Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers

Tze-Haw Huang and Mao Lin Huang
Faculty of Information Technology
University of Technology, Sydney, Australia
{thhuang@it.uts.edu.au, maolin@it.uts.edu.au}

Abstract

This paper proposed a new approach for collecting, analyzing and visualizing co-authoring data of individuals. This approach can be used for understanding the academic collaboration and knowledge domain of individual researchers in a past period through repetitive co-published works. Particularly we extracted the co-authoring data from the DBLP which is one of the largest on-line Computer Science bibliographic databases available on the Internet. To help users to understand the academic collaboration and knowledge domain of individuals, we developed an InterRing visualizer which shows not only the weight of co-authorship of an individual with other researchers in particular academic year, but also the knowledge domain of the individual that was covered by his/her publications published in a past period.

Keywords--- **co-authorship, information visualization, information analysis**

1. Introduction

The social network [7] analysis of co-authorship attracts considerable interest and becoming paramount due to exponential growth of published information in science since last decade. A cutting edge research finding in most cases is based on the cooperation of scientists within research domain and has long been realized that the coauthorship of articles in learned journals provides a window on patterns of collaboration among the academic community. Co-authorship of a paper can be thought of as documenting collaboration between two or more authors, and these collaborations form a "co-authorship network" [6]. Such network reveals the persistent cohesive research collaboration and clustering in the network represents a knowledge domain. Furthermore, the analysis of co-authorship of academic publications could also help to create the Research Quality Framework (RQF) for assessing the research quality of individuals and research groups. Joint authorship often reflects the joint research and the movement of an individual author's research domain.

Co-authorship has been well-studied [1, 6, 8, 9] that reflecting the co-contributions of researchers working towards an academic published paper. Many of these studies used information visualization technique to enhance the cognition process. Figure.1 shows a typical visualization of co-authorship network [2]. In general perspective, it visualizes the dataset in such a way that it approximates the most influential researchers in research domains and their interrelationships among their co-authors. Nevertheless, visualizing co-authorship in a network with nodes representing authors and edges stand for research proximity is a narrow definition for scientific collaboration.

However, most of visualization techniques in bibliography analysis usually use a plain graph of network to present dataset. Each clustering in the network does not exist in isolation; node of such network represents an author and link directly connected to each other if relationship between them can be coupled. Unfortunately, this approach explains general sense of information and did not preserve historical research collaboration as oppose to our technique which details the analyzed author's scientific collaboration and contribution with particular authors as well as inactive authors in the past which possibly suggests the new research community by repetitive published works on a research domain. Furthermore, the network based technique is difficult to work on very large dataset eventually the large collection of nodes will occupy the entire display space with overlapped edges that is difficult to interpret the result.

Our primary scope in this research is focusing on the analysis and visualization of individual's co-authorship network by applying InterRing visualization. We attempt to visualize the academic collaborations and knowledge domain of individual researchers. Our integrated visualizer can show not only the weights of co-authorship of an individual with other researchers in particular publication periods, but also the knowledge domain of the individual that was covered by his/her publications published in a past period. In general, the InterRing visualizes the contributions of co-authors to an analyzed author's past research publications which also

can be further extended to explore the knowledge domain and discovery between research communities.

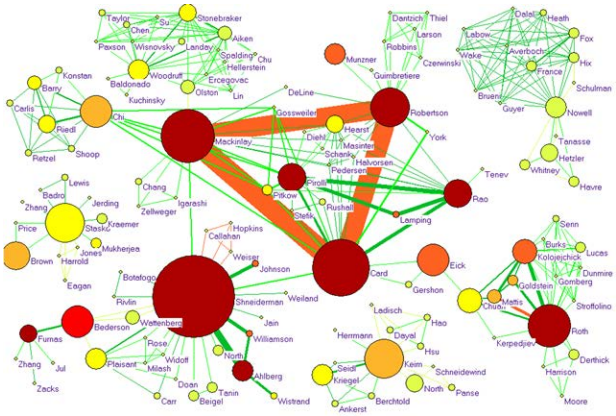

Figure 1: A graph visualization of co-authorship network, collected from [2] with the permission

2. Co-Authorship Data Analysis

The dataset used for analyzing and visualizing co-authorship were ported from the on-line DBLP (http://www.informatik.uni-trier.de/~ley/db/), which is one of the largest Computer Sciences bibliographic data source available on the Internet. It provides information on major computing journals and conference proceedings between year 1936 and 2006. The distinctive advantage of DBLP over another scientific bibliographic database such as CiteSeer is the easily identification of authors. DBLP provides full author names in a publication and CiteSeer uses only the initial of authors which will often cause confusions when multiple result-set has been returned when querying on an author name and also performing insert operation to the database that violates the constraints. Thus, the decision was made to use DBLP as our preference of dataset.

The DBLP engine relies on manual data entry process by original authors or DBLP people whereas CiteSeer uses a collection of URLs that each contains resources of academic papers and actively attempts to retrieve them from these sites on a regular basis. In comparison DBLP is, however, restricted to a limited set of domains and still remains quite dependent on manual process of data entry. Hence, for some domains it is likely that only partial picture can be drawn.

Number of authors	442,886
Number of papers	678,296
Average authors per paper	2.40
Average papers per author	3.67
Conference accepts most papers	Communications of the ACM, 6892

Table.1: The statistics studied was based on DBLP and as a benchmark. Please note the data collected by DBLP may not be complete.

A summary of statistics of DBLP studied is given in Table.1 which is obviously has highlighted our concern as addressed in Section.1. On average, there were 3.67 authors participated in a paper and each represents a knowledge in a subject domain. Regardless of the contribution of each author the knowledge has exchanged more or less and the frequencies of collaboration determine the well formed research communities.

Figure.2 illustrates that the information on the number of publications collected by DBLP has grown exponentially since 1980 which also indicates the explosion of knowledge and the need to understand the contemporary scientific researches.

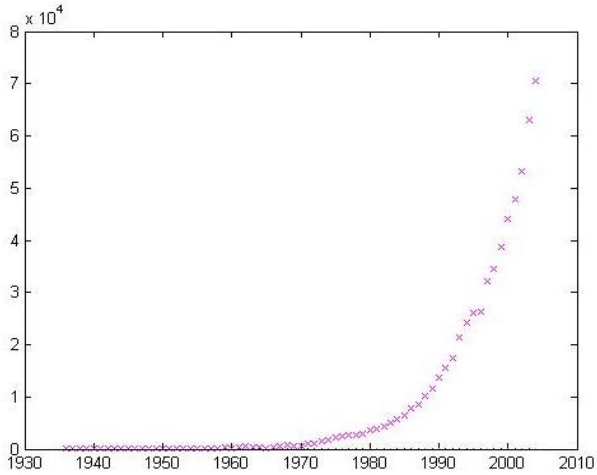

Figure 2 Exponential growths of Computer Science publications since 1936 to date

3. Visualization of Co-Authorship

The advantage of using InterRing to visualize co-authorship instead of network visualization is substantial. We aim to understand the academic collaboration and knowledge domain of individual researchers through the visual representation of co-authoring relationships.

The visualization tool we have developed accepts the name of researcher as a query to extract the co-authoring data from the DBLP dataset consisting of a list of co-authors of joint publications published in certain past years.

The InterRing visualizer will display the outcomes of analyzed author and presenting in a series of concentric rings. In our implementation, each ring represents a particular publication year. From an algorithmic point of view it is often desirable to deal with a small research domain of analyst's current interest rather than the entire research domains. The InterRing adds an extra dimension to the traditional network representation that can be used to project the important feature, such as historical data of the co-authorship.

3.1 Calculation of Individual's Weight

Co-authoring is a common practice in Computer Science. There is no established convention regarding the significance of contribution required to merit and being listed as co-author. Presumably, every author of a joint paper should have made substantial contribution and hence made it difficult to establish a standard of fairness and significance. However, it is customary and based on convention that a researcher produces key ideas and prepares the manuscript certainly deserves to be listed as first author, the remaining co-authors are listed in order of contribution according to their consensus which indicates that ordering is one of approaches to determine the significance that we have used in order to allocate the weight of an author for a year that they have co-published works.

Suppose that there are n co-authors $a_1, a_2 \ldots, a_n$ listed in numerical order in a joint publication p and $c(a_i)$ is the contribution made by a_i to the publication p. We assume that $c(p)=100$ is the total contribution made by all co-authors; q is a deduction of contribution between two authors (note that in our implementation we set $q=2/3$) We have $c(a_{i+1}) = c(a_i) \times q$ and the first author's contribution is $c(a_1) = c(p) \times q$ Thus, the total contribution of publication p can be expressed as

$$c(p) = \sum_{i=1}^{n} c(a_i) = c(a_1) + c(a_2) + \ldots + c(a_n)$$
$$= c(a_1) \times (1 + q + q^2 + q^3 + \ldots + q^{n-1})$$

Therefore, we have

$$c(a_1) = c \div (1 + q + q^2 + q^3 + \ldots + q^{n-1}) \text{ ; and since}$$
$$c(a_i) = c(a_1) \times q^{i-1},$$

Thus, we can calculate the contribution of a particular co-author a_i to the joint publication p by using the following formula:

$$c(a_i) = (c \times q^{i-1}) \big/ (1 + q + q^2 + \ldots + q^{n-1})$$

$$= (c \times q^{i-1}) \big/ [(1 - q^n) \big/ (1 - q)]$$

$$= [c \times (1 - q) \times q^{i-1})] \big/ (1 - q^n)$$

Since we set $c=100$ and $q=2/3$. Thus, we can get

$$c(a_i) = \frac{100 \times (2/3)^{i-1}}{3 - 3 \times (2/3)^n} \% \qquad (1)$$

By applying the formula (1), we can easily calculate the contribution of an author a_i in a research paper in percentage.

Suppose that we are going to count a set of m papers $P_m = \{p_1, p_2, \ldots p_m\}$ as the contribution of a particular academic year, and in that year an individual co-author a_i wrote only k papers $\{p(a_i)_1, p(a_i)_2, \ldots p(a_i)_k\} = P_k$, with other co-authors, where $k \le m$ and $P_k \subseteq P_m$.

If we denote $c(p_j)$ is the overall contribution made by the paper p_j and $c(a_i)_j$ is the contribution made by the author a_i to the paper p_j, then we can easily calculate the weight (or contribution) of an individual in an academic year, in terms of jointing research publications.

$$w(a_i) = \frac{\sum_{j=1}^{k} c(a_i)_j}{\sum_{j=1}^{m} c(p)_j} \qquad (2)$$

The InterRing consists of a series of concentric circles $C_1, C_2, \ldots C_k$ and each circle represents a certain academic year. An InterRing r_i is defined as a circular region between circles C_i *and* C_{i+1}. We partition each ring r_i into n pieces (or Sectors) $s_1, s_2, \ldots s_n$ and assign them to n corresponding co-authors who made contributions of joint publication to a certain academic year. The higher the contribution an author made, the larger the sector he will occupy. The layout of each ring r was drawn from outer to inner with starting point at zero degree displayed in an overlapped manner in the same screen space.

The size of the Sector of a co-author occupied in a ring is derived from the wedge $wg(a_i)=\{C_j, C_{j+1}, \theta\}$, where j is the year of publication and θ is the angle assigned to co-author a_i. The angle θ is the function of the contribution $w(a_i)$ made by a_i in an academic year and can be calculated as below

$$\theta = w(a_i) \text{ x } 2\pi \qquad (3)$$

Figure.3 illustrates the calculation of wedges and the partitioning of Sectors to represent the research contribution of each individual researcher.

Figure.4 shows the InterRing visualization of co-authoring data of a researcher over a past period. Each sector in a ring represents an co-author. In the legend field each co-author has assigned a color for easier identification. The circular rings clearly reveal the academic collaboration and possible movement of research domain. It answers the questions, such as:

- In which year, what researchers have participated in his/her research publications,
- The strength of research collaboration between two researchers over a past period,
- What knowledge domain covered by an individual's research in a past period through the joint publications.

3.2 InterRing Visual Representation

We apply the traditional InterRing drawing method to the ring *R*. We place each co-author $a_i \in A$ on one of the concentric subring r_1, r_2,…, r_k. if a_i has joint publications with current analyzed researcher. The algorithm we used for drawing is stated below

```
Algorithm DrawInterRing (R, δ) {
      Iterator itr = R.getYears();
      radius = δ * itr.length;
      While(itr.hasnext()) {
         SubRing r = (SubRing)itr.next();
         DrawSubRing(r, 0, radius);
         radius = radius - δ;
      }
}
Algorithm DrawSubRing (r, α, radius) {
         Iterator itr = r.getAuthors();
         While(itr.hasnext()) {
            Author a = (Author)itr.next();
            w(α) = a.getWeight();
            θ = w(a) * 2π;
            x1 = radius*cos α; y1 = radius*sin α;
            x2 = radius*cos α + θ; y2 = radius*sin α + θ;
            // Draw a sector based on [x1, y1] and [x2, y2]
            α = α + θ;
         }
      }
   }
```

Where δ is a distance constant between two concentric circles that is defined by width of display space divided by number of academic years. We call the procedure *DrawSubRing(r, 0, radius)* to draw a sequence of sub-rings corresponding to the academic years of analysis. The *radius* is the distance from a circle to the center and the value of *radius* is decremented by δ until the last sub-ring is drawn. The α is an initial degree at the starting point of drawing a sub-ring, and will be incremented by drawing sector by sector, see Figure 3.

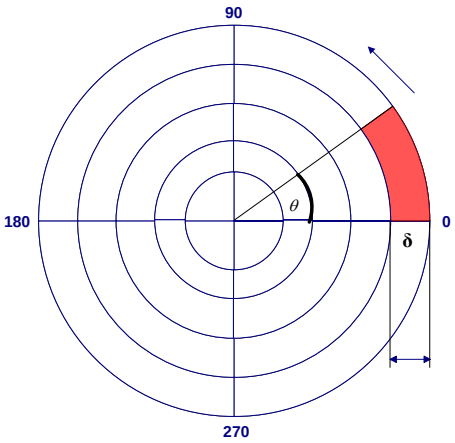

Figure 3: The calculation of Wedges and Sectors to represent the research contribution of an individual.

4. An Example

In this section we will give an example of how to use our InterRing to represent the co-authorship and knowledge domain of an individual researcher through the exploration of DBLP. In section.3 we have discussed how to determine the contribution made by a co-author to a publication *p* and how to calculate the overall weight of a co-author in relation to a particular researcher of joint publications.

For example, Figure.4 shows the analysis result of a selected researcher Dr. M.L Huang for his co-authoring relationships with other 22 researchers in the past 10 years. We can see from the InterRing in Figure 4 that each co-author of Dr. Huang occupies a certain Sector region in a sub-ring r_i corresponding to a particular academic year. We use different colors to help viewers to distinguish sections. The size of a sector indicates the strength of co-authoring relationship between a co-author and the analyzed researcher and the overall contribution made by that co-author in a academic year.

The visualization result clearly tells the user what researchers in which academic year have participated and made contribution in the joint research projects and publications.

The interpretation of the InterRing can also indicate the movement of a researcher's knowledge domain through her/his research collaboration history in a past period. Therefore, the trend of knowledge movement of a researcher could then be possibly predicted by further exploration of his co-authors' specialties.

In addition, in order to explore the research proximity, we could then identify the most recent active co-author, such as Dr Q.V Nguyen by identifying the strongest co-authorship in outer rings. For example, according to Figure.4, Dr. Nguyen has made significant contribution in co-authoring with Dr. Huang since 2002.

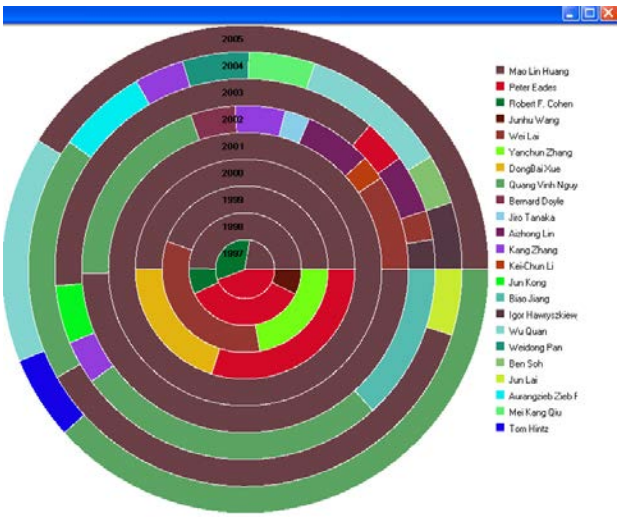

Figure 4: Visualization of an individual's co-authorship network over the past 10 years using InterRing Visualizer

Figure.5 shows the analysis result of Dr. Nguyen. In comparison with Figure.4, This InterRing shows a strong research coherence, high frequency of repetitive collaboration and the establishment of small research community on a similar knowledge domain that has gradually formed since 2002. However, the existence of unexplored research communities can be discovered through the same process.

Figure.6 shows the details the co-authored papers published in a past period between analyzed author and particular co-authors. By analyzing of the past work, the user can understand the common research strength, knowledge sharing and the movement of knowledge domain between co-authors.

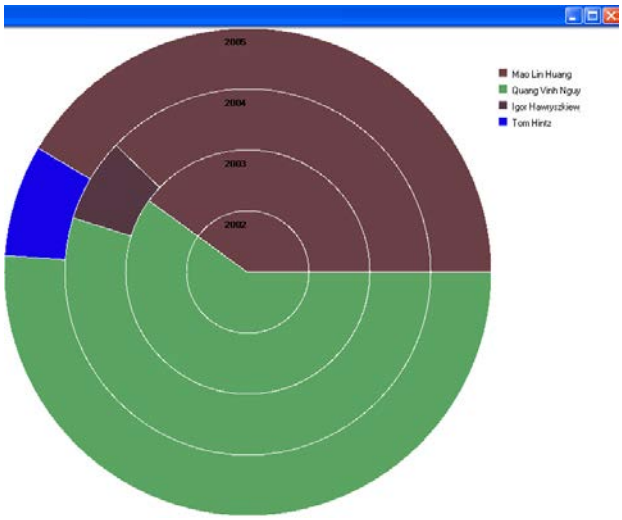

Figure 5 Visualizing an individual's co-authorship network

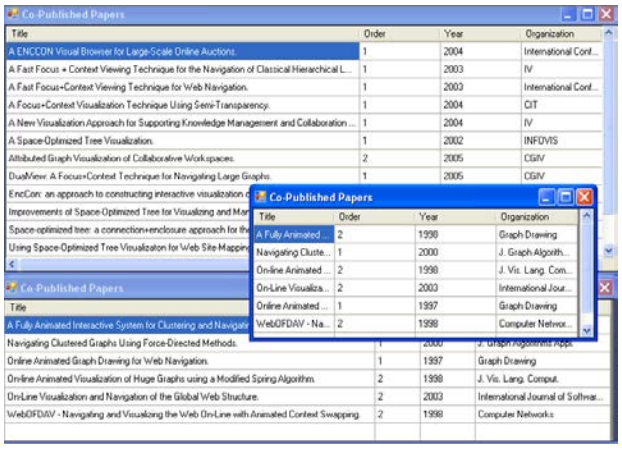

Figure 6: The detail of joint research publications between co-authors

Figure7 shows the integrated visualization of co-authoring relationship; while the InterRing is used to show the co-authoring relational structure, the open textual windows display the detail of joint publications of two co-authors of a interested sector clicked by the user.

5. Conclusions and Future Directions

The primary aim of this research is to visualize the co-authorship relationship and contribution distribution of joint papers that towards the visualization of the entire knowledge and research domain of a particular researcher in the past years. In our future work, as fundamental aspect of knowledge discovery is capturing the knowledge created by researchers and their coauthors. Also, the model of visualization tool that we have initially developed has capability to be extensible in order to support for adding 1) algorithms 2) graphs and 3) various datasets.

We will attempt to integrate the various layout algorithms such as radial and spring drawings for future knowledge domain visualization into our tool. These layout algorithms help to easily present clustering data across knowledge domains. In fact, our interring drawing is more effectively in presenting the multi-dimensions information.

In conclusion this paper has presented a new approach to the analysis of co-authorship network via interring drawing instead of traditional network approach. The developed methodology is able to capture the past scientific collaboration of analyzed author which will not be otherwise visualized in traditional network. We have also discussed the process to identify the research community which is concluded via the comparison of visualization results. The visualized result can also be used to interpret the RQF that helps the government and research organizations to determine the funding based on research quality of a scientist through scientific collaboration.

References

[1] C. Chen and R. J. Paul, "Visualizing a Knowledge Domain's Intellectual Structure", IEEE Comput. 34(3), pp65-71, March, 2001.

[2] W. Ke, K. Borner, and L. Viswanath, "Major Information Visualization Authors, Papers and Topics in the ACM Library", IEEE Symposium on Information Visualization (INFOVIS'04).

[3] K. Börner, L. Dall'Asta,, W. Ke and A. Vespignani, "Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams", Complexity, vol. 10, no. 4, pp57-67, 2005

[4] C. Chen and L. Carr, "A Semantic-Centric Approach to Information Visualization", International Conference on Information Visualization, pp 18-23, 1999

[5] T.K. Landauer, P.W. Foltz and D. Laham, "Introduction to Latent Semantic Analysis", Discourse Processes, 25, pp259-284, 1998

[6] M. E. J. Newman, "Coauthorship Networks and Patterns of Scientific Collaboration", Proceedings of the National Academy of Sciences, 2004, 101: 5200-5205.

[7] S. Wasserman and K. Faust, "Social Network Analysis ",Cambridge University Press, Cambridge, 1994

[8] M. Newman, “Scientific Collaboration Networks: I. Network Construction and Fundamental Results”, Physical Review E, 64(1):016131, 2001.

[9] Yoshikane. Fuyuki, Nozawa. Takayuki and Tsuji. Keita, ”Comparative Analysis of Co-authorship Networks Considering Authors' Roles in Collaboration: Differences between the Theoretical and Application Areas”, ISSI 2005, July, 2005, vol.2, p.509-516.

[10] C. Cotta, J.J. Merelo, “The Complex Network of Evolutionary Computation Authors: an Initial Study”, Physics/0507196, 2005

[11] E. G. Berkowitz and M. R. Elkhadiri, “Creation of a Style Independent Intelligent Autonomous Citation Indexer to Support Academic Research”, Proceedings 15th Midwest Artificial Intelligence and Cognitive Science Conference, pp. 68-73, 2004

[12] A. Goldenberg and A. Moore, “Bayes Net Graphs to Understand Coauthorship Networks KDD”, Workshop on Link Discovery: Issues, Approaches and Applications, 2005

[13] Boanerges Aleman-Meza, Meenakshi Nagarajan, Cartic Ramakrishnan, Amit Sheth, Budak Arpinar, Li Ding, Pranam Kolari, Anupam Joshi, and Tim Finin, “Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection”, WWW 2006, Edinburgh, Scotland

[14] P.A. Chirita, A. Damian, W. Nejdl and W. Siberski, “Search Strategies for Scientific Collaboration Networks”, CIKM 2005, Bremen, Germany

[15] B. Lee, M. Czerwinski, G. Robertson, B. B. Bederson, “Understanding Eight Years of InfoVis Conferences Using PaperLens”, INFOVIS'04, Vol 0, pp216.3

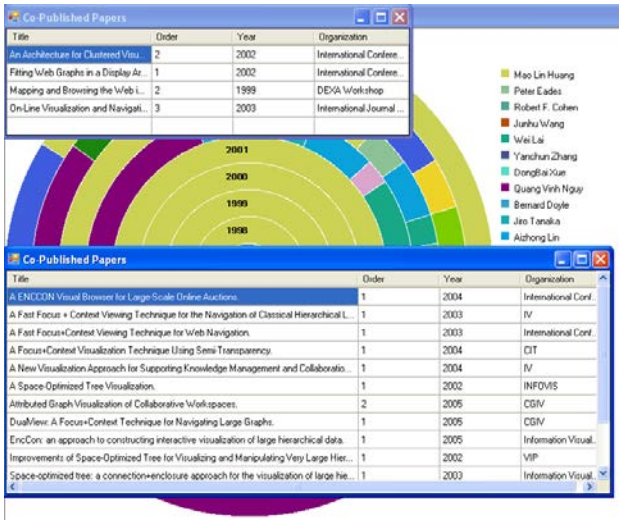

Figure 7 Snapshot of integrated visualization