Network-Based Information Filtering Algorithms: Ranking and Recommendation

Matúš Medo

1 Introduction

After the Internet and the World Wide Web have become popular and widely available, the electronically stored online interactions of individuals have fast emerged as a challenge for researchers and, perhaps even faster, as a source of valuable information for entrepreneurs. We now have detailed records of informal friendship relations in social networks, purchases on e-commerce sites, various sorts of information being sent from one user to another, online collections of web bookmarks, and many other data sets that allow us to pose questions that are of interest from both academical and commercial point of view. For example, which other users of a social network you might want to be friend with? Which other items you might be interested to purchase? Who are the most influential users in a network? Which web page you might want to visit next? All these questions are not only interesting per se, but the answers to them may help entrepreneurs provide better service to their customers and, ultimately, increase their profits.

All the questions posed above have many different ways to be approached that belong to the field of information filtering [1]. The goal of information filtering is to eliminate the redundant or unsuitable information and thus overcome the information overload. In our case, information filtering helps users to choose from an abundant number of possibilities (available products, potential friends, etc.), those that are most likely to be of interest or use for them. Common approaches to this task are based on mathematical statistics, machine learning, and artificial intelligence [2, 3]. They formulate a parametric mathematical model which is calibrated using the readily available data and then use to predict unknown user opinions.

M. Medo (✉)
Physics Department, University of Fribourg, CH-1700 Fribourg, Switzerland
e-mail: matus.medo@unifr.ch

A. Mukherjee et al. (eds.), *Dynamics On and Of Complex Networks, Volume 2*,
Modeling and Simulation in Science, Engineering and Technology,
DOI 10.1007/978-1-4614-6729-8_16,

In this chapter we discuss a different class of algorithms that all make use of a network representation of the data. The current classical example of such an algorithm is PageRank which, while having a far-reaching history [4], has been reinvented and popularized by the founders of Google where it serves up to now as the key element of their Internet search engine [5]. As we shall see below, this algorithm is closely related to random walks that play an important role in physics. (In the case of PageRank, of course, we do not face a random walk in physical space but a random walk on a network consisting of web pages and directed links among them.) These network-based methods can be used alone or in combination with other information filtering techniques, giving rise to hybrid methods [6].

We focus here on two important information filtering tasks—ranking and recommendation. By ranking we mean producing a general list of available items (users or objects) that captures some inherent quality of them. Finding influential users or exceptional web pages belongs to this. By recommendation we mean preparing a specific "recommendation list" for each individual user, each list containing items that are likely to be appreciated by the given user. Finding potential friends or items to purchase belongs here. In addition to traditional unipartite networks where only nodes of one kind are present (such as the network of web sites connected by hyperlinks or a network of users connected by friendship relations), we will often make use of bipartite networks where nodes of two kinds are present. For example, a network connecting users with the items that they have purchased is bipartite because every link connects a user with an item while links between users or between items are entirely absent. For a review of networks and network analysis that do not directly contribute to ranking and recommendation yet they can help to understand the structure of the data in hand, see the survey of complex networks measurements in [7]. For a general overview of dynamical processes on complex networks, see [8].

2 Ranking

When we want to rank nodes of a network, there are obviously many approaches, each of them suiting a different purpose. The simplest possible ranking is by node degree (or, in the case of a directed network, node in-degree) which is based on the assumption that "important" nodes are those that are referred by many other nodes. Many other measures of node importance exist, based either on local or global properties of the given network [9]. In this section, we discuss the importance rankings where score of a node is directly computed by random walk or where score spreads among the nodes in a manner akin to random walk.

2.1 PageRank

When given a directed unipartite network, PageRank [5] is arguably the most famous method to produce a general ranking of the network's nodes. The method is based on the circular idea *"A node is important if it is pointed by other important nodes"* which can be applied to many different situations, including ranking of web sites (an important site is referred by important sites), scientific journals (articles from an important journal are cited by articles from important journals), and people (an important person is referred/trusted by important people). For a review of past research in this direction and the use of this circular idea in various disciplines, see [4].

We begin with a general exposition of the approach, denoting the importance/score of node i as h_i and the nonnegative strength of the link pointing from node i to node j as w_{ij} ($i = 1, \ldots, N$ where N is the number of nodes in the network). The above circular thesis can now be formalized as

$$h_j = \sum_i \frac{w_{ij}}{\sum_k w_{ik}} h_i \tag{1}$$

where the division with $\sum_k w_{ik}$ ensures that the importance of node i is distributed among the nodes pointed by it with each node receiving part proportional to w_{ij}. To simplify our notation, we introduce normalized weights $P_{ij} := w_{ij} / \sum_j w_{ij}$. Now we can write $h_j = \sum_i P_{ij} h_i$ which can be further simplified by matrix notation to get

$$\boldsymbol{h} = \mathsf{P}^{\mathrm{T}} \boldsymbol{h}. \tag{2}$$

This matrix form shows that the sought-for vector $\boldsymbol{h}$ is the right eigenvector of P^{T} associated with eigenvalue 1. Since P^{T} is now a column-normalized matrix (also called stochastic matrix), the Frobenius-Perron theorem applies and states that 1 is its largest eigenvalue. A solution to Eq. (2) thus always exists, and when matrix P is irreducible, this solution is unique. (A matrix is irreducible if and only if in the directed graph that the matrix represents there exists a directed path between any two vertices.) The uniqueness is of course up to multiplication of $\boldsymbol{h}$ by a constant factor which allows us to impose the normalization condition $\sum_i h_i = 1$. Note that Eq. (2) is similar to the eigenvector centrality measures that are common in the analysis of social networks [10, 11]. In that case, however, one does not employ a normalized matrix P but the network's adjacency matrix itself and searches for a vector $\boldsymbol{x}$ satisfying $\mathsf{A}^{\mathrm{T}} \boldsymbol{x} = \lambda \boldsymbol{x}$ where λ is a number.

In addition to the redistribution point of view described above, a random-walk view can often provide useful insights. The normalized weights P_{ij} can be interpreted as probabilities of moving from node i to node j and, consequently, h_i as the probability of being at node i. An initial probability distribution $\boldsymbol{h}^{(0)}$ transforms gradually by $\boldsymbol{h}^{(n+1)} = \mathsf{P}^{\mathrm{T}} \boldsymbol{h}^{(n)}$ until a stationary probability distribution corresponding to the largest eigenvalue of P^{T} is established. If this eigenvalue is degenerated, the stationary solution is not unique. The rate of convergence of this iterative method is determined by the magnitude of the second-largest eigenvalue of P^{T}.

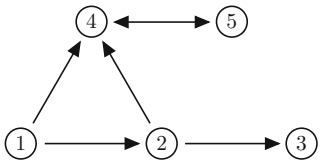

Fig. 1 PageRank computation for a toy network. When $\alpha = 1$ (no random teleportation), scores of nodes 1, 2, and 3 go with iterations to zero, but no stationary distribution exists because one of the eigenvalues is -1 and causes ceaseless alternations of score of nodes 4 and 5. When $\alpha = 0.85$ (the usual value adopted for web site ranking), the resulting score vector is $\boldsymbol{h} = (0.04, 0.06, 0.07, 0.43, 0.40)$. When $\alpha = 0$ (no link-following), all nodes have the same score $1/5$

Our treatment up to now was fully general and applies to any redistribution of h_i values over a weighted network given by weights w_{ij}. Depending on the nature of the problem and the input data, one needs to choose the weights so that the resulting importance vector $\boldsymbol{h}$ contains the information that we are interested in. In the case of PageRank, which was designed to produce the importance score for web sites, the input data consists of a directed network of web sites where a hyperlink from site A to site B can be interpreted as a sign of approval of site B by site A. Since no additional strength information is attached to hyperlinks, the network of hyperlinks is represented by its adjacency matrix A where $A_{ij} = 1$ if there is a link pointing from node i to node j (the network is directed and hence this matrix is not symmetric in general). Weights P_{ij} thus should be the same for all sites j referred by a given node i which, respecting the weight normalization condition, leads to $P_{ij} = A_{ij}/k_i^o$ where k_i^o is the out-degree of node i. Since this is ill defined for nodes with no out-going links ("dangling nodes"), one usually assumes that if $k_i^o = 0$, $P_{ij} = 1/N$ for all j.

One can easily see that even when the problem of nodes with zero out-degree is solved, the resulting solution can easily be pathological in some sense. If the network contains a component without out-going links (so-called bucket; see nodes 4 and 5 in Fig. 1), this part of the network would act as a trap for the random-walk process. Would concentrate there, and it would thus give us little useful information. The inventors of PageRank overcame the problem by postulating that links leading from a node are followed only with certain probability α [5]. With the complementary probability $1 - \alpha$, teleportation (jump) occurs, ending at a randomly chosen node of the network. The corresponding transition matrix (also called Google matrix) is

$$\mathsf{G} = \alpha \mathsf{P}^{\mathrm{T}} + (1 - \alpha)\mathsf{T} \tag{3}$$

where T is the teleportation matrix with all elements equal $1/N$. The parameter α (also called damping) and $1 - \alpha$ determines the weight given to link-following and teleportation, respectively. Since α is the probability of following an out-going link, one can easily compute that the average number of links followed in a row is $\sum_{k=0}^{\infty} k\alpha^k(1 - \alpha) = \alpha/(1 - \alpha)$. In the original PageRank paper, α was proposed to be set around 0.85 which corresponds to following five or six hyperlinks in a row and then jumping to a random page [5]. The value of α is closely related to the

convergence rate of the iterative PageRank computation (the lower the value, the faster the convergence; see [12] for more details). While PageRank was originally devised for directed networks, one can apply it also to undirected networks [13, 14]. The teleportation parameter then plays a crucial role—without it, PageRank score on an undirected network reduces to node degree.

Alternatively, one can replace the uniform teleportation matrix with $\mathbf{1}_N \boldsymbol{v}^{\mathrm{T}}$ where $\mathbf{1}_N$ is an N-dimensional vector of ones and $\boldsymbol{v}$ is a normalized N-dimensional vector which allows us to give preference to some nodes. This provides an important additional degree of freedom and allows one to, for example, devise a topic-specific ranking as described in [15]. A complementary point of view is presented, for example, in [16] where an inverse problem of finding matrix elements G_{ij} from some partial knowledge of node-pair preferences ("we want the score of node i to be higher than that of node j") is studied.

Using the definition of G given in Eq. (3), the PageRank equation $\mathsf{G}\boldsymbol{h} = \boldsymbol{h}$ can be written as $\alpha \mathsf{P}^{\mathrm{T}} \boldsymbol{h} + (1-\alpha)\mathbf{1}_N/N = \boldsymbol{h}$, leading to

$$\boldsymbol{h} = \frac{1-\alpha}{N} (\mathsf{I}_N - \alpha \mathsf{P}^{\mathrm{T}})^{-1} \mathbf{1}_N = \frac{1-\alpha}{N} \sum_{k=0}^{\infty} (\alpha \mathsf{P}^{\mathrm{T}})^k \mathbf{1}_N \tag{4}$$

where I_N is an $N \times N$ identity matrix and $\mathbf{1}_N$ is an N-dimensional vector of ones. Here both the inverse and the series expansion exist as long as $\alpha < 1$. While these formulas for computing $\boldsymbol{h}$ can be easily applied for small systems, a critical advantage of PageRank lies in the fact that the above-mentioned iterative method for finding $\boldsymbol{h}$ is in practice very effective even for very large systems. Thanks to that, PageRank serves as an important input for the Google's ranking of web sites where scores are computed for several billions of pages (for more information on the data mining for the WWW, see [17, 18]). Even for the enormous size of the WWW, only a few tens of iterations are sufficient to compute PageRank to a required precision [19]. The iterative method is also easy to parallelize and, in addition, one can write $\boldsymbol{h}^{(n+1)} = \alpha \mathsf{P}^{\mathrm{T}} \boldsymbol{h}^{(n)} + (1-\alpha)\mathbf{1}_N/N$ and thus benefit from the sparsity of P. In comparison with that, directly multiplying $\mathsf{G}\boldsymbol{h}^{(n)}$ is computationally much more expensive because G has no zero entries.

Another advantage of PageRank is that it is robust to spamming and malicious behavior. This robustness is rooted in the inability of web site administrators to create new hyperlinks pointing to their sites. If they simply create fake new web sites pointing to the site whose status they want to enhance, the artificially created web sites themselves have low scores (because no one points at them) and contribute little to the score of the target site. Of course, various sophisticated methods of manipulating the PageRank still exist [20]. The stability of node rankings obtained with PageRank is the central point in [21] where the authors show that PageRank is particularly prone to noisy data when the network is random (and thus the degree distribution, which is crucial for the ranking's stability, decays exponentially). By contrast, a small number of super-stable nodes whose ranking is particularly resistant to perturbations emerge in scale-free networks.

2.2 Variants of PageRank

From the conceptual point of view, an interesting generalization of PageRank has been proposed in [22] where spreading of the score was separated into branching (due to out-degree) and damping (due to the damping parameter α). In the case of PageRank, damping is exponential because with each propagation step, another multiplication with α is added. The authors show that a power-law damping of the form $1/[(t+1)(t+2)]$ where t denotes the number of steps is equivalent to a so-called TotalRank which is obtained simply by integrating the α-dependent PageRank score over α. Importantly, a linear damping can produce results very close to those obtained with PageRank while requiring fewer iterations to be computed. An important variant of PageRank, EigenTrust, has been proposed to compute trust values in distributed peer-to-peer systems [23]. EigenTrust, which replaces uniform teleportation matrix with random jumps to a set of pre-trusted peers, can be easily computed in a distributed way and is thus suitable for deployment in distributed P2P systems. A very different perspective was adopted in [24] where a class of quantum PageRank algorithms was proposed based on quantized Markov chains.

Almost at the same time as PageRank, another important algorithm based on random walks and circular reasoning was developed independently. It is called HITS (Hypertext-Induced Topic Search), and by contrast to PageRank, it assigns two distinct scores to each node—authority score x_i and hub score y_i [25]. The basic thesis is that a good hub is pointed to by good authorities and vice versa. In mathematical terms, this can be written as

$$\boldsymbol{x}^{(n+1)} = \mathsf{A}^{\mathrm{T}}\boldsymbol{y}^{(n)}, \quad \boldsymbol{y}^{(n+1)} = \mathsf{A}\boldsymbol{x}^{(n+1)} \tag{5}$$

Consequently, one can write $\boldsymbol{x}^{(n+1)} = \mathsf{A}^{\mathrm{T}}\mathsf{A}\boldsymbol{x}^{(n)}$ and $\boldsymbol{y}^{(n+1)} = \mathsf{A}\mathsf{A}^{\mathrm{T}}\boldsymbol{y}^{(n)}$, showing that the stationary authority and hub vectors are the dominant eigenvectors of $\mathsf{A}^{\mathrm{T}}\mathsf{A}$ and $\mathsf{A}\mathsf{A}^{\mathrm{T}}$, respectively. Since these two matrices are not stochastic matrices as it was the case for PageRank, when finding them by iterations, one has to implement additional normalization of the score vectors. In [26], HITS has been generalized to bipartite graphs with the goal to weaken the score reinforcement among the connected nodes and which improve the algorithm's robustness to noisy links. See an extensive review of eigenvector methods for web information retrieval in [27].

In [28], PageRank has been applied to citations among scientific papers (which naturally constitute a directed unweighted network) to assess the relative importance of papers. The authors argued that readers of scientific papers typically follow paths of length two, corresponding to the damping parameter $\alpha = 0.5$ much lower than the original value of 0.85. Albeit the PageRank score of papers was found to be highly correlated with the number of citations (similarly as the PageRank score of web sites is correlated with the number of incoming hyperlinks), significant outliers from this trend were found and identified as seminal publications. This is because the PageRank score redistribution allows a paper with moderate citation count score high thanks to high citation counts of the papers that cite it. As later argued in [29],

time decay is of crucial importance in the analysis of citation networks because, unlike hyperlinks in the WWW, citations basically cannot be updated after a paper is published. There is also an increasing evidence that time plays an eminent role in the growth of citation networks—see [30] for a recent account. See also [31] for a general overview of our knowledge about citation networks.

The effect of aging of publications is included in the CiteRank algorithm [32] where the uniform teleportation matrix is replaced with $\mathbf{1}_N \boldsymbol{\varrho}^{\mathrm{T}}$ where $\varrho_i = \exp[-t_i/\tau]$, t_i is the age of paper i, and τ is a characteristic decay time. Interestingly, when the correlation between the CiteRank score and the number of recently gained citations is investigated, the optimal damping parameter α is found to be close to the value of 0.5 which was before only hypothesized on the basis of reading habits of researchers. The authors consequently show that apart from selecting papers that contribute most to the current research, CiteRank is particularly successful in selecting papers of long-lasting interest.

Similarly, the network of scientific journals with links weighted by the number of times an article from journal i cites an article from journal j is again suitable for PageRank-like computation of journal status [33]. Albeit the number of citations does not directly enter here, the resulting ranking of journals is similar to that obtained with the so-called impact factor (which is essentially the average number of citations of recent papers in a given journal). The observed differences in these two measures allowed the authors to introduce the categories of popular journals (which have high impact factors but their citations come from lesser journals, hence the resulting PageRank score of the popular journals is comparatively small) and prestigious journals (which have moderate impact factor but their citations come from journals with high PageRank score, thus allowing the prestigious journals to score high too). A publicly available web site SJR runs a slightly different algorithm based on citations among journals to rank scientific value of journals and countries (see www.scimagojr.com) [34].

What is perhaps of even a greater interest to researchers than rankings of papers and journals are rankings of the researchers themselves. The simplest approach to achieve that would be to create a directed networks of authors where links are created according to who cites whom and weight these links according to the citation frequency for a given pair of authors. To better represent the diffusion of scientific credit in such a network, the authors in [35] propose additional weights reflecting the number of authors of the citing and cited paper, respectively. If the citing paper A was authored by n_A authors and the cited paper B was authored by n_B authors, $n_A n_B$ independent links pointing from an author of paper A to an author of paper B are created, each with weight $1/(n_A n_B)$. The credit of individuals is then redistributed over the weighted author–author network in a usual twofold way: part $1 - q$ of i's credit goes to the authors cited by i and part q of i's credit is distributed to all authors according to their productivity. For authors with zero out-strength, it is their whole credit what is distributed to all authors in the network. It is then observed how the resulting ranking of authors changes in time and significant correlations are found between highly ranked authors and important scientific prizes being given to them. A very similar algorithm has been used to rank professional tennis players [36].

Another possible approach to the ranking of researchers is by running a PageRank variant on a so-called coauthorship network which is an undirected network where researchers are connected if they have authored a paper together (it is again natural to weight the connection by the number of papers authored together) [37]. Co-citation networks where authors are connected if they were cited together by a paper were also used as input for a PageRank-based algorithm to obtain a ranking of authors [14].

PageRank has been used also to measure the importance of species in the network of ecological relationships where the loss of a single species can trigger a cascade of extinctions [38]. Upon a minor modification of the input network by introducing a root node which is pointed to by each species and which points back to all "primary producers" (species that do not rely on any other species and produce biomass from inorganic compounds) and setting the damping parameter to one (because nutrients cannot randomly jump among nodes in a food web), the authors were able to use the standard PageRank formula. The obtained importance ranking of species was shown to be very effective in choosing nodes leading to the fastest collapse of the food web, outperforming rankings by betweenness and closeness centrality.

A root node pointed by and pointing to all nodes was used also later in [39] where the PageRank algorithm was used to quantify user influence in a directed social network. It is useful to realize that such a root node in fact serves as a teleportation probability: it leads from a given node to the root node and then in the next step to a randomly chosen normal node. This teleportation probability is node dependent: jump to the ground node occurs with a 50% probability for a node with only one original out-going link, but the probability is only 1% for a node with 99 original out-going links. In addition, this root node causes the transition matrix to be irreducible and primitive which guarantees existence and uniqueness of a stationary solution. Based on the tests on data obtained from the social bookmarking service "Delicious.com," the authors of [39] argue that their variant of PageRank is particularly suitable for social networks as it better detects influential users and it is more resistant to manipulations and noisy data.

2.3 Random Walks with Sources and Sinks

As we have seen above, PageRank is built on a process where the initial node occupancy distribution $\boldsymbol{h}^{(0)}$ is gradually washed away by the random walk and an equilibrium distribution $\boldsymbol{h}^{(\infty)}$ emerges. In some cases, there exist nodes that act as sources or sinks—they constantly emit or absorb, respectively, "particles" that are transported over the network [40]. To allow for termination of the random walk, it is assumed that sources not only emit new particles but also absorb particles arriving in them. Denoting the set of source/sink nodes as S and the set of remaining (transient) nodes as T where $|T| := M$ and thus $|S| = N - M$, we can write the transition matrix in the form

$$\mathsf{P} = \begin{pmatrix} \mathsf{P}_{SS} & \mathsf{P}_{ST} \\ \mathsf{P}_{TS} & \mathsf{P}_{TT} \end{pmatrix} \tag{6}$$

where we have sorted the nodes so that the first $N - M$ nodes are from S and the next M nodes are from T. If S is the set of sinks, then $\mathsf{P}_{ST} = \mathsf{0}$ and $\mathsf{P}_{SS} = \mathsf{I}_{N-M}$. We can now ask what is the probability $F_{ij}(t)$ that a particle originating at $i \in T$ gets absorbed in $j \in S$ in t steps or less, avoiding all other sink nodes on its path. This absorption can either occur in one step, with the probability P_{ij}, or the particle can first go to another transient node k and then be absorbed from there in $t - 1$ steps or less. Together we have

$$F_{ij}(t) = P_{ij} + \sum_{k \in T} P_{ik} F_{kj}(t-1) \tag{7}$$

where, of course, $F_{kj}(0) = 0$ for all k and j. This formula can be written in a matrix form as $\mathsf{F}(t) = \mathsf{P}_{TS} + \mathsf{P}_{TT}\mathsf{F}(t-1)$ where $\mathsf{F}(t)$ is an $M \times (N - M)$ matrix of absorption probabilities. The stationary solution F thus fulfills $\mathsf{F} = \mathsf{P}_{TS} + \mathsf{P}_{TT}\mathsf{F}$, and one can express it as

$$\mathsf{F} = (\mathsf{I}_M - \mathsf{P}_{TT})^{-1}\mathsf{P}_{TS} \tag{8}$$

In the simplest case when $\mathsf{P}_{TT} = \mathsf{0}$ (all links from transient nodes lead directly to sink nodes), we obtain $\mathsf{F} = \mathsf{P}_{TS}$ as expected. One can show that the inverse $(\mathsf{I}_N - \mathsf{P}_{TT})^{-1}$ exists if for every $i \in T$ and $j \in S$, there is a directed path from i to j [40].

The dual problem of particle diffusion from sources can be solved analogously, leading to the average number of times, $H_{ij}(t)$, that a particle originating at a source node i visits a transient node j in t steps or less, without being absorbed in a source node. The final result reads

$$\mathsf{H} = \mathsf{P}_{ST}(\mathsf{I}_M - \mathsf{P}_{TT})^{-1}. \tag{9}$$

Unlike F, a particle can visit a transient node j repeatedly and therefore H_{ij} can be greater than one. The described picture can be generalized to include the possibility of particle dissipation also in transient nodes [40]. There is a close relation between random walks with sinks/sources and currents in electric networks—for details, see [41, 42].

PageRank augmented with sinks was shown to increase the diversity of top ranked items [43]. After the top ranked object is found by ordinary PageRank computation, it is turned into a sink and the second object is selected from the remaining transient nodes as the one that has the longest time to absorption. The selected node is then turned into a sink too, and the third object is again found by the absorption time criterion. Since the expected number of visits of node j when starting with node i is $V_{ij} = [(\mathsf{I}_M - \mathsf{P}_{TT})^{-1}]_{ij}$, the expected absorption time of node i is $t_i = \sum_j V_{ij} = (\mathsf{V}\mathbf{1}_M)_i$. The absorption time maximization leads to the preference for nodes that are far away in the given network from the nodes already selected for the top of the ranking, which provides a stimulus to the diversity of results.

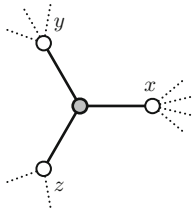

Fig. 2 In random walk, the occupancy probability of the central node in the next time step is $\frac{x}{5} + \frac{y}{4} + \frac{z}{3}$ (where x, y, z are the current occupancy probabilities of the neighboring nodes, respectively). In heat diffusion, the temperature of the central node in the next time step is $\frac{x}{3}+\frac{y}{3}+\frac{z}{3}$ (where x, y, z are the current temperatures of the neighboring nodes, respectively)

We finally note a close connection between random walk and heat diffusion. In random walk, the occupancy probability of a node in the next time step depends on the current occupancy probabilities and degrees of its neighbors. By contrast, in heat diffusion, the temperature of a node in the next time step depends on the current temperatures of its neighbors and the degree of the given node (see Fig. 2 for an illustration). In mathematical terms, while the transition matrix of random walk reads $P_{ij} = A_{ij}/k_i$ and thus P^{T} is column normalized, the matrix converting the current vector of temperature values into a next time step vector reads $O_{ij} = A_{ij}/k_j$ and thus O^{T} is row normalized.

Further connections can be found by studying the emission and absorption processes described above. If we fix a sink node j, the probabilities of absorption in j for particles starting in node i, F_{ij}, satisfy the discrete heat equation on the network. This is easy to see on an unweighted undirected network—given a transient node i and its set of neighbors $\mathcal{N}_i$, we can write similarly as in Eq. (7)

$$F_{ij} = \frac{1}{k_i} \sum_{k \in \mathcal{N}_i} F_{kj}$$

That is, the probability of being absorbed in node j when starting in node i is simply the average over these absorption probabilities when starting in neighbors of node i. The boundary condition is given by the sure absorption in j when starting in j and impossible absorption in j when starting in another sink node (corresponding to the boundary probability values 1 and 0, respectively). Generalization to a weighted or undirected network is straightforward. This duality is illustrated on a toy network in Fig. 3.

2.4 Other Algorithms

Node betweenness in a network is calculated as the fraction of the shortest paths between node pairs that pass through a selected node. If the node lies on many shortest paths, it is assumed to be important for information spreading over the

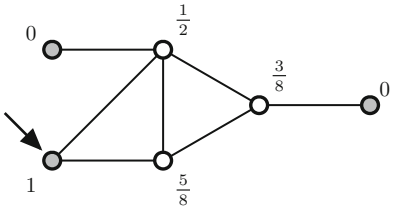

Fig. 3 Random walk with absorption in sink nodes (*shaded with gray*): the probability of being absorbed in the arrow-marked node is shown for each node. These probabilities solve the heat equation with the boundary condition given by the temperature of sink nodes fixed at one (for the *arrow-marked node*) and zero (for all other sink nodes), respectively. For example, the absorption probability $5/8$ for one of the transient nodes can be obtained by averaging the absorption probabilities 1, $1/2$, and $3/8$ of the neighboring nodes

network (e.g., it connects extensive clusters). However, node betweenness considers only the shortest paths and thus neglects a significant part of the network's topology. Random-walk betweenness improves this by considering paths of essentially all lengths, albeit still giving more weight to short ones [42]. It is based on a simple assumption—if random walk starts in node i and ends (gets absorbed) in node j, its contribution to the betweenness of node k is given by the average "net" number of visits of this node during the random walk, $n_k^{(ij)}$. The net number of visits means that passing through a node and then passing through it again from an opposite direction cancel out. Also, if various realizations of random walk are equally likely to pass through a node in opposite directions, these two directions cancel. The resulting betweenness of k is then obtained by averaging the number of visits over start-end node pairs (i, j)

$$b_k = \frac{\sum_{i<j} n_k^{(ij)}}{\frac{1}{2}N(N-1)} \tag{10}$$

where N is the number of nodes in the network. Alternatively, one can obtain the same result building on the electric current injected and removed in a node pair with the contribution to betweenness of node k given by the current passing through this node. The further development is similar to that presented in Sect. 2.3 and ultimately allows to find betweenness values for all nodes in time $O((E + N)N^2)$. This betweenness measure is shown to outperform not only the shortest-path betweenness but also the flow betweenness [42]. With a similar goal, several network flows were typologized and studied by simulations in [44].

A very recent second-order centrality also makes use of random walks but with three distinctions [45]. Firstly, it can be computed in a distributed manner with nodes having only information of who are their neighbors. Secondly, it relies on "unbiased" random walk where the stationary occupancy probability is equal for all nodes regardless of their degree (this is achieved by a Metropolis-Hastings algorithm where step from node i to a neighboring node j is accepted with the probability k_i/k_j for $k_j > k_i$ and always for $k_j \leq k_i$). Finally, it is based on the standard

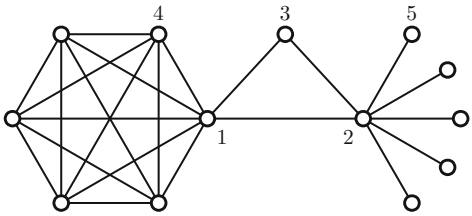

Fig. 4 A toy network for the computation of node centrality (see results in Table 1)

deviation σ_i of the return times to a given node i. The basic idea is that a node with a central position in the network is visited more regularly than peripheral nodes (those are visited in "clusters" with closely grouped subsequent visits interrupted by longer periods when the random walk is in a different part of the network). In addition to numerical stochastic computation of this centrality, various analytical results can be derived and used to better calibrate the numerical implementation.

The network of citations among scientific papers has the special property of being directed and acyclic (the acyclicity is due to citations pointing from a newer paper to older ones). This acyclicity allows one to use the probability of passing through a given node instead of the more traditional occupancy probability. In [46], the probability of passing through node i when the random walk starts in node j, G_{ij}, was proposed to quantify the influence of node i on node j. By summing over j, one consequently obtains the aggregate impact of node i which may be in turn used to rank the nodes. Since aggregate impact of node i correlates with the i's progeny size (by i's progeny we mean the set of nodes from which i can be reached by random walk respecting directions of links), one can better distinguish outstanding nodes by comparing the two characteristics. This passing probability framework has been also used to introduce a new node similarity which is based on the assumption that two nodes are similar if they are both influenced by the same nodes.

To better illustrate performance of the presented methods, we use them to compute node centrality in the network shown in Fig. 4. (Unlabeled nodes have standing identical with that of node 4 or 5.) For the shortest path centrality (also called betweenness centrality), we count also shortest paths where a node lies on the path's beginning or the path's end. For the PageRank score, we use the usual damping value $\alpha = 0.85$. For the random-walk centrality, we follow the prescription given in [42]. For the second-order centrality, we convert the standard deviation of return times σ_i into a centrality value $1/\sigma_i$ (recall that small σ_i is expected for centrally placed nodes). The results summarized in Table 1 show that there are considerable differences between respective centrality measures. While measures agree on a high centrality value of node 1 and a low centrality value of node 5, respectively, big differences exist in assessment of nodes 2, 3, and 4. In particular, eigenvector centrality puts emphasis on the tightly connected part of the network (represented by the complete 6-graph in our toy network) and values little node with low-degree neighbors (in our case, node 2). Random-walk centrality awards the central position of node 3 more than other tested measures which is a direct

Table 1 Centrality values for the network shown in Fig. 4. Values are normalized so that the average centrality is one in all cases

	Node				
Measure	1	2	3	4	5
Degree	1.98	1.98	0.57	1.41	0.28
Shortest path	2.59	3.14	0.66	0.66	0.66
Eigenvector	2.03	0.62	0.52	1.84	0.12
PageRank	1.71	2.65	0.68	1.12	0.47
Random walk	2.31	2.69	1.09	0.84	0.55
Second order	2.23	2.23	0.87	1.17	0.36

consequence of including not only the shortest paths in computation. One can note that degree centrality and second-order centrality rank nodes identically—the value difference between nodes 3 and 4 is however smaller in the case of second-order centrality which is again due to its random-walk origin being able to appreciate the central location of node 3.

3 Recommendation

The task of recommender systems is to utilize past evaluations of items by users to select further items that could be appreciated by the users. We often speak about personalized recommendations because a good recommender system should be able to recognize preferences of individuals and select the object to be recommended accordingly. Thanks to the availability of large-scale data on user behavior and the ever-increasing power of computers at our disposal, the field of recommendation grows rapidly. Nowadays, one can hardly imagine a successful e-commerce site without a sophisticated recommender system (think of Amazon.com) and companies organize competitions aiming to improve their recommendation methods (as it was prominently done by Netflix by their NetflixPrize) [47]. Approaches used to produce recommendations range from variants of the simple thesis "recommend to a user what was already appreciated by similar users" to complicated mathematical models and machine learning techniques [48–50]. The problem of link prediction is closely related to recommendation with the task being to identify possible missing or future links in a given network [51].

In this section, we aim to discuss the use of random walks in recommendation. First of all, similarity measures based on random walks can be used in similarity-based (sometimes called memory-based) collaborative filtering algorithms. Denoting the rating of object α given by user i as $r_{i\alpha}$ and the average rating of user i as μ_i, the generic form of collaborative filtering using user similarity is

$$\tilde{r}_{i\alpha} = \mu_i + \frac{\sum_{j \in R_\alpha} s_{ij} (r_{j\alpha} - \mu_j)}{\sum_{j \in R_\alpha} |s_{ij}|} \tag{11}$$

where $\tilde{r}_{i\alpha}$ is the expected (predicted) rating of object α by user i and R_α is the set of users who have already rated object α. User similarity s_{ij} (or object similarity $s_{\alpha\beta}$ for an item-based variant of collaborative filtering) is usually computed using the standard Pearson similarity or cosine similarity. Our interest now is in random-walk-based similarity measures that can be used instead of traditional ones.

Assuming that random walk starts in node i, one can introduce the average first passage time for node j, $T(j|i)$. The symmetrized quantity $C(i,j) := T(j|i) + T(i|j)$, the average commute time, was shown to act as a distance on the graph and can be further transformed into $\sqrt{C(i,j)}$, a so-called Euclidean Commute Time Distance [52]. In addition, both $C(i,j)$ and $\sqrt{C(i,j)}$ can serve as node similarity measures and in turn effectively used for collaborative filtering. While one can compute $C(i,j)$ on a node-by-node basis using the sink-node machinery described in Sect. 2.3, it is computationally more efficient to employ the formula

$$C(i,j) = 2E\left(l^+_{ii} + l^+_{jj} - 2l^+_{ij}\right) \tag{12}$$

where l^+_{ij} is an element of the Moore-Penrose pseudoinverse L^+ of the network's Laplacian matrix $\mathsf{L} = \mathsf{D} - \mathsf{A}$ (here D is the degree matrix with elements $d_{ij} = k_i\delta_{ij}$) [52]. Pseudoinverse is applied because L cannot be inverted (zero is one of its eigenvalues) and can be computed as $\mathsf{L}^+ = (\mathsf{L} - \mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}/N)^{-1} + \mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}/N$.

A simple node similarity measure based on local random walk was proposed in [53]. Denoting the probability that a random walker starting at node i is located at node j after t time steps as $\pi_{ij}(t)$, the similarity of nodes i and j was proposed in the form

$$s^{LRW}_{ij}(t) = \frac{1}{2E}\left(k_i\pi_{ij}(t) + k_j\pi_{ji}(t)\right) \tag{13}$$

where E is the total number of edges in the graph. Multiplication with node degree (k_i and k_j, respectively) gives more weight to nodes with high degree and compensates for the increased dispersion of random walk at those nodes (if many links lead from x, π_{xy} can be low). The obtained quantity can be summed over different t, leading to "superposed" similarity $s^{SRW}_{ij}(t) = \sum_{\theta=1}^{t} s^{RW}_{ij}(\theta)$. Numerical evaluation on five distinct real networks showed that s^{LRW} and s^{SRW} in most cases outperform traditional similarity metrics in accuracy and are less computationally demanding than other well-performing methods [53]. A method for random-walk computation of paper similarity was proposed specifically for scientific citation data [54]. When computing similarity of papers i and j, two two-step random walks are combined. One aims "downstream" to papers cited by both i and j, thus reflecting the opinion of the authors of i and j. The other aims "upstream" to papers citing both i and j, thus reflecting the opinion of the readers of i and j. It is then shown that this novel similarity measure is able to identify the backbone of the citation network, leading to accurate characterization of hierarchical structure of the scientific development and its classification into fields and subfields.

Due to sparsity of the input data, traditional similarity measures based on overlapping neighborhoods can fail to accurately assess node similarity. To alleviate this problem, it was suggested to transform the similarity matrix into a PageRank-

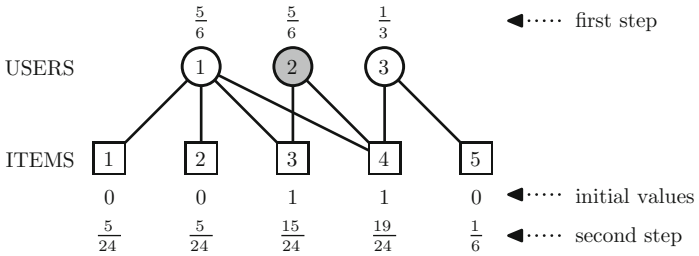

Fig. 5 Illustration of random-walk recommendation for user 2. Items collected by user 2 are initially assigned unit resource which then spreads uniformly to users connected with these items and finally back to the item side. Items with the highest resulting resource amount are then recommended to the given user. In this case, items 1 and 2 score best (items 3 and 4 have higher resulting values but are ignored as they have been already collected by user 2)

like form P by normalization and addition of random jumps and then use $\mathsf{P}(1-\alpha\mathsf{P})^{-1}$ as a new similarity matrix where similarity values are assigned also to item pairs that have not been evaluated by any users [55]. Here $\alpha \in (0,1]$ is the probability of continuing the random walk, and thus $1/\alpha$ is the characteristic number of steps over which similarity is transferred.

Apart from using random walks to quantify node similarity, there are also recommendation methods that are directly based on random walks. In [56], the authors consider the bipartite user-item network where links connect users with the items they collected or appreciated. Note that explicit ratings given by users to items play no role here—the method only requires the knowledge of items that have been collected/favored by individual users. Assuming that each item collected by a given user i is assigned a unit initial resource, this resource is spread uniformly from the collected items to the users connected with them and then in the second step back to items connected with those users (see Fig. 5 for an illustration). The final amount of resource on respective items is then interpreted as their recommendation score and items with the highest score are then recommended to user i (already collected items are of course excluded). The reasoning behind this spreading process is that it selects items that have been collected by users who share some interests with the given user i. At the same time, if user i has collected an extremely popular item α or if a collected item has been co-collected by an extremely active user j, the information signal is weak because the overlap between i and j as well as between i and α is rather small in those cases. The random-walk-based even spreading of the resource is thus a reasonable approach to quantify the resulting recommendation scores.

The transition matrices from objects to users and vice versa have the form $U_{i\alpha} = A_{i\alpha}/k_\alpha$ and $V_{\alpha i} = A_{i\alpha}/k_i$, respectively, where k_α is the degree of item α (the number of users who collected it) and k_i is the degree of user i (the number of items collected by this user). The vector with object recommendation scores for user i then reads $\tilde{\boldsymbol{h}}_i = \mathsf{V}\mathsf{U}\boldsymbol{h}_i$ where $(\boldsymbol{h}_i)_\alpha = A_{i\alpha}$ encodes which items have been actually collected by user i. One can introduce $\mathsf{W}^P := \mathsf{V}\mathsf{U}$ and show that

$$W^P_{\alpha\beta} = \frac{1}{k_\beta} \sum_{i=1}^{U} \frac{A_{i\alpha} A_{i\beta}}{k_i} \tag{14}$$

where indices α and β are used to enumerate items, i enumerates users, and U is the total number of user nodes. One can also spread the initial resource over $2n$ steps in the bipartite network by $(\mathsf{W}^{\mathsf{P}})^n \boldsymbol{h}_i$, but the result converges fast to a vector whose elements are proportional to object degree k_α and hence conveys little information for personalized recommendation.

This basic method has been subsequently generalized in multiple ways. For example, it was proposed to assign the initial amount of resource to items not uniformly but depending on the item degree as k_α^θ [57]. Best results were achieved with $\theta \approx -1$ when the produced recommendations were both more accurate and more personalized. To better answer the need for diversity in recommendations, a hybrid algorithm was proposed which combines the random-walk algorithm with heat spreading [58]. As we have already seen, heat diffusion differs from random walk in normalization of their matrices and thus the matrix of heat diffusion reads $W^H_{\alpha\beta} = (1/k_\alpha) \sum_{i=1}^{U} A_{i\alpha} A_{i\beta} / k_i$. The best performing hybrid of the two has the form

$$W^{P+H}_{\alpha\beta} = \frac{1}{k_\alpha^{1-\lambda} k_\beta^{\lambda}} \sum_{i=1}^{U} \frac{A_{i\alpha} A_{i\beta}}{k_i} \tag{15}$$

where the parameter λ controls the balance between the contribution of random walk and heat spreading. This method was shown to simultaneously increase accuracy and diversity of recommendations.

A combination of random walk and heat diffusion for data with explicit ratings was presented in [59] where recommendation scores obtained by each respective process are multiplied to obtain the final recommendation score. In addition, the employed random walk is self-avoiding, i.e., there is no possibility to return to the initial item node after two steps. If user evaluations are given in an integer scale (a very typical case nowadays), a multichannel spreading can be employed where the states of the random walk are represented not only by the current item but also by the rating given to this item [60]. If, for example, a five-level rating scale is used, 5×5 connections are created between any two items. However, this approach suffers from aggravating the data sparsity problem (the same amount of data is used to construct many more connections between (item, rating) pairs) which limits its performance.

Spreading over a bipartite network is considered also in [61] where the bipartite user-item network is augmented with social links among users (this kind of data is often produced in online gaming). The random walk starting at the user for which recommendations are being made follows a social link to another user with probability α or a link to an item with probability $1 - \alpha$ where it is absorbed. Items are then ranked according to the fraction of random walks absorbed in them. A different mechanism of heat diffusion on an item–item network was used to produce recommendations by representing items liked and disliked by a given user as nodes

with fixed temperature 1 and 0, respectively [62]. From the remaining nodes, those with the highest resulting temperature are then chosen to be recommended to the given user. See [50, Chap. 6] for other related works and more detailed information.

4 Conclusion

We attempted here to give an overview of applications of random walks to information filtering, focusing on the tasks of ranking and recommendation in particular. Despite the amount of work done in these two directions, multiple important research challenges still remain open. Due to the massive amounts of available data, scalability of algorithms is of critical importance. Even when full computation is possible, one can think of potential approaches to update the output gradually when new data arrives. To achieve that, one can use or learn from perturbation theory which is a well-known tool in physics. We have seen that results based on random walks often correlate strongly with mere popularity (represented by degree) of nodes in the network. Such bias toward popularity may be beneficial for an algorithm's accuracy, but it may also narrow our view of the given system and perhaps create a self-reinforcing loop further boosting popularity of already popular nodes. We thus need information filtering algorithms that converge less to the center of the given network. Random walks biased by node centrality or time information about nodes and links could provide a solution to this problem. As a beneficial side effect, this line of research could yield algorithms pointing us to fresh and promising content instead of highlighting old victors over and over again.

Acknowledgement This work was partially supported by the Swiss National Science Foundation Grant No. 200020-132253. I wish to thank a number of wonderful friends and colleagues who helped to shape many of the ideas presented here.

References

[1] U. Hanani, B. Shapira, P. Shoval, Information filtering: Overview of issues, research and systems. User Model. User Adapted Interact. **11**, 203–259 (2001)
[2] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. (Springer, New York, 2001)
[3] I.H. Witten, E. Frank, M.A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn. (Morgan Kaufmann/Elsevier, San Francisco, 2011)
[4] M. Franceschet, PageRank: Standing on the shoulders of giants. Comm. ACM **54**, 92–101 (2011)
[5] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine. Comput. Network ISDN Syst. **30**, 107–117 (1998)
[6] R. Burke, Hybrid web recommender systems, in *The Adaptive Web: Methods and Strategies of Web Personalization*, ed. by P. Brusilovsky, A. Kobsa, W. Nejdl (Springer, Heidelberg, 2007)

[7] L. Costa, F. da, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Characterization of complex networks: A survey of measurements. Adv. Phys. **56**, 167–242 (2007)
[8] A. Barrat, M. Barthelemy, A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, New York, 2008)
[9] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Huang, Complex networks: structure and dynamics. Phys. Rep. **424**, 175–308 (2006)
[10] S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994)
[11] P. Bonacich, P. Lloyd, Eigenvector-like measures of centrality for asymmetric relations. Soc. Network **23**, 191–201 (2001)
[12] P. Berkhin, A survey on PageRank computing. Internet Math. **2**, 73–120 (2005)
[13] N. Perra, S. Fortunato, Spectral centrality measures in complex networks. Phys. Rev. E **78**, 036107 (2008)
[14] Y. Ding, E. Yan, A. Frazho, J. Caverlee, PageRank for ranking authors in co-citation networks. J. Am. Soc. Inform. Sci. Tech. **60**, 2229–2243 (2009)
[15] A. Hotho, R. Jäschke, C. Schmitz, G. Stumme, Information retrieval in folksonomies: search and ranking, in *Lecture Notes in Computer Science*, vol. 4011, ed. by Y. Sure, J. Domingue, pp. 84–95 (2006)
[16] A. Agarwal, S. Chakrabarti, S. Aggarwal, Learning to rank networked entities, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)* (ACM, New York, 14–23 2006)
[17] A.N. Langville, C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, Princeton, 2006)
[18] L. Bing, *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data* (Springer, Heidelberg, 2007)
[19] T.H. Haveliwala, Efficient computation of PageRank. Technical Report, Stanford University Database Group, http://ilpubs.stanford.edu:8090/386/ (1999)
[20] A. Cheng, E. Friedman, Manipulability of PageRank under Sybil strategies, in *Proceedings of the First Workshop on the Economics of Networked Systems (NetEcon06)*, Ann Arbor, 2006
[21] G. Ghoshal, A.-L. Barabsi, Ranking stability and super-stable nodes in complex networks. Nat. Comm. **2**, 394 (2011)
[22] R. Baeza-Yates, P. Boldi, C. Castillo, Generalizing PageRank: damping functions for link-based ranking algorithms, in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)* (ACM, New York, 2006)
[23] S.D. Kamvar, M.T. Schlosser, H. Garcia-Molina, The eigentrust algorithm for reputation management in P2P networks, in *Proceedings of the 12th International Conference on World Wide Web (WWW'03)* (ACM, New York, 2003)
[24] G.D. Paparo, M.A. Martin-Delgado, Google in a quantum network. Sci. Rep. **2**(444), arXiv.org/abs/1112.2079. http://www.nature.com/srep/2012/120608/srep00444/full/srep00444.html?WT.mc_id=FBK_SciReports (2012)
[25] J.M. Kleinberg, Authoritative sources in a hyperlinked environment. J. ACM **46**, 604–632 (1999)
[26] H. Deng, M.R. Lyu, I. King, A generalized Co-HITS algorithm and its application to bipartite graphs, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)* (ACM, New York, 2009)
[27] A.N. Langville, C.D. Meyer, A survey of eigenvector methods for web information retrieval. SIAM Rev. **47**, 135–161 (2005)
[28] P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with Google's PageRank algorithm. J. Informetrics **1**, 8–15 (2007)
[29] S. Maslov, S. Redner, Promise and pitfalls of extending Google's PageRank algorithm to citation networks. J. Neurosci. **28**, 11103–11105 (2008)
[30] M. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks. Phys. Rev. Lett. **107**, 238701 (2011)

[31] F. Radicchi, S. Fortunato, A. Vespignani, Citation networks, in *Models of Science Dynamics, Understanding Complex Systems*, ed. by A. Scharnhorst, et al. (Springer, Berlin Heidelberg, 2012)
[32] D. Walker, H. Xie, K.K. Yan, S. Maslov, Ranking scientific publications using a model of network traffic. J. Stat. Mech. **6**, P06010 (2007)
[33] J. Bollen, M.A. Rodriguez, H. Van de Sompel, J. Status. Scientometrics **69**, 669–687 (2006)
[34] B. Gonzlez-Pereiraa, V.P. Guerrero-Bote, F. Moya-Anegn, A new approach to the metric of journals scientific prestige: The SJR indicator. J. Informetrics **4**, 379–391 (2010)
[35] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Diffusion of scientific credits and the ranking of scientists. Phys. Rev. E **80**, 056103 (2009)
[36] F. Radicchi, Who is the best player ever? A complex network analysis of the history of professional tennis. PLoS ONE **6**, e17249 (2011)
[37] E. Yan, Y. Ding, Discovering author impact: A PageRank perspective. Inform. Process. Manag. **47**, 125–134 (2011)
[38] S. Allesina, M. Pascual, Googling food webs: can an eigenvector measure species' importance for coextinctions? PLoS Comput. Biol. **5**, e1000494 (2009)
[39] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case. PLoS ONE **6**, e21202 (2011)
[40] A. Stojmirović, Y.-K. Yu, Information flow in interaction networks. J. Comput. Biol. **14**, 1115–1143 (2007)
[41] P.G. Doyle, J.L. Snell, Random walks and electric networks. *Carus Mathematical Monographs*, vol. 22 (Mathematical Association of America, Washington, 1984)
[42] M.E.J. Newman, A measure of betweenness centrality based on random walks. Soc. Network **27**, 39–54 (2005)
[43] G.-L. Lin, H. Peng, Q.-L. Ma, J. Wei, J.-W. Qin, Improving diversity in Web search results re-ranking using absorbing random walks, in *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC'10)* (IEEE, 2116–2421 2010)
[44] S.P. Borgatti, Centrality and network flow. Soc. Network **27**, 55–71 (2005)
[45] A.-M. Kermarrec, E. Le Merrer, B. Sericola, G. Trdan, Second order centrality: Distributed assessment of nodes criticity in complex networks. Comput. Comm. **34**, 619–628 (2011)
[46] S. Gualdi, M. Medo, Y.-C. Zhang, Influence, originality and similarity in directed acyclic graphs. EPL **96**, 18004 (2011)
[47] J.B. Schafer, J.A. Konstan, J. Riedl, E-commerce recommendation applications. Data Min. Knowl. Discov. **5**, 115–153 (2001)
[48] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**, 734–749 (2005)
[49] F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (eds.), *Recommender Systems Handbook* (Springer, New York, 2011)
[50] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems. Phys. Rep. **519**, 1–49. arXiv.org/abs/1202.1112 (2012)
[51] L. Lü, T. Zhou, Link prediction in complex networks: a survey. Phys. A **390**, 1150–1170 (2011)
[52] 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**, 355–369 (2007)
[53] W. Liu, L. Lü, Link prediction based on local random walk. EPL **89**, 58007 (2010)
[54] S. Gualdi, C.H. Yeung, Y.-C. Zhang, Tracing the evolution of physics on the backbone of citation networks. Phys. Rev. E **84**, 046104 (2011)
[55] H. Yildirim, M.S. Krishnamoorthy, A random walk method for alleviating the sparsity problem in collaborative filtering, in *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08)* (ACM, New York, 2008)
[56] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation. Phys. Rev. E **76**, 046115 (2007)

[57] T. Zhou, L.-L. Jinag, R.-Q. Su, Y.-C. Zhang, Effect of initial configuration on network-based recommendation. EPL **81**, 58004 (2008)
[58] 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. USA **107**, 4511–4515 (2010)
[59] M. Blattner, B-rank: A top N recommendation algorithm, in *Proceedings of the 1st International Multi-Conference on Complexity, Informatics and Cybernetics*, pp. 336–341, 2010
[60] Y.-C. Zhang, M. Medo, J. Ren, T. Zhou, T. Li, F. Yang, Recommendation model based on opinion diffusion. EPL **80**, 68003 (2007)
[61] A.P. Singh, A. Gunawardana, C. Meek, A.C. Surendran, Recommendations using absorbing random walks, in *Proceedings of the North East Student Colloquium on Artificial Intelligence*, 2007
[62] Y.-C. Zhang, M. Blattner, Y.-K. Yu, Heat conduction process on community networks as a recommendation model. Phys. Rev. Lett. **99**, 154301 (2007)