Forcoa.NET: An Interactive Tool for Exploring the Significance of Authorship Networks in DBLP Data

Zdenek Horak, Milos Kudelka, Vaclav Snasel, Ajith Abraham
Department of Computer Science
FEI, VSB - Technical University of Ostrava
17. listopadu 15, 708 33, Ostrava - Poruba, Czech Republic
{zdenek.horak, milos.kudelka, vaclav.snasel}@vsb.cz
ajith.abraham@ieee.org

Hana Rezankova
Department of Statistics and Probability
University of Economics, Prague
W. Churchill sq. 4, 130 67 Praha 3, Czech Republic
rezanka@vse.cz

***Abstract*—This paper presents an online analysis tool called Forcoa.NET, which is built over the DBLP dataset of publications from the field of computer science. The developed tool is focused on the analysis and visualization of the co-authorship relationship based on the intensity and topic of joint publications. The visualization of co-authorship networks allows to describe the author and his/her current surroundings while still incorporating the historical aspect. The analysis is based on using the forgetting function to hold the information relevant to the selected date. After this analysis, we are capable of computing several measures, which can describe different aspects of user behaviour from the point of view of scientific social network.**

I. Introduction

During the analysis of authors involved in the scholarly publication of articles over a long period of time, the actuality and clarity of the view is often lost. Our aim is to create a view, containing only information relevant for the selected time. For the selection of relevant information we have used the forgetting function, where the basic idea is - the important information prevail, the non-important ones vanish. This process is formally expressed using the stability measure based on the use of forgetting curve [9]. The stability measure characterizes the behaviour of the author/user in the network (if the author publishes regularly and in the long term. We have also performed several experiments with different types of the Forgetting function [10].

The stability is used in two ways - the stability of authors (the vertices of the network) and the stability of the relations between authors (the edges of the network). The stability is the basic measure in our point of view, but in this article we present several other measures based on the stability.

II. Related work

The analysis of general complex networks is well-described in [2] and [3]. Liu et al. [11] provides a good overview of Social Network Analysis, co-authorship networks and a combination of both. They also compared the results of the analysis using classical SNA and PageRank and its modification AuthorRank, respectively. Further coefficients can be found; e.g. in Newman [12]. Hart [7] provided an interesting survey on co-authorship and its models, grounds, etc. Han et al. [6] introduced the concept of supportiveness, which captures co-authorship ties in a non-symmetric way. Huang and Huang [8] addressed 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 along with a comparison of weighted and unweighted variants of SNA coefficients used to identify important authors. Opsahl et al. [13] discusses two aspects of weighted networks the number of ties versus their strength and presents an approach combining these two aspects. Some of our proposed measures benefit from this combination too. The architecture of weighted networks is investigated in [1] along with several measures constructed as an extension of classic network analysis coefficients for weighted networks.

DBL-Browser[1] is a desktop application allowing browsing of the data in the DBLP dataset. System can also visualize relations between authors and between conferences. DBLife[2] is an online tool which crawles various Internet sites and gathers information about authors and their publications. Authors of [14] presented the DBConnect system focused on mining community-related information from the DBLP dataset based on the social network analysis approach. ArnetMiner[3] is probably the most advanced scientific network analysis software. It is capable of analysing authors, conferences, topics and their relations. Results can be visualized using network. This system is also capable of computing several measures and their charts.

In comparison with ArnetMiner, our system is strongly focused on the automatic analysis of interaction between authors and the evolution of their relationships. We do not measure quality of anything, instead we present several measures describing the role and behaviour of authors in the network. In the following Sections, we present the user interface of the proposed system, the mathematical background behind our analysis and several results achieved by our system.

This work was supported by the Czech Science Foundation under the grant no. GA201/09/0990.

[1] http://dbis.uni-trier.de/DBL-Browser/
[2] http://dblife.cs.wisc.edu/
[3] http://arnetminer.org

978-1-4577-1131-2/11/$26.00 ©2011 IEEE

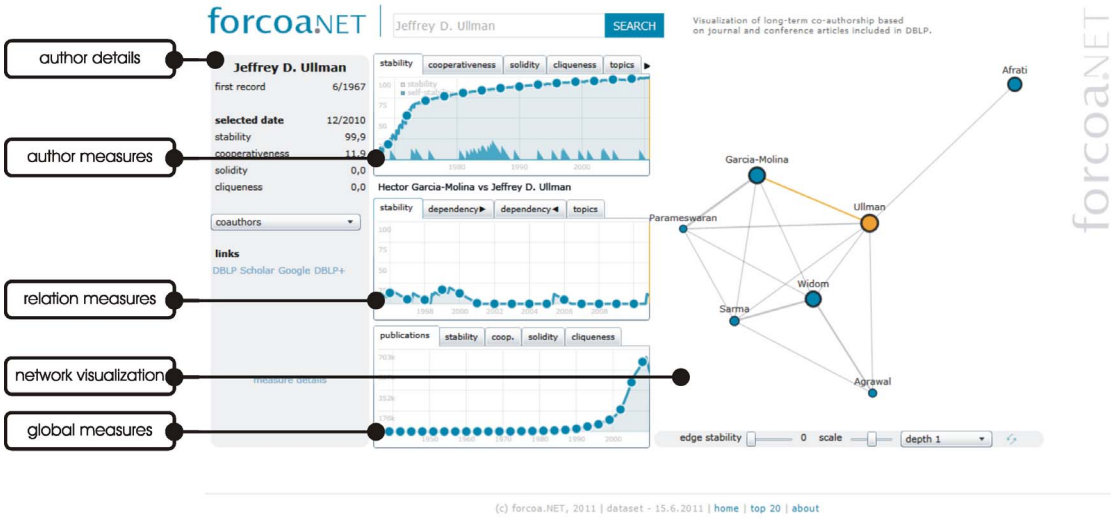

Fig. 1. Forcoa.NET screenshot with basic features highlighted

III. Forcoa.NET Brief description

A. Dataset

The Forcoa.NET system is built over the DBLP dataset[4]. The preprocessing of this dataset is described in our previous paper [9]. Currently, the dataset contains information about 913,534 authors from the field of computer science and 5,192,020 interactions between these authors.

B. Architecture

The design and development of the system was driven by the requirement for the visualization of co-authorship data. A key requirement was the need for the visualization of an author and his/her surroundings in the context of their publication activities, including a simple animation for the visualization of historical data. To fulfill these requirements, one has to work with a large amount of data. Therefore the architecture is adapted to distribute the computation between server and client. The server performs all computations related to the data aggregation, while the client computes all visualization-related data. The communication between server and client is based on Web-services; client application is based on the Silverlight technology (see Figure 2).

C. User interface

Figure 1 presents a screenshot of the Forcoa.NET user interface with one particular author (Jeffrey D. Ullman[5]) selected. In the left part of the interface is a panel containing author details, such as first record in the network and values of several metrics w.r.t. selected date. This panel contains also a combobox with coauthors and direct links pointing to details about author from different sites.

[4]http://dblp.uni-trier.de/xml/

[5]Jeffrey D. Ullman has been selected as having the highest stability in our system.

The right part of the interface contains the visualization of the authors social network with current author highlighted. The network can be filtered using some minimum edge weight (see below) or can be switched to different network view. The default network view contains coauthors to depth 1. The view can be switched to depth 2, but also to the so-called dependency or independency network (showing only dependent or independent coauthors; for detailed explanation see below).

The middle part contains three groups of panels. The first group contain values of author measures (such as stability, cooperativenes, topics, etc.) over the time period. If you select an edge with a coauthor, the second group of panels will appear. These panels contains values of relation measures (stability of the relation, dependency between authors and the topics of the relation). By clicking somewhere in the timeline you can switch the view to different point in time. The third group of panels contains global values of the whole dataset (such as number of publications, distribution of particular measures over the authors.

IV. Forcoa.NET Notions

In the following sections, we use the duality between authors (nodes of the network) and their relationships (edges of the network). These concepts are used interchangeably.

A. Forgetting curve

We understand the term social network as an undirected weighted graph. During the calculations of edge and vertex weights, we use time-dependent information and values related to forgetting. The forgetting curve (see Figure 3a) 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 equation $R = e^{-\frac{t}{S}}$, where R (memory retention) is the probability of recalling information at time t since the last recall, e is the Euler number, t is the

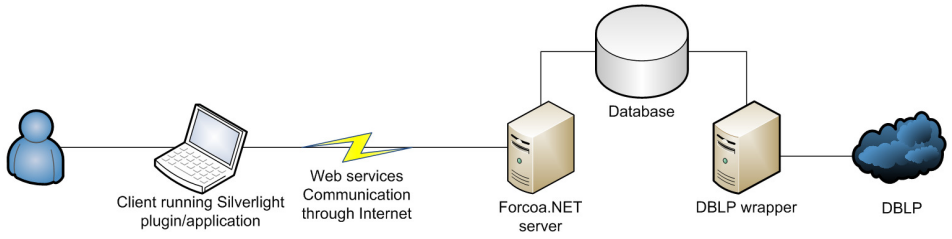

Fig. 2. Forcoa.NET architecture

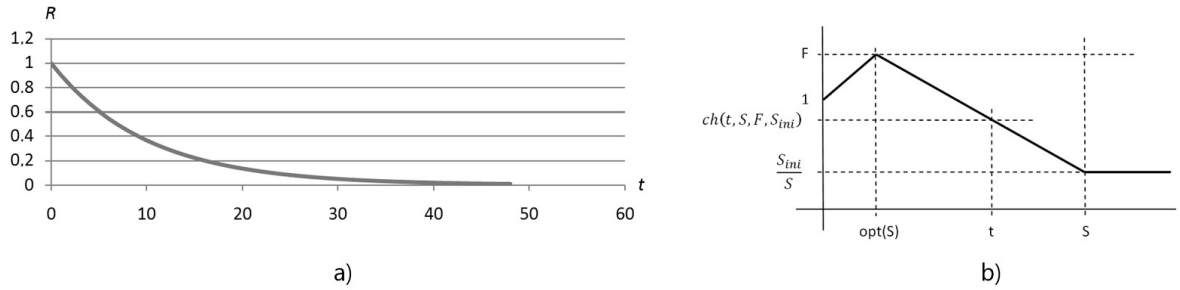

Fig. 3. a) Forgetting curve, b) calculation of $ch(t, S, F, S_{ini})$ in time t

time since the last recall and S (relative strength of memory; stability) is the approximated time since the last recall for which the information is stored in memory. 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 information for the first time, then the time of storing information in memory is $S_{ini} > 0$ and this default value is constant. An important feature of long-term memory is that after 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 t of recall. 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 big effect on the learning. On the other hand, the 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. Therefore for the calculation of S_{new} we have to consider several things. For the purpose of this paper we can simplify the process to the equation $S_{new} = ch(t, S, F, S_{ini}) \cdot S$. Function ch expresses the abstraction of the remembering / forgetting process. One choice of the ch function (based on linear functions) is illustrated in Figure 3b. The detailed description of our approach can be found in [9].

V. Measures

In comparison to classical SNA measures (such as centrality, clustering coefficient, etc.) we focus on the usage of edge and vertex stability (similar approach as in [13]). Every our presented measure describes a particular type of vertex (author) behaviour through the time.

A. Stability

Interactions between particular pairs of vertices continuously take place in the social network. If we consider these interactions as an experience stored in memory, then the ties between the 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 the vertices and the tie between them are. Therefore we can understand the social network as a set of variously stable vertices and ties. The properties of vertices and 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 an analogy to the learning and forgetting of reproduced information. For each vertex and tie we define two time-changing characteristics: **Edge stability** ES is the estimated time for which the tie remains active (since given time t), while Active edge is a tie, for which holds that $ES > 0$ in time t. **Vertex stability** VS is the estimated time for which the vertex remains active (since time t), while Active vertex is a vertex, having $VS > 0$ in time t.

Remark: Due to technical reasons, we use logarithmic scale for the stability measure.

B. Self-stability

As self-stability of an author, we denote the stability of the edge from the author to himself/herself (so-called loop).

This edge stores information about publications where no other coauthor took place.

C. Cooperativeness

The basic motivation is to describe the vertex in relation to important (stable) vertices in its surroundings. The current stability of the vertex, which may be rather high, is not solely dependent on the number of ties or the strength of these ties (e.g. authors with high stability may have no co-authors). As an important tie for its surroundings we consider a tie having an adjacent vertex with high stability. The cooperativeness also takes into account the stability of the relation between these vertices:

$$\textbf{Cooperativeness}(v) = \sum_i \sqrt{ES(e_i) \cdot VS(v_i)},$$

where e_i and v_i are edges and vertices adjacent to the vertex v.

D. Solidity

The aim of the solidity is to give priority to strong ties over the weak ones. As a consequence, the solidity describes vertices whose ties to their surroundings are strong. Solidity takes into account only ties having at least some minimal stability:

$$\textbf{Solidity}(v, stab) = \sum_i (ES(e_i) - stab),$$

where ei are edges adjacent to the vertex v, for which holds $ES(e_i) > stab$. In our experiment we have used $stab = 12$ (as a default value of stability S_{ini}).

E. Dependency

An interesting feature of co-authors may be their relationship in the stronger/weaker terminology. This is a measurement of a relationship with another author in the context of the group of his/her co-authors. This value indicates how stable (w.r.t. to the stability) the ties between the vertex and the neighbors of the second vertex are (incl. the tie between the inspected vertices). This is a non-symmetric measure; at least one common neighbor is the precondition for nonzero value.

$$\textbf{Dep}(v_1, v_2) = \frac{ES(v_1, v_2) + \sum_{e_i \in CE(v_1, v_2)} ES(e_i) \cdot \mathbf{R}(e_i)}{\sum_{e_i \in E(v_1)} ES(e_i)},$$

$$\mathbf{R}(e_i) = \frac{ES(v_1, v_i)}{ES(e_i) + ES(v_1, v_i)},$$

where $CE(v_1, v_2)$ is a set of edges between vertex v_1 and vertices v_i, which are adjacent to the vertex v_2, $\mathbf{R}(e_i)$ is a relative dependency of the vertex v_1 on the vertex v_2 through the vertex v_i and $E(v_1)$ is a set of all edges adjacent to the vertex v_1.

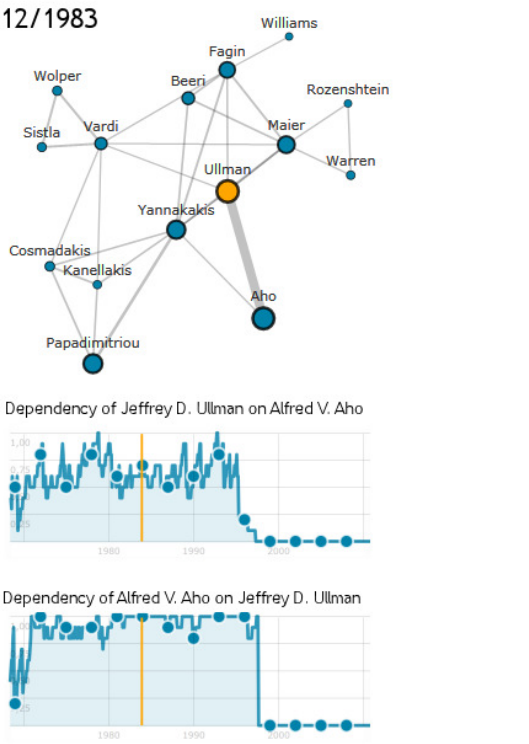

Fig. 5. Forcoa.NET network (for depth 2) of Jeffrey D. Ullman on 12/1983 (top) the evolution of the dependency value between Jeffrey D. Ullman and Alfred V. Aho (bottom)

F. Topics

For each author or relationship between authors we are able to identify topics which can describe this author or relation. From [4] we have obtained a list of topics which were later detected in the titles of the articles. For each author we have aggregated the topics from his articles. The topics of the relation between authors are detected as a topics of joint articles of these authors.

a) Remark: Mentioned measures together with their historic development is illustrated in Figure 4. This figure focused again on Jeffrey D. Ullman is divided into four parts, where each one of them represents one particular date. The first row contains visualization of the surroundings of Jeffrey D. Ullman in years 1968, 1984, 2005 and 2010. On the second row you can find the development of author topics over the years. Last row contains charts of aforementioned measures calculated for this author having highlighted dates corresponding to the network state from the first row. It can be easily seen, that the publication activity is still stable, but the coauthorship model has changed from regular cooperations with several authors (1984) through publications with large group of authors (2005) upto now. Using the Forcoa.NET tool you can also online watch the animation of authors history.

Figure 5 illustrates the dependency measure. Note, that our dependency is strongly based on author interactions, which means, that two coauthors have their dependency based on the

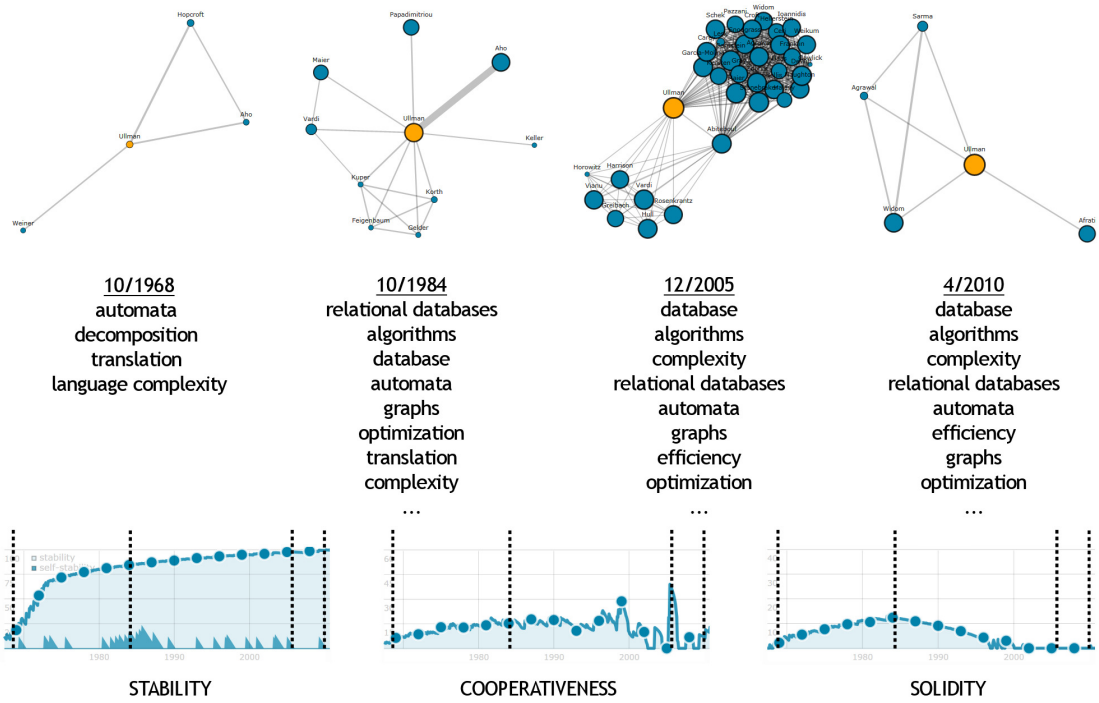

Fig. 4. Jeffrey D. Ullman in Forcoa.NET history

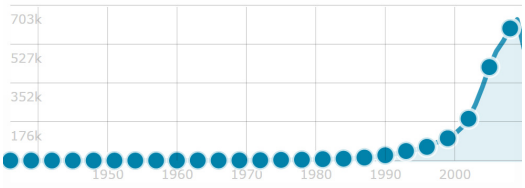

Fig. 6. Number of publications in the DBLP dataset for different years

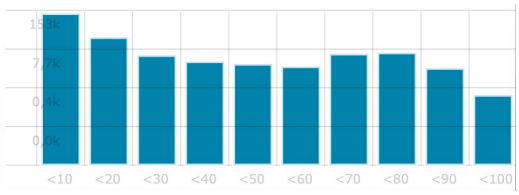

Fig. 7. Stability histogram - number of authors in the dataset according to different stability ranges (using log-plot)

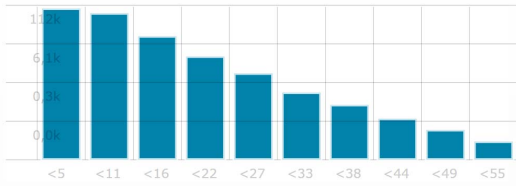

Fig. 8. Cooperativeness histogram - number of authors in the dataset according to different cooperativeness ranges (using log-plot)

number and stability of their publications with other coauthors. The figure shows the Forcoa.NET network of Jeffrey D. Ullman from the year 1983. It is apparent, that besides the strong coauthorship relation with Alfred V. Aho he had also relations to other coauthors (contrary to Mr. Aho). Therefore the dependency chart shows higher dependency of Mr. Aho on Mr. Ullman than vice versa.

G. Dataset statistics

To illustrate the properties of the DBLP dataset and the authors in computer science, we can present several statistics. The chart in Figure 6 illustrates the number of publications in the DBLP dataset over the years. Note that we consider conference proceedings and journal records only. The completeness of the DBLP dataset also remains a question. It can be easily seen, that the number of publications in the field of computer science per year started to grow in the 1990's (23 thousand publications) with rapid increase from the year 2000 (100 thousand publications) with recent value of 600 thousand publications in the year 2008.

Figure 7 shows the distribution of the stability among all authors in the dataset (w.r.t. to the end of the year 2010). It can be easily seen, that there are more than 150 thousand of authors (out of the total number of 913,534 authors) with the stability lower than 10, but there are only 281 authors having the stability between 90 and 100.

Figure 8 contains the histogram of the cooperativeness measure. The largest portion of authors (more than 110 thousand) has the cooperativeness value lower than 5 and the value decreases logarithmically.

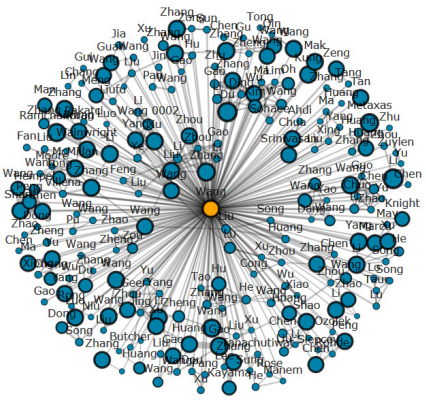

Fig. 9. Forcoa.NET Wei Wang network (12/2010)

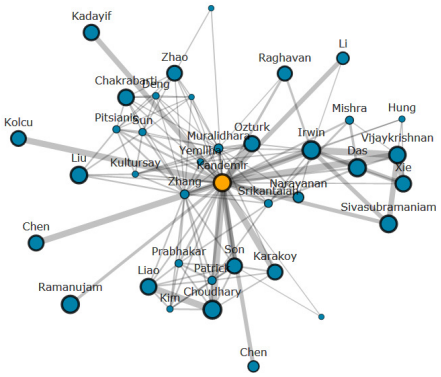

Fig. 10. Forcoa.NET Mahmut T. Kandemir network (12/2010)

H. Author types

Using the presented visualization of authors and their surroundings we have noticed, that there exists several types of authors. Therefore we have employed several statistical methods to confirm our hypothesis. Using factor analysis and K-means clustering we have identified several features present in the behaviour of authors. Due to the limited scope of this paper we cannot discuss these results in detail, we will only show some of them.

There exists a group of authors having many relations to other stable authors, but these relations are relatively weak (with low stability). As an example we can mention Wei Wang, her Forcoa.NET network is shown in Figure 9.

Another group consists of authors with large number of strong relations (they publish often together). As an example we show the network of Mahmut T. Kandemir in Figure 10.

Finally we illustrate relatively large groups of authors which publish in the long term together. This situation is illustrated on the network of Jose J. Pazos Arias in Figure 11.

VI. Conclusion

In this article we have described a novel tool for the study of behaviour of computer science authors. We present several measures similar to the classical SNA measures which can be used for the description of the closest surroundings of authors. All the information presented in this article can be easily verified using the www.forcoa.net site. Readers are welcomed to explore the online portal and any feedback would be greatly appreciated. In the future we would like to further extent the range of our analysis, mostly from the point of view of communities.

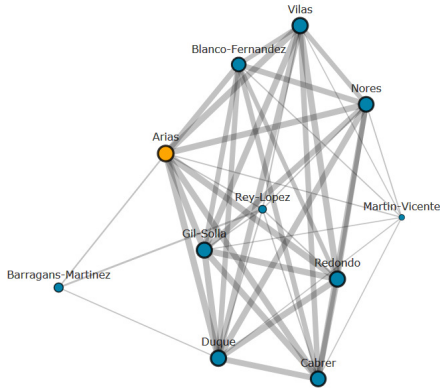

Fig. 11. Forcoa.NET Jose J. Pazos Arias network (12/2010)

References

[1] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, *The architecture of complex weighted network*, Proceedings of the National Academy of Sciences of the United States of America, vol. 101, pp. 3747-3752 (2004)

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

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

[4] J. Diederich, W.T. Balke, U. Thaden, *Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp*, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 505–505 (2007)

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

[6] Y. Han, B. Zhou, J. Pei, Y. Jia, *Understanding Importance of Collaborations in Coauthorship Networks*, SIAM Int. Conference on Data Mining, pp. 1112-1123 (2009)

[7] 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)

[8] T.H. Huang, M.L. Huang, *Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers*, Int. Conference on Computer Graphics, Imaging and Visualization, pp. 18-23 (2006)

[9] M. Kudelka, Z. Horak, V. Snasel, A. Abraham, *Social Network Reduction Based on Stability*, International Conference on Computational Aspects of Social Networks (2010)

[10] M. Kudelka, Z. Horak, V. Snasel, P. Kromer, J. Platos, A. Abraham, *Social and swarm aspects of co-authorship network*, Logic Journal of IGPL (2011)

[11] X. Liu, J. Bollen, M.L. Nelson, H. Van de Sompel, *Co-authorship networks in the digital library research community*, Information Processing & Management, vol. 41 (2005)

[12] M. Newman, *Who is the best connected scientist? A study of scientific coauthorship networks*, Complex networks, pp. 337-370 (2004)

[13] T. Opsahl, F. Agneessens, J. Skvoretz, *Node centrality in weighted networks: Generalizing degree and shortest paths*, Social Networks, vol. 32, pp. 245-251 (2010)

[14] O.R. Zaiane, J. Chen, R. Goebel, *DBconnect: mining research community on DBLP data*, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 74–81 (2007)