Constraint-Based Mining of Sets of Cliques Sharing Vertex Properties

Pierre-Nicolas Mougel[1,2], Marc Plantevit[1,3], Christophe Rigotti[1,2], Olivier Gandrillon[4], and Jean-François Boulicaut[1,2]

[1] Université de Lyon, CNRS, INRIA
[2] INSA-Lyon, LIRIS Combining, UMR5205, F-69621, France
firstname.lastname@liris.cnrs.fr
[3] Université Lyon 1, LIRIS Combining, UMR5205, F-69622, France
[4] Université Lyon 1, Centre de Génétique Moléculaire et Cellulaire, UMR5534, F-69622, Villeurbanne, France
lastname@cgmc.univ-lyon1.fr

Abstract. We consider data mining methods on large graphs where a set of labels is associated to each vertex. A typical example of such graphs is a social network of collaborating researchers where additional information represent the main publication targets (preferred conferences or journals) for each author. We investigate the extraction of sets of dense subgraphs such that the vertices in all subgraphs of a set share a large enough set of labels. As a first step, we consider here the special case of dense subgraphs that are cliques. We proposed a method to compute all *maximal homogeneous clique sets* that satisfy user-defined constraints on the number of separated cliques, on the size of the cliques, and on the number of labels shared by all the vertices. The empirical validation illustrates the scalability of our approach and it provides experimental feedback on two real datasets, more precisely an annotated social network derived from the DBLP database and an enriched biological network of protein-protein interactions. In both cases, we discuss the relevancy of extracted patterns thanks to available domain knowledge.

Keywords: graph mining, network analysis, pattern discovery, constraint-based mining

1 Introduction

Many data can be represented by means of graphs where vertices represent entities and edges represent relationships between entities. For instance, graphs provide a natural representation of important real life networks such as biological networks and social networks. In the last years, such network data became increasingly available and network/graph mining indeed turned out to be one of the most studied and challenging tasks for the data mining community. Many researchers developed approaches to mine graphs in two different and complementary ways. On the one hand, some methods focus on macroscopic graph properties (e.g., degree distribution, diameter) [9] or graph partitioning [12]. On the other hand, many proposals concern the extraction of more sophisticated properties within a pattern discovery setting. In particular, local pattern mining in graphs has received a lot of attention, leading to the introduction of new problems (like support counting in case of non-relational graphs [7, 8]) and resulting in new algorithms to mine collections of graphs [38, 25, 15, 36], single graphs [18, 27], and time-evolving graphs [6, 3, 32].

Most of the existing methods work on graph data only. However, as mentioned, for instance in [21], more informative graphs are often given that can be represented as graphs with feature vectors associated to each vertex. In [21], the authors take advantage of the complementarity of the information carried by edges and features, and extract dense subgraphs such that the vertices in one subgraph share a large enough set of features. In this paper, we consider that datasets denote graphs where a set of labels is associated to each vertex. Such a set can represent a set of Boolean property values and can also be used to encode a discrete feature vector (when feature domains are transformed to be disjoint). Over such graphs, we investigate the extraction of *sets of dense subgraphs* such that the vertices in *all subgraphs* of a set share a large enough set of labels. As a first step, we consider here the special case of dense subgraph known as a clique.

Following a pattern discovery setting within the constraint-based mining framework, we introduce the problem of extracting *maximal homogeneous clique sets* which are sets of cliques that satisfy constraints on the number of separated cliques, on the clique sizes and on the number of labels shared by all the vertices. We propose an efficient algorithm to extract maximal homogeneous clique sets in a complete way. It should be noticed that enumerating all cliques is a time consuming operation (due to the large number of cliques), and that potentially enumerating all the clique sets is obviously even more difficult. Let us now motivate our research thanks to applications from system biology and social network analysis.

Example 1 (Mining Protein-Protein Interaction Graphs). Protein-protein interaction databases contain a large number of interactions. These interactions can be modeled as a protein-protein interaction graph where vertices are proteins and two vertices are connected if the two proteins are known to interact. In biology, a functional module is a cluster of proteins that interact together in a specific cellular process. Thus, cliques often model functional modules. Protein-protein interaction graph can be enriched by gene expression data since proteins are products of genes. In our framework, situations where genes are overexpressed can be considered as labels. [15, 21, 35] consider the same kind of cross fertilization and enable to discover meaningful patterns. Maximal homogeneous clique sets, can help the expert to discover different functional modules that share a large enough set of biological situations. This kind of knowledge is highly valuable since it gives rise to links between these modules. These links can provide evidences of cooperative or competitive actions of groups of genes overexpressed in the same biological situations.

Example 2 (Social Network Analysis). One of the most studied task in social network analysis is the discovery of communities [37]. A community is, for instance, a group of people who share some interests and are connected by strong social interaction. Communities can be modeled as dense subgraphs (e.g., cliques) in which vertices share a large enough set of properties. Extracting maximal homogeneous clique sets, can help to identify sets of communities that share the same interests. For instance, in a co-authorship network (e.g., DBLP[5]), a maximal homogeneous clique set can describe groups of co-authors that are active on the same topics. It can be used, for instance, to design program committees and selection processes.

We provide an experimental feedback in this two contexts (protein interaction network and co-authorship network) on real datasets, showing that using constraints enable to focus on small collections of meaningful patterns. Furthermore, we show that the extraction process is scalable: it can be performed on large graphs with hundreds of thousands of vertices.

The rest of the paper is organized as follows. Section 2 introduces the definitions and the mining methods. Our experiments are reported in Section 3. Related work is discussed in Section 4. Section 5 briefly concludes.

2 Maximal homogeneous clique set

In this section, we define the patterns of interest and introduce a correct method to compute them.

2.1 Pattern definition

Definition 1. *(Dataset)*

Let $\mathcal{L}$ be a set of labels, a dataset *is a pair $\langle G, f\rangle$, where $G = \langle \mathcal{V}, \mathcal{E}\rangle$ is a simple undirected graph (vertices $\mathcal{V}$ and edges $\mathcal{E}$), and f is a total function $f : \mathcal{V} \to 2^{\mathcal{L}}$ associating a set of labels to each vertex.*

A set C of vertices is called a *clique* in G if the subgraph of G induced by C is complete. The collection of all cliques in G is denoted $\mathcal{C}_G$

Definition 2. *(Homogeneous Clique Set) Let α, β, and κ be three strictly positive integers, a* Homogeneous Clique Set *(HCS) in dataset $\langle G, f\rangle$ is a collection M of cliques $\{C_1, \ldots, C_n\} \subseteq \mathcal{C}_G$ such that the three following constraints $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$ and $\mathbb{C}^{sep}$ are satisfied:*

- $\mathbb{C}^{lab}_{\alpha}$ *: $|\bigcap_{C\in M}(\bigcap_{v\in C} f(v))| \geq \alpha$, i.e., the vertices share at least α labels;*
- $\mathbb{C}^{clique}_{\kappa,\beta}$ *: M contains at least κ cliques of size at least β;*
- $\mathbb{C}^{sep}$ *: for all C, C' in M, with $C \neq C'$, we have $C \cup C' \notin \mathcal{C}_G$, i.e., the cliques in M are* separated.

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

It should be noticed that the cliques in a HCS are not required to be maximal cliques in G because this would be too restrictive. Furthermore, the $\mathbb{C}^{sep}$ constraint is needed to avoid a large clique of G to be split and counted as several cliques in a HCS.

For a dataset, the collection of HCS is likely to be large, so we focus on the maximal ones, that are in some sense the most specific.

Definition 3. *(Maximal Homogeneous Clique Set and partial order $\preceq$) A Maximal Homogeneous Clique Set (MHCS) is a HCS which is maximal w.r.t. the partial order $\preceq$ defined as follows. Given M_1 and M_2 two HCS, $M_1 \preceq M_2$ if and only if for all $C_1 \in M_1$ there exists $C_2 \in M_2$ such that $C_1 \subseteq C_2$.*

In general, antisymmetry does not hold for $\preceq$ in any collection of sets, but in the special case of a collection of HCS $\preceq$ is a partial order as state by the following theorem.

Theorem 1. *The relation $\preceq$ is a partial order in a collection of HCS.*

Proof. The relation is trivially reflexive and transitive. To show antisymmetry, consider M_1 and M_2 two HCS such that $M_1 \preceq M_2$ and $M_2 \preceq M_1$. Suppose that $M_1 \neq M_2$, then there exists C in M_1 that is different from all sets in M_2. And since $M_1 \preceq M_2$, there exists C' in M_2 such that $C \subset C'$. As, $M_2 \preceq M_1$, there exists C'' in M_1 such that $C' \subseteq C''$. So, $C \subset C''$, and as elements of a HCS must be separated, this cannot hold. □

The *maximal homogeneous clique set problem* consists in finding all the MHCS in a given dataset.

2.2 Finding all MHCS

Let $vertices(M)$ be the set of vertices appearing in a collection M of sets of vertices, $vertices(M) = \bigcup_{C \in M} C$. The following theorem states that if M is a MHCS, then knowing $vertices(M)$ is sufficient to determine M.

Theorem 2. *Given M a MHCS, then M is the collection of all maximal cliques in the subgraph G_M of G induced by $vertices(M)$.*

Proof (sketch). Let S be the collection of all maximal cliques in G_M and suppose that $M \neq S$. Since S contains all maximal cliques, and M contains cliques that must be separated, then $S \not\subset M$. So, there exists $D \in S$ such that $D \notin M$.

If D is a superset of a clique C in M, since cliques in M are separated, replacing C by D in M leads to a pattern that satisfies $\mathbb{C}^{sep}$ (and also $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$), and thus M is not a maximal HCS.

If D is not a superset of a clique in M, since D is a maximal clique in G_M then all cliques in $M \cup \{D\}$ are separated. So $M \cup \{D\}$ satisfies $\mathbb{C}^{sep}$ (and also $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$), and again M is not a maximal HCS. □

For a dataset $\langle G, f\rangle$, f can be encoded as a binary relation $\mathcal{R} \subseteq \mathcal{V} \times \mathcal{L}$, defined as $x\mathcal{R}y \Leftrightarrow y \in f(x)$, and relating each vertex to its labels. The MHCS are related to the so-called *closed sets* over $\mathcal{R}$ as states by the next theorem.

Let us consider the mappings g and h, defined as follows, $g : 2^{\mathcal{V}} \rightarrow 2^{\mathcal{L}}, g(X) = \{y \in \mathcal{L} | \forall x \in X, \mathcal{R}(x, y)\}$ and $h : 2^{\mathcal{L}} \rightarrow 2^{\mathcal{V}}, h(Y) = \{x \in \mathcal{V} | \forall y \in Y, \mathcal{R}(x, y)\}$. These mappings define a *Galois connection* between $2^{\mathcal{V}}$ and $2^{\mathcal{L}}$ (e.g., see [39, 28]), a set of vertices $V \subseteq \mathcal{V}$ (resp. of labels $L \subseteq \mathcal{L}$) is said *closed* in $\mathcal{R}$ if $V = h(g(V))$ (resp. $L = g(h(L))$), and when restricted to closed sets, the mappings g and h are anti-isomorphisms (i.e., bijections that assign to A, B, s.t. $A \subseteq B$, the images A', B' s.t. $B' \subseteq A'$).

Theorem 3. *Given M a MHCS, then $vertices(M)$ is closed in $\mathcal{R}$.*

Proof (sketch). Let $V = vertices(M)$, $L = g(V)$, $V' = h(L)$, and suppose that V is not closed in $\mathcal{R}$, then $V \subset V'$. Let M' be the collection of all maximal cliques in the subgraph of G induced by V'. This collection satisfies $\mathbb{C}^{sep}$. Since M is a MHCS, then by Theorem 2 M is the collection of all maximal cliques in the subgraph of G induced by V, and thus $M \preceq M'$, and M' satisfies $\mathbb{C}^{clique}_{\kappa,\beta}$ because M satisfies it. Since $V' = h(L)$, the vertices in V' share at least as many labels as the vertices in V. As M satisfies $\mathbb{C}^{lab}_{\alpha}$, M' satisfies it also. So M is not a maximal HCS. □

Let $maxCliques(G, V)$ be the collection of maximal cliques in the subgraph of G induced by V. Theorems 2 and 3 lead to the two following correct ways to find all MHCS:

- Find the closed sets of vertices V in $\mathcal{R}$, such that $maxCliques(G, V)$ satisfies $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$ and $\mathbb{C}^{sep}$. Among them retain the maximal ones. For each of this maximal set V, output $maxCliques(G, V)$.
- Or alternatively, using the Galois connection, find the minimal closed sets of labels L in $\mathcal{R}$, such that $maxCliques(G, h(L))$ satisfies $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$ and $\mathbb{C}^{sep}$, and then for these minimal L, output $maxCliques(G, h(L))$.

Constraint properties

Definition 4. *(Set of vertices and set of labels satisfying the constraints) A set of vertices V (resp. set of labels L), V (resp. L)* satisfies *the constraints $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$ or $\mathbb{C}^{sep}$ if and only if the collection $maxCliques(G, V)$ (resp. the collection $maxCliques(G, h(L))$) satisfies the same constraints.*

Let us consider these constraints and their properties of monotonicity (i.e., if A satisfies a constraint then all supersets of A satisfy it) and anti-monotonicity (i.e., if A satisfies a constraint then each subset of A satisfies it). The following properties are straightforward.

- $\mathbb{C}^{lab}_{\alpha}$ is monotonic (resp. anti-monotonic) w.r.t. the sets of labels (resp. the sets of vertices);

- $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ is anti-monotonic (resp. monotonic) w.r.t. the sets of labels (resp. the sets of vertices).

Additionally, the conjunction $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ can be expressed in a relaxed form by means of the constraint $\mathbb{C}^{vert}$ defined as follows: $\mathbb{C}^{vert}$ holds for M if $|vertices(M)| \geq \beta + \kappa - 1$ (i.e., in order to contain at least κ separated cliques of size at least β, M must be built over at least $\beta + \kappa - 1$ vertices). The anti-monotonicity (resp. monotonicity) w.r.t. the sets of labels (resp. the sets of vertices) of $\mathbb{C}^{vert}$ are obvious as well. Since checking $\mathbb{C}^{vert}$ is easy and do not require to know the maximal cliques, we can use it first, and then check $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ only when $\mathbb{C}^{vert}$ is satisfied.

Algorithm and implementation

So, to extract the MHCS, we can choose between enumerating sets of vertices or sets of labels, while pushing the constraints when applicable. In the targeted application domains, as shown in Section 3, $\mathcal{L}$ is likely to be smaller than $\mathcal{V}$, so in the rest of the paper we consider the enumeration of sets of labels in $\mathcal{R}$. Notice that, in the context of extracting the so-called *formal concepts* in gene expression data, enumerating the sets of labels (representing in this context *set of biological situations*) has been shown to be an interesting approach [26, 31], and turns out to be feasible in practice, even when approaches based on the enumeration of sets of genes was not. However, an experimental comparison remains to be done, to confirm the interest of this strategy for MHCS.

To compute the MHCS through the enumeration of closed sets of labels, we can easily reuse most depth-first or levelwise closed set mining algorithms, handling $\mathbb{C}^{vert}$ as a standard *support* constraint (anti-monotonic), $\mathbb{C}^{lab}_{\alpha}$ as a monotonic constraints, and $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ as another anti-monotonic constraint. In our current implementation we use an algorithm similar to Closet [29]. The conjunction $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ is *pushed partially*, by pruning the current branch if it is not satisfied. We do not push the monotonic $\mathbb{C}^{lab}_{\alpha}$ constraint, that is only checked in a passive way, but it could be used actively in future developments using, for instance, the ExAnte data reduction technique [5].

Since we want only the minimal closed sets of labels satisfying the constraints (the ones corresponding to the maximal HCS), when a closed set L that fulfills the requirements is found then the current branch is pruned, and L and $maxCliques(G, h(L))$ are stored. And finally, when the exploration terminates, to guarantee the minimally of the closed sets, a test is performed over these sets in a post-processing step, and $maxCliques(G, h(L))$ is output for each minimal set L.

To extract the maximal cliques (function *maxCliques*), we implemented the algorithm of [34] that have an *optimal* worst-case time complexity.

3 Experiments

We evaluate our algorithm on two real-world datasets. On the first one, a social network dataset, we report a qualitative study and quantitative results. On the second one, a biological dataset, we present a detailed qualitative interpretation of a relevant pattern. In these experiments, we found original patterns with strong added value which would not be found with usual local pattern mining tasks. Quantitative experiments show that the algorithm scales up on large datasets. All experiments were performed on a PC running GNU/Linux with a 3 GHz Core 2 Duo CPU and 8 GB of main memory installed (no more than 700 MB used by the extraction process). Our experimental setting aims at answering the following questions: Does our pattern definition provide new kind of knowledge and make sense? Can domain experts easily exploit the collection of patterns? Does our approach scale well on large datasets?

3.1 Mining DBLP data

The social network dataset is built from the public DBLP database. This database contains a rather exhaustive bibliographic information from most of computer science conferences and journals. It has been extensively used as an experimental dataset by many researchers. Notice also that scientific collaboration networks have similar properties than social networks [24]. Our dataset is built using all conferences since 2000 included. A vertex represents an author and an edge between two authors means that they have coauthored at least *two papers*[6]. A vertex is labeled with the conference names in which the corresponding author has published (e.g., ICDM, KDD).

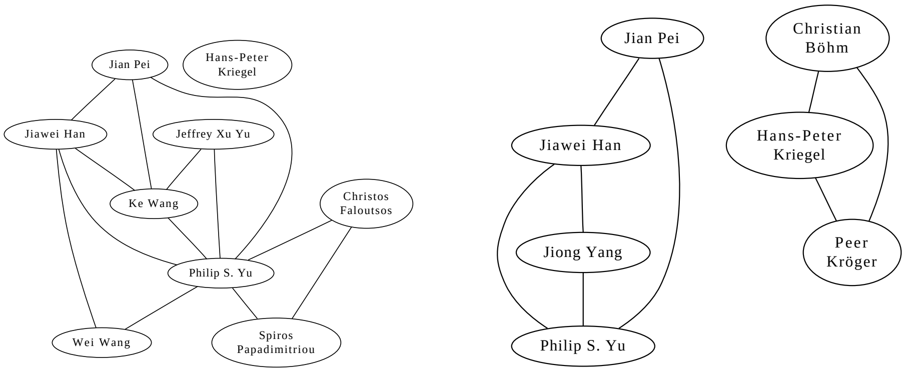
In the first experiment, for an author, in its label list, we retain only the conferences in which she/he has published at least twice (to discard one-shot involvements). Authors with an empty remaining label list are removed. The resulting graph has 117,526 vertices (authors), 467,691 edges (coauthor relationship of at least two papers) and there are 3,257 different labels (conferences). In this dataset, we search for MHCSs with at least 3 cliques of 3 vertices and 6 common labels ($\alpha = 6$, $\beta = 3$ and $\kappa = 3$). 80 patterns respect those constraints. Among them, 32 patterns are related to at least one of the following data-mining conferences: ICDM, KDD or SDM. We focus on two of these patterns, presented Figure 3.1. The pattern in Figure 1(a) contains 5 cliques:
{Jian Pei, Jiawei Han, Ke Wang, Philip S. Yu}, {Jeffrey Xu Yu, Ke Wang, Philip S. Yu}, {Christos Faloutsos, Philip S. Yu, Spiros Papadimitriou}, {Jiawei Han, Philip S. Yu, Wei Wang}, {Hans-Peter Kriegel}.

The first one contains four authors: Jiawei Han, Jian Pei, Philip S. Yu and Ke Wang. It is well known that these prolific authors have been working together over the past decade. Furthermore, we can see that the vertex corresponding to Philip S. Yu has a particular role, looking like a *hub* for 4 of the cliques. Discovering such local hubs in a subnetwork can be important and useful.

The other pattern, presented Figure 1(b), contains 3 cliques:
{Jian Pei, Jiawei Han, Philip S. Yu}, {Jiawei Han, Jiong Yang, Philip S. Yu}, {Christian Böhm, Hans-Peter Kriegel, Peer Kröger}

[6] We do not set an edge between two authors that have coauthored only one paper, since we think that this cannot be interpreted as a real collaboration.

The third clique is not connected to the two others, Christian Böhm, Hans-Peter Kriegel and Peer Kröger are all working in the same university located in Germany, whereas the two other groups are formed by people located in North America (working in the same universities at some time). This kind of local structure is particularly interesting since it exhibits groups that are not known to interact, but that share similar interests.

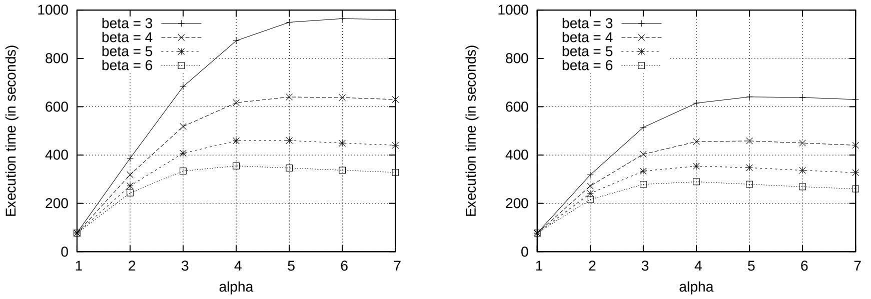


(a) 5 cliques concerning conferences ICDE, ICDM, KDD, PAKDD, SDM, and ACM SIGMOD.

(b) 3 cliques concerning conferences EDBT, ICDE, ICDM, SDM, ACM SIGMOD, and SSDBM

Fig. 1. Two patterns extracted with $\alpha = 6$, $\beta = 3$, and $\kappa = 3$

Now, we report the CPU times and the numbers of patterns obtained in a series of quantitative experiments with different settings for α, β and κ. In these experiments, in order to get a larger dataset, for an author we retain all conferences where she/he has published, and no author is removed. The resulting dataset contains 479,067 vertices, 773,613 edges and 3,607 different labels.

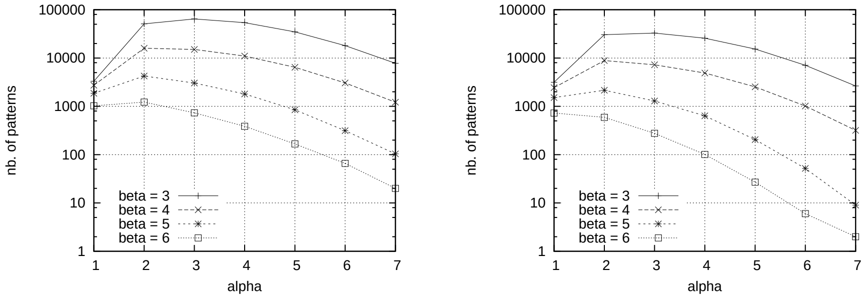


(a) $\kappa = 2$

(b) $\kappa = 3$

Fig. 2. Evolution of CPU time w.r.t. α, β and κ

Regarding main memory usage, we consider the maximal memory usage during each extraction. This maximal value never exceed 700 MB, and is about 657 MB on average over all extractions with a standard deviation of 7 MB. Concerning time performances, Figures 2(a) and 2(b) show that the extractions can be made in reasonable time, even when constraints are weakly selective. The worst case is obtained for $\alpha = 6$, $\beta = 3$, and $\kappa = 2$ and requires less than 17 minutes. As the algorithm starts by enumerating all closed sets having at least α labels, time performances depend mostly on α for small values of α. When α increases, pruning the search space thanks to $\mathbb{C}^{vert}$ and $\mathbb{C}^{clique}_{\kappa,\beta}$ is more effective and thus β and κ have more impact on time performances. Regarding the number of output patterns, Figures 3(a) and 3(b) show that it shrinks fast when parameter values increase. For $\alpha > 1$, when β increases by two, the number of patterns decreases by at least one order of magnitude.



(a) $\kappa = 2$

(b) $\kappa = 3$

Fig. 3. Evolution of the number of patterns (log scale) w.r.t. α, β and κ

3.2 Mining biological data

In this experiment, we built a dataset by using two databases: STRING [14] and SQUAT [19]. STRING aggregates data on protein-protein interactions from different sources (i.e., genomic data, co-expression, literature). Genes encode the proteins, so such interaction data can be read as gene-gene interactions (interactions of the proteins that are encoded by the genes). Among these interactions, we only retained interactions with a *confidence*[7] higher or equal to 400 (default STRING selection threshold). SQUAT is a Boolean (discretization over 0/1) gene expression database containing results from SAGE experiments (the discretization process is explained in [2]). SQUAT was created to support post-genomic data analysis processes for several species, and contains, for thousands of genes, the sets of biological situations (termed *libraries*) where these genes are overexpressed. In our experiments, we used only *Human species* genes. SQUAT contains information about only a subset of STRING genes, and thus we removed from STRING the proteins encoded by genes that do not correspond to genes in SQUAT (using *HUGO names* for the mapping).

So, to sum up, from STRING we obtained a graph where vertices represent genes and edges represent interactions between these genes (interactions

[7] This confidence is a measure provided by STRING. Low confidence means that there are not so many evidences that the interaction exists.

between the proteins encoded by these genes), and using SQUAT we associated to each vertex a set of labels representing the set of biological situations (SAGE libraries) where this gene is overexpressed. The resulting dataset contained 4,988 vertices (genes), 70,126 edges (interactions) and 486 different labels (biological situations).

In [13] and [35], the authors argue that combining gene expression and protein interactions leads to promising results. The main interest of our MHCS approach when mining such a combination of information is the following. Since the cliques do not need to be interconnected, it should enable to extract MHCS containing groups of genes that are functionally unrelated except for proteins making some *bridges*, for example, a transcription factor that activates genes that fall in different functional categories, a pattern that we indeed could extract (see Figure 4). This pattern was extracted using the following parameters: $\alpha = 3$, $\beta = 3$, and $\kappa = 2$.

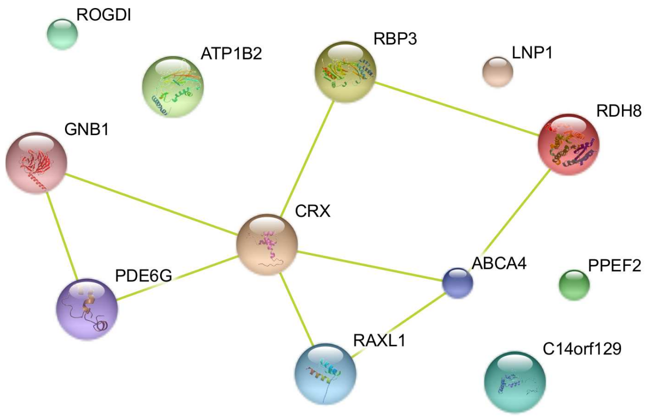


Fig. 4. STRING interaction graph of the genes forming a MHCS with 2 cliques of 3 genes overexpressed in 3 situations

First, we have been studying the relevancy of the pattern using the L2L tool [23]. It is immediately apparent that the best p-values was obtained for the “Visual perception” category of the GO biological processes with a highly significant score ($p = 6.17 \cdot 10^{-11}$). This was due to the six following genes products that are related to eye development and vision: PDE6G, PPEF2, ABCA4, RBP3, RDH8, and CRX. One should note that RAXL1 also harbor retina-related functions (see below) that were not detected by L2L. We then investigated the nature of the three libraries (i.e., labels) found within the pattern. It turned out that all three libraries were made from normal retina[8], which is perfectly rele-

[8] The three libraries are:
SAGE_Retina_Peripheral_normal_B_4Peri

vant given the nature of the extracted genes. We observed that the CRX gene behaved as a hub in the pattern, being the more densely connected vertex. We therefore investigated the nature of CRX. Using the hyperlink from SQUAT to Entrez Gene, we could find the following description of the function of CRX: *"The protein encoded by this gene is a photoreceptor-specific transcription factor which plays a role in the differentiation of photoreceptor cells"*. We then turned to examine the two cliques: (1) CRX, ABCA4, RAXL1 and (2) CRX, GNB1, PDE6G. Both cliques contain, and are associated by, the CRX gene product. This confirms its role as a hub, consistent with its transcriptional factor function. The first clique associates the Retinal-specific ATP-binding cassette transporter (ABCA4) as well as the Retina and anterior neural fold homeobox like 1 (RAXL1). RAXL1 encodes a transcription factor and ABCA4 encodes a membrane protein susceptible to transport retinal. This is therefore a clique centered upon the retinal functions of the proteins it harbors. The second clique associates Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta 1 (GNB1) and the Retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit gamma (PDE6G). The first gene encodes a G-protein and the second one encodes the effector molecule in G-protein-mediated phototransduction in vertebrate rods and cones. This clique appears as a quite homogeneous clique involved in signal transduction and under the control of the CRX transcription factor. So the final interpretation is that we have extracted information regarding the overexpression for a photoreceptor-specific transcription factor in retinal cells, together with 12 genes, most of them harboring known function in the retina. This motif is centered upon CRX, which act as a connection between the two cliques, one centered on G-protein-mediated signal transduction, the other unrelated. This demonstrates the relevancy and the actionability of patterns discovered by the MHCS extraction method.

4 Related work

Two kinds of approaches have been proposed to mine graphs whose vertices are described by set of labels. On the one hand, clustering methods were proposed in [12, 13, 35]. On the other hand, some proposals consider a local pattern discovery approach, generally in a constraint-based mining setting. In [21], the authors introduce the problem of cohesive pattern mining in feature vector graphs (each vertex is associated to a feature vector that represents properties of this vertex). Cohesive patterns are subgraphs that satisfy a subspace cohesion constraint (i.e., they share a large enough set of features), a density constraint, and a connectivity constraint. [20] extends cohesive pattern mining to quantitative features. These approaches do not enable to discover patterns that involve several subgraphs. It should be noticed that, in [3], properties are also associated to vertices of a time-evolving graph. The authors define a method to discover rules from

SAGE_Retina_normal_B_4cRet

SAGE_Retinal_Pigment_Epithelium_normal_B_4MacRPE local graph patterns that characterize the evolution of the graph. However, this approach was not defined to capture several subgraphs within the same pattern.

The problem of mining maximal homogeneous clique sets can be seen as a constraint-based pattern mining task on two data sources (e.g., graph data and transactional data). In this way, our approach is similar to [10] where the authors define a generic framework to extract patterns under a rich set of constraints. They exploit the cross-fertilization of data sources by mining micro-array data for patterns that must also satisfy some constraints on other datasets (e.g., a similarity matrix computed from textual information). However, this approach was not defined for graph data mining.

[15, 11] consider different problems on the same kind of datasets. [15] mines for cross-graph quasi-cliques, and in [11], the authors tackle the problem of redescription mining, that aims at finding subgroups having several descriptions. We think that such approaches are complementary to the one proposed in this paper.

A maximal homogeneous clique set can be seen as a set of patterns (cliques) which satisfy both local constraints and constraints that require to consider several local patterns. In [33], the authors propose the so-called *exception rules* that combine three local patterns (three different rules). Recently, several generic approaches - pattern teams [17], constraint-based pattern set mining [30], constraint programming for n-ary patterns [16] - aim at selecting patterns from the initial set of local patterns to return a smaller and more valuable set according to the context.

Since its introduction, research in pattern mining has aimed at discovering more valuable knowledge nuggets. From the simple frequency constraint, many primitives have been defined and several classes of constraints are now well understood. Pattern domains have become more sophisticated and meaningful. Recently, researchers have considered heterogenous and distributed data combined with domain knowledge to discover implicit relation between concepts from different domains, providing some novel insight into the problem domain [1, 4, 22]. Our work can be seen as being at the frontier between this *bisociative* knowledge discovery and constraint-based pattern mining.

5 Conclusion

In this paper, we considered the combined mining of a graph and of a binary relation associating sets of labels to the vertices. We proposed to search for patterns called maximal homogeneous clique sets, that are sets of cliques such that all vertices in a pattern shared a large enough set of labels. We described how the selection criteria on these patterns can be used as constraints in a complete extraction method. Finally, we reported experiments, showing that these extractions can be made on real datasets, and can lead to meaningful patterns.

Acknowledgments. This work is partly funded by the French contract ANR-07-MDCO-014 Bingo2 (Knowledge Discovery For and By Inductive Queries).

References

1. BISON: Bisociation networks for creative information discovery (http://www.bisonet.eu/)
2. Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.F., Gandrillon, O.: Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biology 3(12), 1–16 (2002)
3. Berlingerio, M., Bonchi, F., Bringmann, B., Gionis, A.: Mining graph evolution rules. In: European Conf. on Machine Learning and Princ. and Pract. of Knowl. Disc. in Databases (ECML/PKDD). pp. 115–130 (2009)
4. Berthold, M.R., Dill, F., Kötter, T., Thiel, K.: Supporting creativity: Towards associative discovery of new insights. In: Pacific-Asia Conf. on Knowl. Discov. and Data Mining (PAKDD). pp. 14–25 (2008)
5. Bonchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: Exante: Anticipated data reduction in constrained patterns mining. In: In Proc. of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. pp. 59–70 (2003)
6. Borgwardt, K.M., Kriegel, H.P., Wackersreuther, P.: Pattern mining in frequent dynamic subgraphs. In: Int. Conf. on Data Mining (ICDM). pp. 818–822 (2006)
7. Bringmann, B., Nijssen, S.: What is frequent in a single graph? In: Pacific-Asia Conf. on Knowl. Discov. and Data Mining (PAKDD). pp. 858–863 (2008)
8. Calders, T., Ramon, J., Dyck, D.V.: Anti-monotonic overlap-graph support measures. In: Int. Conf. on Data Mining (ICDM). pp. 73–82 (2008)
9. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Comput. Survey 38(1) (2006)
10. Crémilleux, B., Soulet, A., Klema, J., Hébert, C., Gandrillon, O.: Discovering knowledge from local patterns in sage data. In: Data Mining and Medical Knowledge Management: Cases and Applications, pp. 251–267. IGI Publishing (2009)
11. Gallo, A., Miettinen, P., Mannila, H.: Finding subgroups having several descriptions: Algorithms for redescription mining. In: SIAM Data Mining Conf. (SDM). pp. 334–345 (2008)
12. Ge, R., Ester, M., Gao, B.J., Hu, Z., Bhattacharya, B., Ben-Moshe, B.: Joint cluster analysis of attribute data and relationship data: The connected k-center problem, algorithms and applications. ACM Trans. Knowl. Discov. Data (TKDD) 2(2), 1–35 (2008)
13. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. In: Int. Conf. on Intelligent Systems for Molecular Biology (ISMB). pp. 145–154 (2002)
14. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C.: STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research 37, 412–416 (2009)
15. Jiang, D., Pei, J.: Mining frequent cross-graph quasi-cliques. ACM Trans. Knowl. Discov. Data (TKDD) 2(4), 1–42 (2009)
16. Khiari, M., Boizumault, P., Crémilleux, B.: Constraint programming for mining n-ary patterns. In: Principles and Practices of Constraint Programming (CP) (2010)
17. Knobbe, A.J., Ho, E.K.Y.: Pattern teams. In: Princ. and Pract. of Knowl. Disc. in Databases (PKDD). pp. 577–584 (2006)
18. Kuramochi, M., Karypis, G.: Finding frequent patterns in a large sparse graph. Data Min. Knowl. Discov. (DMKD) 11(3), 243–271 (2005)

19. Leyritz, J., Schicklin, S., Blachon, S., Keime, C., Robardet, C., Boulicaut, J.F., Besson, J., Pensa, R.G., Gandrillon, O.: Squat: A web tool to mine human, murine and avian sage data. BMC Bioinformatics 9(1), 378 (2008)
20. Miyoshi, Y., Ozaki, T., Ohkawa, T.: Frequent pattern discovery from a single graph with quantitative itemsets. In: Int. Conf. on Data Mining (ICDM) Workshops. pp. 527–532 (2009)
21. Moser, F., Colak, R., Rafiey, A., Ester, M.: Mining Cohesive Patterns from Graphs with Feature Vectors. In: SIAM Data Mining Conf. (SDM). pp. 593–604 (2009)
22. Mozetic, I., Lavrac, N., Podpecan, V., Novak, P.K., Motain, H., Petek, M., Gruden, K., Toivonen, H., Kulovesi, K.: Bisociative knowledge discovery for microarray data analysis. In: Int. Conf. on Computational Creativity. pp. 190–199 (2010)
23. Newman, J.C., Weiner, A.M.: L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biology 6(9), 81 (2005)
24. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review 69(2) (2004)
25. Nijssen, S., Kok, J.N.: Frequent graph mining and its application to molecular databases. In: Systems, Man and Cybernetics (SMC). vol. 5, pp. 4571–4577 (2004)
26. Pan, F., Cong, G., Tung, A.K.H., Yang, J., Zaki, M.J.: Carpenter: finding closed patterns in long biological datasets. In: Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. pp. 637–642. Washington (USA) (August 2003)
27. Papadopoulos, A.N., Lyritsis, A., Manolopoulos, Y.: Skygraph: an algorithm for important subgraph discovery in relational graphs. Data Min. Knowl. Discov. (DMKD) 17(1), 57–76 (2008)
28. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient mining of association rules using closed itemset lattices. Information Systems 24(1), 25–46 (1999)
29. Pei, J., Han, J., Mao, R.: Closet: An efficient algorithm for mining frequent closed itemsets. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 21–30 (2000)
30. Raedt, L.D., Zimmermann, A.: Constraint-based pattern set mining. In: SIAM Data Mining Conf. (SDM) (2007)
31. Rioult, F., Boulicaut, J.F., Crémilleux, B., Besson, J.: Using transposition for pattern discovery from microarray data. In: 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 73–79. San Diego (USA) (June 2003)
32. Robardet, C.: Constraint-based pattern mining in dynamic graphs. In: Int. Conf. on Data Mining (ICDM). pp. 950–955 (2009)
33. Suzuki, E.: Undirected exception rule discovery as local pattern detection. In: Local Pattern Detection. pp. 207–216 (2004)
34. Tomita, E., Tanaka, A., Takahashi, H.: The worst-case time complexity for generating all maximal cliques and computational experiments. Theor. Comput. Sci. (TCS) 363, 28–42 (October 2006)
35. Ulitsky, I., Shamir, R.: Identification of functional modules using network topology and high-throughput data. BMC Systems Biology 1(1) (2007)
36. Wang, J., Zeng, Z., Zhou, L.: Clan: An algorithm for mining closed cliques from large dense graph databases. In: Int. Conf. on Data Engineering (ICDE). p. 73 (2006)
37. Wasserman, S., Faust, K.: Social network analysis: methods and applications. Cambridge University Press (1994)
38. Yan, X., Han, J.: gSpan: Graph-Based Substructure Pattern Mining. In: Int. Conf. on Data Mining (ICDM). pp. 721–724 (2002)

39. Zaki, M.J., Ogihara, M.: Theoretical foundations of association rules. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 1–8 (1998)