How to Become a Group Leader? or Modeling Author Types based on Graph Mining

George Tsatsaronis[1], Iraklis Varlamis[2], Sunna Torge[1], Matthias Reimann[1], Kjetil Nørvåg[3], Michael Schroeder[1], and Matthias Zschunke[1]

[1] Biotechnology Center (BIOTEC), Technische Universität Dresden, Germany
george.tsatsaronis@biotec.tu-dresden.de
[2] Dept. of Informatics and Telematics, Harokopio University of Athens, Greece
[3] Dept. of Computer and Information Science, NTNU, Norway

Abstract. Bibliographic databases are a prosperous field for data mining research and social network analysis. The representation and visualization of bibliographic databases as graphs and the application of data mining techniques can help us uncover interesting knowledge regarding how the publication records of authors evolve over time. In this paper we propose a novel methodology to model bibliographical databases as *Power Graphs*, and mine them in an unsupervised manner, in order to learn basic author types and their properties through clustering. The methodology takes into account the evolution of the co-authorship information, the volume of published papers over time, as well as the impact factors of the venues hosting the respective publications. As a proof of concept of the applicability and scalability of our approach, we present experimental results in the *DBLP* data.

Keywords: Power Graph Analysis, Authors' Clustering, Graph Mining

1 Introduction

Currently, vast amounts of scientific publications are stored in online databases, such as *DBLP* or *PubMed*. These databases store rich information such as the publications titles, author(s), year, and venue. Less often they provide the abstract, or the full publications' content and references. The exploitation of additional features, such as co-authorship information, may help us create novel services for bibliographic databases.

In this direction, new online services that process metadata have appeared, such as *ArnetMiner*[9] or *Microsoft Academic Search*[1]. Services that visualize co-authorship information are also available, such as the *"Instant graph search"*[2], which presents the existent co-authorship paths connecting two authors, or the *"Social graph"*[2], which presents all the co-authors of a single author in a star topology. However, to the best of our knowledge, there is currently no methodology available that models the evolution of the authors' publication profile and

[1] http://academic.research.microsoft.com/

[2] Co-author Path and Graph in Microsoft Academic Search.

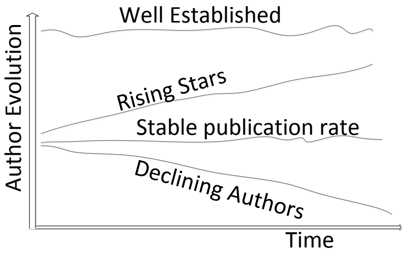


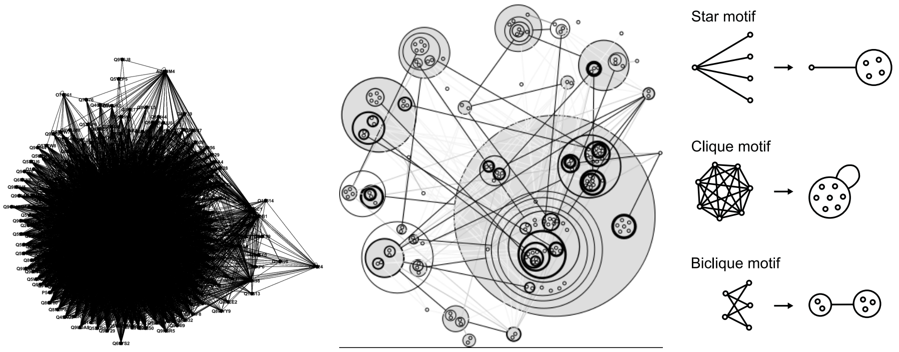
Fig. 1. Motivation: Four basic author "evolution motifs" over time.

detects different types of authors in a bibliographical database, based on their evolution and standing over time.

Figure 1 presents, in a simplified manner, four intuitive author evolution motifs over time. We first distinguish between authors with an *ascending behavior* (*rising stars*), who show high increase in the amount and impact of their published work, as well as on the amount of collaborations with other researchers, and authors with a *descending behavior* (*declining authors*), who have a declining rate and impact of publications. Furthermore, in case the evolution is more *static*, we distinguish between authors that produce constantly a large amount of work over that time frame (*well established*) and authors that produce fewer publications, in a lower but stable rate (*stable publication rate*).

Motivated by the aforementioned motifs, this work addresses the author modeling problem using unsupervised learning, since supervised learning is not applicable due to the absence of manually annotated data for this task. First, we define four basic features that capture the authors' publication profile. Secondly, we monitor the evolution of these features over time and generate respective *evolution indices* per author. Finally, we use these indices to cluster authors with similar evolution profile into respective groups. The exact properties of each author evolution motif are deduced from the analysis of the final clusters.

Our methodology introduces two novel ideas. The first is the application of *Power Graph Analysis* [5] to the co-authorship graphs constructed from bibliographical data, which is performed for the first time, to the best of our knowledge, in bibliographical analysis. The use of *Power Graphs* allows fast and large-scale clustering experiments since they can compress by even up to 40% the information of the original co-authorship graph, as we show in our experiments, in a lossless information manner. Also, they can visualize more efficiently than the original graphs the co-authorship information by identifying several motifs, e.g., cliques and bi-cliques. The second is the introduction of a set of features for each author, which is based on the *Power Graphs* structure, the number, and the impact of her publications. Through the features' monitoring over time, the respective *evolution indices* are computed, based on which the authors' clustering is conducted. Finally, the indices are employed in the analysis of the resulting clusters as descriptors of the *authors' dynamics*. The contributions of this work



(a) Huge biological "fur ball" network.

(b) After the application of *Power Graphs.*

(c) Basic motifs recognized by *Power Graph Analysis.*

Fig. 2. Figure2(a) shows an example of a huge biological network. Figure 2(b) shows the corresponding *Power Graph.* The three basic motifs recognized by *Power Graphs* are shown in Figure 2(c): *Star*, *Clique* and *Biclique*. *Power Nodes* are sets of nodes and *Power Edges* connect *Power Nodes.* A *Power Edge* between two *Power Nodes* signifies that all nodes of the first set are connected to all nodes of the second set.

can, thus, be summarized into the following: (a) a novel methodology for modeling the *dynamics* of authors' publication profiles, and clustering them into groups, (b) transfer of the *Power Graph Analysis* methodology from the field of bioinformatics, to the field of bibliographical databases analysis, in order to visualize and process co-authorship graphs, and, and, (c) empirical analysis and demonstration of the applicability of our approach in a large bibliographical data set (*DBLP*). The rest of the paper is organized as follows: Section 2 presents some preliminary concepts and discusses related work. Section 3 introduces our methodology and in Section 4 we present our experimental findings. Section 5 concludes and provides pointers to future work.

2 Preliminaries and Related Work

2.1 Visualizing Graphs with Power Graphs

In the bioinformatics field, networks play a crucial role, but their efficient visualization is difficult. Biological networks usually result in *"fur balls"*, from which little insight can be gathered. In the direction of providing an efficient methodology for visualizing large and complex networks, such as protein interaction networks, the authors in [5] introduce *Power Graph Analysis*, a methodology for analyzing and representing efficiently complex networks, without losing information from the original networks. The analysis is based on identifying *re-occurring network motifs* using several abstractions. The three basic motifs recognized by *Power Graphs* are shown in Figure 2. These are the *Star*, the *Clique* and the

Biclique, and constitute the basic abstractions when transforming the original graph into a *Power Graph* with *Power Nodes*, i.e., sets of nodes, connected by *Power Edges*. *Power Graphs* offer up to 90% compression of the original network structure [5], allowing for efficient visualization. Figure 2 shows an example of a *"fur ball"* network, and its transformation after the application of *Power Graph Analysis*.

Power Graphs have been successfully applied in bioinformatics, as the networks are rich in the aforementioned motifs[5]. Co-author networks in the bibliographic analysis are implicitly built on such motifs and, thus, perfectly suited for applying *Power Graph Analysis*. A publication is either considered as *Clique* of all authors or as *Biclique* with first and last authors on one and all other authors on the other end. Motivated by this, in this paper we apply *Power Graphs Analysis* into co-authorship graphs extracted from bibliographical databases, in order to define authors' features. As shown in the experimental section, the resulting *Power Graphs* allow for a very efficient visualization of the co-authorship graph.

2.2 Mining Graphs from Bibliographical Databases

Graph-based mining methods in bibliographic databases usually create a graph from author names, venues, or papers' topics, apply a graph partitioning algorithm to locate interesting sub-graphs, and present results in the form of node clusters, e.g., authors by topic, or through visualization of the graph, e.g., co-authors of a single author in a star topology. The various methods for representing bibliographical databases as graphs, can be divided in two categories: (i) those that use n-partite graphs, which contain for example authors, conferences, or topics as nodes, and edges that connect different node types representing relations, and, (ii) those that use graphs with a single node type and edges that may vary in meaning depending on the application. An example of the former category can be found in [8], where bipartite models connecting conferences to authors are employed to rank authors and conferences. Tripartite graph models for authors-conferences-topics have also been introduced in the past [9]. In the later category, e.g., the work in [2], nodes correspond to authors, and edges represent citation or co-authorship relation. The resulting representation can be used to rank authors, find author communities, measure *author centrality* [3, 4, 6], or find special relations between authors, such as advisor-advisee [10]. Regarding the evolution analysis of graphs, snapshot-based approaches, e.g., an author-paper graph per year, are frequently used [1, 8]. In these approaches, predefined measures from each snapshot are extracted and monitored over time. In this work we present an approach which differs from the aforementioned in several points; (a) the co-authorship graph is used only as a basis for defining collaboration-related authors' features , (b) the authors' clustering process analyzes the evolution of these features over time, as well as the changes in the volume and impact of the authors' publications, and, (c) the clustering aims at identifying author evolution motifs and uncover their characteristics, and not to detect author communities.

Paper ID	Authors
p1	a1, a3
p2	a1, a4
p3	a1, a5
p4	a2, a3
p5	a2, a4
p6	a2, a5
p7	a3, a6
p8	a3, a7
p9	a3, a8
p10	a3, a9
p11	a3, a10
p12	a20, a31
p13	a20, a32
p14	a21, a31
p15	a21, a33
p16	a22, a31
p17	a22, a34

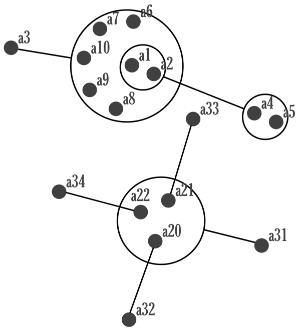


Fig. 3. A sample co-authorship *Power Graph*.

3 Approach

The suggested methodology comprises the following steps: (1) creation of the co-authorship *Power Graphs* in different time points, i.e., years, within a given time frame (Section 3.1), (2) computation of each feature per time point, and of each features' *evolution index* per author (Section 3.2), and, (3) clustering of authors based on their *evolution indices* (Section 3.3).

3.1 Co-authorship Graphs with *Power Graphs*

The initial co-authorship graph contains authors as nodes and weighted edges that connect pairs of authors. Given a paper with k authors, an edge connecting each pairwise combination of the k authors is added in the graph. Given a specific time point t_i (e.g. a year), the initial co-authorship graph accumulatively contains all the publication records from the beginning until t_i. Edges' weights represent the number of papers that the two authors have co-authored until t_i.

The original graph is converted to a *Power Graph* as explained in [5]. An example of bibliographic records (papers and authors), and the resulting co-authorship *Power Graph* is depicted in Figure 3, where for reasons of simplicity we assume that all papers have exactly two authors. Authors *a1* and *a2* have exactly the same co-authors (*a3*, *a4*, *a5*). The same holds for authors *a3*, *a4* and *a5*. This is depicted by a *bi-clique* in the *Power Graph*. Author *a3* has collaborated with *a1*, *a2* and *a6* to *a10*. So *a3* forms a *star* with his co-authors, who form a pair of nested *Power Nodes* (set *a1*, *a2* is inside the greater *Power Node*). Finally, all the co-authors of *a31* form a *star*. The transformation of the original co-authorship graphs into *Power Graphs* offers three very important advantages: (i) it performs a first-level clustering of the authors based on their co-authorship information, (ii) it compresses the original graph, without losing information, and, (iii) it allows for a more efficient visualization of the co-authorship information.

3.2 Authors' Features and *Evolution Indices*

After constructing the co-authorship *Power Graphs*, the next step is to define the features based on which authors' evolution may be measured. Given a *Power Graph* G_i in time point t_i, we define the following features for every author a_k: (1) the size of the *Power Node* to which a_k belongs (S_i) (if a_k is not member of a *Power Node* then S_i=0), (2) the sum of the *Power Nodes'* sizes with which a_k's *Power Node*, or any *Power Node* containing her *Power Node*, is connected (C_i), (3) the number of papers authored by a_k (P_i) until t_i, and, (4) the aggregated impact of a_k's publications until t_i (I_i) [3]. The intuition behind each of the aforementioned features is straightforward; S_i measures the number of the most frequent co-authors an author has at time point t_i (size of her *clique*), C_i measures the size of her extended *clique*, i.e., the co-authors of her co-authors, P_i measures the number of publications up to t_i, and, finally, I_i measures the total impact of her work.
More formally, for an author a_k who belongs to *Power Node* V_k, in the *Power Graph* for time point t_i, S_i is defined as:

$$S_{ik} = we_{V_k,V_k} \cdot |V_k| \tag{1}$$

where $|V_k|$ is the size of *Power Node* V_k, and we_{V_k,V_k} is the weight of the edge connecting *Power Node* V_k with itself. This later weight shows the *strength* of the clique formed by the authors in *Power Node* V_k. Let V_k be connected to n other *Power Nodes*. Then, C_i, for author a_k is defined as:

$$C_{ik} = \sum_{m=1..n} we_{V_k,V_m} \cdot |V_m| \tag{2}$$

where V_m is any *Power Node* connected to V_j with an edge of weight we_{V_j,V_m}. P_i is defined as the number of papers produced by author a_k up to time point t_i. If time points are years, then P_{ik} denotes the number of papers that author a_k has produced from the beginning of her career up to t_i. Finally, I_i is the sum of the impact factors of the authors publications up to t_i. More formally, if the author has written n papers up to t_i, then I_i is defined as follows:

$$I_{ik} = \sum_{m=1..n} IF_m \tag{3}$$

where IF_m is the impact factor of the venue or the journal where paper m was published.

[3] We assign to each paper the impact factor of the venue or journal, in which each paper was published. For our experiments we used the list maintained by *Citeseer*. Historical impact factors are not taken into account due to the lack of respective data

The next step is to define an *Evolution Index* for each of these four features (S_{ik}, C_{ik}, P_{ik}, and I_{ik}), which capture the way the feature values evolve over time. For this reason, given a time span $T : [t1, t2]$, for which we monitor authors, we build a *Power Graph* G_i for each $t_i \in T$ (e.g., for each year). Our aim is to capture the *dynamics* of author a_k in each of the four feature dimensions. For the definition of each of the four *Evolution Indices* we employ a function, which we call *change*. *Change* measures the ratio of the change of any of the features S_{ik}, C_{ik}, P_{ik}, and I_{ik} from time point t_{i-1} until time point t_i. *Change* for feature S_{ik} of author a_k is defined as:

$$Schange_{ik} = \frac{S_{ik} - S_{(i-1)k}}{S_{ik}} \tag{4}$$

The above equation captures the ratio of *change* occurred in a_k's *Power Node* from t_{i-1} to t_i. If a_k's clique has grown from t_{i-1} to t_i then the respective *Power Node* V_k size will increase, and $Schange_{ik}$ will be positive (i.e., in contrast to 0 where there is no change). In a similar manner, $Cchange_{ik}$ captures the ratio of the change in the author's connectivity with other cliques, $Pchange_{ik}$ captures the change with regards to the volume of the papers produced from t_{i-1} until t_i, and $Ichange_{ik}$ the change in her publications' impact.
We can now define the *Evolution Index* (*EI*) for the S feature ($S.EI$), given T, as shown in the following equation:

$$S.EI_{Tk} = \max_{t_i \in T} Schange_{ik} \cdot S_{t_2 k} \cdot \sum_{t_i \in T} Schange_{ik} \tag{5}$$

where t_2 is the final time point in the examined time frame T. Equation 5 captures the evolution of the author over the time frame T, and measures the *dynamics* of the author in the dimension of feature S, as it takes into account the maximum occurred change, the standing of the authors according to S in the final time point, and the sum of all occurred changes. Similarly, the $C.EI_{Tk}$, $P.EI_{Tk}$, and $I.EI_{Tk}$ *Evolution Indices* can be defined for C, P and I features respectively.

3.3 Clustering Authors Using *Bisecting K-Means*

In this section we demonstrate the use of the *Evolution Indices* in the author clustering task. The clustering algorithm that we employ is *bi-secting K-Means*, which delivers high clustering performance, better than *K-Means*, and other agglomerative techniques [7]. *Bisecting K-Means* starts with a single cluster containing all data points, and iteratively selects a cluster and splits it into two clusters until the desired number of clusters is reached or a quality criterion *coherence* is met. Its time complexity is linear to the number of data points.

In our case, each author a_k is a single data point with four dimensions, which correspond to the four evolution indices. Finding authors' clusters is, consequently, formulated as a typical clustering problem, which can be solved using *bisecting K-Means*. Authors are represented as vectors in the four dimensional

```
   Input: Database of papers $D$, time frame $T[t_1, t_2]$, number of desired clusters $k$
   Output: A clustering solution of authors into $k$ groups
1  Construct Power Graph at $t_0$, i.e., the previous time point from $t_1$
2  foreach time point $t_i \in T$ do
3      Construct Power Graph at $t_i$
4      foreach author $a_k \in D$ do
5          Measure and store $Schange_{ik}, Cchange_{ik}, Pchange_{ik}, Ichange_{ik}$
6  Measure and store $S.EI_{Tk}, C.EI_{Tk}, P.EI_{Tk}, I.EI_{Tk}$
7  Put all authors $a_k \in D$ into a single cluster
8  Pick a cluster to split
9  Find 2 sub-clusters using the basic K-Means algorithm (bisecting step)
10 Repeat step 9 for ITER times and choose the best split
11 Repeat steps 8, 9, and 10 until number of clusters is $k$
```

Algorithm 1: Clustering authors of a bibliographical database using *Power Graphs* and *bisecting K-Means*.

space, so the *cosine similarity* measure can be used for measuring similarity between two data points, or between a data point and a cluster centroid. The centroid of a cluster K of authors ($\boldsymbol{C}$), is defined as follows:

$$\boldsymbol{C} = \frac{1}{|K|} \sum_{a_k \in K} \boldsymbol{a_k} \tag{6}$$

Algorithm 1 describes our methodology of organizing authors in a bibliographic database into k groups. If the bibliographical database contains m papers written by n distinct authors, and the resulting *Power Graph* at any time point contains maximum *pn Power Nodes*, then the complexity of the algorithm can be summarized into $O(|T| \cdot (m + n^2 logn + pn) + n)$, which, even for millions of data points, makes the task computationally feasible.

4 Evaluation and Results

In order to demonstrate the organization of authors into categories, and analyze the properties, we experiment using our methodology with the *DBLP Computer Science Bibliography* database. The database comprises $925,324$ distinct author names, and $1,601,965$ publication entries (papers). For our experiments we select the authors, and their publications, that have in total a minimum of 5 publications by 2010. We then apply our methodology using two different experimental set-ups: (a) we use time frame $T = [2000, 2010]$ to cluster the authors into four categories, and analyze the clusters' properties, and, (b) we use time frame $T = [2000, 2005]$ and examine whether our method organizes successfully authors into clusters, by evaluating the behavior of each cluster in the next five years, i.e., $[2006, 2010]$. For our experiments we construct the *Power Graphs* from the original co-authorship graphs, for the periods $[2000, 2010]$, and $[2000, 2005]$ respectively. Table 1 reports statistics from the construction of the *Power Graphs* for all years in $[2000, 2010]$, where the edge reduction rates reached up to 41%.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
#Nodes	53,679	60,732	67,530	75,490	83,797	92,453	100,301	108,051	113,548	117,619	119,767
#Edges	146,562	174,951	207,691	244,132	285,517	331,618	377,586	427,991	473,601	518,158	552,956
#Power Nodes	22,396	25,634	28,730	32,584	36,640	40,972	45,040	49,069	52,060	54,414	56,046
#Power Edges	87,346	103,864	122,279	143,698	168,736	197,421	227,889	261,704	293,246	323,583	349,538
#Edge Reduction Rate	0.404	0.406	0.411	0.411	0.409	0.404	0.396	0.388	0.38	0.375	0.367

Table 1. Number of nodes and edges in the original co-authorship graphs per year, and in the constructed *Power Graphs*.

	Cluster Centroids			
Cluster ID	Feature S	Feature C	Feature P	Feature I

1	29.96	7.16	35.25	14.32
2	64.01	12.8	18.17	8.18
3	6.12	1.37	31.82	6.29
4	5.56	11.21	23.63	13.79

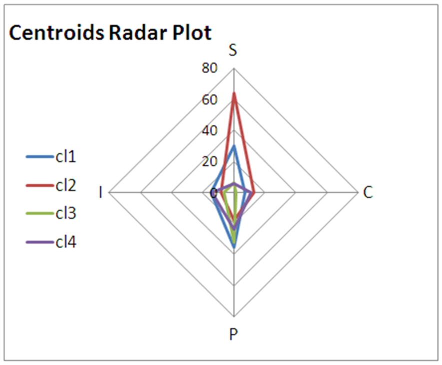
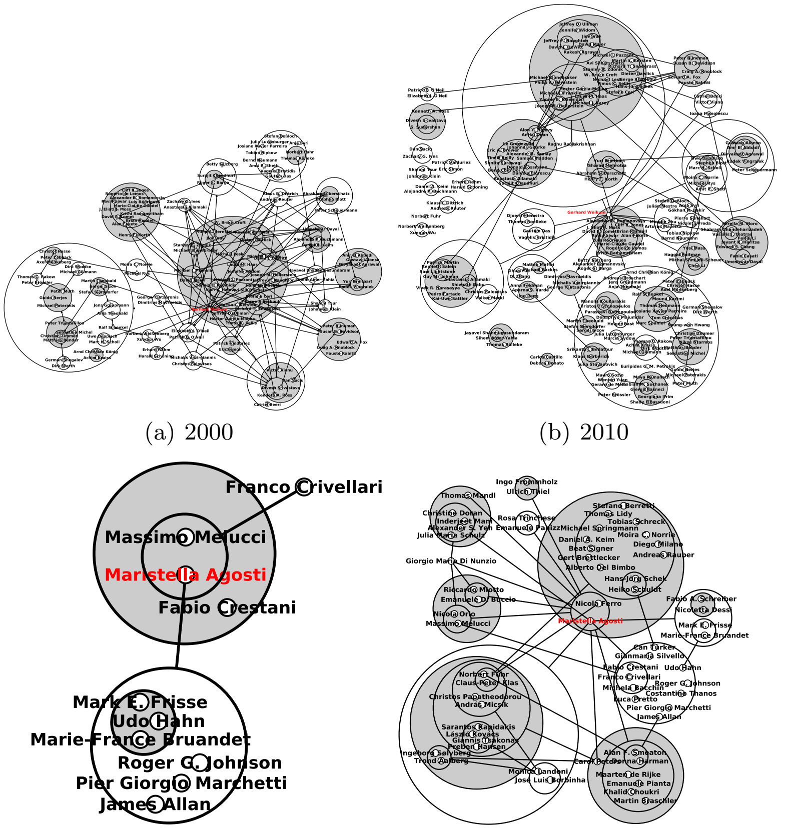


Fig. 4. The centroids of the four basic clusters, and their radar plot according to the four feature dimensions.

4.1 Clustering *DBLP* Authors

In Figure 4 we show the results of experiment (a). The number of author clusters was set to 4, in an effort to identify the basic intuitive motifs shown in Figure 1. We performed a series of experiments with an increasing number of *ITER* in *bi-secting K-means* (from 5 to 500). As expected, higher values produced more stable clusters. The left part of the figure shows the centroids of the four clusters (cluster 1 to 4), and their respective values in the four dimensions (S, C, P, and I). The right part of the figure shows a radar plot of the four centroids in the four different dimensions. Each polygon represents a cluster centroid, and expands more towards a specific direction, if the respective feature value is high.

Cluster 2 contains the most *connected* authors (highest S). The majority of the authors in this cluster are *well established* with great dynamics in expanding their collaborations (C). The explanation is that most of the authors in cluster 2 are group leaders or professors and have many collaborations. One can certainly trace *rising stars* inside that cluster, since its points have huge dynamics in collaborations. Some examples of authors in that cluster are *Christos H. Papadimitriou*, and *Maristella Agosti*. Cluster 1 is certainly the group with the most candidate *rising stars*. The authors in this cluster show the best dynamics in paper publishing (P), and also really good dynamics publications' impact (I).



(a) 2000

(b) 2010

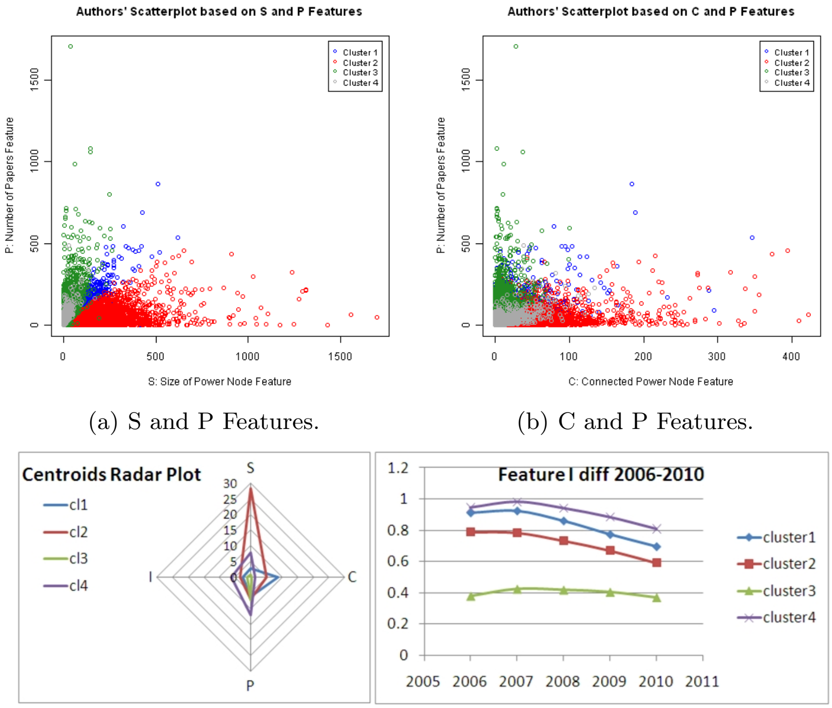
(c) 2000

(d) 2010

Fig. 5. Evolution of *Power Graphs* for two authors over two time points. Grayed circles denote a *Clique* and white a *Biclique*.

Some of the authors in cluster 1 are already well established, but the common characteristic of all authors in cluster 1 is their high dynamics in three dimensions (S, P, and I). Some examples of authors in cluster 1 are *Gerhard Weikum*, and *George Buchanan*. The third most interesting cluster is 4. Authors here certainly publish much and have great potential, but they still need to work their I and S features. Clusters 3 and 4 contain stable publishing authors, without special dynamics though. Finally, in cluster 3 the main motif is that of isolated authors (low S and C values). This cluster also contains *declining authors*, which have ceased expanding their collaborations, but their dynamics in paper publishing (P) remain high.

In Figure 5 we show an example of the *Power Node* evolution between 2000 and 2010 for two authors: *Gerhard Weikum* from cluster 1, and *Maristella Agosti*



(a) S and P Features.

(b) C and P Features.

(c) Prediction of impact in 2006-2010.

Fig. 6. Figures 6(a) and 6(b): Authors' scatter plots based on two feature combinations. Figure 6(c): Clustering authors in 2000-2005, and predicting impact in 2006-2010.

from cluster 2[4]. The figure also shows their connected *Power Nodes*. Figures 5a and b, show how S and C evolved for *Gerhard Weikum*, and Figures 5c and d, for *Maristella Agosti*: the author of cluster 2 had larger changes in her *Power Node* and its neighborhood, compared to the author of cluster 1. This example demonstrates the general motif of clusters 1 and 2 in the radar plot of Figure 4 regarding S and C.

In Figures 6a and 6b we show the authors' scatter plot based on two feature combinations; S and P, and P and C. The figure demonstrates an example of the features' ability to separate the authors. As shown, S and P can separate cluster 1 from the rest, while C and P can separate cluster 3 from 4. Similar findings were observed in all the remaining features' pairs. Figure 6c shows the results of experiment (b). The left part shows the radar plot of the authors' cluster centroids for $2000 - 2005$. The clustering predicts that authors in cluster 4 have large dynamics in increasing their publications' impact factor (I). The right part shows the yearly increase in authors' impact factor for $2006 - 2010$. It

[4] Zoom is possible in the electronic edition.

verifies that authors in cluster 4 had the largest difference (on average over all cluster points) per year on the impact feature compared to the authors of the rest clusters.

5 Conclusions

In this paper we introduced a novel methodology for the organization of authors into basic clusters, using *Power Graph Analysis*. We defined evolution indices over features that capture the connectivity and strength of the authors' co-operations, as well as their publications' volume and impact over time. We demonstrated the applicability of our approach to capture the dynamics of authors using the evolution of the four defined features by clustering authors in *DBLP* with *bi-secting K-Means*. It is in our next plans to explore and interpret authors' clustering using several different values for the k parameter. We also plan to explore the application of our methodology for comparing institutions, based on the notion of the centroid of the institutions' authors, as well as comparing scientific venues, or individual authors. In this direction, the comparison methodology would consider the publications of the respective institutions, venues, or authors, and follow the methodology in this paper. Similar entities, e.g., institutions, would result in similar clusters, for the same value of k, and the comparison could be feasible by placing the clusters at the same radar plot.

References

1. C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee. Exploring the computing literature using temporal graph visualization. In *Visualization and Data Analysis*, pages 45–56, 2003.
2. W. Ke, K. Borner, and L. Viswanath. Major information visualization authors, papers and topics in the acm library. In *INFOVIS*, pages 216.1–9, 2004.
3. X. Li, C. Foo, K. Tew, and S. Ng. Searching for rising stars in bibliography networks. In *Database Systems for Advanced Applications*, pages 288–292. 2009.
4. M. A. Nascimento, J. Sander, and J. Pound. Analysis of sigmod's co-authorship graph. *SIGMOD Rec.*, 32:8–10, 2003.
5. L. Royer, M. Reimann, B. Andreopoulos, and M. Schroeder. Unraveling protein networks with power graph analysis. *PLoS Computational Biology*, 4(7), 2008.
6. A. F. Smeaton, G. Keogh, C. Gurrin, K. McDonald, and T. Sødring. Analysis of papers from twenty-five years of sigir conferences: what have we been doing for the last quarter of a century? *SIGIR Forum*, 36:39–43, 2002.
7. M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, pages 109–110, 2000.
8. Y. Sun, T. Wu, Z. Yin, H. Cheng, J. Han, X. Yin, and P. Zhao. Bibnetminer: mining bibliographic information networks. In *SIGMOD '08*, pages 1341–1344. ACM, 2008.
9. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, pages 990–998, 2008.
10. C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *KDD*, pages 203–212, 2010.