Local rules associated to k-communities in an attributed graph

Henry Soldano[1,2], Guillaume Santini[1], Dominique Bouthinon[1]

[1]LIPN, Université Paris 13, Sorbonne Paris Cité, France
[2]Atelier de Bio-Informatique, ISYEB, Museum d'Histoire Naturelle, Paris, France

MANEM, ASONAM, 2015

Plan

Mining Patterns in attributed networks

Context

Increasing interest in knowledge discovery in linked data, with a focus on connectivity structure (searching for frequent labelled subgraphs, detecting communities).

- social networks as co-author graphs
- biological networks as gene interaction graphs

and, more recently a focus in attributed networks:

- Each vertex is described in some pattern language (e.g annotation of a gene)

Mining Patterns in attributed networks

Context

Increasing interest in knowledge discovery in linked data, with a focus on connectivity structure (searching for frequent labelled subgraphs, detecting communities).

- social networks as co-author graphs
- biological networks as gene interaction graphs

and, more recently a focus in attributed networks:

- Each vertex is described in some pattern language (e.g annotation of a gene)

Knowledge Discovery Problem

Given a graph whose vertices are labelled by attribute values, find interesting patterns :

dense subgraph(s) $\times$ attribute pattern (Mougel et al 2012, Silva et al 2012)

or

relation between such patterns, as implication/association rules.

From Abstract to Local patterns in attributed networks

Searching for Abstract Knowledge (Soldano and Santini, ECAI 2014)

- Define an abstract lattice of (subgraph, attribute pattern) pairs, where the subgraph is made of highly connected parts of the pattern subgraph (for instance made of k-cliques), plus derived abstract implication rules

Searching for Local Knowledge (This work)

- Investigate (subgraph, attribute pattern) pairs, where the subgraph is highly connected (for instance focussing on one connected component of the pattern subgraph), plus derived local implication rules

From Abstract to Local implications

Implication validity relies on inclusion of (standard, abstract or local) extensions.

Let $G = (O, E)$ be an attributed network.

- Valid on 2^O
  Any vertex which has q also has w

$$q \rightarrow w \text{ iff } \mathrm{ext}(q) \subseteq \mathrm{ext}(w)$$

- Valid on abstraction A (vertex subsets of G made of union of triangles). Any triangle which has q also has w

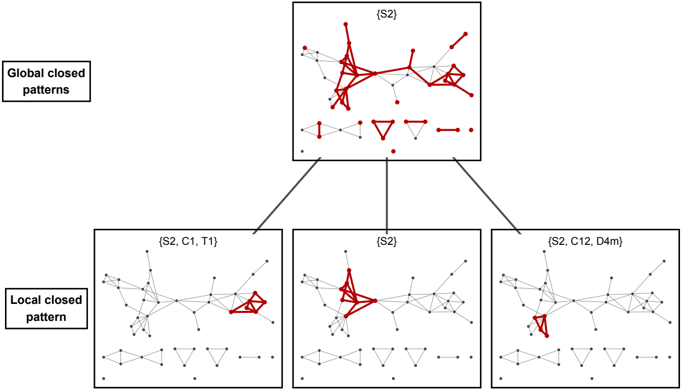
$$\Box q \rightarrow \Box w \text{ iff } \mathrm{ext}_A(q) \subseteq \mathrm{ext}_A(w)$$

- Valid on confluence F (connected vertex subsets of G). Any connected vertex subset containing i which has q also has w

$$\Box_i q \rightarrow \Box_i w \text{ iff } \mathrm{ext}_i(q) \subseteq \mathrm{ext}_i(w)$$

Example: 3-communities in a friendship network

A network of teenage friends in Scotland and their lifestyle.



$$\Box_{t_1}\ \text{S2} \rightarrow \Box_{t_1}\ \text{S2-C1-T1}$$

The community that contains t_1 and has a regular sporting activity (S2), also does not smoke Cannabis nor Tobacco (C1, T1).

Plan

Support-closed patterns in Data Mining and FCA

Let L be a pattern language and O a set of objects in which patterns may occur

Definition (Support-closed patterns)

- $t \equiv_O t'$ iff $ext(t) = ext(t')$
- The maximal elements of the equivalence classes are the support-closed patterns.

Support-closed patterns in Data Mining and FCA

Let L be a pattern language and O a set of objects in which patterns may occur

Definition (Support-closed patterns)

- $t \equiv_O t'$ iff $ext(t) = ext(t')$
- The maximal elements of the equivalence classes are the support-closed patterns.

When the pattern language is a lattice, there is a closure operator f such that in each equivalence class

- the closed pattern $c = f(t)$ is the unique support-closed element equivalent to t,
- the implication rules $t \rightarrow c \backslash t$ hold on O.

Support-closed patterns in Data Mining and FCA

Let L be a pattern language and O a set of objects in which patterns may occur

Definition (Support-closed patterns)

- $t \equiv_O t'$ iff $ext(t) = ext(t')$
- The maximal elements of the equivalence classes are the support-closed patterns.

When the pattern language is a lattice, there is a closure operator f such that in each equivalence class

- the closed pattern $c = f(t)$ is the unique support-closed element equivalent to t,
- the implication rules $t \rightarrow c \backslash t$ hold on O.

$f(t) = int \circ ext(t)$

Given $e \subseteq O$, $\mathrm{int}(e)$ is obtained by intersecting the elements of e.

Support-closed patterns in Data Mining and FCA

Let L be a pattern language and O a set of objects in which patterns may occur

Definition (Support-closed patterns)

- $t \equiv_O t'$ iff $ext(t) = ext(t')$
- The maximal elements of the equivalence classes are the support-closed patterns.

When the pattern language is a lattice, there is a closure operator f such that in each equivalence class

- the closed pattern $c = f(t)$ is the unique support-closed element equivalent to t,
- the implication rules $t \rightarrow c \backslash t$ hold on O.

$f(t) = int \circ ext(t)$

Given $e \subseteq O$, $\mathrm{int}(e)$ is obtained by intersecting the elements of e.

The equivalence classes form a (concept) lattice of (e, c) pairs

Interior Operators and Abstractions, JETAI 2002 to ICFCA 2011

Projection

$p \ : \ M \rightarrow M$ is an interior operator or a projection on $(M, \leq)$ *iff* :

- $p(x) \leq x$ (intensivity)
- $x \leq y \ \Rightarrow \ p(x) \leq p(y)$ (monotonicity)
- $p(x) = p(p(x))$ (idempotence)

Interior Operators and Abstractions, JETAI 2002 to ICFCA 2011

Projection

$p : M \rightarrow M$ is an interior operator or a projection on $(M, \leq)$ *iff* :

- $p(x) \leq x$ (intensivity)
- $x \leq y \Rightarrow p(x) \leq p(y)$ (monotonicity)
- $p(x) = p(p(x))$ (idempotence)

Extensional abstraction reduces support sets to abstract support sets

Let $A = p[2^O]$ whose elements are called abstract groups

- $p \circ ext(t)$ is the abstract support set of t,

Interior Operators and Abstractions, JETAI 2002 to ICFCA 2011

Projection

$p \ : \ M \rightarrow M$ is an interior operator or a projection on $(M, \leq)$ *iff* :

- $p(x) \leq x$ (intensivity)
- $x \leq y \ \Rightarrow \ p(x) \leq p(y)$ (monotonicity)
- $p(x) = p(p(x))$ (idempotence)

Extensional abstraction reduces support sets to abstract support sets

Let $A = p[2^O]$ whose elements are called abstract groups

- $p \circ ext(t)$ is the abstract support set of t,
- $f(t) = int \circ p \circ ext(t)$ is an abstract closed pattern

Interior Operators and Abstractions, JETAI 2002 to ICFCA 2011

Projection

$p : M \to M$ is an interior operator or a projection on $(M, \leq)$ *iff* :

- $p(x) \leq x$ (intensivity)
- $x \leq y \Rightarrow p(x) \leq p(y)$ (monotonicity)
- $p(x) = p(p(x))$ (idempotence)

Extensional abstraction reduces support sets to abstract support sets

Let $A = p[2^O]$ whose elements are called abstract groups

- $p \circ ext(t)$ is the abstract support set of t,
- $f(t) = int \circ p \circ ext(t)$ is an abstract closed pattern
- $\Box t_1 \to \Box t_2$ iff $p \circ ext(t_1) \subseteq p \circ ext(t_2)$ means:
  if an abstract group shares pattern t_1 then the group shares t_2
- We obtain an (abstract) lattice of (e, c)*pairs*

Graph abstractions

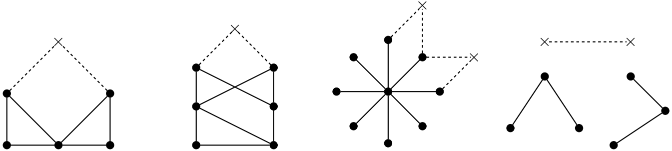
Let $G = (V, E)$ be a graph and $G_e = (e, E(e))$ be the subgraph induced by the vertex subset e.

We can build a graph abstraction by

- defining a property $P(x, e)$ on a vertex x of G_e such that the truth of P is preserved when increasing the subgraph by adding new vertices and corresponding edges.

$p(e)$ is the greatest subset $e' \subseteq e$ such that $P(x, e')$ is true for x in e'.

Graph abstractions, ECAI 2014



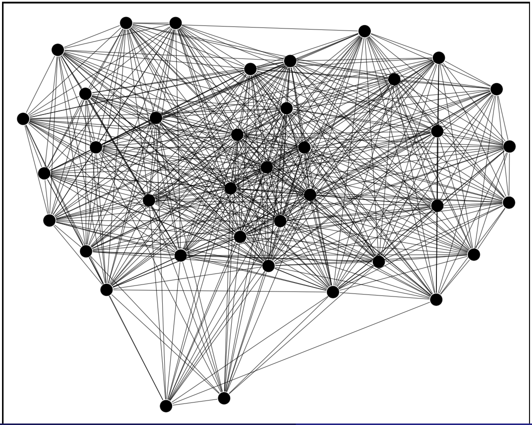
(a) (b) (c) (d)

$e' = p(e)$, i.e. e' belongs to the graph abstraction, iff for all x within $G_{e'}$:

- (a) x belongs to a triangle, $(3 - clique)$
- (b) x belongs to a 2-club of size at least 6 $(2 - club \geq 6)$
- (c) x has degree at least 8 or is connected to a vertex y of degree at least 8 $(near - star(8))$
- (d) x belongs to a connected component whose size is at least 3 $(cc \geq 3)$.

A degree $\geq$ 16 pattern in a DBLP network

- 45131 authors labelled with DM and DB conferences and journals (1990–2011) and 228,188 co-authoring links (A. Bechara Prado and coll. 2013)
- From VLDBJ with support 1276 and abstract support 38, we obtain $\Box$ VLDBJ $\rightarrow$ $\Box$ ICDE, SIGMOD, VLDB



What does that means ?

What does that means ?

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

[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, Jeffrey 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)

Graph abstractions in multiplex networks

Natural extension to multiplex networks:

Average degree among layers

e belongs to the graph abstraction iff for all x, the average degree of x in the G_e^i is such that $\bar{d}(x) \geq k$

To belong to a graph pattern in several layers

e belongs to the graph abstraction iff for all x, x belongs to a triangle in at least k layers

Plan

Pre-confluences and confluences (ICFCA 2014)

Definition

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 lattice is a pre-confluence with a minimum

Lemma

For any $x, y \in F^m$ their least upper bound does not depend on m:

1. *$x \vee_F y$ is the least element of $F^x \cap F^y$*

A pre-confluence is a union of lattices in which joins coincide

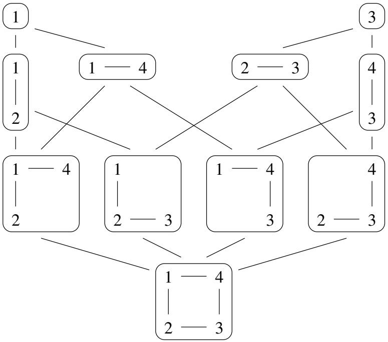
Definition

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 is a confluence of T with $\perp_T$ as minimum.

The set of connected vertex subsets of a graph

The pre-confluence F of connected vertex subsets of $G = (\{1, 2, 3, 4\}, \{12, 23, 34, 14\})$ containing 1 or 3:



F also is a confluence of $T = 2^{1234}$

A confluence is associated to a set of interior operators $p_m : T^m \rightarrow F^m$ s.t. $p_m(t)$ is the greatest subset of t in F containing m:

$p_1(13) = 1$, $p_3(13) = 3$, $p_1(123) = p_3(123) = 123$

The following result generalizes a previous result on abstractions:

Proposition

Let F_1 and F_2 be two confluences of T then, $F_{12} = F_1 \cap F_2$ is a confluence of T

Example

Let G be a graph and

- F be the set of connected vertex subsets of graph G
- A be the set of vertex subsets made of triangles of G
- F_A is the set of connected vertex subsets of graph G made of triangles

Local closures (Soldano, ICFCA 2015)

Let F be a confluence of $X = 2^O$ and m a minimal object subset in F
Consider $F^m = p_m[X^m]$ and $L_{\mathrm{int}(m)}$, i.e. patterns that occurs in m:

Proposition

1. $f_m = \mathrm{int} \circ p_m \circ \mathrm{ext}$ *is a closure operator on* $L_{int(m)}$

$p_m(\mathrm{ext}(q))$ is the local support set of q in F that contains m.
f_m is the *local closure operator* with respect to m.

Example

$f_i(q))$ is the most specific pattern that occurs in the connected component of the pattern q subgraph that contains vertex i.

Local closure operators

The local support sets form a pre-confluence:

Theorem

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

$h(e)$ is the local support set of $\mathrm{int}(e)$ that contains $m \leq e$
$h[F]$ is a pre-confluence isomorphic to the set P of local concept pairs:

Definition

$P = \{(e, I) \mid e = p_m \circ \mathrm{ext}(I), I = \mathrm{int}(e), m \leq e\}$ *is called a local concept pre-confluence*

Plan

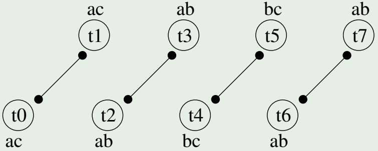
Local implications

$p_m \circ \mathrm{ext}(q) \subseteq p_m \circ \mathrm{ext}(w)$, rewrites as a local implication $\Box_m q \rightarrow \Box_m w$.

The set of $\Box_m c \rightarrow \Box_m l$ local implications, where c is a (global) closed pattern and l a local closed pattern, with $c \subset l$, represents (a basis for) the local knowledge associated to the confluence F.

Example

Attributed graph G, and confluence F of connected vertex subsets of G with size at least 2.

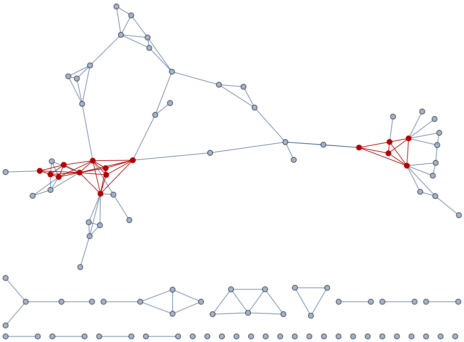


4 local concepts $(\{t_0, t_1\}, ac)$, $(\{t_2, t_3\}, ab)$, $(\{t_4, t_5\}, bc)$, $(\{t_6, t_7\}, ab)$
Various local implications rules as
$r_1 : \Box_{t_1} a \rightarrow \Box_{t_1} ac$, $r_2 : \Box_{t_1} ac \rightarrow \Box_{t_1} ac$.
r_1 is more informative than $r_2 \Rightarrow r_2$ is eliminated.

DBLP: DMKD,IDArev $\rightarrow_{268924}$ DMgroup

A local rule $c \rightarrow_i l$ with

- c an abstract closed pattern in the degree ≥ 4 abstraction A
- i is a vertex of the left connected component of the red subgraph induced by $ext_A(c)$
- l is the corresponding local closed pattern



Plan

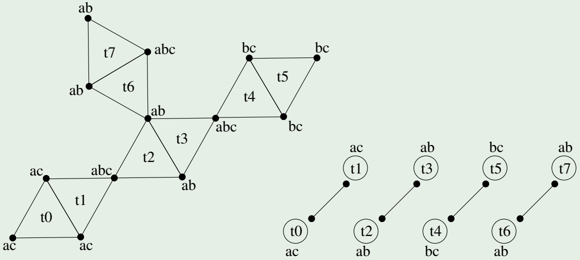
Deriving a graph from a set of vertex subsets

Until now: A local support set is a connected component of some pattern subgraph. What if, given some pattern, interesting local vertex subsets overlap ?

Example (3-communities)

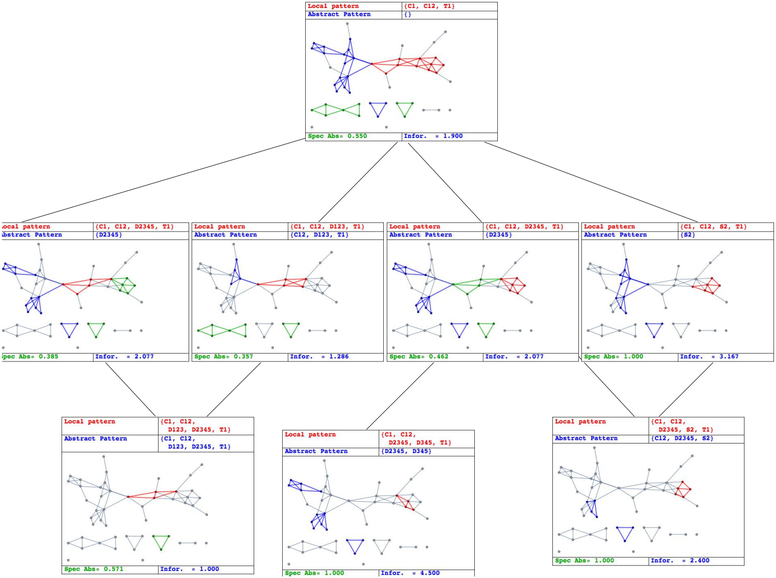
$T \subseteq 2^O$= triangles in $G \Rightarrow$

$G_T = (T, E_T)$ where $(t, t') \in E_T$ iff t and t' are adjacent in G.



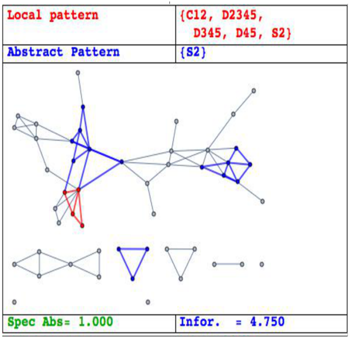
F^T= Confluence of 3-communities

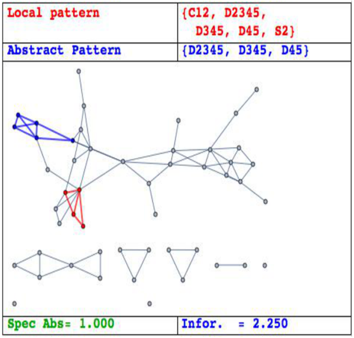
The pre-confluence of size $\geq$ 4 3-communities (part)



The pre-confluence of size $\geq$ 4 3-communities (zoom)

An element (e, I) of the pre-confluence and two local rules.





$\Box_m$ D45 $\rightarrow \Box_m$ C12-D45-S2
$\Box_m$ S2 $\rightarrow \Box_m$ C12-D45-S2

Summary

We have defined local concepts and local rules:

- In a local concept (e, l), the local support set e is the greatest object subset in F including some minimal object subset m in which l occurs.
- A local concept pre-confluence is associated to a basis of local implications each relating a closed pattern c to a local closed pattern $l = \mathrm{int}(e)$ associated to some minimal object subset m.
- In attributed graphs, local concepts and local implications rely on "highly connected" subgraphs induced by attribute patterns.
- Local rules (c, e, l) are enumerated using ParaminerLC, a variant of PARAMINER (Negrevergne et al, 2014)
  Only *maximally informative* rules are selected:
  - (c, e, l) is such that $l \neq c$
  - (c, e, l) eliminates (c', e, l) if $c < c'$