Ageing of Edges in Collaboration Networks and its Effect on Author Rankings

Dalibor Fiala[1], Gabriel Tutoky[2], Peter Koncz[2], Ján Paralič[2]

[1]University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic
email: dalfia@kiv.zcu.cz

[2]Technical University of Košice, Letná 9, 04001 Košice, Slovakia
email: {gabriel.tutoky, peter.koncz, jan.paralic}@tuke.sk

Abstract: In this paper we show that assigning weights to the edges in a collaboration network of authors, according to a decreasing exponential function depending on the time elapsed since the publication of a common paper, may add valuable information to the process of ranking authors based on importance. The main idea is that a recent collaboration represents a stronger tie between the co-authors than an older one and, therefore, reduces the weight of potential citations between the co-authors. We test this approach, on a well-known data set and with an established methodology of using PageRank-based ranking techniques and reference sets of awarded authors and demonstrate that edge ageing may improve the ranking of authors.

Keywords: collaborations; citations; PageRank; scholars; rankings

1 Introduction

The ranking of authors of scholarly literature based on various bibliometric aspects has been popular in recent years for the purpose of the research assessment of individual scientists and related activities, such as prize and grant awarding, hiring, or promotion. It may indeed be considered a valuable complementary tool to manual research evaluation. In our previous research [1, 2], we were concerned with applying the recursive Google PageRank algorithm by Brin and Page [3] to bibliographic networks in order to rank researchers by prestige that is based on both the citation and the collaboration network of authors [4]. We have shown that such a ranking scheme may be more objective and fair. On the other hand, in another study [5] we have tested the effect of edge ageing in a collaboration social network [6] of teenagers. The main idea is that the strength of a collaboration tie (i.e. the edge weight) diminishes within the course of time. The goal of this present short paper is to show that we are able to combine both approaches (bibliographic PageRank and edge ageing) to produce author rankings that may

better reflect reality. There was other related work, prior to [1], for example, the pioneering studies on the usage of PageRank in bibliographic networks [7, 8, 9] and many more appeared later, e.g. [10, 11, 12, 13]. From the numerous papers on scientific collaboration networks, we discuss herein, two of the most well-known works – [14] and [15]. And, of course, the natural extension of each collaboration network analysis is a visualization process, e.g. [16] or [17], which is, however, beyond the scope of this article.

2 Methods and Data

The importance of nodes in a directed graph is often assessed by means of the number of in-coming edges and more advanced techniques are recursive in that they study the significance of these in-linking nodes by inspecting their in-links and so on. The importance, in this sense, is sometimes called prestige. One of these methods is the well-known PageRank algorithm by Brin and Page [3] originally conceived for the ranking of webpages, which may, nevertheless, be applied to any directed graph. Thus, let $G = (V, E)$ be a directed graph of citations between authors, with V as the set of vertices (authors) and E as the set of edges (with all citations between two authors in the same direction merged to one edge). So if author v cites author u (once or multiple times), there is exactly one edge $(v, u) \in E$. Then the PageRank score PR(u) of author u is computed recursively and the result depends on the scores of all citing authors, in the following way:

$$\mathrm{PR}(u) = \frac{1-d}{|V|} + d \sum_{(v,u)\in E} PR(v)\Omega \tag{1}$$

where d is the damping factor (set to 0.85 in the original web experiments), and Ω is either $D_{out}^{-1}(v)$ (with D_{out} being the out-degree of v) like in the standard PageRank by [3] or $\sigma_{v,u} \Big/ \sum_{(v,k)\in E} \sigma_{v,k}$ like in the bibliographic PageRank by Fiala et al. [1], where $\sigma_{v,k} = w_{v,k} \Big/ \left(\left[(c_{v,k}+1) / (b_{v,k}+1) \right] \sum_{(v,j)\in E} w_{v,j} \right)$ and w, b, and c are various coefficients that determine the weights of edges between authors. The parameters w, b, and c are themselves based on the topology of both the citation and collaboration network of authors and for details we refer to [1] or to [2]. Here, let us say, that these coefficients help assign less weight to citations from colleagues and that the strength of relationship between two authors does not rely solely on the number of joint papers but also on other factors such as the number of coauthors in those papers. Based on the combination of the above parameters, Fiala [2] called the seven new PageRank (PR) variants tailored for bibliographic networks in this way: (*PR) collaboration*, *publications*, *allCoauthors*, *allDistCoauthors*, *allCollaborations*, *coauthors*, and *distCoauthors*.

As far as collaboration networks are concerned, we can additionally consider several approaches to the projection of collaborations among authors like binary weights, counting of co-occurrences, Newman's weights determination (see Eq. 2) and the *edge weights ageing principle*. We discuss all of these methods in detail in [5]. Briefly, a collaboration network of authors is a two-mode network [18] with two types of nodes – authors and their publications. In this network, actors are connected together via their common publications, but there does not exist, a direct connection between them (or between publications). This network needs to by projected onto a one-mode network with a single type of nodes – authors. See Figure 1 for an example of such a projection.

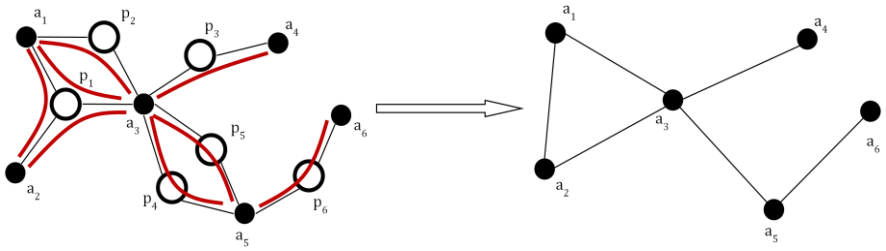

Figure 1
Projection of an author collaboration two-mode network
(a – authors, p – publications) onto a one-mode network

A projected relation (collaboration) between two authors depends on several factors, e.g. the number of co-authors of a single publication or the number of common publications of two selected authors, etc. An expression of that relationship in a binary form can be expressed as 1, if two authors have a common publication and 0 if not. A better approach is that the relation weight is expressed in the interval <0, 1> where 0 represents no relation and 1 represents the maximum strength of the relation. The first factor – number of co-authors of a single publication (N_a) can be expressed by the formula proposed by [19]:

$$w_{ij}^{p} = \frac{1}{N_a - 1} \tag{2}$$

or by the formulas proposed in [5], one of which, is an exponential expression with the decay parameter α_e, which can be adjusted with respect to the particular type of collaboration network. In this way, it can influence the shape of the exponential curve:

$$w_{ij}^{p} = \alpha_e^{2-N_a} \tag{3}$$

where w_{ij}^{p} is a projected weight of a single publication *p* between authors *i* and *j*.

Additionally, we can consider that collaboration strength between two authors is related to time-based factors – year of publication or more accurately to the time spent between their common publications and their frequency. We suppose that if the frequency of common publications of two authors is higher and/or their common publications are newer then their collaboration strength is larger. And vice versa, if the frequency of common publications is lower and/or their common publications are older, then the collaboration strength of those authors is smaller. In [5] we propose time dependent weights in the representation of a one-mode projected affiliation network – a kind of ageing of the ties (edges). This should be considered as a similar approach to the one presented in [20] or [21] where authors considered ageing of the nodes in the context of citation networks. They describe a node's age as influence on the probability of connecting the current node to new nodes in the network.

The ageing of the ties (collaborations) among authors can be described as an evolutionary process depicted in Figure 2 where $e_1, e_2, \ldots, e_m$ are events (common publications) on which authors i and j collaborate together. This ageing allows for the termination of sporadic and insignificant relations (e.g. when two authors have rarely participated in common publications) and, by contrast, if the relations are periodically repeated (we assume that these relations are significant), they are "highlighted" in the network and their weight is increased (see events e_1 through e_3 in Figure 2). After the creation of the first collaboration (based on the first event e_1), a tie with a value of 1 comes into being which decreases exponentially in time according to the formula

$$w_{ij}(t+\delta) = \left\{ \begin{array}{ll} w_{ij}(t)\, e^{-\theta \Delta t} & \text{when } w_{ij}(t)\, e^{-\theta \Delta t} > \varepsilon \\ 0 & \text{otherwise} \end{array} \right\} \tag{4}$$

where $w_{ij}(t)$ is the weight of the collaboration in time t and $w_{ij}(t+\delta)$ is the collaboration weight after time δ. The value ε is the threshold value of a minimum collaboration weight and the factor θ is called the ageing factor which designates the "ageing speed" and is described by $\theta = \frac{\ln 2}{t_{1/2}}$ where $t_{1/2}$ is the time span after which the weight of ties decreases by 50% in the ageing process. For more details see [5].

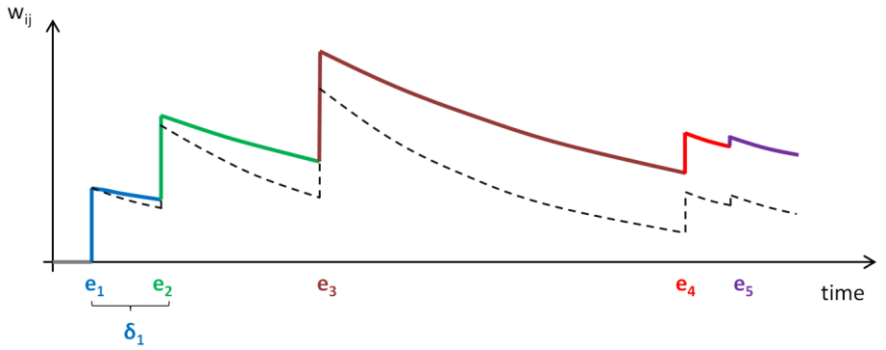

Figure 2

Evolution in time of a single collaboration tie between authors *i* and *j* for two various ageing factors (Events e_1, .., e_5 represent common publications between authors *i* and *j*. At the time when a publication is produced, the initial strength of the tie between the authors (value w_{ij}) is determined by the projection method and subsequently the tie strength decreases as time passes since the creation of the new publication.)

As for the data examined in our analysis, we investigated the same data set as in the study by Fiala [2], where there was a citation graph of 205,780 computer science publications from the period 1996-2005 with 276,957 citations between them. This paper citation graph was subsequently transformed to an author citation graph with 187,016 authors and 1,471,312 citations between them (with no self-citations allowed). The authors were represented by their surnames and given names initials and we did not perform any name disambiguation or unification. Self-citations of authors were removed before we started our analysis. The resulting (two-mode) collaboration network thus consisted of 392,796 nodes (publications + authors) and 492,284 edges (publication-author connections).

3 Results and Discussion

We applied those seven ranking techniques mentioned in the section on methods with ageing factors set on (i.e. "new methods") to the same author citation network described in [2]. The result of this application was seven author rankings which we compared to the rankings from the above study, in which edge weight ageing did not occur (i.e. "old methods"). The comparison was twofold: first, we calculated five statistics (minimum, maximum, median, mean value, and standard deviation) for the "new" and "old" rankings using the ranks of Codd[1] and Turing[2]

[1] Codd Award, http://www.sigmod.org/sigmod-awards/sigmod-awards#innovations
[2] Turing Award, http://awards.acm.org/homepage.cfm?srt=all&awd=140

awardees and plotted them as lines in Figure 3 and Figure 4 and, second, created boxplots for each pair of "new" and "old" rankings (with the first quartile, median, and third quartile of ranks forming the bar and the minimum and maximum ranks being the whiskers) displayed in Figure 5 and Figure 6.

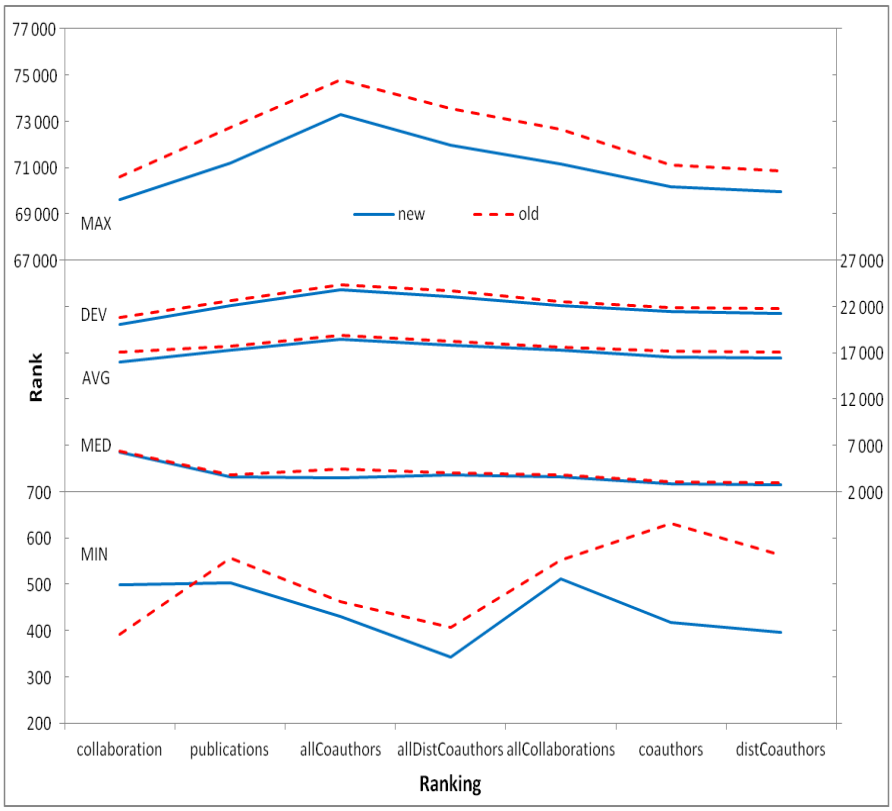

Figure 3

Codd Award and five statistics of new and old methods

(MAX – maximum rank, MIN – minimum rank, MED – median rank, AVG – mean rank, DEV – standard deviation of ranks produced by a particular ranking method. The lower the values of MAX, MIN, MED, and AVG, the better the ranking. Even for DEV, lower values mean a more compact ranking, which is a desirable property.)

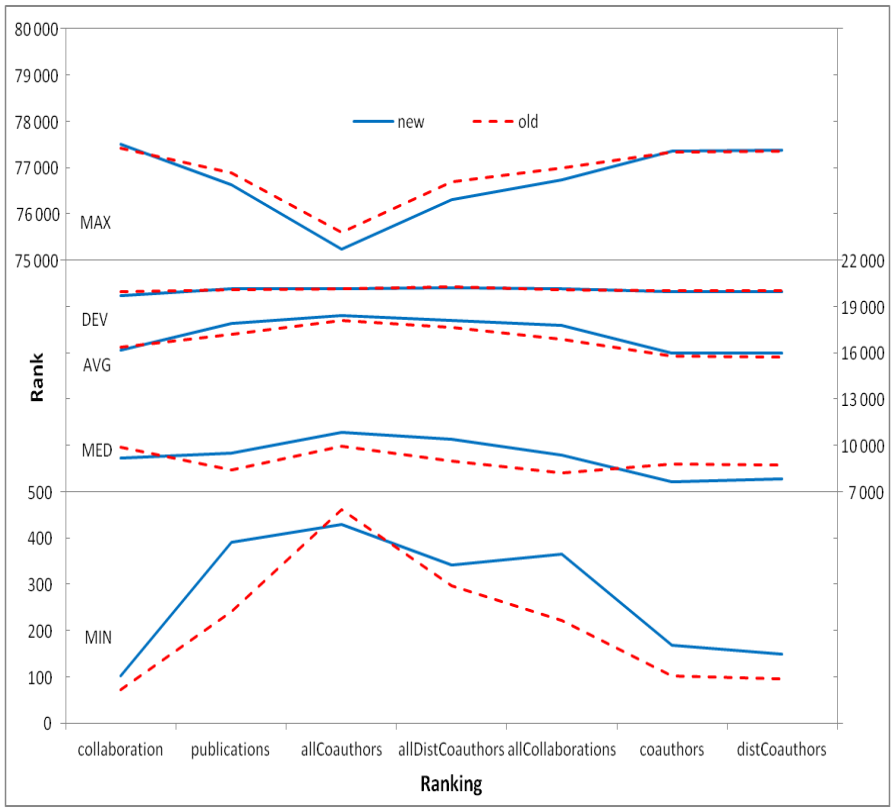

Figure 4

Turing Award and five statistics of new and old methods

(MAX – maximum rank, MIN – minimum rank, MED – median rank, AVG – mean rank, DEV – standard deviation of ranks produced by a particular ranking method. The lower the values of MAX, MIN, MED, and AVG, the better the ranking. Even for DEV, lower values mean a more compact ranking, which is a desirable property.)

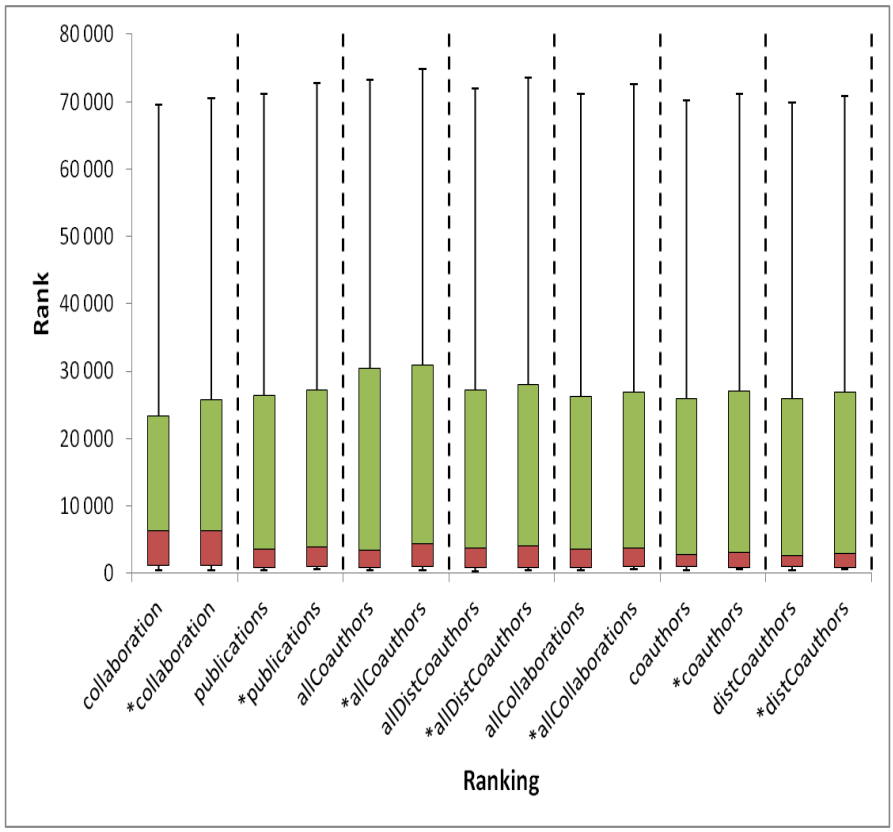

Figure 5

Codd Award boxplots (* = old method)

(The bottom of each bar represents the 25th percentile and the top is the 75th percentile. The median rank of each ranking method lies on the boundary between the red and the green rectangle. The minimum and maximum ranks achieved are marked with the whisker lines. The smaller a box, the greater the trend to produce better ranks for the awarded authors.)

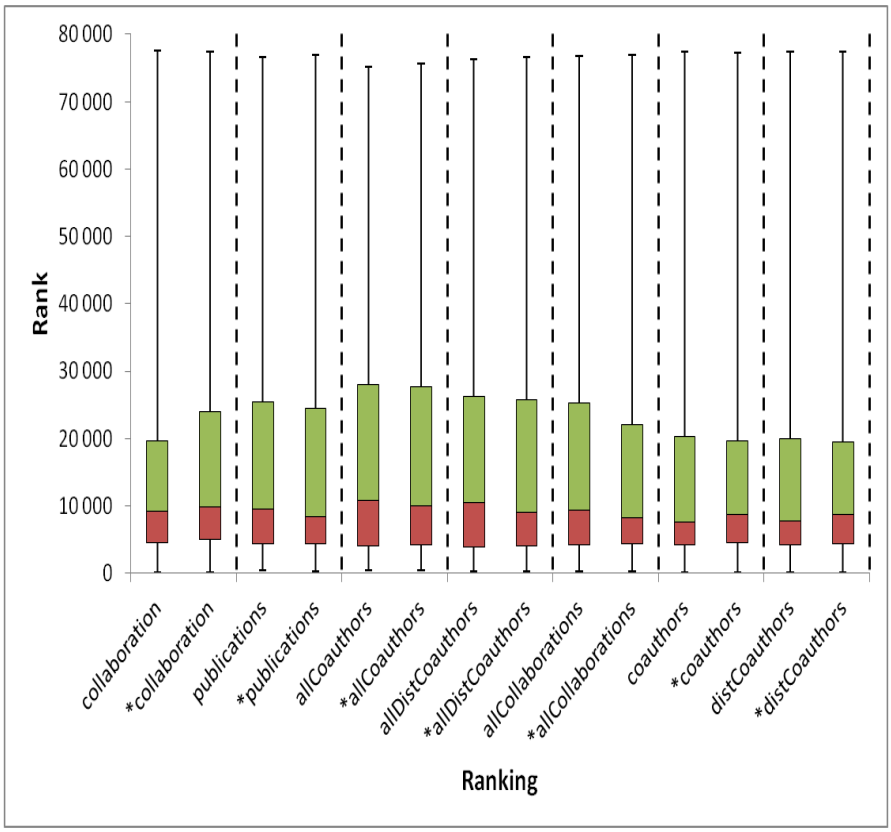

Figure 6

Turing Award boxplots (* = old method)

(The bottom of each bar represents the 25th percentile and the top is the 75th percentile. The median rank of each ranking method lies on the boundary between the red and the green rectangle. The minimum and maximum ranks achieved are marked with the whisker lines. The smaller a box, the greater the trend to produce better ranks for the awarded authors.)

The blue solid line represents the new methods (with ageing factor enabled) and the red dashed line represents the old methods from [2]. Since lower ranks mean better ranks (i.e. position 1 is better than position 100), we can see in Figure 5 that the new methods outperform the old ones, most remarkably in the terms of the maximum and minimum rank achieved (MAX and MIN) and only slightly regarding the other three statistics (median rank – MED, mean rank – AVG, and standard deviation of ranks – DEV). However, unlike Codd Award this phenomenon is much less visible in Figure 5 with Turing Award winners. Only the MAX statistics is clearly better there for the new methods. DEV is approximately the same and the remaining three are worse for the new methods.

Similar trends may be observed with the boxplots. While the bars tend to be positioned lower (towards better ranks) with the new methods for Codd Award winners in Fig. 5, it is almost the opposite for Turing Award laureates in Fig. 6. A first explanation of this observation may be the different nature of these awards and some specific features of the data set underlying our analysis. While the Codd Award, intended for database researchers, seems to be well modelled by the citation patterns in our sample of Web of Science data, the Turing Award for achievements in general computer science is less so. It appears that much more than the analysis of citation and collaboration networks is needed to identify the winners of such a broadly defined award. A further investigation into the different outcomes of the application of the ranking methods based on edge weight ageing to Codd Award and Turing Award winners will be necessary in the future.

The main weakness of the ageing of ties is the right determination of the θ factor used in Eq. 4. This factor must be estimated considering the data set and the time range of the publications and the awards considered. If we determine θ to yield a too "slow" ageing, the method does not have the required influence on the results and, conversely, when a too "fast" ageing is produced, the method suppresses nearly all of the ties in the network and we will not have enough ties sufficiently distributed across all possible strengths. As an optimum value for θ we have chosen $t_{1/2} = 730$ days, which means that the strength of all ties in the network decreases by a half of their value during 2 years. This value was determined based on previous experiments similar to the experiments described in [5] for estimating θ, which are not described in this article.

Conclusion

In this short paper, we showed that the ageing of edges and their weights in the collaboration networks of authors, might add some useful information to the process of assessing the importance of individual authors. More precisely, by weighing the edges in a collaboration network of authors according to a decreasing exponential function, that depends on the time elapsed since a common publication was produced, we are able to differentiate between fresh and obsolete collaborations and to decide on the strength of relationship between two authors. This factor may then be input in a ranking scheme as in [1] or [2] that ranks authors in a citation network, taking into account the information from the corresponding collaboration network. We demonstrated this approach on the same data set and methodology used by [2] and showed that edge ageing improved the ranking of researchers when Codd Award winners were used as a reference set of outstanding scholars but did not make the ranking better when Turing Award winners were used. We argue that this contrast may be the result of the distinct nature of these awards (a specialized versus a general award) that are reflected differently by the underlying data set. We think examining the influence of the ageing of edges, in collaborative networks, on the ranking of authors is worth of further research.

Acknowledgement

This work was supported by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090 and in part by the Ministry of Education of the Czech Republic under grant MSMT MOBILITY 7AMB14SK090. This work was also supported by the Slovak Research and Development Agency under the contract No. SK-CZ-2013-0062 and by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12. Thanks are due to M. Dostal and C. Havrilová for their useful comments as well.

References

[1] Fiala, D., Rousselot, F., & Ježek, K. (2008) PageRank for Bibliographic Networks. *Scientometrics*, 76(1) 135-158

[2] Fiala, D. (2012) Time-Aware PageRank for Bibliographic Networks. *Journal of Informetrics, 6*(3) 370-388

[3] Brin, S., & Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia, 107-117

[4] Newman, M. E. J. (2001) Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality. *Physical Review E*, *64*, art. no. 016132

[5] Tutoky, G. (2011) Discovery and Exploitation of Knowledge in Collaboration Social Networks. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, *3*(4) 28-36

[6] Wasserman, S., & Faust, K. (1994) Social Network Analysis. Cambridge University Press, Cambridge, UK

[7] Liu, X., Bollen, J., Nelson, M. L., & Van De Sompel, H. (2005) Co-Authorship Networks in the Digital Library Research Community. *Information Processing and Management*, *41*(6) 1462-1480

[8] Bollen, J., Rodriquez, M. A., & Van De Sompel, H. (2006) Journal Status. *Scientometrics*, *69*(3) 669-687

[9] Chen, P., Xie, H., Maslov, S., & Redner, S. (2007) Finding Scientific Gems with Google's PageRank Algorithm. *Journal of Informetrics*, *1*(1) 8-15

[10] Ma, N., Guan, J., & Zhao, Y. (2008) Bringing PageRank to the Citation Analysis. *Information Processing and Management*, *44*(2) 800-810

[11] Yan, E., & Ding, Y. (2009) Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107-2118

[12] Ding, Y. (2011) Applying Weighted PageRank to Author Citation Networks. *Journal of the American Society for Information Science and Technology*, *62*(2) 236-245

[13] Yan, E., & Ding, Y. (2011) Discovering Author Impact: A PageRank Perspective. *Information Processing and Management*, *47*(1) 125-134

[14] Newman, M. E. J. (2001) The Structure of Scientific Collaboration Networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2) 404-409

[15] Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002) Evolution of the Social Network of Scientific Collaborations. *Physica A*, *311*(3-4) 590-614

[16] Shannon, P., Markiel, A., Ozier, O., Baliga N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11) 2498-2504

[17] Horváth, A. (2013) The Cxnet Complex Network Analyser Software. *Acta Polytechnica Hungarica*, *10*(6) 43-58

[18] Latapy, M., Magnien, C., & Vecchio, N. D. (2008) Basic Notions for the Analysis of Large Two-Mode Networks. *Social Networks*, *30*(1) 31-48

[19] Newman, M. E. J. (2004) Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks. *Lecture Notes in Physics*, *650*, 337-370

[20] Hajra, K. B., & Sen, P. (2005) Aging in Citation Networks. *Physica A*, *346*(1-2) 44-48

[21] Zhu, H., Wang, X., & Zhu, J.-Y. (2003) Effect of Aging on Network Structure. *Physical Review E*, *68*(5) art. no. 056121