2010 International Conference on Computational Aspects of Social Networks

Social Network Reduction Based on Stability

Milos Kudelka, Zdenek Horak, Vaclav Snasel and Ajith Abraham
VSB Technical University Ostrava
Ostrava, Czech Republic
Email: kudelka@inflex.cz, zdenek.horak.st4@vsb.cz, vaclav.snasel@vsb.cz, ajith.abraham@ieee.org

***Abstract*—The analysis of social networks is concentrated especially on uncovering hidden relations and properties of network members (vertices). Most of the current approaches are focused mainly on different network types and different network coefficients. On one hand, the analysis can be relatively simple; on the other hand some complex approaches to network dynamics can be used. This paper introduces a novel aspect of network analysis based on the so-called Forgetting Curve. For network vertices and edges, we define two coefficients, which describe their role in the network depending on their long-term behavior. Using one of these parameters we reduce the network to smaller components. We provide some experimental results using DBLP[1] dataset. Our research illustrates the usefulness of the proposed approach.**

***Keywords*-social network reduction, memory, stability, complexity reduction, co-authorship network, visualization**

I. Motivation

Generally, the term memory is understood as committing, storing and recalling of experiences. The role of memory is crucial because it stores and recalls all the information we need for our normal lives. All stimuli and situations in which we find ourselves are compared to their traces in memory, which allows us to recognize the meaning of these stimuli and situations. Recalling information is either a reproduction or re-memorization of already known information. The process of forgetting is opposite to the process of recalling. To forget something means not to lose the particular memory trace, but replace it with a new experience. Nothing is forgotten, it just cannot be recalled, because it has lost its meaning. There are two factors causing information to be forgotten. The first one is the extinction of the unused memory trace and the second is the interference of new experiences - the replacement of less important information by the more important ones.

The goal of our research is to apply known and proven methods of learning and forgetting into the field of social networks. The human brain stores information, which is fixed in the memory by its frequent usage, but which can – when not used – also fade from the memory. This process is very complex. However many experiments have already been done (see [4]), which lead to fairly exact description of functions involved in the memorization and the forgetting of information. We wish to review the social network as a human brain, which learns and forgets information. The reason is that the vertices of the network are people having these functions in their brains. In the following text, we understand under the term social network an undirected weighted graph. During the calculations of edge and vertex weights, we use time-dependent and the forgetting of information-related values. For our experiments, we use our own hypothesis that in removing vertices and edges with a low weight we can reduce the network but still maintain the important ties.

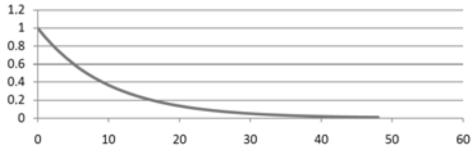

Figure 1. Forgetting Curve

II. Forgetting Curve

Ebbinghaus proposed the forgetting curve in 1885. The forgetting curve (see [4]) defines the probability that a person can recall information at time t since previous recall. It can describe long-term memory and is usually presented using the following equation.

$$R = e^{-\frac{t}{S}}$$

R (memory retention) the probability of recalling information at time t since the last recall.
e Euler number (aprox. 2.718).
t time since the last recall.
S (relative strength of memory - stability) approximated time since the last recall for which is the information stored in memory.

Remark There are also different approaches for the computation of the forgetting curve (see for example [18]) but the conclusions are always very similar - the forgetting process is much faster in the beginning (see fig. 1).

The computation depends on the type of memory, especially on the estimated time S (this value is not constant in the long term). For simplicity, assume that if we work with the information for the first time, then the time of storing

[1] http://www.informatik.uni-trier.de/~ley/db/

978-0-7695-4202-7/10 $26.00 © 2010 IEEE
DOI 10.1109/CASoN.2010.120

information in memory is $S_{ini} > 0$ and this default value is constant.

An important feature of long-term memory is that after reproduced information recall in the time $t > 0$, the time of storing information in memory S changes. The change is dependent on the previous time S and on the time of recall t. Ideally, the reproduced recall multiplies this time (in comparison with the previous value) by factor $F > 1$.

The other important feature of long-term memory is, that immediate reproduced recall (too quick) of information has no bigger effect on the learning. On the other hand, the reproduced recall too lately (in time near S) causes substantial forgetting. There is an optimal time between these two extreme situations in which the reproduced information recall causes a high level of remembering (and consequently the maximum increase of time S by factor F).

In the ideal case (reproducing the information in optimal time), the remembering of information is gradual and very effective - after each recall, the time of storing information in memory S (remembering) is multiplied by factor F.

A. Reproduced information recall

For updated S_{new}, after new information recall should hold:

1) If $t > S$ then $S_{new} = S_{ini}$ (information is considered as new)
2) If $t \rightarrow S$ then $S_{new} \rightarrow S_{ini}$ (late recall is considered as almost new information) .
3) If $t \rightarrow 0$ then $S_{new} \rightarrow S$ (early recall has almost no influence)
4) If $t \rightarrow \text{opt}(S)$ then $S_{new} \rightarrow F\cdot S$, where $\text{opt}(S)$ is the function returning optimal time for recalling the information and F is the factor of optimal improvement.

Remark For reproduced information recall is $R = 1$. This follows from the fact, that $t = 0$ at this moment. For the factor of optimal improvement holds, that when the information is recalled at optimal time, the value of S is multiplied by two (depending on the type of memory). Therefore we can assume that $F \in (1; 2\rangle$.

B. Calculation of S_{new}

We have to consider three things:

1) The function $\text{opt}(S)$ for the calculation of optimal information recall time.
2) The choice of optimal improvement factor F.
3) Function $f(t, S, F)$ for calculation of S_new.

Function opt(S) Available sources present the optimal time for reproduced information recall in the range of 10–30% of time S. The setting of this function is dependent on the type of memory (e.g. $\text{opt}(S) = 0.2\cdot S$).

The factor F **of optimal improvement** The factor F is involved in the computation of time S for which the information is held in memory (is remembered). This factor is again dependent on the type of memory. For the calculation of S with the same type of memory the value of F is constant (e.g. $F = 1.2$)

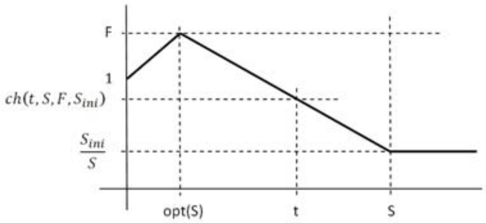

Figure 2. Calculation of S in time t

The function $ch(t, S)$ The value of S_{new} is dependent on the type of memory, on the time of repetitive information recall and on the previous value of S (this incorporates the history of learning mentioned information). For the calculation of S_{new} we need to design the function of $ch(t, S, F, S_{ini})$ for the calculation of the coefficient of change of the value S. Then holds:

$$S_{new} = ch(t, S, F, S_{ini})\cdot S$$

Available sources contains various approaches for the computation of value of function $ch(t, S, F, S_{ini})$. For example we will use simple relation based on linear functions (see fig. 2):

1) If $0 \leq t \leq \text{opt}(S)$ then $ch(t, S, F, S_{ini}) = 1 + (F - 1)\cdot \frac{t}{\text{opt}(S)}$
2) If $\text{opt}(S) \leq t \leq S$ then $ch(t, S, F, S_{ini}) = F - (F - \frac{S_{ini}}{S})\cdot \frac{t-\text{opt}(S)}{S-\text{opt}(S)}$
3) If $t > S$ then $ch(t, S, F, S_{ini}) = \frac{S_{ini}}{S}$

III. Forgetting of social network

We assume that interactions between particular pairs of vertices take place in the social network continuously. If we consider these interactions as an experience stored in memory, then the ties between two vertices of the network are more stable, if this network learns these interactions. As a result we assume that the more interactions occur between the two vertices, the more stable is the tie between them. Therefore we can understand the social network as a set of variously stable ties.

Remark Interaction between two vertices as well as information leaves traces in memory. This trace is dependent on how often these interactions take place (as an analogy to the reproduced information recall). If we will understand the network as an analogy to the human brain, then the memorization of ties in the network will correspond to the degree of remembering the information in the brain.

A. Edge Retention & Stability

The properties of ties change over time, depending on how often and in what time two vertices interact. For the calculation of the properties of ties we use the Forgetting

Curve. It is the analogy to the learning and forgetting of reproduced information - reproduced interaction. For each tie we define three time-changing characteristics.

Definition: Edge Retention Edge Retention ER expresses the probability that a reproduced interaction will take place in given time t between two vertices connected by given edge.

Definition: Edge Stability Edge Stability ES is the estimated time for which the tie between vertices remains active (since given time t).

Definition: Active Edge Active Edge is a tie, for which holds that $ES > 0$ in given time t.

B. Vertex Retention & Stability

Like the retention and stability of ties we can define the same vertex characteristics and use the forgetting curve in their calculations again.

Definition: Vertex Retention Vertex Retention VR expresses the probability that a reproduced interaction will take place in given time t between this vertex and any other vertex.

Definition: Vertex Stability Vertex Stability VS is the estimated time for which the vertex remains active (since given time t).

Definition: Active Vertex Active Vertex is a vertex, for which holds that $VS > 0$ in given time t.

IV. Related work

The analysis of general complex networks is well-described in [2] and [3]. Liu et al. in [13] provides a good overview of Social Network Analysis, co-authorship networks and their combination. They also compared the results of the analysis using classical SNA coefficients (such as closeness, betweeness, etc.) and PageRank and its modification AuthorRank, respectively. Newman's [15] work is also organized in a similar way, but contains additional coefficients such as the number of papers per author, number of authors per paper, distances between authors, etc.

Hart [8] provided an interesting survey on co-authorship (although from a different field of science), the reasons why the authors work together, what are the benefits of working together, what tasks are usually shared among co-authors, different co-authorship models and the name ordering protocols. Han et al. [7] introduced the concept of *supportiveness*, which captures co-authorship ties in a non-symmetric way.

A visualization of social network is a very important part of the whole SNA, as good visualization can quickly provide good insight into the network structure, its vertices and their properties. The evolution of this visualization from simple hand-drawing images up to complex computer generated schemes was illustrated by Freeman in [6]. Ye et al. [19] discussed the visualization of co-authorship networks using a minimum spanning tree of the largest component, filtering unstable links between vertices using threshold and highlighting important groups as cliques. Huang and Huang [11] addresed two main problems of most visualization techniques - the problematic application in large-scale networks and the difficulty to incorporate historical data in one artifact.

Elmacioglu and Lee [5] presented statistics calculated from the DBLP dataset about conference papers and their authors. They also provided comparison of weighted and unweighted variants of SNA coefficients used to identify important authors in the network. An interesting approach to the visualization of co-authorship networks constructed from the DBLP dataset using overlapping groups can be found in Santamaria and Theron [17].

Barabasi et al. [1] is focused on the evolution of the social network of co-authorship, respectively on the evolution of its characteristic properties. Co-authorship can be also considered as a suitable area for link prediction (see for example [14], [16])

The application of network analysis on DBLP data is not limited to co-authorship networks, but may also be used to identify different communities sharing common interests and to follow their evolution from emerging communities to their vanishing point (see Huang et al. [10]). The analysis of the citation network leads to the well-known H-index (see Hirshman [9]). Liu et al. [12] used diversity to identify important vertices in the network and performs an experiment on the DBLP dataset. The greater difference between the neighbors of a particular vertex, the greater is the diversity.

Our proposed approach is strongly based on the historical data and differs in using the Forgetting curve.

V. DBLP Dataset Experiments

For our experiments, we need time-dependent data to calculate the retention and stability of the forgetting curve. In April 2010, we downloaded the DBLP dataset in XML[2] and preprocessed it for further usage. First of all, we selected all conferences held by IEEE, ACM or Springer, which gave us 9,768 conferences. For every conference we identified the month and year of the conference.

In the next step we extracted all authors having at least one published paper in the mentioned conferences (as authors or co-authors). This gave us 443,838 authors. Using the information about authors and their papers we were able to create a set of cooperations between these authors consisting of 2,054,403 items. An important fact is that *cooperation* is understood to be the co-authorship of one paper. Using the information about the conference date, we accompanied

[2] Available from http://dblp.uni-trier.de/xml/

Table I
TOP 10 VERTICES AND EDGES BY STABILITY

#	AUTHOR(vertex)	CO-AUTHORS (edge)	
1	Christos H. Papadimitriou	Irith Pomeranz	Sudhakar M. Reddy
2	Moshe Y. Vardi	Enrico Macii	Massimo Poncino
3	Serge Abiteboul	Evangelos Kranakis	Danny Krizanc
4	Martin Wirsing	Feng Bao	Robert H. Deng
5	Hector Garcia-Molina	Divyakant Agrawal	Amr El Abbadi
6	Philip S. Yu	Maurizio Rebaudengo	Matteo Sonza Reorda
7	Amir Pnueli	Louise E. Moser	P. M. Melliar-Smith
8	John H. Reif	Xiao Zhou	Takao Nishizeki
9	Paul G. Spirakis	Patrick Girard	Christian Landrault
10	Ugo Montanari	Orna Kupferman	Moshe Y. Vardi

these cooperations by time information. We also ignored the ordering of author names as it is impossible to investigate the particular ordering protocol (by alphabet, by contribution, etc.) and hence all co-authors are given equal importance.

A. Weighting Edges and Vertices

We computed the weight of edges and vertices as their stability in time t. We divided the entire recorded publication period of conferences (the first record from 1963) into one-month time periods. If during one month an author has published a paper with another co-author in at least one conference (held by IEEE, ACM or Springer), then we set one interaction for the both authors (vertices) and the tie between author and co-author (edge) for this month. For each vertex and edge we obtain a list of months in which the interactions occurred. Then we applied the forgetting curve to compute the retention and stability of every author and tie in a specified month.

We have truncated the selected time period to December 2008 to obtain the most complete dataset. Of course, the weight (stability of vertices and edges) changes in every month, but the following calculations are made until the end of year 2008. At that time only 122 289 authors and 248 519 ties were active (the other had a stability equal to zero). The first ten authors and co-authorship ties according to the stability to the end of year 2008 are shown in Table I. Stability does not depend only on the number of interactions (and the number of publications consequently) but also depends on how often and how regularly these interactions occur. The calculation of retention and stability for each vertex and edge has linear time complexity with the number of interactions and constant space complexity. Consequently, the calculation is very effective even for large networks.

For further explanation and visualization, we selected an author who has the most entries in DBLP - Philip S. Yu. Figure 3 shows the evolution of retention and stability of this author from the year 1986. As evident, initially the retention decreased rapidly because of the Forgetting curve. Subsequently thanks to his high and regular publication activity, the value settled down and the stability grew almost continuously. In Figure 4, we can see the evolution of stability of the edge with co-author Haixun Wang (the most stable edge of Philip S. Yu at the end of 2008).

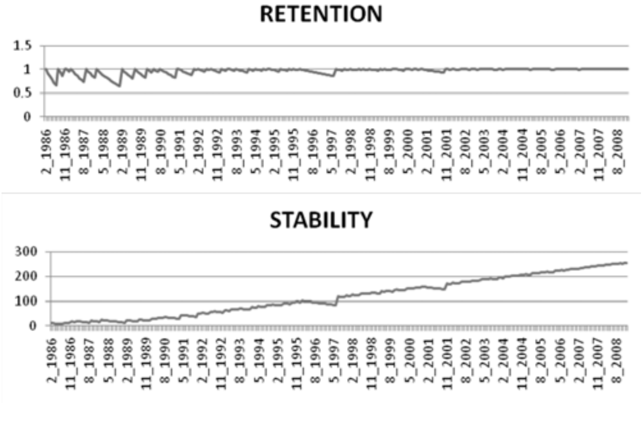

Figure 3. Retention and Stability evolution of Philip S. Yu

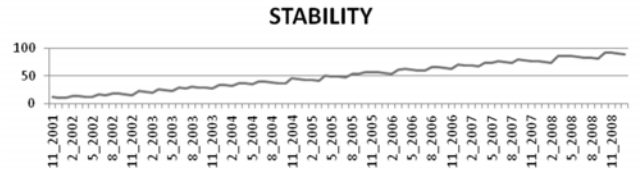

Figure 4. Evolution of Stability between Philip S. Yu and Haixun Wang

B. Network Reduction

The problem of the analysis of networks such as DBLP from the point of view of co-authorship is that the search for deeper relationships and network visualization is complicated due to the large portion of noise caused by many authors and relationships that are insignificant in the long-term. Therefore various methods filtering vertices and edges are used. As an example we considered only authors having at least a specific number of publications or considering only authors and co-authorship which are not older than a certain predefined limit. In our approach the filter works in a natural way. Just remember that the stability coefficient contains the history. This is because the stability in time t is dependent on the previous time. Following the stability, we can describe the publication trend of an author or his relationship (co-authorship) and the value of stability in time t then shows how stable (significant) the author or his relationship is.

Figure 5 contains the sub network focused on the co-authors of Philip S. Yu. The network contains all co-authors and all ties among these authors during the whole publication period to the end of the year 2008 (every author in this sub network has at least one publication with Philip S. Yu in a conference held by IEEE, ACM or Springer). Ten authors with the highest stability in the network are labeled by their names. The Figure on the right side shows the same network, but reduced using a minimum stability of 12. Technically speaking, we have removed all ties with stability less than 12 (i.e. ties which, if not reproduced, cease

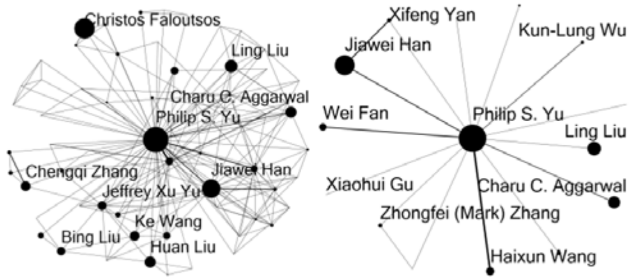

Figure 5. Reduction of network of Philip S. Yu co-authors

Table II
NUMBER OF COMPONENTS AND THEIR SIZE AFTER THE REDUCTION TO STABILITY 12

Vertices	633	119	89	45	43	37	35	34	33	30	29
Components	1	1	1	1	1	1	1	1	1	1	1
Vertices	27	21	19	18	17	16	15	14	13	12	11
Components	1	2	3	3	5	3	9	8	7	14	13
Vertices	10	9	8	7	6	5	4	3	2	**13,372**	
Components	14	25	38	74	122	185	452	875	2,003	**3,867**	

to be active after 12 months). Then we have removed all vertices which lost all ties to Philip Yu. The size of the vertex expresses the stability of the authors while the weight of the edge expresses the stability of represented co-authorship. We can see for example two 3-cliques[3]. It is also apparent at first sight that one of them is more stable (has stronger edges).

C. Network components

The same principle as the previous sample we have used for the whole network on the end of year 2008. Initially, the network had 122,289 active vertices and 248,519 active edges. First of all we have reduced the network to a minimum stability of 12. This means we have removed all edges with a stability lower than 12. Then we have removed all vertices, which lost all their edges. The result can be seen in the Table II. We have identified a total number of 3,867 components, there remained in the network 13,372 vertices only (10.9% active authors).

Figure 6 contains the largest component after the reduction. It contains 633 authors. We can see that the component is composed of backbones containing authors with higher stability. These authors have groups of authors with lower stability in their neighborhood. The detailed inspection reveals that the component contains small loops and cliques only. There are several other authors in the neighborhood of Philip S. Yu, but this region is not in the center of the component.

The figure 7 shows the second largest component fo the reduced network, which contains 119 authors. By its nature, it is similar to the first component. Notice the strong ties between the author Wen Gao and other authors. However, in the center of the component is a different author (Hanging Lu).

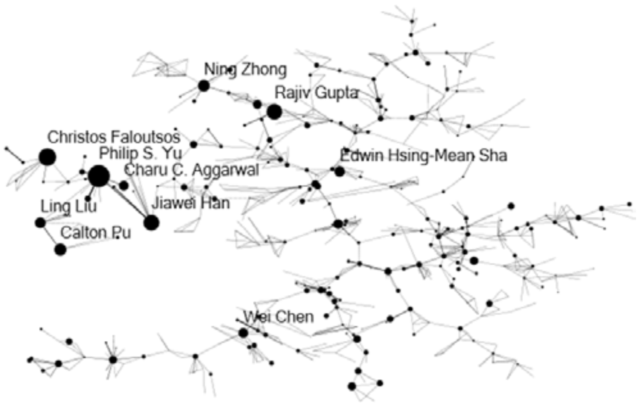

Figure 6. The largest component of the network after the reduction to stability 12

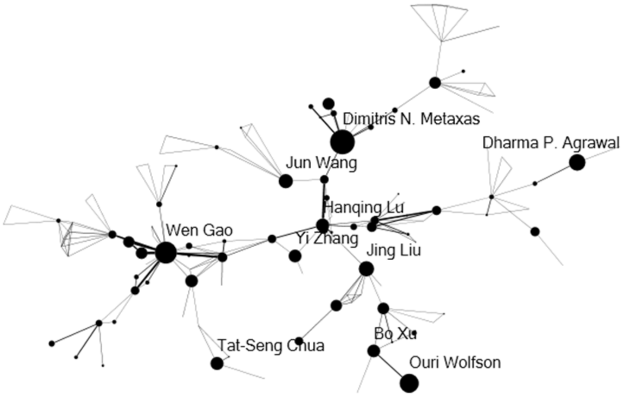

Figure 7. The second largest component of the network after the reduction to stability 12

Table III
NUMBER OF COMPONENTS AND THEIR SIZE AFTER THE REDUCTION TO STABILITY 24

Vertices	16	15	14	13	12	11	10	9
Components	1	0	0	0	1	0	7	3
Vertices	8	7	6	5	4	3	2	**2018**
Components	5	2	8	15	53	134	551	**780**

Figure 8 shows the largest component with sixteen vertices after the reduction of the network to the stability of 24. We have removed all edges with stability lower than 24 and vertices without adequate edges. The results of the reduction are shown in table III. We have identified 780 components, there remained 2,018 vertices after the reduction only (1.65% of active authors). We can find three 3-Cliques with very stable ties. Furthermore we can see the author with 12 very stable ties in the center of the component.

[3] Clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge

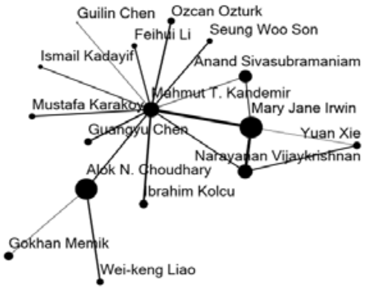

Figure 8. The largest component of the network after the reduction to stability 24

VI. Conclusion

In this article, we introduced two parameters for the vertices and edges of social network - the retention and stability. For the calculation of these parameters we used the Forgetting Curve, which is a well-known approach as a result of experiments with human memory. We view the Forgetting Curve as an heuristic, which allows us to effectively analyze the stability of elements of a social network and also to reduce the network to the most important components. The network works in a similar way as the human brain, which forgets information and also learns new things. Therefore the results of the network analysis vary over time. However the important properties remain and change little. Future research will focus primarily on the comparison with other existing approaches.

Acknowledgement

This work was supported by the Czech Science Foundation under the grant no. 102/09/1494.

References

[1] AL Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert and T. Vicsek: *Evolution of the social network of scientific collaborations*, Physica A: Statistical Mechanics and its Applications, vol. 311, pp. 590–614, 2002

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.U. Hwang: *Complex networks: Structure and dynamics*, Physics Reports, vol. 424, pp. 175–308, 2006

[3] L.F. Costa, FA Rodrigues, G. Travieso and P.R.V. Boas: *Characterization of complex networks: A survey of measurements*, Advances in Physics, vol. 56, pp. 167–242, 2007

[4] H. Ebbinghaus, H.A. Ruger and C.E. Bussenius: *Memory: A contribution to experimental psychology*, 1885/1913

[5] E. Elmacioglu and D. Lee: *On six degrees of separation in DBLP-DB and more*, ACM SIGMOD Record, vol. 34, pp. 33-40, 2005

[6] L.C. Freeman: *Visualizing social networks*, Journal of social structure, vol. 1, 2000

[7] Y. Han, B. Zhou, J. Pei and Y. Jia: *Understanding Importance of Collaborations in Co-authorship Networks*, Proceedings of the 2009 SIAM International Conference on Data Mining, pp. 1112–1123, 2009

[8] R.L. Hart: *Co-authorship in the academic library literature: a survey of attitudes and behaviors*, The Journal of Academic Librarianship, vol. 26, pp. 339–345, 2000

[9] JE Hirsch: *An index to quantify an individual's scientific research output*, Proceedings of the National Academy of Sciences, vol. 102, pp. 16569–16572, 2005

[10] Z. Huang, Y. Yan, Y. Qiu and S. Qiao: *Exploring Emergent Semantic Communities from DBLP Bibliography Database*, Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, pp. 219–224, 2009

[11] T.H. Huang and M.L. Huang: *Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers*, Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation, pp. 18–23, 2006

[12] L. Liu, F. Zhu, C. Chen, X. Yan, J. Han, P. Yu and S. Yang: *Mining Diversity on Networks*, Database Systems for Advanced Applications, pp. 384–398, 2010

[13] X. Liu, J. Bollen, M.L. Nelson, and H. Van de Sompel: *Co-authorship networks in the digital library research community*, Information Processing & Management, vol. 41, pp. 1462–1480, 2005

[14] J. O'Madadhain, J. Hutchins, P. Smyth: *Prediction and ranking algorithms for event-based network data*, ACM SIGKDD Explorations Newsletter, vol. 7, pp. 23–30, 2005

[15] M. Newman: *Who is the best connected scientist? A study of scientific coauthorship networks*, Complex networks, pp. 337–370, 2004

[16] M. Pavlov and R. Ichise: *Finding Experts by Link Prediction in Co-authorship Networks*, 2nd Internation ExpertFinder Workshop, pp. 42–55, 2007

[17] R. Santamarıa and R. Theron: *Overlapping Clustered Graphs: Co-authorship Networks Visualization*, Smart Graphics, pp. 190–199, 2008

[18] J.T. Wixted and E.B. Ebbesen: *Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions*, Memory and Cognition, vol. 25, pp. 731–739, 1997

[19] Q. Ye, B. Wu and B. Wang: *Visual Analysis of a Co-authorship Network and Its Underlying Structure*, Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 4, pp. 689–693, 2008