing data.

The data division framework is universal for most of issues for prediction. For instance, the prediction of dynamics such as spreading and cascading failure can be investigated via this framework. The dynamical processes are run until they reach stationary states in networks. The simulation time series data can also divided into a E^T and E^P to estimate the accuracy of prediction methods.

2.3. Evaluation metrics

Prediction needs to be validated. Even though there are numerous evaluation metrics for prediction, one has to choose the most proper ones according to different context. Hofman et al. [8] introduced a basic procedure as shown in Fig. 1. In this subsection, we will introduce several most basic and commonly used metrics for prediction evaluation. Some other metrics for specific prediction problems, we will introduce them in the corresponding sections.

Classification metrics. Numerous prediction problems are binary classification type problems, namely to distinguish what will truly happen from all possibilities. Therefore, the classification metrics in statistics are employed to measure the accuracy of the prediction methods. Within this framework, the *true positive* (TP) measures the number of real future events that are correctly predicted. The *true negative* (TN) focus on the events that are not happening in the future and computes the number of such events that are not included in the prediction. When making predictions, it is possible to fall into two different kinds of error: predicting events that will not happen in the future (known as the type I error); and not predicting events that will happen in the future (known as the type II error). The number of type I and type II errors are respectively denoted as the *false positive* (FP) and *false negative* (FN). These quantitative are further used to design the accuracy metrics called *sensitivity* and *specificity*. The sensitivity reflects the ability of the prediction method in avoiding false positives, which can be written as $TN/(TN + FP)$ in mathematical form. The specificity represents the ability of the prediction method in avoiding false negative, which is simply computed as $TP/(TP + FN)$. A well-performed prediction method is expected to have both high sensitivity and specificity.

Area under the receiver operating characteristic curve. The receiver operating characteristic (ROC) curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. When using normalized units, the area under the ROC curve (AUC) is equal to the probability that a predictor can rank a real future event e_r higher than a randomly chosen event e_c that will not happen in the future. It can be simply calculated as

$$AUC = \int_{-\infty}^{\infty} (TPR(T) - FPR(T))dT, \tag{1}$$

where T is the threshold. AUC ranges from [0, 1]. $AUC = 1$ represents a perfect prediction, while $AUC = 0.5$ is the result of random guessing, representing a worthless prediction.

AUC index can also be approximated in a less computational consuming way. The basic idea is to conduct in total n times of comparisons to directly estimate the probability that e_r is ranked higher than e_c. If, among n times of comparisons, e_r is ranked higher than e_c for n_1 times and they are with the same rank for n_2 times, then AUC can be estimated by

$$AUC = \frac{n_1 + n_2/2}{n}. \tag{2}$$

This approximation is usually used in the link prediction problem where directly computing AUC is overly time-consuming.

Precision and recall. Precision and recall are another type of notable metrics in the literature. In information retrieval, precision and recall are usually defined based on TP, FP and FN. Specifically, precision is calculated as $TP/(TP+FP)$ while recall is calculated as $TP/(TP + FN)$ (the same as sensitivity). These two measures are both independent of TN, which is generally

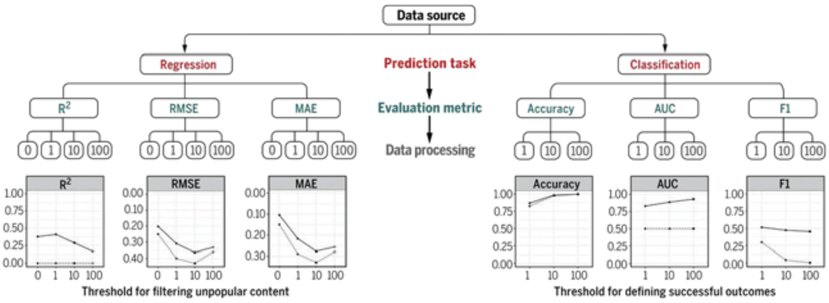

Fig. 1. Basic process of prediction. R^2 is coefficient of determination; AUC is area under the ROC curve; RMSE is root mean squared error; MAE is mean absolute error; F1 score is the harmonic mean of precision and recall. After [8].

unknown and much large in number. Precision and recall have other variants. One important variant is in recommender system where the calculations need to take into account the length of the recommendation list. This variant will be discussed in the corresponding section.

Correlation coefficients. In some problems, the accuracy of a prediction method is estimated by computing the correlation coefficients between the predictor and the variable that needs to be predicted. A high correlation naturally indicates a high accuracy. A typical example is predicting the spreading ability of a node in complex networks with the topological induces. There are three mainstream correlation coefficients, i.e. Pearson coefficient, Spearman coefficient and Kendal's tau coefficient. The Pearson correlation coefficient is defined for two paired sequences (X_i, Y_i) with length N. Mathematically, it computed as

$$r_p = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}, \tag{3}$$

where $\bar{X}$ and $\bar{Y}$ are respectively the mean value of the sequences X and Y. r_p ranges from -1 to 1, with -1 and 1 respectively indicating a completely negative and positive correlations. Though the Pearson coefficient is widely used, it has a obvious drawback. It is very sensitive to extreme values in X and Y. As the heavy-tailed distributions are found in numerous systems, this drawback cannot be neglected.

The second coefficient is the Spearman coefficient. It is defined for the ranks of sequences X and Y. Denoting x_i and y_i respectively as the ranks for the components X_i and Y_i, the Spearman coefficient can be written as

$$r_s = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}. \tag{4}$$

Apparently, the Spearman coefficient is the Pearson coefficient of the ranks of two sequences X and Y. It reflects the monotonic relatedness of two sequences. There is a simple way to calculate Spearman coefficient as

$$r_s = 1 - \frac{6\sum_{i=1}^{N}(x_i - y_i)^2}{N(N^2 - 1)}. \tag{5}$$

The Kendal's tau coefficient also measures the rank correlation between two sequences X and Y. Specifically, it counts the difference between the number of concordant pairs and the number of discordant pairs between these two sequences. The formula reads

$$\tau = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} sgn[(x_i - x_j)(y_i - y_j)]}{N(N - 1)}, \tag{6}$$

where $sgn(x)$ is the sign function, which returns 1 if $x > 0$; 1 if $x < 0$; and 0 for $x = 0$. τ value also ranges between -1 and 1. As Kendal's tau requires a large number of pair comparison, its computational complexity is higher than the previous two coefficients.

3. Node-oriented microscopic prediction of network structure

In the literature, a mathematical framework is proposed for predicting the properties of individual based on the scale of complex systems. Kleiber's law suggests that for the majority of animals, the 0.75 power of body mass is the most reliable basis for predicting the ratio of metabolism [9]. West's observation also summaries similar scaling prediction in quantity arising from complex systems [10]. As the big data era comes, powerful computers made it possible to analyze data for complex systems containing a large number of components: from man-made systems and social systems such as the World

Table 1
Linear and nonlinear growth mechanisms. Here $P(k)$ is degree distribution.

Name	$\prod(i)$	$P(k)$	Ref.
Preferential attachment	$\propto k_i$	k^{-3}	[12]
Asymptotically linear	$\propto ak_i$	$k^{-\gamma} \begin{cases} \gamma \to 2 \text{ if } a \to \infty \\ \gamma \to \infty \text{ if } a \to 0 \end{cases}$	[13]
Asymptotically linear	$\propto b + k_i$	$k^{-\gamma} \begin{cases} \gamma \to 2 \text{ if } b = 0 \\ \gamma \to \infty \text{ if } b \to \infty \end{cases}$	[14,15]
Multiplicative node fitness	$\propto \eta_i k_i$[a]	$\sim \frac{k^{-1-C}}{\ln(k)}$	[16]
Additive–multiplicative fitness	$\propto \eta_i(k_i - 1) + \varsigma_i$	$\sim \frac{k^{-1-m}}{\ln(k)}$; $(m \in [1, 2])$	[16]
Nonlinear	$\propto k_i^{\alpha}$	–	[13]

[a] If give a uniform fitness distribution and $C = 1.255$.

Wide Web and the Internet, online social networks, networks of movie actors, scientific collaboration networks, epidemic networks, and the sex web to biological systems such as the metabolic network, protein interaction networks, cell-signaling and food webs. In general, the complex networks consist of nodes that interact with each other, so that the interactions have both some level of regularity and randomness. We will introduce node-oriented microscopic prediction of network structure. In this subsection, the content includes model-based node degree prediction, extrapolation-based node popularity prediction, node influence prediction, and discovering the hidden nodes.

3.1. Model-based node degree prediction

One of the basic assumptions of evolving network models is that the total number of links in a growing network is a linear function of the network size, that is, of the total number of nodes. This linear growth does not change the mean degree (i.e. the number of connections per nodes) of the network. Besides, there are also growth models with asymptotically linear or node attributes extend the research. However, in many real growing networks, the node degree evolves with time. The recent works take into account the temporal effects of the network growth. The framework of a simple growing network can be described that at time $t + 1$ a new node is introduced to connect to an existing node i with the probability $\prod(i)$ and the node i has k links before time t i.e. degree. It predicts a strong relationship between a node's age and its degree. In this subsection, we mainly classify related works into degree prediction without aging effects and with aging effects.

3.1.1. Network growth models without aging effects

We first introduce several classic network generation models. In a random network like classic Erdős–Rényi (ER) model [11], a node is introduced to connect to another node i with a constant probability $\prod(i) = p$ $(0 \leq p \leq 1)$, and thus the degree distribution follows Poisson. In small world networks, each link is rewired to connect two randomly selected nodes with a constant probability $\prod(i) = p$ $(0 \leq p \leq 1)$ and finally, degree distribution also follows Poisson. However, the empirical results from real networks show that these real networks are growing with new nodes added to the existing networks, and finally the degree of the network follows power-law, as opposed to the Gaussian or Poisson degree distributions of random networks. The famous model for the growth of networks with the mechanism called preferential linking, for instance the Barabasi–Albert model [12]. Mathematically, $\prod(i) \propto k$. Thus, this linear type of growth is usually considered to be a natural feature of growing networks, and is a basis of other linear and nonlinear growth as $\prod(i) = b + ak^{\alpha}$, see Table 1.

3.1.2. Degree growth with aging effects

Despite the preferential attachment provides a common framework for many theoretical models and empirical data sets, it neglects the temporal effects of network growth. Many real systems display the number of links increases faster than the number of nodes, thus the increase of the average degree in growing network describes the corresponding scaling relations for the accelerated growth. If one considers constant initial attractiveness b in directed networks [15], at time t a new node is added and points to a number of random nodes in the network. Additionally there are $c_0 t^{\theta}$ links are generated and each node is directed to a higher in-degree nodes i with asymptotically linear preferential attachment $\prod(k_{in}) \propto b + k_{in}$. The average degree of network is $\langle k \rangle \sim t^{\theta}$ and the corresponding degree distribution is $P(k) = k^{-\gamma}$, here $\gamma = 1.5$ if $\theta \to 1$, and $\gamma \to 2$ if $\theta \to 0$. If one considers internal links with a probability [17], the average degree of network is $\langle k \rangle = at + 2b$ and corresponding degree distribution also preserves power-law form.

In another model, new nodes are added to the network with a constant probability, and the selected nodes connect to b existing nodes in the network with preferential attachment ($\prod(i) \propto bk_i$). Additionally, the number $aN(t)$ of links is a percentage a of the nodes $N(t)$ that are present in the network are chosen. The probability that a link connects nodes i and j is expressed as $\propto aN(t)k_i k_j$. As a result, the average degree of the network is $\langle k \rangle = at + 2b$ but the power-law degree distribution displays a segmentation at a critical degree. In addition, in some systems, the probability that a new node connects to a node i is not only proportional to the degree k_i, but also depends on its age. In the gradual aging model [15], a new node is introduced to connect to an existing node i with the probability $\prod(i) \propto k_i(t-t_i)^{-\upsilon}$, where υ is a tunable parameter, t_i is the age of node i. The observation of empirical data sets reveals papers and actors gradually lose their ability to attract more citation and collaboration.

In the mentioned methods, if the degree growth of an observed network is agreed with a proposed model, one can use the model to fit the real degree growth and then use the model to predict the future degree. But usually, the degree growth of real networks deviate from these proposed models. Medo et al. [18] proposed a fitness model for growing networks based on the empirical observations. A new node at time t is introduced to connect to an existing node i with the probability

$$\prod(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)}. \tag{7}$$

Here $R_i(t)$ is the relevance of node j at time t which is can be observed in the real network. According to the preferential attachment (PA), the temporal degree that the node i owned at time $t+\Delta t$ is predicted by $\Delta k_i(t+\Delta t)_{(PA)} = \Delta L(t+\Delta t)k_i(t)/L(t)$, where $\Delta L(t+\Delta t)$ is the number of links added to the network during Δt, $L(t)$ is the number of cumulated links at time t. If, in the reality, the node i actually owns the degree $\Delta k_i(t+\Delta t)$ during Δt, the relevance can be defined as the ratio between the actual and the expected degree during Δt,

$$R_i(t, \Delta t) = \frac{\Delta k_i(t+\Delta t)}{\Delta k_i(t+\Delta t)_{(PA)}} = \frac{\Delta k_i(t+\Delta t)L(t)}{k_i(t)\Delta L(t+\Delta t)}. \tag{8}$$

Ren et al. [19] used the relevance model in the bipartite networks to characterize how the popularity of online contents evolved over time, and found that the popularity of the online contents typically exceeded theoretical preferential popularity ($R_i \gg 1$) in the early lifespan, and later restricted to the classic preferential popularity increase mechanism ($R_i \approx 1$).

3.2. Extrapolation-based node popularity prediction

The large number of online contents including video, photo, music sharing, blogs, wikis, social bookmarking, collaborative portals, and news aggregators highlight the challenge of predicting how much attention any of it will finally attract. Thus, predicting the popularity of the online contents not only deepens our understanding of complex systems but also has significant implications for marketing and traffic control to policy-making and risk management. The database of online contents' past history produces a big amount of time-stamped data, making it possible to study the dynamics of the online popularity and how it evolves over time on a global scale [20,21].

Predicting the popularity has been widely studied in the literature focusing on videos [22,23], music [24], news [25], and other online social collective dynamics [26]. Cha et al. [27] observed a highly linear correlation between the number of video views on early days and later days on YouTube. Borghol et al. [28] suggested that a strong linear growth law is the most important factor of prediction of popularity [29]. Shen et al. [30] used the reinforcement Poisson mechanism based on the well-known "rich-get-richer" phenomenon to predict the popularity dynamics. Instead of a stronger presence of the rich-get-richer phenomenon, Vasconcelos et al. [31] showed a lower correlation between the early and late popularities in Foursquare. Chen et al. [32] took the view of video popularity lifespan and found that the relative popularity of the online content is dependent on its age and its intrinsic attributes.

Furthermore, the prediction of popularity is by no means restricted to "rich-get-richer" behaviors, and should take into account the exogenous attributes of the online contexts [33]. Accordingly, Ratkiewicz et al. [34] examined the popularity of Wikipedia topics and Web pages and presented an evolving model that combines the classic preferential growth mechanism with the influence of exogenous factors. The exogenous attributes are incorporated in modeling and predicting evolving popularity dynamics in user-generated videos [35], the citation of scientific release [18,36], and the activity of scientists [37].

In addition, there are also some novel prediction approaches for item popularity like considering local clustering behavior of users [38] and using machine learning techniques to explore item attributes in different dimensions [39]. In this subsection, we will review some typical prediction methods of item popularity based on "rich-get-richer" mechanism and item attributes.

3.2.1. Linear models

The very popular items are thought to result from a positive feedback mechanism leading to rich-get-richer. Motived by this, Szabo et al. [40] proposed a linear model to predict the popularity. At first, one can give two definitions. Reference time t_r is the time at which they intend to predict the popularity of an item whose age with respect to its upload (promotion) time. Indicator time t_i is that when in the life cycle of the item they performed the prediction ($t_i < t_r$). Thus, the linear model is proposed as,

$$\begin{aligned} \ln N(t_r) &= \ln[r(t_i,t_r)N(t_i)] + \xi(t_i,t_r) \\ &= \ln r(t_i,t_r) + \ln N(t_i) + \xi(t_i,t_r), \end{aligned} \tag{9}$$

where $N(t)$ is the popularity of the item at time t. $r(t_i, t_r)$ accounts for the linear relationship between the log-transformed popularities at different times. ζ is a noise term describing the randomness one can observed in the data. There is also an alternative description for the observed correlations: let t_i vary in the model, the popularity at the given time t_r should be described by the following formula,

$$\ln N(t_r) = \ln N(t_0) + \sum_{\tau=t_0}^{t_r} \eta(\tau), \tag{10}$$

$\eta(\tau)$ is the noise following an arbitrary, fixed distribution, and τ is taken in small, discrete time steps. t_0 is an early point in time after the birth time of an item.

3.2.2. Weighted linear models

Preferential attachment is a well-known mechanism of the linear growth law which assumes that the probability a node to attract a new link is proportional to its cumulative degree. Zeng et al. [41] considered that items which are popular at time t are expected to have better chances to become more popular. This implies that the cumulative degree of an item $k_\alpha(t)$ and $\Delta k_\alpha(t)$ is a great predictor of its future popularity increase. Mathematically, the prediction of the popularity of a given target item α,

$$\begin{aligned} s_\alpha(t, T_p) &= (1-\lambda)k_\alpha(t) + \lambda \Delta k_\alpha(t, T_p) \\ &= k_\alpha(t) - \lambda k_\alpha(t - T_p); \end{aligned} \tag{11}$$

When $\lambda = 0$, the predictor simplifies to the total popularity method; When $\lambda = 1$, the prediction equals to the recent popularity. The method can be extended to the weighted popularity prediction in the bipartite networks as the following form,

$$s_\alpha(t^*, T_p) = \sum_i (A_{i\alpha}(t^*) - A_{i\alpha}(t^* - T_p))k_i(t^*)^\gamma . \tag{12}$$

Here $A_{i\alpha}$ is the interaction between an item α and a user i in a bipartite network, for example, in a rating system, a user i rate 5-stars to a movie α. γ is a weight parameter which quantifies the activeness of users. If it is extended to the user social network, $k_i(t^*)$ can transform to the influence of user i as I_i. In addition, one can consider the mechanism that the influence of a link exponentially decays with time [42]. An aging function is accordingly introduced to calculate the prediction of popularity as,

$$s_\alpha(t) = \sum_i A_{i\alpha}(t) e^{\gamma(T_{i\alpha} - t)}, \tag{13}$$

where $T_{i\alpha}$ denotes the time at which user i selected an item α. γ is a positive parameter which controls the decay speed. A larger γ indicates a faster decay, and $\gamma = 0$ corresponds to the cumulative popularity without any decay.

3.2.3. Reinforced Poisson models

Shen et al. [30] proposed a generative stochastic framework applying a reinforced Poisson process to predict the item popularity. The reinforced Poisson process takes account three important factors: (1) Fitness of an item which can characterize its intrinsic competitiveness against other items; (2) The relaxation function corresponds to the aging effect on the ability to attract new attention; (3) A reinforcement mechanism according to the well-known rich-get-richer phenomenon. Taking a citation network as an example, and the reinforced Poisson process is defined as the rate function $x(t)$ for a given paper,

$$x(t) = \lambda k(t) f(t; \theta), \tag{14}$$

where λ is the fitness of a paper, $k(t)$ is the total number of citations received until time t, $f(t; \theta)$ is the relaxation function that characterizes the ability to attract new attention which affected by the aging parameters θ. Consider the aging of papers and assume a log-normal relaxation function in citation network as well as in other online contents, one can obtain,

$$f(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma t} \exp(-\frac{(\ln t - \mu)}{2\sigma^2}). \tag{15}$$

The parameters μ and σ can be calculated by maximizing the logarithmic likelihood.

3.2.4. Self-avoiding mass diffusion models

Zeng et al. [43] devised a self-avoiding mass diffusion (SAMD) method which outperformed extrapolation in the long run and predicted future popular items long before they become prominent. It is noted that the diffusion processes have been applied to design the recommendation algorithms which can be regarded as a prediction problem [44–47]. Consider a bipartite network $A_{i\alpha}$ built by the interaction between an item α and a user i, the initial condition and the diffusion processes is modified as follows:

(1) The initial condition is as $m_i^{(\alpha)}(0) = a_{i\alpha}(t) - a_{i\alpha}(t - T_h)$.

(2) The diffusion process is from a user to a user only considered connections between time $(t - T_h)$ and t, thus, the diffusion function is defined as,

$$W'_{ij}(t, T_h) = \frac{1 - \delta_{ij}}{\Delta k_i(t - T_h; t)} \sum_{\beta=1}^{Z} \frac{\Delta a_{i\beta}(t - T_h; t)\Delta a_{j\beta}(t - T_h; t)}{\Delta k_\beta(t - T_h; t)}; \tag{16}$$

where $\Delta a_{j\beta}(t - T_h; t)$=$a_{j\beta}(t)$-$a_{j\beta}(t - T_h)$ and $\Delta k_i(t - T_h; t)$=$k_i(t)$-$k_i(t - T_h)$. since $t \geq t - T_h$, $\Delta k_i(t - T_h; t)$ and $\Delta k_\beta(t - T_h; t)$ more than 0. Finally, the predicted ranking of popular items is given by

$$s_\alpha(t, T_f, T_h) = \mathrm{rank}(M^{'(\alpha)}(t, T_h)); \tag{17}$$

$$M^{'(\alpha)}(t, T_h) = \sum_{i}^{N} \Delta a(t - T_h; t) m_i^{'(\alpha)}(1); \tag{18}$$

$$\vec{m}_i^{'(\alpha)}(1) = \mathrm{W}'(t, T_h) \cdot \vec{m}_i^{'(\alpha)}(0). \tag{19}$$

The higher the diffusion score, the more similar are the users who have already collected item α. Thus, the item could have higher popularity.

3.3. *Predicting the influence of nodes*

Recently, more and more attention has been paid to predicting individual influence in networks. With an effective algorithm to predicting spreader influence [48–50], we can, for instance, hinder spreading in the case of diseases or accelerate spreading in the case of information dissemination. This problem can be interpreted as predicting the spreading influence of nodes based on its structural properties.

So far, various centrality measures have been applied to predict or identify the node of influence in complex networks. Related classical centrality measures include the degree as the number of neighbors a node connects with, the closeness centrality [51] as the reciprocal of the sum of the geodesic distances to all other nodes, betweenness centrality [52] as the number of shortest paths through a certain node, eigenvector centrality [53] as the component of the eigenvector to the largest eigenvalue of the adjacency matrix, k-shell [54] as the node location in a network.

Lately, a lot of works try to design efficient algorithms that outperform the classical centrality methods. For example, some algorithms focus on directly modifying the basic centrality measures including degree [55], closeness, and betweenness [56,57]. Some works focus on improving the k-shell method by removing the degeneracy of the method [58,59]. Some others try to cut down the computational complexity of eigenvector [60]. Moreover, the concept of path diversity is used to improve the ranking of spreaders [61]. Some methods are also designed in directed networks to identify the influential nodes such as LeaderRank, which is shown to outperform the well-known PageRank method in both effectiveness and robustness [62]. Reviews [48,49] introduced methods in identifying and predicting vital nodes and compared well-known methods on disparate real networks. Another review [50] surveyed the existing methods in ranking nodes of both static and evolving networks.

In real life, many networks are inherently evolving. For example, friends are added and removed in online social networks; the topology of the Internet changes with time; and contacts between mobile devices depend on the time of day. Ghoshal et al. [63] observed that PageRank ranking is sensitive to topological perturbation in random networks. In contrast, in scale-free networks the emergence of super-stable nodes whose ranking becomes independent of perturbations, which due to the fat-tailed nature of the degree distribution. Thus, the classical or extended centralities could manifest different spreader topology in a network, which leads to different efficacy and applicability for identifying vital nodes and predicting the influence of spreaders [64].

In this subsection, we first discuss the effect of tunable network topology on the accuracy of the four centrality methods for predicting the node of significance or influence in undirected networks, then introduce solutions for the dynamical networks with tunable network topology and the complexity of computation in large-scale networks. Finally, we review the age bias of metrics on predicting the node significance in directed growing networks and will introduce predictability of scientific discovery.

3.3.1. *Influence of nodes in dynamic networks*

Normally, an undirected network $G = (V, E)$ with N nodes and M links could be represented by an adjacent matrix $\mathbf{A} = \{a_{ij}\} \in R^{n,n}$, where $a_{ij} = 1$ $(i \neq j)$ if node i and node j are connected, and $a_{ij} = 0$ otherwise. Many topology measures have been proposed to identify the node influence. Here we introduce the procedure of identifying the influence of nodes in undirected networks for simplicity. The topology measurements will generate a ranking list of nodes. In principle, the ranking from an effective ranking method should be as close as possible to the ranking based on the real spreading process. one can employ spreading model to simulate the spreading process on networks such as the susceptible–infected–recovered (SIR) model [65]. In the SIR model, all nodes are regarded as initially susceptible except one infectious node. At each step,

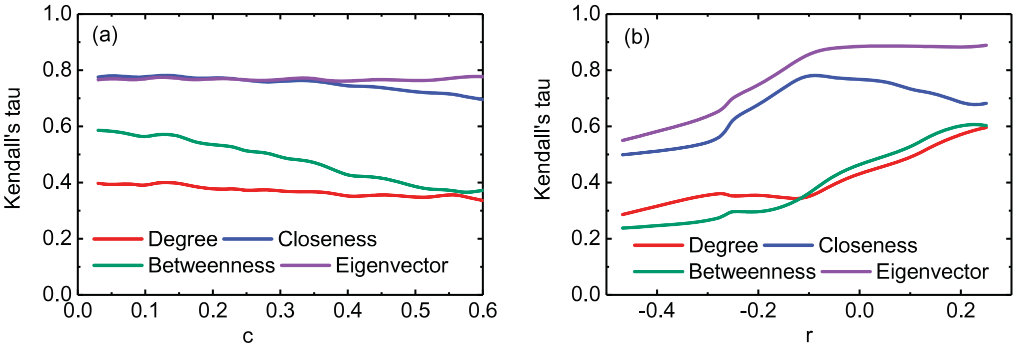

Fig. 2. Tunable network topology affects accuracy of node centrality. (a) Variation of centrality measures accuracy with different clustering coefficient in HK model. (b) The accuracy analysis of four centrality methods on the scale-free network model with tunable assortative coefficient r.

the infected nodes will spread the virus to susceptible neighbors with a certain infectious rate (β), and an infected node will recover after a few spreading steps. The spreading influence of node i is denoted as S_i^β, which is quantified in terms of the total prevalence of the epidemic process, i.e. the fraction of nodes being infected when the infection starts at node i. Based on this, one could obtain the true spreading influence of nodes and a corresponding ranking can be generated by the topology measures. Kendall's tau coefficient τ [66] is used to measure the correlation between the topology-based ranking and the spreading-based ranking. A higher Kendall's tau value τ indicates a more accurate identification of the nodes' spreading influence. Node centrality measurements are based on characterizing the network topology structure in a certain perspective. Actually, the real networks are evolving with time, the evolution of the network topology structure would affect the accuracy of the node centrality. Thus, we investigate the performance of centrality methods for a growing scale-free network model with tunable network topology structure.

We first discuss effects of tunable network clustering and employ the Holme–Kim model [67] to construct scale-free networks with tunable clustering. The HK (Holme–Kim) model is introduced as follows, (1) Preferential attachment (PA). The newly added node connects to the existing node i preferentially, which is described as the same as the preferential attachment. (2) Triad formation (TF): If a link between j and i was added in the previous PA step, then add one more edge from j to a randomly chosen neighbor of i with a probability.

For comparing with the accuracy of the four centrality measures, we simulate the susceptible–infected–recovered (SIR) spreading on the tunable clustering networks and calculate the accuracy as the correlation between the centrality value of nodes and their spreading coverage in the network with SIR model. Fig. 2a shows that degree centrality and the betweenness centrality are more accurate in networks with lower clustering, while the eigenvector centrality performs well in high clustering networks, and the accuracy of the closeness centrality keeps stable in networks with tunable clustering. In addition, the accuracy of the degree centrality and the betweenness centrality are more reliable in the spreading process with the high infectious rates than that of the eigenvector centrality and the closeness centrality.

We also investigate the performance of centrality methods for a growing scale-free network model with tunable assortative coefficient [68]. This network model (namely TASF model) is defined as: (1) The newly added node connects to the existing node i preferentially, which is described as the same as BA model; (2) This node selects a neighbor node s of the node i with probability $k^\alpha(i)/\sum_{j\in\Gamma_i} k^\alpha(j)$, where α is the tunable parameter and Γ_i is the neighbor node set of node i. The accuracy in the TASF networks with positive assortative coefficient has different trend from the ones with negative assortative coefficient. Fig. 2b illustrates that the accuracy analysis of four centrality methods on the scale-free network model with tunable assortative coefficient r and different infectious rate β. One can find that when the network changes from disassortative to assortative, i.e. value of r from negative to positive, the accuracy of the degree centrality and the betweenness centrality trends to be larger, but different of the eigenvector centrality and the closeness centrality, whose accuracy at first increases to peak point and then descends. In summary, the assortative coefficient presenting degree–degree correlation significantly influences the accuracy of centrality.

In addition, one can find that the traditional centrality methods are by no means easy to be applied to predict node influence in the dynamic networks. Many real networks are an inherently evolving, and the structure of the network operates affects the performance of prediction. Therefore, it is necessary to introduce methods to predict node influence in dynamic networks. A dynamic network can be represented as a series of static networks in each period. As revealed in Fig. 3, the problem for predicting the average network centrality values of the nodes is as follows: a dynamic network can be observed during k past time intervals and indicated as $G_{1,k}$, the $G_{k+l,k+l+m}$ will a network in future and be unknown now.

A reasonable solution is to use the average centrality value during k past time intervals to predict the average centrality of nodes in the future $G_{k+l,k+l+m}$. Kim et al. [69] designed a prediction function. Assume the past network $G_{1,k}$, and time windows l and m ($k < l < m$), the problem can be formulated on minimizing the predicted centrality and true centrality by using Polynomial Regression. With this notation, the problem is transformed to minimize the average error between the

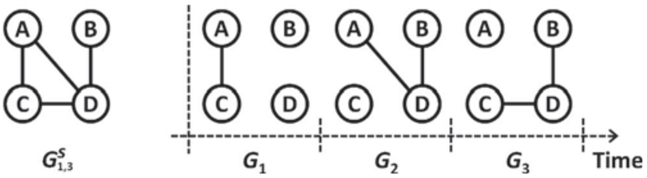

Fig. 3. An example of the dynamic network. (Left) aggregated static network. (Right) time-varying dynamic network. After [69].

guessed centrality values and the true centrality values. Give $G^D_{1,k}$, l and m, one can find $\tilde{C}_{a,b}(u)$ where $a = k+l$, $b = a+(m-1)$ for each $u \in V$ to minimize,

$$Error(G^D_{1,k}, l, m) = \frac{\sum_{u \in V} |C_{a,b}(u) - \tilde{C}_{a,b}(u)|}{|V|} \tag{20}$$

There also exists challenging situation that complex networks would contain millions of nodes and links. Some methods such as closeness and betweenness could better quantify the influence of a node, but they have higher computational complexity due to calculating the shortest paths between all pairs of nodes in a network. Fortunately, one does not need to compute accurately the total centrality, the proposed range-limited betweenness centrality [70,71] can predict betweenness centralities of individual nodes and even have an overlap of 75% for top 100 nodes with betweenness centrality. For the root node i, the initial condition is that $\sigma_{ii} = 1$ for other nodes, $k \neq i$, one can set the $\sigma_{ik} = 0$. The following steps are repeated for each L-range subnetwork, $l = 1, \ldots, L$. The detailed steps are as follows, and an example of the calculation is also shown in Fig. 4.

(1) Build subnetwork $G_l(i)$ using breadth-first search.

(2) Calculate σ_{ik} for all nodes $k \in G_l(i)$ using

$$\sigma_{ik} = \sum_{\substack{j \in G_{l-1}(i) \\ (j,k) \in G_l(i)}} \sigma_{ij}, \tag{21}$$

and set $b^l_l(i|k) = 1$.

(3) Proceed backward through $r = l-1, \ldots, 1, 0$. At first calculate the $l-BCs$ of links $(j, k) \in G_{r+1}(i)(j \in G_r(i), k \in G_{r+1}(i))$ recursively;

$$b^{r+1}_l(i|j, k) = b^{r+1}_l(i|k)\frac{\sigma_{ij}}{\sigma_{ik}}. \tag{22}$$

For nodes $j \in G_r(i)$, one can use above equation and get

$$b^r_l(i|j) = \sum b^{r+1}_l(i|j, k) \tag{23}$$

(4) Finally return to step 1 until the last sub cell $G_L(i)$ is reached. At the end, the cumulative $[l] - BCs$, that is the $B_L(i) = \sum_{l=1} B_l(i)$.

In many real cases such as advertising and news propagation, the spreading only aims to cover a specific group of nodes. Therefore, it is necessary to study the spreading influence of nodes toward localized targets in complex networks. A reversed local path algorithm [72] is devised for this problem. The basic idea is inspired by computing the paths up to length 3 starting from the target nodes to other nodes. The paths with different lengths are aggregated to obtain the final ranking score of a node. Mathematically,

$$I_{RLP} = \sum_{l=0}^{2} \epsilon f {A_{ij}}^{l+1}, \tag{24}$$

where f is a $1 \times N$ vector in which the components corresponding to the target nodes are 1, and 0 otherwise. A is the $N \times N$ adjacency matrix of the network A_{ij}. Here, ϵ is a tunable parameter controlling the weight of the paths with different lengths and is set to be a small value.

3.3.2. *Significance of nodes in growing networks*

As we know, the world overflows with creative works. In reality, it is difficult to measure the significance of works, because the evaluation of the true significance of the work depends on the historical moment, and very much "in the eye of the beholder". Fortunately, thanks to the big data related to the work, we can evaluate significance of the work independently according the network structure, and can predict the work who will potentially win big awards like "Oscar" or "Nobel". Many popular ranking algorithms like Google's PageRank [73] and degree are static in nature which exhibit main

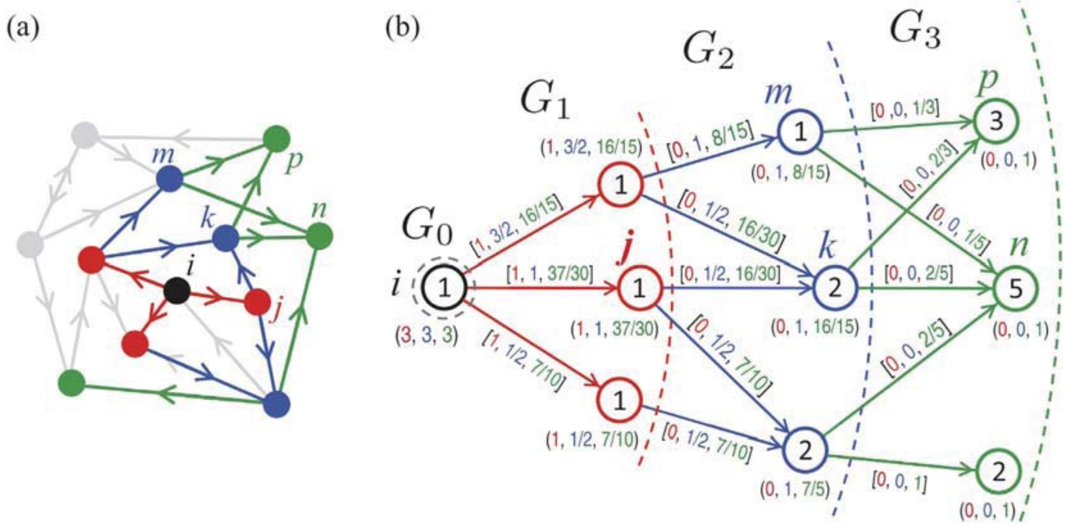

Fig. 4. The calculation of the range-limited betweenness centrality. (a) An example of toy network. The three successive shells of C_3 subnetwork of node i are colored red, blue, green. Gray elements are not part of subnetwork. (b) The $b_l^r(i|j)$ values are corresponding for shells $l = 1$, $l = 2$, and $l = 3$. After [71].

shortcomings when applied to real networks that rapidly evolve over time. Network-based metrics like CiteRank [74] and long gap degree [75] consider time effect. However, a fundamental question is still open: what performance of methods on uncovering significant nodes in the growing networks? Mariani et al. [76] analyzed the relationship between the algorithm's efficacy and properties of the network and showed that realistic temporal effects make PageRank fail in individuating the most valuable nodes for a broad range of model parameters. Mariani et al. [77] also developed a rescaled PageRank centrality with the explicit requirement that paper score is not biased by paper age and identified the Milestone papers [78] and predicted significant papers [77] according to the network of citations among the 449,935 papers published by the American Physical Society (APS) journals between 1893 and 2009.

Furthermore, Medo et al. [79] introduced discoverers as the users in data from real systems who significantly outperform the others in the rate of making discoveries, i.e. in being among the first ones to collect items that eventually become very popular. Furthermore, statistical null models serve this purpose by producing random networks whilst keeping chosen network's properties fixed. While there is increasing interest in networks that evolve in time, we still lack a robust time-aware framework to assess the statistical significance of their observed structural properties. Ren et al. [80] proposed a dynamic null model that preserves both the network's degree sequence and the time evolution of individual nodes' degree values. The proposed model can be used to explore the significance of widely studied network properties such as degree–degree correlations and the relations between popular node centrality metrics.

Recently, the review [50] surveyed the existing ranking algorithms, both static and time-aware, and their applications to evolving networks, and deep understanding of how existing ranking algorithms perform, and which are their possible biases that may impair their effectiveness. Simultaneously, recent advances in predicting the significance of the node in evolving networks have enabled the development of a wide and diverse range of ranking algorithms that take the temporal dimension into account. Here, this subsection will give a comparison of metrics based on network-based metrics according to a growing networks of citation between US movies, with results presented in Fig. 5.

Normally, a citation network $G = (V, E, T)$ with N nodes and M links with time stamp could be described by an adjacent matrix $A = \{a_{ij}^t\} \in R^{N,N}$, where $a_{ij}^t = 1$ if node j is cited by node i (i.e. $i \rightarrow j$) at time t, and $a_{ij} = 0$ otherwise. Now, we introduce a few well-known metrics based on network topology.

Citations is the simplest one, which is defined as the number of times each node is cited as follows,

$$C_i = \sum_j a_{ij}, \tag{25}$$

where the citations could be called in-degree (k_i^{in}) as well. The corresponding out-degree of node i is defined as $k_i^{out} = \sum_i a_{ij}$.

Long Gap The formula for the time lag of a citation is as follows:

$$t = y(i^{out}) - y(i^{in}), \tag{26}$$

where $y(i)$ is an age of node i, and i^{out} and i^{in} are the node on the outgoing and incoming sides of a link, respectively. After calculating the time lag for every link in the network, we count the number of citations with the time lag of at least t year that each node receives. This is the long-gap citation count [75]. It is noted that t equals 25 year in movie citation networks.

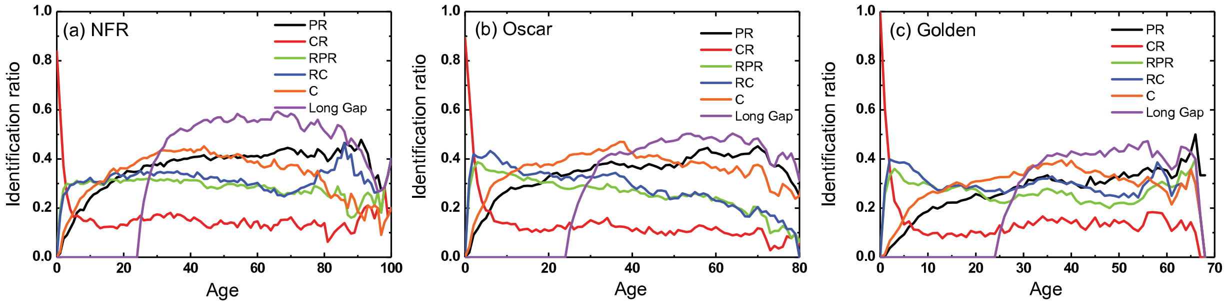

Fig. 5. The identification ratio of five metrics on different age of significant works. We only consider top 5% raking list in movie citation networks and then compute the number of significant works corresponding to their ages.

PageRank The node i of PageRank [73] can be calculated as the stationary solution of the recursive formula as follows,

$$PR_i(s) = \frac{1-d}{N} + d\sum_j \left[\frac{a_{ji}}{k_j^{out}} \mathrm{sgn}(k_j^{out}) + \frac{1}{N}(1 - \mathrm{sgn}(k_j^{out})) \right] PR_j(s-1), \tag{27}$$

where d is damping factor, and $\mathrm{sgn}(x)$ is sign function. Mostly d equals 0.5 which is the usual choice for citation data [81].

CiteRank CiteRank algorithm is a time-dependent variant of PageRank with a teleportation term that decays exponentially with node age, which is intended to favor the recent nodes and thus provide a better representation of nodes' relevance for the current lines of research [74]. The node of CiteRank scores CR_i can be found as the stationary solution of the following set of recursive linear equations,

$$CR_i(s) = (1-d)\frac{\exp(-(t-t_j)/\tau)}{\sum_j \exp(-(t-t_j)/\tau)} + d\sum_j \left[\frac{a_{ji}}{k_j^{out}} \mathrm{sgn}(k_j^{out}) + \frac{1}{N}(1 - \mathrm{sgn}(k_j^{out})) \right] CR_j(s-1), \tag{28}$$

where t_i is the birth date of node i and t is the time at which the scores are computed. We will set $\tau = 2.6$ and $d = 0.5$ in movie citation networks.

Rescaled methods To overcome the well-known PageRank's bias against old nodes in citation data, CiteRank algorithm introduces an exponential penalization for an old node. However, CiteRank score does not allow one to fairly compare papers of different age. The rescaled methods do not depend on paper age [82]. The rescaled methods [77] is derived from Citations and PageRank score respectively and have two steps, we take the rescaled PageRank as an example as follows,

1. Compute PageRank score for each node and label whole nodes in order of decreasing age.

2. For a node i, the mean PageRank score $\mu_i(PR)$, and corresponding standard deviation $\delta_i(PR)$ calculated by the set of nodes $j \in [i - \Delta p, i + \Delta p]$ which are labeled by step 1. Then, the rescaled PageRank score of node i is given as,

$$RPR_i = \frac{PR_i - \mu_i(PR)}{\delta_i(PR)}, \tag{29}$$

where the parameterΔp represents the number of nodes in the averaging window of each node. It is noted that in order to have the same number of nodes in each averaging window, a different definition of the averaging window is needed for the oldest and the most recent Δp nodes. For the oldest and the most recent nodes, one can calculate $\mu_i(PR)$ and $\delta_i(PR)$ over the nodes $j \in [1, 2\Delta p]$ and $j \in [N - 2\Delta p, N]$, respectively. Analogously, the rescaled citation can be computed by the above two steps.

In theory, the calculation of the significance of the node should be independent of the age of nodes, but the above classical or extended metrics could manifest age bias in citation networks which means that some nodes' significance are benefit from their age. The age bias leads to different effectiveness and applicability for identifying and predicting the significance of the nodes.

We now discuss the performance of the metrics for predicting and identifying significant works in a movie citation network. Movie citation network is introduced as follows: Like scientists, artists are often influenced or inspired by prior works, for example, the famous flying bicycle scene in E.T.: The Extra-Terrestrial (1982) is similar to a sequence in The Thief of Bagdad (1924) where characters also fly in front of the moon. The movie citations come in the form of similar quotes, similar settings, or similar movie techniques and so on. Using the movie citations between movies, we can construct a directed network where a node is a movie, and a direct link is a citation. As the above example, we can build a directed link from E.T.: The Extra-Terrestrial (1982) to The Thief of Bagdad (1924). This network consists of 15,425 movies connected by 42,794 citations. The whole movies produced in the United States from 1894 to 2011. The detailed description of the movie citation network can be seen in Ref. [75].

In reality, it is difficult to measure the significance of works, because the evaluation of the true significance of the work depends on the historical moment, and very much "in the eye of the beholder". By definition, we select movies from NFR, Oscar, and Golden Globe three representative awards in the USA filmdom as three significant work lists. The NFR highlights “culturally, historically, or aesthetically significant”, and the requirement of films being released at least 10 years ago. But Oscar and Golden Globe awards are an annual American awards ceremony honoring cinematic achievements in the film industry. In addition, Oscars are awarded from a professional honorary organization, but Golden Globe awards are decided by the wide public attention.

One can consider top 5% ranking list in movie citation networks and calculate the identification ratio between the number of significant works and the works with the top ranking positions. A higher identification ratio means a higher accuracy of the metric. The age bias of metrics for identifying significant works in citation networks is revealed in Fig. 5. The mechanism of their backside should be the main inducement that causes the age bias of metrics for identifying significant works in citation networks.

As we know in citation network, nodes that had been cited many times in the past were more likely to be cited again. Thus the metric of Citations naturally prefer to identifying old nodes, and Long Gap metric takes this advantage to mine hidden significance of old nodes thoroughly. Meanwhile, the famous PageRank algorithm gets to benefit from adaptive and parameter-free and then broadly uses in different areas of science. Compared to PageRank algorithm in favor of old nodes in citation networks, the CiteRank allows an exponential penalization for old nodes, and successes to properly distinguish nodes with early age groups. Further, one way of avoiding that the metric of Citations preferences to old nodes and PageRank fails in the growth of networks owing to the temporal effects, the rescaled methods directly balance age bias of the targeted metrics of Citations and PageRank. These results could give a firmer foundation for age bias of the metrics for identifying significant nodes, which have been studied extensively to describe the dynamics of real evolving networks.

3.4. Prediction of missing nodes

The approaches are used to predict the hidden nodes are either heuristic in nature [83] or rely on rules of network evolution [84]. Recent advance focuses on reconstructing the network using compressive sensing framework to uncover missing nodes in complex networks [85–91]. The basic idea is that treating the system as if there was no hidden node. The neighbors of the hidden node tend to exhibit abnormally dense interaction patterns. According to the multiple time series analysis, some nodes in comparison with those associated with normal nodes do not have hidden nodes in their neighborhoods. To detect a hidden node, it is necessary to identify its neighboring nodes. For an externally accessible node, if there is a hidden node in its neighborhood, the corresponding entry in the reconstructed adjacency matrix will exhibit an abnormally dense pattern or contain meaningless values. In addition, the estimated coefficients for the dynamical and the coupling functions of such an abnormal node typically exhibit much larger variations when different data segments are used, in comparison with those associated with normal nodes that do not have hidden nodes in their neighborhoods. The multiple time-series segments is used to calculate the variance in the reconstructed coefficient vectors for all nodes as follows.

$$\sigma_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{j=1}^{N}(w_{ij}-\hat{w}_{ij})^2}, \tag{30}$$

where T is the number of data segments used in time-series analysis, N the network size and $\hat{w}_{ij}$ the average weights over T simulations. The neighboring nodes of the hidden node are those with abnormally dense connection patterns and significantly larger variances than others. If there are more than one hidden node or time delay, or other situation could be seen in the review [91]. Rossi et al. [83] introduced a categorize techniques for data representation transformations in relational domains that incorporates link transformation and node transformation as symmetric representation tasks to predict the existence of nodes and their relevant features. In some cases, networks in which one knows how the nodes are connected, but the class labels of the nodes or how class labels relate to the network links are missing. Peel et al. [92] then use the relationship between node attributes and network links to accurately predict groups of nodes with similar link patterns.

In social networks, we might know the social and geographical indicators such as age, sex, country of an individual for whom we would like to predict unknown acquaintances. The proposed approach [93] is based on a unified representation of the network data and metadata. The network itself is with an adjacency matrix A where an edge connects two nodes. In the second layer, both the data and the metadata nodes are present, and the connection between them is represented by a bipartite adjacency matrix T. A principled method [93] is used to access both aspects simultaneously, which is constructed for the data and metadata, and a nonparametric Bayesian framework to infer its parameters from annotated data sets. Then this feature can be used to predict the connections to missing nodes when only the metadata are available, as well as predicting missing links. If the metadata correlate well with the network structure, the node membership distribution should place the missing node with a larger likelihood in its correct group. In order to quantify the relative predictive improvement of the metadata information for node i, the predictive likelihood ratio $\lambda_i \in [0, 1]$ is defined as,

$$\lambda_i = \frac{P(a_i|A, T, b, c)}{P(a_i|A, T, b, c) + P(a_i|A, b)}, \tag{31}$$

where $P(a_i|A, T, b, c)$ is the node membership distribution and $P(a_i|A, b)$ is the probability conditioned on the observed partition. For an unobserved node i, they correspond to the ith row of the augmented adjacency matrix, $b = b_i$ and $c = c_i$ are the group memberships of the data and tag nodes respectively.

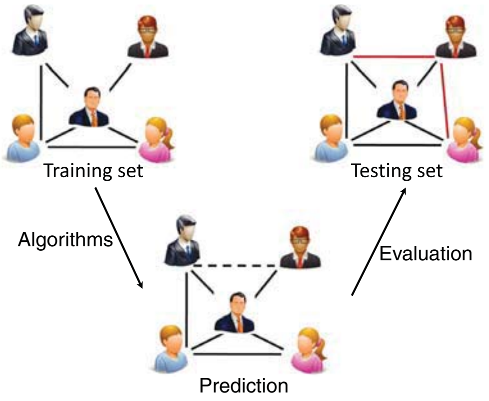

Fig. 6. The basic procedure of link prediction.

4. Interaction-oriented microscopic prediction of network structure

The complex network abstracts from interactions of real-world networks from biological networks such as protein–protein interaction networks, metabolic networks, food webs to information spread, scientific collaboration networks, and online social networks. The interesting questions related the interactions can be posed as: How does the interaction pattern change over time? What are the factors that drive an interaction? How is the interaction between two nodes affected by other nodes? The specific problems that we need to address are to predict the likelihood of an interaction between two nodes, knowing that the probability of interactions between the nodes in the current state of the networks. For instance, the prediction of the actors co-starring in acts and of the collaborations in co-authorship networks, the process of recommending items to users can be considered as a link prediction problem. Interaction-oriented predictions can be used to extract missing links or future links [94], vanishing nodes [95], reciprocal relationships [96], spurious links [97], and so on. The basic procedure of link prediction is shown in Fig. 6. Following the tasks, we will review link prediction in binary and bipartite networks. Then, we will focus on the related works with predicting salient links networks and spurious links specifically.

4.1. Link prediction in complex networks

In this subsection, we will survey an array of methods for link prediction in simple complex networks. There exist a variety of techniques for link prediction, ranging from feature-based classification, matrix property and probabilistic related models [94,98–100]. These methods differ from each other with respect to algorithm complexity, prediction performance, scalability, and generalization ability. We consider three types of models: (1) attributes of nodes and connections based on network structure. The features based on network topology are the most natural for link prediction. (2) Global features of the adjacent matrix using matrix factorization or structural perturbation of matrix, or structural Hamiltonian matrix. (3) probabilistic approaches consist of those based on Bayesian probabilistic models and Markovian approaches.

4.1.1. Features of interaction between node and its neighbors

In fact, many works of link prediction concentrated only on the network topology. Typically, calculation of the similarity is only based on the node neighborhoods and the ensembles of paths between a pair of nodes. Some of the well-known structure-based prediction methods are based common neighbor and preferential attachment, path-based, and random walk [94,98–100]. One can assume an undirected network $G = (V, E)$ and for a node x, let $\Gamma(x)$ denote the set of neighbors of node x.

In common sense, two nodes x and y are more likely to have a link if they have many common neighbors. The simplest measure of this neighborhood overlap is the directed count [101], namely

$$S_{xy}^{CN} = |\Gamma(x) \bigcap \Gamma(y)|. \tag{32}$$

The mechanism of preferential attachment can be employed to generate evolving scale-free networks and also predict many node attributes, where the probability that a new link is connected to the old node x is proportional to the degree of node k_x. Motived by this, the corresponding similarity index can be defined as [102]

$$S_{xy}^{PA} = k_x * k_y. \tag{33}$$

Thus, a lot of methods based on common neighbors and integrated the mechanism of preferential attachment can perform in link prediction well as shown in Table 2.

Table 2
Link prediction methods based on common neighbors and the mechanism of preferential attachment.

Name	Similarity	References
Jaccard index	$S_{xy}^{Jaccard} = \frac{\lvert\Gamma(x)\bigcap\Gamma(y)\rvert}{\lvert\Gamma(x)\bigcup\Gamma(y)\rvert}$	[103]
Salton index	$S_{xy}^{salton} = \frac{\lvert\Gamma(x)\bigcap\Gamma(y)\rvert}{\sqrt{k_x k_y}}$	[101]
Sørenson index	$S_{xy}^{sørensen} = \frac{2\lvert\Gamma(x)\bigcap\Gamma(y)\rvert}{k_x+k_y}$	[104]
Hub promoted index (HPI)	$S_{xy}^{HPI} = \frac{2\lvert\Gamma(x)\bigcap\Gamma(y)\rvert}{\min(k_x,k_y)}$	[105]
Hub depressed index (HDI)	$S_{xy}^{HDI} = \frac{2\lvert\Gamma(x)\bigcap\Gamma(y)\rvert}{\max(k_x,k_y)}$	[106]
Leicht–Holme–Newman index (LHN1)	$S_{xy}^{LHN1} = \frac{2\lvert\Gamma(x)\bigcap\Gamma(y)\rvert}{k_x k_y}$	[106]
Adamic-Adar index (AA)	$S_{xy}^{AA} = \sum_{z\in\Gamma(x)\bigcap\Gamma(y)} \frac{1}{\log k_z}$	[102]
Resource allocation (RA)	$S_{xy}^{RA} = \sum_{z\in\Gamma(x)\bigcap\Gamma(y)} \frac{1}{k_z}$	[107,108]

4.1.2. Path based methods

In the real world, the friends of a friend can become a friend. The fact suggests that the path distance between two nodes in a social network can influence the formation of a link between them. The shorter the distance, the higher the chance that it could happen [109]. Mathematically

$$S_{xy} = \min(|p_{x\to y}|). \tag{34}$$

Katz [110] index is based on the all possible paths, which directly sums over the collection of paths and is exponentially damped by the length to give the shorter paths more weights. The mathematical expression reads

$$\begin{aligned} S^{katz} &= \beta A + \beta^2 A^2 + \beta^2 A^2 + \cdots + \beta^k A^k + \cdots \\ &= (I - \beta A)^{-1} - I, \end{aligned} \tag{35}$$

where β must be lower than the reciprocal of the largest eigenvalue of adjacency matrix A which ensures the convergence of the above Eq. (35).

Being alike Katz index, **Leicht–Holme–Newman Index (LHN2)** [106] is defined as,

$$\begin{aligned} S^{LHN2} &= \gamma(I + \beta A + \beta^2 A^2 + \beta^2 A^2 + \cdots + \beta^k A^k + \cdots) \\ &= \gamma(I - \beta A)^{-1}, \end{aligned} \tag{36}$$

where γ and β are free parameters controlling the balance between the two components of the similarity

Local path index [108,111] considers three order length and is defined as,

$$S^{LP} = A^2 + \varepsilon A^3. \tag{37}$$

Furthermore, the dynamics of a random walker on the network is encoded by a transition probability matrix with elements of the form, $p(i \to j)$ measuring the probability that a walker passes from i to j. Motived by this concept, the dynamics of a random walker can be used to the link prediction. Thus the transition probability matrix is defined as $P = D^{-1}A$, where diagonal matrix is calculated as $D_{ii} = \sum_j A_{ij}$, and corresponding to transition state at the step t which interprets as χ_t. The process of a random walker is described,

$$\chi(t+1) = P\chi(t). \tag{38}$$

The concept of hitting time comes from random walks on a network [100,112]. For two nodes, x and y in a network, the hitting time $H_{x,y}$ is defined as the expected number of steps required for a random walk starting at x to reach y. Shorter hitting time denotes that the nodes are similar to each other, so they have a higher chance of connecting in the future.

$$S_{xy}^{HT} = E(\min\{\chi(t) = y|\chi(0) = x\}), \tag{39}$$

where variable $x(t) = y$ denotes that a random walker is at node y at time t. Based on transition probability matrix, the similarity of two nodes between x and y can be redefined as

$$S_{xy}^{HT} = 1 + \sum_{w\in\Gamma(x)} p_{x\to w} S_{wy}^{HT}. \tag{40}$$

According to the definition, the hitting time measure is usually asymmetric. Commute Time counts the expectation of steps used to reach node x from y, and those nodes are needed to reach node y from x. Mathematically,

$$S_{xy}^{CT} = S_{xy}^{HT} + S_{yx}^{HT}. \tag{41}$$

Commute Time also be obtained by

$$S_{xy}^{CT} = M(L_{xx}^{+} + L_{yy}^{+} + L_{xy}^{+}), \tag{42}$$

here L^{+} is the pseudo-inverse of matrix $L = D - A$. Cosine similarity based on L^{+} is,

$$S_{xy}^{\cos} = \frac{v_x^T v_y}{\sqrt{(v_x^T v_x)(v_y^T v_y)}} = \frac{L_{xy}^{+}}{\sqrt{L_{xx}^{+} L_{yy}^{+}}}, \tag{43}$$

where $v_x = e_x\sqrt{L^{+}}$ and ex is a vector of 0 except the entries corresponding to node x that is filled with 1.

According to the definition of the random walk, if the walker is allowed to return to the starting point with a probability of $1 - \lambda$, where $\lambda \in [0, 1]$, then the concept is formally defined as random walk with restart (RWR) [73], whose updating equation is described as follows,

$$\chi(t+1) = \lambda P \chi(t) + (1 - \lambda)e_x. \tag{44}$$

Thus, keep updating x until convergence, the stationary distribution node x can meet

$$\chi_x = (1 - \lambda)(I - \lambda P^T)^{-1} e_x. \tag{45}$$

Finally, the similarity measurement based on random walk with restart between node x and y is

$$S_{xy}^{RWR} = \chi_x(y). \tag{46}$$

4.1.3. Global features of the adjacent matrix

We will survey link prediction using matrix factorization, structural perturbation of matrix, or structural Hamiltonian matrix. We first review link prediction via matrix factorization [113,114]. Suppose that matrix factorization is given by

$$A = U \sum V^T, \tag{47}$$

where R is the rank of A, U and V are orthogonal matrices of sizes M and N respectively, and is a diagonal matrix $\sum$ of singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R > 0$. It is well known that the best rank-k approximation of A is then given by,

$$A \approx U_k \sum_k V_k^T \tag{48}$$

where U_k and V_k comprise the first kth columns of U and V and $\sum_k$ is the kth principal sub-matrix of $\sum$. A matrix of scores for predicting future links can be calculated as,

$$S = \sum_k \sigma_k u_k v_k^T, \tag{49}$$

where u_k and v_k are the kth columns of U and V respectively.

Kunegis et al. [114] generalized dimensionality reduction methods to solve the link prediction problem. Let A and B be two adjacency matrices of the training and test set for the link prediction, and have the same node set, a spectral transformation functions F that maps A to B with minimal error is given by the solution to the following optimization problem:

$$\begin{aligned} &\min\nolimits_F \|F(A) - B\|_F \\ &s.t.\ F \in S \end{aligned} \tag{50}$$

where $\|.\|_F$ denotes the Frobenius norm. Here, the constraint ensures that the function F belongs to the family of spectral transformation functions (S). If give a symmetric matrix $A = U\Lambda U^T$, one can get $F(A) = UF(\Lambda)U^T$ for such an F, where $F(A)$ applies the corresponding function on reals to each eigenvalue separately.

Structural perturbation method (SPM) [115] is based on the hypothesis that the features of a network are stable if a small fraction of edges is randomly removed. In SPM, a small fraction of edges ΔA is removed from the network. Obviously, $A = A^R + \Delta A$. Thus adjacent matrix A^R of the remaining network is decomposed into,

$$A^R = \sum_{k=1}^{N} \lambda_k u_k u_k^T, \tag{51}$$

where λ_k and u_k are the eigenvalue and the corresponding eigenvector for A^R respectively. After perturbation, the eigenvalue λ_k is corrected to be $\lambda_k + \Delta\lambda_k$ and its corresponding eigenvector is corrected to be $u_k + \Delta u_k$,

$$(A^R + \Delta A)(u_k + \Delta u_k) = (\lambda_k + \Delta\lambda_k)(u_k + \Delta u_k), \tag{52}$$

multiply by Δu_k^T and neglecting second-order terms $u_k^T \Delta A \Delta u_k$ and $\Delta \lambda_k u_k^T \Delta u_k$, one can obtain

$$\Delta \lambda_k \approx \frac{u_k^T \Delta A u_k}{u_k^T u_k}. \tag{53}$$

Use the perturbed eigenvalues and keep eigenvectors unchanged, then perturbed matrix can be obtained by

$$\tilde{A} = \sum_{k=1}^{N} (\lambda_k + \Delta \lambda_k) u_k u_k^T. \tag{54}$$

The similarity of nodes i and j is given by the corresponding value of the matrix $\tilde{A} = \{\tilde{a}_{ij}\}$.

The algorithmic framework of structural Hamiltonian can simply summary that the probability of links in a network is estimated according to a predefined structural Hamiltonian [116]. The existence score of a non-observed link is quantified by the conditional probability of adding the focal link to the network while the spurious probability of an observed link is quantified by the conditional probability of deleting the link. The basic idea is considered the high-order loops, and then a structural Hamiltonian is defined as

$$H(A) = -\sum_{k=3}^{k_c} \beta_k \ln(\mathrm{Tr} A^k) \tag{55}$$

β_k are the temperature parameters and the optimal k_c. When $k > 2$, the number of loops of length k that start and end at node i is $[A^k]_{ij}$. Because of $A = U^T \Lambda U$ and $\mathrm{Tr}(A^k) = \mathrm{Tr}(U^T \Lambda^k U) = \mathrm{Tr}(\Lambda^k U^T U) = \mathrm{Tr}(\Lambda^k) = \sum_{i=1}^{N} \lambda_i^k$. Thus, the structural Hamiltonian can be rewritten as

$$H(A) = -\sum_{k=3}^{k_c} \beta_k \ln(\sum_{i=1}^{N} \lambda_i^k). \tag{56}$$

Give an observed network A^o, and the probe set A^P, one can see $A = A^O + A^P$. The parameter to maximize the probability of the appearance of A^O accords to

$$P(A^O) = \frac{1}{Z} \exp[-H(A^o)], \tag{57}$$

where $Z = \sum_{A' \in M} \exp(-H(A'))$. The conditional probability of the appearance of the link (x, y) based on the observed network,

$$S_{xy} = \frac{1}{Z_{xy}} \exp(-H(\overset{\sim}{\mathrm{A}}(x, y))), \tag{58}$$

where $\tilde{A}(x, y)$ is the network s that the observed network by adding the link (x, y), and Z_{xy} is a normalization factor which and plays no role in producing the prediction. In the spurious link identification problem. The spurious can be estimated by the conditional probability of the absence of this link where equation $\tilde{A}(x, y)$ is corresponding to the observed network A^o by removing the link (x, y).

4.1.4. Probabilistic models

Bayesian approaches A Bayesian network encodes probabilistic relationships among distinctions of interest in an uncertain-reasoning problem. The Bayesian approach to learning Bayesian networks amounts to searching for network-structure hypotheses with high relative posterior probabilities. Bayesian networks have been shown to provide a good representation language for statistical patterns in real-world domains. By learning a Bayesian network from data, we can obtain a deeper understanding of our domain and the statistical dependencies in it.

Suppose a domain of discrete variables $\{x_1, x_2, \ldots, x_n\} = U$, and a database of cases $\{C_1, C_2, \ldots, C_n\} = D$, one wishes to determine the joint distribution $P(C|D, x_i)$ of a new case C. Given the database and current state of information ξ. Rather than reason about this distribution directly, the data is regarded as a random sample from an unknown Bayesian network structure Bs with unknown parameters. Use B_s^h to denote the hypothesis that the data is generated by network structure, one can assume the hypotheses corresponding to all possible network structures form a mutually exclusive and collectively exhaustive set [117],

$$p(C|D, \xi) = \sum_{\mathrm{all} B_s^h} p(C|D, B_s^h, \xi) p(B_s^h|D, \xi). \tag{59}$$

In practice, it is impossible to sum over all possible network structures. Consequently one can attempt to identify a small subset H of network-structure hypotheses that account for a large fraction of the posterior probability of the hypotheses.

Rewriting the previous equation, $p(C|D, \xi)$ is

$$p(C|D,\xi) \approx \frac{1}{\sum_{B_s^h \in H} p(B_s^h|D,\xi)} \sum_{B_s^h \in H} p(C|D, B_s^h, \xi) p(B_s^h|D,\xi). \tag{60}$$

From this relation, one can see that only the relative posterior probabilities of hypotheses matter.

The Bayesian approach is not only an approximation for $P(C|D, \xi)$ but a method for learning network structure. When $|H| = 1$, one can learn a single network structure: the MAP (maximum a posteriori) structure of U. When $|H| > 1$, one can learn a collection of network structures. Heckerman et al. [118] proposed probabilistic entity-relationship model not only for the attributes in a relational model, but also for the relational structure itself. Yu et al. [119] proposed a family of stochastic relational models (SRM) for the link prediction. The key idea of SRM is to model the stochastic structure of links via a tensor interaction of multiple Gaussian processes (GP). Consider pairwise asymmetric links l between nodes, the local measurements of a real-valued latent relational function t can be derived as: the network is $U \times U \to R$, and each link $l_{i,n}$ is solely dependent on its latent value $t_{i,n}$, modeled by the likelihood $p(l_{i,n}|t_{i,n})$. Let the relational processes be characterized by hyperparameter $\theta = \{\theta_\Sigma, \theta_\Omega\}$, θ_Σ for the GP kernel function on U and θ_Ω for the GP kernel function on U, a SRM defines a Bayesian prior $p(t|\theta)$ for the latent variables t. Thus, the link likelihood under such a prior is

$$p(R_{\prod}|\theta) = \int \mathit{\Pi}_{(i,n) \in \prod} p(r_{i,n}|t_{i,n}) p(t|\theta) dt, \tag{61}$$

$\theta = \{\theta_\Sigma, \theta_\Omega\}$ where $R_\Pi = \{r_{i,n}\}_{(i,n)\imath\Pi}$. The hyperparameter θ is estimated by maximizing the evidence, which is an empirical Bayesian approach to learn the relational structure of data. Once θ is determined, the link for a new pair of entities can be predicted by marginalization over the a posteriori $p(t|R_l, \theta)$.

Stochastic block model (SBM) [97] can capture the community structure, where nodes are partitioned into groups and the probability that two connected nodes depend solely on the groups to which they belong. Assume that the observed network is a realization of an underlying probabilistic model, either because the network itself is the result of a stochastic process. The set of generative models m could conceivably generate the networks, and $p(M|A)$ the probability that $M \in m$ is the model that is closed to the observed network A. If get a new observation of the network, the outcome would, in general, be different from A; Using Bayes theorem, the probability $p(X = x)$ for an arbitrary network property X is

$$p(X = x|A) = \frac{\int_m p(X = x|M) p(A|M) dM}{\int_m p(A|M') dM'}, \tag{62}$$

where $p(X = x|M)$ is the probability that $X = x$ in the network generated with model M, and where $p(A|M)$ is the probability that model M gives rise to A among all possible adjacency matrices, and $p(M)$ is a priori probability that model M is the correct one. Within the family of stochastic block models, one can evaluate the likelihood of each model M because the probability of any two nodes i and j being connected depends only on the groups to which they belong.

$$\mathcal{L}_{Bm}(A|P, Q) = \prod_{\alpha \le \beta} Q_{\alpha\beta}^{l_{\alpha\beta}} (1 - Q_{\alpha\beta})^{r_{\alpha\beta} - l_{\alpha\beta}}, \tag{63}$$

where $l_{\alpha\beta}$ is the number of links in A between nodes in groups α and β of the partition P, and $r_{\alpha\beta}$ is the maximum number of such links (that is, the number of pairs of nodes such that one node is in α and the other is in β).

Markovian approaches A parameterized probabilistic model (**PPM**) of network evolution [120] displayed that the structure of a network probabilistically changes over time. The basic idea behind the model is as follows: If you have a friend who has a strong influence on you, your association will be highly affected by the friend's association. Based on the hypothesis: $\phi(t)$ is the edge label function at time t, and changes over time. A Markov model in which $\phi(t + 1)$ depends only on $\phi(t)$. An edge label is copied from node l to node m randomly with probability w_{lm} as time evolves. Node k has a strong influence on node i, and there is an edge between node k and node j. Following the above hypothesis, there will likely be an edge between node i and node j. Similarly, if there are no edges between k and j, there will likely be no edge between i and j. There are two possible ways for $\varphi_{ij}^{(t)}$ to assume a particular edge label. One possibility is that node k has copied an edge label to node i or to node j. The other is that $\varphi_{ij}^{(t+1)} = \varphi_{ij}^{(t)}$ and nothing has happened (indicating that a copy happened somewhere else in the network). Following the above discussion, the probability $\varphi_{ij}^{(t+1)}$ of an edge existing between node i and node j at time $t + 1$ can be written as

$$\varphi_{ij}^{(t+1)} = \frac{1}{N-1}\Big(\sum_{k \ne ij} w_{kj}\varphi_{ki}^{(t)} + w_{ki}\varphi_{kj}^{(t)}\Big) + \Big(1 - \frac{1}{N-1}\Big(\sum_{k \ne i,j} w_{kj} + w_{ki}\Big)\Big)\varphi_{ij}^{(t)}, \tag{64}$$

for the case when the copy happens if k copies its label to node i, then k should already have an edge with j. If k copies its label to node j, it should already have an edge with i.

The relational Markov network (**RMN**) is the relational counterpart of Markov Networks [121]. Let V denotes a set of discrete random variables, and v is an instantiation of the variables in V. A Markov network for V defines a joint distribution over V through an undirected dependency network and a set of parameters. For a network G, if $C(G)$ is the set of cliques (not necessarily maximal), the Markov network defines the distribution,

$$p(v) = \frac{1}{Z} \prod_{c \in C(G)} \phi_c(v_c), \tag{65}$$

where Z is the standard normalizing factor, V_c is the node set of the clique c, and ϕ_c is a clique potential function. RMN specifies the cliques using the notion of a relational clique template.

Wang et al. [122] developed a local probabilistic model namely **MRF** for link prediction. The process contains three steps. The first step is to find a collection of central neighborhood sets. Given two nodes x and y, their central neighborhood sets can be found in many ways. The most natural way is to find the shortest path between x and y and then all the nodes along this path can belong to one central neighborhood set. Assume that the set Q contains all the nodes that are present in any of the central neighborhood set. The second step is to obtain the training data for the MRF model, which is taken from the event log of the social network. Typically a social network is formed by a chronological set of events where two or more actors in the network participate. In case of co-authorship network, co-authoring an article by two or more persons in the network is an event. This collection consisted of the set of events is further refined to include only the nodes belonging to the set Q. Assume this collection is the set V_α. In the final step, an MRF model (say, M) is trained from the training data. This training process is translated to a maximum entropy optimization problem which is solved by iterative scaling algorithm. The $P_M(Q)$ is the probability distribution over the power set of Q. Thus, the link prediction can be changed to solve the following optimization problem,

$$P_M(Q) = \arg\max_{p \in P} H(p), \tag{66}$$

where $H(p) = -\sum_x p(x) \log p(x)$. Once the model $P_M(Q)$ is built, one can use inference to estimate the joint probability between the nodes x and y. The advantage of a local mode is that the number of variables in the set V_Q is small, so exact inference is feasible.

Empirical evidence indicates that many real networks are hierarchically organized. Clauset et al. [123] focused on the hierarchical structure inherent in social and biological networks and proposed the hierarchical structure model (**HSM**) to predict missing links. Each internal node r is associated with a probability P_r and the connecting probability of a pair of nodes is equal to $p_{r'}$ where r' is the lowest common ancestor of these two nodes. Give a real network G and a dendrogram D, and let E_r be the number of edges in G whose endpoints have r as their lowest common ancestor in D, L_r and R_r are the number of leaves in the left and right subtrees rooted at r respectively. Then the likelihood of the dendrogram D together with a set of p_r is

$$\mathcal{L}(D, P_r) = \prod_r P_r^{E_r}(1 - p_r)^{L_r R_r - E_r}. \tag{67}$$

For a fixed D, it is obvious that $p_r^* = E_r/(L_r R_r)$, then maximizes $\mathcal{L}(D, P_r)$. Therefore, according to the maximum likelihood method by Eq. (67) and with a fixed D, it is easy to determine P_r that best fits the network G.

4.2. Link prediction in bipartite networks for recommendation

Over the last decade, the rapid growth of information in both online and offline leads to an information overload problem [124,125]. Lots of online surfers would confuse that which one is the best when searching, reading, shopping, entertaining, or even dating [19,126,127]. As shown in Fig. 7, we present an example of a bipartite network which is constituted by the records of consumer purchases. One can recommend an object to a user based on his/her similar past purchases with others or the purchase records of those who have some resemblance to him/her. In a broad view, the personalized recommendations can serve for each consumer depending on the past purchases of the consumer as well as information relating to the similarity of other consumers or items. With this concept, many online business platforms such as Alibaba, Amazon are reported to develop sophisticated information filtering systems to boost their online sales. Therefore, the recommendation (i.e. link prediction in bipartite networks) is reduced to the problem that estimating the valuation for products that have not been seen by consumers in the fact that millions of products exist in online business platform [128,129].

Due to its significance for economy and society, designing an effective recommendation system has received wide attractions in many branches of science such as computer science, information science and interdisciplinary physics [128–130]. One of the most promising information filtering algorithms is the collaborative filtering (CF) [131,132]. The CF makes work according to the database of the users' past history of purchases and the product searching records to offer the personalized recommendation. The generic CF is classified into two broad groups which were memory-based and model-based methods as shown in Fig. 8 [130,133]. The memory-based methods predict missing information and recommend products based on similarity measures between users and products [132,134]. The model-based algorithms use the collection of the user and object information to learn an information filtering model by clustering [135], Bayesian [136], matrix factorization [137,138] and machine learning techniques [139,140].

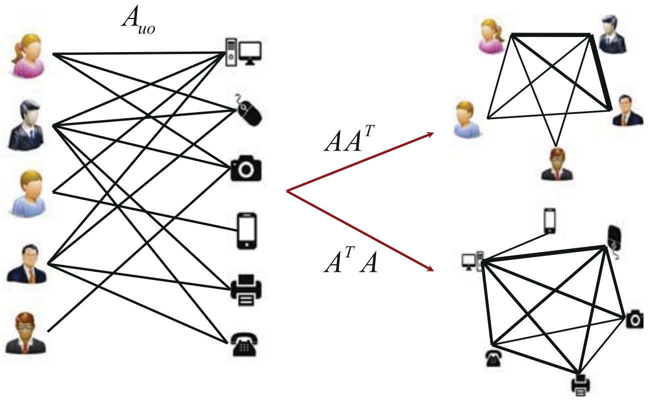

Fig. 7. A recommender can be modeled as in a bipartite network. The recommendation has become a promising and effective way to filter out the irrelevant information and provides personalized suggestions according to the track of past purchases of users. The basic idea in the quantification of similarity between two users is based on the number of objects which have been chosen by both users in the past. It is also possible to define a similarity between two objects based on the number of users who have chosen them.

Being different from the perspective computer sciences, the interdisciplinary physics approaches adapted the complex network theory and various classical physics processes have provided some new insights and solutions for the challenges in the active field of the recommendation [129,141,142], for instance, a diffusion process analogous to the heat conduction process across a bipartite complex network [143], a network-based inference method considering the resource allocation dynamics on bipartite complex networks [44], the opinion diffusion [144] and the gravity principle [145] being extended in the recommendation, the information core and information backbone [146,147] shed some light on the in-depth understanding of the recommendation. Further, the review [129] highlighted a prospect of physicists to a comprehensive guide to recommendation algorithms.

The related CF and interdisciplinary algorithms have already been successfully applied to many well-known recommendation platforms. Meanwhile, many recent works have been devoted to study the expansion of both algorithms, for instance hybrid method [148,149], biased-heat conduction [150,151], multi-channel diffusion [152], preferential diffusion [45,46], hybrid diffusion [47], direct random walks method based on CF [153], hypergraph model with social tag [154,155], multi-linear interactive matrix factorization [156]. These algorithms would further improve the efficiency of the recommendation.

By referring to them, there could exist a simple general formula behind CF, interdisciplinary physics algorithms as well as the extension methods. Motivated by this idea, Ren et al. [157] proposed a simple general model in which employing the dynamics of the random walk in bipartite networks and then derive an analytical expression for tunable parameters of the transition probability matrix. When taking into account the degree information, the process of random walkers can be equivalent to the representative algorithms such as the CF [131,132], heat conduction method [144], network-based inference method [44], hybrid method [148].

Recently, some advance gives promising solutions which aggregate complex network analysis and machine learning techniques. For instance, a novel method named the Multi-Linear Interactive Matrix Factorization (MLIMF) model the interactions between users and the factors (e.g. emotions, locations, the time, movie genres, movie directors), which may have the significant influence on the user's decision process [156]. In addition, a superior latent collaborative retrieval model (TIIREC) integrate the possible item-based information into basic latent collaborative retrieval model [158]. The proposed model can be easily generalized to deal with many other tasks involving to model ternary interaction among entities such as collaborative image annotation, personalized search. The purpose of recommendation can be simply defined as generating a personalized ranking list of objects to fit a particular user's tastes with respect to a given query. To achieve this goal, the proposed approaches define a scoring function $f(\cdot)$ to represent the relevance of a given triple (query, user, object) $\in Q \times U \times O$, where Q, U, O denote the set of queries, users, objects respectively.

In practice, only the top-k retrieved objects could draw users' attention. Thereby, the learned scoring function $f(\cdot)$ should promote users' interesting objects to high position as much as possible for a particular query. Analogous to matrix factorization approaches, the relationship between each pair entities can be measured by the dot-product of their latent factor vector (LCR). Formally, LCR's parameter space includes matrices $S \in R^{|Q|\times n}$, $V \in R^{|U|\times n}$, $T \in R^{|O|\times n}$, which denotes the feature matrix of queries, users, and objects, respectively. To preciously evaluate a user's preference on an object with respect to a given query, LCR additionally allocates each user u an encoder matrix $U_u \in R^{n\times n}$. The scoring function $f(\cdot)$ of LCR model can be then given as follows:

$$f_{LCR}(q, u, o) = S_q U_u T_o^T + V_u T_o^T, \tag{68}$$

where S_q represents the row of S corresponding to query q. V_u is the row of V corresponding to user u, and T_o denotes the row of T corresponding to object o.

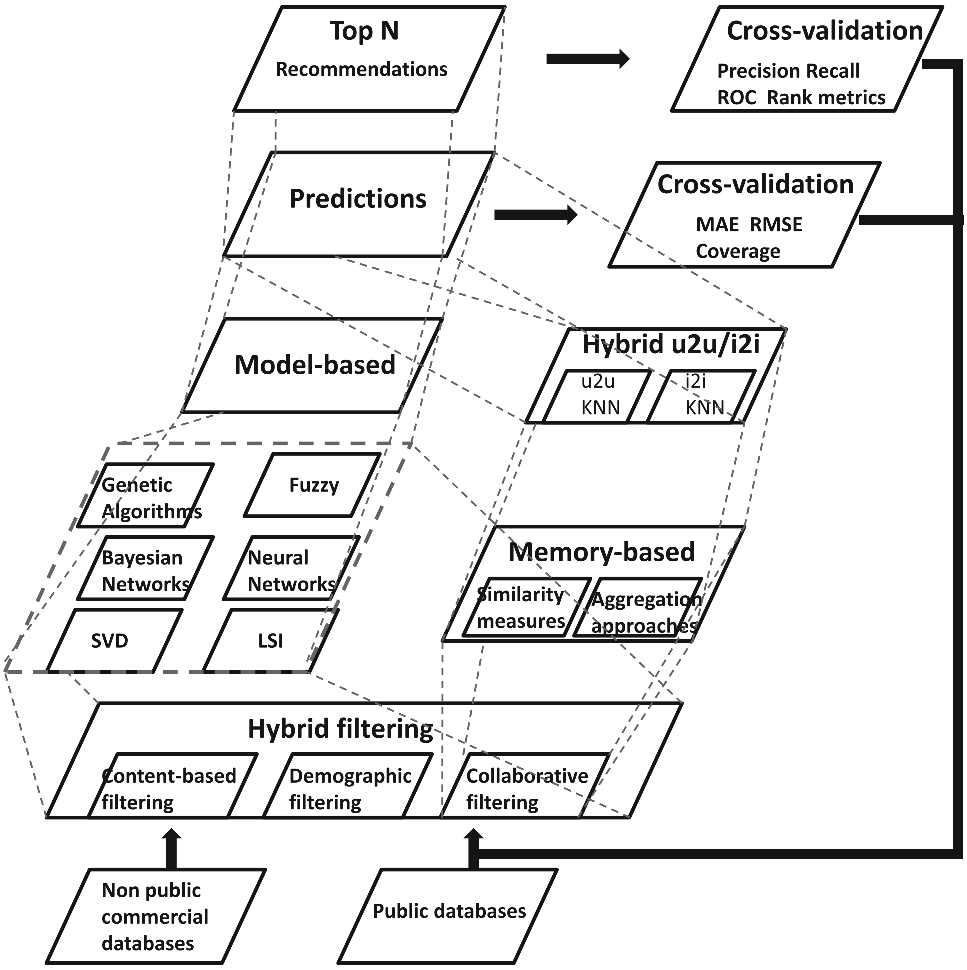

Fig. 8. Traditional models of recommendation and their relationships from the perspective of computer sciences. After [130].

Then objects' collaborative information based on the assumption is regarded objects as the media node could leverage the similar objects with rich descriptive terms to improve the performance of retrieval systems on estimating the ranks of sparse objects. Consequently, TIIREC can be modified as:

$$f = f_{LCR} + S_q A_o V_u^T, \tag{69}$$

where A_o is the linear transformation matrix of object a.

4.3. Predicting salient links and extracting network backbone

Predicting salient links or extracting network backbone in the complex network is one of the most challenging tasks. Many concepts have been proposed to address this problem such as centrality statistics, coarse graining [159–161], which consists of grouping nodes based on topological role in the complex network. In complex networks, the distribution of link weights are usually broadly and few edges account for most of the total weight, so there is a viable option that keeping just the largest weights and set a threshold. If the weights are larger than a predefined threshold, the links could be preserved [162–165]. These remaining links can be regarded as salient links or network backbone. Selecting links with a predefined threshold toward the largest weights would destroy the heterogeneity in the distribution of link weights. This is a crucial feature of complex weighted networks.

Meanwhile, there are a few methods capable of filtering the information on the links so as to respect the multiscale structure of complex weighted networks. Such techniques include a two-stage algorithm [166], a method based on a multilevel network analysis [167], and based on the immediate neighborhood of each node [168]. However, global thresholding methods have a shortcoming which introduces a predefined threshold. To avoid that, one may construct a maximum spanning tree [169], whereas many links as possible are removed such to maintain the connectedness of the network and to keep the largest possible total weight on the remaining links. This traditional technique is also not ideal, as it reduces the network to a tree, whereas cycles are very important structural features of complex networks. In this subsection, we will review three classifications of methods based on centrality statistics, global threshold methods and maximum spanning tree.

4.3.1. Centrality statistics

Betweenness centrality b and link salience s [170] is based on the notion of shortest paths in weighted networks. Given a weighted network defined by the weight matrix w_{ij} and a shortest path that originates at node x and terminates at node y it is convenient to define the indicator function $\sigma_{ij}(y, x) = 1$ if link $i \rightarrow j$ is on the shortest path from x to y, otherwise 0. Edge-betweenness is defined according to

$$b_{ij} = \frac{1}{N^2} \sum_{x,y} \sigma_{ij}(y, x). \tag{70}$$

A shortest path tree $T(x)$ rooted at node x can be represented as a matrix with elements.

$$T_{ij}(x) = \begin{cases} 1 & \text{if } \sum_{y} \sigma_{ij}(y, x) > 0; \\ 0 & \quad \text{otherwise} \end{cases} \tag{71}$$

and then salience s_{ij} of link $i \rightarrow j$ is given by

$$s_{ij} = \frac{1}{N} \sum_{x} T_{ij}(x). \tag{72}$$

There are many criteria to extract the backbone of complex networks on the basis of the topology and dynamics. The hierarchical backbone can be characterized in terms of the concepts of hierarchical degree according to the interpretation of the network weight matrix as a transition matrix, which expresses the total weights of virtual edges established along successive transitions [171]. Glattfelder et al. [167] interpreted such networks as systems in which mass is created at some nodes and transferred to the nodes upstream. The amount of mass flowing along a link from node i to node j is given by the scalar quantity associated with the node j, times the weight of the link, $W_{ij}v_j$. The backbone corresponds to the subnetwork in which a preassigned fraction of the total flow of the system is transferred. Use a unique process-based approach based on the Boolean model to decompose a network into motifs and then apply this technique to two cell-cycle networks, Wang et al. [172] found that each of these networks contains a giant backbone motif spanning all the network nodes that provide the main functional response for two cell-cycle networks. The backbone is, in fact, the smallest network capable of providing the desired functionality.

The extraction of network backbone also can use mutual information from nonlinear time series analysis and betweenness from complex network theory. The discovery of backbone of the climate network accords to uncovering novel pathways of global energy and dynamical information flow in the climate system [173]. Tumminello et al. [174] extracted a subgraph that can be embedded on a surface of the genus which is a topologically invariant property of a surface defined as the largest number of nonisotopic simple closed curves that can be drawn on the surface without separating it, i.e., the number of handles in the surface. The key idea of the method that networks with different degrees of complexity can be constructed by iteratively linking the most strongly connected nodes under the constraint of generating graphs that can be embedded on a surface of a given genus.

In addition, some predicting measures are derived from insights about the topological connectivity and the nonlinear dynamics of the networks. Witthaut et al. [175] integrated the overall network topology with the load distribution resulting from the collective network dynamics and present nonlocal relations to identify a network's response to link failures. On this basis, two network-based strategies were proposed to identify critical links as quantifying the redundant capacity of the network and estimate the flow rerouting through developing a renormalized linear response theory.

4.3.2. Global thresholding methods

Global thresholding methods based on the local identification of the statistically relevant weight heterogeneities can filter out the network backbone and salient links in weighted networks with a strong disorder, preserving structural properties and hierarchies at all scales. The practical procedure can describe that the preserved edges represent statistically significant deviations with respect to a null model for the local assignment of weights to edges. Serrano et al. [168] built this framework to extract the multiscale backbone of complex weighted networks. Marotta et al. [176] extended to detect the backbone of the weighted bipartite network of the Japanese credit market relationships. The procedure can describe as follows, an undirected weighted network with nonnegative link weights between node i and node j. $p_{ij} = w_{ij}/\sum_{l=1}^{k_i} w_{il}$ denotes the fractional edge weight between node i and node j. This edge distribution at node i is compared to a null model in which $k_i - 1$ points are thrown down on the unit interval to create a random distribution of k_i weights that sum to one. An edge is declared to be significant if the probability of observing an edge fraction p larger than p_{ij} under the null model is less than some fixed value, i.e. $P(p > p_{ij}) < \alpha$ for a fixed α. The null model is defined as anomalous fluctuations which provides the expectation for the disparity measure of a given node in a pure random case.

Zhang et al. [177] proposed a globally and locally adaptive network backbone (GLANB) extraction method by synthetically considering the topological structure, i.e. the weights of the links and the degrees of the nodes. The GLANB measures the statistical importance SI_{ij} of link (i, j) by using a null model to calculate the probability. The involvement I_{ij} is compatible with the null hypothesis and defined as edge betweenness or other centrality. The statistical importance SI_{ij} is defined as

$$SI_{ij} = 1 - \int_0^{I_{ij}} f(x; k_i)dx, \tag{73}$$

where $k_i \geq 2$ is the degree of node i. In this study, the involvement follows a uniform distribution, which is similar to what the disparity filter method has performed for the normalized weights of the links as described as Eq. (73). Thus,

$$SI_{ij} = (1 - I_{ij})^{k_i - 1}. \tag{74}$$

To control the impact of the degree on the statistical importance, a parameter is added to the formula as follows:

$$SI_{ij} = (1 - I_{ij})^{(k_i - 1)^c}. \tag{75}$$

If $c = 0$ then the statistical importance SI_{ij} is determined only by I_{ij} and is not affected directly by the degree. As c increases, the impact of the degree becomes larger. The smaller the value of SI_{ij} is, the more significantly the link (i, j) is not compatible with a random distribution. The GLANB can identify a backbone of a network by setting the significance level α for the SI_{ij} based on the distribution of SI, or identify the hierarchical backbones by setting different significance levels since the backbone under high significance level will contain the backbone under low significance level.

Radicchi et al. [178] proposed the Global Statistical Significance (GloSS) method, which satisfies these constraints that the edge between nodes i and j with observed weight w_{ij}. The degrees and strengths of i and j are k_i, k_j, s_i, and s_j. This can be formalized by means of a Bayesian approach. Give the degrees and strengths of its end nodes, the probability to observe weight $w_{ij} \neq 0$ on the edge reads

$$p(w_{ij}|s_i, k_i, s_j, k_j) = p_{obs}(w_{ij}) \frac{p(s_i, s_j|w_{ij}, k_i, k_j)}{p(s_i, s_j|k_i, k_j)}. \tag{76}$$

The denominator on the right-hand side is a normalization factor, while $P_{obs}(w_{ij})$ is a well-defined number. In order to estimate the term in the numerator one must take into account that w_{ij}, k_i, k_j are given, and so the free variables contributing to s_i and s_j are the weights of the remaining $k_i - 1$ and $k_j - 1$ connections between nodes i and j, respectively. These weights can be treated as independent random variables in the null model, with the only restrictions that $\sum_{k \neq j} w_{ik} = s_i - w_{ij}$ and $\sum_{k \neq i} w_{jk} = s_j - w_{ij}$. Finally, an edge is declared to be significant if the probability of observing an edge fraction p larger than $p(w_{ij}|s_i, k_i, s_j, k_j)$ under the null model is less than some fixed value, i.e. $P(p > p(w_{ij}|s_i, k_i, s_j, k_j)) < \alpha$ for a fixed α. Despite its apparently high complexity, the computation of the significance level can be carried out numerically in a fast and accurate way.

Foti et al. [179] developed the backbone extraction namely backbone extraction locally adaptive network sparsification (LANs) that did not rely on any particular null model but used the empirical distribution of similarity weight to determine and then retain statistically significant edges. Assume the degree of a node to be the number of positively weighted edges incident to that node. For each node i and all neighbors j. The fraction of non-zero edges with weight less than or equal to p_{ij},

$$f(ij) = \frac{1}{k_i} \sum_{l=1}^{k_i} \delta(p_{il} \leq p_{ij}), \tag{77}$$

where $\delta(\cdot)$ is the indicator function. For each edge, this gives the probability of choosing an edge at random of fractional weight less than or equal to p_{ij}. If $1 - f(ij)$ is less than a predetermined significance level α, the edge is locally significant and include it in the backbone network. Thus, the salient links are the ones that are statistically significant at the level α and cannot be explained by random variation.

4.3.3. *Maximum spanning trees*

Scellato et al. [180] extracted the backbone of a network called Maximum Centrality Spanning Trees (MCSTs), i.e. maximum weight spanning trees where the edge weight is defined as the centrality of the edge. The edge information centrality is a measure relating the edge importance to the ability of the network to respond to the deactivation of the edge itself. The network performance, before and after a certain edge is deactivated, is measured by the efficiency of the network G. The information centrality of a link C_α^I is defined as the relative drop in the network efficiency caused by the removal from G of the edge l.

$$C_\alpha^I = 1 - \frac{eff_{G'}}{eff_G}. \tag{78}$$

The efficiency of the network,

$$eff_G = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}, \tag{79}$$

where G' is the network with N nodes and $l-1$ edges obtained by removing edge α from the original network G.

Transport properties in random and scale-free networks are studied by analyzing the betweenness centrality B distribution $P(B)$ in the minimum spanning trees (MSTs) and infinite incipient percolation clusters (IIPCs) of the networks. Choi et al. [181] studied the transport property in complex networks through the scaling behavior of $P(B)$ on two different transport backbones, MST and IIPC. From the numerical analyses, it is found that $P(B)$ measured on the transport backbones scales as $P(B) \sim B^{-\delta}$. The transport backbones of the constructed networks are extracted. IIPC is obtained in the following way. Starting from the given network, a link is randomly chosen to be removed. Then, $\mathbb{k} = \langle k^2 \rangle / \langle k \rangle$ is checked. If $\mathbb{k} > 2$, the removing process of links is continued. Otherwise, the removing process is stopped because the largest cluster in the network becomes IIPC [164,182] at $\mathbb{k} = 2$. To extract MSTs of constructed scale-free and random networks, a weight w_{ij} to each link between nodes i and j is assigned by three assignment schemes. The simplest one is the random weight assignment scheme in which w_{ij} is a random number in the interval $[0, 1]$ [183]. In the second weight assignment scheme, $w_{ij} = 1/(k_i k_j)$. An example of this kind of weighted networks is a scientist collaboration network. In the third weight assignment scheme, $w_{ij} = k_i k_j$. An example of the third kind is the airport network. Then, MSTs are extracted by using Prim's algorithm [184] or consider their equilibrium properties and transfer to an optimization problem [185].

Kim et al. [186] demonstrated that the spanning tree separation(STS) was a link salience measure able to identify the backbone of networks. Link salience shares the underlying edge centrality idea with the STS: the more pathways go through an edge, the more important it is. The STS uses spanning trees for pathways. Consider a simple network $G = (V, E)$ with N nodes and M edges, the number of spanning trees containing an edge e_{ij} is

$$STS(G)_{e_{ij}} = b_{ii} + b_{jj} - 2b_{ij}. \tag{80}$$

With $B = \{b_{ij}\} = Det(K_\alpha).K_\alpha^{-1}$. K_α is the augmented Kirchhoff matrix. It equals to the determinant of the Kirchhoff matrix K (the degree matrix minus the adjacency matrix) of D, with the rows and columns corresponding to the element being deleted. The important links represent the informative backbone of a network.

4.3.4. *Extracting network backbone in bipartite networks*

Online systems exist a group of core users who carry most of the information for recommendation [147,187,187]. Zhang et al. [147] proposed two types of criterion combining the time-aware and topology-aware to extract the backbone which contains the essential information for the recommender systems. In order to examine whether it is abundant (or even misleading) information in the online user–object bipartite networks, two categories of link removal algorithms are given: time-aware strategy and topology-aware strategy. The time-aware strategy uses the time information to assign a score for each pair of connected nodes, which is directly defined as their relevance with the underlying assumption that a relevant connection is likely to be a part of the information backbone for the recommendation. Here are four typical algorithms:

(1) System oldest removal (SOR): The link appeared earliest among all the remaining links is removed.

(2) System newest removal (SNR): The link appeared latest among all the remaining links is removed.

(3) Individual oldest removal (IOR): The oldest link for each target user is removed.

(4) Individual newest removal (INR): The newest link for each target user is removed.

Topology-aware algorithms use the network structure to compute the relevance of each link $i\alpha$. Here four typical algorithms are defined:

(1) Most popular removal (MPR): The popularity of a link $l_{i\alpha}$ is defined as $k_i k_\alpha$, where k_i is degree of user i. One can calculate the popularity of all the remaining links and remove the most popular links.

(2) Least popular removal (LPR): The most unpopular links will be removed.

(3) Most rectangles removal (MRR): A rectangle is defined as a subgraph consisting of four links from two users to two items. One can calculate the number of rectangles that each link belongs to, then remove the link with most rectangles.

(4) Fewest rectangles removal (FRR): One can remove the link with fewest rectangles. In order to make all the algorithms comparable, all links should be removed in 50 macro-steps. Therefore, around 2 percent links will be chosen in each macro-step.

Alternatively, the strategy based on heat conduction method [144] and network-based inference method [44] can be described in a more intuitive way [187]. The initial resources placed on objects are first evenly divided among neighboring users and then evenly divided among those users' neighboring objects. In a real network, there can be a lot of neighboring users who have common objects with the target user. Only a few of the most relevant neighboring users should be taken into account in the diffusion. By doing this, there will be less computation in the recommendation and the noisy information from the less relevant users can be reduced. Accordingly, Zeng et al. [187] proposed the k-Nearest Neighbor Mass Diffuse (KNNMD) method in which only the k nearest neighbors of the target users will be considered. Four different ways can be used to identify the most relevant neighbors: (1) Random. When the resources are located at the user side, the random method randomly selects k users as the neighbors; (2) Degree-based. The degree-based method selects k users with the largest degrees as neighbors; (3) Resource-based. The resource-based method selects k users with the largest received resources as the neighbors; (4) Similarity-based ones. It is need to compute the similarities between the target user and other users.

To be able to extract network backbone from weighted bipartite networks. For example, in a country–product trading bipartite network, the Revealed Comparative Advantage [188] is used to parameterize the volume of trade interaction in order to determine whether a trade interaction can be significant or not. One can calculate whether a country's share of a product's world market, is larger or smaller than the product's share of the entire whole market. Mathematically,

$$RCA_{cp} = \frac{v_{cp}/\sum_c v_{cp}}{\sum_p v_{cp}/\sum_c \sum_p v_{cp}}, \tag{81}$$

where v_{cp} is the volume of trade interaction between a country c and a product p. The natural cutoff used to determine whether a trade interaction has revealed comparative advantage is $RCA \geq \lambda$. At this point, the country's share of that product's market is equal or larger than the product's share of the whole market.

4.4. Discovering spurious links

Many observed networks have been discovered to have a class of spurious links. The existence of such spurious links often leads to confusion and misrepresentation in the complex networks. To solve this problem, Guimera et al. [97] quantified the ability to discover spurious links by adding random links to the true network, and then ranking the link reliabilities (again, in decreasing order), finally calculating the probability that a false positive (i.e. $A^O_{ij} = 1$ and $A^T_{ij} = 0$) is ranked lower than a true positive (i.e. $A^O_{ij} = 1$ and $A^T_{ij} = 1$). This process is similar to the framework of the link prediction. So, most of the approaches of link prediction can be used to discover the spurious links such as utilizing link prediction to discover spurious links in the case of protein interaction networks [189] and the stochastic block model to predict spurious links in heterogeneous military network [190].

Recently, the set of attributes in the context information can be developed to clean the spurious links [191]. Accordingly, Zeng et al. [192] proposed a hybrid method that combines similarity-based index and edge-betweenness centrality. The method can effectively eliminate the spurious interactions while leaving the network connected and preserving the network's functionalities. The hybrid index is proposed to combine the similarity-based and the centrality-based approaches. The underlying idea is that a link is a true one either if it connects similar nodes or if it has a central position in the network. This strategy avoids the removal of important links so that the network's properties and functions are preserved with the small drawback of failing to identify a few spurious interactions.

$$R^{hyb}_{ij} = \lambda \frac{R^{CN}_{ij}}{\max(R^{CN})} + (1-\lambda)\frac{R^{BC}_{ij}}{\max(R^{BC})}, \tag{82}$$

where $\lambda \in [0, 1]$ is a tunable parameter.

Besides monopartite networks, the solution can be extended to bipartite networks [193]. The basic idea is that the local diffusion processes are used to measure the inter-similarity (the similarity between nodes of different kinds) in bipartite networks. The inter-similarity reveals asymmetry if the diffusion is applied in different directions. Accordingly, a bi-directional hybrid diffusion method is shown to achieve higher accuracy than the existing diffusion methods in predicting spurious links in bipartite networks.

5. Macroscopic prediction of network structure

5.1. Community prediction

5.1.1. Community detection

Communities are usually groups of nodes having higher probability of being connected to each other than to members of other groups as shown in Fig. 9. The rich set of interactions between individuals in the society results in the community structure corresponding to an actual community such as a group of people brought together by a common interest, a common location or workplace or family ties. In the past decade, there are many related works and surveys that refer to network community detection in complex networks. Fortunato et al. [194] presented a comprehensive review in the area of community detection for undirected networks from a statistical physics perspective, and recently gave a user guide for community detection in [195]. Another survey [196] is to review the methods and algorithms proposed by the wider research community to deal with the clustering in directed networks. A variety of basic measures and metrics are available that can tell us about small-scale structure in networks, such as correlations, connections and recurrent patterns, the review [197,198] discussed community detection on medium and large scales where we are working with larger or denser networks. Networks that can have big size and introduces algorithmic methods for community detection and the development of such methods has been a highly active area of research in the past few years. Thus, the detailed surveys of the metrics proposed for community detection and evaluation can also be introduced in an amount of review literature [199–204].

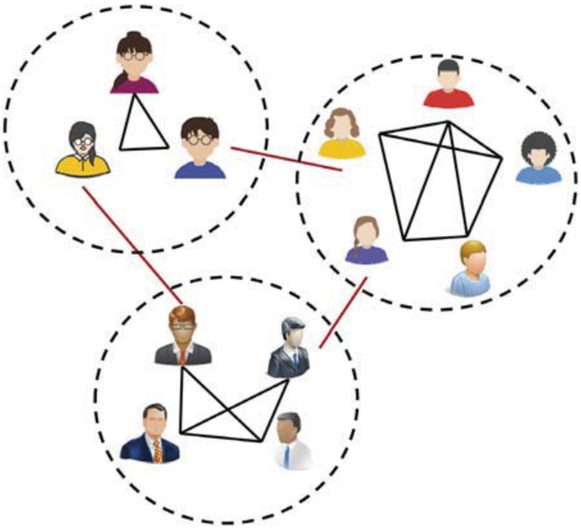

Fig. 9. An example of community detection.

5.1.2. Evolution of community structure

Consider the real situation that frequent changes in the activity and communication patterns of individuals, the associated social and communication networks evolve constantly. The most referred methods governing the underlying community evolution is limited. We give attention to the recent works which highlighted the community detection in evolving networks or the evolving community. Yang et al. [205] investigated the evolution properties of the new community members for dynamic networks. Onnela et al. [206] analyzed the communication patterns of millions of mobile phone users, and observed a coupling between interaction strengths and the network's local structure, which significantly slows the diffusion process, resulting in dynamic trapping of information in communities. An alternative aim is to reduce the computational complexity needed to track the evolving behavior of large networks [207].

Since the network topology is changing continuously, the quantitative description of social networks has to be dynamic. How does network structure affect diffusion? Recent studies [208] suggested that the answer depends on the type of contagion. Hence, the spread within highly clustered communities is enhanced, while diffusion across communities is hampered. Karan et al. [209] tried to capture the essential part of the dynamics and observed that community structure might have evolved from an earlier simpler configuration and what could be its future configuration based on a set of measurable parameters, each with clear meaning in the context of social interaction and social dynamics.

To better understand the relationship between structural communities, due to purely topological connectivity, and the functional clusters, due to the interplay between structure and dynamics. A way based on random walk dynamics can be used to predict and identify the emergence of functional modules in collective phenomena [210]. Palla et al. [211] developed a method based on clique percolation to uncover community evolution. The basic idea is on networks capturing the collaboration between scientists and the calls between mobile phone users. Results showed that large groups persist longer if they are capable of dynamically altering their membership, suggesting that an ability to change the composition results in better adaptability. The behavior of small groups displays the opposite tendency, the condition for stability being that their composition remains unchanged. The knowledge of the time commitment of the members to a given community can be used for estimating the community's lifetime. Young et al. [212] introduced an intuitive model that describes both the emergence of community structure and the evolution of the internal structure of communities in growing social networks. The model comprises two complementary mechanisms: the first mechanism accounts for the evolution of the internal link structure of a single community, and the second mechanism coordinates the growth of multiple overlapping communities. Hebert et al. [213] developed a complex network organization model where connections are built through growing communities, whereas past efforts typically tried to arrange random links in a scale-free, modular and self-similar manner. This model shows that these universal properties are a consequence of preferential attachment at the level of communities: the scale-free organization is inherited by the lower structural levels. Mirshahvalad et al. [214] examined the effect of resampling the original network on community detection and suggested the more dependencies one can maintain in the resampling scheme, the earlier one can predict structural change.

Besides social networks, some works focus on other networks. Tzekina et al. [215] focused on the evolution of trade "islands" in a world trade network in which countries are linked with directed edges weighted according to the fraction of total dollars sent from one country to another. International oil trade is a subset of global trade and there exist oil trade communities. These communities evolve over time and provide clues of international oil trade patterns. Zhong et al. [216] investigated the communities in the oil trade networks and analyzed their evolutionary properties in terms of community number, community scale, distribution of countries, quality of partitions, and stability of communities. Guellet et al. [217] proposed a generic predictor based on local information to predict the effects of different forms of structural stress on the robustness of the metabolic network. Mucha et al. [218] devised a generalized framework of network quality functions to study the community structure of arbitrary multislice networks, which are combinations of individual networks coupled through links that connect each node in one network slice to itself in other slices.

5.2. Topological evolution

5.2.1. Observation of topological evolution

Dynamical networks evolve over time, the main statistical quantities like characterizing networks of degree distribution, clustering coefficient, path length, betweenness centrality, and clusters are corresponding to evolve with time. A lot of works focus on the observation of topological evolution in dynamical networks such as the Internet networks, Facebook, APS, Enron and Wiki data sets [205]. It is well known that the Internet autonomous system (AS)-level topology grows in an exponential form obeying the famous Moore's law. Zhang et al. [219] empirically studied the evolution of AS networks and theoretically predict that the size of the AS-level Internet will double every 5.32 years. Some of its structural properties remain unchanged like that the size of a k-core or k-density with larger k is nearly stable over time [219–221]. Liu et al. [222] gave an attention to k-core with largest k (namely core of network) in online social network and empirically investigated the evolving characteristics of the network core such as degree distribution, clustering coefficient, path length, betweenness centrality. Kossinets et al. [223] analyzed a dynamic Email social network in which interactions between individuals are inferred from time-stamped e-mail headers, and found that network evolution is dominated by a combination of effects arising from network topology itself and the organizational structure in which the network is embedded. In the absence of global perturbations, average network properties appear to approach an equilibrium state, whereas individual properties are unstable. By analyzing the growth of Facebook, the probability of contagion is tightly controlled by the number of connected components in an individual's contact neighborhood, rather than by the actual size of the neighborhood by analyzing the growth of Facebook [55]. Li et al. [224] built a network of listed companies in the Chinese stock market based on common shareholding data from 2003 to 2013 and analyzed the evolution of topological characteristics of the network (e.g., average degree, diameter, average path length and clustering coefficient) with respect to the time sequence. There are some works focus on the evolution patterns of user–object bipartite networks in a large time span [225,226].

5.2.2. Topological evolution according to dynamics

The dynamics of human activity and item popularity is a crucial issue in social media networks. By analyzing the dynamics of human activity and item popularity in social media networks, Zhang et al. [227] proposed an evolving model for such networks, in which the evolution is driven only by two-step random walk. Numerical experiments verify that the model can qualitatively reproduce the distributions of user activity and item popularity observed in empirical networks. Employing the mean-field approach, a detailed theory is proposed to predict the dynamics of the Minority Game system subject to pinning control for various network topologies [228]. Bornholdt et al. [229] studied the topological evolution of an asymmetrically connected threshold network by a simple local rewiring rule: quiet nodes grow links, active nodes lose links. Demetrius et al. [230] introduced the concept of network entropy as a characteristic measure of network topology which can predicted that the evolutionarily stable states of evolved networks will be characterized by extremal values of network entropy. Taking into account the correlation between nodes' degrees and their corresponding data values in the original time series, Manshour et al. [231] showed that topological quantities can also be used to predict the Hurst exponent with an exception for anti-persistent fractional Gaussian noises in complex networks.

Based on a set of measured time series only, the scaling law can be used to predict the node degree, and a set of hub nodes to be predicted in an efficient way, Wang et al. [232] devised a framework can be stated as: Given an unknown network and a set of measured time series from the network, one can infer certain properties of the network based solely on the time series. Firstly, one can estimate the degree k_l of an arbitrary node l. This can be done by disabling any node that is connected to node l. When denote this node by m which is disabled, the degree of node l becomes $k_l - 1$ and its average fluctuation becomes

$$(\Delta x_l')^2 = \frac{\sigma^2}{2c(k_l-1)}\left(1+\frac{1}{\langle k\rangle}\right), \tag{83}$$

which can be measured. c is coupling strength. It should be emphasized that this can be done without explicit knowledge about the network structure and dynamics. Taking the ratio between the original fluctuation Δx_l^2 and $\Delta x_l'^2$, l yields

$$\frac{\Delta x_l^2}{\Delta x_l'^2} = \frac{\sigma^2(1+1/\langle k\rangle)/(2ck_l)}{\sigma^2(1+1/\langle k\rangle)/(2c(k_l-1))}, \tag{84}$$

which gives

$$k_l = \frac{1}{1-\Delta x_l^2/\Delta x_l'^2}. \tag{85}$$

After k_l has been estimated, the degree of any node j in the network can be determined according to scaling law,

$$k_j = k_l \frac{\Delta x_l^2}{\Delta x_j^2}. \tag{86}$$

The advantage of this method is that the errors in the predictions of k_i and c can be eliminated. The prediction error E_k of different degrees for the three types of dynamics is reported generally less than 5% [232].

Even if the network structure at a future time point is not available, one can still predict its properties. Sikdar et al. [233] proposed a standard forecast model of time series to predict the properties of a temporal network such as number of active nodes, average degree, clustering coefficient at a future time instance. Let the size of the window be w, one wants to predict the value of the time series at time t. Consider the time series of the previous w time steps consisting of the values between time steps $t-1-w$ to $t-1$ and fit the regressive model to it and obtain its value at time step t, the procedure for forecasting at every value of t is repeated. Thus, the time step t is the test point and the series of points $t-1-w$ to $t-1$ form the training set. One can imagine this process as a sliding window of size w which is used for learning the auto-regressive equation and the point that falls immediately outside the window is the unknown that is to be predicted. Elementary models of time series forecasting could be categorized into Auto-regressive (AR) and Moving average (MA) models [234]. In case of an auto-regressive model of order p, AR(p) the value of the time series at time step t is given as,

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + e_t + c, \tag{87}$$

where α_i is a parameter, e_t is the white noise error term and c is a constant. Similarly, in case of moving average model of order q, $MA(q)$ the value of the time series at time step t is given as,

$$y_t = \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q} + \mu + e_t + c, \tag{88}$$

where β_i is a parameter, e_t is white noise error term and μ is the expectation of y_t. These two models could be combined into Auto-regressive–moving-average (ARMA (p, q)) [234] where the value of the time series at time step t is given as:

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q} + e_t + c. \tag{89}$$

However, the time series show evidences of nonstationarity and short term dependencies and these models are insufficient and hence Auto-Regressive-Integrated-Moving-Average (ARIMA) model [235] is used for forecasting. The initial differencing step in ARIMA model is used to reduce the non-stationarity. On fitting an ARIMA(p, d, q) model to a time series, an auto-regressive equation of the form is,

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q} + c. \tag{90}$$

Hence one can take a time series corresponding to a network property and fit a regressive model to it. Thus, a regressive equation for that series can be used in forecasting. In order to predict a value at a future time point, one can divide the data in smaller parts and perform the predictions on these smaller stretches.

5.3. *Trend prediction*

5.3.1. *Trend prediction in online social interactions*

Recent developments in digital technology have made possible the collection and analysis of massive amount of human social data and the ensuing discovery of a number of strong online behavioral patterns. Individual interactions are more predictable when users act on their own rather than when attending group activities. By analyzing historical data about 70 years of the domain of American movie, Sreenivasan et al. [236] investigated how novelty influences the revenue generated by a film, and find a relationship that resembles the Wundt–Berlyne curve. Sharda et al. [237] trained a neural network to process pre-release data, such as quality and popularity variables, and classify movies into nine categories according to their anticipated income, from "flop" to "blockbuster". For test samples, the neural network classifies only 36.9% of the movies correctly, while 75.2% of the movies are at most one category away from correct. By applying the sentiment of blog stories on movies, the finding is that positive sentiment is indeed a better predictor for movie success when applied to a limited context around references to the movie in weblogs, posted prior to its release [238]. Asur et al. [239] devised a prediction system for the revenue of movies extracted from the volume of Twitter mentions. Oghina et al. [240] predicted IMDB movie ratings according to extract the movie comments from social media such as tweets from Twitter and comments from YouTube. In addition, one can use a model of social media to predict popularity of news in terms of the number of the volume of comments on online news stories [241,242] or the number of future visitation [243].

Based on large numbers of Google search queries, Ginsberg et al. [244] studied the correlation between the relative frequency of certain queries and influenza activity in each region of the United States and give a new way to predict influenza epidemics in areas with a large population of web search users. Besides using the correlation between social media and predictor. There are some prediction based on the persistence dynamics of trends in online social interaction. Tan et al. [245] treated the popularity of online videos as time series over the given periods and proposed a novel time series model for popularity prediction. The proposed model is based on the correlation between early and future popularity series. Wang et al. [246] introduced a stochastic dynamic model that takes into account the persistence dynamics of trends in social media. The proposed model predicts the distribution of trend durations as well as the thresholds in popularity that lead to the emergence of given topics as trends within social media. Yeh et al. [247] proposed the central location tracking method for the data trend and potency by considering the occurrence order of the observed data. This approach aims at obtaining better predictability and fewer estimation errors for small sample sets. A natural application of this would be the prediction of the society's reaction to a new product in the sense of popularity and adoption rate.

Mestyan et al. [248] built a minimalistic predictive model for the financial success of movies based on collective activity data of online users. The popularity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in a well-known online encyclopedia (Wikipedia). A multivariate linear regression model can be used to predict the box office revenue y. The general form of a regression model at time before release, based on a set of predictor variables s is,

$$y = \sum_{j\in s} \alpha_j(t)x_j(t) + C_s(t) + \varepsilon_s(t), \tag{91}$$

where $\alpha_j(t)$ is time varying parameters of the linear regression model, $C_s(t)$ is a constant and $\varepsilon_s(t)$ is the noise term.

Individual interactions are more predictable when users act on their own rather than when attending group activities. Wang et al. [249] used entropy to measure the randomness of a user's activities. The estimated probabilities for all states $p(i)$ have the property that $\sum_{i=1}^{M} p(i) = 1$. If these probabilities do not change with time, the randomness of user's possible states can be measured by the uncorrelated entropy,

$$H_1 = -\sum_{i=1}^{M} p(i)\log p(i). \tag{92}$$

Notice that if each state is equally probable, this uncorrelated entropy is maximal and equals to

$$H_0 = \log M(i). \tag{93}$$

To measure the randomness of the sequence from knowledge of the previous states, the conditional entropy is defined,

$$H_2(i|j) = -\sum_{j=1}^{M} p(j) \sum_{j=1}^{M} p(i|j)\log p(i|j). \tag{94}$$

Finally the predictability of the user's activity sequence by using the mutual information is

$$I = H_1(i) - H_2(i|j). \tag{95}$$

For each user, the inequalities $0 \le H_2 \le H_1 \le H_0$ are satisfied. I is equal to the amount of information one can gain about the next state by knowing the current state. If there is no second order correlation between state sequences, H_1 is equal to H_2, and I takes the minimum value of 0. If the next state is completely determined by the previous state, or in other words the user activity is completely predictable, I takes the maximum value of H_1.

5.3.2. Trend prediction in the stock market

Stock trend prediction is regarded as one of the most challenging tasks of financial prediction. Bollen et al. [250] analyzed moods of Tweets and based on their investigations they could predict daily up and down changes in Dow Jones Industrial Average values with an accuracy of 87.6%. Some qualitative method is developed for the prediction of stock market trend including using the concept of dynamical Bayesian factor graph [251], the adaptive time-weighted rule voting model [252], the ensemble version of empirical mode decomposition and adding look-ahead bias [253]. With this regard, many machine learning approaches are also used to improve the prediction results. These approaches mainly focus on two aspects: regression problem of the stock price and prediction problem of the turning points of stock price, for instance a new feature construction approach for status box [254], hybridizing fractal feature selection method and support vector machine [255].

6. Applications of prediction

6.1. Prediction in biology networks

6.1.1. Predicting salient nodes in biology networks

Essential nodes in a gene network, a protein network, a metabolic network or a neuronal network are such like genes or proteins are required for the survival of an organism under certain conditions, and the functions they encode are therefore considered a foundation of life. A list of essential proteins are collected from a lot of databases [256–258]. For instance, a protein in protein interaction network is considered as an essential protein if it is marked as essential in one database.

There are about two subclasses in computational approaches. The first classifications are topological centrality measures such as Degree Centrality, Betweenness Centrality, Closeness Centrality, Subgraph Centrality, Eigenvector Centrality, Information Centrality, Bottle Neck, Density of Maximum Neighborhood Component, Local Average Connectivity-based method, Range-Limited Centrality, L-index, Leader Rank, Normalized α-Centrality, and Moduland-Centrality which are introduced in Ref. [49]. The recent advances extend identification of essential nodes by using the topological features of protein–protein interaction (PPI) networks such as based on edge clustering coefficient [259,260], based on the integration of protein–protein [261,262], based on a new combination of local interaction density and protein complexes [263], or the basic idea is that each protein can be viewed as a material particle which creates a potential field around itself and the interaction of all proteins forms a topological field over the network [264]. Besides, some extend topological centrality measures can be applied to both weighted and unweighted biological networks [265].

The second classifications are multi-information fusion measures, i.e., a combination of topological centrality measures and other biological information of proteins, such as protein complexes [266,267], gene ontology terms of proteins [268,269], gene expression data [261,270,271], orthologous information [269], and overlapping essential modules [272]. However, the effectiveness of fusion strategies or mechanisms has not been sufficiently discussed. Therefore, it is critical to design suitable network-level methods integrated appropriately with biological information for prediction of essential proteins.

6.1.2. *Predicting the interaction in biology networks*

Understanding the complex interplay between genes or proteins requires integration of data from a wide variety of sources such as gene expression, genetic linkage, protein interaction, and protein structure among others. Predicting protein–protein interactions is a key role for many areas of biomedical research. Protein networks have been used to identify new disease genes, identify disease-related subnetworks and network-based disease classification [273]. Predictions of physical and functional links between cellular components are often based on correlations between experimental measurements, such as gene expression.

Several bioinformatics methods have been developed to formulate predictions about the functional role of genes and proteins, including their role in diseases [274]. Stark et al. [275] investigated the development of high throughput assays to identify the behavior of proteins, sugars, lipids, and other metabolites in cellular interactions. Rual et al. [276] described pairwise interactions in the human protein–protein interaction network. Bonneau et al. [277] considered relative changes in 72 transcription factors and 9 environmental factors to predicts dynamic transcriptional responses accurately in a free living cell. Yu et al. [278] demonstrated that map of the yeast interactome network provides high-quality binary interaction information. Guo et al. [279] introduced an approach called partial Granger causality to reliably reveal interaction patterns in multivariate data with exogenous inputs and latent variables in the frequency domain. Braun et al. [280] developed a logistic regression model which was trained using the data from these reference sets to combine the assay outputs and then calculate the probability that any newly identified interaction pair is a true biophysical interaction once it has been tested in the tool kit. Snijder et al. [281] presented the hierarchical interaction structure which outperforms commonly used methods in the inference of functional interactions between genes measured in large-scale experiments. Motter et al. [282] proposed an alternative, network-based strategy that aims to restore biological function by forcing the cell to either bypass the functions affected by the defective gene, or to compensate for the lost function. Barzel et al. [283] exploited the fundamental properties of dynamical correlations to develop a method to predict molecular interactions in E. coli networks. The method receives as input the observed correlations between node pairs and uses a matrix transformation to turn the correlation matrix into a highly discriminative silenced matrix, which enhances only the terms associated with direct causal links. The method enhanced the discriminative power of the correlations by twofold, yielding $>50\%$ predictive improvement over traditional correlation measures and 6% over mutual information. Yan et al. [284] applied a control framework to the connectome of the nematode Caenorhabditis elegans. The proposed model can predict the involvement of each C. elegans neuron in locomotor behaviors and the importance of individual neurons in C. elegans locomotion.

6.2. *Prediction in scientific networks*

Recently, Clauset et al. [285] and Zeng et al. [286] surveyed the interdisciplinary field of the science of science and what it allow us to predict scientific discovery and success. Researchers use big data on published works and scientific careers to explore quantitative patterns in the science of science. For example, modern bibliographic databases allow researchers to study citation counts, which provide a convenient measure of scientific impact. Interactions in scientific activities like citation and collaboration mined by the bibliographic databases can construct complex networks. One can use the approaches of predicting the influence of nodes surveyed. However, the citation dynamics of scientific papers appears nonlinear and this nonlinearity has far-reaching consequences, such as diverging citation distributions and runaway papers [287]. Besides the approaches based on general structure and dynamics in complex networks, we main survey two aspects in the field of science in prospect of complex networks. The first one is that predicting impact of the scientific discovery. The other is a topic of influence of scientific researchers.

6.2.1. *Predicting impact of scientific discovery*

Newman [82] considered the "first-mover" advantage in scientific citation network under which the first papers in a field will receive citations at a rate enormously higher than papers published later. Moreover papers are expected to retain this advantage in perpetuity they should receive more citation indefinitely, no matter how many other are published after them.

It is assumed that on average each paper published cites previous papers, which are chosen in proportion to the number k of citations they already have plus a positive constant r. The average number of citations a paper can calculated as:

$$\hat{k}(t) = r(t^{-\frac{1}{\alpha-1}} - 1), \tag{96}$$

where time t is measured using the rescaled time coordinate $t = i/n$. i is ranking position by descending the publication time and n is total papers. Papers published early on should on average receive far more citations than those published later, even allowing for the facts that later papers has less time to accumulate citations. The substantial first-mover advantage is verified in scientific citation network [288].

Wang et al. [4] considered three fundamental mechanisms capturing the temporal citation dynamics of individual papers. Preferential attachment captures the well documented fact that highly cited papers are more visible and are more likely to be cited again than less-cited contributions. Aging captures the fact that each paper's novelty fades by the effect of long-term decay which described by a log-normal survival probability. Fitness captures the inherent differences between papers, accounting for the perceived novelty and importance of a discovery. Combining these three factors, the probability that paper i is cited at time t after publication is

$$\Pi_i(t) \sim \eta_i c_i(t) P_i(t). \tag{97}$$

Eq. (97) can be used to predict the cumulative number of citations acquired by paper i at time t after publication,

$$c_i^t = m\left[e^{\frac{\beta\eta_i}{A}\Phi\left(\frac{\ln t-\mu_i}{\sigma_i}\right)} - 1\right], \tag{98}$$

where

$$\Phi(x) = \frac{\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\,dy}{\sqrt{2\pi}}, \tag{99}$$

$\Phi(x)$ is the cumulative normal distribution. m measures the average number of reference each new papers contains. β captures the growth rate of the total number of publications. A is a normalization constant. These three parameters are global parameters, having same value for all publications. Finally, when $t \to \infty$ in Eq. (98), one can predict the total number of citations a paper acquires during its lifetime.

$$c_i = m(e^{\frac{\beta\eta_i}{A}}). \tag{100}$$

And finally it can drive the impact of scientific journal. These three fundamental mechanisms also can be considered to improve other approaches such as a preferential mechanism to the PageRank algorithm when aggregating resource from different nodes to enhance the effect of similar nodes [289], nonlinearity to the PageRank algorithm when aggregating resources from different nodes to further enhance the effect of important papers [290].

Citations in peer-reviewed articles are generally accepted measures of scientific impact. Online social networks such as Twitter, blogs or social bookmarking tools provide the possibility to construct innovative article-level or journal-level metrics to evaluate impact and influence. However, the relationship of these new metrics to traditional metrics such as citations is introduced in [291]. Social impact measures based on tweets are proposed to complement traditional citation metrics. The proposed metric can measure uptake of research findings and to filter research findings resonating with the public in real time. Tweets can predict highly cited articles within the first 3 days of article publication.

In the science of science, there exists an interesting and attractive phenomenon named "sleeping beauty" [292]. The above introduced models has obvious limitations: It cannot account for "sleeping beauty". Because some papers far exceed the predictions made by simple preferential attachment. A Sleeping Beauty (SB) in science refers to a paper whose importance is not recognized for several years after publication or regarded as that a paper that is little cited (sleeps) for a long period of time and then becomes much cited (awakened) as shown in Fig. 10a. Burrell et al. [293] discussed that the question therefore arises as to whether such awakenings can be explained or expected purely by the random nature of the model or whether they are so unlikely that an alternative explanation should be sought. A systematic analysis of nearly 25 million publications in the natural and social sciences over a time span longer than a century found that sleeping beauties occur in all fields of study as shown in Fig. 10b.

Ke et al. [294] introduced a parameter-free method to quantify to what extent a paper is an SB. Given a paper, c_t is defined as the number of citations received in the tth year after its publication; t indicates the age of the paper. The index B is measured at time $t = T$, and that the paper receives its maximum number c_{t_m} of yearly citations at time $t_m \in [0, T]$ as shown in Fig. 10a. The beauty coefficient value B for a given paper is based on the comparison between its citation history and a reference line. Consider the straight line ℓ_t that connects the points $(0, c_0)$ and (t_m, c_{t_m}) in the time-citation plane. This line is described by the equation,

$$\ell_t = \frac{c_{t_m} - c_0}{t_m} t + c_0, \tag{101}$$

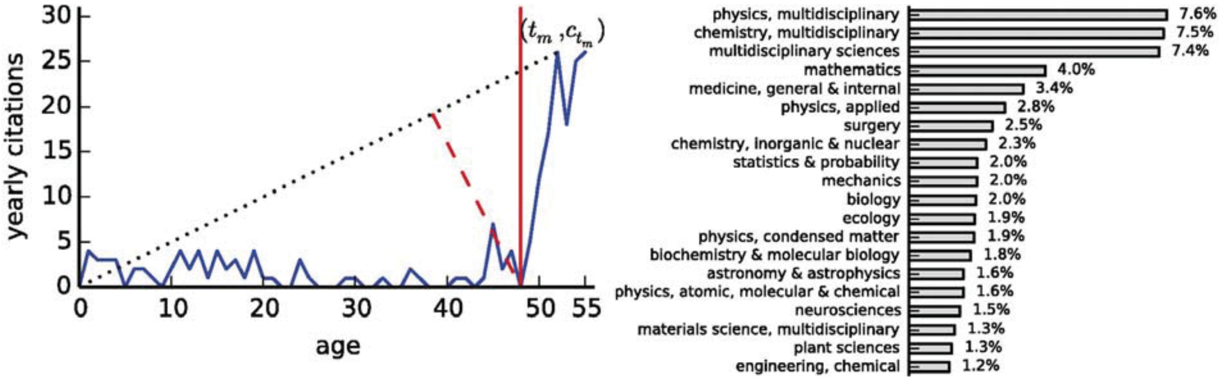

Fig. 10. (a) Illustration of the definition of the beauty coefficient B and the awakening time of a paper. (b) Top 20 disciplines producing SBs in science. Papers with beauty coefficient is in the top 0.1% of the entire WoS database, and one can compute the fraction of those papers that fall in a given subject category. After [294].

where $(c_{t_m} - c_0)/t_m$ is the slope of the line, and c_0 the number of citations received by the paper in the year of its publication. For each $t \le t_m$, then one can compute the ratio between $\ell_t - c_t$ and $\max\{1, c_t\}$. Summing up the ratios from $t = 0$ and $t = t_m$, the beauty coefficient is defined as,

$$B = \sum_{t=0}^{t_m} \frac{\ell_t - c_t}{\max\{1, c_t\}}. \tag{102}$$

Thus $B \ge 0$ and $B = 0$ for papers with $t_m = 0$ means papers with citations growing linearly with time. The awakening time t_a is defined as the time t at which the distance d_t between the point (t, c_t) and the reference line ℓ_t reaches its maximum:

$$t_a = \arg(\max_{t \le t_m} \{d_t\}). \tag{103}$$

The d_t is given by

$$d_t = \frac{|(c_{t_m} - c_0)t - t_m c_t + t_m c_0|}{\sqrt{(c_{t_m} - c_0)^2 + t_m^2}}. \tag{104}$$

We have reviewed representative approaches to predict the impact of a paper or "sleeping beauty". But, how to predict the representative work for individual researcher is another important yet uneasy problem. So far, the representative work of a researcher is usually selected as his/her most highly cited paper or the paper published in top journals. The representative work of a scientist is considered as an important paper in his/her area of expertise. Accordingly, Niu et al. [295] propose a self-avoiding preferential diffusion process to generate personalized ranking of papers for each scientist and discover their representative works [295]. The citation data from American Physical Society (APS) is used to validate the proposed method, which shows that the self-avoiding preferential diffusion method can rank the Nobel prize winning paper in each Nobel laureate's personal ranking list higher than the citation count and PageRank methods, indicating the effectiveness of the method.

6.2.2. Predicting influence of scientific researchers

Bibliometric measures of individual scientific achievement are of particular interest if they can be used to predict future achievement. One can easily understand citation indicators at the time of prediction, namely the number of papers, the total number of citations, the career length, the average number of published papers per year, the average annual citations, the annual citations at the time of prediction, the average citations per paper. Yin et al. [296] took a large-scale quantitative analysis on how time affects citations, and developed a new theoretical framework to reconcile the interplay between temporal decay of citations and the growth of science. More specifically, Mazloumian et al. [297] devised an approach to predicting scholars' scientific impact, which estimated for scholar s the citations to a certain subset of his papers (selected by time-window w) in k subsequent years as shown in Fig. 11. The citation indicators $X = \{x_k\}$ is defined as,

$$c_i = \alpha_{s[i]} + \sum_k \rho_k \log(x_k) + \varepsilon_i, \tag{105}$$

where ρ_k is the coefficient of citation indicator x_k. $\alpha_{s[i]}$ is the intercept estimated for scholar s. The well-known indexes like the h index, the m index, and the g index [298,299]. h index: Natural number h for which the scientist has h papers with at least h citations. M index: as a scientist's h index value divided by the time (years) elapsed from the first publication of the

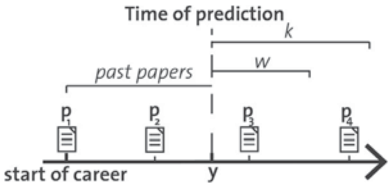

Fig. 11. Illustration of papers published in a scientist career. After [297].

focal scientist. *G* index is proposed by [300] and defined as a scientist's g index value it the highest number g of papers that receives g^2 or more citations. Penner et al. [301] found the performance of h-index strongly depends upon career age for the purpose of predicting a scientist's future. The h-index and similar metrics can capture only past accomplishments, not future achievements. To solve this flaw, Acuna et al. [3] attempted to predict the h-index of each scientist t years ahead by linear regression model as follows

$$h(t+\Delta t) = \beta_0(\Delta t) + \beta_h(\Delta t)h(t) + \beta_{\sqrt{n_p}}(\Delta t)\sqrt{n_p(t)} + \beta_t(\Delta t)t + \beta_j(\Delta t)j(t) + \beta_q(\Delta t)q(t). \tag{106}$$

Here $h(t)$ is the current h-index; $n_{p(t)}$ is the number of publications authored or coauthored; $j(t)$ is the number of distinct journals of the publications; $q(t)$ is the number of papers published in high impact journals; T_n is number of articles written; h is current h-index; y is years since publishing first article; j is number of distinct journals published in; q is number of articles in high impact factor journals like Nature, Science, Nature Neuroscience, Proceedings of the National Academy of Sciences and Neuron.

Sinatra et al. [302] took a data-driven view to analyze the evolution of individual scientific impact and found that the highest-impact work in a scientist's career is randomly distributed within a lot of work. That is, the highest-impact work can be, with the same probability, anywhere in the sequence of papers published by a scientist career. The random-impact rule allows us to develop a quantitative model, which systematically untangles the role of productivity and luck in each scientific career. The model assumes that each scientist selects a project with a random potential p and improves on it with a factor Q_i, resulting in a publication of impact Q_ip. The parameter Q_i captures the ability of scientist i to take advantage of the available knowledge in a way that enhances ($Q_i > 1$) or diminishes ($Q_i < 1$) the potential impact p of a paper. The model predicts that truly high-impact discoveries require a combination of high Q and luck p and that increased productivity alone cannot substantially enhance the chance of a very high impact work. The Q model provides an analytical expression of these traditional impact metrics and can be used to predict their future time evolution for each individual scientist, being also predictive of independent recognitions, like Nobel prizes. The number of citations 10 years after publication c_{10} prompting us to write the impact $c_{10,i\alpha}$ of paper a published by scientist i as

$$c_{10,i\alpha} = Q_i p_\alpha, \tag{107}$$

where p_α is the potential impact of paper α in the sequence of papers published by a scientist i. The parameter Q_i is an individual parameter for a scientist i. The parameter Q_i captures the ability of scientist i to take advantage of the available information, enhancing (or diminishing) the impact of paper α. Q_i is considered to be constant throughout a scientist's career.

Coauthorship is a generic and significant presence in the domain of science. Analyzing a data set of over 100,000 publications from the field of Computer Science, Sarigol et al. [303] studied how centrality in the coauthorship network differs between authors who have highly cited papers and those who do not. The results show that a Machine Learning classifier, based only on coauthorship network centrality metrics measured at the time of publication, is able to predict with high precision whether an article will be highly cited five years after publication. This one also can predict that if a paper is authored by an author with a top 10% betweenness centrality, degree centrality, k-core centrality and eigenvector centrality, then the paper will be among the top 10% most cited papers five years after publication.

Qi et al. [304] used the publication data from American Physical Society (APS) journals to analyze the influence of outstanding scientists on young collaborators' career development, and found evidence that a young researcher tends to have a more proficient career research productivity in the future if he/she has collaborated with outstanding scientists in his/her early career. The positive effect of outstanding scientists on young collaborators is actually becoming stronger over time.

Analyzing teamwork from more than 50 million papers, patents, and software products spanned from 60 years, Wu et al. [305] demonstrated across this period that larger teams developed recent, popular ideas, while small teams disrupted the system by drawing on older and less prevalent ideas. Attention to work from large teams came immediately, while advances by small teams succeeded further into the future. Differences between small and large teams magnify with impact. Small teams have become known for disruptive work and large teams for developing work. Differences in topic and research design account for part of the relationship between team size and disruption, but most of the effect occurs within people, controlling for detailed subject and article type.

Jia et al. [306] performed a large-scale analysis of extensive publication records to analyze research interest evolution of individual scientists and developed a random-walk-based model, allowing us to accurately reproduce the empirical observations that there is a high degree of regularity underlying scientific research and individual careers. Assuming two topic vectors based on the first and last m papers of the scientist (g_i and g_f respectively), capturing the research interest at the earliest and the latest stages of the career. Using the complementary cosine similarity between g_i and g_f, the interest change J of a scientist along the career is quantified as:

$$J = 1 - \frac{g_i g_f}{\|g_i\| \cdot \|g_f\|}. \tag{108}$$

The equation captures research interest change resulting from change of topics or from change of engagement in topics, providing an effective quantification on the extent of change. $J = 0$ indicates that the author studied the same set of topics at the two stages of the career. $J = 1$ corresponds to a complete interest change.

6.3. *Prediction in economic–social networks*

Complex networks can be usefully applied in the economic, although there is limited data available with which to develop our understanding. However concepts from statistical physics make it possible to reconstruct details of a economic network from partial sets of information [307]. In economic networks, nodes can represent firms, banks, or even countries, and where links between the nodes represent their mutual interactions, as ownership, credit–debt relationships or international trade. These evolving interactions can be represented by network dynamics that are bound in space and time and can change with the environment and also evolve with the nodes [308]. Recently, the field of economic complexity develop data-driven approaches to forecast the evolution of a dynamical economic system by the concept of the network representation. The network space built by the country database of exported products could condition the activity of the economic entity [309,310]. The recent advances have been substantiated to illustrate the correlation between economic growth and the country's productive network space. Specially, a country–product network $G(C, P, L)$ is constructed by the matrix M_{cp}, where C and P represent the set of countries and products, and L is trading interactions which exist only between countries and products [311]. By combining tools from network science, a robust and stable relationship between a country's productive structure and its economic growth has been established. Based on the concept of complex networks, we will review the recent advance of economic complexity.

6.3.1. *Diversity and ubiquity*

Diversity and ubiquity are respectively crude approximations of the variety of capabilities available in a country or required by a product. Diversity is related to the number of products that a country is connected to. Ubiquity is related to the number of countries that a product is connected to. Thus, ubiquity and diversity [309,312,313] are simply by summing over the rows or columns of that matrix M_{cp}. Formally,

$$k_c = \sum_p M_{cp}; \quad k_p = \sum_c M_{cp}. \tag{109}$$

6.3.2. *Eigenvector-based complexity index*

At first, the method of reflections consists of iteratively calculating the average value of the previous-level properties of a node's neighbors and is defined as the set of the observable [309,312]. The method of reflections as the recursive set of observable is defined as,

$$k_c^{(n)} = k_c \sum_p M_{cp} k_p^{(n-1)}; \quad k_p^{(n)} = k_p \sum_p M_{cp} k_c^{(n-1)}. \tag{110}$$

For $n \geq 1$, the $k_c^0 = k_c$ and $k_p^0 = k_p$. As Ref's analysis [314,315], the algorithm converges to constant value after approximate 16 steps. To generate a more accurate measure of the number of capabilities available in a country, or required by a product, Refs. [316,317] corrects the information that diversity and ubiquity carry by using each one to correct the other. For countries, this requires us to calculate the average ubiquity of the products that it exports, the average diversity of the countries that make those products and so forth. For products, this requires us to calculate the average diversity of the countries that make them and the average ubiquity of the other products that these countries make. One can use the matrix

$$M_{cc'} = M_{cp} diag(\frac{1}{k_p}) M_{cp}^T diag(\frac{1}{k_c}); \quad M_{pp'} = M_{cp}^T diag(\frac{1}{k_c}) M_{cp} diag(\frac{1}{k_p}). \tag{111}$$

The matrix $M_{cc'}$ connects country c with country c' according to the number of products that are exported by both. The matrix $M_{pp'}$ connects product p to product p'. The eigenvector e of the matrix $M_{cc'}$ and $M_{pp'}$ are associated with the second largest eigenvalue. Hausmann et al. [316,317] argued that the eigenvector with the second largest eigenvalue that captures the largest amount of variance in the system. Hence, the ECI is defined as:

$$ECI_c = \frac{\vec{K}_c - \left\langle \vec{K}_c \right\rangle}{s(\vec{K}_c)}; \quad ECI_p = \frac{\vec{K}_p - \left\langle \vec{K}_p \right\rangle}{s(\vec{K}_p)} \tag{112}$$

where, $\langle\rangle$ stands for the average value, $s()$ represents the standard deviation. $\vec{K}$ is eigenvector of the matrix $M_{cc'}$ and $M_{pp'}$ with the second largest eigenvalue. The non-monetary metrics not only reveals the status of global regional economic development, but also can be fit for predicting regional economic growth. Gao et al. [318,319] quantified the economic complexity of China's provinces through analyzing 25 years' firm data, and furthermore constructed the industry space by the data describing the evolution of China's economy between 1990 and 2015 which can capture the industrial diversification of Chinese provinces.

6.3.3. *Fitness and complexity index*

The basic idea is to define an iteration process which couples the fitness of a country to the complexity of a product and then obtain the fixed point values [314,315,320–327]. For the fitness, this is proportional to the sum of the products exported weighted by their complexity. For the fitness of a country the situation is more subtle. To a first approximation, the complexity of a product is inversely proportional to the number of countries which export it. This iterative method is composed of two steps in each iteration. First compute the intermediate variables,

$$\widetilde{F}_c^{(n)} = \sum_p M_{cp}\widetilde{Q}_p^{(n-1)}; \quad \widetilde{Q}_p^{(n)} = \frac{1}{\sum_c M_{cp}\frac{1}{\widetilde{F}_c^{(n-1)}}}. \tag{113}$$

At each step the fitness and complexity are normalized by their average value.

$$F_c^{(n)} = \frac{\widetilde{F}_c^{(n)}}{\left\langle \widetilde{F}_c^{(n)} \right\rangle}; \quad Q_p^{(n)} = \frac{\widetilde{Q}_p^{(n)}}{\left\langle \widetilde{Q}_p^{(n)} \right\rangle}. \tag{114}$$

The initial conditions are $F_c^{(0)} = 1$ for each country c, $Q_p^{(0)} = 1$ for each product p. Pugliese et al. [328] and Wu et al. [329] investigated the convergence properties of the algorithm. Mariani et al. [330] compared two metrics, Fitness–Complexity and the method of reflections, and furthermore proposed a generalization of the Fitness–Complexity metric. In addition, fitness and complexity index has been successfully applied to the cases including India [331] and Netherlands [332]. Besides, Stojkoski et al. [333] reported that not only goods but also services are important for predicting the rate at which countries will grow. Cimini [334] characterized the scientific fitness of each nation that is, the competitiveness of its research system and the complexity of each scientific domain by the fitness and complexity index can be able to assess quantitatively the advantage of scientific diversification.

6.3.4. *Nestedness*

When extracting network topology from aggregated economic data, the architecture of trade network significantly exhibits a nested structure which is a statistical property of interaction data presented in matrix form. The nested structure has recently gathered much attention as a metric for characterizing ecological and economic systems [335]. In a perfectly nested matrix, the entries in each successive row are a strict subset of those in the previous row, while the entries in each successive column are a strict subset of those in the previous column. In ecological system, a nested pattern in mutualistic networks promotes biodiversity and preserves structural stability [336,337]. Nested patterns are also widely present in economic systems and suggested that the dynamics of nestedness could predict the evolution of industrial ecosystems [338,339].

Firstly, we introduce one of most popular nestedness measurements namely nested overlap and decreasing fill (NODF) [340–342]. In the beginning of calculation, the rows and the columns of a matrix are swapped and rank-ordered by the sum of the presences in each of these rows and columns respectively. The transformed matrices are then ready to be processed by the flowing equation

$$NODF_{ij} = \sum_{i<j} \begin{cases} 0 & \text{if } k_i = k_j; \\ \dfrac{\sum_i M_{ij}M_{ik}}{\min(k_i, k_j)} & \text{otherwise.} \end{cases} \tag{115}$$

$$\eta = \frac{\sum_{i<j}(NODF_{ij}^{row} + NODF_{ij}^{column})}{\frac{n(n-1)}{2} + \frac{m(m-1)}{2}}; \tag{116}$$

Here k is the sum of the presences in each of the row and column respectively, i.e. degree. The NODF measure takes values between 0 (unnested) and 1 (perfectly nested). Fig. 12 gives a example of calculation of nestedness.

The international trading system is mainly modeled with a single network in the previous works, such as the monopartite product space network and the bipartite country–product network. Economic complexity embedded in a dynamic interaction of a large number of different agents can reflect the development of nations. The limitations have constrained related research to capture the rich process resulting from global effects, not the effects of individual economic agents and their interactions.

In order to better capture the more detailed dynamics, Ren et al. [343] characterized the international trading system with a multi-layer network with each layer representing the transnational trading relations of a product. This framework

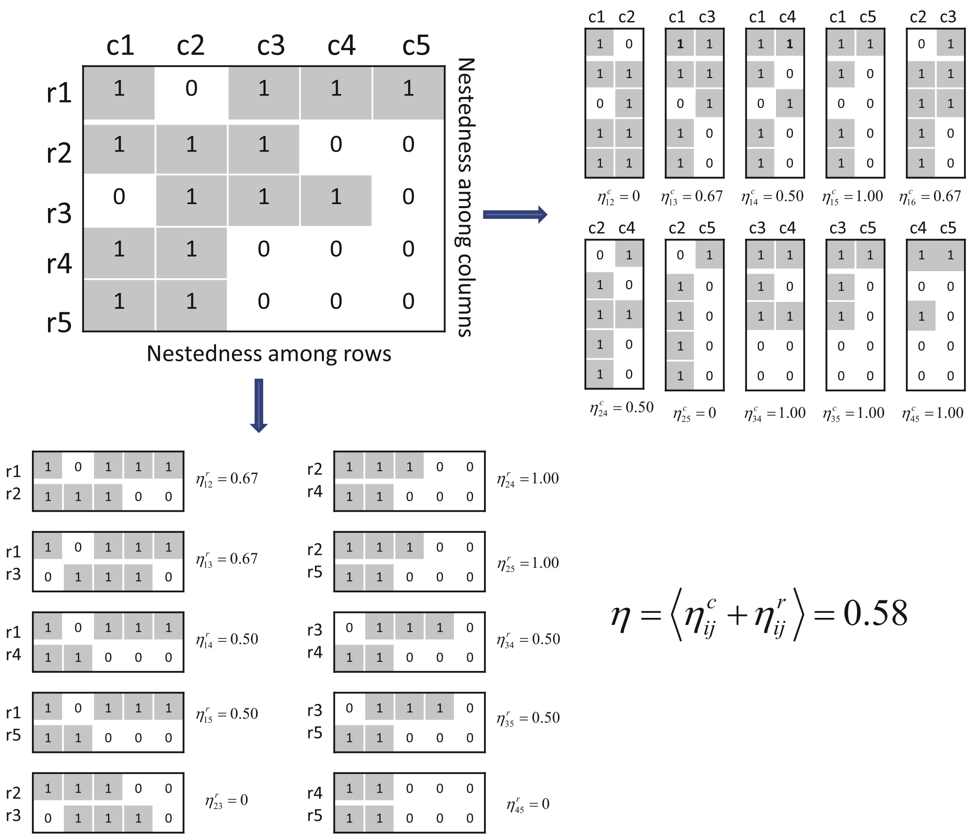

Fig. 12. The calculation of nestedness. After [340].

immediately reveals the nested structure in each layer and accordingly develop an accurate and robust measure of the complexity of products. The nestedness in each layer of the international trading multilayer network with the NODF metric and report their distribution as shown in Fig. 13. One can see that the nestedness of layers varies significantly. For instance, the nestedness of Bovine, Motorcycles, and Medical Instruments is 0.12, 0.37 and 0.56 respectively, which locates in the low, medium and high areas of nestedness distribution. Furthermore, nestedness gives a reasonable ranking of product complexity. Specifically, the sectors (8, 7, 5, 6) include machinery and chemicals have high nestedness value, while the sectors (3, 4, 2, 0, 1) of raw materials (such as beverages and tobacco, food, crude inedible materials, mineral fuels, lubricants, animal and vegetable oils, fats and waxes) have lower nestedness value. Further, the nestedness can be extended to measure the competitiveness of countries and the resultant metric has a higher ability than the existing metrics in predicting a country's future economics.

6.3.5. *Data driven methods*

The massive amount of the data of the individual activity available from the online services also provides an unprecedented opportunity to understand the economic development. Choi et al. [344] used search engine data to forecast near-term values of economic indicators. Examples include automobile sales, unemployment claims, travel destination planning and consumer confidence. Bollen et al. [250] showed large-scale collections of daily Twitter posts can be used to predict the stock market. Mestyan et al. [248] built a minimalistic predictive model for the financial success of movies based on collective activity data of online users. Blumenstock et al. [345] mapped mobile phone metadata and results showed that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Motived by this, the distribution of wealth of an entire nation can accurately be calculated by the predicted attributes of millions of individuals. Analyzing the travel patterns of 500,000 individuals in Cote d'Ivoire using mobile phone call data records, Lu et al. [346] showed that individual trajectories of mobile phone is highly dependent on historical behaviors, and that the maximum predictability is not only a fundamental theoretical limit for potential predictive power, but also an approachable target for actual prediction accuracy.

Eagle et al. [347] analyzed the most complete record of a national communication network with national census data. These data make possible a population-level investigation of the relation between the structure of social networks and access to socioeconomic opportunity. They developed two new metrics to capture the social and spatial diversity of communication

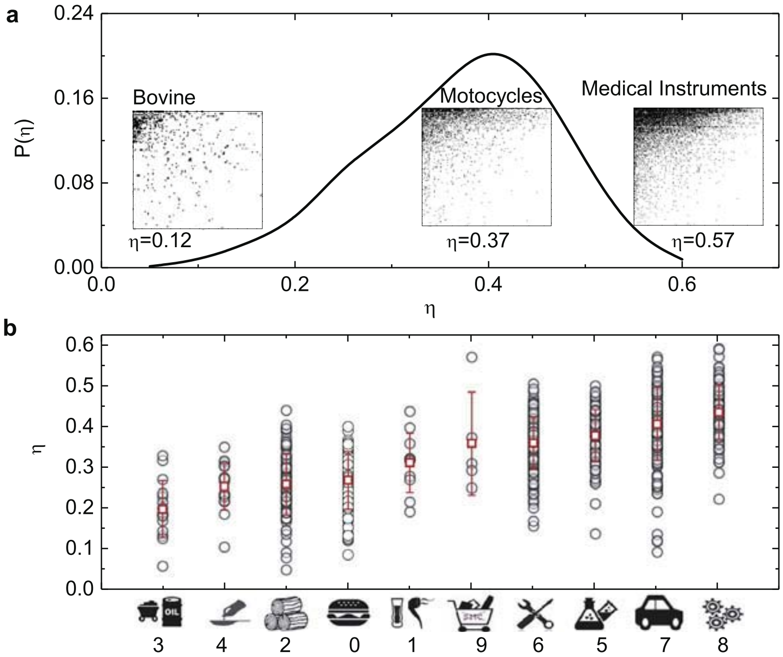

Fig. 13. (a) Three representations of matrices which are corresponding to the world trade networks of Bovine, Motorcycles, and Medical Instruments. The presences represent trade interaction between two countries. (b) Average nestedness of 786 products. The 786 products are classified into 10 sectors by Standard International Trade Classification (SITC). For each sector one can compute $\langle\eta(s)\rangle = \sum_{i\in s}\eta(i)/N(i \in s)$ where i is a product of sector s, $\eta(i)$ is nestedness value of product i, $N(i \in s)$ is the number of products belonging to the sector s. The black points correspond to products belonging to the according sections. The red points are the mean nestedness value of products belonging to sections and the red lines are the error bar. After [343].

interactions within an individual's social network and found that the diversity of individuals' relationships is strongly correlated with the economic development of communities. The topological diversity is defined as a function of the Shannon entropy,

$$H(i) = -\sum_{j=1}^{k} p_{ij}\log(p_{ij}), \tag{117}$$

where k is the number of i's contacts and p_{ij} is the proportion of i's total call volume that involves j. Then the social diversity $D_{social}(i)$ is defined as the Shannon entropy associated with individual i's communication behavior, normalized by k:

$$D_{social} = \frac{H(i)}{\log(k)}. \tag{118}$$

The above measure of topological diversity does not take into account the geographic diversity in the calling patterns within a community. A similar measure is defined as spatial diversity, $D_{spatial}(i)$, by replacing call volume with the geographic distance spanned by an individual's ties to the 1992 telephone exchange areas in the UK. High diversity scores imply that an individual splits her time more evenly among social ties and between different regions. The relationship between social network diversity and socioeconomic rank as shown in Fig. 14, which suggest that diversity of individual's relation is strongly correlated with the economic development of regions.

6.4. Prediction technical–social networks

We live in an increasingly interconnected world of techno-social systems, in which infrastructures composed of different technological layers are interoperating within the social component that drives their use and development. People check their e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Individual movements in public places may be captured by video cameras, and medical records stored as digital files. People may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these digital traces on the Internet that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies [348,349]. Modern techno-social systems consist of large-scale physical infrastructures (such as transportation systems and power distribution grids) embedded in a dense web of communication and computing infrastructures whose dynamics and evolution which are defined and driven by the human behavior [126].

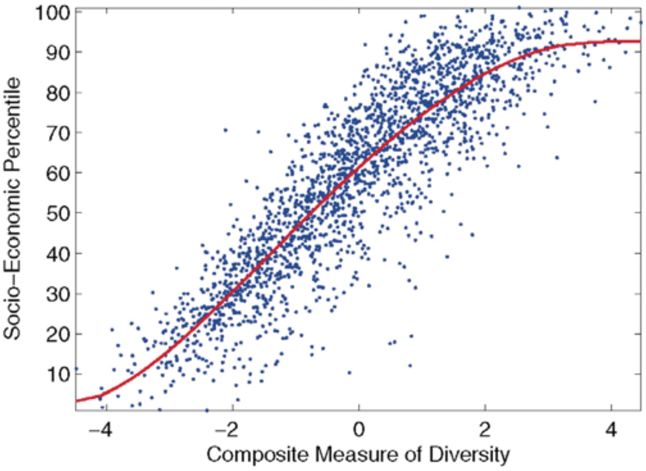

Fig. 14. The relationship between social network diversity and socioeconomic rank. Socioeconomic rank is based on the 2004 UK government's Index of Multiple Deprivation (IMD). After [347].

To predict the behavior of such systems, it is necessary to start with the mathematical description of patterns found in real-world data. Lim et al. [350] purported to predict the location of ethnic violence in the former Yugoslavia with a complex agent-based model. Brockmann et al. [351] showed that popular Web sites for currency tracking collect a massive number of records on money dispersal that can be used as a proxy for human mobility. This work opened a new way to the general exploitation of proxy data for human interaction and mobility [352]. Further, Brockmann et al. [353] analyzed disease spread via the "effective distance" rather than geographical distance, wherein two locations that are connected by a strong link are effectively close. The approach was successfully applied to predict disease arrival times or disease source [354] and presented a stochastic computational framework for the forecast of global epidemics that considers the complete worldwide air travel infrastructure complemented with census population data. Analogously, modern mobile phones and personal digital assistants combine sophisticated technologies such as Bluetooth, Global Positioning System, and WiFi, constantly producing detailed traces on our daily activities. For instance, by applying an information-theoretic method to the spatiotemporal data of cell-phone locations, Song et al. [355] found that human mobility patterns are remarkably predictable. Inspired by this, Takaguchi et al. [356] addressed a similar predictability question in a different kind of human social activity: conversation events. The predictability in the sequence of one's conversation partners is defined as the degree to which one's next conversation partner can be predicted when given the current partner. Gonzalez et al. [357] used mobile phone data to track the movements of 100,000 people over a 6-month time span. Furthermore, it is now possible to use sensors and tags that produce data at the microscale of one-to-one interactions [348,349]. Another example is using Twitter to predict electoral outcomes [358], however with its biases and limitations [359,360]. Interesting studies have appeared treating the use of social media indicators to predict the scientific impact of research articles from the pre-print sharing web [361] and Twitter mentions [291].

Predicting the behavior of a complex system requires a joint quantitative description of the system's structure and dynamics. Barzel et al. [362] bridged topology and dynamics, predicting that a complex system's response to perturbations is driven by a small number of universal characteristics. Nonlinear prediction as a way of distinguishing chaos from random fractal sequences [363]. An extremely challenging problem of significant interest is to predict catastrophes in advance of their occurrences. Boers et al. [364] introduced the concept of network divergence on directed networks derived from a non-linear synchronization measurement to predict extreme events. Wang et al. [365,366] presented a general approach to predicting catastrophes in nonlinear dynamical systems under the assumption that the system equations are completely unknown and only time series reflecting the evolution of the dynamical variables of the system are available. In the review [91], the recent advances focused on compressive sensing was surveyed a diverse array of problems based reconstruction of nonlinear and complex networked systems such as prediction of catastrophic bifurcations, forecasting future attractors of time-varying nonlinear systems, detection of hidden nodes.

7. Predictability and feedback

Complex networks have already become a ubiquitous way of representing many real systems in which the pattern of interactions between a system's components is itself complex. There is a wave of sensational discovery in complex networks, aiming at not only exploring the structural interaction of the network itself but also predicting the dynamics taking place on the network. A range of application, from biology, society, economics to technical–social systems, depend on our ability to foresee evolution and dynamics of the systems, raising a fundamental question: To what degree is dynamics of complex systems predictable? Here we survey the limits of predictability in complex systems by studying the structure of complex networks.

7.1. Predictability in human behaviors

What is the role of randomness in human behavior and to what degree are individual human actions predictable? Song et al. [355] quantified the interplay between the regular and thus predictable and the random and thus unforeseeable, probing through human mobility the fundamental limits that characterize the predictability of human dynamics. By measuring the entropy of each individual's trajectory, and found a 93% potential predictability in user mobility across the whole user base. To evaluate the predictability, the entropy accounts for both the relative frequency of the visited locations and the order of the visits.

$$S_i = -\sum_{T_i' \subset T_i} P(T_i')\log_2[P(T_i')], \tag{119}$$

where $P(T_i')$ is the probability of finding a subsequence T_i' in T_i. Based on this measure of entropy, one can estimate the upper bound of the success rate in predicting the future location of the mobile phone user immediately after T_i. The maximum predictability Π is defined as a limiting case of Fano's inequality (a relation derived from calculation of the decrease in a noisy information channel). That is, if a user with entropy S moves between N locations, then the predictability is,

$$\Pi \leq \Pi^{\max}(S, N), \tag{120}$$

where $\Pi^{\max}$ is given by,

$$S = H(\Pi^{\max}) + (1 - \Pi^{\max})\log_2(N - 1), \tag{121}$$

with the binary entropy function

$$H(\Pi^{\max}) = -\Pi^{\max}\log_2(\Pi^{\max}) - (1 - \Pi^{\max})\log_2(1 - \Pi^{\max}). \tag{122}$$

Take an example, a user with $\Pi^{\max} = 0.2$, this means that at least 80% of the time the individual chooses his location in a manner that appears to be random, and only in the remaining 20% of the time can we hope to predict his whereabouts. In other terms, no matter how good our predictive algorithm, we cannot predict with better than 20% accuracy the future whereabouts of a user with $\Pi^{\max} = 0.2$. Therefore, $\Pi^{\max} = 0.2$ represents the fundamental limit for each individual's predictability.

Inspired by this work, Takaguchi et al. [356] addressed a similar predictability question in a different kind of human social activity: conversation events. The predictability in the sequence of one's conversation partners is defined as the degree to which one's next conversation partner can be predicted a given current partner. The predictability of conversation events for each individual is based on the longitudinal data of face-to-face interactions collected from two company offices in Japan. The conversation events are predictable to a certain extent: knowing the current partner decreases the uncertainty about the next partner by 28.4% on average. Much of the predictability is explained by long-tailed distributions of inter-event intervals. However, a predictability also exists in the data, apart from the contribution of their long-tailed nature. In addition, an individual's predictability is correlated with the position of the individual in the static social network derived from the data. Individuals confined in a community – in the sense of an abundance of surrounding triangles – tend to have low predictability, and those bridging different communities tend to have high predictability. The above previously studies have shown that human movement is predictable to a certain extent at different geographic scales. The existing prediction techniques exploit only the past history of the person taken into consideration as input of the predictors. De et al. [367] showed that by means of multivariate nonlinear time series prediction techniques it is possible to increase the predictability by considering movements of friends, people, or more in general entities, or characterized by high mutual information as inputs. Using this framework, Sekara et al. [368] explored the complex interplay between social and geospatial behavior, documenting how the formation of cores is preceded by coordination behavior in the communication networks and demonstrating that social behavior can increase the predictability.

Our daily social-media experience seemingly ordinary items like videos, news, publications unexpectedly gain an enormous amount of attention. Miotto et al. [369] proposed a method that, give some information on the items, and then quantifies the predictability of events, i.e., the potential of identifying in advance the most successful items. Applying this method to different data, ranging from views in YouTube videos to posts in Usenet discussion groups, one can invariantly find that the predictability increases for the most extreme events. This is done by formulating a simple prediction problem which allows for the computation of the optimal prediction strategy. The problem is limited to provide a binary (yes/no) prediction whether an item will be an extreme event or not (attention passes a given threshold). Predictability is then quantified as the quality of the optimal strategy. Predictions are based on information on items which generally lead to a partition of the items in groups $g \in \{1, \ldots, G\}$ that have the same feature. Since the membership to a group g is the only thing that characterizes an item, predictive strategies can only be based on the probability of having E for that group $P(E|g)$. Therefore, one can use the quality of prediction of the optimal strategy to quantify the predictability Π (i.e., the potential prediction) of the system for the given problem and information. By geometrical arguments one can obtain from

$$\Pi = \sum_{g} \sum_{h<g} \frac{P(g)P(h)(P(E|h) - P(E|g))}{P(E)(1 - P(E))}, \tag{123}$$

where $p(g)$ is the probability of group g and g is ordered by decreasing $P(E|g)$, i.e., $h < g \Rightarrow P(E|h) > P(E|g)$.

Miotto et al. [369] applied this method to four different systems: views of YouTube videos, comments in threads of Usenet discussion groups, votes to Stack-Overflow questions. The empirical finding is that in all cases the predictability increases for more extreme events (increasing threshold). This finding can extend to earthquake, population movements following large-scale disasters may be significantly more predictable than previously thought [370]. In addition, Guimera et al. [371] investigated to what extent it is possible to make predictions of a justice's vote based on the other justices' votes in the same case and found that justices are significantly more predictable than one would expect from an ideal situation in which justice decisions are uncorrelated. The predictability of a justice with respect to the predictability in an equivalent ideal court provides a quantitative proxy for stable justice correlations, which ultimately reflect a priori attitudes toward the law. Penner et al. [301] analyzed a large set of careers distributed across 3 disciplines including physics, biology and mathematics, and found that although future measures of impact are correlated with past measures, the current state of the art models simply do not do a good enough job of predicting future impact to be used with confidence in the career advancement decision process. People need to not only understand the success and attrition rates of scientific careers, but also, it is critical to grasp the limits-of-prediction.

7.2. Predictability in economic complexity

As introduced in section (Applications of prediction), Economic complexity has provided new perspectives to cast economic prediction into the conceptual scheme of forecasting the evolution of a dynamical system. Cristelli et al. [315] argued that a recently introduced non-monetary metrics for country competitiveness (fitness) allows for quantifying the hidden growth potential of countries. This comparison defines the fitness–income plane where one can observe that country dynamics presents strongly heterogeneous patterns of evolution as shown in Fig. 15. The flow in some zones is found to be laminar while in others a chaotic behavior is instead observed. These two regimes correspond to very different predictability features for the evolution of countries: in the former regime, the strong predictable pattern while the latter scenario exhibits a very low predictability. The usual tool regressions used in economics are no more the appropriate strategy to deal with such a heterogeneous scenario and new concepts, borrowed from dynamical systems theory are mandatory. Therefore a data-driven method is proposed as the selective predictability scheme in which the degree of predictability of the economic dynamics depends on the specific position in the income–fitness plane.

7.3. Predictability in nonlinear dynamics

Nonlinear forecasting has recently been shown to distinguish between deterministic chaos and uncorrelated (white) noise added to periodic signals, and can be used to estimate the degree of chaos in the underlying dynamical system. The correlation between predicted and actual values measured may decrease with time i.e. a property synonymous with chaos. Tsonis et al. [363] showed that by determining the scaling properties of the prediction error as a function of time, and used nonlinear prediction to distinguish between chaos and random fractal sequences. Dynamical interdependence or generalized synchrony implies predictability, and such predictions were described in [372] which uses local polynomial maps of a driving system to predict the behavior of a unidirectionally coupled chaotic response system. Unfortunately the application of these ideas to experimental systems with arbitrary coupling is not straightforward. By defining the predictability of each system based on a knowledge of the other system, Schiff et al. [373] derived a measure of dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. Boers et al. [364] introduced the concept of network divergence, which is based on the non-linear synchronization measure event synchronization and complex network theory to forecast spatially extensive extreme rainfall in the eastern Central Andes. The measure network divergence introduced here is designed to assess the predictability of extreme events in significantly interrelated time series. Methods for forecasting time series are a critical aspect of the understanding and control of complex networks. When the model of the network is unknown, nonparametric methods for prediction have been developed. Hamilton et al. [374] considered how to make use of a subset of the system equations, if they are known, to improve the predictive capability of forecasting methods.

In addition, prediction of physical and functional links between cellular components are often based on correlations between experimental measurements, such as gene expression. However, correlations are affected by both direct and indirect paths, confounding our ability to identify true pairwise interactions. Barzel et al. [283] exploited the fundamental properties of dynamical correlations in networks to develop a method to silence indirect effects. The method enhanced the discriminative power of the correlations by twofold predictive improvement over traditional correlation measures. Extend to ecosystems, ecosystems are subjected to chronic disturbances, such as harvest, pollution, and climate change. The capacity to forecast how species respond to such press perturbations is limited by our imprecise knowledge of pairwise species interaction strengths and many direct and indirect pathways along which perturbations can propagate between species. Network complexity (size and connectance) has thereby been seen to limit the predictability of ecological

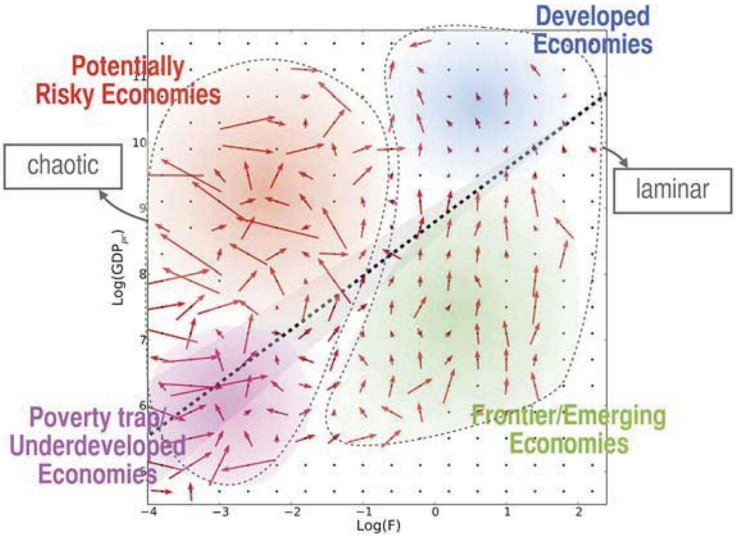

Fig. 15. The dynamics of the evolution of countries in the fitness–income plane. A coarse graining of the dynamics highlights two regimes for the dynamics of the evolution of countries in the fitness–income plane. There exists a laminar region in which fitness is the driving force of the growth and the only relevant economic variable in order to characterize the dynamics of countries. The evolution of countries in this region is highly predictable. There is also a second regime, which appears to be chaotic and characterized by a low level of predictability. After [315].

systems. Iles et al. [375] demonstrated a counteracting mechanism in which the influence of indirect effects declines with increasing network complexity when species interactions are governed by universal allometric constraints. These observations corroborate the feasibility of systematic experimental implementation of synthetic rescues. Indeed, the main difficulties expected in verifying our predictions, namely the inaccuracies in matching real genetic and environmental conditions as well as potential side effects of rescue deletions due to, e.g. unknown function, are substantially alleviated by the robustness and flexibility of the rescue interactions [282].

7.4. *Predictability in epidemics*

The epidemic pattern predictability is quantitatively determined and traced back to the occurrence of epidemic pathways defining a backbone of dominant connections for the disease spreading. Colizza et al. [376] thought the predictability of the epidemic evolution with respect to the inherent stochastic dynamics of the disease transmission. To address predictability of epidemics in complex networks, Colizza et al. [354] devised a set of quantitative measures able to characterize the level of predictability of the epidemic pattern. It is clear that to make the forecast more realistic, it is necessary to introduce more details in the disease dynamics. Loecher et al. [377] enhanced predictability of epidemic outbreaks in scale-free networks by replacing the node degree with the random walk centrality. Tsonis et al. [378] constructed the networks of the surface temperature field for El Niño and for La Niña years and found that the El Niño network possesses significantly fewer links and lower clustering coefficient and characteristic path length than the La Niña network, which indicates that the former network is less communicative and less stable than the latter. The underlying cause is that predictability of temperature should decrease during El Niño years.

8. Summary and outlook

Prediction in science is quantitative forecast of a system's future development under specific conditions. For numerous real systems such as ecosystems, airlines and social and economic organizations, they are composed of a large number of components with complex and evolving interact with each other. Networks are an effective tool describing the interaction structure in these complex systems. Thus, the prediction problem in many real systems can be formulated as predictions in complex networks, ranging from prediction of links at microscopic level to prediction of collective trend at macroscopic level. The rapid development of information technology brings in the big data era, providing an unprecedented opportunity to investigate real complex systems with high quality empirical data. In this context, there is a recent trend of developing prediction methods for different real networks.

In this review, we summarize recent major progress regarding the prediction problems in complex networks, aiming to cover issues in both structural and dynamical aspects. We first introduce the general framework for investigating prediction problems in complex networks. The works attempting to predict the microscopic properties of networks are classified as node-oriented predictions and link-oriented predictions. Specifically, we review methods for predicting nodes' future popularity in growing networks, as well as methods for predicting future links in networks. For macroscopic prediction, the works on predicting the evolution of community structure and some other topological metrics are reviewed. We also discuss several key applications of the prediction methods in real systems, including biology networks, scientific networks, economic–social networks and technical–social networks. Finally, we review the effort in understanding the predictability of different systems.

Despite considerable efforts, numerous issues in network prediction remain challenging. Together with the development of the theoretical works, many prediction methods have been used in practice. For instance, the link prediction methods have been widely used in online recommendation by guessing the future connections between users and products, reconstructing biological and social networks by adding missing links and removing spurious links. However, we remark that one has to be careful in selecting methods. In fact, the prediction will have some feedback effect on the system. If the prediction methods are not properly adopted, the system may evolve or be reconstructed into a distorted state. The recommendation methods, for example, will guide users' selection. If the recommendation methods that mainly recommend popular items are iteratively used, the attention of most online users will be guided to a limited number of products, forming an undesired structure for online retailers. When link prediction methods are used in network reconstruction, some missing links may be added to the network by mistake. It is possible that these links will significantly alter the structural properties of the networks, leading to a misunderstanding of the real networks. Therefore, when selecting prediction methods, one has to consider multi-dimension evaluation. Instead of focusing only on the prediction accuracy, one has to pay attention to its feedback effect, so as to avoid critical mistakes.

In the literature dealing with time series prediction, a large part of the prediction methods are based on regression or machine learning approaches. Though some methods are very high in accuracy, the underlying reasons for the high performance are mostly unclear. Recently, there is a trend of establishing mechanistic models for prediction. The general approach is to first understand the key factors as well as the equation driving the evolution of the system. These mechanistic models are usually tested with rescaling analysis. If the time series collapse into a single curve after the considered factors are scaled out, the mechanistic model already capture all the key factors driving the evolution of the system. Then the empirical data is used to fit parameters, which can be used for predicting future time series. So far, these mechanistic models are developed only in several systems such as scientific publication systems. The mechanistic models in other systems are remain to be developed.

Cold start is a classic problem in prediction. It means that the prediction cannot be very accurate if the available information is limited. This problem is particular important for link prediction(including also recommendation) problems in which the links of new nodes (users or items in the case of recommendation) are difficult to predict. For an online retailer, improving the prediction accuracy for new users will keep these users using a web site and attract more new users. Solution to cold start problem in link prediction problems are now mainly based on global similarity metrics or incorporating external information (e.g. multi-layer networks). Cold start problem is also a challenge in trend prediction. In citation networks, for instance, many methods for future citation prediction use 5 years citation history of a paper, which hinders its application for newly published papers. In general, more prediction methods for solving the cold start problems still need to be developed.

An emerging problem in prediction is predicting the dynamics taking place on networks. So far, the prediction of the spreading dynamics has attracted much attention. Many related issues such as prediction of future spreading coverage, prediction of future infected nodes, and the predictability of spreading have been investigated in the literature. The prediction of congestion in the transportation process on networks is also discussed from the perspective of traffic management. The network dynamics are actually many, such as percolation, synchronization and opinion formation. The prediction of the future evolution of these dynamics given the early dynamics evolution is observed is a very meaningful problem in both theory and application, which asks for future effort along this research direction.

In practice, once prediction is made, it will inevitably have feedback effect of the future evolution of the systems. In stock markets, if the price of a stock is predicted to be increasing, more people will buy this stock, which will result in a further increase of the stock price. In recommender system, the recommendation of a product to many users will lead to a huge increase of future selections of this product, which will further increase its probability of being recommended in the future. This feedback effect can have either positive or negative consequence to the systems, depending on the type of system itself. The increase of stock price might be favorable for those people who hold the stock, the overwhelming attention to a single product in online system might be harmful to the system as it dramatically destroy the diversity of the system. In the long run, what will happen is that all people will connect to a small number of products. Therefore, the feedback of prediction is also an important problem for investigation. This problem, however, is neglected in many prediction research. We remark here that more future effort is required to put in problem.

Predicting the future based on the current state or historical data is a crucial task in many applications. In this review article, we summarize recent progress in the field of prediction in complex networks. Though prediction techniques are sufficiently mature for some problem, reliable prediction approaches are still missing in many systems. 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. In the future work, if one can identify the fundamental mechanism despite the high stochastic in real systems, more effective prediction tools will be designed and the accurate predictions at both microscopic and macroscopic levels as well as for the dynamical processes taking place on the systems will be much more feasible. We believe that breakthroughs will soon appear in the future.

Acknowledgments

We acknowledge Dr. Manuel Sebastian Mariani for fruitful discussion. This work is partially supported by the National Natural Science Foundation of China (Grant No. 61603046), the Natural Science Foundation of Beijing (Grant No. L160008), Hangzhou Normal University Research Funding Project, Qiantang River Talents Plan (Class D projects).

References

[1] A.-L. Barabási, Network science: Luck or reason, Nature 489 (7417) (2012) 507–508.
[2] B.R. Jasny, R. Stone, Prediction and its limits, Science 355 (6324) (2017) 468–469.
[3] D.E. Acuna, S. Allesina, K.P. Kording, Future impact: Predicting scientific success, Nature 489 (7415) (2012) 201–202.
[4] D. Wang, C. Song, A.-L. Barabási, Quantifying long-term scientific impact, Science 342 (6154) (2013) 127–132.
[5] C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility, Nat. Phys. 6 (10) (2010) 818–823.
[6] F. Altarelli, A. Braunstein, L. Dall'Asta, R. Zecchina, Optimizing spread dynamics on graphs by message passing, J. Stat. Mech. Theory Exp. 2013 (09) (2013) P09011.
[7] Y. Hu, S. Ji, L. Feng, S. Havlin, Y. Jin, Optimizing locally the spread of influence in large scale online social networks, 2015. ArXiv preprint arXiv:1509.03484.
[8] J.M. Hofman, A. Sharma, D.J. Watts, Prediction and explanation in social systems, Science 355 (6324) (2017) 486–488.
[9] M. Kleiber, et al. The fire of life. An introduction to animal energetics, 1961.
[10] G. West, Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies, Penguin, 2017.
[11] P. Erdos, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci 5 (1) (1960) 17–60.
[12] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[13] P.L. Krapivsky, S. Redner, F. Leyvraz, Connectivity of growing random networks, Phys. Rev. Lett. 85 (21) (2000) 4629.
[14] S.N. Dorogovtsev, J.F.F. Mendes, Exactly solvable small-world network, Europhys. Lett. 50 (1) (2000) 1.
[15] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks with aging of sites, Phys. Rev. E 62 (2) (2000) 1842.
[16] G. Bianconi, A.-L. Barabási, Competition and multiscaling in evolving networks, Europhys. Lett. 54 (4) (2001) 436.
[17] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A 311 (3) (2002) 590–614.
[18] M. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks, Phys. Rev. Lett. 107 (23) (2011) 238701.
[19] Z.-M. Ren, Y.-Q. Shi, H. Liao, Characterizing popularity dynamics of online videos, Physica A 453 (2016) 236–241.
[20] R.M. Bond, C.J. Fariss, J.J. Jones, A.D. Kramer, C. Marlow, J.E. Settle, J.H. Fowler, A 61-million-person experiment in social influence and political mobilization, Nature 489 (7415) (2012) 295–298.
[21] C. Lynch, Big data: How do your data grow?, Nature 455 (7209) (2008) 28–29.
[22] M.J. Salganik, P.S. Dodds, D.J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market, Science 311 (5762) (2006) 854–856.
[23] M.-S. Shang, L. Lü, Y.-C. Zhang, T. Zhou, Empirical analysis of web-based user-object bipartite networks, Europhys. Lett. 90 (4) (2010) 48006.
[24] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proc. Natl. Acad. Sci. 105 (41) (2008) 15649–15653.
[25] F. Wu, B.A. Huberman, Novelty and collective attention, Proc. Natl. Acad. Sci. 104 (45) (2007) 17599–17601.
[26] A.-L. Barabási, The origin of bursts and heavy tails in human dynamics, Nature 435 (7039) (2005) 207–211.
[27] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, S. Moon, I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system, in: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC '07, ACM, New York, NY, USA, 2007, pp. 1–14.
[28] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, A. Mahanti, The untold story of the clones: Content-agnostic factors that impact youtube video popularity, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, ACM, New York, NY, USA, 2012, pp. 1186–1194.
[29] K. Lerman, R. Ghosh, Information contagion: An empirical study of the spread of news on digg and twitter social networks, ICWSM 10 (2010) 90–97.
[30] H.-W. Shen, D. Wang, C. Song, A.-L. Barabási, Modeling and predicting popularity dynamics via reinforced Poisson processes, 2014. ArXiv preprint arXiv:1401.0778.
[31] M. Vasconcelos, J. Almeida, M. Gonçalves, D. Souza, G. Gomes, Popularity dynamics of foursquare micro-reviews, in: Proceedings of the Second ACM Conference on Online Social Networks, COSN '14, ACM, New York, NY, USA, 2014, pp. 119–130.
[32] L. Chen, Y. Zhou, D.M. Chiu, A lifetime model of online video popularity, in: Computer Communication and Networks, ICCCN, 2014 23rd International Conference on, IEEE, 2014, pp. 1–8.
[33] S. Fortunato, A. Flammini, F. Menczer, Scale-free network growth by ranking, Phys. Rev. Lett. 96 (21) (2006) 218701.
[34] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, A. Vespignani, Characterizing and modeling the dynamics of online popularity, Phys. Rev. Lett. 105 (15) (2010) 158701.
[35] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, A. Mahanti, Characterizing and modelling popularity of user-generated videos, Perform. Eval. 68 (11) (2011) 1037–1055.
[36] Y.-H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, PLoS One 6 (9) (2011) e24926.
[37] M. Medo, Statistical validation of high-dimensional models of growing networks, Phys. Rev. E 89 (3) (2014) 032801.
[38] J. Liebig, A. Rao, Predicting item popularity: Analysing local clustering behaviour of users, Physica A 442 (2016) 523–531.
[39] M.H. Latif, H. Afzal, Prediction of movies popularity using machine learning techniques, Int. J. Comput. Sci. Netw. Secur. 16 (8) (2016) 127.
[40] G. Szabo, B.A. Huberman, Predicting the popularity of online content, Commun. ACM 53 (8) (2010) 80–88.
[41] A. Zeng, S. Gualdi, M. Medo, Y.-C. Zhang, Trend prediction in temporal bipartite networks: the case of movielens, netflix, and digg, Adv. Complex Syst. 16 (2013) 1350024.
[42] Y. Zhou, A. Zeng, W.-H. Wang, Temporal effects in trend prediction: identifying the most popular nodes in the future, PLoS One 10 (3) (2015) e0120735.
[43] A. Zeng, C.H. Yeung, Predicting the future trend of popularity by network diffusion, Chaos 26 (6) (2016) 063102.
[44] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation, Phys. Rev. E 76 (4) (2007) 046115.
[45] T. Zhou, L.-L. Jiang, R.-Q. Su, Y.-C. Zhang, Effect of initial configuration on network-based recommendation, Europhys. Lett. 81 (5) (2008) 58004.
[46] L. Lü, W. Liu, Information filtering via preferential diffusion, Phys. Rev. E 83 (6) (2011) 066119.
[47] F.-G. Zhang, A. Zeng, Information filtering via heterogeneous diffusion in online bipartite networks, PLoS One 10 (6) (2015) e0129459.
[48] S. Pei, H.A. Makse, Spreading dynamics in complex networks, J. Stat. Mech. Theory Exp. 2013 (12) (2013) P12002.

[49] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, Phys. Rep. 650 (2016) 1–63.
[50] H. Liao, M.S. Mariani, M. Medo, Y.-C. Zhang, M.-Y. Zhou, Ranking in evolving complex networks, Phys. Rep. (2017).
[51] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (4) (1966) 581–603.
[52] K.-I. Goh, E. Oh, B. Kahng, D. Kim, Betweenness centrality correlation in social networks, Phys. Rev. E 67 (1) (2003) 017101.
[53] S.P. Borgatti, Centrality and network flow, Social Networks 27 (1) (2005) 55–71.
[54] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (11) (2010) 888–893.
[55] J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg, Structural diversity in social contagion, Proc. Natl. Acad. Sci. 109 (16) (2012) 5962–5966.
[56] C. Dangalchev, Residual closeness in networks, Physica A 365 (2) (2006) 556–564.
[57] J. Zhang, X.-K. Xu, P. Li, K. Zhang, M. Small, Node importance for dynamical process on networks: A multiscale characterization, Chaos 21 (1) (2011) 016107.
[58] A. Zeng, C.-J. Zhang, Ranking spreaders by decomposing complex networks, Phys. Lett. A 377 (14) (2013) 1031–1035.
[59] J.-G. Liu, Z.-M. Ren, Q. Guo, Ranking the spreading influence in complex networks, Physica A 392 (18) (2013) 4154–4159.
[60] R. Poulin, M.-C. Boily, B.R. Mâsse, Dynamical systems to define centrality in social networks, Social Networks 22 (3) (2000) 187–220.
[61] D.-B. Chen, R. Xiao, A. Zeng, Y.-C. Zhang, Path diversity improves the identification of influential spreaders, Europhys. Lett. 104 (6) (2014) 68006.
[62] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PLoS One 6 (6) (2011) e21202.
[63] G. Ghoshal, A.-L. Barabási, Ranking stability and super-stable nodes in complex networks, Nat. Commun. 2 (2011) 394.
[64] R.A.P. Da Silva, M.P. Viana, L. da Fontoura Costa, Predicting epidemic outbreak from individual features of the spreaders, J. Stat. Mech. Theory Exp. 2012 (07) (2012) P07005.
[65] M.J. Keeling, P. Rohani, Modeling Infectious Diseases in Humans and Animals, Princeton University Press, 2008.
[66] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.
[67] P. Holme, B.J. Kim, Growing scale-free networks with tunable clustering, Phys. Rev. E 65 (2) (2002) 026107.
[68] Q. Guo, T. Zhou, J.-G. Liu, W.-J. Bai, B.-H. Wang, M. Zhao, Growing scale-free small-world networks with tunable assortative coefficient, Physica A 371 (2) (2006) 814–822.
[69] H. Kim, J. Tang, R. Anderson, C. Mascolo, Centrality prediction in dynamic human contact networks, Comput. Netw. 56 (3) (2012) 983–996.
[70] M. Ercsey-Ravasz, Z. Toroczkai, Centrality scaling in large networks, Phys. Rev. Lett. 105 (3) (2010) 038701.
[71] M. Ercsey-Ravasz, R.N. Lichtenwalter, N.V. Chawla, Z. Toroczkai, Range-limited centrality measures in complex networks, Phys. Rev. E 85 (6) (2012) 066103.
[72] Y. Sun, L. Ma, A. Zeng, W.-X. Wang, Spreading to localized targets in complex networks, Sci. Rep. 6 (2016).
[73] S. Brin, L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, Comput. Netw. 56 (18) (2012) 3825–3833.
[74] D. Walker, H. Xie, K.-K. Yan, S. Maslov, Ranking scientific publications using a model of network traffic, J. Stat. Mech. 2007 (06) (2007) P06010.
[75] M. Wasserman, X.H.T. Zeng, L.A.N. Amaral, Cross-evaluation of metrics to estimate the significance of creative works, Proc. Natl. Acad. Sci. 112 (5) (2015) 1281–1286.
[76] M.S. Mariani, M. Medo, Y.-C. Zhang, Ranking nodes in growing networks: When pagerank fails, Sci. Rep. 5 (2015).
[77] M.S. Mariani, M. Medo, Y.-C. Zhang, Identification of milestone papers through time-balanced network centrality, J. Inform. 10 (4) (2016) 1207–1223.
[78] http://journals.aps.org/prl/50years/milestones.
[79] M. Medo, M.S. Mariani, A. Zeng, Y.-C. Zhang, Identification and impact of discoverers in online social systems, Sci. Rep. 6 (2016) 34218.
[80] Z.-M. Ren, M.S. Mariani, Y.-C. Zhang, M. Medo, A time-respecting null model to explore the structure of growing networks, 2017. ArXiv preprint arXiv:1703.07656.
[81] P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with googles pagerank algorithm, J. Inform. 1 (1) (2007) 8–15.
[82] M. Newman, The first-mover advantage in scientific publication, Europhys. Lett. 86 (6) (2009) 68001.
[83] R.A. Rossi, L.K. McDowell, D.W. Aha, J. Neville, Transforming graph data for statistical relational learning, J. Artificial Intelligence Res. 45 (2012) 363–441.
[84] B. Bringmann, M. Berlingerio, F. Bonchi, A. Gionis, Learning and predicting the evolution of social networks, IEEE Intell. Syst. 25 (4) (2010) 26–35.
[85] R.-Q. Su, W.-X. Wang, Y.-C. Lai, Detecting hidden nodes in complex networks from time series, Phys. Rev. E 85 (6) (2012) 065201.
[86] R.-Q. Su, Y.-C. Lai, X. Wang, Y. Do, Uncovering hidden nodes in complex networks in the presence of noise, Sci. Rep. 4 (2014).
[87] Z. Shen, W.-X. Wang, Y. Fan, Z. Di, Y.-C. Lai, Reconstructing propagation networks with natural diversity and identifying hidden sources, Nat. Commun. 5 (2014).
[88] Z. Shen, S. Cao, W.-X. Wang, Z. Di, H.E. Stanley, Locating the source of diffusion in complex networks by time-reversal backward spreading, Phys. Rev. E 93 (3) (2016) 032301.
[89] X. Han, Z. Shen, W.-X. Wang, Z. Di, Robust reconstruction of complex networks from sparse data, Phys. Rev. Lett. 114 (2) (2015) 028701.
[90] Z.-L. Hu, X. Han, Y.-C. Lai, W.-X. Wang, Optimal localization of diffusion sources in complex networks, R. Soc. Open Sci. 4 (4) (2017) 170091.
[91] W.-X. Wang, Y.-C. Lai, C. Grebogi, Data based identification and prediction of nonlinear and complex dynamical systems, Phys. Rep. 644 (2016) 1–76.
[92] L. Peel, Active discovery of network roles for predicting the classes of network nodes, J. Complex Netw. 3 (3) (2014) 431–449.
[93] D. Hric, T.P. Peixoto, S. Fortunato, Network structure, metadata, and the prediction of missing nodes and annotations, Phys. Rev. X 6 (3) (2016) 031038.
[94] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.
[95] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A.A. Nanavati, A. Joshi, Social ties and their relevance to churn in mobile telecom networks, in: Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology, ACM, 2008, pp. 668–677.
[96] J. Hopcroft, T. Lou, J. Tang, Who will follow you back?: reciprocal relationship prediction, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 1137–1146.
[97] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. 106 (52) (2009) 22073–22078.
[98] M. Al Hasan, M.J. Zaki, A survey of link prediction in social networks, in: Social Network Data Analytics, Springer, 2011, pp. 243–275.
[99] F. Gao, K. Musial, C. Cooper, S. Tsoka, Link prediction methods and their accuracy for different social networks and network metrics, Sci. Program. 2015 (2015) 1.
[100] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Assoc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.
[101] G.G. Chowdhury, Introduction to Modern Information Retrieval, Facet Publishing, 2010.
[102] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Netw. 25 (3) (2003) 211–230.
[103] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.
[104] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1–34.
[105] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.

[106] E.A. Leicht, P. Holme, M.E. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2) (2006) 026120.
[107] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin, Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, Phys. Rev. E 75 (2) (2007) 021102.
[108] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (4) (2009) 623–630.
[109] M. Al Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, in: SDM06: Workshop on Link Analysis, Counter-Terrorism and Security, 2006.
[110] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[111] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (4) (2009) 046122.
[112] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, IEEE Trans. Knowl. Data Eng. 19 (3) (2007) 355–369.
[113] D.M. Dunlavy, T.G. Kolda, E. Acar, Temporal link prediction using matrix and tensor factorizations, ACM Trans. Knowl. Discov. Data 5 (2) (2011) 10.
[114] A.K. Menon, C. Elkan, Link prediction via matrix factorization, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 437–452.
[115] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, H.E. Stanley, Toward link predictability of complex networks, Proc. Natl. Acad. Sci. 112 (8) (2015) 2325–2330.
[116] L. Pan, T. Zhou, L. Lü, C.-K. Hu, Predicting missing links and identifying spurious links via likelihood analysis, Sci. Rep. 6 (2016).
[117] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Mach. Learn. 20 (3) (1995) 197–243.
[118] D. Heckerman, C. Meek, D. Koller, Probabilistic entityrelationship models, in: PRMs and Plate Models. In SRL Workshop, ICML, 2004.
[119] K. Yu, W. Chu, S. Yu, V. Tresp, Z. Xu, Stochastic relational models for discriminative link prediction, in: Advances in Neural Information Processing Systems, 2007, pp. 1553–1560.
[120] H. Kashima, N. Abe, A parameterized probabilistic model of network evolution for supervised link prediction, in: Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE, 2006, pp. 340–349.
[121] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 2014.
[122] C. Wang, V. Satuluri, S. Parthasarathy, Local probabilistic models for link prediction, in: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE, 2007, pp. 322–331.
[123] A. Clauset, C. Moore, M. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (2008) 98–101.
[124] D.J. Watts, A twenty-first century science, Nature 445 (7127) (2007) 489–489.
[125] J.B. Schafer, J.A. Konstan, J. Riedl, E-commerce recommendation applications, in: Applications of Data Mining to Electronic Commerce, Springer, 2001, pp. 115–153.
[126] A. Vespignani, Predicting the behavior of techno-social systems, Science 325 (5939) (2009) 425–428.
[127] A. Vidmer, A. Zeng, M. Medo, Y.-C. Zhang, Prediction in complex systems: The case of the international trade network, Physica A 436 (2015) 188–199.
[128] P.B. Kantor, L. Rokach, F. Ricci, B. Shapira, Recommender Systems Handbook, Springer, 2011.
[129] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Phys. Rep. 519 (1) (2012) 1–49.
[130] J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez, Recommender systems survey, Knowl.-Based Syst. 46 (2013) 109–132.
[131] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, ACM, 1994, pp. 175–186.
[132] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th International Conference on World Wide Web, ACM, 2001, pp. 285–295.
[133] J.S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1998, pp. 43–52.
[134] D. Goldberg, D. Nichols, B.M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry, Commun. ACM 35 (12) (1992) 61–70.
[135] L.H. Ungar, D.P. Foster, Clustering methods for collaborative filtering, in: AAAI Workshop on Recommendation Systems, vol. 1, 1998, pp. 114–129.
[136] L. Ungar, D.P. Foster, A formal statistical approach to collaborative filtering, in: CONALD98, 1998.
[137] Y. Azar, A. Fiat, A. Karlin, F. McSherry, J. Saia, Spectral analysis of data, in: Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing, ACM, 2001, pp. 619–626.
[138] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer (8) (2009) 30–37.
[139] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[140] R. Keshavan, A. Montanari, S. Oh, Matrix completion from noisy entries, in: Advances in Neural Information Processing Systems, 2009, pp. 952–960.
[141] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47–97.
[142] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Modern Phys. 81 (2) (2009) 591–646.
[143] Y.-C. Zhang, M. Blattner, Y.-K. Yu, Heat conduction process on community networks as a recommendation model, Phys. Rev. Lett. 99 (15) (2007) 154301.
[144] Y.-C. Zhang, M. Medo, J. Ren, T. Zhou, T. Li, F. Yang, Recommendation model based on opinion diffusion, Europhys. Lett. 80 (6) (2007) 68003.
[145] J.-H. Liu, Z.-K. Zhang, C. Yang, L. Chen, C. Liu, X. Wang, Gravity effects on information filtering and network evolving, PLoS One 9 (3) (2014) e91070.
[146] W. Zeng, A. Zeng, H. Liu, M.-S. Shang, T. Zhou, Uncovering the information core in recommender systems, Sci. Rep. 4 (2014).
[147] Q.-M. Zhang, A. Zeng, M.-S. Shang, Extracting the information backbone in online system, PLoS One 8 (5) (2013) e62624.
[148] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J.R. Wakeling, Y.-C. Zhang, Solving the apparent diversity-accuracy dilemma of recommender systems, Proc. Natl. Acad. Sci. 107 (10) (2010) 4511–4515.
[149] A. Fiasconaro, M. Tumminello, V. Nicosia, V. Latora, R.N. Mantegna, Hybrid recommendation methods in complex networks, Phys. Rev. E 92 (1) (2015) 012811.
[150] A. Stojmirovic, Y.-K. Yu, Information flow in interaction networks, J. Comput. Biol. 14 (8) (2007) 1115–1143.
[151] J.-G. Liu, T. Zhou, Q. Guo, Information filtering via biased heat conduction, Phys. Rev. E 84 (3) (2011) 037101.
[152] M.-S. Shang, C.-H. Jin, T. Zhou, Y.-C. Zhang, Collaborative filtering based on multi-channel diffusion, Physica A 388 (23) (2009) 4867–4871.
[153] J.-G. Liu, K. Shi, Q. Guo, Solving the accuracy-diversity dilemma via directed random walks, Phys. Rev. E 85 (1) (2012) 016118.
[154] Z.-K. Zhang, C. Liu, A hypergraph model of social tagging networks, J. Stat. Mech. Theory Exp. 2010 (10) (2010) P10005.
[155] Z.-K. Zhang, T. Zhou, Y.-C. Zhang, Tag-aware recommender systems: a state-of-the-art survey, J. Comput. Sci. Tech. 26 (5) (2011) 767–777.
[156] L. Yu, C. Liu, Z.-K. Zhang, Multi-linear interactive matrix factorization, Knowl.-Based Syst. 85 (2015) 307–315.
[157] Z.-M. Ren, Y. Kong, M.-S. Shang, Y.-C. Zhang, A generalized model via random walks for information filtering, Phys. Lett. A 380 (34) (2016) 2608–2614.
[158] L. Yu, J. Huang, G. Zhou, C. Liu, Z.-K. Zhang, TIIREC: A tensor approach for tag-driven item recommendation with sparse user generated content, Inform. Sci. 411 (2017) 122–135.
[159] B.J. Kim, Geographical coarse graining of complex networks, Phys. Rev. Lett. 93 (16) (2004) 168701.
[160] S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, U. Alon, Coarse-graining and self-dissimilarity of complex networks, Phys. Rev. E 71 (1) (2005) 016127.
[161] D. Gfeller, P. De Los Rios, Spectral coarse graining of complex networks, Phys. Rev. Lett. 99 (3) (2007) 038701.

[162] I. Farkas, D. Ábel, G. Palla, T. Vicsek, Weighted network modules, New J. Phys. 9 (6) (2007) 180.
[163] R. Moreno-Bote, N. Parga, Role of synaptic filtering on the firing response of simple model neurons, Phys. Rev. Lett. 92 (2) (2004) 028102.
[164] Z. Wu, L.A. Braunstein, S. Havlin, H.E. Stanley, Transport in weighted networks: partition into superhighways and roads, Phys. Rev. Lett. 96 (14) (2006) 148702.
[165] J.J. Ramasco, B. Gonçalves, Transport on weighted networks: When the correlations are independent of the degree, Phys. Rev. E 76 (6) (2007) 066106.
[166] P.B. Slater, A two-stage algorithm for extracting the multiscale backbone of complex weighted networks, Proc. Natl. Acad. Sci. 106 (26) (2009) E66–E66.
[167] J.B. Glattfelder, S. Battiston, Backbone of complex networks of corporations: The flow of control, Phys. Rev. E 80 (3) (2009) 036104.
[168] M.Á. Serrano, M. Boguná, A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proc. Natl. Acad. Sci. 106 (16) (2009) 6483–6488.
[169] P. Macdonald, E. Almaas, A.-L. Barabási, Minimum spanning trees of weighted scale-free networks, Europhys. Lett. 72 (2) (2005) 308.
[170] D. Grady, C. Thiemann, D. Brockmann, Robust classification of salient links in complex networks, Nat. Commun. 3 (2012) 864.
[171] L. da Fontoura Costa, The hierarchical backbone of complex networks, Phys. Rev. Lett. 93 (9) (2004) 098702.
[172] G. Wang, C. Du, H. Chen, R. Simha, Y. Rong, Y. Xiao, C. Zeng, Process-based network decomposition reveals backbone motif structure, Proc. Natl. Acad. Sci. 107 (23) (2010) 10478–10483.
[173] J.F. Donges, Y. Zou, N. Marwan, J. Kurths, The backbone of the climate network, Europhys. Lett. 87 (4) (2009) 48007.
[174] M. Tumminello, T. Aste, T. Di Matteo, R.N. Mantegna, A tool for filtering information in complex systems, Proc. Natl. Acad. Sci. USA 102 (30) (2005) 10421–10426.
[175] D. Witthaut, M. Rohden, X. Zhang, S. Hallerberg, M. Timme, Critical links and nonlocal rerouting in complex supply networks, Phys. Rev. Lett. 116 (13) (2016) 138701.
[176] L. Marotta, S. Miccichè, Y. Fujiwara, H. Iyetomi, H. Aoyama, M. Gallegati, R.N. Mantegna, Backbone of credit relationships in the Japanese credit market, EPJ Data Sci. 5 (1) (2016) 10.
[177] X. Zhang, Z. Zhang, H. Zhao, Q. Wang, J. Zhu, Extracting the globally and locally adaptive backbone of complex networks, PLoS One 9 (6) (2014) e100428.
[178] F. Radicchi, J.J. Ramasco, S. Fortunato, Information filtering in complex weighted networks, Phys. Rev. E 83 (4) (2011) 046101.
[179] N.J. Foti, J.M. Hughes, D.N. Rockmore, Nonparametric sparsification of complex multiscale networks, PLoS One 6 (2) (2011) e16431.
[180] S. Scellato, A. Cardillo, V. Latora, S. Porta, The backbone of a city, Eur. Phys. J. B 50 (1) (2006) 221–225.
[181] W. Choi, H. Chae, S.-H. Yook, Y. Kim, Classification of transport backbones of complex networks, Phys. Rev. E 88 (6) (2013) 060802.
[182] R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Resilience of the internet to random breakdowns, Phys. Rev. Lett. 85 (21) (2000) 4626.
[183] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks, Proc. Natl. Acad. Sci. USA 101 (11) (2004) 3747–3752.
[184] R.K. Ahuja, T.L. Magnanti, J.B. Orlin, Network flows: theory, algorithms, and applications, 1993.
[185] M.C. Lagomarsino, P. Jona, B. Bassetti, Logic backbone of a transcription network, Phys. Rev. Lett. 95 (15) (2005) 158701.
[186] J. Kim, T. Wilhelm, Spanning tree separation reveals community structure in networks, Phys. Rev. E 87 (3) (2013) 032816.
[187] W. Zeng, M. Fang, J. Shao, M. Shang, Uncovering the essential links in online commercial networks, Sci. Rep. 6 (2016) 34292.
[188] B. Balassa, Trade liberalisation and "revealed" comparative advantage, Manchester Sch. 33 (2) (1965) 99–123.
[189] Y. Hulovatyy, R.W. Solava, T. Milenković, Revealing missing parts of the interactome via link prediction, PLoS One 9 (3) (2014) e90073.
[190] C. Fan, Z. Liu, X. Lu, B. Xiu, Q. Chen, An efficient link prediction index for complex military organization, Physica A 469 (2017) 572–587.
[191] M.L. Lee, W. Hsu, V. Kothari, Cleaning the spurious links in data, IEEE Intell. Syst. 19 (2) (2004) 28–33.
[192] A. Zeng, G. Cimini, Removing spurious interactions in complex networks, Phys. Rev. E 85 (3) (2012) 036101.
[193] P. Zhang, A. Zeng, Y. Fan, Identifying missing and spurious connections via the bi-directional diffusion on bipartite networks, Phys. Lett. A 378 (32) (2014) 2350–2354.
[194] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[195] S. Fortunato, D. Hric, Community detection in networks: A user guide, Phys. Rep. 659 (2016) 1–44.
[196] F.D. Malliaros, M. Vazirgiannis, Clustering and community detection in directed networks: A survey, Phys. Rep. 533 (4) (2013) 95–142.
[197] M.E. Newman, Communities, modules and large-scale structure in networks, Nat. Phys. 8 (1) (2012) 25.
[198] M.E. Newman, Detecting community structure in networks, Eur. Phys. J. B 38 (2) (2004) 321–330.
[199] S.E. Schaeffer, Graph clustering, Comput. Sci. Rev. 1 (1) (2007) 27–64.
[200] M.A. Porter, J.-P. Onnela, P.J. Mucha, Communities in networks, Notices Amer. Math. Soc. 56 (9) (2009) 1082–1097.
[201] M. Coscia, F. Giannotti, D. Pedreschi, A classification for community discovery methods in complex networks, Stat. Anal. Data Min. 4 (5) (2011) 512–546.
[202] S. Parthasarathy, Y. Ruan, V. Satuluri, Community discovery in social networks: Applications, methods and emerging trends, in: Social Network Data Analytics, Springer, 2011, pp. 79–113.
[203] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, ACM Comput. Surv. 45 (4) (2013) 43.
[204] T. Chakraborty, A. Dalmia, A. Mukherjee, N. Ganguly, Metrics for community analysis: A survey, ACM Comput. Surv. 50 (4) (2017) 54.
[205] K. Yang, Q. Guo, S.-N. Li, J.-T. Han, J.-G. Liu, Evolution properties of the community members for dynamic networks, Phys. Lett. A 381 (11) (2017) 970–975.
[206] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, Structure and tie strengths in mobile communication networks, Proc. Natl. Acad. Sci. 104 (18) (2007) 7332–7336.
[207] J. Sun, J.P. Bagrow, E.M. Bollt, J.D. Skufca, Dynamic computation of network statistics via updating schema, Phys. Rev. E 79 (3) (2009) 036116.
[208] L. Weng, F. Menczer, Y.-Y. Ahn, Virality prediction and community structure in social networks, Sci. Rep. 3 (2013) 2522.
[209] R. Karan, B. Biswal, A model for evolution of overlapping community networks, Physica A 474 (2017) 380–390.
[210] M. De Domenico, Diffusion geometry unravels the emergence of functional clusters in collective phenomena, Phys. Rev. Lett. 118 (16) (2017) 168301.
[211] G. Palla, A.-L. Barabasi, T. Vicsek, Quantifying social group evolution, Nature 446 (7136) (2007) 664–668.
[212] J.-G. Young, L. Hébert-Dufresne, A. Allard, L.J. Dubé, Growing networks of overlapping communities with internal structure, Phys. Rev. E 94 (2) (2016) 022317.
[213] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, L.J. Dubé, Structural preferential attachment: Network organization beyond the link, Phys. Rev. Lett. 107 (15) (2011) 158702.
[214] A. Mirshahvalad, O.H. Beauchesne, É. Archambault, M. Rosvall, Resampling effects on significance analysis of network clustering and ranking, PLoS One 8 (1) (2013) e53943.
[215] I. Tzekina, K. Danthi, D.N. Rockmore, Evolution of community structure in the world trade web, Eur. Phys. J. B 63 (4) (2008) 541–545.
[216] W. Zhong, H. An, X. Gao, X. Sun, The evolution of communities in the international oil trade network, Physica A 413 (2014) 42–52.
[217] O. Güell, F. Sagués, M.A. Serrano, Predicting effects of structural stress in a genome-reduced model bacterial metabolism, Sci. Rep. 2 (2012).

[218] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, Science 328 (5980) (2010) 876–878.

[219] G.-Q. Zhang, G.-Q. Zhang, Q.-F. Yang, S.-Q. Cheng, T. Zhou, Evolution of the Internet and its cores, New J. Phys. 10 (12) (2008) 123027.

[220] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, A model of Internet topology using k-shell decomposition, Proc. Natl. Acad. Sci. 104 (27) (2007) 11150–11154.

[221] C. Orsini, E. Gregori, L. Lenzini, D. Krioukov, Evolution of the Internet k-dense structure, IEEE/ACM Trans. Netw. 22 (6) (2014) 1769–1780.

[222] J.-G. Liu, Z.-M. Ren, Q. Guo, D.-B. Chen, Evolution characteristics of the network core in the Facebook, PLoS One 9 (8) (2014) e104028.

[223] G. Kossinets, D.J. Watts, Empirical analysis of an evolving social network, Science 311 (5757) (2006) 88–90.

[224] J. Li, D. Ren, X. Feng, Y. Zhang, Network of listed companies based on common shareholders and the prediction of market volatility, Physica A 462 (2016) 508–521.

[225] Q. Guo, L. Ji, J.-G. Liu, J. Han, Evolution properties of online user preference diversity, Physica A 468 (2017) 698–713.

[226] J. Liu, L. Hou, Y.-L. Zhang, W.-J. Song, X. Pan, Empirical analysis of the clustering coefficient in the user-object bipartite networks, Internat. J. Modern Phys. C 24 (08) (2013) 1350055.

[227] P. Zhang, M. Li, L. Gao, Y. Fan, Z. Di, Characterizing and modeling the dynamics of activity and popularity, PLoS One 9 (2) (2014) e89192.

[228] J.-Q. Zhang, Z.-G. Huang, Z.-X. Wu, R. Su, Y.-C. Lai, Controlling herding in minority game systems, Sci. Rep. 6 (2016).

[229] S. Bornholdt, T. Rohlf, Topological evolution of dynamical networks: Global criticality from local dynamics, Phys. Rev. Lett. 84 (26) (2000) 6114.

[230] L. Demetrius, T. Manke, Robustness and network evolutionan entropic principle, Physica A 346 (3) (2005) 682–696.

[231] P. Manshour, Complex network approach to fractional time series, Chaos 25 (10) (2015) 103105.

[232] W.-X. Wang, Q. Chen, L. Huang, Y.-C. Lai, M.A.F. Harrison, Scaling of noisy fluctuations in complex networks and applications to network prediction, Phys. Rev. E 80 (1) (2009) 016116.

[233] S. Sikdar, N. Ganguly, A. Mukherjee, Time series analysis of temporal networks, Eur. Phys. J. B 89 (1) (2016) 11.

[234] C. Chatfield, The Analysis of Time Series: An Introduction, CRC press, 2016.

[235] G.E. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time Series Analysis: Forecasting and Control, John Wiley & Sons, 2015.

[236] S. Sreenivasan, Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords, Sci. Rep. 3 (2013).

[237] R. Sharda, D. Delen, Predicting box-office success of motion pictures with neural networks, Expert Syst. Appl. 30 (2) (2006) 243–254.

[238] G. Mishne, N.S. Glance, et al. Predicting movie sales from blogger sentiment, in: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, pp. 155–158.

[239] S. Asur, B.A. Huberman, Predicting the future with social media, in: Web Intelligence and Intelligent Agent Technology, WI-IAT, 2010 IEEE/WIC/ACM International Conference on, vol. 1, IEEE, 2010, pp. 492–499.

[240] A. Oghina, M. Breuss, M. Tsagkias, M. de Rijke, Predicting imdb movie ratings using social media, in: European Conference on Information Retrieval, Springer, 2012, pp. 503–507.

[241] M. Tsagkias, W. Weerkamp, M. De Rijke, Predicting the volume of comments on online news stories, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, 2009, pp. 1765–1768.

[242] M. Tsagkias, W. Weerkamp, M. De Rijke, News comments: Exploring, modeling, and online prediction, Adv. Inf. Retr. (2010) 191–203.

[243] C. Castillo, M. El-Haddad, J. Pfeffer, M. Stempeck, Characterizing the life cycle of online news stories using social media reactions, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2014, pp. 211–223.

[244] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (7232) (2009) 1012–1014.

[245] Z. Tan, Y. Wang, Y. Zhang, J. Zhou, A novel time series approach for predicting the long-term popularity of online videos, IEEE Trans. Broadcast. 62 (2) (2016) 436–445.

[246] C. Wang, B.A. Huberman, Long trend dynamics in social media, EPJ Data Sci. 1 (1) (2012) 2.

[247] C.-W. Yeh, D.-C. Li, A trend prediction model from very short term data learning, Expert Syst. Appl. 37 (2) (2010) 1728–1733.

[248] M. Mestyán, T. Yasseri, J. Kertész, Early prediction of movie box office success based on Wikipedia activity big data, PLoS One 8 (8) (2013) e71226.

[249] C. Wang, B.A. Huberman, How random are online social interactions?, Sci. Rep. 2 (2012).

[250] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.

[251] L. Wang, Z. Wang, S. Zhao, S. Tan, Stock market trend prediction using dynamical Bayesian factor graph, Expert Syst. Appl. 42 (15) (2015) 6267–6275.

[252] M. Podsiadlo, H. Rybinski, Financial time series forecasting using rough sets with time-weighted rule voting, Expert Syst. Appl. 66 (2016) 219–233.

[253] D.C. Furlaneto, L.S. Oliveira, D. Menotti, G.D. Cavalcanti, Bias effect on predicting market trends with EMD, Expert Syst. Appl. 82 (2017) 19–26.

[254] X.-d. Zhang, A. Li, R. Pan, Stock trend prediction based on a new status box method and adaboost probabilistic support vector machine, Appl. Soft Comput. 49 (2016) 385–398.

[255] L.-P. Ni, Z.-W. Ni, Y.-Z. Gao, Stock trend prediction based on fractal feature selection and support vector machine, Expert Syst. Appl. 38 (5) (2011) 5569–5576.

[256] H.-W. Mewes, D. Frishman, K.F. Mayer, M. Münsterkötter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, V. Stümpflen, MIPS: analysis and annotation of proteins from whole genomes in 2005, Nucleic Acids Res. 34 (2006) D169–D172.

[257] J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, et al., SGD: Saccharomyces genome database, Nucleic Acids Res. 26 (1) (1998) 73–79.

[258] R. Zhang, Y. Lin, DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes, Nucleic Acids Res. 37 (2008) D455–D458.

[259] J. Wang, M. Li, H. Wang, Y. Pan, Identification of essential proteins based on edge clustering coefficient, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (4) (2012) 1070–1080.

[260] X. Zhang, J. Xu, W.-x. Xiao, A new method for the discovery of essential proteins, PLoS One 8 (3) (2013) e58763.

[261] M. Li, H. Zhang, J.-x. Wang, Y. Pan, A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data, BMC Syst. Biol. 6 (1) (2012) 15.

[262] J.B. Pereira-Leal, B. Audit, J.M. Peregrin-Alvarez, C.A. Ouzounis, An exponential core in the heart of the yeast protein interaction network, Mol. Biol. Evol. 22 (3) (2004) 421–425.

[263] J. Luo, Y. Qi, Identification of essential proteins based on a new combination of local interaction density and protein complexes, PLoS One 10 (6) (2015) e0131418.

[264] M. Li, Y. Lu, J. Wang, F.-X. Wu, Y. Pan, A topology potential-based method for identifying essential proteins from PPI networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 12 (2) (2015) 372–383.

[265] Y. Tang, M. Li, J. Wang, Y. Pan, F.-X. Wu, CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks, Biosystems 127 (2015) 67–72.

[266] J. Ren, J. Wang, M. Li, H. Wang, B. Liu, Prediction of essential proteins by integration of PPI network topology and protein complexes information, Bioinform. Res. Appl. (2011) 12–24.

[267] M. Li, Y. Lu, Z. Niu, F.-X. Wu, United complex centrality for identification of essential proteins from PPI networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 14 (2) (2017) 370–380.

[268] M. Hsing, K.G. Byler, A. Cherkasov, The use of gene ontology terms for predicting highly-connected'hub'nodes in protein-protein interaction networks, BMC Syst. Biol. 2 (1) (2008) 80.

[269] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, Y. Pan, Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks, BMC Syst. Biol. 6 (1) (2012) 87.

[270] Q. Xiao, J. Wang, X. Peng, F.-x. Wu, Y. Pan, Identifying essential proteins from active PPI networks constructed with dynamic gene expression, BMC Genomics 16 (3) (2015) S1.

[271] S.J. Furney, M.M. Albà, N. López-Bigas, Differences in the evolutionary history of disease genes affected by dominant or recessive mutations, BMC Genomics 7 (1) (2006) 165.

[272] B. Zhao, J. Wang, M. Li, F.-x. Wu, Y. Pan, Prediction of essential proteins based on overlapping essential modules, IEEE Trans. Nanobiosci. 13 (4) (2014) 415–424.

[273] T. Ideker, R. Sharan, Protein networks in disease, Genome Res. 18 (4) (2008) 644–652.

[274] M.G. Kann, Protein interactions and disease: computational approaches to uncover the etiology of diseases, Brief. Bioinform. 8 (5) (2007) 333–346.

[275] J. Stark, R. Callard, M. Hubank, From the top down: towards a predictive biology of signalling networks, Trends Biotechnol. 21 (7) (2003) 290–293.

[276] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, et al., Towards a proteome-scale map of the human protein-protein interaction network, Nature 437 (7062) (2005) 1173–1173.

[277] R. Bonneau, M.T. Facciotti, D.J. Reiss, A.K. Schmid, M. Pan, A. Kaur, V. Thorsson, P. Shannon, M.H. Johnson, J.C. Bare, et al., A predictive model for transcriptional control of physiology in a free living cell, Cell 131 (7) (2007) 1354–1365.

[278] H. Yu, P. Braun, M.A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, et al., High-quality binary protein interaction map of the yeast interactome network, Science 322 (5898) (2008) 104–110.

[279] S. Guo, J. Wu, M. Ding, J. Feng, Uncovering interactions in the frequency domain, PLoS Comput. Biol. 4 (5) (2008) e1000087.

[280] P. Braun, M. Tasan, M. Dreze, M. Barrios-Rodiles, I. Lemmens, H. Yu, J.M. Sahalie, R.R. Murray, L. Roncari, A.-S. De Smet, et al., An experimentally derived confidence score for binary protein-protein interactions, Nature Methods 6 (1) (2009) 91–97.

[281] B. Snijder, P. Liberali, M. Frechin, T. Stoeger, L. Pelkmans, Predicting functional gene interactions with the hierarchical interaction score, Nature Methods 10 (11) (2013) 1089–1092.

[282] A.E. Motter, N. Gulbahce, E. Almaas, A.-L. Barabási, Predicting synthetic rescues in metabolic networks, Mol. Syst. Biol. 4 (1) (2008) 168–168.

[283] B. Barzel, A.-L. Barabási, Network link prediction by global silencing of indirect correlations, Nature Biotechnol. 31 (8) (2013) 720–725.

[284] G. Yan, P.E. Vértes, E.K. Towlson, Y.L. Chew, D.S. Walker, W.R. Schafer, A.L. Barabási, Network control principles predict neuron function in the Caenorhabditis elegans connectome, Nature 550 (7677) (2017) 519–523.

[285] A. Clauset, D.B. Larremore, R. Sinatra, Data-driven predictions in the science of science, Science 355 (6324) (2017) 477–480.

[286] A. Zeng, S. Zhesi, Z. Jianlin, W. Jinshan, F. Ying, W. Yougui, S.H. Eugene, The science of science: From the perspective of complex systems, Phys. Rep. 714 (2017) 714–715.

[287] M. Golosovsky, S. Solomon, Uncovering the dynamics of citations of scientific papers, 2014. ArXiv preprint arXiv:1410.0343.

[288] M. Newman, Prediction of highly cited papers, Europhys. Lett. 105 (2) (2014) 28002.

[289] J. Zhou, A. Zeng, Y. Fan, Z. Di, Ranking scientific publications with similarity-preferential mechanism, Scientometrics 106 (2) (2016) 805–816.

[290] L. Yao, T. Wei, A. Zeng, Y. Fan, Z. Di, Ranking scientific publications: the effect of nonlinearity, Sci. Rep. 4 (2014).

[291] G. Eysenbach, Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact, J. Med. Internet Res. 13 (4) (2011).

[292] A.F. Van Raan, Sleeping beauties in science, Scientometrics 59 (3) (2004) 467–472.

[293] Q.L. Burrell, Are "sleeping beauties" to be expected?, Scientometrics 65 (3) (2005) 381–389.

[294] Q. Ke, E. Ferrara, F. Radicchi, A. Flammini, Defining and identifying sleeping beauties in science, Proc. Natl. Acad. Sci. 112 (24) (2015) 7426–7431.

[295] Q. Niu, J. Zhou, A. Zeng, Y. Fan, Z. Di, Which publication is your representative work?, J. Inform. 10 (3) (2016) 842–853.

[296] Y. Yin, D. Wang, The time dimension of science: Connecting the past to the future, J. Inform. 11 (2) (2017) 608–621.

[297] A. Mazloumian, Predicting scholars' scientific impact, PLoS One 7 (11) (2012) e49246.

[298] J.E. Hirsch, An index to quantify an individual's scientific research output, Proc. Natl. Acad. Sci. USA 102 (46) (2005) 16569.

[299] J.E. Hirsch, Does the h index have predictive power?, Proc. Natl. Acad. Sci. 104 (49) (2007) 19193–19198.

[300] L. Egghe, An improvement of the h-index: The g-index, in: ISSI, 2006.

[301] O. Penner, R.K. Pan, A.M. Petersen, K. Kaski, S. Fortunato, On the predictability of future impact in science, Sci. Rep. 3 (2013).

[302] R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, Quantifying the evolution of individual scientific impact, Science 354 (6312) (2016) aaf5239.

[303] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, F. Schweitzer, Predicting scientific success based on coauthorship networks, EPJ Data Sci. 3 (1) (2014) 9.

[304] M. Qi, A. Zeng, M. Li, Y. Fan, Z. Di, Standing on the shoulders of giants: the effect of outstanding scientists on young collaborators careers, Scientometrics 111 (3) (2017) 1839–1850.

[305] L. Wu, D. Wang, J.A. Evans, Large teams have developed science and technology; small teams have disrupted it, 2017.

[306] T. Jia, D. Wang, B.K. Szymanski, Quantifying patterns of research-interest evolution, Nat. Hum. Behav. 1 (2017) 0078.

[307] G. Caldarelli, A. Chessa, F. Pammolli, A. Gabrielli, M. Puliga, Reconstructing a credit network, Nat. Phys. 9 (2013) 125–126.

[308] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, D.R. White, Economic networks: The new challenges, Science 325 (5939) (2009) 422–425.

[309] C.A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The product space conditions the development of nations, Science 317 (5837) (2007) 482–487.

[310] C.A. Hidalgo, Economic complexity: From useless to keystone, Nat. Phys. 14 (1) (2018) 9.

[311] G. Caldarelli, M. Cristelli, A. Gabrielli, L. Pietronero, A. Scala, A. Tacchella, A network analysis of countries export flows: firm grounds for the building blocks of the economy, PLoS One 7 (10) (2012) e47278.

[312] C.A. Hidalgo, R. Hausmann, The building blocks of economic complexity, Proc. Natl. Acad. Sci. 106 (26) (2009) 10570–10575.

[313] J. Felipe, C. Hidalgo, Economic diversification implications for Kazakhstan, in: J. Felipe (Ed.), Development and Modern Industrial Policy in Practice: Issues and Country Experiences, 2015, pp. 160–196.

[314] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, L. Pietronero, A new metrics for countries' fitness and products' complexity, Sci. Rep. 2 (2012) 723.

[315] M. Cristelli, A. Tacchella, L. Pietronero, The heterogeneous dynamics of economic complexity, PLoS One 10 (2) (2015) e0117174.

[316] R. Hausmann, C.A. Hidalgo, S. Bustos, M. Coscia, A. Simoes, M.A. Yildirim, The Atlas of Economic Complexity: Mapping Paths to Prosperity, Mit Press, 2014.

[317] D. Hartmann, M.R. Guevara, C. Jara-Figueroa, M. Aristarán, C.A. Hidalgo, Linking economic complexity, institutions, and income inequality, World Dev. 93 (2017) 75–93.

[318] J. Gao, T. Zhou, Quantifying China's regional economic complexity, 2017. ArXiv preprint arXiv:1703.01292.

[319] J. Gao, B. Jun, A. Pentland, T. Zhou, C.A. Hidalgo, et al. Collective learning in China's regional economic development, 2017. ArXiv preprint arXiv:1703.01369.

[320] M. Cristelli, A. Tacchella, A. Gabrielli, L. Pietronero, A. Scala, G. Caldarelli, Competitors communities and taxonomy of products according to export fluxes, Eur. Phys. J. Spec. Top. 212 (1) (2012) 115–120.

[321] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, L. Pietronero, Economic complexity: conceptual grounding of a new metrics for global competitiveness, J. Econom. Dynam. Control 37 (8) (2013) 1683–1691.
[322] M. Cristelli, A. Gabrielli, A. Tacchella, G. Caldarelli, L. Pietronero, Measuring the intangibles: A metrics for the economic complexity of countries and products, PLoS One 8 (8) (2013) e70726.
[323] M. Cristelli, A. Tacchella, L. Pietronero, An overview of the new frontiers of economic complexity, in: Econophysics of Agent-Based Models, Springer, 2014, pp. 147–159.
[324] A. Zaccaria, M. Cristelli, A. Tacchella, L. Pietronero, How the taxonomy of products drives the economic development of countries, PLoS One 9 (12) (2014) e113770.
[325] M. Cristelli, A. Tacchella, A. Zaccaria, L. Pietronero, Growth scenarios for sub-Saharan countries in the framework of economic complexity, 2014.
[326] E. Pugliese, G.L. Chiarotti, A. Zaccaria, L. Pietronero, Complex economies have a lateral escape from the poverty trap, PLoS One 12 (1) (2017) e0168540.
[327] O. Angelini, M. Cristelli, A. Zaccaria, L. Pietronero, The complex dynamics of products and its asymptotic properties, PLoS One 12 (5) (2017) e0177360.
[328] E. Pugliese, A. Zaccaria, L. Pietronero, On the convergence of the fitness-complexity algorithm, Eur. Phys. J. Spec. Top. 225 (10) (2016) 1893–1911.
[329] R.-J. Wu, G.-Y. Shi, Y.-C. Zhang, M.S. Mariani, The mathematics of non-linear metrics for nested networks, Physica A 460 (2016) 254–269.
[330] M.S. Mariani, A. Vidmer, M. Medo, Y.-C. Zhang, Measuring economic complexity of countries and products: which metric to use?, Eur. Phys. J. B 88 (11) (2015) 293.
[331] E. Pugliese, G.L. Chiarotti, A. Zaccaria, L. Pietronero, Economic Complexity as a Determinant of the Industrialization of Countries: The Case of India, Cambridge Press, 2015.
[332] A. Zaccaria, M. Cristelli, R. Kupers, A. Tacchella, L. Pietronero, A case study for a new metrics for economic complexity: The Netherlands, J. Econ. Interact. Coord. 11 (1) (2016) 151–169.
[333] V. Stojkoski, Z. Utkovski, L. Kocarev, The impact of services on economic complexity: Service sophistication as route for economic growth, PLoS One 11 (8) (2016) e0161633.
[334] G. Cimini, A. Gabrielli, F.S. Labini, The scientific competitiveness of nations, PLoS One 9 (12) (2014) e113470.
[335] M.D. König, C.J. Tessone, Y. Zenou, Nestedness in networks: A theoretical model and some applications, Theor. Econ. 9 (3) (2014) 695–752.
[336] U. Bastolla, M.A. Fortuna, A. Pascual-García, A. Ferrera, B. Luque, J. Bascompte, The architecture of mutualistic networks minimizes competition and increases biodiversity, Nature 458 (7241) (2009) 1018–1020.
[337] R.P. Rohr, S. Saavedra, J. Bascompte, On the structural stability of mutualistic systems, Science 345 (6195) (2014) 1253497.
[338] S. Bustos, C. Gomez, R. Hausmann, C.A. Hidalgo, The dynamics of nestedness predicts the evolution of industrial ecosystems, PLoS One 7 (11) (2012) e49393.
[339] A. Garas, C. Rozenblat, F. Schweitzer, The network structure of city-firm relations, 2015. ArXiv preprint arXiv:1512.02859.
[340] M. Almeida-Neto, P. Guimarães, P.R. Guimarães, R.D. Loyola, W. Ulrich, A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement, Oikos 117 (8) (2008) 1227–1239.
[341] W. Ulrich, M. Almeida-Neto, N.J. Gotelli, A consumer's guide to nestedness analysis, Oikos 118 (1) (2009) 3–17.
[342] S.J. Beckett, C.A. Boulton, H.T. Williams, FALCON: a software package for analysis of nestedness in bipartite networks, F1000Research 3 (2014).
[343] Z.-M. Ren, A. Zeng, Y.-C. Zhang, Bridging nestedness and economic complexity in multilayer international trading networks, Preparing.
[344] H. Choi, H. Varian, Predicting the present with google trends, Econ. Rec. 88 (s1) (2012) 2–9.
[345] J. Blumenstock, G. Cadamuro, R. On, Predicting poverty and wealth from mobile phone metadata, Science 350 (6264) (2015) 1073–1076.
[346] X. Lu, E. Wetter, N. Bharti, A.J. Tatem, L. Bengtsson, Approaching the limit of predictability in human mobility, Sci. Rep. 3 (2013).
[347] N. Eagle, M. Macy, R. Claxton, Network diversity and economic development, Science 328 (5981) (2010) 1029–1031.
[348] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, Computational social science, Science 323 (5915) (2009) 721–723.
[349] G.N. Gilbert, Computational Social Science, vol. 21, Sage, 2010.
[350] M. Lim, R. Metzler, Y. Bar-Yam, Global pattern formation and ethnic/cultural violence, Science 317 (5844) (2007) 1540–1544.
[351] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature 439 (7075) (2006) 462–465.
[352] D. Brockmann, F. Theis, Money circulation, trackable items, and the emergence of universal human mobility patterns, IEEE Pervasive Comput. 7 (4) (2008).
[353] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science 342 (6164) (2013) 1337–1342.
[354] V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani, The role of the airline transportation network in the prediction and predictability of global epidemics, Proc. Natl. Acad. Sci. USA 103 (7) (2006) 2015–2020.
[355] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021.
[356] T. Takaguchi, M. Nakamura, N. Sato, K. Yano, N. Masuda, Predictability of conversation partners, Phys. Rev. X 1 (1) (2011) 011008.
[357] M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782.
[358] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with twitter: What 140 characters reveal about political sentiment, ICWSM 10 (1) (2010) 178–185.
[359] D. Gayo Avello, P.T. Metaxas, E. Mustafaraj, Limits of electoral predictions using twitter, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, 2011.
[360] D. Gayo-Avello, I wanted to predict elections with twitter and all I got was this lousy paper, in: A Balanced Survey on Election Prediction using Twitter Data, 2012.
[361] T. Brody, S. Harnad, L. Carr, Earlier web usage statistics as predictors of later citation impact, J. Assoc. Inf. Sci. Technol. 57 (8) (2006) 1060–1072.
[362] B. Barzel, A.-L. Barabási, Universality in network dynamics, Nat. Phys. 9 (10) (2013) 673–681.
[363] A. Tsonis, J. Elsner, Nonlinear prediction as a way of distinguishing chaos from random fractal sequences, Nature 358 (6383) (1992) 217–220.
[364] N. Boers, B. Bookhagen, H.M. Barbosa, N. Marwan, J. Kurths, J. Marengo, Prediction of extreme floods in the eastern Central Andes based on a complex networks approach, Nat. Commun. 5 (2014) ncomms6199.
[365] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, C. Grebogi, Predicting catastrophes in nonlinear dynamical systems by compressive sensing, Phys. Rev. Lett. 106 (15) (2011) 154101.
[366] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, M.A.F. Harrison, Time-seriesbased prediction of complex oscillator networks via compressive sensing, Europhys. Lett. 94 (4) (2011) 48006.
[367] M. De Domenico, A. Lima, M. Musolesi, Interdependence and predictability of human mobility and social interactions, Pervasive Mob. Comput. 9 (6) (2013) 798–807.
[368] V. Sekara, A. Stopczynski, S. Lehmann, Fundamental structures of dynamic social networks, Proc. Natl. Acad. Sci. 113 (36) (2016) 9977–9982.
[369] J.M. Miotto, E.G. Altmann, Predictability of extreme events in social media, PLoS One 9 (11) (2014) e111506.
[370] X. Lu, L. Bengtsson, P. Holme, Predictability of population displacement after the 2010 Haiti earthquake, Proc. Natl. Acad. Sci. 109 (29) (2012) 11576–11581.
[371] R. Guimerà, M. Sales-Pardo, Justice blocks and predictability of us supreme court votes, PLoS One 6 (11) (2011) e27188.
[372] A.C.A. Hope, A simplified Monte Carlo significance test procedure, J. R. Stat. Soc. 30 (3) (1968) 582–598.

[373] S.J. Schiff, P. So, T. Chang, R.E. Burke, T. Sauer, Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble, Phys. Rev. E 54 (6) (1996) 6708.
[374] F. Hamilton, T. Berry, T. Sauer, Predicting chaotic time series with a partial model, Phys. Rev. E 92 (1) (2015) 010902.
[375] A.C. Iles, M. Novak, Complexity increases predictability in allometrically constrained food webs, Amer. Nat. 188 (1) (2016) 87–98.
[376] V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani, The modeling of global epidemics: Stochastic dynamics and predictability, Bull. Math. Biol. 68 (8) (2006) 1893–1921.
[377] M. Loecher, J. Kadtke, Enhanced predictability of hierarchical propagation in complex networks, Phys. Lett. A 366 (6) (2007) 535–539.
[378] A.A. Tsonis, K.L. Swanson, Topology and predictability of el nino and la nina networks, Phys. Rev. Lett. 100 (22) (2008) 228502.