Mining potential research synergies from co-authorship graphs using power graph analysis

Iraklis Varlamis*

Department of Informatics and Telematics,
Harokopio University of Athens,
89, Harokopou St., Athens 17671, Greece
E-mail: varlamis@hua.gr
*Corresponding author

George Tsatsaronis

Bioinformatics Group,
Biotechnology Center (BIOTEC),
Technical University of Dresden,
Tatzberg 47/49, Dresden 01307, Germany
E-mail: george.tsatsaronis@biotec.tu-dresden.de

Abstract: Bibliographic databases are a prosperous field for data mining research and social network analysis. They contain rich information, which can be analysed across different dimensions (e.g., author, year, venue, and topic) and can be exploited in multiple ways. The representation and visualisation 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, candidate reviewers for a paper 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 visualisation tool from the biological domain, we are able to provide comprehensive visualisations that help us uncover hidden relations between authors and suggest potential synergies between researchers or groups.

Keywords: bibliographic databases; co-authorship graphs; graph analysis; similarity metrics; graph mining; graph-based recommendations.

Reference to this paper should be made as follows: Varlamis, I. and Tsatsaronis, G. (2012) 'Mining potential research synergies from co-authorship graphs using power graph analysis', *Int. J. Web Engineering and Technology*, Vol. 7, No. 3, pp.250–272.

Biographical notes: Iraklis Varlamis is a Lecturer at the Department of Informatics and Telematics of Harokopio University of Athens. He received his PhD in Computer Science from Athens University of Economics and Business, Greece. He is a former member of the DB-NET (http://www.db-net.aueb.gr/ Head: Professor Vazirgiannis) and WIM (http://wim.aueb.gr Head: Associate Professor Vassalos) research groups. His research interests vary from data-mining and the use of semantics in web mining to graph analysis in social networks. He has published several articles in international journals and conferences, concerning web document clustering, the use of semantics in web

link analysis and web usage mining, word sense disambiguation using thesauruses, etc. More information is available at http://www.dit.hua.gr/~varlamis.

George Tsatsaronis holds a PhD in Text Mining from the Department of Informatics, Athens University of Economics and Business. He is currently a Senior Postdoctoral Fellow at the Bioinformatics group of the Technical University of Dresden (BIOTEC, focusing in text mining techniques for the biomedical domain. In the past, he was a Postdoctoral Fellow at the Department of Computer and Information Science, Database Systems Group, Norwegian University of Science and Technology, in the field of distributed text mining. His research interests include word sense disambiguation, text retrieval, classification, and clustering, statistical natural language processing, measures of lexical semantic relatedness and similarity, semantic smoothing kernels, and supervised and unsupervised techniques for embedding knowledge resources in text applications.

1 Introduction

Currently, vast amounts 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 (Doms and Schroeder, 2009), which specialises 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* (http://www.arnetminer.org/) (Tang et al., 2008) or *Microsoft Academic Search* (http://academic.research.microsoft.com/). Authors and venues ranking, organisation by year or topic, author profiles extraction and authors name disambiguation are only some of the services provided on top of these databases. Other services visualise co-authorship information, e.g., the 'instant graph search'[1], 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.

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 visualising 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* (Royer et al., 2008), 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 visualisation 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* 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 visualised in an efficient manner. Finally, we show how we can predict possible research synergies between authors from this final augmented graph.

The contribution of this work can be summarised in the following:

- The reduction of the co-authorship graph's complexity, with the use of *power graph analysis* techniques. The use of these techniques significantly reduces the complexity of the problem, in comparison to traditional co-authorship analysis techniques.
- The efficient application of data mining techniques in the resulting *power graph*, which results in the identification of potential research synergies between authors that share interests and co-authors but have not yet collaborated.

The rest of the paper is organised as follows: Section 2 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 3 introduces our approach for mining potential research synergies from co-authorship graphs. Section 4 demonstrates our findings stemming from the application of our approach to bibliographic data. Finally, Section 5 concludes and provides pointers to future work.

2 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

c results are presented either in the form of node clusters, e.g., authors by topic, conferences by topic, or using a graph visualisation 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 visualise 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 visualisation options, in the following we summarise the most important research works in the field and illustrate the different alternatives in each process.

2.1 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)

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 Zaiane et al. (2007) a bipartite model that connects conferences to authors is proposed. Tripartite graph models for authors-conferences-topics have also been introduced in the past (Tang et al., 2008; Zaiane et al., 2007). In these cases, the topics information is extracted from the paper titles and the resulting tripartite models expand the authors-topics model presented in Rosen-Zvi et al. (2004). Finally, in Sun et al. (2008), the authors perform domain specific author and conference ranking by analysing 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 (Ke et al., 2004), 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*) (Huang and Huang, 2006, 2007), but can also be employed for measuring *author centrality* (Nascimento et al., 2003; Smeaton et al., 2002), 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 (Han et al., 2009). In this direction, in our previous work

(Tsatsaronis et al., 2009), we presented an application of co-author *hypergraph* creation and clustering of its nodes.

2.2 Mining bibliographical graphs

Bibliographic data organisation 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 Rodriguez et al. (2002), 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 visualise communities of researchers (Zaiane et al., 2007; Ichise et al., 2005; Tsatsaronis et al., 2011a). In Tsatsaronis et al. (2011a), we used power graphs to represent co-authorship networks, which were created using bibliographic data. This significantly reduced the size of the graph and increased the scalability of our community mining algorithm. Using power graph analysis techniques in the same bibliographical data (DBLP data) and data mining techniques on the evolving co-authorship power graph, we proposed a novel methodology for clustering authors based on their publication and cooperation profile (Tsatsaronis et al., 2011b).

In our previous work (Tsatsaronis et al., 2009), we experimentally studied the use of a novel semantic relatedness measure for the thematic organisation of research papers in an attempt to improve the effectiveness of retrieval in bibliographic data. In particular, we used the *OMIOTIS* measure (Tsatsaronis et al., 2010), which captures the semantic relatedness between text segments, and with its application we enabled the thematic organisation of the bibliographic data stored in online databases.

In the direction of organising 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) (Angelova and Weikum, 2006), *concept base vector space models* (Shimano and Yuakawa, 2008), in order to assign research papers into appropriate categories. The combined utilisation of metadata and full-text information for classifying bibliographic records into appropriate subject classes (Montejo-Raez et al., 2005) has also been proposed in the past.

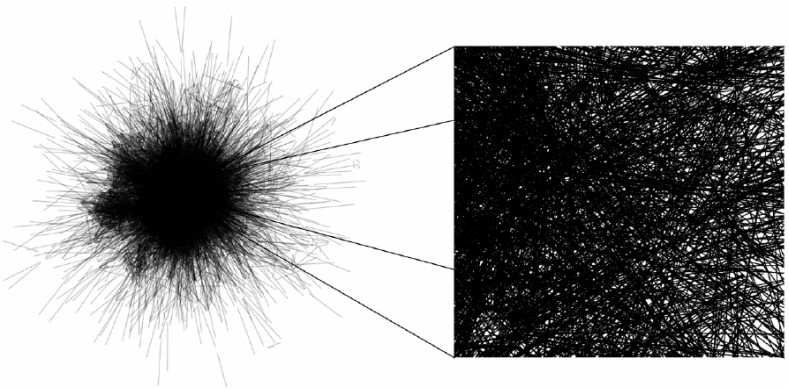
Our work is complementary to the above research directions. In this paper, we propose a novel method to construct a co-authorship graph and mine author communities, in order to identify potential future research synergies. For this purpose, we use the similarity between authors belonging to different communities. In the graph creation step, we visualise the co-authorship communities using a technique transferred from the biomedical domain, namely *power graph analysis* (Royer et al., 2008), 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. Similarity is computed using an unsupervised semantic relatedness measure (i.e., a measure which compares concepts instead of terms), which results to similarity edges between *power nodes* that denote similarity in authors' interests in conceptual level.

2.3 Visualising graphs with power graphs

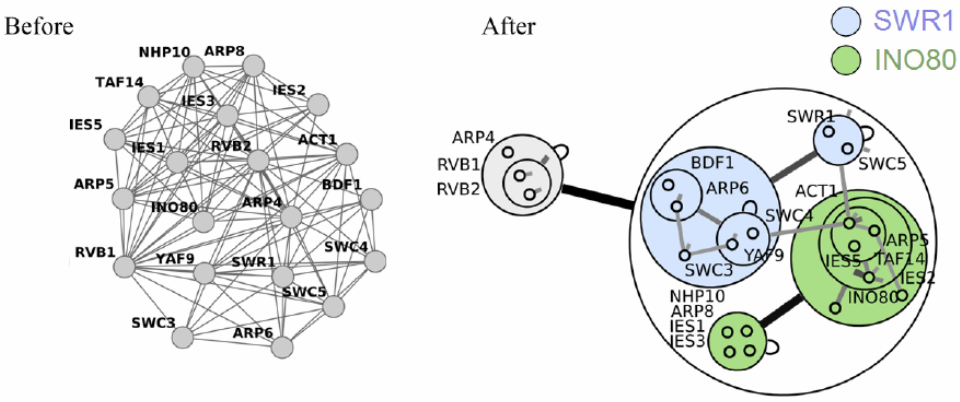
In biology and bioinformatics studies, networks play a crucial role. Yet, their analysis and representation is a difficult problem. Recent experimental and computational progress

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* analysed and published regularly. A common way to access the information in a network is though direct visualisation, but this often fails as it just results in 'fur balls' [see Figure 1(a)] 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.

Figure 1 (a) huge biological 'fur ball' network (b) before and after the application of *power graphs* (see online version for colours)



(a)



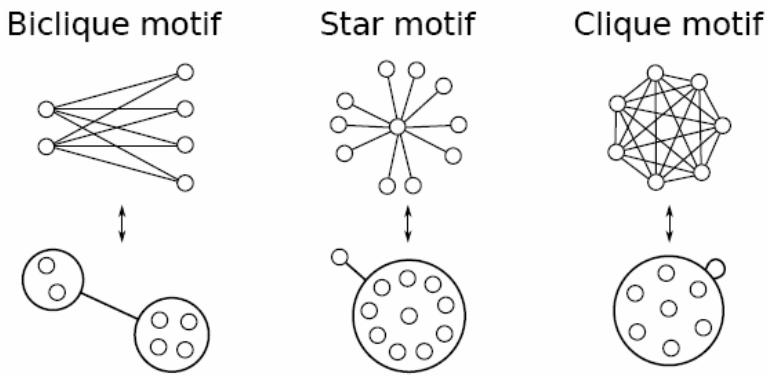
(b)

Notes: An example of a huge biological network is shown in Figure 1(a). 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*.

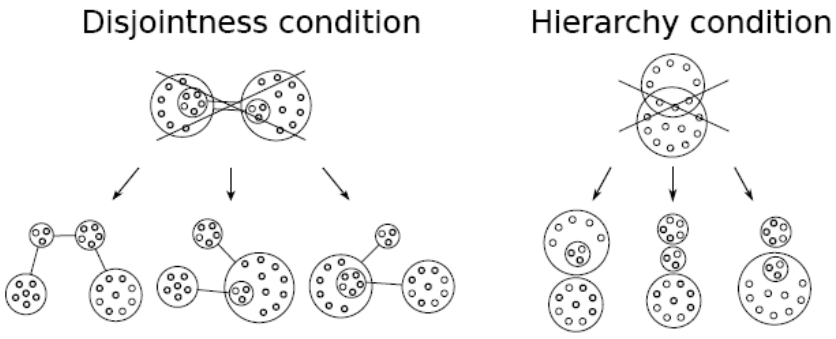
In the direction of providing an efficient methodology for visualising large and complex biological networks, like the graphs constructed from bibliographical data may be, without loosing information, the authors in Royer et al. (2008) present a novel

methodology for analysing and representing such networks, introducing *power graphs*. *Power graphs* are a lossless representation of networks which reduces network complexity by explicitly representing *re-occurring network motifs*. Moreover, *power graphs* can be clearly visualised, as they compress up to 90% of the edges in biological networks and are applicable to all types of networks such as *protein interaction*, *regulatory*, or *homology networks*. In Royer et al. (2008), the authors demonstrate the usefulness of *power graph analysis* on five detailed biological examples ranging from *protein-ligand binding* to *regulatory networks* and *homology networks*. Figure 1 shows two examples of 'fur balls' in biological networks (Royer, 2010): in the original network, it is difficult to visualise and understand the patterns of interaction between proteins due to the huge number of nodes and edges, whereas in the second example, which is in small scale, the application of *power graphs* reduces the number of nodes and edges and results into a visualisation, where the shared protein complexes can be identified more easily.

Figure 2 Basic motifs recognised by *power graphs*, (a) *power graph* semantics: biclique, star, and clique motifs. *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. (b) *Power graph* conditions and their equivalent decompositions



(a)



(b)

The three basic motifs recognised by *power graphs* are shown in Figure 2(a). 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*.

In the following, we will define *power graphs* formally. Given a graph $G = (V, E)$, where V is the set of nodes or vertices and $E \subseteq V \times V$ is the set of edges that are unordered pairs of distinct nodes, a *power graph* $G' = (V', E')$ is a graph defined on the *power set* of nodes $V' \subseteq \mathfrak{P}(V)$, whose elements (*power nodes*) are connected to each other by *power edges*: $E' \subseteq V' \times V'$. The two power nodes of a power edge must be disjoint or identical: $\forall (u, v) \in E' : (u \cap v = \emptyset) \vee (u = v)$. A *power edge* is a set of edges. Hence, *power graphs* are defined both on the power set of nodes $V' \subseteq \mathfrak{P}(V)$, as well as on the *power set* of edges $E' \subseteq \mathfrak{P}(E)$. The set of nodes V in G is the union of all *power nodes* v'. Hence, $V = \cup_{v' \in V'} v'$. The set V' of all *power nodes* is required to be minimal, i.e., each $v' \in V'$ must participate in at least one *power edge* $e' \in E'$, or be a singleton set.

In Figure 2(a), we can now understand in detail the types of motifs recognised by *power graph analysis*. If two *power nodes* are connected by a *power edge* in G', this signifies that in G all nodes of the first *power node* are connected to all nodes of the second *power node*, thus, the two sets form a complete connected bipartite subgraph. It does not imply that the nodes inside each power node are connected among each other. A special case of biclique is the star where one of the two *power nodes* is a singleton node. If a *power node* in G' is connected to itself by a reflexive *power edge*, this means that all nodes inside the *power node* are connected to each other by edges in G, thus, the set is a complete connected subgraph.

Since *power graphs* are drawn in the plane, two conditions are required:

a disjointness of power edges

b hierarchy of power nodes.

Condition (a) means that each edge $e \in E$ of the original graph G is represented by one and only one *power edge* $e' \in E'$ in the power graph G' of G. Condition (b) means that any two power nodes $v_1, v_2 \in V'$ in G' are either disjoint, or one is included in the other. Figure 2(b) shows the possible decompositions to fulfil these two conditions. Relaxing the previous two conditions leads to *abstract power graphs* that are difficult to visualise and interpret.

The quality of a *power graph* is measured using edge reduction [equation (1)].

$$\Re = \frac{(|E| - |E'|)}{|E|} \tag{1}$$

Edge reduction quantifies the amount by which the number of edges in G' is smaller than in G, relatively to the number of edges in G. Edge reduction is essentially the *compression rate* achieved by the *power graph* transformation. It assesses the improvement of a *power graph* representation over the original graph, without considering the meaning of the indicated structures.

Algorithm 1 The description of the *power graph analysis* algorithm

Input: A graph $G = \{V, E\}$, Minimum similarity $s_{\min}$, Weight w_{uv} for each edge (u, v) in E

Output: A *power graph* $G' = \{V', E'\}$

Initialise C and C' to empty sets and M to an empty numeric matrix
Add for each node v in V the singleton cluster $\{v\}$ to C and to C'
Update M with $s(U, W)$ for each pair of clusters (U, W) in C
while $|C'| > 1$ and $s_{\max}(M) \geq s_{\min}$ **do**
    Find (U, W) with the maximum $s_{\max}$ in M
    Merge U and W, and add the new cluster to C and C'
    Update M with the similarities of the new cluster to the rest
Add neighbourhood $N(U)$ of each cluster U to C, if $s(N(U)) > s_{\min}$
Initialise V' and E' to empty sets, and L to an empty list
For each node $v \in V$ add a singleton set $\{v\}$ to V'
foreach (U, W) in C **do**
    if $U \cap W = \emptyset$ *and* $(U \cup W, U \times W)$ is a *sub-graph in G* **then**
        Add *power edge* (U, W) to L
        Compute the weight of the edge (U, W)
    if *$U = W$ and the U-induced graph in G is a clique* **then**
        Add *power edge* (U, U) to L
        Compute the weight of the edge (U, U)
while $L \neq \emptyset$ **do**
    Remove the *power edge* (U, W) with the largest weight
    if $\exists S \in V' : U \cap S \neq \emptyset$ *but* $U \not\subset S$ *and* $S \not\subset U$ **then**
        Add to L the candidate power edges $(U \setminus S, W)$ and $(U \cap S, W)$
    else if $\exists S \in V' : W \cap S \neq \emptyset$ *but* $W \not\subset S$ and $S \not\subset W$ **then**
    Add to L the candidate power edges $(U, W \setminus S)$ and $(U, W \cap S)$
    else if $\exists S \in V' : (U \times W) \cap (S \times T) \neq \emptyset$ **then**
    if $U \subset S$ **then**
        Add to L the candidate *power edge* $(U, W \setminus T)$
    else if $U \subset T$ **then**
    Add to L the candidate *power edge* $(U, W \setminus S)$
    else if $W \subset S$ **then**
    Add to L the candidate *power edge* $(U \setminus T, W)$
    else if $W \subset T$ **then**
    Add to L the candidate power edge $(U \setminus S, W)$

33 **else if** (*U*, *W*) *is a clique* **then**
34 Add *power node U* to *V′* and *power edge* (*U*, *U*) to *E′*
35 **else**
36 Add *power nodes U* and *W* to *V′* and *power edge* (*U*, *W*) to *E′*
37 **foreach** *edge* (*u*, *v*) ∈ *E not yet covered by any* power edge *in E′* **do**
38 Add the singleton *power edge* ({*u*}, {*v*}) to *E′*
39 Return *G′* = (*V′*, *E′*)

In the following, we introduce formally the algorithm with which the near-minimal *power graph* representation can be produced from an input graph. The algorithm, described in Algorithm 1, supports weighted graphs and requires as input the original graph *G*, and a minimum similarity threshold $s_{\min}$. The role of $s_{\min}$ will be justified sufficiently during the detailed explanation of the algorithm that follows. The algorithm consists of a first phase that collects candidate *power nodes* and a second phase that uses these to search for *power edges*. In the first phase (lines 1–10 of Algorithm 1), candidate *power nodes* are identified with hierarchical clustering (Eisen et al., 1998) based on neighbourhood similarity. A candidate power node is a set of nodes that have neighbours in common. The similarity of two neighbourhoods is a *generalised Jaccard Index* (Rasmussen, 1992) on weighted sets. The similarity score is shown in equation (2).

$$s\big(N(u), N(v)\big) = \frac{\sum_{x \in N(u) \cap N(v), x \neq u, v} \min\big(w_{xu}, w_{xv}\big) + \alpha}{\sum_{x \in N(u) \cup N(v), x \neq u, v} \min\big(w_{xu}, w_{xv}\big)} \tag{2}$$

where $N(u)$ is the neighbourhood of cluster u, w_{xu} the weight of x in the weighted neighbourhood of u, and α the clique contribution to the similarity. The value of α is given by equation (3), in case $\min(w_{uv}, w_{vu}, w_{uu}, w_{vv}) > 0$, and otherwise is 0.

$$\alpha = \frac{1}{2}\big(w_{uv} + w_{vu} + w_{uu} + w_{vv}\big) \tag{3}$$

The weight of a neighbour n in the weighted neighbourhood of cluster c is the average over all nodes in cluster c [equation (4)].

$$w_{nc} = \frac{\sum_{x \in c} w_{nx}}{|c|} \tag{4}$$

For the identification of stars and other highly asymmetric bicliques, we add for each node v two sets to the candidate *power nodes*: its neighbourhood set $N(v)$ and the set of common neighbours of the nodes in $N(v)$, $\cap_{v' \in N(v)} N(v')$, that contains at least v. Each of these clusters u is only added if its accumulated neighbourhood similarity is above the given threshold: $s(N(u)) > s_{\min}$, where $s_{\min}$ is given as input to Algorithm 1.

In the second phase (lines 11–38 of Algorithm 1) power edges are searched. The *minimal power graph* problem is to be seen as an optimisation problem to find the *power graph* achieving the highest edge reduction. The *greedy power edge* search follows the heuristic of making the local optimum decision at each step with the aim of finding the global optimum, or at least to come close to it. Among the candidate *power nodes* found

in phase one, each pair that forms a complete connected (bipartite) subgraph in G is a candidate *power edge*. The candidates abstracting the most edges are added successively to the *power graph*. If necessary, candidates are decomposed, e.g., Figure 2(b).

3 Approach

3.1 Co-authorship graphs with power graphs

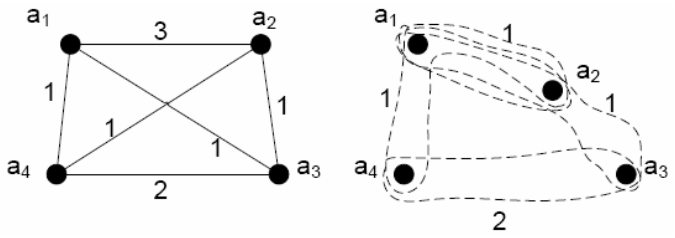
The first decision regarding the creation of co-authorship graphs is on the type and meaning of edges. When a paper has k authors, the two representation alternatives are either to add a *hyperedge* connecting the k author nodes, or to add simple edges connecting each pairwise combination of the k authors. Since *power graphs* do not support *hyperedges* we work with the second alternative. However, using *hypergraphs* and a *hypergraph* partitioning algorithm (Selvakkumaran and Karypis, 2003) 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 Han et al. (2009), with very interesting results.

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 $we = \{v_1, v_2, w_{v_1,v_2}\} \in WE$ represents that authors v_1 and v_2 have co-authored w_{v_1,v_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* $whe = \{v_1, ..., v_n, w_{v_1,...,v_n}\} \in WHE$ represents that authors $v_1, \ldots, v_n$ have co-authored $w_{v_1,...,v_n}$ papers. An example of a bibliographic record and the resulting co-authorship graph and *hypergraph* is depicted in Table 1 and Figure 3, respectively.

Table 1 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

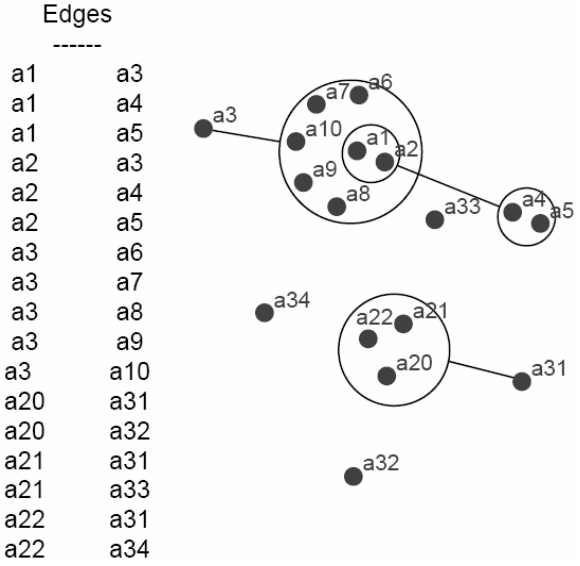
Figure 3 The graph and hypergraph co-authorship models



3.2 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 *biclique* 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 biclique 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 from each set).

Figure 4 A sample co-authorship *power graph*



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 materialised. 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 *bicliques* in the co-authorship graph is the most straightforward indication of a potential research synergy. Each author in one set of the *biclique* 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 graph analysis* supports the *power node inclusion motif* (Royer, 2010), when a *power node* (in our case a group of authors) contains another *power node* and some more distinct authors. The outer *power node* corresponds to a group of authors who have co-authored many papers, whereas the inner *power node* is a subset of the outer comprising authors who co-author more. Searching for potential synergies between authors in the inner node and the additional authors in the outer node

can be based on the similarity of interests and provides interesting results, as shown in the experimental section. Figure 4 gives an example of co-authorship information and the corresponding *power graph*.

When the size of each *power node* in the *biclique* 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.

3.3 Algorithmic description

The algorithmic description of our method given a publication database *D* is presented in Algorithm 2. 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 *bicliques*. 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.

Algorithm 2 The enhanced power graph creation algorithm

Input: Database of papers D, empty graph G = {V, WE}
Output: A list of candidate power nodes CPN

```
1  foreach Paper p ∈ D do
2      foreach Author a ∈ p.authors do
3          foreach Author b ∈ p.authors, b ≠ 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 ∈ PE do
11     if pe.node1 ∈ PN then
12         CPN.add(node1);
13     if pe.node2 ∈ PN then
14         CPN.add(node2);
```

```
15  foreach Power Node pn ∈ PN do
16      foreach node n ∈ pn.nodes do
17          pn_temp = a new empty power node;
18          if n ∉ PN then
19              pn_temp.add(n);
20              CPN.add(pn_temp);
```

3.4 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 neighbours 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 minimising the *power graph* structure as an optimisation problem. Since the details of the used *power graph* algorithm implementation are not known (http://www.biotec.tu-dresden.de/research/schroeder/powergraphs/) we can simply assume that its complexity is relative to the complexity of the hierarchical algorithm [$O(n^2\log(n))$ if the priority-queue *HAC* algorithm is implemented (Manning et al., 2008)], 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 cooperate in the future. The selected *power nodes* can be highlighted in the visualisation 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.

3.5 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 cooperated. 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 cooperated 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 distil the candidate *power nodes* and find more promising matches.

4 Evaluation and results

4.1 Experimental setup

In order to provide a demonstration of our method, we employed the *DBLP* (http://www.informatik.uni-trier.de/~ley/db/) *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 (see Figure 5 for a projected view of DBLP entries). The visualisation 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 visualisations: one that contains all the *power nodes* and *edges* and one that contains only the strongest edges.

Figure 5 Sample entries from the DBLP database (see online version for colours)

title	year	pubkey	authorname
Speed up interactive image retrieval.	2009	journals/vldb/ShenJTHZ09	Kian-Lee Tan
Speed up interactive image retrieval.	2009	journals/vldb/ShenJTHZ09	Xiaofang Zhou
Speed up interactive image retrieval.	2009	journals/vldb/ShenJTHZ09	Heng Tao Shen
Speed up interactive image retrieval.	2009	journals/vldb/ShenJTHZ09	Zi Huang
Speed up interactive image retrieval.	2009	journals/vldb/ShenJTHZ09	Shouxu Jiang
B-tries for disk-based string management.	2009	journals/vldb/AskitisZ09	Justin Zobel
B-tries for disk-based string management.	2009	journals/vldb/AskitisZ09	Nikolas Askitis
Localized monitoring of kNN queries in wireless s...	2009	journals/vldb/YaoTL09	Ee-Peng Lim
Localized monitoring of kNN queries in wireless s...	2009	journals/vldb/YaoTL09	Yuxia Yao
Localized monitoring of kNN queries in wireless s...	2009	journals/vldb/YaoTL09	Xueyan Tang
Hierarchically compressed wavelet synopses.	2009	journals/vldb/SacharidisDS09	Timos K. Sellis
Hierarchically compressed wavelet synopses.	2009	journals/vldb/SacharidisDS09	Antonios Deligiannakis
Hierarchically compressed wavelet synopses.	2009	journals/vldb/SacharidisDS09	Dimitris Sacharidis

Data processing is done as described in the previous sections:

a we create the initial co-authorship graph from the selected subset of publications

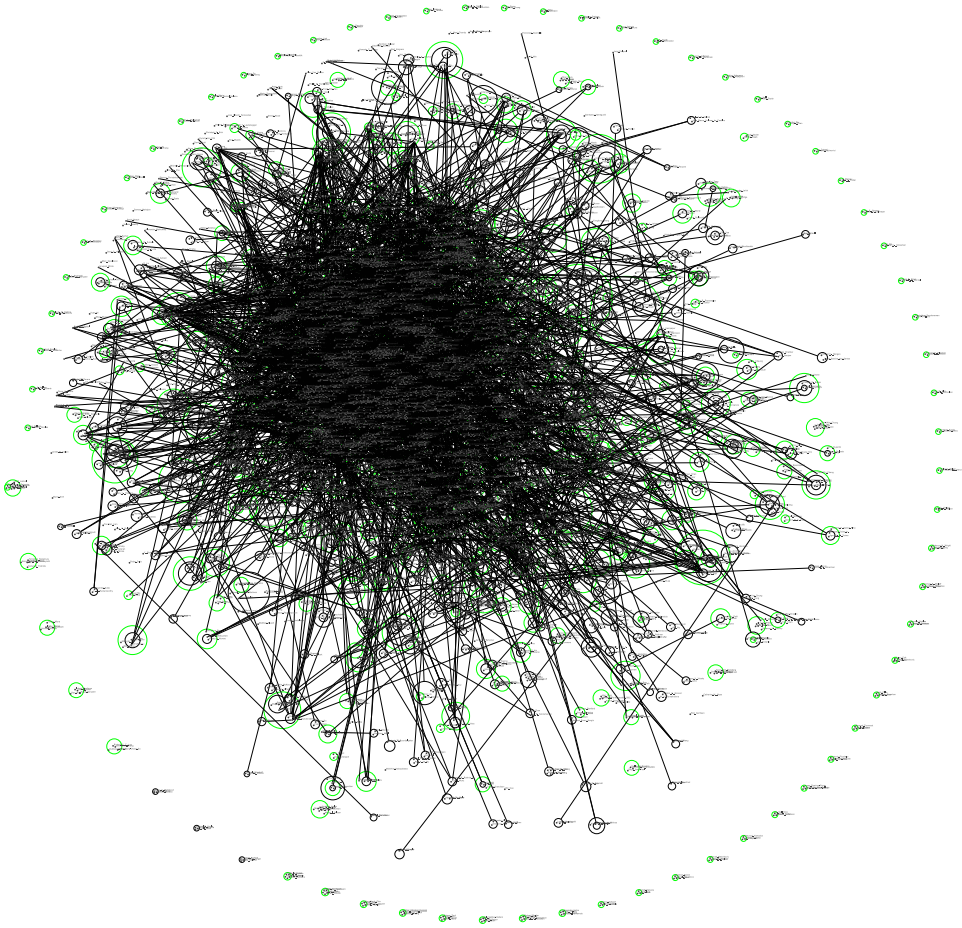
b we generate the *power graph* from the initial graph

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.

In our experiments, we used the command line version of power graph analysis tool (http://www.biotec.tu-dresden.de/schroeder/group/powergraphs/ev_clt_usage.html), which is written in Java and exports graph visualisation in portable network graphics (*PNG*) and scalable vector graphics (*SVG*) formats.

Figure 6 Authors *power graph* (top-5 database conferences) (see online version for colours)

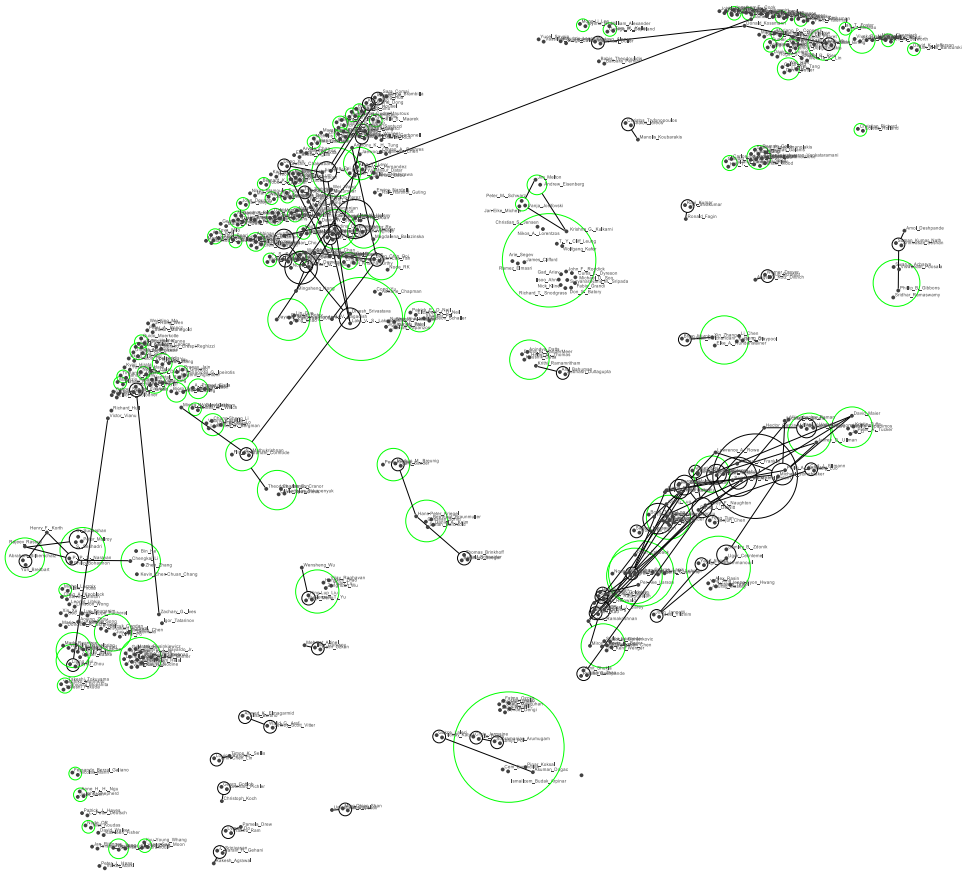


4.2 Results on the DBLP data

4.2.1 The database conferences

In the first experiment, we process the DBLP publications from the top-5 conferences in Databases[3], namely: *SIGMOD* (http://www.sigmod.org/about-sigmod), *VLDB* (http://www.vldb.org), *PODS* (http://www.sigmod.org/the-pods-pages), *ICDE* (http://www.icde.org/), *EDBT/ICDT* (http://icdt.tu-dortmund.de/). 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).

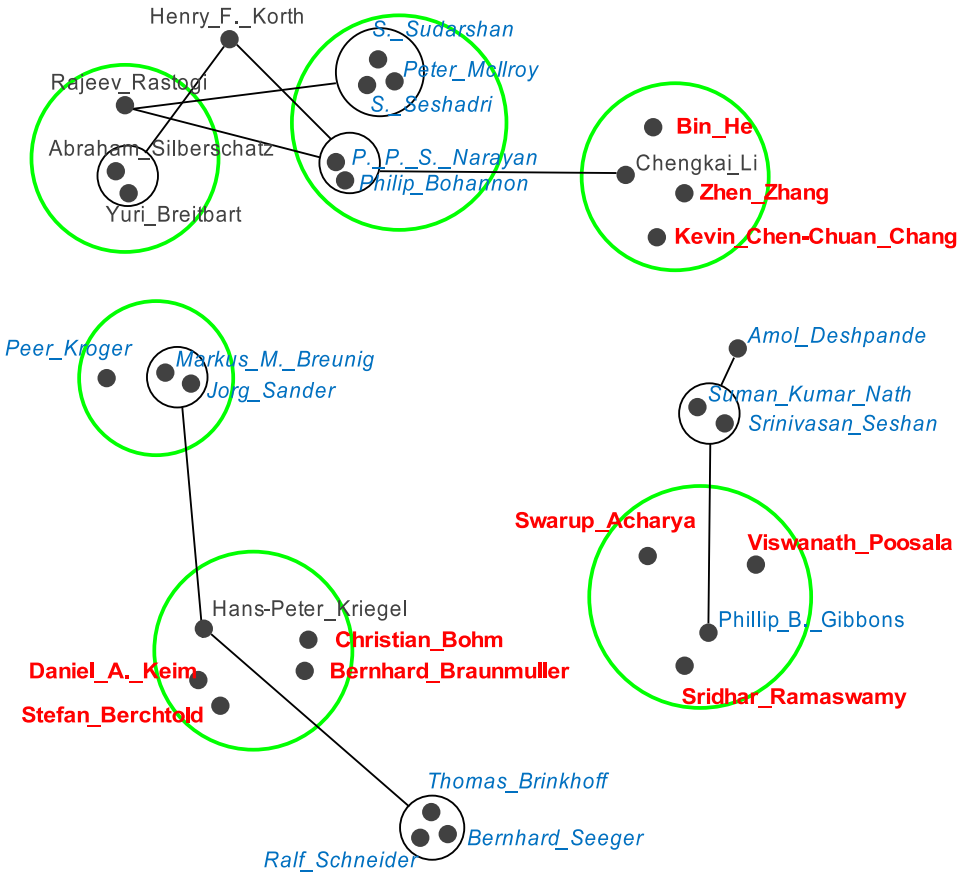
Figure 7 Pruned author-*power graph* (top-5 database conferences) (see online version for colours)



After applying the *power graph* algorithm, the resulting graph shown in Figure 6 contains 1,601 power nodes and 8,572 power edges. In this graph, *power nodes* contain authors that frequently co-author (have written more than *threshold* papers together), whereas

power edges connect individual authors or groups of authors to other groups, with whom they co-author (again above the same *threshold*) but less frequently. A modified version of the *power graph*, where the weakest *power edges* have been pruned away (using a higher *threshold* is presented in Figure 7 and uncovers interesting substructures of the original graph.

Figure 8 Part of the pruned author-*power graph* (top-5 database conferences) (see online version for colours)



Note: Star motif

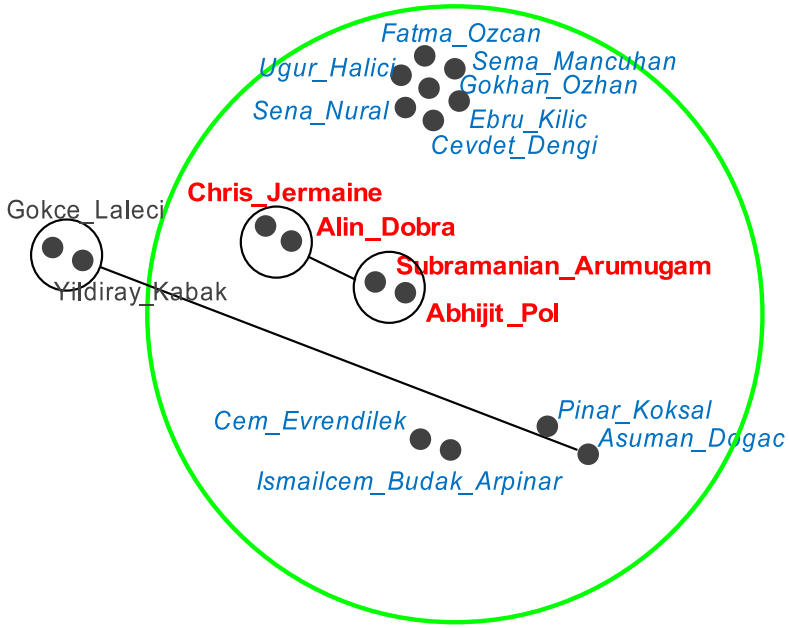
In Figures 8 and 9, we zoom on the *power graph* in order to present some of these structures. The potential research synergies must be searched in cases like the ones we highlight:

a in Figure 8, the co-authors of an author that belongs in a start motif (e.g., the four co-authors of *H.P. Kriegel*, in bold face font), may cooperate with the authors in the *power node* of the star motif (e.g., authors in italics)

b in Figure 9, 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 cooperate with all other authors in the outer *power node* (e.g., authors in italics).

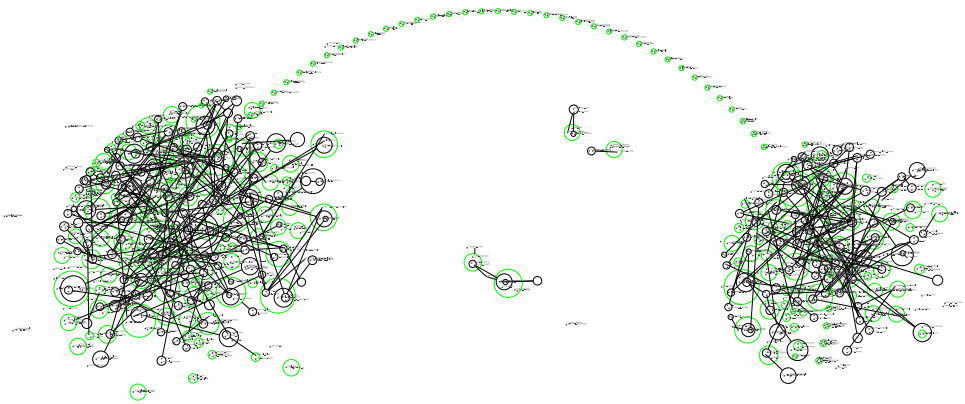
An additional piece of information that we can easily draw from *power graphs* are the author *cliques* (e.g., the co-authors of *P. Kriegel* in Figure 8) that correspond to authors who cooperate frequently. The cliques are easily distinguished from groups of authors that have cooperatively written several papers (e.g., the group of authors on the top of the big *power node* in Figure 9), which are closely placed in the graph but do not form a *power node*.

Figure 9 Part of the pruned author-*power graph* (top-5 database conferences) (see online version for colours)



Note: *Power node* nesting motif

Figure 10 Pruned author *power graph* (SIGIR and SIGGRAPH conferences) (see online version for colours)



4.2.2 Multi-disciplinary graphs

In the second experiment, we attempt to visualise 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* (http://www.siggraph.org/) and *SIGIR* (http://www.sigir.org/), 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 10 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).

4.2.3 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* (http://omiotis.hua.gr) measure (Tsatsaronis et al., 2010) and the methodology we presented in Tsatsaronis et al. (2009). For each candidate pair of authors, we measure the average semantic relatedness between their published works in the respective conferences, then we sort candidate author pairs in decreasing similarity score. The candidate pairs are selected as described in Section 4.2.1, taking care to remove candidates that have already collaborated in the past.

The top results for the database conferences subset are presented in Table 2. 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 optimisation*.

Table 2 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

5 Conclusions

In this paper, we have introduced a novel approach for the organisation 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 visualisation 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 organisation. 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.

Acknowledgements

Authors would like to thank Matthias Reimann and Michael Schroeder who generously provided us access to the power graph analysis tools and original figures.

References

Angelova, R. and Weikum, G. (2006) 'Graph-based text classification: learn from your neighbors', in *SIGIR*, pp.485–492.

Doms, A. and Schroeder, M. (2009) 'Semantic search with gopubmed', in Bry, F. and Maluszynski, J. (Eds.): *Semantic Techniques for the Web*, pp.309–342, Springer-Verlag, Berlin, Heidelberg.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl. Acad. Sci. USA*, Vol. 95, No. 25, pp.14863–14868.

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

Huang, T-H. and Huang, M.L. (2006) '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, CGIV '06*, pp.18–23.

Huang, T-H. and Huang, M.L. (2007) 'Visualization of individual's knowledge by analyzing the citation networks', *International Conference on Computer Graphics, Imaging and Visualization*, pp.465–470.

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

Ke, W., Borner, K. and Viswanath, L. (2004) 'Major information visualization authors, papers and topics in the ACM library', in *Proceedings of the IEEE Symposium on Information Visualization*, p.216.1, IEEE Computer Society, Washington, DC, USA.

Manning, C.D., Raghavan, P. and Schtze, H. (2008) *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.

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

Nascimento, M.A., Sander, J. and Pound, J. (2003) 'Analysis of Sigmod's co-authorship graph', *SIGMOD Rec.*, September, Vol. 32, pp.8–10.

Rasmussen, E. (1992) 'Clustering algorithms', in *Information Retrieval – Data Structures and Algorithms*, pp.419–442, Prentice Hall.

Rodriguez, S., Oliveira, I. and de Souza, J. (2002) 'Competence mining for virtual scientific community creation', *International Journal of Web Based Communities*, Vol. 1, No. 1, pp.90–102.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004) 'The author-topic model for authors and documents', in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI '04*, pp.487–494, AUAI Press, Arlington, Virginia, USA.

Royer, L. (2010) 'Unraveling the structure and assessing the quality of protein interaction networks with power graph analysis', PhD thesis, Technical University of Dresden, Dresden, Germany.

Royer, L., Reimann, M., Andreopoulos, B. and Schroeder, M. (2008) 'Unraveling protein networks with power graph analysis', *PLoS Computational Biology*, Vol. 4, No. 7, p.e1000108.

Selvakkumaran, N. and Karypis, G. (2003) '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, ICCAD '03*, p.726, IEEE Computer Society, Washington, DC, USA.

Shimano, T. and Yuakawa, T. (2008) 'An automated research paper classification method for the IPC system with the concept base', in *Proc. of the NTCIR-7 Workshop Meeting*.

Smeaton, A.F., Keogh, G., Gurrin, C., McDonald, K. and Sødring, T. (2002) 'Analysis of papers from twenty-five years of sigir conferences: what have we been doing for the last quarter of a century?', *SIGIR Forum*, September, Vol. 36, pp.39–43.

Sun, Y., Wu, T., Yin, Z., Cheng, H., Han, J., Yin, X. and Zhao, P. (2008) 'Bibnetminer: mining bibliographic information networks', in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08*, pp.1341–1344, ACM, New York, NY, USA.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2008) 'Arnetminer: extraction and mining of academic social networks', in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, pp.990–998, ACM, New York, NY, USA.

Tsatsaronis, G., Reimann, M., Varlamis, I., Gkorgkas, O. and Nørvåg, K. (2011a) 'Efficient community detection using power graph analysis', in *Proceedings of the 9th Workshop on Large-scale and Distributed Informational Retrieval, LSDS-IR '11*, pp.21–26, ACM, New York, NY, USA.

Tsatsaronis, G., Varlamis, I., Torge, S., Reimann, M., Nørvåg, K., Schroeder, M. and Zschunke, M. (2011b) 'How to become a group leader? Or modeling author types based on graph mining', in *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries, TPDL'11*, pp.15–26, Springer-Verlag, Berlin, Heidelberg.

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

Tsatsaronis, G., Varlamis, I., Stamou, S., Nørvåg, K. and Vazirgiannis, M. (2009) 'Semantic relatedness hits bibliographic data', in *WIDM*, pp.87–90.

Zaiane, O.R., Chen, J. and Goebel, R. (2007) '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, WebKDD/SNA-KDD '07*, pp.74–81, ACM, New York, NY, USA.

Notes

1 Co-author Path in Microsoft Academic Search
2 Co-author Graph in Microsoft Academic Search
3 As provided by Microsoft Academic Search in http://academic.research.microsoft.com