Evolution of Co-Authors Communities Formed by Terms on DBLP

Alisa Babskova, Pavla Dráždilová, Jan Martinovič, Václav Svatoň, and Václav Snášel

VŠB – Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
17. listopadu 15/2172, 708 33 Ostrava, Czech Republic
{alisa.babskova.st,pavla.drazdilova,jan.martinovic}@vsb.cz
{vaclav.svaton.st,vaclav.snasel}@vsb.cz

Abstract. The DBLP Computer Science Bibliography server provides bibliographic information on major computer science journals and proceedings. DBLP indexes more than 2.1 million articles and contains titles of articles, their authors, years of publication etc. Downloadable DBLP dataset is very interesting resource for evolution analysis of co-author networks. The paper deals with subgraphs of the authors from DBLP with common interests. The common interest of the authors is defined by terms, which are extracted from the titles of articles. The subgraphs are extracted for each year separately based on the published years. These subgraphs represent the communities of co-authors, for which is observed their development in time. That new view of these communities of the co-authors offer a new way for analysis and measurement of article datasets.

1 Introduction

The aim of this paper was to develop a methodology for finding, tracking, analysing and evaluating the development of the groups of authors who deal with the areas specified by chosen terms. We can see whether this area is still developing, expires, is stable or promising. The results of this paper could be used by researchers to point their professional interest.

Our work has been inspired by papers in which the authors tried to analyse dynamic aspects of communities. Authors present in the paper [5] a framework for modelling and detecting community evolution over time. They proposed the community matching algorithm which efficiently identifies and tracks similar communities over time. A series of significant events and transitions is defined to characterize the evolution of networks in the terms of its communities and individuals. The authors also propose two metrics called stability and influence metrics to describe the active behaviour of the individuals. They present experiments to explore the dynamics of communities on the Enron email and DBLP datasets.

In the paper [14] authors construct word association network from DBLP bibliography records based on word concurrence relationship in titles and analyse statistical distribution of edge frequency. The authors find that frequency distribution of the word also satisfy power-law distribution.

V. Snášel, K. Richta, J. Pokorný (Eds.): Dateso 2013, pp. 109–118, ISBN 978-80-248-2968-5.

The paper [3] was written to address the question which communities will grow rapidly, and how do the overlaps among the pairs of communities change over time. In the paper were used two large sources of data: friendship links and community membership on LiveJournal, and co-authorship and conference publications in DBLP. Authors of this work studied how the evolution of these communities relates to properties such as the structure of the underlying social networks.

In the article [8] authors show an interesting metrics for evaluating communities evolving in time. For their experiment they consider data sets of the monthly list of articles in the Cornell University Library e-print condensed matter archive and the record of phone calls between the customers of a mobile phonecompany. About this metrics we will talk more in Section 4. In this article, the proposed metrics are used for evaluation of communities of co-authors that were extracted from DBLP dataset (see Section 5).

The study of the dynamic evolution is relatively new subject in the research of the social communities. The research of this paper is focused to study the communities extracted from the DBLP dataset and their dynamic grow in time. The short introduction to the social network is described in the Section 2 and general concept of the DBLP is shown in the Section 3. The Section 4 contains description of dynamic metrics and in the Section 5 is shown practical example of using these metrics on communities of co-authors from DBLP. Also in Section 5 is described algorithm for Extraction of Communities of Co-authors in time.

2 Social Networks

A social network (SN) is a set of people or groups of people with similar pattern of contacts or interactions such as friendship, co-working, or information exchange [10]. The World Wide Web, citation networks, human activity on the internet (email exchange, consumer behaviour in e-commerce), physical and biochemical networks are some examples of social networks. Social networks are usually represented by graphs, where nodes represent individuals or groups and lines represent relations among them. Mathematicians and some computer scientists usually describe these networks by means of graph theory [7].

Social network analysis (SNA) is a collection of methods, techniques and tools that aim to analyse the social structures and relational aspects of these structures in a social network [11]. The study of social networks is a quite old discipline. Many studies oriented to the analysis of social networks have been provided. The datasets used in these studies are obtained by using questionnaires. In contrast to previous SNA research, contemporary provided, and more structured approaches, are based on the automated way of research. In the late 1990s, development of new information and communication technologies (such as internet, cellular phones) enabled the researchers to construct large-scale networks using the data collections stored in e-mail logs, phone records, information system logs or web search engines.

Community detection is an important aspect in discovering the complex structure of social networks. A community is defined as a subset of nodes within the network such that connections between the nodes are denser than connections with the rest of

the network [10]. Community structure can be defined using modules (classes, groups or clusters etc.).

3 Digital Bibliography Library Project

DBLP (Digital Bibliography Library Project) is a computer science bibliography database hosted at University of Trier, in Germany. It was started at the end of 1993 and listed more than 2.1 million publications in January 2013. These articles were published in Journals such as VLDB, the IEEE and the ACM Transactions and Conference proceedings [4]. DBLP has been a credible resource for finding publications, its dataset has been widely investigated in a number of studies related to data mining and social networks to solve different tasks such as recommender systems, experts finding, name ambiguity, etc. Even though, DBLP dataset provides abundant information about author relationships, conferences, and scientific communities, it has a major limitation that is its records provide only the paper title without the abstract and index terms.

Many experts focuses on the task of finding persons with high level of experience on a specific topic. To achieve this objective researchers approached this task mainly in three different ways. The first group applied an information retrieval techniques to solve it [1], the authors of this paper proposed a weighted language model, which introduces a document prior probability to measure the importance of the document written by an expert. The second group approached this task using social network analysis metrics [12], in this study a large online help seeking community, the Java Forum, was analysed using social network analysis methods and a set of network-based algorithms, including PageRank and HITS. While the third group used a hybrid approach of information retrieval and social network analysis for finding academic experts [13]. In [13] the authors created a local information document for each person to measure his initial level of experience on a topic using information retrieval models. Then they applied propagation on the graph of experts to update his level of expertise according to his relations with the other nodes. In the article [2], the authors focused on the detection of communities with the use of spectral clustering. This algorithm was used in the article [6] to find the communities in a subnetworks that were defined by the selected terms (from the whole DBLP).

4 Dynamic network analysis

Dynamic network analysis (DNA) varies from traditional social network analysis. DNA could be used for analysis of the non static information of nodes and edges of social network. DNA is a theory in which relations and strength of relations are dynamic in time and the change in the one part of the system is propagated through the whole system, and so on. DNA opens many possibilities to analyse and study the different parts of the social networks. We can study behaviour of individual communities, persons or the whole graph of the social network. The paper is focused to analyse the behaviour of communities extracted from DBLP and divided by time. The proposed approach which use dynamic metrics is inspired by work of Palla et al. [8].

The auto-correlation function $C(t)$ is used to quantify the relative overlap between two states of the same community $A(t)$ at t time steps apart:

$$C(t) = \frac{|A(t_0) \cap A(t_0+t)|}{|A(t_0) \cup A(t_0+t)|}, \tag{1}$$

where $|A(t_0) \cap A(t_0+t)|$ is the number of common nodes (members) in $A(t_0)$ and $A(t_0+t)$, and $|A(t_0) \cup A(t_0+t)|$ is the number of nodes in the union of $A(t_0)$ and $A(t_0+t)$.

The stationarity of community is defined as the average correlation between subsequent states:

$$\zeta = \frac{\sum_{t=t_0}^{t_{max}-1} C(t,t+1)}{t_{max}-t_0}, \tag{2}$$

where t_0 denotes the birth of the community, and t_{max} is the last step before the extinction of the community. Thus, $(1-\zeta)$ represents the average ratio of members changed in one step.

Authors of the paper [8] found that the auto-correlation function decays faster for the larger communities, showing that the membership of the larger communities is changing at a higher rate. In contrast, they said that small communities change at a smaller rate with their composition being more or less static. The stationarity was used to quantify static aspect of community evolution.

5 Evolution of Co-authors Communities

To create our experiments and to count dynamic metrics we generate DBLP subgraphs of selected terms for each year in which this term occurs. Generating of these subgraphs of DBLP authors is described in the following section. This final set of subgraphs is input for our experiments and to count dynamic metrics.

5.1 Extraction of Communities of Co-authors in Time

For the experiments we used a data collection of publications and their authors from the DBLP server [1]. When processing XML dataset we analysed records for the following publication types: *article*, *inproceedings* and *incollection*. During the experiment 2,055,469 articles (set *Articles*), 1,182,363 authors (set *Authors*) were indexed and 308,933 terms from titles of articles (set *Terms*) were extracted. A set of *Terms* contains both terms lemmatized by Porter's algorithm [9] and their forms without lemmatization. For each article we store informations about authors, key for DBLP collection, date when it was added to the DBLP collection and publication year. For an author we register his ID, simplified name for information retrieval, special form of his name for the DBLP collection, number of articles and links to the most important terms of the author. Furthermore, we use a matrix of articles and their terms $M^{Articles \times Terms}$.

[1] DBLP dataset: http://dblp.uni-trier.de/xml/ - downloaded October 2012

```
Example of Article
*******
Key: reference/social/SlaninovaMDOS10
Date: {1/1/2010 12:00:00 AM}
Id: 876067
MDate: {11/13/2011 12:00:00 AM}
Authors Count: 5
```

Before creating a subgraph, we need to determine the set of terms, which we will be searching for. These terms represent articles we are interested in. We will denote this set as *Query*. It can contain both terms with or without the lemmatization. We use both forms because anyone can come across the need to look up words in their original form. As an example, the word *modularity* in social networks means something different than the base form *modul* obtained by the lemmatization. After we identified the terms, we need to get the articles defined by these terms. These articles $Articles_Q$ are determined by the non-zero values in the matrix M in those columns, that match the searched terms (OR query). If we want to select only those articles in whose titles contains all entered terms (AND query), then we must remove such articles from the set $Articles_Q$ which have some of the term missing in the title.

The set of the years in which the articles were published in the set $Articles_Q$ we denote as Y. From the set of articles $Articles_Q$ we select set of authors $Authors_y$ who published together, for each year $y \in Y$. Now for every year $y \in Y$ we create graph $G_y(Authors_y, E)$, where E represents strength of authorship.

Dynamic metrics described in the Section 4 are generally metrics used to evaluate the characteristics of the community. About such community, we have to know that it changes over time and also we should have information on how the community looked at each time step of its existence. Therefore to get the information about the communities and their changes in time from subgraph of the authors, we need to execute a series of steps which are described below.

Algorithm for Finding Component Evolution in Time

(I) *Creating the longest continuous consecutive time chain of graphs G_y*
Input graphs may have different time intervals between them. But for the next step we need to choose the longest consecutive time period with one year interval.
For example:
Input graphs: G_{1998}, G_{1999}, G_{2002}, G_{2003},G_{2004}, G_{2005},G_{2006},G_{2007}, G_{2012}.
For processing we use this set of graphs: G_{2002}, G_{2003},G_{2004}, G_{2005},G_{2006},G_{2007}.

(II) *Finding connected components of the subgraph*
Graphs from the previous step are non connected. We search for all the connected components to get components for each year with which we will continue to work.

(III) *Create chain of the connected components across all time steps*
1. We choose the first largest component c from the graph in the first time step.
2. According to the following rules we select next component (follower) in the next time step based on the current component c. We denote this component as similar component. We are looking for the components which has the biggest number of the same nodes as the current component c and for selection we have to choose one of the following options:

(a) If only one similar component is found we denote it as follower.
(b) If more than one similar components are found we denote the biggest one as follower.
(c) If no component is found we choose as a follower the biggest existing component in this time step.

3. Step 2 is repeated for each time step except the last one.

Basically we are talking about the components that consist of the DBLP authors and links between them which are formed on the basis of the common interest - the same terms in the titles of their articles. Therefore we can say that our components are the communities of the co-authors. Due to the above described algorithm, we prepare the set of consecutive components. We assume that this set represents the development of one community over time.

This idea allows us to calculate dynamic metrics described in the Section 4. Recall that the auto-correlation is calculated for each of the two states of the same community, followed with computed value of stationarity.

5.2 Experiments

To demonstrate experiments, we choose terms: ”elearning”, ”elearning teach blackboard”, ”elearning teach moodle”, ”mysql”, ”oracle”, ”social network”, ”dynamic social network”, ”social network analysis”. Basic properties of the communities found for each set are described in the table, where we present the count of time steps for each community.

Terms	Count of time steps	Year from	Year to
elearning	12	2001	2012
elearning teach blackboard	43	1971	2013
elearning teach moodle	43	1971	2013
mysql	5	2008	2012
oracle	33	1981	2013
social network	17	1997	2013
dynamic social network	10	2003	2013
social network analysis	17	1997	2012

Table 1. Communities of co-authors developed in time

Evolution of communities of co-authors in the time are demonstrated in the Figures 1 and 2. These figures show changes of counts of members of each community in time.

In Figure 1 on the left, we can see a development of the three communities, which published in similar areas, namely ”social network”, ”dynamic social network” and ”social network analysis”. If we look at the change of the curves of authors in communities that deal with ”social network” and ”social network analysis”, we will notice that curves from 1997 to 2009 look similar. In 2009, we can notice a great interest in the generic term ”social network”. According to information shared by Facebook provider

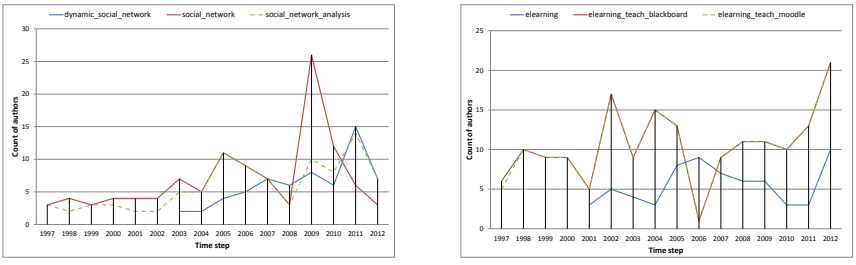


Fig. 1. Evolution of communities of co-authors for the terms *"social network", "dynamic social network", "social network analysis" and "elearning", "elearning teach blackboard", "elearning teach moodle"*

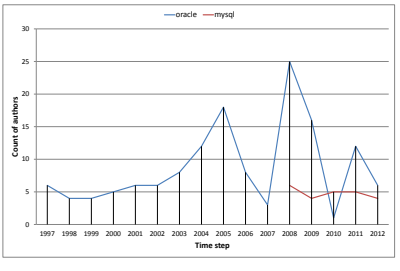


Fig. 2. Evolution of communities of co-authors for the terms *"mysql", "oracle"*

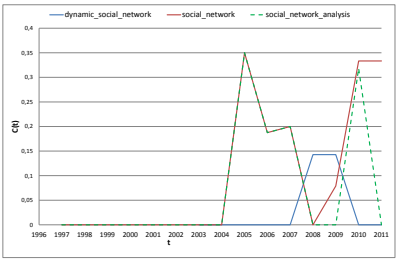
in 2009 [2], there was the largest detected increase of new users on Facebook. In 2009, around 150 million new users have joined the social networking site Facebook. In the following years, the number of newly connected users varied from 5 to 50 millions per year.

Since 2009, interest in generic term "social network" began to decline strongly. On the other hand, interest in terms "dynamic social network" and "social network analysis" had increased. At the same time, these two curves began to grow similarly.

We would like to draw attention to an important property of value of auto-correlation. Auto-correlation is always computed for the community in a time interval t to the change of the community in the following time slot $(t+1)$. Because of this property, we show the results until 2011 in Figure 3 since the value of auto-correlation for 2012 can be calculated correctly only at the end of 2013.

On the left side of Figure 3, we present auto-correlation values for communities "social network", "dynamic social network" and "social network analysis". The higher the

[2]Number of active users at Facebook over the years, http://news.yahoo.com/number-active-users-facebook-over-230449748.html



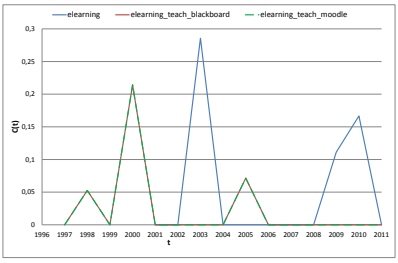


Fig. 3. Auto-correlations of communities of co-authors for terms *"social network", "dynamic social network", "social network analysis" and "elearning", "elearning teach blackboard", "elearning teach moodle"*

auto-correlation is, the more authors in the community in these time periods had stable interest in publishing together with someone else. This means in our case publishing together in the same area of interest that was initialized by the terms. According to the auto-correlation curves, there was stable interest in "social network" in 2004 which then continuously decreased until 2009. From 2009 onwards we can see a stable growth of interest in publishing in "social network". From 2007 to 2010, there is evident growth of interest in the field of "dynamic social network". However, it is smaller than that of the generic term "social network". For the community "social network analysis", we can follow a similar stability evolution of the authors who published in the area of "social networks".

We can create the same analysis for the auto-correlation curves of communities formed by terms "elearning", "elearning teach blackboard" and "elearning teach moodle", shown on the right side of the Figure 3 . In this analysis can be noticed an interesting factor that from 2011 to 2012, the community which deals with "elearning" has the largest value of auto-correlation. We could say that it is experiencing a period of steady state of authors who publishes in this area.

In the Figure 4, we show the values of stationarity for all communities, which we analysed in our experiments. In general, this value characterizes the degree of variability of community in time. The larger the value of stationarity is, the more the community is stable and static. On the other hand, the smaller value indicates a community more dynamic and more changeable in time. In the Figure 4, we see that the largest value of stationarity has the community publishing about "social network", "oracle", "social network analysis", "elearning". But if we look at the data, we may notice that the communities dealing with "social network" a "social network analysis" are relatively young, and therefore their values of stationarity are higher than in the older communities. Communities dealing with "dynamic social network", "elearning teach blackboard" a "elearning teach moodle" are more dynamic in the sense that only a few authors have published in this area for a time.

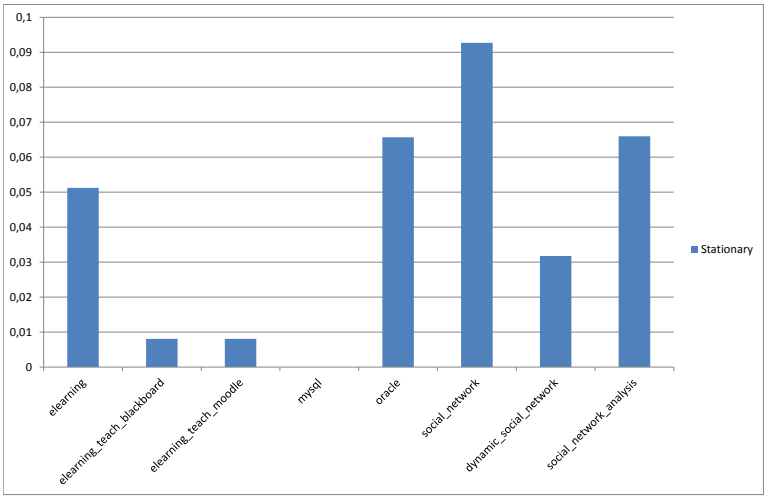


Fig. 4. Stationarity of communities of co-authors for terms *"elearning", "elearning", "elearning teach blackboard", "elearning teach moodle", and "social network", "dynamic social network", "social network analysis"*

6 Conclusion

The research presented in this paper is oriented to analysis of communities of co-authors evolution formed by terms on DBLP. In the paper, the analysis of evolution of co-authors in the communities was presented, with the focus to their growth. The method for evaluation of the stability of authors' interests in the communities extracted from DBLP was described. Moreover, the method for identification of dynamic or static communities in the time was presented. Experiments have been demonstrated on the network of co-authors. Naturally, presented methods can be used for other different networks and another types of communities.

The step Number III is one of the most important steps in the algorithm presented in the Section 5.1, because it defines which components represent an image of one component in different time periods. In future, we want to enrich our experiments by changing this step of the presented algorithm. Together with condition for a particular user incorporated into this step, it gives a completely different view on the issue of selecting the components. Analysis of the evolution of community formed around a user brings the opportunity to research and analyse not only dynamic properties of the community itself but also the possibility of studying the characteristics of the users or the analysis of evolution in individual cases.

Acknowledgment

This work was supported by SGS, VSB – Technical University of Ostrava, Czech Republic, under the grant No. SP2013/167 Analysis of Users' Behaviour in Complex Networks.

References

1. H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. *2008 Eighth IEEE International Conference on Data Mining*, pages 163–172, 2008.
2. P. Drazdilova, J. Martinovic, and K. Slaninova. Spectral clustering: Left-right-oscillate algorithm for detecting communities. In M. Pechenizkiy and M. Wojciechowski, editors, *New Trends in Databases and Information Systems*, volume 185 of *Advances in Intelligent Systems and Computing*, pages 285–294. Springer Berlin Heidelberg, 2013. 10.1007/978-3-642-32518-2_27.
3. J. K. Lars Backstrom, Dan Huttenlocher. Group formation in large social networks: membership, growth, and evolution. *Science*, pages(9):44–54, 2006.
4. M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. *LNCS*, 2476:1–10, 2002.
5. F. S. Mansoureh Takaffoli, Justin Fagnan and O. Zaiane. Tracking changes in dynamic information networks. *2011 International Conference on Computational Aspects of Social Networks CASoN*, pages 94–101, 2011.
6. S. Minks, J. Martinovic, P. Drazdilova, and K. Slaninova. Author cooperation based on terms of article titles from dblp. In *IHCI2011*, 2011.
7. M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):58, 2003.
8. G. Palla, A. lászló Barabási, T. Vicsek, and B. Hungary. Quantifying social group evolution. *Nature*, 446:664–667, 2007.
9. M. F. Porter. Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
10. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks, Feb 2004.
11. J. Scott. *Social Network Analysis*. Newbury Park CA: Sage, 1992.
12. J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM.
13. J. Zhang, J. Tang, and J. Li. Expert finding in a social network. *Advances in Databases Concepts Systems and Applications*, 4443:1066–1069, 2007.
14. Y. Q. Zhixing Huang, Yan Yan and S. Qiao. Exploring emergent semantic communities from dblp bibliography database. *2009 International Conference on Advances in Social Network Analysis and Mining*, pages 219–224, 2009.