Logic Journal of IGPL Advance Access published February 8, 2011

Social and swarm aspects of co-authorship network

MILOŠ KUDĚLKA, ZDENĚK HORÁK, VÁCLAV SNÁŠEL,
PAVEL KRÖMER, JAN PLATOŠ and AJITH ABRAHAM, *Department of Computer Science, VŠB-TU Ostrava, Ostrava, Czech Republic.*
E-mail: milos.kudelka@vsb.cz; zdenek.horak.st4@vsb.cz; vaclav.snasel@vsb.cz; pavel.kromer@vsb.cz; jan.platos@vsb.cz; ajith.abraham@ieee.org

Abstract

The analysis of social networks is concentrated mainly on uncovering hidden relations and properties of network nodes (vertices). Most of the current approaches are focused on different network types and different network coefficients. This article introduces a social network analysis based on the so-called Forgetting Curve and Swarm Intelligence inspired by the Ant Colony Optimization. We analyse a co-authorship network and identify two types of ties among its nodes. The Forgetting Curve and Swarm Intelligence are used to model the dynamics of such a network.

Keywords: Social network reduction, dynamic networks, memory, stability, ant colony optimization.

1 Introduction

Publication activities make up a large, complex and steadily growing social network. In such a network, the evolution and dynamics contain much hidden information and implicit knowledge. In this article, we focus on the analysis of two types of ties among authors, namely social ties and ties indicating common professional interests.

One of the major problems in the studies of complex dynamic systems is the description of the behaviour of a whole system with the help of the behaviour of its parts [1]. In this work, we define the behavioural patterns of groups on the basis of social and professional interactions. Social ties are long-term relations. They are created during joint studies, joint work assignments, etc. Professional ties are based on professional interests and they can change in time faster than social relations. The computer science bibliography database (DBLP[1]) is a good example of large dynamic network that contains a lot of implicit information.

We model the social and professional ties using two different methods. Social binds are modelled using the Forgetting Curve and professional relations with the help of social insects behaviour inspired by Ant Colony Optimization. The Forgetting Curve [2] defines the probability that a person can recall information at time t since previous recall. Authors joint publications are interpreted as a reminder and recall of information stored in long-term memory. The result is information about the stability of ties among authors. Regular and long-term joint cooperation is understood as a nearly permanent tie.

The relations based on common professional interests are modelled with the help of methods based on emulation of the behaviour of social insects [1, 3]. The weights of the relations between authors are maintained by the application of pheromones. Each joint publication of two authors causes an increase of the amount of pheromones in the tie between and strengthens the path linking

[1] http://dblp.uni-trier.de/db/index.html.

doi:10.1093/jigpal/jzr002

the two authors. The main aim of this work is to apply known and proven methods of learning and forgetting and swarm intelligence in the field of social network analysis. We introduce novel properties of network nodes and edges and new algorithms for their maintenance.

2 Related work

The analysis of general complex networks is well-described, e.g. in [9]. Liu *et al*. [13] provided a good overview of Social Network Analysis (SNA), co-authorship networks and the application of SNA to co-authorship network analysis. They also compared the results of the analysis using classical SNA coefficients (such as closeness, betweeness, etc.) and PageRank and its modification AuthorRank, respectively.

Hart [12] provided an interesting survey on co-authorship and overview of 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*. [11] introduced the concept of *supportiveness*, which captures co-authorship ties in a non-symmetric way. Elmacioglu and Lee [10] presented statistics calculated from the DBLP data set about conference papers and their authors. They also provided a 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 data set using overlapping groups can be found in Santamaria and Theron [16].

Barabasi *et al*. [8] 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 [14, 15]. A more general approach to the time aspect of social networks is discussed in [18]. This method simultaneously models social network structure and predicts social actions of users. The article also contains a brief survey of dynamic social networks analysis. [18] also contains a particular application of the time-constrained probabilistic factor graph to the field of mining the advisor–advisee relationships from the DBLP data set.

The static analysis of network structure is also an important part of the SNA. Widely used methods of network structure mining include Hyperlink Induced Topic Search (HITS) and PageRank. Both methods were developed for and succesfully applied to the web search and analysis of the trustworthiness of web documents, but they can be used to explore the structure of general networks as well [4, 5].

PageRank is based on the so-called random surfer model. The random surfer starts her session at a random web document. In every document, the surfer can either follow any of the hyperlinks linking the page to the rest of the World Wide Web, or with the probability d, teleport to another randomly chosen web document (including documents not linked from the current one). The concept of teleportation allows documents to be reached that are not linked from any other document on the web [4]. PageRank of a document corresponds to the probability that a random surfer visits the document.

In this article, we introduce a hybrid intelligent approach combining forgetting and swarm intelligence. It is strongly based on the historical data and it uses two different approaches for updating the weights of ties and nodes in the network. Broadly speaking, we define an alternative model of complex network browsing based on the concepts of forgetting and social insect behavioural patterns. The two approaches are used to extract two types of relations from the network—the long-term social links and short-term professional links. The random surfer analogy is hereby replaced by the ant analogy. The method follows the recent trend of hybridization of intelligent methods in order to cope with complex real world data sets and problems [19, 20].

3 Forgetting curve

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 e.g. [2]), which lead to fairly exact description of functions involved in the memorization and the forgetting of information. We 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 investigate the hypothesis that by removing vertices and edges with low weight we can reduce the network but still maintain the important ties.

The forgetting curve defines the probability that a person can recall information at time t since previous recall. It can describe long-term memory and it is usually expressed 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.

There are also different approaches to the computation of the forgetting curve (see for example [17]) but the conclusions are always very similar—the forgetting process is much faster in the beginning.

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

An important feature of the long-term memory is that after reproduced information recall at 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 the long-term memory is that immediate (too early) reproduced recall of information has no bigger effect on the learning. On the other hand, the too late reproduced recall (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. For updated S_{new}, after new information recall should hold:

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

For reproduced information recall is $R=1$. It is caused by 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 $F \in (1;2\rangle$. For the calculation of S_{new}, we have to consider three things:

Function opt(S) for the calculation of optimal information recall time. 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. $F=1.2$)

The function $ch(t,S)$ **for calculation of** S_{new}**.** 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 $ch(t,S,F,S_{\text{ini}})$ for the calculation of the coefficient of change of the value S. Then holds: $S_{\text{new}}=ch(t,S,F,S_{\text{ini}})\cdot S$.

There are various approaches for the computation of the value of function $ch(t,\ S,F,S_{\text{ini}})$. In this work we use simple evaluation based on linear functions (see Figure 1):

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

3.1 *Forgetting of social network*

We assume that the interactions between particular pairs of vertices take place in the social network continuously. If we understand these interactions as an experience stored in the 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 differently stable ties.

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

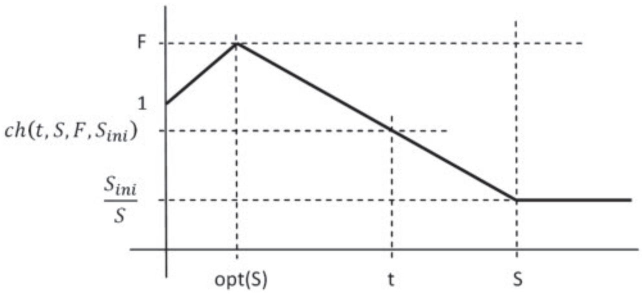


FIG. 1. Calculation of S in time t.

recall). If we understand the network as an analogy to the human brain, then the memorization of ties in the network corresponds to the degree of remembering the information in the brain.

The properties of ties change over time, depending on how often and in what time the 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 vertex and tie we defined the following time-changing characteristics: *Edge Stability ES* is the estimated time for which the tie between vertices remains active (since given time t). *Active Edge* is a tie, for which holds that $ES > 0$ in given time t. *Vertex Stability VS* is the estimated time for which the vertex remains active (since given time t). *Active Vertex* is a vertex, for which holds that $VS > 0$ in given time t.

4 Swarm intelligence and ant colony optimization

Swarm Intelligence is a research area dealing with the design of multiagent systems inspired by the intelligent behaviour of social insects. It is concerned with models of global behavioural patterns rather than with the design of a sophisticated controller that would govern the application. A swarm system consists of many unsophisticated agents that cooperate in order to achieve desired behaviour [7].

Ant Colony Optimization (ACO) [3] is a popular meta-heuristic method based on certain behavioural patterns of foraging ants. Emulation of ants' behaviour can be used as a probabilistic computational technique for solving complex problems that can be reduced to finding optimal paths in graphs [3]. An artificial ant k placed in vertex i moves to node j with probability p_{ij}^k:

$$p_{ij}^k = \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{l\in N_i^k}(\tau_{il}^{\alpha}\eta_{il}^{\beta})} \tag{1}$$

where N_{ik} represents the neighbourhood of ant k in node i (i.e. nodes that are available to move on), τ_{ij} represents the amount of pheromones placed on arc a_{ij} and η_{ij} corresponds to a priori information reflecting the cost of passing arc a_{ij}. After the ants finish their movement forward, they return to the nest with food. The tour of ant k is denoted as T^k. The length of T^k, called C^k, or the amount of food collected called L_k (i.e. the solution quality), is used to specify the amount of pheromones to be placed by ant k on each arc on the trail that led to the food source:

$$\Delta\tau_{ij}^k = \begin{cases} \frac{1}{C^k} & \text{if arc } (i,j) \text{ belongs to } T^k \\ 0 & \text{otherwise} \end{cases}, \quad \tau_{ij} = \tau_{ij} + \sum_{k=1}^{m}\Delta\tau_{ij}^k \tag{2}$$

After all the ants finish one round of their movement, the pheromones evaporate (i.e. the amount of pheromones on each arc is reduced). It can be expressed using the equation $\tau_{ij} = (1-\rho)\tau_{ij}$. The coefficients α, β and ρ are general parameters of the algorithm. This basic version of the ACO algorithm is called an ant system (AS).

In this work, we use a swarm-intelligent algorithm based strongly on the ACO method. We interpret each joint publication of two authors in a co-authorship network as if an ant would pass the link between the authors and update its pheromones. Moreover, when the two authors do not publish a joint article in a time period, the pheromones on the link between them evaporate and their connection becomes weaker.

5 Experiments with the DBLP data set

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 data set in XML[2] and pre-processed it for further usage. First of all, we selected all conferences held by IEEE, ACM or Springer, which gave us 9768 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 article 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.

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 article 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 data set.

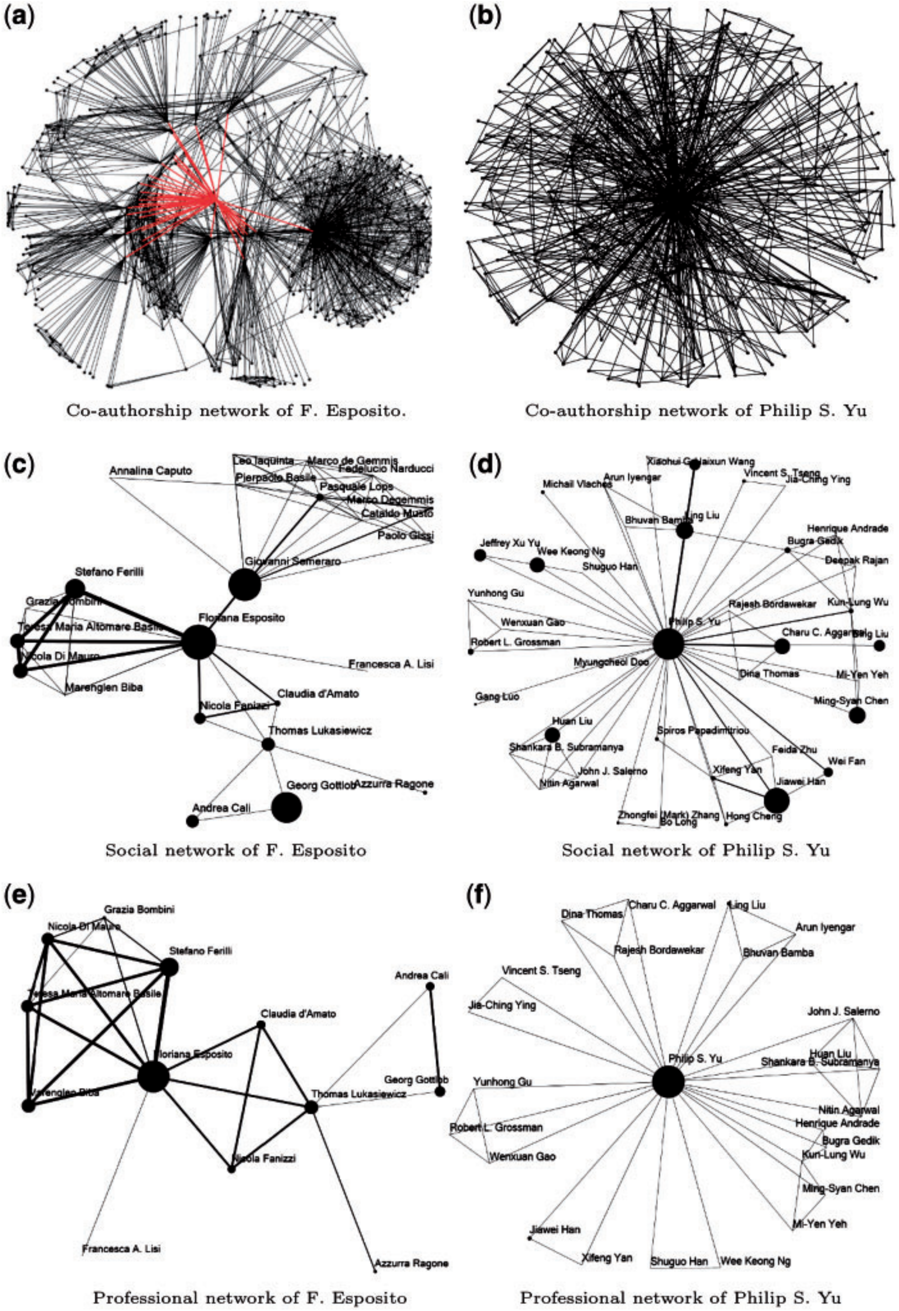
The ant colony metaphor was applied to the co-authorship network as follows: each joint publication of two authors triggers an ant passing between the two nodes. When an ant traverses the arc between two authors, the amount of pheromones on both, author nodes and the link in between, is increased by one. The whole ACO implementation is:

1. Initialization: initialize the value of pheromones to zero for all vertices and arcs in the network
2. Ant movment: each month, when an ant passes the arc, the value of pheromones is increased $\tau_{ij} = \tau_{ij} \cdot C$ where $C > 0$ is an constant. We have used $C = 1.2$.
3. For all objects visited by an ant, let $\tau_{ij} = S_{\text{ini}}$ iff $\tau_{ij} < S_{\text{ini}}$. This step assigns an extra amount of pheromone to objects that were interacting for the first time $S_{\text{min}} = 12$.
4. Evaporation: for each object that is not visited by an ant in given month, let $\tau_{ij} = \tau_{ij}(1 - \delta)$ where $0 < \delta < 1$. We have used $\delta = 0.25$.
5. For each idle (i.e. not visited) object, let $\tau_{ij} = 0$ iff $\tau_{ij} = S_{\text{min}}$. This step erases longer unused connections and removes co-authors that are no longer actively linking with author.

The values of the coefficients C, S_{ini} and S_{min} were set with respect to the settings of Forgetting Curve. The first joint work of two authors creates a tie that is remembered by the system for at least 12 months.

The DBLP data set has been used to observe the dynamics of social and professional ties of two authors in the DBLP. Floriana Esposito and Philip S. Yu were selected after the analysis of activity (number of records in the database) and stability of authors in the network. Floriana Esposito is an author who has been active since 1990 and who has a lot of strong ties (in terms of stability) and Philip S. Yu is an author with highest number of records in the DBLP data set and who has a number of strong co-authors. The two level network of co-authors of Floriana Esposito (F. Espositos co-authors and their co-authors) is shown in Figure 2a. The links between F. Esposito and her co-authors are shown in red (online). Figure 2c shows the network of co-authors of F. Esposito found by the Forgetting Curve. The five most important co-authors and five strongest ties in reduced network are listed in Table 1. This set of authors and ties represents in our model the social interactions (long-term social network) of F. Esposito. Figure 2e displays the network of co-authors of F. Esposito

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



Co-authorship network of F. Esposito.

Co-authorship network of Philip S. Yu

Social network of F. Esposito

Social network of Philip S. Yu

Professional network of F. Esposito

Professional network of Philip S. Yu

FIG. 2. The networks of Floriana Esposito and Philip S. Yu.

TABLE 1. First five co-authors and first five strongest ties in the social network of Floriana Esposito

No.	Co-author	Weight	Collaboration	Weight
1	F. Esposito	174	F. Esposito - S. Ferilli	101
2	G. Semeraro	164	S. Ferilli - T. M. Altomare Basile	82
3	G. Gottlob	155	S. Ferilli - N. Di Mauro	82
4	S. Ferilli	107	T. M. Altomare Basile - N. Di Mauro	82
5	T. M. Altomare Basile	82	F. Esposito - T. M. Altomare Basile	75

TABLE 2. First five co-authors and first five strongest ties in the professional network of F. Esposito

No.	Co-author	Weight	Collaboration	Weight
1	F. Esposito	26	F. Esposito - S. Ferilli	16
2	S. Ferilli	16	F. Esposito - M. Biba	12
3	T. Lukasiewicz	12	S. Ferilli - M. Biba	12
4	M. Biba	12	F. Esposito - N. Di Mauro	12
5	N. Di Mauro	12	F. Esposito - T. M. Altomare Basile	12

TABLE 3. First five authors and first five strongest ties in the social network of Philip S. Yu

No.	Co-author	Weight	Collaboration	Weight
1	P. S. Yu	266	H. Wang - Philip S. Yu	78
2	Jiawei Han	222	C. C. Aggarwal - P. S. Yu	61
3	Ling Liu	155	Xifeng Yan - Jiawei Han	52
4	Ming-Syan Chen	151	P. S. Yu - Jiawei Han	45
5	C. C. Aggarwal	144	P. S. Yu - Kun-Lung Wu	45

TABLE 4. First five authors and first five strongest ties in the professional network of Philip S. Yu

No.	Co-author	Weight	Collaboration	Weight
1	Philip S. Yu	288	Y. Gu - R. L. Grossman	14
2	Ling Liu	59	John J. Salerno - Philip S. Yu	12
3	Jiawei Han	59	Huan Liu - Philip S. Yu	12
4	Huan Liu	17	Nitin Agarwal - Philip S. Yu	12
5	R. L. Grossman	14	R. L. Grossman - Philip S. Yu	12

reduced by the swarm intelligence. The five most important co-authors and five strongest ties in the network of F. Esposito reduced by ants are listed in Table 2.

Figure 2b displays the network of co-authors of Philip S. Yu, the DBLP author with highest number of records. Figure 2d and Figure 2f displays the social network of P. S. Yu and the professional network of P. S. Yu respectively. The list of five most important co-authors of P. S. Yu and his five most emphasized ties in the social network are listed in Table 3. The same information from his professional network is shown in Table 4.

6 Conclusion

In this article, we studied two aspects of a complex and time-dependent co-authorship network—the DBLP. With the help of the Forgetting Curve We have mined the long-term (social) ties that connect authors in an almost permanent manner. Next, we have used the swarm intelligence and ant colony optimization to detect professional links that indicate recent cooperation of two authors. Such a professional link does not guarantee future joint authorship between the two authors. However, a strong social link yields high probability of a future joint publication.

We have sought for social and professional networks of two members of the DBLP. Floriana Esposito and Philip S. Yu were selected with the aim to investigate social and professional network of two different authors. Floriana Esposito represents a social type of author. While F. Esposito has strong ties (in the sense of stability), Phillip S. Yu is an example of an active author with a high number of colaborators.

In both cases, the algorithm delivered different social and professional network. As apparent from the tables, the social network contains authors with high weight and strong ties (i.e. long-term partners). The weights of both, co-authors and co-authorship links in social network, varies significantly. The professional network contains authors and connections with approximately the same weight. They are recent collaborators of the studeied authors.

Funding

This work was supported by the Czech Science Foundation under the grant no. 102/09/1494 and 201/09/0990.

References

[1] D. M. Gordon. *Ant Encounters: Interaction Networks and Colony behaviour*. Princeton University Press, 2010.

[2] H. Ebbinghaus. *Memory: A Contribution to Experimental Psychology*. Teachers College, Columbia University, 1885/1913.

[3] M. Dorigo and T. Stützle. *Ant Colony Optimization*. MIT Press, 2004.

[4] I. H. Witten, M. Gori, and T. Numerico. *Web Dragons: Inside the Myths of Search Engine Technology*. Morgan Kaufmann, 2006.

[5] S. Brin and L. Page. 'The anatomy of a large-scale hypertextual web search engine.' 1998, Available at http://infolab.stanford.edu/~backrub/google.html.

[6] J. M. Kleinberg. 'Authoritative sources in a hyperlinked environment.' *Journal of the ACM*, **46**, 604–632, 1999.

[7] C. Blum and D. Merkle. *Swarm Intelligence: Introduction and Applications*. Springer, 2008.

[8] A. L. 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*, **311**, 590–614, 2002.

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

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

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

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

[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*, **41**, 1462–1480, 2005.

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

[15] M. Pavlov and R. Ichise. *Finding Experts by Link Prediction in Co-authorship Networks*. In *Second Internationsl ExpertFinder Workshop*, pp. 42–55. CEUR Workshop Proceedings, 2007.

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

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

[18] C. Tan, J. Tang, J. Sun, Q. Lin, F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1049–1058. ACM, 2010.

[19] A. Abraham, E. Corchado, and J. M. Corchado. 'Hybrid learning machines.' *Neurocomputing*, **13–15**, 2729–2730, 2009.

[20] E. Corchado, A. Abraham, and A. de Carvalho. 'Editorial: Hybrid intelligent algorithms and applications.' *Information Science*, **180**, 2633–2634, 2010.

Received 15 November 2010