Time-aware Ranking in Dynamic Citation Networks

Rumi Ghosh*, Tsung-Ting Kuo‡, Chun-Nan Hsu*, Shou-De Lin‡ and Kristina Lerman*
*Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA
†Department of Computer Science, University of Southern California, Los Angeles, CA 90007, USA
‡Department of Computer Sci. & Info. Engg., National Taiwan University, Taipei 106, Taiwan
§Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
Email: chunnan@iis.sinica.edu.tw

***Abstract*—Many algorithms have been developed to identify important nodes in a complex network, including various centrality metrics and PageRank, but most fail to consider the dynamic nature of the network. They therefore suffer from recency bias and fail to recognize important new nodes that have not had as much time to accumulate links as their older counterparts. This paper describes the Effective Contagion Matrix (ECM), a solution to address recency bias in the analysis of dynamic complex networks. The idea of ECM is to explicitly consider the temporal order of links and chains of links connecting to a node with some temporal decay factors. We tested ECM with three large real world citation networks on the task of predicting papers' future importance. We compared ECM's performance with two static metrics, degree-centrality and PageRank, and two time-aware metrics, age-based PageRank and CiteRank. We show that ECM is more appropriate for predicting future citations and PageRank scores with regard to new citations. We also describe a procedure to estimate ECM's parameters from the data. Combining all five scores into a ν-SVR regression model of future citations improves the predictive performance further.**

***Keywords*-data mining; network analysis; dynamic networks; citation networks;**

I. Introduction

The structure of many complex networks is not static, but evolves over time as underlying microscopic processes create or destroy nodes and edges. While many algorithms have been proposed to analyze network structure to identify important nodes or hidden groups, they often fail to take the dynamic nature of the network into account. The PageRank algorithm [1], for example, ranks Web pages by analyzing the structure of hyperlinks between them. Although PageRank has been enormously successful both technically and commercially, it has limitations. While PageRank treats the Web as a static network, in reality its structure changes over time as new pages and hyperlinks are created or destroyed. As a result, PageRank will generally judge newer pages, which have not had as much time to accumulate hyperlinks, to be less important than older pages [2], even though Web users are often interested in the recency of information [3]. Another example of a dynamic network is the citation network, which grows every year as new scientific articles are published which cite existing articles. Like the Web, the structure of the citation network can be analyzed to rank scientists [4] and find important scientific papers [5]–[7]. Here too, static metrics like PageRank [8], [9] and citations count are known to be biased against recent papers, although just like Web users, researchers seeking important scientific articles are more interested in recent documents.

While several time-aware metrics have been proposed to address the recency bias of PageRank [2], they fail to take the dynamic nature of the network into account. Typically, these approaches extend PageRank simply by initiating the random walk from a node that is chosen with probability that depends on its age [5]. The random walk, however, is carried out on the static graph. We claim that by ignoring the temporal order of links, these approaches lose important information about the structure of dynamic networks. Recently, Lerman et al. [10] proposed a new centrality metric for dynamic networks. This metric generalizes α-Centrality metric [11], [12], which measures centrality of a node by the number of paths, of any length, that connect it to other nodes. The dynamic centrality metric exploits an intuition that in order for a message sent by one node in a network to reach another after some period of time, there must exist a path that connects the source and destination nodes through intermediaries at different times. However, like PageRank, this metric too is biased in favor of older nodes which had more time to accumulate links.

In this paper we propose a time-aware version of dynamic centrality which properly discounts older papers while still taking the dynamic nature of the network into account. We evaluate our approach on large real-world scientific papers citation networks by seeing how well it predicts papers' future importance. We show that time-aware dynamic centrality metrics are more appropriate for identifying important papers that will attract more citations in the future.

II. Time-aware Ranking

Given a dynamic network $G(V(t), E(t))$, for example, a citation network where $V(t)$ are papers and $E(t)$ are directed links indicating the direction of citation, consider t to be the smallest time interval in which there is no change in the topology of the dynamic network. In the citation network, we take t to be one year. Figure 1 shows an example citation network as a reel to emphasize the temporal

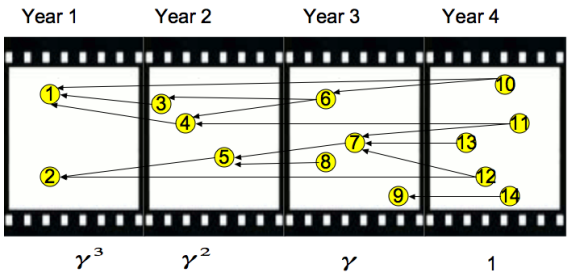


Figure 1. An example citation network.

nature of citations. Each frame within the reel corresponds to a year and the nodes within the frame represent the papers published that year. Let the citation network under consideration comprise of papers published over the period of time $t_1 < \cdots < t_k < \cdots < t_N$. At the end of year t_N, there will be M papers. Therefore, a paper p_j, published at time t_1, that is cited by paper p_i, published at time t_2, then necessarily $t_1 < t_2$. Let $t(p_i)$ be the year paper p_i is published. The adjacency matrix corresponding to this citation network comprising of all papers published up to year t_n is an $M \times M$ matrix A_n is:

$$\begin{aligned} A_n[i,j] &= 1 \ \textit{if paper } p_i \textit{ cites } p_j \textit{ and } t(p_i) \leq t_n \\ &= 0 \ \textit{otherwise} \end{aligned} \tag{1}$$

e

A researcher can find papers by following citations links back in time from a particular paper. Let the probability of following a citation link be α. Then $\alpha A_N[i,j]$ gives the probability of reaching the paper p_j following the citations of paper p_i ($t(p_i) \leq t_N$). The term $\alpha^2 A_N A_{N-1}[i,j]$ gives the expectation of reaching paper p_j from paper p_i through a one-hop paths. Each one-hop path can be described as a chain in which paper p_i cites some paper p_k, which in turn cites paper p_j. Continuing with this process the n-hop paths from paper p_i to paper p_j would be given by $\alpha^n \Pi_{k=1}^{n} A_{N+1-k}$. Therefore, the expectation of following citation links from paper p_i and reaching p_j is given by $C_{N,\alpha}[i,j]$, where $C_{N,\alpha}$ is the *contagion matrix* described by:

$$\begin{aligned} C_{N,\alpha} &= \alpha^{N-1} A_N \cdots A_3 A_2 \\ &+ \alpha^{N-2} A_N A_{N-1} \cdots A_3 \\ &+ \cdots + \alpha^2 A_N A_{N-1} + \alpha A_N \end{aligned} \tag{2}$$

The term $C_{N,\alpha}[i,j]$ gives the number of paths from paper p_i to p_j attenuated by their length.

The more paths there are from paper p_i to p_j, the higher the likelihood that a researcher will find p_j by following citation chains from p_i. The expression above is similar to the α-Centrality [13] and the Katz score [14] metrics used in social network analysis. These metrics rank nodes by the total number of paths connecting them to other nodes, exponentially attenuated by their length, so that shorter paths contribute more to ranking than longer paths.

In a citation network, we assume that no paper cites itself and there are no citations of papers in the same year. Since, A_N is a lower triangular matrix (with diagonal elements 0), $C_{N,\alpha}$ reduces to

$$C_{N,\alpha} = \sum_{i=1}^{N-1} \alpha^i A_N^i \tag{3}$$

A. Retained Adjacency Matrix (RAM)

The contagion matrix weighs all edges equally, whether they link to newer papers or older papers. Therefore, there will be more paths to older papers, giving them greater importance. However, people often prefer more recent information [3]. We capture this effect by defining a time-aware adjacency matrix, called *retained adjacency matrix* (RAM), that takes the recency of cited papers into account. We use parameter $\gamma < 1$ to give a greater weight to a more recent paper, and the weight attached to a paper decreases as the paper ages. If v is the value associated with a citation link for a paper published in year t_n, a scaled down value $\gamma^{n_i} v$ is the value associated with a citation link paper published in year t_{n-n_i}. Thus, a greater weight is given to a paper published in year t_n than to earlier papers. Given a time interval $[t_1, t_N]$, the retained adjacency matrix of the network[1] up to time t_n is:

$$\begin{aligned} R_{n,\gamma}(i,j) &= \gamma^{N-n_i} \ \textit{if } p_i \textit{ cites } p_j \textit{ and } t(p_i) = t_{n_i} \leq t_n \\ &= 0 \ \textit{otherwise}. \end{aligned} \tag{4}$$

where parameter γ is the retention probability.

B. Effective Contagion Matrix (ECM)

Using retained adjacency matrix (Eq. 4), we can rewrite the contagion matrix (Eq. 2 and 3) as:

$$EC_{N,\alpha,\gamma} = \sum_{i=1}^{N-1} \alpha^i R_N^i \tag{5}$$

This *effective contagion matrix* (ECM) measures the number of citation chains between papers, with the chains attenuated not only by their length (with parameter α), but also by the age of the citing papers (with parameter γ). Thus, older papers are de-emphasized in the ECM matrix.

We use ECM to score papers at the end of a time period $[t_i, t_N]$ and rank them according to their scores. The score of a paper p_j is given by $EC_N(j) = \sum_i EC_N(i,j)$. The greater the number of papers citation chains incident on p_j, the greater its influence.

[1]The time of reference or the time when calculations are being performed is t_N.

C. Efficiently Computing ECM

By storing the matrix as an adjacency list and through efficient bookkeeping, the time and space complexity of the algorithm can be reduced considerably. We have devised the *Block Expansion Algorithm*, to compute the ECM rankings, with a runtime complexity of $O(|E||V|N)$. The algorithm is an incremental version that assumes that data becomes available year-by-year. Alternatively, assuming we have complete citation information up to year N, we can simply apply the definitions above to compute rankings up to year N.

III. Related Time-aware Metrics

In this section, we review related metrics that we will compare with in our experimental evaluation of ECM.

A. Degree Centrality

Centrality determines node's importance in a network. This measure is dependent on the network structure. The simplest centrality metric, degree centrality, measures the number of edges that connect a node to other nodes in a network.

B. Time-aware PageRank Metrics

PageRank [1] is well known for its use in ranking Web search results. In PageRank, the importance of a Web page i is measured as the probability of a random visit to the page:

$$\text{PR}_i = \frac{\alpha}{T} + (1-\alpha) \sum_{i=1, j \neq i}^{k} \frac{\text{PR}_j}{L_j}$$

where T is the total number of Web pages, α is the probability of leaving a page (usually set as 0.15), each j is a Web page that contains a hyperlink to i, and L_j is the number of outgoing hyperlinks in j.

Age-based PageRank [2] is a variant of PageRank that applies an exponential degradation function $f(\text{age}) = (1 + a \cdot \exp^{-b \cdot \text{age}})$ of a node's age in the computation of its PageRank score:

$$\text{PR}_i = \frac{\alpha}{T} + (1-\alpha) f(\text{age}_i) \sum_{i=1, j \neq i}^{k} \frac{\text{PR}_j}{L_j},$$

where age_i is the age of page i.

CiteRank [5] was designed specifically for ranking papers in a citation network. CiteRank performs a random walk on an aggregated citation graph, but initiates the walk from a recent paper chosen with probability that depends on its age. Authors estimated parameters of the random walk by fitting papers' CiteRank score to the number of citations accrued by papers over some time period. Let W be a transfer matrix with elements $W_{ij} = 1/L_j$ if paper j cites i and 0 otherwise. The probability that a researcher follows the citation links to encounter a paper is defined as

$$\vec{T} = I \cdot \vec{\rho} + (1-\alpha) W \cdot \vec{\rho} + (1-\alpha)^2 W^2 \cdot \vec{\rho} + \ldots$$

where $\rho_i = \exp^{-\text{age}_i/\tau_{dir}}$ is the probability of initially selecting paper i, age_i is the age of the paper and τ_{dir} characteristic decay time.

C. Regression Models

It is also possible to combine different ranking metrics by constructing a regress model from the citation network. In this paper, we consider ν-SVR, a support vector regression model [15]. The idea is to fit a function that maps a given paper to the number of citations to the paper in the future. The input feature vector representing a paper consists of its scores by different importance metrics.

IV. Experimental Results

A. Data Sets and Metrics

We considered a citation data set that consists of articles uploaded to the theoretical high energy physics (Hep-Th) section of the *arXiv* preprints server from 1992 to April, 2003 (http://snap.stanford.edu/data/cit-HepTh.html). Each article is identified by a unique number, with first two digits representing the year of submission. Data was cleaned by removing citations to articles that appeared in the future, as well as citations of the article to itself. There are 1,044 citations in the Hep-Th data set in this category.

We partitioned the data by year to construct snapshots of the dynamic network in consecutive years. The citations made by papers uploaded to *arXiv* during some year form the edges of the snapshot for that year. A year may not be an optimal partition of the data, since a small number of articles published in one year cite others published in the same year, but it is a convenient time scale to measure scientific production and interaction between researchers. We also considered the phenomenology section (Hep-Ph) from the same source (http://snap.stanford.edu/data/cit-HepPh.html) and processed the data similarly.

The American Physical Society (APS) data set is one of the largest citation networks available (https://publish.aps.org/datasets). This data set consists of 450,000 articles published in *Physical Review Letters*, *Physical Review*, and *Reviews of Modern Physics* and dates back to 1893. We removed 408 articles without publication dates and 615 forward citations. "APS Part" is a subset of the Physical Review citation network that contains only papers published in the recent 20 years (1989-2009). During that period, more than 64% of all 116 years of papers were published, an exponential boom of publications. Table I shows the statistics of these citation network data sets.

We compared 11 metrics in our experimental evaluation as shown in Table II. To reduce the cost of matrix multiplication in the computation of ECM, an additional parameter t was introduced as a threshold to reset values in the matrix that were below t. We also considered three support vector regression models, one of them combines the above metrics except both RAM and ECM, one except RAM, and one

Table I
BASIC STATISTICS OF THE DATA SETS.

Type	Hep-Th	Hep-Ph	APS Part	APS
Node	27770	34546	290286	449678
Link	352807	421578	2605644	4707689
Year	11	11	20	116

Table II
METRICS COMPARED.

Metric	**Remark**
D	degree = ID + OD
ID	in-degree
OD	out-degree
PR	PageRank [1]
ABPR	Age-based PageRank [2]
CR	CiteRank [5]
ECM	Effective Contagion Matrix (Sec. II-B)
RAM	Recency Adjacency Matrix (Sec. II-A)
SVR	ν-SVR [15] w/ D,ID,OD,PR,ABPR
SVRECM	ν-SVR [15] w/ D,ID,OD,PR,ABPR,ECM
SVRRAM	ν-SVR [15] w/ D,ID,OD,PR,ABPR,RAM

except ECM. For all models, we chose linear kernel and 100 as its cost, tuned by minimizing the mean square error on the training sets.

B. Correlation of Rankings

Each metric ranks the oldest 90% of the papers in a data set, based on the citations between them. Then we used the citations from the remaining 10% to the old papers to measure how well the metrics rank the importance of the old papers by the following criteria:

- `Cite`: compute the Spearman correlation coefficient between the ranked list by the scores produced by each metric and the ranked list by future citation counts. This is similar to the evaluation criterion given in the CiteRank paper [5].
- `FutNew`: compute the PageRank scores of the network containing all papers but only new citations, then evaluate the Spearman correlation coefficient between the ranked list by the scores from each metric and the ranked list of the PageRank scores. This is the criterion used in the FutureRank paper [7].
- `FutAll`: Similar to `FutNew`, but evaluate the PageRank scores for the entire network that contains all papers and all citations.

These criteria were chosen to measure if a metric can rank a paper by its potential of attracting *new* citations. In contrast, some of the previous works defined an importance node as the status of a node in the *current* network.

We now report the experimental results using the above three evaluation criteria for each citation network data set.

Hep-Th Table III shows the evaluation results using the three criteria the ten importance metrics. Initially, we used the parameters suggested by the authors of the corresponding metrics. These parameters led to good performance for certain criteria but not all. For example, usually we set $\alpha = 0.15$ for PageRank. These parameter settings were derived for various purposes that are not necessarily the same as the criteria given here. We thus tuned another set of parameters for each metric to maximize the average of all criteria and gave the performance results in multiple rows for comparison. We performed a similar tuning for ECM and RAM but we selected one that maximizes the average of `Cite` and `FutNew` only because to obtain the best performance in terms of `FutAll`, γ needs to be set to close to one, which will reduce RAM to be equivalent to ID (in-degree) and its performance for the other two criteria will be as poor as ID. Then we integrated the scores produced by the metrics with the balanced performing settings of the parameters to train three SVR models.

The results show that SVRECM performed the best for matching the frequency of future citations (`Cite`) and matching the PageRank scores in the new network (`FutNew`). Combining existing metrics, SVR performed reasonably well but still worse than those SVR models with either ECM and RAM. Other than the three supervised ensemble models, ECM performed the best followed by RAM and then CiteRank, suggesting that degrading weights of aging citations to remove recency bias is effective as both ECM and RAM performed well. That ECM outperformed RAM suggests that considering chains of citations is useful for ranking the potential of papers. Age-based PageRank and PageRank performed much better than all other metrics when the PageRank scores were measured based on the entire network. This suggests that in fact the PageRank order of the papers does not change much as more papers were added to the network, where old papers will enjoy biased preference. Yet though Age-based PageRank represents an attempt to address this issue, its performance in terms of `Cite` and `FutNew` is low, worse than the degree centrality metrics.

Interestingly, in-degree also performed well in terms of `FutAll` while out-degree, when combined with in-degree, performed better than in-degree along in terms of `Cite` and `FutNew`. Therefore, out-degree may correlate with the potential of a paper being cited in the future. The correlation might due to out-degree's role in forming citation chains modeled explicitly by ECM.

Figure 2 shows the results obtained by partitioning papers in different proportions. Recall that the results shown in Table III were obtained by using the oldest 90% of the papers for ranking and the remaining 10% for evaluation. In terms of `Cite`, ECM, CiteRank and Degree improved their performance and hit peaks when the partition is 70% to 30% because for these metrics to work, we need a sufficiently

Table III
COMPARISON ON CORRELATION RESULTS FOR HEP-TH DATA SET.

Metric	Parameter	Cite	FutNew	FutAll
D	N/A	0.5500	0.4802	0.4786
ID	N/A	0.4643	0.4169	0.8617
OD	N/A	0.4037	0.3372	0.0007
PR	$\alpha = 0.15$	0.2760	0.2620	0.9737
PR	$\alpha = 0.48$	0.2889	0.2763	0.9709
ABPR	$\alpha = 0.15$ $a = 0.3, b = 0.005$	0.2775	0.2635	0.9741
ABPR	$\alpha = 0.48$ $a = 3.0, b = 0.0001$	0.2932	0.2778	0.9754
CR	$\alpha = 0.48, \tau_{dir} = 1$	0.6003	0.5812	0.4629
CR	$\alpha = 0.31, \tau_{dir} = 1.6$	0.5946	0.5739	0.5364
ECM	$\alpha = 0.1, \gamma = 0.3$ $t = 0.01$	0.6460	0.6008	0.4805
RAM	$\gamma = 0.3$	0.6187	0.5719	0.6137
SVR	$c = 100$	0.6486	0.6061	0.5540
SVRECM	$c = 100$	0.6685	0.6222	0.4654
SVRRAM	$c = 100$	0.6679	0.6211	0.4924

large citation network for ranking and sufficiently many papers in the future to provide abundant opportunities of new citations. For PageRank and Age-based PageRank, their performance degrades gradually. The trend is similar for FutNew. Contrastingly, PageRank and Age-based PageRank improve in terms of FutAll while other metrics degrade. We plotted similar charts for the results obtained by partitioning papers by years and obtained similar curves.

Hep-Ph Table IV shows the results for the Hep-Ph data set. For the sake of conciseness, only results using balanced parameters are given. The winners and losers are similar to those for Hep-Th, except that though ECM still outperformed RAM, SVRRAM performed better than SVRECM. We also plotted charts similar to those given in Figure 2 and observed similar curves.

American Physical Society data set We then compared the metrics with the Physical Review citation network, which spans more than a hundred years and is larger than the other two data sets by an order of magnitude. Due to its large size, it is difficult to search for optimal parameters by exhaustively enumerating combinations. Instead, we developed an approach to estimate optimal parameters by fitting a power law distribution curve.

To estimate α, we find the distribution of citation chains that span consecutive years. In other words, we set $\gamma = 0$, so that no older citations are retained. N_j gives the total number of chains of length j that start in year t_{n-j+1} and end in year t_n. Assuming that the probability of picking a chain is proportional to the probability of transmitting a message along the chain, N_j decays geometrically with α. Therefore, the probability of choosing a citations chain of length j is given by α^j. The expected number of citation chains is $E(N_j) = \alpha E(N_{j-1})$.

Table IV
COMPARISON OF CORRELATION RESULTS FOR HEP-PH DATA SET.

Metric	Parameter	Cite	FutNew	FutAll
D	N/A	0.5521	0.4966	0.5515
ID	N/A	0.4857	0.4464	0.9007
OD	N/A	0.3267	0.2830	-0.0400
PR	$\alpha = 0.48$	0.3381	0.3204	0.9754
ABPR	$\alpha = 0.48$ $a = 3.0, b = 0.0001$	0.3468	0.3286	0.9763
CR	$\alpha = 0.31, \tau_{dir} = 1.6$	0.5819	0.5703	0.5785
ECM	$\alpha = 0.1, \gamma = 0.3$ $t = 0.01$	0.6482	0.6106	0.5475
RAM	γ = 0.3	0.6275	0.5890	0.6789
SVR	$c = 100$	0.6358	0.6032	0.6441
SVRECM	$c = 100$	0.6776	0.6420	0.5784
SVRRAM	$c = 100$	0.6808	0.6447	0.5704

To estimate γ, we assume that citation retention probability decays geometrically with time [8]. Let C_k^j be the number of papers at time $j - k$ cited by papers at time j. Since the number of citations increases in time, we calculate $W_k^j = C_k^j / \sum_k C_k^j$, the fraction of papers appearing at time $j - k$ that are cited by papers at time j. Taking the average of W_k^j for all j, gives the expected fraction of citations in a given paper to papers published k years before it, $E(W_k)$. Therefore according to our hypothesis, $E(W_k) = \gamma E(W_{k-1})$.

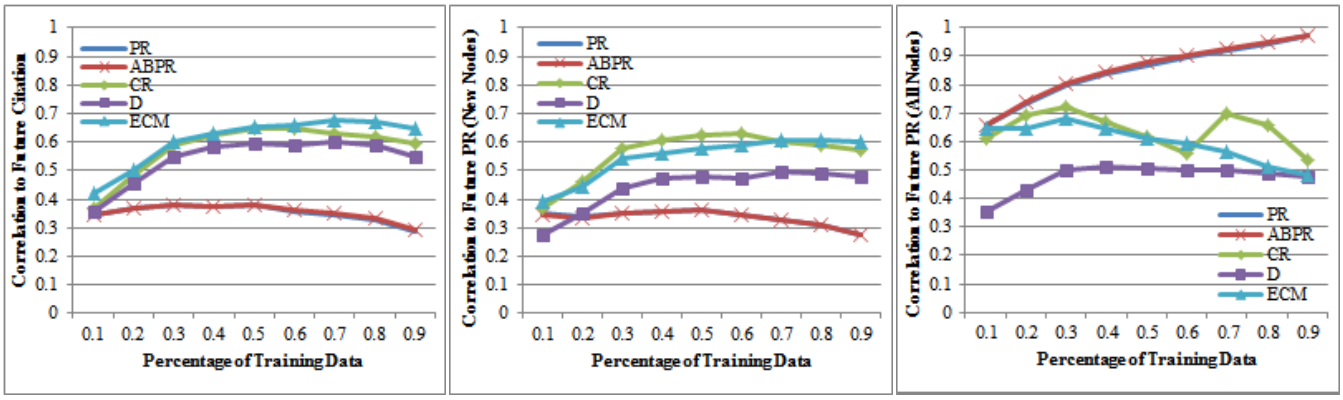
To obtain a manageable set for initial testing, we extracted a subset of papers published in the most recent 20 years, which is about half the size of the total network. We call this data set "APS Part." We estimated the parameters for ECM by applying the approach described above. For other metrics, we selected the parameters that yielded the best balanced performance for them.

Table V shows the results for this subset. The estimated parameters are given in the second row for ECM and RAM. The results show that estimated parameters actually improved the performance for ECM for this data set, but the estimated γ is not optimal for RAM. With authors-tuned parameters for this data set, CiteRank performed better than ECM and RAM here. The overall best performer in terms of Cite and FutNew is still SVRECM.

We then estimated the parameters and repeated the empirical comparison for the entire network. Table VI shows the results. With more data, the performance is improved for most of cases but the winners and losers are similar to the results for its subset.

C. Predicting Top Highly Cited Articles

Correlation between ranked lists reveals the quality of ranking for an entire set of papers but usually we are interested more in how well a metric can identify top papers



(a) Cite (b) FutNew (c) FutAll

Figure 2. Comparison of importance metrics on Hep-Th data set with different partition of data.

Table V
COMPARISON OF CORRELATION RESULTS FOR "APS PART" DATA SET.

Metric	Parameter	Cite	FutNew	FutAll
D	N/A	0.4543	0.4038	0.4838
ID	N/A	0.3171	0.2883	0.8783
OD	N/A	0.3136	0.2695	-0.1186
PR	$\alpha = 0.48$	0.2017	0.1933	0.9615
ABPR	$\alpha = 0.48$ $a = 3.0, b = 0.0001$	0.1902	0.1801	0.9404
CR	$\alpha = 0.55, \tau_{dir} = 8$	0.5912	0.5799	0.3933
ECM	$\alpha = 0.1, \gamma = 0.3$ $t = 0.01$	0.5270	0.4987	0.4116
ECM	$\alpha = 0.00058, \gamma = 0.9$ $t = 0.0004$	0.5365	0.5061	0.5263
RAM	$\gamma = 0.5$	0.5154	0.4847	0.5992
RAM	$\gamma = 0.9$	0.4200	0.3885	0.8125
SVR	$c = 100$	0.6152	0.5925	0.4314
SVRECM	$c = 100, t = 0.0004$ $\alpha = 0.00058, \gamma = 0.9$	0.6273	0.5992	0.4154
SVRRAM	$c = 100, \gamma = 0.5$	0.6272	0.5972	0.4195

Table VI
COMPARISON OF CORRELATION RESULTS FOR "APS" DATA SET.

Metric	Parameter	Cite	FutNew	FutAll
D	N/A	0.3980	0.3674	0.5840
ID	N/A	0.2757	0.2582	0.8988
OD	N/A	0.2822	0.2534	0.0221
PR	$\alpha = 0.48$	0.1822	0.1770	0.9618
ABPR	$\alpha = 0.48$ $a = 3.0, b = 0.0001$	0.2658	0.2594	0.9571
CR	$\alpha = 0.55, \tau_{dir} = 8$	0.6423	0.6366	0.2094
ECM	$\alpha = 0.1, \gamma = 0.3$ $t = 0.01$	0.5883	0.5705	0.3596
ECM	$\alpha = 0.007, \gamma = 0.71$ $t = 0.001$	0.6034	0.5851	0.4540
RAM	$\gamma = 0.6$	0.5472	0.5290	0.5515
RAM	$\gamma = 0.71$	0.5444	0.5256	0.5755
SVR	$c = 100$	0.6548	0.6461	0.2842
SVRECM	$c = 100, t = 0.001$ $\alpha = 0.007, \gamma = 0.71$	0.6734	0.6610	0.3118
SVRRAM	$c = 100, \gamma = 0.71$	0.6730	0.6606	0.2969

that will be cited frequently in the future. More specifically, we want to evaluate whether a metric can predict the top 10% highly cited papers by ranking them as high as possible but ignore how the other papers are ordered.

This task is similar to evaluating the efficacy of an information retrieval system. Usually, the area under the receiver operating characteristic curve (AUC) score is the most popular criterion for the task. However, it was shown that AUC may fail to faithfully reflect the actual quality when the AUC scores are pooled together to evaluate a retrieval system for multiple independent retrieval tasks [16]. AUC is not robust against outlier results. When two disjoint sets of queries are considered, its value for the union of the two sets does not always lie between its value for the two sets of queries. Finally, AUC does not always decrease as the threshold relaxed to include the entire retrieval list. To address these issues, a new evaluation method called the threshold average precision (TAP-k) was proposed [16]. We will adopt this new method to evaluate the metrics on their performance for predicting top 10% of highly cited articles.

TAP-k was designed as a score to evaluate a metric's ranking results for *multiple* independent lists. To compute TAP-k, first we need to determine x, the largest cutoff threshold that produces a median of k false positives over all the output ranked lists. For each list, let P be the total number of positives, in our case, this is the number of top 10% highly cited papers. Let P_x be the precision of the list with cutoff x. That is, the ratio of the number of papers obtained a score $\geq x$ by the metric and the number of papers that are in the set of the top 10% highly cited papers. Define

Table VII
TAP-K RESULTS FOR ALL DATA SETS USING TOP 5 PERCENT CITED PAPER AS POSITIVES.

Metric \ k	5	10	20	100
D	0.0620	0.0771	0.0933	0.1403
ID	0.0517	0.0579	0.0691	0.1090
OD	0.0033	0.0052	0.0073	0.0161
PR	0.0093	0.0106	0.0151	0.0295
ABPR	0.0040	0.0068	0.0086	0.0228
CR	0.0055	0.0070	0.0092	0.0301
ECM	0.1195	0.1339	0.1511	0.2054
RAM	0.1066	0.1330	0.1511	0.2012

Table VIII
TAP-K RESULTS FOR ALL DATA SETS USING TOP 10 PERCENT CITED PAPER AS POSITIVES.

Metric \ k	5	10	20	100
D	0.0496	0.0691	0.1006	0.1426
ID	0.0353	0.0495	0.0738	0.1045
OD	0.0019	0.0033	0.0048	0.0140
PR	0.0054	0.0072	0.0110	0.0269
ABPR	0.0024	0.0042	0.0055	0.0207
CR	0.0031	0.0041	0.0060	0.0256
ECM	0.0969	0.1196	0.1311	0.1875
RAM	0.0933	0.1149	0.1341	0.1853

Table IX
TAP-K RESULTS FOR ALL DATA SETS USING TOP 20 PERCENT CITED PAPER AS POSITIVES.

Metric \ k	5	10	20	100
D	0.0609	0.0792	0.0937	0.1373
ID	0.0478	0.0531	0.0601	0.0935
OD	0.0017	0.0023	0.0039	0.0147
PR	0.0057	0.0079	0.0103	0.0240
ABPR	0.0030	0.0036	0.0056	0.0180
CR	0.0029	0.0033	0.0063	0.0289
ECM	0.0729	0.1090	0.1250	0.1833
RAM	0.0745	0.1049	0.1200	0.1633

APC_x as the sum of the precisions at each rank above the cutoff. For each ranked list, we can compute the weighted average of APC_x and P_x by

$$\text{APCP}_x = \frac{P \cdot \text{APC}_x + P_x}{P + 1}$$

TAP-k is the average of APCP_x over all ranked lists by the metric. A large TAP-k score indicates a better overall performance. TAP-k will penalize ranked lists that are cut short prematurely in an attempt to boost its precision and ranked lists with scores that only reflect the rank but not the quality or importance of the retrieved items (in our case, the papers). See [16] for details.

Again, we applied each metric to rank the oldest 90% of the papers. Among these papers, the top $N\%$ of the papers that were cited the most frequently by the remaining 10% new papers were considered as the true records to be retrieved. The best performing parameters given in Section IV-B for each data set were applied for each metric. Table VII–IX shows the results for retrieving top $N = 5\%, 10\%, 20\%$ highly cited papers, respectively, in terms of TAP-k for $k = 5, 10, 20$ and 100. ECM performs the best by a large margin in most of the cases, with a few exceptions where RAM is the best performer. The results show that ECM and RAM not only ranked highly cited papers higher, but also provided scores that reflected the potential of the papers better.

V. RELATED WORK

Ranking scientific publications is an important application for dynamic network analysis. A long line of bibliometrics research attempted to define objective metrics for identifying important scientific papers, researchers, publication venues, and institutions. The now-accepted measures for evaluating the impact of papers and individual researchers include citations count and h-index [17]. Article citations provide important evidence for ranking scientific papers. Previous works treated citation networks as static networks that aggregate all citations links created over some time period. Chen *et al.* [6] implemented PageRank algorithm on an aggregated network to find influential papers. Radicchi *et al.* [4] divided the entire data period into homogeneous intervals containing equal numbers of citations and applied a PageRank-like algorithm to rank papers and authors within each time slice, thereby, enabling them to study how an author's influence changes in time. In order to address ranking algorithms' bias for older papers, Walker *et al.* [5] introduced *CiteRank*, a modified version of PageRank, that explicitly takes paper's age into account. Sayyadi and Getoor [7] described *FutureRank*, an algorithm that predicts paper's PageRank scores some time in the future. FutureRank implicitly takes time into account by partitioning data in time, and using data in one period to predict paper's ranking in the next. Similar to the approach in [4], FutureRank combines influence rankings computed on the papers and authors networks into a single score. This score is shown to correlate well with the paper's PageRank score computed on citations links that will appear in *the future*. In addition to these metrics, EventRank [18] is also a modification of PageRank that takes into account a temporal sequence of events, e.g., spread of an email message, in order to calculate importance of nodes in a network. This approach takes into account the effect of the *dynamic process* on ranking.

These approaches are somewhat related: our metrics can be said to estimate the expected value of all temporal sequences taking place on the network, the effect of the

dynamic network topology, while no previous work took the temporal order of citation edges into account.

VI. Conclusion

In this paper, we present two new time-aware metrics, ECM and RAM, to rank the publications in a citation network. RAM considers direction citations and degrade its weight as years pass by with a parameter γ. ECM compounds this factor by also considering chains of citations and introducing the other parameter α to penalize the length of the chains. We used four criteria to evaluate their effectiveness as an indicator of a paper's potential of attracting future citations. We performed experimental comparison using these criteria and reported the results here. We summarize our findings as follows.

- If the goal is to rank papers by their probability of being cited in the future, regression models trained by integrating various unsupervised metrics as the features perform the best for all data sets.
- If the goal is to identify future highly cited papers, the ECM score provides the most reliable performance, followed by RAM.
- PageRank and Age-based PageRank are not suitable as indicators of a paper's potential of attracting future citations. They reflect the importance of a paper in a static network.
- Considering citation chains usually help as ECM outperforms RAM and other metrics in most cases.
- For a huge data set, fitting a power law curve can effectively produce well-performing parameters.

Computing ECM involves multiplications of large matrices that could be time-consuming for a large network. Our future work includes to develop a more efficient algorithm to compute ECM and a more efficient and effective algorithm to estimate optimal parameters.

References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University, Tech. Rep., 1998.

[2] R. Baeza-Yates, F. Saint-Jean, and C. Castillo, "Web structure, dynamics and page quality," in *String Processing and Information Retrieval*, ser. Lecture Notes in Computer Science, A. H. F. Laender and A. L. Oliveira, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, September 2002, vol. 2476, ch. 12, pp. 453–461. [Online]. Available: http://dx.doi.org/10.1007/3-540-45735-6_12

[3] K. Berberich, M. Vazirgiannis, and G. Weikum, "Time-aware authority ranking," *Internet Mathematics*, vol. 2, no. 3, pp. 301–332, January 2005. [Online]. Available: http://www.metapress.com/content/y067112167681312

[4] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, "Diffusion of scientific credits and the ranking of scientists," *Physical Review*, vol. E80, pp. 056 103+, Sep 2009. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.80.056103

[5] D. Walker, H. Xie, K.-K. Yan, and S. Maslov, "Ranking scientific publications using a simple model of network traffic," Dec 2006, coRR, abs/physics/0612122. [Online]. Available: http://arxiv.org/abs/physics/0612122

[6] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with google's pagerank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, January 2007. [Online]. Available: http://arxiv.org/abs/physics/0604130

[7] H. Sayyadi and L. Getoor, "Future rank: Ranking scientific articles by predicting their future pagerank," in *2009 SIAM Int. Conf. on Data Mining (SDM09)*, 2009. [Online]. Available: http://linqs.cs.umd.edu/basilic/web/Publications/2009/sayyadi:sdm09/sayyadi_futureRank_sdm09.pdf

[8] S. Redner, "Citation statistics from 110 years of physical review," *Physics Today*, vol. 58, no. 6, pp. 49–54, 2005. [Online]. Available: http://dx.doi.org/10.1063/1.1996475

[9] S. Maslov and S. Redner, "Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks," Jan. 2009. [Online]. Available: http://arxiv.org/abs/0901.2640v1

[10] K. Lerman, R. Ghosh, and J. H. Kang, "Centrality Metric for Dynamic Networks," in *Proceedings of KDD workshop on Mining and Learning with Graphs (MLG)*, Jun. 2010. [Online]. Available: http://arxiv.org/abs/1006.0526

[11] P. Bonacich, "Eigenvector-like measures of centrality for assymetric relations," *Social Networks*, vol. 23, pp. 191–201, 2001.

[12] R. Ghosh and K. Lerman, "The structure of heterogeneous networks," in *Proc. of the 1st IEEE Social Computing Conf.*, 2009. [Online]. Available: http://arxiv.org/abs/0906.2212

[13] P. Bonacich, "Power and centrality: A family of measures," *The American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987. [Online]. Available: http://dx.doi.org/10.2307/2780000

[14] L. Katz, "A new status derived from sociometric analysis," *Psychometrika*, vol. 18, pp. 39–43, 1953.

[15] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, pp. 1207–1245, May 2000. [Online]. Available: http://portal.acm.org/citation.cfm?id=1139689.1139691

[16] H. D. Carroll, M. G. Kann, S. L. Sheetlin1, and J. L. Spouge, "Threshold average precision (tap-k): A measure of retrieval designed for bioinformatics," *Bioinformatics*, vol. 26, no. 14, pp. 1708–1713, 2010.

[17] J. E. Hirsch, "An index to quantify an individual's scientific research output," *PNAS*, vol. 102, no. 46, pp. 16 569–16 572, 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0507655102

[18] J. O'Madadhain and P. Smyth, "Eventrank: a framework for ranking time-varying networks," in *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*. New York, NY, USA: ACM, 2005, pp. 9–16. [Online]. Available: http://dx.doi.org/10.1145/1134271.1134273