Formal Concept Analysis of Attributed Networks

Henry Soldano, Guillaume Santini, and Dominique Bouthinon

1 Introduction

Our purpose is to investigate and analyze attributed graphs. In this article we discuss how recent extensions of Formal Concept Analysis apply to this problem. We consider undirected graphs $G(O, E)$ where E is the edge set, and O the vertex set. The vertices are labeled by a description in an attribute pattern language L with a lattice structure, typically $L = 2^I$ where I is a set of binary attributes. Note that we may consider such an attributed graph both as a graph whose vertices are labeled with subsets of I and as set of objects each described by such a subset of I and that may be related together by edges.

The former view leads to consider the methodology used to investigate graphs, in particular social and complex networks. Most of the work in this area consider unlabeled networks and is concerned by what may be said about the topological structure of the network. A large set of measures have been proposed to analyze these networks, and two main ways have been proposed to extract interesting subgraphs. The first way consider the network as made of a *core*, i.e., a dense subgraph whose vertices are highly connected, together with its *periphery*, made of vertices highly connected to the core, but poorly connected between them [6].

H. Soldano (✉)
Université Paris 13, Sorbonne Paris Cité, L.I.P.N UMR-CNRS 7030, F-93430, Villetaneuse, France

Museum National d'Histoire Naturelle, ISYEB - UMR 7205 CNRS MNHN UPMC EPHE, F-75005, Paris, France
e-mail: henry.soldano@lipn.univ-paris13.fr

G. Santini • D. Bouthinon
Université Paris 13, Sorbonne Paris Cité, L.I.P.N UMR-CNRS 7030, F-93430, Villetaneuse, France
e-mail: guillaume.santini@lipn.univ-paris13.fr; dominique.bouthinon@lipn.univ-paris13.fr

R. Missaoui et al. (eds.), *Formal Concept Analysis of Social Networks*,
Lecture Notes in Social Networks, DOI 10.1007/978-3-319-64167-6_6

The second way considers the network as made of a number of dense subnetworks, called *communities* whose vertices are highly connected within the community and poorly connected to vertices of other communities [8]. Finally, the two views may be combined, for instance, by considering the network as made of communities each having some core/periphery structure [16].

Regarding the notion of core, there have been various ways to define it, starting from the k-core of a network which is the greatest subnetwork whose vertices all have degree at least k in the subnetwork [17]. By changing the topological property, but keeping this idea of a greatest subnetwork whose vertices share the property within the network, we obtain various core definitions [3]. A core may also be defined as the greatest subnetwork made of a subset of a family of small, connected subnetworks. The simplest example is the k-clique core that is only made of k-cliques. When $k = 3$, the core is made of triangles which are known to be an important substructure in social networks analysis [29].

Concerning the idea of communities, it has been extensively investigated mostly as an optimization problem: how to optimally partition the network in subnetworks maximizing some measure. A second view of communities derives from some strong structural property that has to be satisfied within a community. We will further call them *structural communities*, or simply communities, as we only consider these kind of communities in the remaining part of this article. The main example is the k-community approach that divides a k-clique core (as defined above) in connected subnetworks each satisfying a stronger property (see below) [13] .

From the first point of view, adding attributes to the vertices means that each attribute pattern induces a subgraph whose vertices satisfy the pattern. Each such subgraph could then be investigated, extracting its core and communities. The question is then how to summarize and select relevant information from such a set of results.

The second view considers the attributed graph first as a table representing a set of objects described by attributes, and then considers that edges may relate objects. This leads to the use of standard methodology of data analysis by adapting it to dealing with topological information. The whole purpose of this article is to discuss how Formal Concept Analysis, which was originally concerned with data tables, may be extended in order to take into account the topological information. The main idea we propose here is to consider as parameters the notion of cores and structural communities relevant to the data to analyze and adapt accordingly the FCA methodology.

Regarding the reduction of a graph to its core, this may be obtained by defining an interior operator p on the vertex powerset 2^O. This approach, based on a previous work on abstraction in Formal Concept Analysis [24], produces *abstract closed patterns* structured in an *abstract concept lattice* together with a basis of *abstract implications* written $\Box q \rightarrow \Box w$. All concept extents are then images of the interior operator. When considering attributed networks, and given some core definition, such an interior operator reduces a vertex subset e to the vertices forming the core of the subgraph $G(e)$ induced by e [23]. It is then called a *graph abstraction* operator. Fig. 1 represents an attributed graph G, the subgraph induced by pattern a (in plain

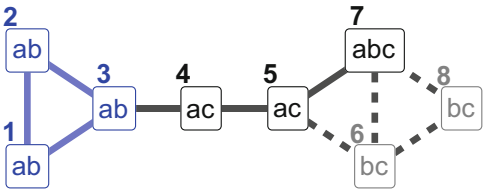

Fig. 1 The pattern *a* subgraph is displayed with *plain lines*, the corresponding 3-clique abstract subgraph is displayed in *blue lines*. The associated abstract closed pattern is *ab*

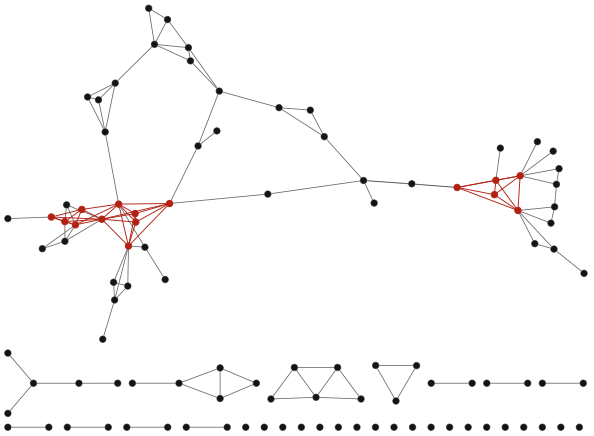

Fig. 2 The DMKD,IDArev pattern subgraph in the DBLP co-authoring experiment. The *red vertices* and *edges* represent the subgraph induced by the degree ≥ 4 abstract extension

lines), and the 3-clique core of this subgraph, i.e., its *abstract subgraph* (in blue lines). Note that all vertices of the core share attribute *b*. This means that *ab* is an abstract closed pattern and that the abstract implication $\Box a \rightarrow \Box ab$ holds, i.e., any triangle in *G* whose vertices share *a* also share *b*.

Recent works in attributed graph mining are interested in searching for local patterns made of a constraint on a subset of attributes together with a density constraint on a vertex subset, and this using various notions of maximality [11, 18]. In a companion article [21], we have defined *local closed patterns* corresponding to maximal attribute patterns each associated with one dense subgraph, allowing to extract *local implications*, particular to specific dense groups of objects. For that purpose Formal Concept Analysis (FCA) had to be extended in order to take into account this notion of locality.

The simplest example is obtained by considering a subgraph made of various connected components, and associating to each connected component a local closed pattern, i.e., the most specific pattern shared by the vertices of this connected component. More generally, local closed patterns may be associated with the connected components of abstract subgraphs. The family of such connected components forms a partial order called a *cc-confluence* while the corresponding *local concepts* have a weaker structure called a *pre-confluence*. As an example, in Fig. 2 we display a pattern subgraph extracted from a DBLP co-authoring network labeled by journal and conference names, together with its abstract subgraph (in bold and red vertices and lines) when considering a 4-core abstraction. The abstract subgraph has two connected components, i.e., two structural communities of scientists. Again we

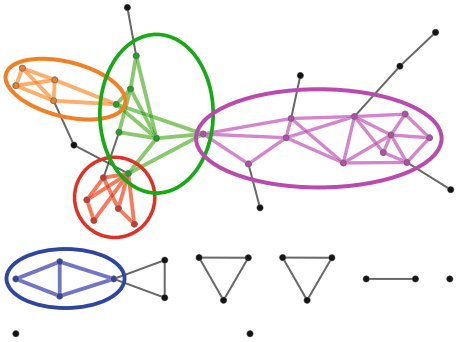

Fig. 3 The original Friendship network of a group of West Scotland pupils. The pupils and edges forming 3-communities of size at least 4 are displayed in various colors

may associate to this structure a set of implications, called *local implications*. In the previous example we found a local implication $\Box_i q \rightarrow \Box_i w$ stating that in the connected component containing vertex i of the degree ≥ 4 abstract subgraph of pattern q, all vertices also share pattern w.

Connected components of abstract subgraphs as represented in *cc*-confluences do not always completely capture the idea of communities as considered in social network analysis. As discussed in [21], we may, however, enlarge the local closed patterns approach by deriving a new graph G_T from G whose vertex set T is a set of vertex subsets of G. Figure 3 displays a graph whose vertices represent pupils of a school in the West of Scotland, whose edges represent friendship relations, and whose vertex attributes concern substance use and sporting activity.[1] As a running example we consider the subgraph induced by the empty pattern, i.e., the whole graph. By applying a 3-clique graph abstraction restricted to connected components containing at least 4 vertices,[2] we obtain a subgraph made of the bold and colored edges and vertices. This abstract subgraph is made of two connected components, therefore leading to two local concepts. However, the largest connected component is clearly made of distinct dense parts, i.e., *communities*, we would like to consider when defining local closed patterns. Fortunately, when considering k-communities [13] we can solve this problem by applying the cc-confluence approach to a new graph derived from the original graph. More precisely, a k-community is a vertex subset in a graph G that corresponds to a connected component in a derived graph G_T. The vertices of G_T are k-cliques in G and an edge relates two vertices whenever the corresponding k-cliques share $k-1$ vertices in G. Each colored subgraph in Fig. 3 defines such a 3-community.

The last task is to define interestingness measures to rank abstract or local patterns and implications. Regarding patterns, we will search for patterns whose abstract (resp. local) subgraph is a large part of the whole pattern subgraph, i.e., which preserves the core (resp. communities) definitions. The corresponding measure is called *specificity*. Regarding the abstract and local implications, we search for

[1] http://www.stats.ox.ac.uk/~snijders/siena/s50_data.htm.

[2] We further call the 3-*clique and cc-*≥ 4 abstraction.

implications which are informative, i.e., which did not hold as standard implications and therefore bring some new information about our data. The corresponding measure is called *Informativity*. A preliminary discussion of both measures was presented in [25]. We propose here definitions of *specificity* and *informativity* both at the abstract and local level and experiment them on two real attributed networks.

Section 2 describes the attributed graphs used in our experiments. Section 3 presents abstract concept lattices, abstract implications, and graph abstractions together with associated interestingness measures. Section 4 defines local concept pre-confluences, related local implications, *cc*-confluences, and interestingness measures. In Sect. 5 we show how we extract the set of 3-communities associated with pattern subgraphs by using derived *cc*-confluences, and we display the local concept ordering of the attributed network of teenage friendship displayed in Fig. 3. In Sect. 6 we briefly discuss the implementation used in our experiments.

2 Datasets

In this section we will consider experiments with two datasets. In both cases the data are described as a graph $G = (O, E)$. Vertices of this graph are have labels from 2^I, where I is a set of items, i.e., binary attributes. Since objects are not always described by binary attributes, the binarization preprocessing is described when necessary.

2.1 Teenage Friends and Lifestyle Study

The dataset is denoted by *s50-1* and is a standard attributed graph dataset.[3] It represents 148 friendship relations between 50 pupils of a school in the West of Scotland, and labels concern the substances used (tobacco, cannabis, and alcohol) and sporting activity. The values of the corresponding variables are ordered. The binarization process consists in defining variables representing the value intervals. T stands for Tobacco consumption and has values 1 (no smoking), 2 (occasional), and 3 (regular). C stands for cannabis consumption and has values 1 (never tries) to 4, D stands for alcohol consumption and has values 1 (does not drink) to 5, and S stands for sporting activity and has two values 1 (occasional) and (2) regular. Binary attributes represent intervals, for instance, C234 means that the value of C is at least 2 and therefore represents the interval $[2, 4]$. In Table 1 we present the binary attributes we have defined. Attribute subsets represent intersection of intervals. For example pattern {D123, D2345} requires that the value of D lies within the interval $[1, 3] \cap [2, 5] = [2, 3]$. Note that, for the sake of simplicity we do not distinguish the two highest values of attributes T, C, and D.

[3] http://www.stats.ox.ac.uk/~snijders/siena/s50_data.htm.

Table 1 The binary attributes used to label the vertices in the Teenage Friendship network

Tobacco	Cannabis	Alcohol	Sport
T1, T23	C1, C12, C234, C34	D1, D12, D123, D2345, D345, D45	S1, S2

2.2 A DBLP Dataset

This is the DBLP dataset as described in [4]. There are 45,131 vertices, 228,188 edges, and 555 connected components. Vertices are authors that have published at least one paper in one among 29 journals or conferences of the Database and Datamining communities[4] during the 1/1990–2/2011 period. An edge links two authors whenever they are coauthors of at least one article. The conferences are clustered in three clusters: DB (databases), DM (data mining), and AI (artificial intelligence) according to a conference ranking site categorization.[5] The binary attributes are the journal and conference names together with the three clusters. An attribute has value 1 if the author has published in the corresponding journal or conference or cluster.

3 Abstract Closed Patterns in Attributed Networks

3.1 Closed Patterns

In this section we introduce the necessary definitions and terminology we use in the article. Note that the terminology is somewhat non-standard in FCA. Indeed, as we need to interleave interior operators with extensional and intensional operators, the standard X'' notation to represent closed elements is not so convenient, so we rather denote, respectively, by ext and int the extensional and intensional operators. Furthermore, in this introductory paragraph we relate FCA to closed pattern mining.

A standard pattern mining procedure consists in considering the set of occurrences of patterns, belonging to some pattern language L with a lattice structure,[6] within an object set O (see, for instance, [5]). This language is partially ordered following a general-to-specific ordering, and each object o is described as a

[4]Conferences: KDD, ICDM, ECML/PKDD, PAKDD, SIAM DM, AAAI, ICML, IJCAI, IDA, DASFAA, VLDB, CIKM, SIGMOD, PODS, ICDE, EDBT, ICDT, SAC ; Journals: IEEE TKDE, DAMI, IEEE Int. Sys., SIGKDD Exp., Comm. ACM, IDA J., KAIS, SADM, PVLDB, VLDB J., ACM TKDD.

[5]http://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html.. DB = {VLDB, SIGMOD, PODS, ICDE, ICDT, EDBT, DASFAA, CIKM}; DM= {SIGKDD Explorations, ICDM, PAKDD, ECML/PKDD, SDM}; AI= {IJCAI, AAAI, ICML, ECML/PKDD}.

[6]We recall that in a lattice any pair of elements (x, y) has a greatest lower bound $x \wedge y$ (or *meet*) and a least upper bound (or *join*) $x \vee y$.

particular pattern $d(o)$. A pattern q occurs in object o whenever $d(o)$ is more specific (i.e., larger) than q. The set of occurrences $\text{ext}(q)$ of a pattern q is called its *extension*. An intension function $\text{int}(e)$ returns the most specific pattern associated with the extension e. This means that we relate a pattern q to the most specific pattern with same extension by applying the *closure operator* $\text{int} \circ \text{ext}$ to q. $\text{int} \circ \text{ext}(q)$ is then called a closed pattern. The pattern language L typically is 2^I where I is a set of binary attributes (aka items). With no loss of generality we will further use the powerset 2^I as a pattern language while what follows also applies to wider languages as pattern structures[10]. When $L = 2^I$ the closure operator on patterns then simply intersects the object descriptions of the extension of the entry pattern. This means that when considering patterns with same extension as equivalent, closed patterns are the representatives of the equivalences classes. Such a class has therefore a maximum but also minimal elements, called *minimal generators*. When the patterns belong to 2^I, the min–max basis of implications[14] is defined as follows:

$m = \{g \rightarrow f \backslash g \mid f \text{ is a closed pattern}, g \text{ is a generator}, f \neq g, \text{ext}(g) = \text{ext}(f)\}$

This basis represents all the implications $t \rightarrow t'$ that hold on O, i.e., such that $\text{ext}(t) \subseteq \text{ext}(t')$. This precisely means that all these implications may be derived from the min–max basis. Obviously all non-trivial implications, i.e., implications such that $t' \not\subseteq t$, may be inferred from an implication $l \rightarrow r$ of m where $l \subseteq t$ and $r \cup l \supseteq t'$.

Finally, note that the enumeration of closed patterns is in general restricted to frequent patterns, i.e., patterns whose extension is larger than some threshold. In FCA, such a constraint leads to iceberg lattices [27].

3.2 Abstract Closed Patterns

We summarize here how abstraction is applied in FCA by constraining the extensional space. We first recall the definitions of closure operators and *interior operators*, the latter being further used to restrict the pattern extensions to be *abstract extensions*. In what follows all ordered sets are finite, and in particular any topped meet-semilattice (resp. pointed join-semilattice) is a lattice.

Definition 1 Let U be an ordered set and $f : U \rightarrow U$ a self map such that for any $x, y \in U$, f is monotone, i.e., $x \leq y$ implies $f(x) \leq f(y)$ and idempotent, i.e., $f(f(x)) = f(x)$, then:

- If f is extensive, i.e., $f(x) \geq x$, f is called a closure operator
- If f is intensive, i.e., $f(x) \leq x$, f is called a dual closure operator, an interior operator, or also a projection.

In the first case, an element such that $x = f(x)$ is called a closed element.

Ranges of interior operators on lattices are called *abstractions* and are characterized by the following Proposition:

Proposition 1 (see [24]) *A subset A of $X = 2^O$ is the range $p[X]$ of some interior operator p on X, if and only if for any elements x, y in A, their join $x \cup y$ also belongs to A and A contains the empty set. The interior operator is related to its range as follows:*

$$p(x) = \sup_{\{a \in A | a \subseteq x\}} a.$$

Let then p be the interior operator associated with some abstraction A, $p(x)$ be the greatest element of A contained in x. Closed pattern analysis has been recently extended to *abstract* closed pattern analysis by noticing that applying an interior operator on the extensional space 2^O we obtain again a closure operator on the pattern language 2^I [15, 24]:

Proposition 2 *Let $X = 2^O$ and $L = 2^I$, p be an interior operator on 2^O, and $A = p[X]$ be the associated abstraction, we have that $(int, p \circ ext)$ is a Galois connection on (A, L), i.e.,:*

$f = \text{int} \circ p \circ \text{ext}$ is a closure operator on L,

The *abstract extension* of pattern q is defined as $p \circ \text{ext}(q)$. A new equivalence relation is then defined such that $q \equiv_A w$ whenever $p \circ \text{ext}(q) = p \circ \text{ext}(w)$, each equivalence class of which corresponds to some abstract extension in A. There is then a unique *abstract support closed* pattern, i.e., a most specific pattern among all patterns sharing the same abstract extension, which is obtained as $f(q) = \text{int} \circ p \circ \text{ext}(q)$. $f(q)$ is then called an *abstract closed pattern*. This leads to the definition of abstract concepts organized in a concept lattice:

Corollary 1 ([24]) *The set of (abstract extension, abstract closed pattern) pairs $(e = ext(c), c = int(e))$, ordered following A, is a lattice called an abstract concept lattice.*

Note that, as p is monotone, whenever $\text{ext}(q) \subseteq \text{ext}(w)$, i.e., $q \rightarrow w$ is valid, we also have $\text{ext}_A(p) = p \circ \text{ext}(q) \subseteq \text{ext}_A(w) = p \circ \text{ext}(w)$. The latter inclusion states the validity of an *abstract implication* we will rewrite as $\Box^A q \rightarrow \Box^A w$.

This way we obtain *abstract min–max basis* with the same definition as earlier in this section except that ext_A replaces ext and therefore abstract implications relate minimal elements (i.e., A-generators) to maximal element (the abstract closed pattern, or A-closed pattern) of the same abstract equivalence class. We have then the following definition:

Definition 2 The *abstract min–max basis* m_A of valid abstract implications is defined as

$m_A = \{\Box^A g \rightarrow \Box^A f \backslash g \mid f$ is an A-closed pattern, g is a A-generator, $f \neq g$, $\text{ext}_A(g) = \text{ext}_A(f)\}$.

In the same way as in the standard min–max basis case, all implications $\Box^A t \rightarrow \Box^A t'$ that hold on A, i.e., such that $\text{ext}_A(t) \subseteq \text{ext}_A(t')$, may be inferred from the abstract min–max basis.

3.3 Graph Abstractions

These ideas have been applied to attributed graphs by defining graph abstractions [23]. The set of objects O is then the set of vertices of a graph $G = (O, E)$, and each vertex o is labeled by an attribute pattern $d(o) \in 2^I$.

A graph abstraction is an abstraction of 2^Odefined through a characteristic property P such that $P(x, e)$ expresses some minimal connectivity requirement of the vertex x within the subgraph G_e induced by some vertex subset e.

Proposition 3 *Let P be such that*

- $P(x, e)$ *implies* $x \in e$ *and*
- $e \subseteq e'$ *and* $P(x, e)$ *implies* $P(x, e')$,

and let q be a mapping defined by $q(e) = \{x \in e | P(x, e)\}$, *then the mapping p defined by* $p(e) = \text{fixedpoint}(q, e)$ *is an interior operator on* 2^O.

Consider a subgraph G_e, where $p(e)$ represents the greatest vertex subset of e inducing a subgraph whose vertices all satisfy the associated characteristic property. This subgraph $G_{p(e)}$ will be further called the *abstract subgraph* of G_e. We give hereunder examples of graph abstractions, defined through their characteristic property and exemplified in Fig. 4.

1. degree $\geq k$. The degree $\geq k$-abstract subgraph of a graph is its k-core [17].
2. k-club $\geq s$: x has to belong to at least one k-club of size at least s in G_e. This is a relaxation of the notion of clique[1]: a k-club is a subset c of vertices such that there is a path of length $\leq k$ between any pair of vertices in G_c. A triangle, a 3-clique, is a 1-club of size 3 (Fig. 4a). Figure 4b represents a 2-club of size 6 and therefore a 2-club$\geq$ 6 abstract group.
3. nearStar(k, d): x has to have degree at least k or there must be a path of length at most d between x and some y with degree at least k. For instance, the simplest nearStar$(8, 1)$ abstract group is a central node connected with eight nodes. Such an abstraction is useful when we want the abstraction to preserve hubs [2] (i.e., high degree vertices) together with their (low degree) neighbors (see Fig. 4c).
4. $cc \geq s$: x has to belong to a connected component of size at least s in G_e (see Fig. 4d).

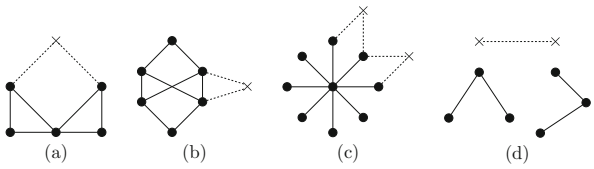

Fig. 4 Graph abstractions corresponding to various vertex characteristic properties. In each graph *plain circles* and *plain lines* form the abstract subgraph, *crosses* and *dotted lines* represent the vertices and edges out of the abstract subgraph. (**a**) x has to belong to a 3-clique, (**b**) x has to belong to a 2-club of size at least 6, (**c**) x has to be connected to a vertex y such that the degree of y is at least 6, i.e., to a nearstar(6,1), (**d**) x has to belong to a connected component whose size is at least 3

Finally, it is interesting to note that we can combine two (or more) abstractions A_1 and A_2 in two ways, defining a new composite abstraction either stronger or weaker than both A_1 and A_2. For instance, we may want to consider an abstract subgraph where vertices both have a degree larger than some k and belong to a connected component exceeding a minimal size s. On the contrary, we may want an abstract subgraph such that at least one of the two characteristic properties is satisfied by all the vertices. This would be the case, for instance, if we want to keep both vertices that have a degree larger than, say 10, and vertices in a star, i.e., connected to a hub which degree is at least 50. The following proposition states that we can freely combine abstractions in both directions.

Proposition 4 *Let P_1 and P_2 two characteristic properties of abstractions defined on the same object set O, and let $P_1 \wedge P_2$ and $P_1 \vee P_2$ be defined as follows:*

- $P_1 \wedge P_2(x, e) = P_1(x, e) \wedge P_2(x, e)$
- $P_1 \vee P_2(x, e) = P_1(x, e) \vee P_2(x, e)$

Both $P_1 \wedge P_2$ and $P_1 \vee P_2$ are characteristic properties of abstractions.

3.4 Interestingness Measures on Abstract Patterns and Implications

3.4.1 Specificity of Abstract Patterns

We are now interested in measuring knowledge brought by abstract closed patterns and abstract implications [25]. For that purpose we first generalize hereunder the *structural correlation* measure introduced by A. Silva and co-authors [19], originally introduced to relate a subgraph to its content in terms of quasi-cliques and rename it as *specificity*.

Definition 3 Let q be a pattern, A an abstraction of some powerset of objects O, the specificity of q with respect to A is defined as:

$$S_A(q) = \frac{|\ \mathrm{ext}_A(q)\ |}{|\ \mathrm{ext}(q)\ |}$$

Consider, for instance, a 3-clique abstraction. Whenever $S_A(q)$ is close to 1, the pattern q subgraph is mainly made of triangles. To the contrary, whenever $S_A(q)$ is close to 0, the pattern q subgraph almost displays no triangles, which means quite isolated vertices. We relate this way a pattern q to the measure of how selecting vertices satisfying this pattern preserves the topological property associated with the abstraction.

Example 1 Figure 1 displays a graph each vertex of which is described by an itemset. We observe then that:

- $\text{ext}(a) = e = \{1, 2, 3, 4, 5, 7\}$ induces the subgraph $G(e)$ (blue+black).
- $\text{ext}_A(a) = \{1, 2, 3\}$ as $4, 5, 7$ do not belong to any 3-clique in $G(e)$.
- $\text{int} \circ \text{ext}_A(a) = ab \cap ab \cap ab = ab$ is an abstract closed pattern.
- $S_A(a) = 1/2, S_A(ab) = 3/4$.

Note that among the patterns of some equivalence class of $\equiv_A$ the abstract closed pattern c has maximal specificity:

For any t, if $ext_A(t) = ext_A(c)$ then $S_A(c) \geq S_A(t)$

3.4.2 Informativity of Abstract Implications

Apart from measuring through specificity what is specific to the pattern in its abstract view, we are also interested when considering abstract implications in how informative they are. For that purpose we consider abstract implications whose left and right patterns are equivalent in the abstract space A, i.e., have same abstract extension, as in the min–max abstract implication basis defined above. Whenever these patterns are also equivalent in the original space 2^O intuitively the implication is uninformative. Assume, for instance, that $a \rightarrow abc$ is valid, then validity of the abstract implication with same left and right members does not bring any new information. On the contrary, assume that $\square^A a \rightarrow \square^A abc$ is valid while $a \rightarrow abc$ has only confidence 0.5, i.e., $\text{ext}(abc) = 0.5 * \text{ext}(a)$, then clearly the abstract implication brings some information.

Definition 4 Let q be a pattern, A an abstraction of 2^O, the *informativity* of the valid implication $r : \square^A q \rightarrow \square^A w$ is defined as:

$$I_A(r) = 1 - \frac{| \text{ext}(qw) |}{| \text{ext}(q) |}$$

Informativity has a range between 0 and 1 and estimates the probability of not having w whenever we have q in graph G. This quantity has value 0 whenever $q \rightarrow w$ holds and has limit 1 whenever $| \text{ext}(qw) |$ approaches 0, i.e., restricting the extension of patterns to elements of A concentrates the extension of q to the very few sharing also w.

Intuitively, the informativity of an abstract implication measures what we discovered when we observed that q and qw share the same abstract support.

Example 2 Following Example 1 illustrated in Fig. 1 consider the abstract implication $r : \square a \rightarrow \square ab$. This abstract implication has the following semantics: "a 3-clique of G whose vertices share pattern a also share pattern b," and its informativity is therefore $I_A(r) = 1 - 1/2 = 0.5$.

Note that implications of the abstract min–max basis which relate minimal elements g of an abstract equivalence class to the corresponding abstract closed pattern c have maximal informativity:

Let $g \equiv_A t \equiv_A t' \equiv_A c$ then $I_A(\square^A g \rightarrow \square^A c) \geq I_A(\square^A t \rightarrow \square^A t')$.

Table 2 Top-15 abstract closed patterns in the Teenage Friendship network ranked according to their 3-clique and cc≥4 specificity

N°	$\square^A$c	$\lvert ext^A(c)\rvert$	$\lvert ext(c)\rvert$	$S_A(c)$
1	$\square^A$D45-C34	5	7	0.714
2	$\square^A$C12	27	42	0.643
3	$\square^A\emptyset$	32	50	0.64
4	$\square^A$C12-T1	21	36	0.583
5	$\square^A$D45	9	17	0.529
6	$\square^A$D345	15	29	0.517
7	$\square^A$C1-T1	17	33	0.515
8	$\square^A$D123-C12-T1	15	30	0.5
9	$\square^A$D345-C12	10	21	0.476
10	$\square^A$D2345	21	45	0.467
11	$\square^A$C12-S2	15	33	0.455
12	$\square^A$D45-C12-S2	4	9	0.444
13	$\square^A$D2345-C12	16	37	0.432
14	$\square^A$S2	16	37	0.432
15	$\square^A$D123-C1-T1	11	27	0.407

3.5 *Experiments*

Some experiments on the two datasets described in Sect. 2 have been performed and discussed in [23]. We discuss hereunder new experiments in particular regarding the interestingness measures.

We firs consider the Teenage Friendship network s50. Among the 50 pupils 38 belong to triangles (i.e., 3-cliques). As there are no isolated triangles, the abstract subgraph when considering only connected components with size at least 4 is reduced to 32 pupils. The corresponding abstraction is therefore *3-clique and* $cc \geq 4$.

Table 2 displays the top 15 patterns according to the corresponding specificity. We observe that specificity is clearly non-monotonic with respect to abstract extension size and that among top patterns we find both small (and therefore general) and large patterns. We further discuss the pattern with highest specificity D45-C34, which corresponds to pupils with high alcohol and cannabis consumption, in Sect. 5 where we search for communities.

Table 3 displays the top 15 abstract implications according to the 3-*clique and* $cc \geq 4$ abstract informativity. Again, informativity is clearly nonmonotonic with respect to extension size. The first and third implications have the same abstract pattern as their rightmost member: it concerns the same abstract subgraph, whose pupils have in common the behavior D45-C12-S2, but is obtained either by reducing the pattern D345-S2 subgraph or the pattern D45-S2 subgraph. Obviously the former subgraph corresponds to a higher informativity as it includes the latter subgraph. As a matter of fact, the third implication is *redundant* with the first one and could be removed with no information loss. This leads to reduce the abstract min–max basis by eliminating all such redundant rules. We will apply such an idea

Table 3 Top-15 abstract implications in the Teenage Friendship network ranked according to their 3-clique and cc≥4 informativity

N°	$\Box^A$g → $\Box^A$c	$\lvert ext(c)\rvert$	$\lvert ext(g)\rvert$	$I_A(r)$
1	$\Box^A$D345-S2 → $\Box^A$D45-C12-S2	9	22	0.591
2	$\Box^A$C234 → $\Box^A$D45-C34	7	14	0.5
3	$\Box^A$D45-S2 → $\Box^A$D45-C12-S2	9	13	0.308
4	$\Box^A$T1-D345 → $\Box^A$D345-C1-T1	14	18	0.222
5	$\Box^A$D23 → $\Box^A$D23-C1-T1	23	28	0.179
6	$\Box^A$C1-D345 → $\Box^A$D345-C1-T1	14	17	0.176
7	$\Box^A$C34 → $\Box^A$D45-C34	7	8	0.125
8	$\Box^A$D2345-S2 → $\Box^A$D2345-C12-S2	29	33	0.121
9	$\Box^A$T1-D2345 → $\Box^A$D2345-C1-T1	29	33	0.121
10	$\Box^A$C1-D2345-S2 → $\Box^A$D2345-C1-S2-T1	22	25	0.12
11	$\Box^A$C1-S2- → $\Box^A$C1-S2-T1	25	28	0.107
12	$\Box^A$C12-D45 → $\Box^A$D45-C12-S2	9	10	0.1
13	$\Box^A$D12 → $\Box^A$D12-C1-T1	19	21	0.0952
14	$\Box^A$C1-D2345 → $\Box^A$D2345-C1-T1	29	32	0.0938
15	$\Box^A$D123 → $\Box^A$D123-C12-T1	30	33	0.0909

when defining a basis for local implications in Sect. 4.2. The second implication in this ranking concerns pattern D45-C34, mentioned as the highest specificity pattern. The implication states that the 3-*clique and cc* ≥ 4 subgraph of pupils that have pattern C234, which corresponds to a medium-to-high cannabis consumption behavior, selects pupils with the D45-C34 pattern, i.e., those who have both high alcohol and cannabis consumption behavior. When applying no abstraction, the implication only holds on 7 among the 14 pupils that have pattern C234, which results in informativity $1 - 7/14 = 0.5$.

We discuss now some new details on experiments performed on the DBLP dataset. The experiment consisted in applying a degree $\geq k$ abstraction with increasing k-values and we focused in abstract patterns obtained with $k = 16$, which corresponds to a very strong abstraction: in an abstract extension each author is required to have 16 co-authors within the abstract extension. We obtained few abstract closed patterns and in particular the abstract closed pattern VLDBJ, ICDE, SIGMOD, VLDB and the related abstract implication $\Box$ VLDBJ → $\Box$ ICDE, SIGMOD, VLDB. Both the abstract closed pattern and its abstract minimal generator VLDBJ have an abstract extension of 38 among the 1276 VLDBJ authors in the dataset. The implication states that a dense group of co-authors that have published in the Very Large Database Journal also have published in several database conferences. In Fig. 5 we present the corresponding subgraph. Such a very dense co-authoring subgraph within the VLDBJ subgraph is somewhat unexpected. Its abstract specificity $38/1276 \approx 0.085$ is low, but still higher than the 0 value we could expect from such a high abstraction level. The abstract implication has a high informativity about ≈ 0.65 coming from the fact that among the 1276 authors who published in VLDBJ journal only 441 did publish in all the conferences ICDE, SIGMOD, and VLDB.

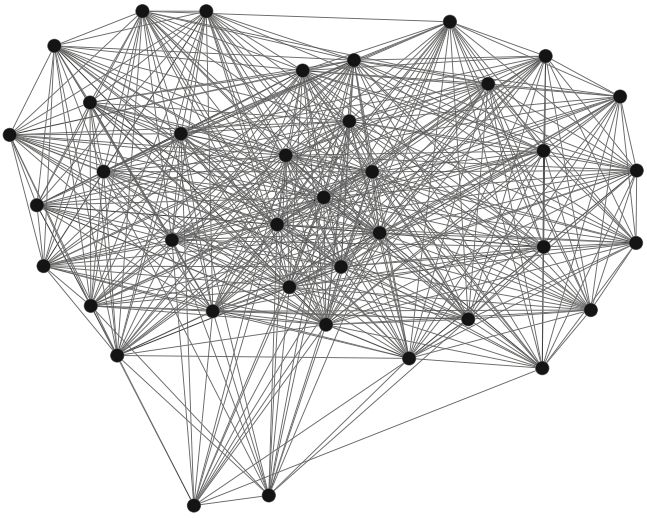

Fig. 5 The subgraph obtained when applying the degree ≥ 16 abstraction to the VLDBJ subgraph in the DBLP co-authoring experiment

[j56] Serge Abiteboul, Rakesh Agrawal, Philip A. Bernstein, Michael J. Carey, Stefano Ceri, W. Bruce Croft, David J. DeWitt, Michael J. Franklin, Hector Garcia-Molina, Dieter Gawlick, Jim Gray, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, Martin L. Kersten, Michael J. Pazzani, Michael Lesk, David Maier, Jerey F. Naughton, Hans-Jörg Schek , Timos K. Sellis, Avi Silberschatz, Michael Stonebraker, Richard T. Snodgrass, Jeffrey D. Ullman, Gerhard Weikum, Jennifer Widom, Stanley B. Zdonik: **The Lowell database research self-assessment**.Commun. ACM 48(5): 111-118 (2005)

Fig. 6 An example of reference with many authors that leads to a high degree subnetwork

We made then some investigations in the DBLP repository, focussing of the 38 authors of the abstract extension, and found an article whose reference is given in Fig. 6 and whose abstract begins as follows:

> A group of senior database researchers gathers every few years to assess the state of database research ...

In some sense the explanation of the pattern we discovered is straightforward. However, the whole purpose of pattern mining is to find unexpected patterns, hidden within large datasets, and interpret them in order to acquire some new knowledge. It is exactly what happens here: we were not aware of these regular meetings of senior database researchers, and we learned something new, though, of course, this knowledge is clearly widely known within the database community.

When considering a weaker abstraction, namely here a degree ≥ 4 abstraction, we obtain more abstract closed patterns sometimes made of several connected

components. Figure 2 in Sect. 1 represents the DMKD, IDArev pattern subgraph together with the subgraph induced by the abstract extension of the pattern. This abstract subgraph is made of two connected components, the one in the right part of the figure is made of ten vertices and we are then interested in knowing whether there is some more specific pattern than the abstract closed pattern DMKD, IDArev, which would be shared by this connected component. Answering such questions means mining at a local level the attributed graph, and this is the subject of the next section.

4 Local Closed Patterns in Attributed Networks

Given some attribute pattern, we are now interested in extracting *local support closed patterns*, i.e., maximal attribute patterns each associated with one dense subgraph, so allowing to extract *local implications* particular to specific dense groups of objects. Recently FCA has been extended to *local closed patterns*: they are obtained by applying a set of *local closure operators* [22]. In the graph case, this means that from the extension of some (closed) pattern c, various dense extensions, called *local extensions* are extracted each associated with a *local closed pattern*, i.e., the most specific pattern l common to the elements of the local extension. Again we obtain a set of *local implications* corresponding to inclusion of local extensions, but now such an implication is only valid in the vicinity of some dense group of vertices.

In [21] we introduced locality in the closure framework with the main motivation of investigating local patterns in attributed graphs. For that purpose we have first to define pre-confluences and confluences which are structures weaker than lattices that have been investigated in FCA [21, 23]. Confluences, in particular, are close to but different from confluent families as defined in [5]. We further denote by E^x the up sets $\{y \in E | y \geq x\}$ of an ordered set E, by E_x its down sets $\{y \in E | y \leq x\}$, and by $\min(E)$ the set of its minimal elements.

First note that ordered sets we consider are all finite. We define a pre-confluence as a finite ordered structure that generalizes the (finite) lattice structure:

Definition 5 A finite ordered set F is a pre-confluence if and only if for any $m \in \min(F)$, $F^m = \{x \in F \mid x \geq m\}$ is a lattice.

A consequence of this definition is that a (finite) lattice is a pre-confluence with a minimum. The structure has a partial join operator:

Proposition 5 *For any $m \in \min(F)$ and any $x, y \in F^m$ their least upper bound is the least element of $F^x \cap F^y$ we further denote by $x \vee_F y$.*

This means that a pre-confluence is a union of lattices in which joins coincide. A particular case is which of a pre-confluence included in a host lattice and which is join preserving:

Definition 6 Let T be a lattice and $F \subseteq T$ be a pre-confluence with as join $\vee_F = \vee_T$, F is called a confluence of T.

An abstraction of T, as defined above is a confluence of T with $\perp_T$ as minimum. We have then the following property when considering 2^O as the host lattice:

Proposition 6 *Let $X = 2^O$ be a lattice, $F \subseteq X$ is a confluence of X if and only if for any $x, y \in F^m$ with $m \in \min(F)$, we have that $x \cup y$ belongs to F.*

A confluence is then associated with a set of interior operators:

Proposition 7 *Let F be a confluence of a lattice X, and $m \in \min(F)$,*

- *$p_m : X^m \rightarrow X^m$, such that $p_m(x) = \vee_{q \in F^m \cap X_x} q$, is an interior operator and $p_m[X^m] = F^m$.*

We are concerned here with extensional confluences, i.e., confluences of $X = 2^O$ [21] that generalize extensional abstractions as graph abstractions. In this case, let x be an element of X greater than or equal to some minimal element m of F, then $p_m(x)$ returns the greatest subset of x in F that includes m. We now define graph confluences, which are the original motivation for defining confluences:

Definition 7 Let $G = (O, E)$ be a graph, and F be the family of vertex subsets inducing connected subgraphs of G. F is a confluence of 2^O called the graph confluence of G.

Proof By definition, any singleton $\{s\}$ induces a connected subgraph of G. Furthermore, the union of two connected vertex subsets that each includes a given vertex singleton $\{s\}$ also is a connected vertex subset. Following Proposition 6, F is then a confluence of 2^O.

The elements of F are simply called the *connected vertex subsets* of O. By abuse of notation we write p_s and F^s rather than $p_{\{s\}}$ and $F^{\{s\}}$. The interior operator p_s projects then any vertex subset e containing vertex s on the connected component of the subgraph G_e induced by e that contains s. The up set F^s is then the set of connected vertex subsets containing s, and the union of all these F^s represents the whole set of connected subgraphs of G.

Example 3 Let $G = (O, E)$ be a graph (displayed at the bottom of Fig. 7) whose vertex set is $O = \{1, 2, 3, 4\}$. Let $F \subseteq 2^O$ be the set of connected vertex subsets of G. F is a confluence whose set of minimal elements is $\min(F) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$. The subset $F^{1+3} = F^1 \cup F^3$ representing connected vertex subsets containing vertices 1 or 3 is also a confluence. Figure 7 displays the diagram of F^{1+3}. □

The extension e of a pattern q may then be projected through interior operators on various smaller *local extensions* $\{e_i\}$ corresponding to the connected components of the pattern subgraph. These interior operators are associated with *local closure operators* [21]:

Proposition 8 *Let F be a confluence of $X = 2^O$, m a minimal element of F, and $L_{int(m)}$ be the down set of the pattern lattice L whose elements q are such that $q \geq \text{int}(m)$, then*

$f_m = \text{int} \circ p_m \circ \text{ext}$ is a closure operator on $L_{\text{int}(m)}$

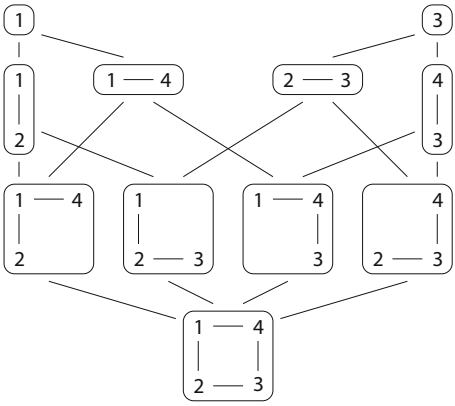

Fig. 7 A square graph (in the *bottom* of the figure) and the Hasse diagram of the confluence F^{1+3} of connected vertex subsets that contain vertices 1 or 3

In a graph confluence, let $e = \text{ext}(q)$, then $p_s(e)$ is the connected component of the pattern subgraph G_e to which the vertex s belongs. Obviously, $p_s(e) = p_v(e)$ for any vertex v in the same connected component. Therefore $f_s(q)$ is a *local closed pattern* w.r.t. any vertex in this connected component, i.e., the most specific pattern shared by the vertices in the connected component.

Now a general result is that the set of local extensions is a pre-confluence:

Theorem 1 *The mapping $h : F \to F : h(e) = p_m \circ \text{ext} \circ \text{int}(e)$ for $m \leq e$ is a closure operator on F and $E = h[F]$ is a pre-confluence.*

$h(e)$ is therefore the local extension of $\text{int}(e)$ that contains $m \leq e$ and $h[F]$ is a pre-confluence isomorphic to the set P of *local concept pairs* defined as follows:

Definition 8 The set of local concept pairs $P = \{(e, l) \mid e = p_m \circ \text{ext}(l), l = \text{int}(e), m \leq e\}$ is called a *local concept pre-confluence*.

To summarize we have defined local concepts as (local extension, local closed patterns) pairs and we have shown that they are organized in a structure with possibly several minimal elements, therefore generalizing the concept lattice definition. In the graph confluence exemplified above the local extensions simply are the connected components of the pattern subgraphs. We will now extend graph confluences by intersecting graph confluences with abstractions.

4.1 Cc-Confluences

We remark now that we can freely intersect confluences:

Proposition 9 *Let F_1 and F_2 be confluences of X, then $F_1 \cap F_2$ is a confluence.*

Since abstractions of X are confluences of X with the bottom element of X as their unique minimal element, the above proposition means we can freely intersect abstractions with confluences to build smaller confluences. Many confluences can then be derived from a graph confluence by intersecting it with some abstractions. We call this family of confluences the *cc-confluences*.

Definition 9 Let F be the graph confluence of some graph G and A be a graph abstraction of G, then the confluence $F \cap A$ is called the cc-confluence of G associated with A.

For instance, considering A as the k-clique abstraction, we obtain the *cc*-confluence of connected subgraphs of G made of k-cliques. Note that *cc*-confluences have an important property: rather than considering the minimal elements m of F when defining local closure operators we can consider vertices as in the graph confluence F. This is because, given any abstract subgraph and any m included in its vertex set, all vertices v of m belong to the same connected component and therefore to the same local extension. This is computationally important as this means that when considering local extensions we only need to consider each vertex in the extension and associate it to the connected component to which it belongs.

4.2 Local Implications

Inclusion of local extensions defines validity of local implications $\Box^A_m q \rightarrow \Box^A_m w$, where m is a minimal element of F_A in the extension of q. Note that, as the local extension of pattern q is obtained by applying an interior operator, which is monotone, to the support set of q, we have that, whenever $\Box^A q \rightarrow \Box^A w$ is valid and $m \subseteq \text{ext}_A(w)$, we also have that $\Box^A_m q \rightarrow \Box^A_m w$ is valid, i.e., we may infer the latter local implication from the former abstract implication.

Example 4 Consider the graph displayed in Fig. 8. The 3-clique *cc*-confluence has as minimal elements $\{123, 567, 678\}$ and rewrites as $F_A = \{123, 567, 678, 5678\}$. The extension of pattern b is equal to its abstract extension 123, 678, and the abstract closed pattern is also b. However, the corresponding abstract subgraph displays two connected components 123 and 678. The vertices of the latter share bc which is consequently its local closed pattern. This leads to a local implication:

- $\Box^A_{678} b \rightarrow \Box^A_{678} bc$

In a cc-confluence the local implication may be indexed with respect to any vertex of the corresponding connected component: a triangle in the same connected

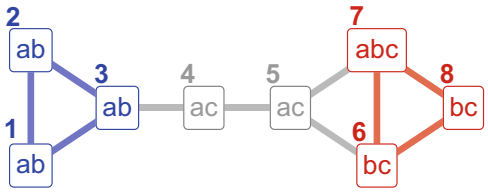

Fig. 8 The pattern b 3-clique abstract subgraph displays two connected components. The *blue one*, on the *left*, is also the pattern b abstract subgraph of motif a leading to the local implication $\Box^A_{123} a \rightarrow \Box^A_{123} ab$

component as 6, when considering the pattern b abstract subgraph, also has c, which rewrites as $\square_6^A b \rightarrow \square_6^A bc$. We will simply say that "a triangle containing 6 and which has b also has c."

We search now for a basis B_{F_A} of valid local implications from which we may infer any local implications. We will consider a basis B_m for a given minimal element m of F and obtain the whole basis $B_{F_A} = \cup B_m$ by joining these bases. Consider a given abstract closed pattern c whose abstract extension has a connected component that contains m, and let $l = f_m(c)$ be the corresponding local closed pattern, with respect to the cc-confluence F_A. This means that the implication $\square_m c \rightarrow \square_m l$ holds. We select then a basis B_m of *informative* ($l \neq c$) and *irredundant* (there is no other implication $\square_m c' \rightarrow \square_m l$ with c' less specific than c in the implication set) ones. From B_m we may infer all local implications associated with m by applying standard axioms in the same way as in the case of the *min–max basis* in the standard closed or abstract framework. The basis $B_{F_A} = \cup B_m$ represents the local knowledge deriving from the reduction of the extensional space from abstraction A to cc-confluence F_A:

Definition 10 The Local Min–max Basis B_{F_A} associated with the cc-confluence F_A is defined as:
$B_{F_A} = \{\square_m^A c \rightarrow \square_m^A l \mid$ where c A-closed , l locally closed , $c \neq l, \text{ext}_m^A(c) = \text{ext}_m^A(l)$,
and for all $c' \subset c$ we have $\text{ext}_m^A(c') \neq \text{ext}_m^A(c)\}$

4.3 Interestingness Measures on Local Patterns and Implications

As in the abstract case, we may measure novelty brought locally [25]. We first extend the specificity measure to local patterns. If the ratio of a local extension to the abstract extension is high this means that the corresponding connected component is the largest part of the abstract subgraph.

Definition 11 Let q be a pattern, F be a cc-confluence and $m \in F$ be such that $m \subseteq \text{ext}_A(q)$, the specificity of q near m is defined by

$$S_F(q, m) = \frac{| \text{ext}_m^A(q) |}{| \text{ext}_A(q) |}$$

We then define the informativity of a local implication by observing that in a valid local implication the left and right parts have same local extensions, while their abstract extensions are different. Therefore, as in the abstract knowledge case, we define the local informativity as the probability, at the abstract level, not to have the right part when the left one is true:

Definition 12 The local informativity of valid local implication $r : \Box_m^A q \rightarrow \Box_m^A qw$ is defined by

$$I_F(r) = 1 - \frac{|\ \mathrm{ext}_A(qw)\ |}{|\ \mathrm{ext}_A(q)\ |}$$

When local informativity is close to 1, it means that the probability to observe w when we have q at the abstract level is low: to be able to deduce w from q is specific of the graph region where m lies.

Example 5 We carry on Example 4 displayed in Fig. 8. The local specificity near 678 of pattern bc is $S_F(bc, 678) = 3 \div 3 = 1$: it does not appear elsewhere in the bc abstract subgraph. Informativity of local implication $\Box_{678}^A b \rightarrow \Box_{678}^A bc$ is $1 - (2 \div 6) = 0,5$: the abstract support of b is 6, but only three vertices share the local closed pattern bc. Regarding the local closed pattern ab, it also has local specificity 1 as ab is only found in the connected component 123, and it leads to a new local implication $\Box_{123}^A b \rightarrow \Box_{123}^A ab$ whose informativity is $0,5$. What we see here is that we obtain new knowledge regarding pattern b which depends on the region of the graph we consider.

Now, when considering the Teenage Friends attributed graph displayed in Fig. 3, clearly the friendship relations are organized in 3-cliques, therefore any stronger abstraction will be poorly informative. However, as mentioned in Sect. 1, when considering the 3-clique abstract graph associated with the empty pattern the unique connected component could be separated in several (overlapping) communities (displayed in Fig. 3 in various colors). We discuss and exemplify in the next section how to apply the local closure strategy to discover such sub-communities in an attributed graph.

5 Local Concepts from a Derived Graph

In what follows, we will consider a family $T \subseteq 2^O$ of vertex subsets, and consider T as the vertex set of a graph $G_T = (T, E_T)$ *derived* from G. The simple graph confluence F of 2^T is then the new extensional space and we will search for the corresponding local closed patterns. The local extensions are afterwards transformed into extensions in 2^O. Let $u : 2^T \rightarrow 2^O$ be such that $u(e_T) = \cup_{t \in e_T} t$. $u(e_T)$ is called the *flattening* of e_T. We consider then the two maps ext_T and $\mathrm{int_T}$ defined as follows:

- $\mathrm{ext}_T : L \rightarrow 2^T$ with $\mathrm{ext}_T(q) = \{t | t \subseteq \mathrm{ext}(q)\}$
- $\mathrm{int}_T : 2^T \rightarrow L$ with $\mathrm{int}_T(e_T) = \mathrm{int} \circ u(e_T)$

$\mathrm{ext}_T(q)$ represents the extension of q in T when considering that q occurs in t whenever q occurs in all elements of t (seen as a subset of O). Conversely $\mathrm{int}_T(e_T)$ represents the greatest pattern in L whose extension in T includes e_T, i.e., whose

extension in O contains, as subsets, the elements of e_T. Now, consider as T the family of k-cliques of G and that $(t_1, t_2) \in E_T$ whenever t_1 and t_2 share $k-1$ vertices in G. A k-community in G [13] is a vertex subset that results from the flattening (in the sense defined above) of some connected component of G_T. The local closed patterns w.r.t. F are then most specific patterns occurring in k-communities of pattern subgraphs of G. This way we obtain local concepts and associated local implications, whose local extensions are these k-communities. Note that, in the derived case, the local concepts do not form a pre-confluence: technically we obtain a pre-confluence of 2^T, but two different local extensions in 2^T may result in the same flattening, corresponding to one 3-community. As a consequence, the local concept order is no more a pre-confluence.

5.1 Experiments on a Derived Graph

Coming back to our Teenage Friendship attributed graph, we have applied this strategy and built the derived graph G_T, where T is the set of 3-cliques of the original attributed graph. In Figs. 9 and 10 we display the ordered set of 3-communities with

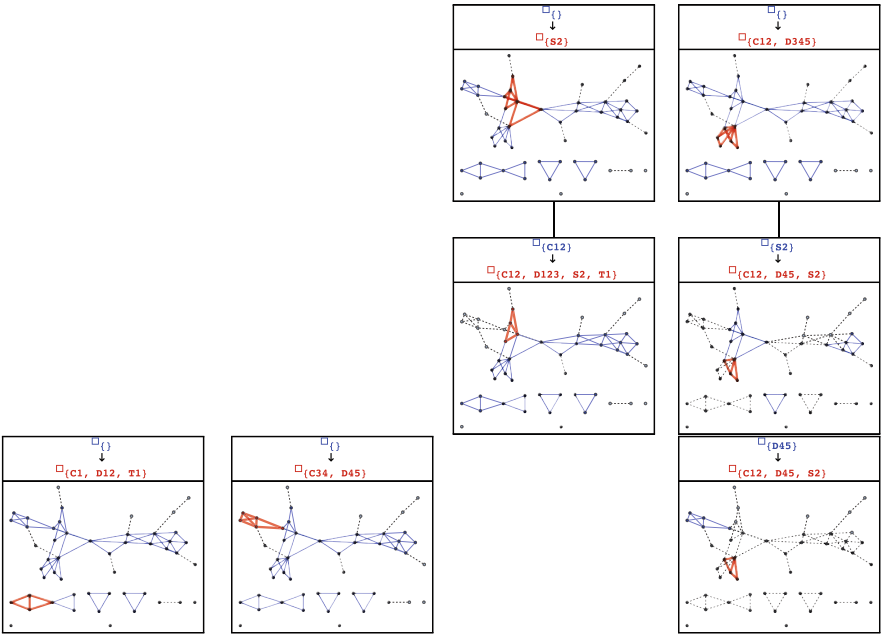

Fig. 9 The ordered set of size ≥ 4 3-communities of the Teenage Friendship network (part-I). The 3-communities are displayed in *red* and *bold lines* from the larger ones on the *top* to the smaller one on the *bottom*.The abstract subgraphs are displayed in *plain lines*. On the *right* at the *bottom* we have a 3-community displayed twice as it is built from two different abstract closed patterns

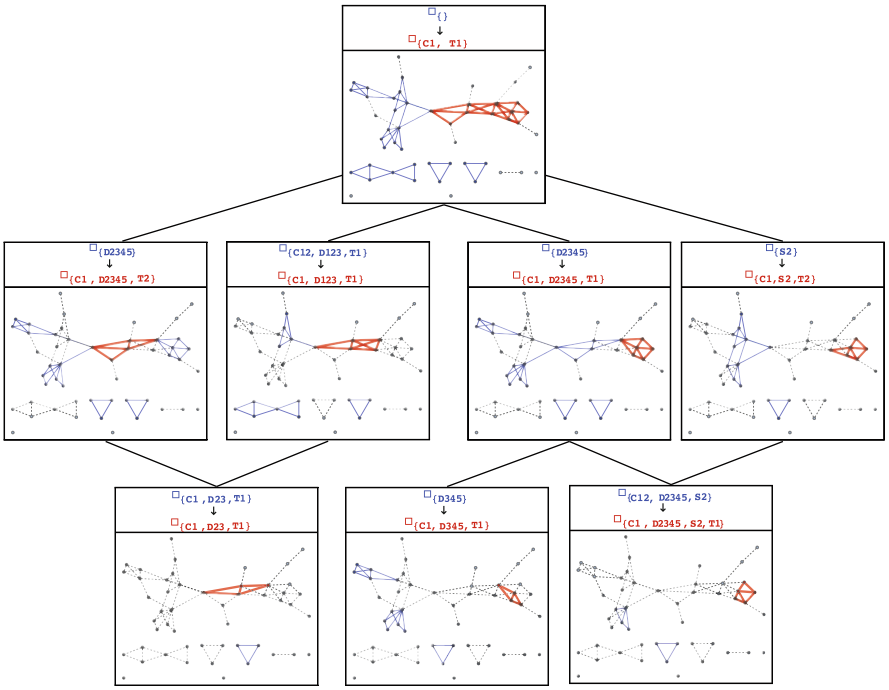

Fig. 10 The ordered set of size ≥ 4 3-communities of the Teenage Friendship network (part-II)

size at least 4.[7] The minimal 3-communities are the lowest ones on both figures. Each element of the pre-confluence represents a (3-community, local closed pattern) pair but may be associated with several non-redundant local implications. This happens for one 3-community displayed on the right at the bottom of Fig. 9 and associated with two local implications each represented in a square. Each square displays in red the 3-community, and in red+green+blue the abstract extension of the abstract closed pattern forming the left part of the implication. In Fig. 10 we have a unique maximal 3-community on the top, and a hierarchy of sub-communities.

We now investigate interestingness of local patterns and local implications. Note that the definitions given in Sect. 4.3 have to be adapted since we have to replace local extensions as defined in Sect. 4 by 3-communities, i.e., flattening of local extensions in the derived graph G_T. Table 4 displays the local closed patterns ranked according to their specificities. Each local closed pattern is indexed by the first triangle, in the lexicographic ordering, leading to the corresponding 3-community. Consider, for instance, the first two local patterns l_1 = D45-C12-S2 and l_2 = D45-C34 which have both specificity 1. This means that the set of pupils triangle having

[7] Formally, this means that we also apply an abstraction to the derived graph to avoid connected components corresponding to 3-communities smaller than four members.

Table 4 Top 15 local closed patterns ranked according to their specificity in the Teenage Friendship network

N°	$\square^A_m l$	$\lvert ext^A_m(l)\rvert$	$\lvert ext^A(l)\rvert$	$S_F(l,m)$
1	$\square^A_{25,31,32}$D2345-C1-S2-T1	5	5	1
2	$\square^A_{21,31,32}$C1-S2-T1	6	6	1
3	$\square^A_{17,19,24}$D23-C1-T1	4	4	1
4	$\square^A_{27,29,30}$D123-C12-S2-T1	4	4	1
5	$\square^A_{22,25,31}$D345-C1-T1	4	4	1
6	$\square^A_{10,11,15}$D45-C12-S2	4	4	1
7	$\square^A_{26,44,7}$D45-C34	5	5	1
9	$8\square^A_{40,45,46}$D12-C1-T1	4	6	0.667
10	$\square^A_{17,19,24}$C1-T1	11	17	0.647
11	$\square^A_{22,25,31}$D2345-C1-T1	6	10	0.6
12	$\square^A_{1,11,14}$D345-C12	6	10	0.6
13	$\square^A_{17,18,19}$D2345-C1-T1	5	10	0.5
14	$\square^A_{46,48,49}$D12-C1-T1	3	6	0.5
15	$\square^A_{17,19,24}$D123-C1-T1	5	11	0.455

D45-C12-S2 (respectively, D45-C34) forms a (unique) 3-community we further refer to as *Community 1* (respectively, *Community 2*). Communities 1 and 2 have in common high alcohol consumption behavior (D45), but differ in that the members of Community 1 do not smoke cannabis (C12) and have a regular sporting activity (S2), while the members of Community 2 have regular cannabis consumption (C34).

Consider now the implications of the local min–max basis, ranked according to their informativities, displayed in Table 5. We have here some implications with high informativity. As an example, the fifth (I_5) and ninth (I_9) local implications concern Community 1 while the second (I_2) concerns Community 2.

As an illustration we consider Community 1 and its two related local implications I_5 and I_9. As associated with the same community their rightmost member is the same local closed pattern D45-C12-S2. However, as they are extracted from different abstract subgraphs corresponding, respectively, to abstract closed patterns S2 and D45, they have different informativities. I_5 has informativity 0.75, while I_9 has informativity $\approx$ 0.56. In Fig. 11 we display Community 1 together with the necessary information to compute the local informativity of I_5.

The high informativity value of I_5 means that what the pupils in this community have in common, i.e., high alcohol consumption, no cannabis consumption and regular sporting activity (D45-C12-S2) is unfrequent among pupils in triangles with regular sporting activity outside of the community.

Table 5 Top 15 local implications of the min–max basis ranked according to their informativity int the Teenage Friendship network

N°	$\square^A_m$c $\rightarrow$ $\square^A_m$1	$\lvert ext^A(l)\rvert$	$\lvert ext^A(c)\rvert$	$I_F(r)$
1	$\square^A_{27,29,30}$C12 $\rightarrow$ $\square^A_{27,29,30}$D123-C12-S2-T1	4	27	0.852
2	$\square^A_{26,44,7}\emptyset \rightarrow \square^A_{26,44,7}$D45-C34	5	32	0.844
3	$\square^A_{46,48,49}\emptyset \rightarrow \square^A_{46,48,49}$D12-C1-T1	6	32	0.813
4	$\square^A_{40,45,46}\emptyset \rightarrow \square^A_{40,45,46}$D12-C1-T1	6	32	0.813
5	$\square^A_{10,11,15}$S2 $\rightarrow$ $\square^A_{10,11,15}$D45-C12-S2	4	16	0.75
6	$\square^A_{22,25,31}$D345 $\rightarrow$ $\square^A_{22,25,31}$D345-C1-T1	4	15	0.733
7	$\square^A_{1,11,14}\emptyset \rightarrow \square^A_{1,11,14}$D345-C12	10	32	0.688
8	$\square^A_{21,31,32}$S2 $\rightarrow$ $\square^A_{21,31,32}$C1-S2-T1	6	16	0.625
9	$\square^A_{10,11,15}$D45 $\rightarrow$ $\square^A_{10,11,15}$D45-C12-S2	4	9	0.556
10	$\square^A_{22,25,31}$D2345 $\rightarrow$ $\square^A_{22,25,31}$D2345-C1-T1	10	21	0.524
11	$\square^A_{17,18,19}$D2345 $\rightarrow$ $\square^A_{17,18,19}$D2345-C1-T1	10	21	0.524
12	$\square^A_{11,19,30}\emptyset \rightarrow \square^A_{11,19,30}$S2	16	32	0.5
13	$\square^A_{17,19,24}\emptyset \rightarrow \square^A_{17,19,24}$C1-T1	17	32	0.469
14	$\square^A_{11,19,30}$C12 $\rightarrow$ $\square^A_{11,19,30}$C12-S2	15	27	0.444
15	$\square^A_{25,31,32}$D2345-C12-S2 $\rightarrow$ $\square^A_{25,31,32}$D2345-C1-S2-T1	5	9	0.444

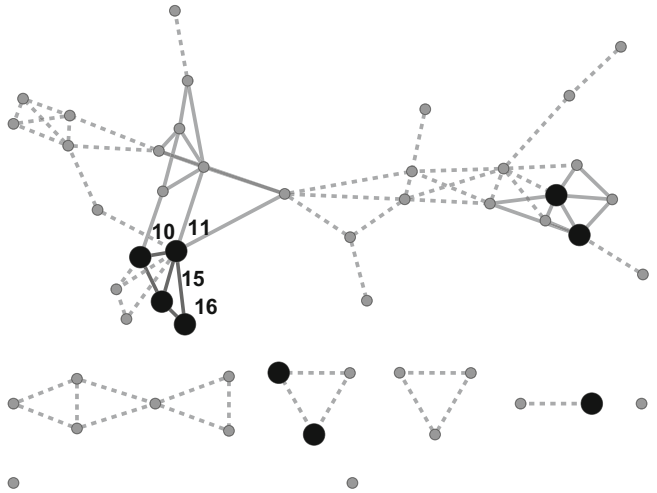

Fig. 11 Community 1 (represented as *bold dots* joined by *bold lines*) composed of five pupils with high alcohol consumption, no cannabis consumption C12 and regular sporting activity (D45-C12-S2). The community has size 4 and is one of the communities of the S2 abstract subgraph which is represented with *gray plain lines* joining 16 vertices. The Informativity of the local implication I_5 associated with community 1, which has S2 as its leftmost member and D45-C12-S2 as its rightmost member, is therefore $1 - 4/16 = 0.75$. Note also the *three black dots* representing vertices which have the S2 pattern but do not belong to its abstract subgraph

6 Implementation

In our experiments we first used the CORON software [28] to compute frequent closed patterns, according to some frequency threshold, then apply a set of PYTHON functions as a post-processing[8] to compute abstract and local patterns and implications. More recently we have implemented an efficient algorithm using a divide and conquer strategy similar to that proposed in [5] and implemented in [12]. This allows in particular to directly apply the frequency constraints at the abstract and local levels. A first version, named ParaminerLC, is experimented in [26] and was designed to handle the 3-communities local knowledge extraction problem. The selection of the implications belonging to the local min–max local basis is performed as a post-processing. Our current implementation in progress is a versatile program enumerating abstract and local frequent closed patterns and local implications with various definition of abstraction and locality.

7 Conclusion

In this article we have addressed problems in which the extensional space, made of the vertex subsets of an attributed network, is constrained according to connectivity properties. We have first considered abstract vertex subsets in which a constraint has to be satisfied by each vertex in the subgraph they induce, as, for instance, a minimum degree constraint. The extensional space is in this case a particular lattice called an abstraction. We have then shown, benefiting from previous work in FCA, how abstract support closed patterns, i.e., maximal patterns among those sharing the same abstract extension, could be obtained using a closure operator. This has resulted in defining a wide class of abstract concept lattices, whose elements are (abstract extension, abstract closed pattern) pairs, each corresponding to a particular abstraction. This way we obtain a global information on how the graph topology is related to the pattern extensions. We have then considered a way to extract local knowledge from an attributed network. For that purpose, using a recent extension of FCA to local extensional spaces, called confluences, we have related each pattern to various local extensions, corresponding to connected components in subgraphs induced by abstract vertex subsets. We obtain this way a set of local concepts, organized in a generalization of the lattice structure called a pre-confluence. Furthermore we have defined both abstract implications and local implications representing knowledge which is valid at the abstract and local levels, i.e., regarding the latter, in the vicinity of particular vertices. For both abstract and local patterns and implications we have proposed proper interestingness measures, namely specificity which measures to what extent the original extensions of patterns

[8]The corresponding software is to be found in https://lipn.univ-paris13.fr/~santini/data/ProjClos.tgz.

are preserved in abstract and local concepts, and informativity, which measures novelty brought by abstract and local implications. Finally we have applied these ideas to enumerate 3-communities in a network. These 3-communities are in fact sub-communities as each is a 3-community in some subnetwork induced by an attribute pattern.

Overall, what we propose here is a new way, brought by recent developments in Formal Concept Analysis, to explore social and complex networks as attributed graphs. As an application, we are currently involved in the ADALAB project which aims at helping the robot scientist EVE [30] to design experiments.[9] In this context, we use our methodology to explore a co-regulation network labeled with information regarding gene expression. Future works concerns, on the extensional side, applying these ideas to attributed directed graphs or multiplex networks. We also consider to use abstract and local extensional constraints while extending the pattern language to a wider class of pattern languages. First, as in [7, 9, 15] by building a meet-semilattice adapted to the mining problem and using interior operators to reduce it to a tractable language. This has been in particular successfully applied to graph mining [10]. Then, as in [5, 20] by considering confluent languages allowing to treat connectivity within the pattern language.

Acknowledgements This work was partially supported by CHIST-ERA grant (AdaLab, ANR 14-CHR2-0001-01).

References

1. Balasundaram, B., Butenko, S., Trukhanov, S.: Novel approaches for analyzing biological networks. J. Comb. Optim. **10**, 23–39 (2005)
2. Barabàsi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999). doi:10.1126/science.286.5439.509. http://www.sciencemag.org/content/286/5439/509.abstract
3. Batagelj, V., Zaversnik, M.: Fast algorithms for determining (generalized) core groups in social networks. Adv. Data Anal. Classif. **5**(2), 129–145 (2011). doi:10.1007/s11634-010-0079-y
4. Bechara Prado, A., Plantevit, M., Robardet, C., Boulicaut, J.F.: Mining graph topological patterns: finding co-variations among vertex descriptors. IEEE Trans. Knowl. Data Eng. **25**(9), 2090–2104 (2013). http://liris.cnrs.fr/publis/?id=5685
5. Boley, M., Horváth, T., Poigné, A., Wrobel, S.: Listing closed sets of strongly accessible set systems with applications to data mining. Theor. Comput. Sci. **411**(3), 691–700 (2010)
6. Borgatti, S.P., Everett, M.G.: Models of core/periphery structures. Soc. Netw. **21**(4), 375–395 (2000). doi:10.1016/S0378-8733(99)00019-2. http://dx.doi.org/10.1016/S0378-8733(99)00019-2
7. Ferré, S., Ridoux, O.: An introduction to logical information systems. Inf. Process. Manag. **40**(3), 383–419 (2004)
8. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75 – 174 (2010)

[9] http://www.adalab.mib.manchester.ac.uk/.

9. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: International Conference on Conceptual Structures (ICCS). LNCS, vol. 2120, pp. 129–142. Springer, Heidelberg (2001)
10. Kuznetsov, S.O., Samokhin, M.V.: Learning closed sets of labeled graphs for chemical applications. In: Kramer, S., Pfahringer, B. (eds.) ILP. Lecture Notes in Computer Science, vol. 3625, pp. 190–208. Springer, Heidelberg (2005)
11. Mougel, P.N., Rigotti, C., Gandrillon, O.: Finding collections of k-clique percolated components in attributed graphs. In: PAKDD(2), Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference, PAKDD 2012, Kuala Lumpur, 29 May–1 June 2012. Lecture Notes in Computer Science, vol. 7302, pp. 181–192. Springer, Heidelberg (2012)
12. Negrevergne, B., Termier, A., Rousset, M.C., Méhaut, J.F.: Paraminer: a generic pattern mining algorithm for multi-core architectures. Data Min. Knowl. Discov. **28**(3), 593–633 (2013). doi:10.1007/s10618-013-0313-2
13. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–818 (2005). doi:10.1038/nature03607
14. Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., Lakhal, L.: Generating a condensed representation for association rules. J. Intell. Inf. Syst. **24**(1), 29–60 (2005)
15. Pernelle, N., Rousset, M.C., Soldano, H., Ventos, V.: Zoom: a nested Galois lattices-based system for conceptual clustering. J. Exp. Theor. Artif. Intell. **2/3**(14), 157–187 (2002)
16. Rombach, M.P., Porter, M.A., Fowler, J.H., Mucha, P.J.: Core-periphery structure in networks. SIAM J. Appl. Math. **74**(1), 167–190 (2014). doi:10.1137/120881683. http://dx.doi.org/10.1137/120881683
17. Seidman, S.B.: Network structure and minimum degree. Soc. Netw. **5**, 269–287 (1983)
18. Silva, A., Meira Jr., W., Zaki, M.J.: Mining attribute-structure correlated patterns in large attributed graphs. Proc. VLDB Endow. **5**(5), 466–477 (2012). http://dl.acm.org/citation.cfm?id=2140436.2140443
19. Silva, A., Meira Jr., W., Zaki, M.J.: Mining attribute-structure correlated patterns in large attributed graphs. Proc. VLDB Endow. **5**(5), 466–477 (2012)
20. Soldano, H.: Closed patterns and abstraction beyond lattices. In: Glodeanu, C.V., Kaytoue, M., Sacarea, C. (eds.) Formal Concept Analysis 12th International Conference, ICFCA 2014, Cluj-Napoca, 10–13 June 2014. Lecture Notes in Computer Science, vol. 8478, pp. 203–218. Springer, Heidelberg (2014). doi:10.1007/978-3-319-07248-7_15. http://dx.doi.org/10.1007/978-3-319-07248-7_15
21. Soldano, H.: Extensional confluences and local closure operators. In: Baixeries, J., Sacarea, C., Ojeda-Aciego, M. (eds.) Formal Concept Analysis - Proceedings of the 13th International Conference, ICFCA 2015, Nerja, 23–26 June 2015. Lecture Notes in Computer Science, vol. 9113, pp. 128–144. Springer, Heidelberg (2015)
22. Soldano, H.: Extensional confluences and local closure operators. In: Baixeries, J., Sacarea, C., Ojeda-Aciego, M. (eds.) Formal Concept Analysis 13th International Conference, ICFCA, Nerja. LNCS, vol. 9113, pp. 128–144. Springer, Heidelberg (2015)
23. Soldano, H., Santini, G.: Graph abstraction for closed pattern mining in attributed network. In: Schaub, T., Friedrich, G., O'Sullivan, B. (eds.) European Conference in Artificial Intelligence (ECAI). Frontiers in Artificial Intelligence and Applications, vol. 263, pp. 849–854. IOS Press, Prague (2014)
24. Soldano, H., Ventos, V.: Abstract Concept Lattices. In: Valtchev, P., Jäschke, R. (eds.) International Conference on Formal Concept Analysis (ICFCA). LNAI, vol. 6628, pp. 235–250. Springer, Heidelberg (2011)
25. Soldano, H., Santini, G., Bouthinon, D.: Abstract and local rule learning in attributed networks. In: Esposito, F., Hacid, M.S., Pivert, O., Ras, Z. (eds.) Foundations of Intelligent Systems 22nd International Symposium, ISMIS, Lyon. LNAI, vol. 9384, pp. 313–323. Springer, Heidelberg (2015)

26. Soldano, H., Santini, G., Bouthinon, D.: Local knowledge discovery in attributed graphs. In: A. Esposito (ed.) 27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI, pp. 250–257. IEEE Computer Society, Vietri sul Mare (2015)
27. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with titanic. Data Knowl. Eng. **42**(2), 189–222 (2002)
28. Szathmary, L., Napoli, A.: Coron: A framework for levelwise itemset mining algorithms. In: Ganter, B., Godin, R., Nguifo, E.M. (eds.) Third International Conference on Formal Concept Analysis (ICFCA'05), pp. 110–113, Lens, Supplementary Proceedings (2005)
29. Watts, D.J., Strogatz, S.H.: Collective dynamics of /'small-world/' networks. Nature **393**(6684), 440–442 (1998). http://dx.doi.org/10.1038/30918
30. Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L.N., De Grave, K., Ramon, J., de Clare, M., Sirawaraporn, W., Oliver, S.G., King, R.D.: Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. J. R. Soc. Interface **12**(104) (2015). doi:10.1098/rsif.2014.1289. http://rsif.royalsocietypublishing.org/content/12/104/20141289