Visualizing Bibliographic Databases as Graphs and Mining Potential Research Synergies

Iraklis Varlamis
Department of Informatics and Telematics
Harokopio University of Athens
Athens, Greece
Email: varlamis@hua.gr

George Tsatsaronis
Department of Computer and Information Science
Norwegian University of Science and Technology
Email: gbt@idi.ntnu.no

***Abstract*—Bibliographic databases are a prosperous field for data mining research and social network analysis. They contain rich information, which can be analyzed across different dimensions (e.g., author, year, venue, topic) and can be exploited in multiple ways. The representation and visualization of bibliographic databases as graphs and the application of data mining techniques can help us uncover interesting knowledge concerning potential synergies between researchers, possible matchings between researchers and venues, or even the ideal venue for presenting a research work. In this paper, we propose a novel representation model for bibliographic data, which combines co-authorship and content similarity information, and allows for the formation of scientific networks. Using a graph visualization tool from the biological domain, we are able to provide comprehensive visualizations that help us uncover hidden relations between authors and suggest potential synergies between researchers or groups.**

I. Introduction

Currently, vast amount of scientific publications are stored in online databases, such as *DBLP*, *arXiv*, and *PubMed*. These databases store rich information such as the publications titles, author(s), year, and venue. Less often they contain the abstract or the full publications' content, and their references. Despite their rich content, bibliographic databases offer limited accessibility and do not efficiently exploit metadata elements. They usually restrict user queries to simple keyword-based search and retrieve scientific publications that contain the query terms in the selected metadata elements. As a result, there is often large semantic gap in bibliographic search engines between users' needs and retrieved results, since access to the full content of the papers, or even the abstracts, is often restricted.

The exploitation of additional semantics such as date, affiliation, citations, co-citations, and co-authorship may further improve search capabilities and create novel services for bibliographic databases. In this direction, semantic enabled search engines for bibliographical data sources, such as *GoPubMed* [1], which specializes in the life sciences, overcome traditional keyword-based search problems and improve search results. In other cases, the increased popularity of social networks analysis had a significant impact on deployed bibliographic databases search services. New databases have been published, offering online services that process publication metadata at the maximum, such as *ArnetMiner*[1][2] or *Microsoft Academic Search*[2]. Authors and venues ranking, organization by year or topic, author profiles extraction and authors name disambiguation are only some of the services provided on top of these databases. Services that visualize co-authorship information are also available, such as the *"Instant graph search"*[3], which presents the existent co-authorship paths connecting two authors, or the *"Social graph"*[4], which presents all the co-authors of a single author in a star topology.

Despite the advantages of the semantic-enabled technologies, an imminent implication of the restricted access to the full articles information is that all research efforts towards mining scientific communities and bibliographic databases are restricted to accessing only the metadata offered by the bibliographic sources. Under these circumstances, mining bibliographical databases in order to extract possible research synergies, identify research trends, and discover scientific thematic cliques or co-authorships, are restricted to the processing of co-authorship or co-citation graphs. In this direction, we propose in this paper a novel methodology for constructing and visualizing co-authorship graphs from bibliographical databases, and show how these graphs can be mined to extract useful information such as possible future research synergies and strong collaboration links.

The suggested method is a two-level approach. At the first level, a co-authorship graph is constructed processing bibliographical data. The graph is then processed using a novel technique called *Power Graph Analysis* [3], which we transfer for the first time, to the best of our knowledge, from the bioinformatics domain to the processing of bibliographical graphs. Through *Power Graph Analysis*, a given graph's nodes may be clustered, through *cliques* and *bicliques* recognition in the initial graph. The resulting *Power Graph* allows a very efficient visualization of the authors graph, while in tandem identifies *cliques* and *bicliques* of co-authors, representing them with *Power Nodes*. At this stage, each *Power Node* is essentially a set of authors that have written several papers together. At the second level, we augment the *Power Graph*

[1] http://www.arnetminer.org/
[2] http://academic.research.microsoft.com/
[3] Co-author Path in Microsoft Academic Search
[4] Co-author Graph in Microsoft Academic Search

with edges between *Power Nodes* that quantify the similarity between the authors' sets, in terms of the similarity of the papers' titles written by the respective author set. This second level, offers a richer representation of the initial co-authorship graph, which is visualized in an efficient manner. Finally, we show how we can predict possible research synergies between authors from this final augmented graph.

The rest of the paper is organized as follows: Section II presents some preliminary concepts regarding the construction of graphs from bibliographical databases and the use of *Power Graphs* in the bioinformatics domain, and discusses related work. Section III introduces our approach for mining potential research synergies from co-authorship graphs. Section IV demonstrates our findings stemming from the application of our approach to bibliographic data. Finally, Section V concludes and provides pointers to future work.

II. Preliminaries and Related Work

The primary focus of this work is the co-authorship information provided by bibliographic databases and the secondary is the short (title) or extended (abstract or full paper) content that pertains to each publication. Citation information is not considered, since this is seldom available in bibliographic databases. Graph-based mining methods in bibliographic databases operate usually in three steps: (a) a graph is created using authors, conferences and papers' topics, (b) application of a graph-based partitioning or ranking algorithm takes place, and, (c) results are presented either in the form of node clusters, e.g., authors by topic, conferences by topic, or using a graph visualization approach, e.g., co-authors of a single author in a star topology. The majority of research works in bibliographic data mining aims at ranking authors or finding authors' communities. In this work we present a different approach which combines graph-based community mining with text mining techniques in order to extract and visualize useful information from bibliographic data, such as potential synergies between researchers or research groups. In order to provide a better understanding of how graphs are created from bibliographic data and what are the visualization options, in the following we summarize the most important research works in the field and illustrate the different alternatives in each process.

A. Constructing Graphs from Bibliographical Databases

Inspired by social network analysis, works on bibliographical databases have proposed different alternatives for modelling bibliographic information using graphs. These can be divided in two main categories: (a) methods that create n-partite graphs, which contain for instance authors, conferences, or topics as nodes, and edges that connect nodes of different type and represent relations (e.g., an author has published a paper in a conference), and (b) methods that create graphs with a single node type and edges that may vary in meaning depending on the application.

In the former category, in [4] a bipartite model that connects conferences to authors is proposed. Tripartite graph models for authors-conferences-topics have also been introduced in the past [2], [4]. In these cases the topics information is extracted from the paper titles and the resulting tripartite models expand the authors-topics model presented in [5]. Finally, in [6] the authors perform domain specific author and conference ranking by analyzing a bipartite author-conference graph using clustering and ranking heuristics.

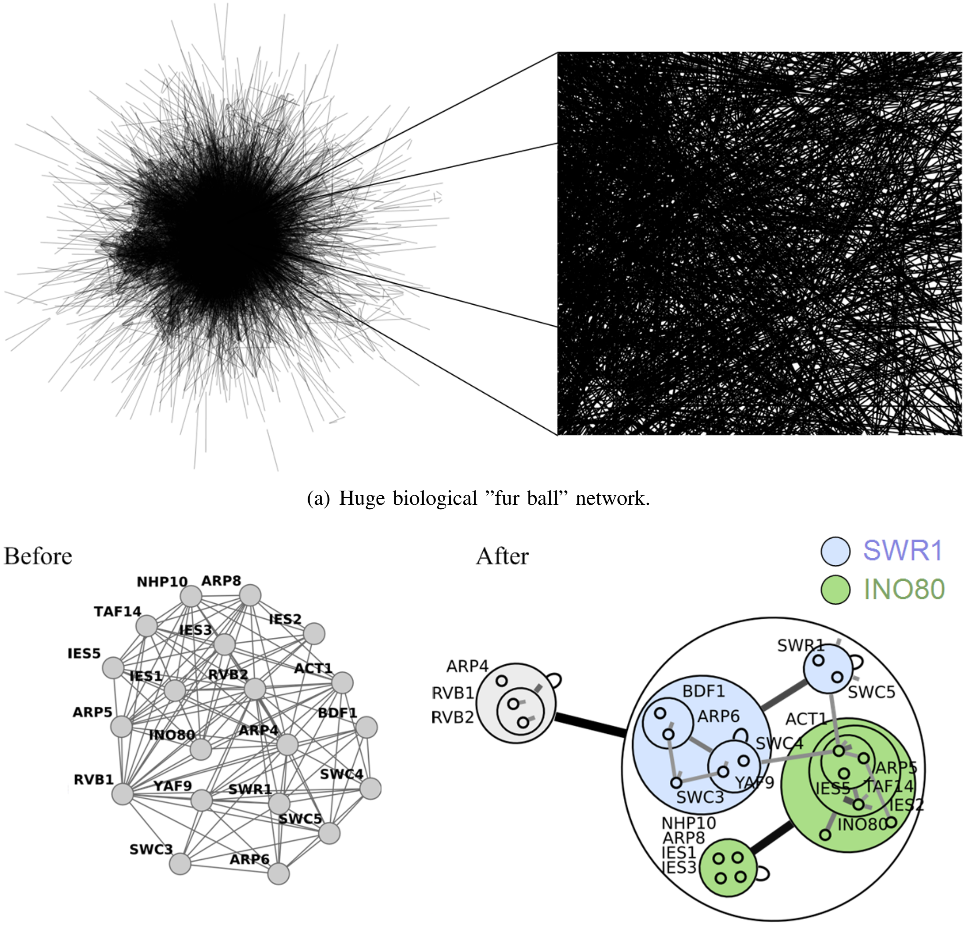
In the latter category, the graph nodes are usually the authors, with the edges representing either citation or co-authorship relations between the connected nodes. The first one is a directed citation graph [7], which is usually employed for ranking authors, whereas the second is an undirected co-authorship graph, which is mainly used for finding author communities (also known as *cliques*) [8], [9], but can also be employed for measuring *author centrality* [10], [11], a type of author importance. Some of the criteria that might be used to weigh the edges of such author graphs are: number of co-authored papers, content similarity between their publications, number of co-citations or couplings, and number of common conferences between the connected authors.

Another interesting type of graphs that can be constructed from bibliographical data are the co-author *hypergraphs*, where each edge (*hyperedge*) corresponds to a publication and connects all the co-authors of the specific publication. Author cliques can then be extracted from the graph [12]. In this direction, in our previous work [13], we presented an application of co-author *hypergraph* creation and clustering of its nodes.

B. Mining Bibliographical Graphs

Bibliographic data organization has attracted the application focus of many data mining research works. In the case of scientific community mining from publication records the challenge is to discover research communities that share common interests. In [14] a method is proposed that relies on the scientists' publication records in order to create scientific communities. Moreover, community mining systems have been proposed in the past, which use bibliographic data in order to discover and visualize researchers' communities [4], [15].

In our previous work [13], we experimentally studied the use of a novel semantic relatedness measure for the thematic organization of research papers in an attempt to improve the effectiveness of retrieval in bibliographic data. In particular, we used the *OMIOTIS* measure [16], which captures the semantic relatedness between text segments, and with its application we enabled the thematic organization of the bibliographic data stored in online databases. In this direction, of organizing bibliographical entries into thematic subsets based on text similarity, other research works have employed standard text classification techniques, e.g., *Bayesian* methods or *Support Vector Machines* (*SVM*) [17], *Concept Base Vector Space Models* [18], in order to assign research papers into appropriate categories. The combined utilization of metadata and full-text information for classifying bibliographic records into



(a) Huge biological ”fur ball” network.

(b) Before and after the application of *Power Graphs*.

Fig. 1. An example of a huge biological network is shown in Figure1(a) [3]. In a smaller scale example (Figure 1(b)) the application of *Power Graphs* demonstrates how shared protein complexes can be easily identified in the produced *Power Graph*.

appropriate subject classes [19] has also been proposed in the past.

Our work is complementary to the above research directions. In this paper we propose a novel method to construct and mine the co-authorship graph, in order to identify potential future research synergies. For this purpose, we use the similarity between authors from different communities, measured on the context of their publications. In the graph creation step, we visualize the co-authorship communities using a technique transferred from the biomedical domain, namely *Power Graph Analysis* [3], and in the second step we enrich the constructed *Power Graph* with similarity edges between *Power Nodes*, based on the text-to-text similarity between the authors' paper titles. Our method is unsupervised, thus obviates the need for collecting training data samples and its performance does not depend on the quality of any type of training examples.

C. Visualizing Graphs with Power Graphs

In biology and bioinformatics studies, networks play a crucial role. Yet, their analysis and representation is a difficult problem. Recent experimental and computational progress yields diverse networks of increased size and complexity. For example there are networks of several types, such as small and large scale *interaction networks*, *regulatory networks*, *genetic networks*, *protein-ligand interaction networks*, and *homology networks* analyzed and published regularly. A common way to access the information in a network is though direct visualization, but this often fails as it just results in *”fur balls”* from which little insight can be gathered. On the other hand, clustering techniques manage to avoid the problems caused by the large number of nodes and even larger number of edges by keeping a coarse-grained level of the networks' information and, thus, abstracting details. But these fail too since, in fact, much of the biological information lies in the details. Similar restrictions hold for bibliographic data and more specifically, for co-authorship graphs. The large number of authors and co-authorship edges makes the visualization task extremely difficult.

In order to provide an efficient methodology for visualizing large and complex biological networks, without loosing information, the authors in [3] present a novel methodology for analyzing and representing such networks, introducing *Power Graphs*. *Power Graphs* are a lossless representation

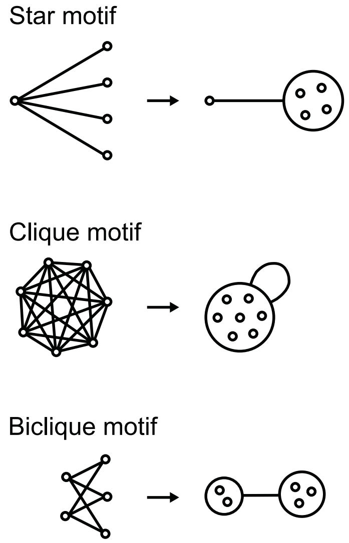


Fig. 2. The three basic motifs recognized by *Power Graphs*: *Star*, *Clique* and *Biclique*. *Power Nodes* are sets of nodes and *Power Edges* connect *Power Nodes*. A *Power Edge* between two *Power Nodes* signifies that all nodes of the first set are connected to all nodes of the second set. Nodes within a *Power Node* are not necessarily connected to each other.

of networks which reduces network complexity by explicitly representing *re-occurring network motifs*. Moreover, *Power Graphs* can be clearly visualized, as they compress up to 90% of the network's edges and are applicable to all types of networks such as *protein interaction*, *regulatory*, or *homology networks*. Figure 1 shows two examples of *"fur balls"* in biological networks: in the second example, which is small scale, the application of *Power Graphs* results into a visualization where the shared protein complexes can be easily identified, whilst in the original network, this was impossible [20]. In Figure 2 the three basic motifs recognized by *Power Graphs* (i.e. *Star*, *Clique* and *Biclique*) are shown. They constitute the basic transformations to cluster the nodes of an original graph into *Power Nodes*, connected with *Power Edges*.

In this paper, it is for the first time, to the best of our knowledge, that *Power Graphs* are applied for mining information from bibliographical graphs. In analogy to protein networks that contain proteins as nodes and edges that represent their interactions, the co-authorship graphs contain information about authors and their co-operations. As a consequence, finding re-occurring structural motifs in co-authorship graphs is a step towards discovering author communities. Additionally, in this work, based on the contextual similarity of publications we discover potential synergies between members from different communities.

III. Approach

A. *Co-authorship Graphs with* Power Graphs

The first decision regarding the creation of co-authorship graphs is on the type and meaning of edges. When a paper has k authors, the two representation alternatives are either to add a *hyperedge* connecting the k author nodes, or to add simple edges connecting each pairwise combination of the k authors. Since *Power Graphs* do not support *hyperedges* we work with the second alternative. However, using *hypergraphs* and a *hypergraph* partitioning algorithm [21] is another option. The second decision refers to the weighting of edges. Although many existing studies on the co-authorship graphs model the co-authorship relation by an undirected and unweighed edge, in this work we want to model the strength of the relation between authors, by adding edge weights. An edge weighting scheme for the author graph has been also employed in [12], with very interesting results.

TABLE I
An example of a bibliographic record.

$paper-id$	$Authors$
p_1	a_1,a_2,a_3
p_2	a_1,a_2,a_4
p_3	a_1,a_2
p_4	a_3,a_4
p_5	a_3,a_4

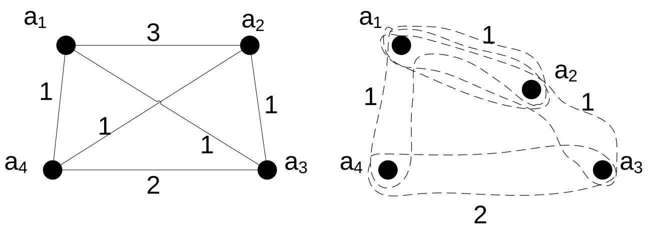


Fig. 3. The graph and hypergraph co-authorship models.

The resulting weighted co-authorship graph is formally modelled as follows. Let the graph $G = (V, WE)$, where V is the set of authors and a weighted edge $we = \{\nu_1, \nu_2, w_{\nu_1,\nu_2}\} \in WE$ represents that authors ν_1 and ν_2 have co-authored w_{ν_1,ν_2} papers. Similarly, this representation can be used if *hyperedges* are used, as follows. Let $G = (V, WHE)$, where V is the set of authors and a weighted *hyperedge* $whe = \{\nu_1, ..., \nu_n, w_{\nu_1,...,\nu_n}\} \in WHE$ represents that authors ν_1,..., ν_n have co-authored $w_{\nu_1,...,\nu_n}$ papers. An example of a bibliographic record and the resulting co-authorship graph and *hypergraph* is depicted in Table I and Figure 3 respectively.

B. Power Edges *Information*

The most important contribution of the *Power Graph* model is its ability to group several nodes into *Power Nodes* and to aggregate edges into *Power Edges*. However, the knowledge that can be extracted from each of the three main *Power Graph* motifs, namely *star*, *clique*, and *bi-clique* may differ. In the *star* motif a *Power Edge* connects an author with a set of co-authors. The *clique* motif corresponds to a clique of authors that frequently publish papers together and the corresponding *Power Edge* is a loop to the *Power Node* itself. Finally, in the *bi-clique* motif, a *Power Node* groups two or more *stars* and as a result the *Power Edge* connects two distinct author sets whose members have published papers together (one author

Edges	
------	
a1	a3
a1	a4
a1	a5
a2	a3
a2	a4
a2	a5
a3	a6
a3	a7
a3	a8
a3	a9
a3	a10
a20	a31
a20	a32
a21	a31
a21	a33
a22	a31
a22	a34

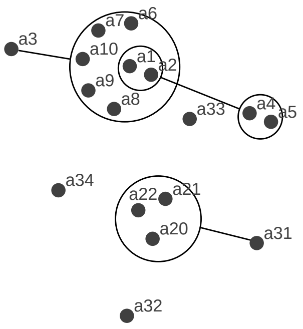


Fig. 4. A sample co-authorship *Power Graph*.

from each set).

The usability of the *clique* motif is limited in our paradigm, since all authors inside the *Power Node* of the clique have already published a joint work in the past, so the synergy has already been materialized. However, a further analysis of a *star* motif will probably reveal potential synergies. As mentioned before, the *Power Node* in the *star* motif contains all the co-authors of a given author. All these authors have a common point of reference and, consequently, if their interests match, they can form cooperations. However, finding *bi-cliques* in the co-authorship graph is the most straightforward indication of a potential research synergy. Each author in one set of the *bi-clique* has co-authored one or more papers with all authors in the other set but not with the authors in his own set. This motif depicts a possible cooperation among authors inside each *Power Node*. Finally, *Power Graphs* support another motif which can be very useful in our application. This is the *Power Node inclusion motif* [20], where a *Power Node* contains another *Power Node* and several more distinct authors. The *inclusion motif* can be also exploited for hidden synergy potentials. Figure 4 gives an example of co-authorship information and the corresponding *Power Graph*.

When the size of each *Power Node* in the *bi-clique* or the size of the *Power Node* in the *star* motif, or the external *Power Node* in the *inclusion* motif is large, then the authors inside the *Power Node* must be examined in terms of similarity of interests, as in these cases possible future research synergies may be found. In our model, similarity of interests is modelled as the similarity between their published works (i.e., between nodes that participated in the *Power Edge* creation). The algorithmic details of the proposed method are presented in the following section.

C. Algorithmic Description

The algorithmic description of our method given a publication database D is presented in Algorithm 1. We assume that authors (the nodes in our graph) are in lexicographical order. The first step in our method (lines 1 to 8) is the creation of the co-authorship graph G from the database of papers D. As explained before, for each paper p in the database, a set of weighted edges is added (or updated) to the co-authorship graph. The second step (line 9) is the application of the *Power Graph Analysis* algorithm to the original graph G and the creation of the *Power Graph PG*, which comprises *Power Nodes* (*pn*) that are either nested or form *cliques*, *stars* and *bi-cliques*. The details of the *Power Graph Analysis* algorithm are available in [3]. The final step of the algorithm comprises the examination of *Power Edges* (lines 10 to 14) and *Power Nodes* (lines 15 to 20) and results in a list of *Power Nodes* which may contain potential research synergies.

```
**Input**: Database of papers D, empty graph G={V,WE}
**Output**: A list of candidate Power Nodes CPN
1  foreach Paper $p \in D$ do
2      foreach Author $a \in p.authors$ do
3          foreach Author $b \in p.authors, b \neq a$ do
4              $V.add(a)$;
5              $V.add(b)$;
6              if $WE.containsKey(E_{(a,b)})$ then
7                  $WE.updateValueOf(E_{(a,b)})$;
8              else $WE.put((E_{(a,b)}, 1))$;
9  $PG\{PN, PE\} = PowerGraph(G)$;
10 foreach Power Edge $pe \in PE$ do
11     if $pe.node_1 \in PN$ then
12         $CPN.add(node_1)$;
13     if $pe.node_2 \in PN$ then
14         $CPN.add(node_2)$;
15 foreach Power Node $pn \in PN$ do
16     foreach node $n \in pn.nodes$ do
17         $pn_{temp}$ = a new empty power node ;
18         if $n \notin PN$ then
19             $pn_{temp}.add(n)$;
20             $CPN.add(pn_{temp})$;
```

Algorithm 1: The Enhanced Power Graph creation algorithm

D. Complexity and Implementation Issues

The computational complexity of our method is explained in the following. We assume that the database contains m papers written by n distinct authors, and that the resulting *Power Graph* contains *pn Power Nodes*. The first step, which is the creation of the initial co-authorship graph ($G = \{V, WE\}$, where V is a set of vertices and *WE* a map of edge-weight values), requires a single scan of the publication database. Given that the size of the graph does not exceed the size of the main memory, the complexity of the first step is $O(m)$.

The second step, relies on the *Power Graph* algorithm, which is a two-phase procedure. In the first phase the algorithm identifies potential *Power Nodes* using a *Jaccard*-based similarity metric on the neighbors of each node and a similarity based hierarchical clustering algorithm. For example, the similarity between two authors is maximum when they have written the same number of papers with the same co-authors. The second phase of the *Power Graph* algorithm performs a greedy search for *Power Edges*, by examining the problem of minimizing the *Power Graph* structure as an optimization problem. Since the details of the used *Power Graph* algorithm implementation are not known[5] we can simply assume that its complexity is relative to the complexity of the hierarchical algorithm ($O(n^2log(n))$ if the priority-queue *HAC* algorithm is implemented [22]), and to the complexity of the greedy power edge search algorithm, which is linear to the number of *Power Nodes* ($O(pn)$).

The final step, performs pairwise comparisons (using the paper title information) between authors in each *Power Node* that is of interest (i.e., *Power Nodes* that are nested, or form *bi-cliques*, or belong to a *star* motif) as described previously. In the worst case the complexity of this step is linear to the total number of *Power Nodes* (*pn*) and *Power Edges* (*pe*), in order for all the possible motifs to be checked. As a result, the complexity of the *Enhanced Power Graph* creation method is $O(m + n^2 \cdot log(n) + pn + pe + pn)$. Given that *Power Graph* reports an 80% reduction to the number of edges and nodes the resulting complexity is $O(m + n^2 \cdot log(n))$.

The output of the algorithm is a set of *Power Nodes* from the original *Power Graph*, which contains authors that can possibly co-operate in the future. The selected *Power Nodes* can be highlighted in the visualization of the *Power Graph*, or given as input to the author matching module, which examines author similarity of interests in terms of their papers' context.

E. A Walkthrough Example

In this section we present a more detailed examination of the example shown in Figure 4. For simplicity, in this example we assume that each paper has exactly two authors and corresponds to a single edge. Authors *a1* and *a2* have exactly the same co-authors (*a3*, *a4*, *a5*) but have never co-operated. The same holds for authors *a3*, *a4* and *a5* who have never worked together. This is depicted by a *bi-clique* in the *Power Graph*. In addition to the preview, author *a3* has collaborated with *a1*, *a2* and *a6* to *a10*. For this reason author *a3* forms a *star* with his co-authors, who form a pair of nested *Power Nodes* (*a1*, *a2* is inside the greater *Power Node*). Finally, all the co-authors of *a31* form a *star*, the *Power Node* of which is of potential interest. All other authors that have co-operated with a single author are ignored. If a threshold value is added on the size of *Power Nodes* to be examined, we will be able to further distill the candidate *Power Nodes* and find more promising matches.

IV. Evaluation and Results

A. Experimental Setup

In order to provide a demonstration of our method, we employed the *DBLP*[6] *Computer Science Bibliography*, which comprises more than 1.5 million publications. The database provides for each indexed paper the authors, title, venue and year of publication. The visualization of the complete graph would not make any sense since the *DBLP* database contains publications from many different research fields. Thus, we have selected subsets of the *DBLP* dataset, which comprise papers published in the same conferences, and the same years. For the graphs' presentation we provide two alternative visualizations: one that contains all the *Power Nodes* and *Edges* and one that contains only the strongest edges.

Data processing is done as described in the previous sections: (a) we create the initial co-authorship graph from the selected subset of publications, (b) we generate the *Power Graph* from the initial graph, and, (c) we prune the weakest components of the *Power Graph* in order to improve the readability of the result. Finally, we present in details the most interesting structures in each power graph.

B. Results on the DBLP Data

1) The database conferences: In the first experiment, we process the *DBLP* publications from the top-5 conferences in Databases[7], namely: *SIGMOD*, *VLDB*, *PODS*, *ICDE*, *ICDT*. The subset contains $11,369$ papers published since 1969. The papers have been written by $10,524$ authors. Several papers have more than two authors, and several author pairs have co-authored more than one paper. In order to reduce the complexity and improve readability of the graph, we omit authors that have written only one paper in any of these conferences. The resulting graph finally contains $3,860$ nodes (authors) and $15,382$ edges (co-authorship entries).

After applying the *Power Graph* algorithm, the resulting graph shown in Figure 5 contains $1,601$ power nodes and $8,572$ power edges. A modified version of the *Power Graph*, where the weakest *Power Edges* have been pruned away is presented in Figure 6 and uncovers interesting substructures of the original graph. In Figures 7 and 8 we zoom on the *Power Graph* in order to present some of these structures. The potential research synergies must be searched in cases like the ones we highlight: (a) in Figure 7 the co-authors of an author who has a *star* motif (e.g., the 4 co-authors of *H.P. Kriegel*, in bold face font), may co-operate with all other authors in the *Power Node* of the *star* motif (e.g., authors in italics), (b) in Figure 8 the authors in a *Power Node* (or *bi-clique*), which is nested in another *Power Node* (e.g., *C.Jermain* and authors in bold face fonts), may co-operate with all other authors in the outer *Power Node* (e.g., authors in italics).

An additional piece of information that we can easily draw from *Power Graphs* are the author *cliques* (e.g., the co-authors of *P. Kriegel* in Figure 7) that correspond to authors who co-operate frequently. The cliques are easily distinguished from groups of authors that have co-operatively written several papers (e.g., the group of authors on the top of the big *Power Node* in Figure 8), which are closely placed in the graph but do not form a *Power Node*.

[5] http://www.biotec.tu-dresden.de/research/schroeder/powergraphs/

[6] http://www.informatik.uni-trier.de/~ley/db/

[7] as provided by Microsoft Academic Search in http://academic.research.microsoft.com

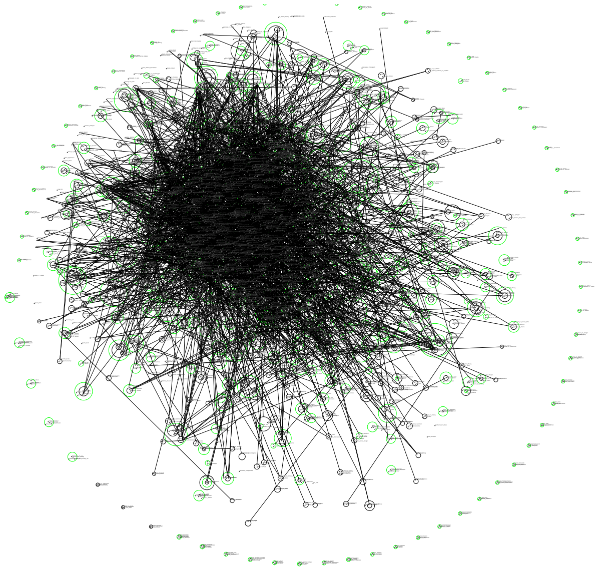


Fig. 5. Authors *Power Graph* (top-5 database conferences).

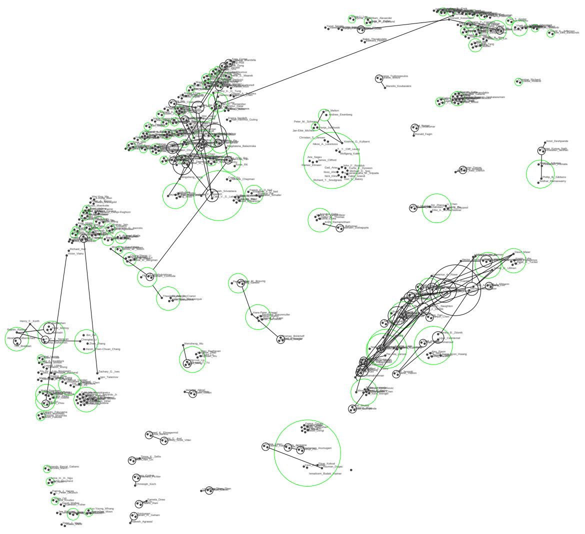


Fig. 6. Pruned author *Power Graph* (top-5 database conferences).

2) Multi-disciplinary graphs: In the second experiment we attempt to visualize the *Power Graph* of two research communities from different disciplines, namely computer graphics and information retrieval. More specifically we select papers that have been published in *SIGGRAPH* and *SIGIR*, the top conferences in computer graphics and information retrieval respectively. The selected subset comprises $3,568$ papers written by $5,386$ authors. Following the same author pruning strategy, we produce a graph that contains $1,303$ nodes (authors) and $2,769$ edges (co-authorship entries). The respective *Power Graph* shown in Figure 9 contains $1,391$ *Power Edges* and 523 *Power Nodes*, and forms two distinct, compact graph regions (*IR* community is on the left and *Graphics* community on the right). All the small subgraphs between the two main regions belong to one or other field, do not connect to the *Power Nodes* of each subgraph and have been placed by the *Power Graph* drawing module in between the two sub graphs only for presentation purposes (i.e., they are not special *Power Nodes* that lie between the two research fields).

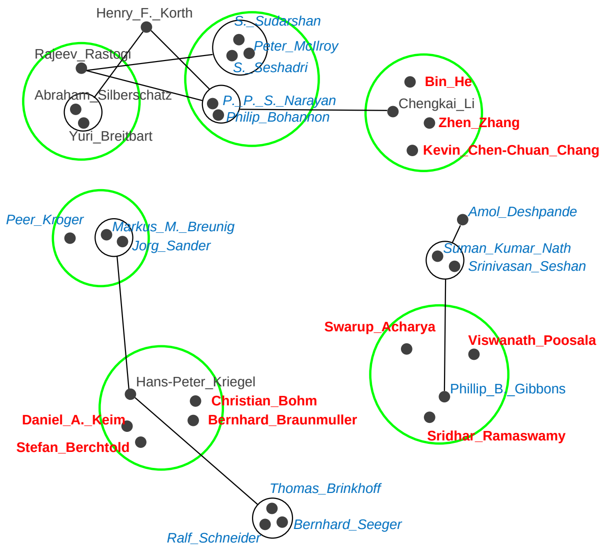


Fig. 7. Part of the pruned author *Power Graph* (top-5 database conferences). Star motif

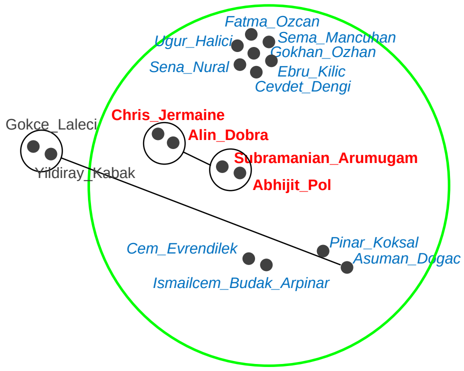


Fig. 8. Part of the pruned author *Power Graph* (top-5 database conferences). Power Node nesting motif.

3) Measuring similarity based on content: In this final experiment, we further examine the cases of possible cooperation between authors by measuring their similarity of interests based on the titles of their publication record. We employ the *OMIOTIS* measure [16] and the methodology we

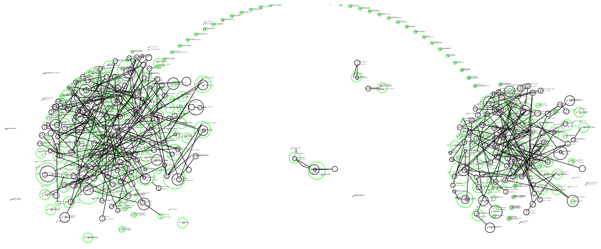


Fig. 9. Pruned author *Power Graph* (*SIGIR* and *SIGGRAPH* conferences).

presented in [13]. For each candidate pair of authors we measure the average semantic relatedness between their published work in the respective conferences, then we sort candidate author pairs in decreasing similarity score. The candidate pairs are selected as described in Section IV-B1, taking care to remove candidates that have already collaborated in the past.

The top results for the database conferences subset are presented in Table II. Authors in the first positions have common co-authors, but have not co-authored a paper yet. A manual examination of their publication record reveals that their interests match. For example the first pair of authors works on *Business Process Modelling*, the second pair works on *Privacy*, and the third on *SQL server's optimization*.

TABLE II
TOP CANDIDATE PAIRS, RANKED BY SIMILARITY

Author A	Author B	similarity
Daniel_Deutch	Anat_Eyal	0.222
Kristen_LeFevre	Alexandre_V,_Evfimievski	0.179
Ming-Chuan_Wu	Steve_Herbert	0.156
Babu_Krishnaswamy	Aleksandras_Surna	0.154
Alon_Y,_Halevy	Chen_Li	0.152
Ming-Chuan_Wu	Aleksandras_Surna	0.148
Jorg_Sander	Daniel_A,_Keim	0.139
Conor_Cunningham	Steve_Herbert	0.138
Sandeepan_Banerjee	Anand_Manikutty	0.136
Yanif_Ahmad	Magdalena_Balazinska	0.135

V. CONCLUSIONS

In this paper, we have introduced a novel approach for the organization and the efficient presentation of bibliographic database contents. The contribution of our approach lies on the use of a graph reduction method that facilitates the efficient visualization of the dense co-authorship graph, the identification of potential research synergies based on the analysis of the *Power Graph*, and the ranking of potential co-author pairs by similarity of interests. More specifically, we have demonstrated how the use of *Power Graph Analysis* can uncover potential future research synergies between authors. This modular approach helps us to avoid the burden of finding the optimal clustering and classification scheme for bibliographic data organization. As a proof of concept, we demonstrated some of the capabilities of our approach in the *DBLP* data and we believe that it can be fruitfully explored in several other data mining tasks. It is on our next plans to apply the same approach to more bibliographic networks as well as to other social networks, and to study the evolution of the graphs over time based on the comparison of different graph snapshots taken in different years.

REFERENCES

[1] A. Doms and M. Schroeder, "Semantic search with gopubmed," in *REWERSE*, 2009, pp. 309–342.

[2] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 990–998.

[3] L. Royer, M. Reimann, B. Andreopoulos, and M. Schroeder, "Unraveling protein networks with power graph analysis," *PLoS Computational Biology*, vol. 4, no. 7, p. e1000108, 2008.

[4] O. R. Zaiane, J. Chen, and R. Goebel, "Dbconnect: mining research community on dblp data," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ser. WebKDD/SNA-KDD '07. New York, NY, USA: ACM, 2007, pp. 74–81. [Online]. Available: http://doi.acm.org/10.1145/1348549.1348558

[5] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, ser. UAI '04. Arlington, Virginia, United States: AUAI Press, 2004, pp. 487–494. [Online]. Available: http://portal.acm.org/citation.cfm?id=1036843.1036902

[6] Y. Sun, T. Wu, Z. Yin, H. Cheng, J. Han, X. Yin, and P. Zhao, "Bibnetminer: mining bibliographic information networks," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '08. New York, NY, USA: ACM, 2008, pp. 1341–1344. [Online]. Available: http://doi.acm.org/10.1145/1376616.1376770

[7] W. Ke, K. Borner, and L. Viswanath, "Major information visualization authors, papers and topics in the acm library," in *Proceedings of the IEEE Symposium on Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 216.1–. [Online]. Available: http://portal.acm.org/citation.cfm?id=1038262.1038825

[8] T.-H. Huang and H. M. L., "Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers," in *Proceedings of the IEEE International Conference on Computer Graphics, Imaging and Visualisation*, ser. CGIV '06, 2006, pp. 18–23.

[9] T.-H. Huang and M. L. Huang, "Visualization of individual"s knowledge by analyzing the citation networks," *International Conference on Computer Graphics, Imaging and Visualization*, vol. 0, pp. 465–470, 2007.

[10] M. A. Nascimento, J. Sander, and J. Pound, "Analysis of sigmod's co-authorship graph," *SIGMOD Rec.*, vol. 32, pp. 8–10, September 2003. [Online]. Available: http://doi.acm.org/10.1145/945721.945722

[11] 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*, vol. 36, pp. 39–43, September 2002. [Online]. Available: http://doi.acm.org/10.1145/792550.792556

[12] Y. Han, B. Zhou, J. Pei, and Y. Jia, "Understanding importance of collaborations in co-authorship networks: A supportiveness analysis approach," in *Proceedings of the Ninth SIAM International Conference on Data Mining*. ACM-SIAM, 2009, pp. 1111–1122.

[13] G. Tsatsaronis, I. Varlamis, S. Stamou, K. Nørvåg, and M. Vazirgiannis, "Semantic relatedness hits bibliographic data," in *WIDM*, 2009, pp. 87–90.

[14] S. Rodriguez, I. Oliveira, and J. de Souza, "Competence mining for virtual scientific community creation," *International Journal of Web Based Communities*, vol. 1, no. 1, pp. 90–102, 2002.

[15] R. Ichise, H. Takeda, and K. Ueyama, "Community mining tools using bibliography data," in *Proc. of the 9th International Conference on Information Visualization*, 2005, pp. 953–958.

[16] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, "Text relatedness based on a word thesaurus," *J. Artif. Intell. Res. (JAIR)*, vol. 37, pp. 1–39, 2010.

[17] R. Angelova and G. Weikum, "Graph-based text classification: learn from your neighbors," in *SIGIR*, 2006, pp. 485–492.

[18] T. Shimano and T. Yuakawa, "An automated research paper classification method for the IPC system with the concept base," in *Proc. of the NTCIR-7 Workshop Meeting*, 2008.

[19] A. Montejo-Raez, L. Urena-Lopez, and R. Steinberger, "Text categorization using bibliographic records: beyond document content," in *Proc. of the 21st Conference of the Spanish Society for NLP*, 2005, pp. 119–126.

[20] L. Royer, *Unraveling the Structure and Assessing the Quality of Protein Interaction Networks with Power Graph Analysis*. Dresden, Germany: PhD Thesis, Technical University of Dresden, 2010.

[21] N. Selvakkumaran and G. Karypis, "Multi.objective hypergraph partitioning algorithms for cut and maximum subdomain degree minimization," in *Proceedings of the 2003 IEEE/ACM international conference on Computer-aided design*, ser. ICCAD '03, 2003, pp. 726–.

[22] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.