Local Pattern Detection in Attributed Graphs

Jean-François Boulicaut[1], Marc Plantevit[2], and Céline Robardet[1(✉)]

[1] Université de Lyon, CNRS, INSA de Lyon, LIRIS UMR5205,
69621 Villeurbanne, France
{jean-francois.boulicaut,celine.robardet}@insa-lyon.fr
[2] Université de Lyon, CNRS, Univ. Lyon1, LIRIS UMR5205,
69622 Villeurbanne, France
marc.plantevit@univ-lyon1.fr

Abstract. We propose to mine the topology of a large attributed graph by finding regularities among vertex descriptors. Such descriptors are of two types: (1) the vertex attributes that convey the information of the vertices themselves and (2) some topological properties used to describe the connectivity of the vertices. These descriptors are mostly of numerical or ordinal types and their similarity can be captured by quantifying their co-variation. Mining topological patterns relies on frequent pattern mining and graph topology analysis to reveal the links that exist between the relation encoded by the graph and the vertex attributes. In this paper, we study the network of authors who have cooperated at some time with Katharina Morik according to the data available in DBLP database. This is a nice occasion for formalizing different questions that can be considered when an attributed graph describes both a type of interaction and node descriptors.

Keywords: Attributed graph mining · Katharina Morik co-authorship

1 Introduction

A timely challenge concerns enriched graph mining to support knowledge discovery. We recently proposed the topological pattern domain [25], a kind of gradual pattern that extends the rank-correlated sets from [6] to support attributed graph analysis. In such graphs, the binary relation encoded by the graph is enriched by vertex numerical attributes. However, existing methods that support the discovery of local patterns in graphs mainly focus on the topological structure of the patterns, by extracting specific subgraphs while ignoring the vertex attributes (cliques [21], quasi-cliques [20,29]), or compute frequent relationships between vertex attribute values (frequent subgraphs in a collection of graphs [16] or in a single graph [5]), while ignoring the topological status of the vertices within the whole graph, e.g., the vertex connectivity or centrality. The same limitation holds for the methods proposed in [18,24,27,28], which identify sets of vertices that have similar attribute values and that are close neighbors. Such approaches only focus on a local neighborhood of the vertices and do not consider the connectivity of the vertex in the whole graph.

S. Michaelis et al. (Eds.): Morik Festschrift, LNAI 9580, pp. 168–183, 2016.
DOI: 10.1007/978-3-319-41706-6_8

To investigate the relations that may exist between the position of the vertices within the graph and their attribute values, we proposed to extract topological patterns that are sets made of vertex attributes and topological measures. Such measures quantify the topological status of each vertex within the graph. Some of these measures are based on the close neighborhood of the vertices (e.g., the vertex degree), while others describe the connectivity of a vertex by considering its relationship with all other vertices (e.g., the centrality measures). Combining such microscopic and macroscopic properties characterizes the connectivity of the vertices and it may be a sound basis to explain why some vertices have similar attribute values.

Topological patterns of interest are composed of vertex properties that behave similarly over the vertices of the graph. The similarity among vertex properties can be captured by quantifying their correlation, which may be positive or negative. To that end, we extend the Kendall rank correlation coefficient to any number of variables, as well as to negative correlation. Whereas this measure is rather theoretically sounded, its evaluation is computationally demanding as it requires to consider all vertex pairs to estimate the proportion of which that supports the pattern. The well known optimization techniques that are used for evaluating the correlation between two variables (and that leads to a theoretical complexity in $O(n \log n)$) do not extend directly when a higher number of variables is considered. We tackled this issue and proposed several optimization and pruning strategies that makes it possible to use this approach on large graphs. We also introduced several interestingness measures of topological patterns that differ by the pairs of vertices that are considered while evaluating the correlation between descriptors: (1) While all the vertex pairs are considered, patterns that are true all over the graph are extracted; (2) When including only the vertex pairs that are in a specific order regarding to a selected numerical or ordinal attribute reveals the topological patterns that emerge with respect to this attribute; (3) Examining the vertex pairs that are connected in the graph makes it possible to identify patterns that are *structurally correlated* to the relationship encoded by the graph. Besides, we designed an operator that identifies the top k representative vertices of a topological pattern.

In this paper, we study the network of authors who have cooperated at some time with Katharina Morik according to the data available in the DBLP database. Doing so, we emphasize powerful mechanisms for detecting new types of local patterns in interaction graphs. Indeed, we formalize different questions that can be considered when an attributed graph describes both interactions and vertex descriptors. This has not yet been studied systematically. It enables also to discuss the need for new post-processing techniques that exploit both the patterns and the graph data. Finally, detecting local patterns in various data types has motivated a lot of research in our group and writing a chapter at this Festschrift occasion is also an implied reference to the domain that gave us the first occasion to spend time and work with our smart colleague [22].

2 Related Work

Graph mining is an active topic in Data Mining. In the literature, there exist two main trends to analyze graphs. On the one hand, graphs are studied at a macroscopic level by considering statistical graph properties (e.g., diameter, degree distribution) [2,7]. On the other hand, sophisticated graph properties are discovered by using a local pattern mining approach. Recent approaches mine attributed graphs which convey more information. In such graphs, information is locally available on vertices by means of attribute values. As argued by Moser et al. [23], "*often features and edges contain complementary information, i.e., neither the relationships can be derived from the feature vectors nor vice versa*".

Attributed graphs are extensively studied by means of clustering techniques (see e.g., [1,8,13,15,19,32]) whereas pattern mining techniques in such graphs have been less investigated. The pioneering work [23] proposes a method to find dense homogeneous subgraphs (i.e., subgraphs whose vertices share a large set of attributes). Similar to this work, Günnemann et al. [14] propose a method based on subspace clustering and dense subgraph mining to extract non redundant subgraphs that are homogenous with respect to vertex attributes. Silva et al. [28] extract pairs of dense subgraphs and Boolean attribute sets such that the Boolean attributes are strongly associated with the dense subgraphs. In [24], the authors propose the task of finding the collections of homogeneous k-clique percolated components (i.e., components made of overlapping cliques sharing a common set of true valued attributes) in Boolean attributed graphs. Another approach is presented in [18], where a larger neighborhood is considered. This pattern type relies on a relaxation of the accurate structure constraint on subgraphs. Roughly speaking, they propose a probabilistic approach to both construct the neighborhood of a vertex and propagate information into this neighborhood. Following the same motivation, Sese et al. [26] extract (not necessarily dense) subgraphs with common itemsets. Note that these approaches use a single type of topological information based on the neighborhood of the vertices. Furthermore, they do not handle numerical attributes as in our proposal. However, global statistical analysis [11] of a single graph considers several measures to describe the graph topology, but does not benefit from vertex attributes. Besides, current local pattern mining techniques on attributed graphs do not consider numerical attributes nor macroscopic topological properties. To the best of our knowledge, our paper represents a first attempt to combine both microscopic and macroscopic analysis on graphs by means of (emerging) topological pattern mining. Indeed, several approaches aim at building global models from local patterns [12], but none of them tries to combine information from different graph granularity levels.

Co-variation patterns are also known as gradual patterns [9] or rank-correlated itemsets [6]. Do et al. [9] use a support measure based on the length of the longest path between ordered objects. This measure has some drawbacks w.r.t. computational and semantics aspects. Calders et al. [6] introduce a support measure based on the Kendall's τ statistical measure. However, their approach is not defined to simultaneously discover up and down co-variation patterns as

does our approach. Another novelty of our work is the definition of other interestingness measures to capture emerging co-variations. Finally, this work is also the first attempt to use co-variation pattern mining in attributed graphs.

3 Topological Vertex Properties

Let us consider a non-directed attributed graph $G = (V, E, L)$, where V is a set of n vertices, E a set of m edges, and $L = \{l_1, \cdots, l_p\}$ a set of p numerical or ordinal attributes associated with each vertex of V. Important properties of the vertices are encoded by the edges of the graph. From this relation, we can compute some topological properties that synthesize the role played by each vertex in the graph. The topological properties we are interested in range from a microscopic level – those that described a vertex based on its direct neighborhood – to a macroscopic level – those that characterize a vertex by considering its relationship to all other vertices in the graph. Statistical distributions of these properties are generally used to depict large graphs (see, e.g., [2,17]). We propose here to use them as vertex descriptors.

3.1 Microscopic Properties

Let us consider here only three topological properties to describe the direct neighborhood of a vertex v:

- The degree of v is the number of edges incident to v ($deg(v) = |\{u \in V, \{u, v\} \in E\}|$). When normalized by the maximum number of edges a vertex can have, it is called the degree centrality coefficient: $\textsc{Degree}(v) = \frac{deg(v)}{n-1}$.
- The clustering coefficient evaluates the connectivity of the neighbors of v and thus its local density:

$$\textsc{Clust}(v) = \frac{2|\{\{u, w\} \in E, \{u, v\} \in E \wedge \{v, w\} \in E\}|}{deg(v)(deg(v) - 1)}$$

3.2 Mesoscopic Property

We also consider the position of each vertex to the center of the graph, that is the distance – the number of edges of a shortest path – to a peculiar vertex. In the following, we call this property the $\textsc{Morik\_number}(\textsc{v})$ as we consider the relative position of the vertices to the vertex that corresponds to Katharina Morik.

3.3 Macroscopic Properties

We consider five macroscopic topological properties to characterize a vertex while taking into account its connectivity to all other vertices of the graph.

- The relative importance of vertices in a graph can be obtained through centrality measures [11]. Closeness centrality $\textsc{Close}(v)$ is defined as the inverse of the average distance between v and all other vertices that are reachable from it. The distance between two vertices is defined as the number of edges of the shortest path between them: $\textsc{Close}(v) = \frac{n}{\sum_{u \in V} |shortest\_path(u,v)|}$.
- The betweenness centrality $\textsc{Betw}(v)$ of v is equal to the number of times a vertex appears on a shortest path in the graph. It is evaluated by first computing all the shortest paths between every pair of vertices, and then counting the number of times a vertex appears on these paths: $\textsc{Betw}(v) = \sum_{u,w} \mathbb{1}_{shortest\_path(u,w)}(v)$.
- The eigenvector centrality measure (EgVect) favours vertices that are connected to vertices with high eigenvector centrality. This recursive definition can be expressed by the following eigenvector equation $Ax = \lambda x$ which is solved by the eigenvector x associated to the largest eigenvalue λ of the adjacency matrix A of the graph.
- The PageRank index [4] is based on a random walk on the vertices of the graph, where the probability to go from one vertex to another is modelled as a Markov chain in which the states are vertices and the transition probabilities are computed based on the edges of the graph. This index reflects the probability that the random walk ends at the vertex itself:

$$\textsc{PageRank}(v) = \alpha \sum_{u} \mathbb{1}_E(\{u,v\}) \frac{\textsc{PageRank}(u)}{deg(u)} + \frac{1-\alpha}{n}$$

  where the parameter α is the probability that a random jump to vertex v occurs.
- Network constraint [30] evaluates to what extent person's contacts are redundant

$$\textsc{Network}(v) = \sum\nolimits_{u|(u,v)\in E} [\frac{1}{deg(v)} + \sum\nolimits_{w|(u,w) \; and \; (v,w)\in E} (\frac{1}{deg(v)} \frac{1}{deg(u)})]^2$$

  When its value is low, the contacts are rather disconnected, whereas when it is high, the contacts are close or strongly tied.

These 9 topological properties characterizes the graph relationship encoded by E. These properties, along with the set of vertex attributes L, constitutes the set of vertex descriptors $\mathcal{D}$ used in this paper.

4 Topological Patterns

Let us now consider topological patterns as a set of vertex attributes and topological properties that behave similarly over a large part of the vertices of the graph. We assume that all topological properties and vertex attributes are of numerical or ordinal type, and we propose to capture their similarity by quantifying their co-variation over the vertices of the graph. Topological patterns are

defined as $P = \{{D_1}^{s_1}, \cdots, {D_\ell}^{s_\ell}\}$, where D_j, $j = 1 \ldots \ell$, is a vertex descriptor from $\mathcal{D}$ and $s_j \in \{+, -\}$ is its co-variation sign. In the following, we propose three pattern interestingness measures that differ in the pairs of vertices considered for their evaluation.

4.1 Topological Patterns over the Whole Graph

Several signed vertex descriptors co-vary if the orders induced by each of them on the set of vertices are consistent. This consistency is evaluated by the number of vertex pairs ordered the same way by all descriptors. The number of such pairs constitutes the so-called support of the pattern. This measure can be seen as a generalization of the Kendall's τ measure. When we consider all possible vertex pairs, this interestingness measure is defined as follows:

Definition 1 ($Supp_{all}$). *The support of a topological pattern P over all possible pairs of vertices is:*

$$Supp_{all}(P) = \frac{|\{(u,v) \in V^2 \mid \forall D_j^{s_j} \in P : D_j(u) \rhd_{s_j} D_j(v)\}|}{\binom{n}{2}}$$

where $\rhd_{s_j}$ denotes $<$ when s_j is equal to $+$, and $\rhd_{s_j}$ denotes $>$ when s_j is equal to $-$.

This measure gives the number of vertex pairs (u, v) such that u is strictly lower than v on all descriptors with sign $+$, and u is strictly higher than v on descriptors with sign $-$.

As mentioned in [6], $Supp_{all}$ is an anti-monotonic measure for positively signed descriptors. This is still true when considering negatively signed ones: adding D_{l+1}^- to a pattern P leads to a support lower than or equal to that of P since the pairs (u, v) that support P must also satisfy $D_{l+1}(u) > D_{l+1}(v)$. Besides, when adding descriptors with negative sign, the support of some patterns can be deduced from others, the latter referred to as symmetrical patterns.

Property 1 (Support of symmetrical patterns). Let P be a topological pattern and $\overline{P}$ be its symmetrical, that is, $\forall D_j^{s_j} \in P$, $D_j^{\overline{s_j}} \in \overline{P}$, with $\overline{s_j} = \{+, -\} \setminus \{s_j\}$. If a pair (u, v) of V^2 contributes to the support of P, then the pair (v, u) contributes to the support of $\overline{P}$. Thus, we have $Supp_{all}(P) = Supp_{all}(\overline{P})$.

Topological patterns and their symmetrical patterns are semantically equivalent. To avoid the irrelevant computation of duplicate topological patterns, we exploit Property 1 and enforce the first descriptor of a pattern P to be signed by $+$.

Mining frequent topological patterns consists in computing all sets of signed descriptors P, but not their symmetrical ones, such that $Supp_{all}(P) \geq minsup$, where $minsup$ is a user-defined minimum support threshold.

4.2 Other Interestingness Measures

To identify most interesting topological patterns, we propose to give to the end-user the possibility of guiding its data mining process by querying the patterns with respect to their correlation with the relationship encoded by the graph or with a selected descriptor. Therefore, we revisit the notion of emerging patterns [10] by identifying the patterns whose support is significantly greater (i.e., according to a growth-rate threshold) in a specific subset of vertex pairs than in the remaining ones. This subset can be defined in different ways according to the end-user's motivations: either it is defined by the vertex pairs that are ordered with respect to a selected descriptor called the class descriptor, or it is equal to E, the set of edges. Whereas the former highlights the correlation of a pattern with the class descriptor, the latter enables to characterize the importance of the graph structure within the support of the topological pattern.

Emerging Patterns w.r.t. a Selected Descriptor. Let us consider a selected descriptor $C \in \mathcal{D}$ and a sign $r \in \{+, -\}$. The set of pairs of vertices that are ordered by C^r is

$$\mathcal{C}_{C^r} = \{(u, v) \in V^2 \mid C(u) \rhd_r C(v)\}$$

The support measure based on the vertex pairs of $\mathcal{C}_{C^r}$ is defined below.

Definition 2 ($Supp_{C^r}$). *The support of a topological pattern P over C^r is:*

$$Supp_{C^r}(P) = \frac{|\{(u, v) \in \mathcal{C}_{C^r} \mid \forall D_j^{s_j} \in P : D_j(u) \rhd_{s_j} D_j(v)\}|}{|\mathcal{C}_{C^r}|}$$

Analogously, the support of P over the pairs of vertices that do not belong to $\mathcal{C}_{C^r}$ is denoted $Supp_{C^{\overline{r}}}(P)$. To evaluate the impact of C^r on the support of P, we consider the growth rate of the support of P over the partition of vertex pairs $\{\mathcal{C}_{C^r}, \mathcal{C}_{C^{\overline{r}}}\}$: $Gr(P, C^r) = \frac{Supp_{C^r}(P)}{Supp_{C^{\overline{r}}}(P)}$

If $Gr(P, C^r)$ is greater than a minimum growth-rate threshold, then P is referred to as emerging with respect to C^r. If $Gr(P, C^r) \approx 1$, P is as frequent in $\mathcal{C}_{C^r}$ as in $\mathcal{C}_{C^{\overline{r}}}$. If $gr(P, C^r) \gg 1$, P is much more frequent in $\mathcal{C}_{C^r}$ than in $\mathcal{C}_{C^{\overline{r}}}$. For example, $Gr(\{h^+, i^-, \textsc{Betw}^+\}, t^+) = 2.31$. The intuition behind this definition is to identify the topological patterns that are mostly supported by pairs of vertices that are also ordered by the selected descriptor.

Emerging Patterns w.r.t. the Graph Structure. It is interesting to measure if the graph structure plays an important role in the support of a topological pattern P. To this end, we define a similar support measure based on pairs that belongs to E, the set of edges of the graph:

$$\mathcal{C}_E = \{(u, v) \in V^2 \mid \{u, v\} \in E\}$$

Based on this set of pairs, we define the support of P as:

Definition 3 ($Supp_E$). *The support of a topological pattern P over the pairs of vertices that are linked in G is:*

$$Supp_E(P) = \frac{2|\{(u,v) \in \mathcal{C}_E \mid \forall D_j^{s_j} \in P : D_j(u) \rhd_{s_j} D_j(v)\}|}{|\mathcal{C}_E|}$$

The maximum value of the numerator is $\frac{|\mathcal{C}_E|}{2}$ since: (1) if $(u,v) \in \mathcal{C}_E$ then $(v,u) \in \mathcal{C}_E$, and (2) it is not possible that $\forall D_j^{s_j} \in P$, $D_j(u) \rhd_{s_j} D_j(v)$ and $D_j(v) \rhd_{s_j} D_j(u)$ at the same time. For instance, the pattern $\{h^+, i^-\}$ is supported by all the twenty possible pairs that are edges, its support is thus equal to 1. The support of P over the pairs of vertices that do not belong to $\mathcal{C}_E$ is denoted $Supp_{\overline{E}}(P)$.

As before, to evaluate the impact of E on the support of P, we consider the growth rate of the support of P over the partition of vertex pairs $\{\mathcal{C}_E, \mathcal{C}_{\overline{E}}\}$: $Gr(P,E) = \frac{Supp_E(P)}{Supp_{\overline{E}}(P)}$.

$Gr(P,E)$ enables to assess the impact of the graph structure on the pattern. Therefore, if $Gr(P,E) \gg 1$, P is said to be *structurally* correlated. If $Gr(P,E) \ll 1$, the graph structure tends to inhibit the support of P.

5 Top k Representative Vertices

The user may be interested in identifying the vertices that are the most representative of a given topological pattern, thus enabling the projection of the patterns back into the graph. For example, the representative vertices of the pattern $\{\text{t}^+, \text{BETW}^-\}$ would be researchers with a relatively large number of IEEE TKDE papers and a low betweenness centrality measure.

We denote by $S(P)$ the set of vertex pairs (u,v) that constitutes the support of a topological pattern P:

$$S(P) = \{(u,v) \in V^2 \mid \forall D_j^{s_j} \in P : D_j(u) \rhd_{s_j} D_j(v)\}$$

which forms, with V, a directed graph $G_P = (V, S(P))$. This graph satisfies the following property.

Property 2. The graph $G_P = (V, S(P))$ is transitive and acyclic.

Proof. Let us consider $(u,v) \in V^2$ and $(v,w) \in V^2$ such that, $\forall D_j^{s_j} \in P$: $D_j(u) \rhd_{s_j} D_j(v)$ and $D_j(v) \rhd_{s_j} D_j(w)$. Thus, $D_j(u) \rhd_{s_j} D_j(w)$ and $(u,w) \in S(P)$. Therefore, G_P is transitive.

As $\rhd_s \in \{<, >\}$, it stands for a strict inequality. Thus, if $(u,v) \in S(P)$, $(v,u) \notin S(P)$. Furthermore, as G_P is transitive, if there exists a path between u and v, there is also an arc $(u,v) \in S(P)$. Therefore, $(v,u) \notin S(P)$ and we can conclude that G_P is acyclic.

As G_P is acyclic, it admits a topological ordering of its vertices, which is, in the general case, not unique. The top k representative vertices of a topological

pattern P are identified on the basis of such a topological ordering of V and are the k last vertices with respect to this ordering. Considering that an arc $(u, v) \in S(P)$ is such that v dominates u on P, this vertex set contains the most dominant vertices on P. The top k representative vertices of P can be easily identified by ordering the vertices by their incoming degree.

Although the support of topological patterns is an anti-monotonic measure, its computation is quadratic in the number of vertices of the graph which prevents the extraction of such patterns on large graph using classical pattern mining algorithms. To overcome this problem, we proposed in [25] an upper bound on this measure that can be computed linearly in the number of vertices. This upper bound takes advantages of the presence of ties in the descriptor values. By pre-computing some indexes on the descriptors, almost all non frequent patterns are pruned without computing their support when the minimum support is high.

The computation of topological patterns is done in an ECLAT-based way [31]. More precisely, all the subsets of a pattern P are always evaluated before P itself. In this way, by storing all frequent patterns in the hash-tree $\mathcal{M}$, the anti-monotonic frequency constraint is fully-checked on the fly. We compute the upper bound on the support to prune non-promising topological patterns. When this upper bound is greater than the minimum threshold, the exact support is computed. Another optimization is based on the deduction of the support from already evaluated patterns: A pair of vertices that supports a pattern P can support pattern PA^+ or pattern PA^-, or none of them. Thus, another upper bound on $Supp_{all}(PA^-)$ is $Supp_{all}(P) - Supp_{all}(PA^+)$. Note that these patterns have already been considered before the evaluation of PA^-. So, to be stringent, we bound the support by taking the minimum between this value and the upper bound. When computing the support of the pattern, the top k representative vertices are also identified.

6 Studying Katharina Morik's Network

In the following, we propose to use TopGraphMiner to study the scientific co-authorship network of Katharina Morik. After presenting the attributed graph we generate from the DBLP digital library[1], we provide qualitative results that show the implication of Katharina in the machine learning community.

6.1 Katharina Morik's Co-authorship Network

The co-authorship graph is built from the DBLP digital library. Regarding Katharina's bibliography, we select all the conference venues and journals in which Katharina has at least one DBLP entry[2]. We gather all the publications in these conference venues and journals since their foundation, and derived a graph where the vertices stand for the authors and edges link two authors who

[1] http://dblp.uni-trier.de/.
[2] http://www.dblp.org/search/index.php?query=author:katharina_morik.

co-authored at least one paper in this corpus. To each vertex, we associate the number of publications in each of these 53 selected conferences or journals as vertex properties. We then removed isolated vertices, that is to say, authors who has no co-author in the selected publications. The resulting attributed graphs involves 81 222 vertices and 466 152 undirected edges. Notice that, even if this attributed graph is generated based on Katharina's publications, her co-authors only represent 0.1 % of the vertices of the whole graph, while the vertices whose distance to Katharina is at most 2 represent less than 2 % of the whole set of vertices. The average Morik number is 4.05 and 4033 authors have no path to Katharina (infinite Morik number). There are 1428 connected components.

Figure 1 presents this co-authorship graph restricted to the authors that are at most at a distance of 2 from Katharina and that have a degree value greater than 20. Applying the community detection Chinese Whisper algorithm [3], we obtain 68 communities whose most salient are represented on the figure. The purple community, that gathers 177 authors including Katharina, is very dense (1096 edges). It brings together well identified researchers in data mining, machine learning and data bases. The other main communities are labeled on the graph. Our goal is to analyse this graph with regard to several questions:

– Are there any interesting patterns among publications?

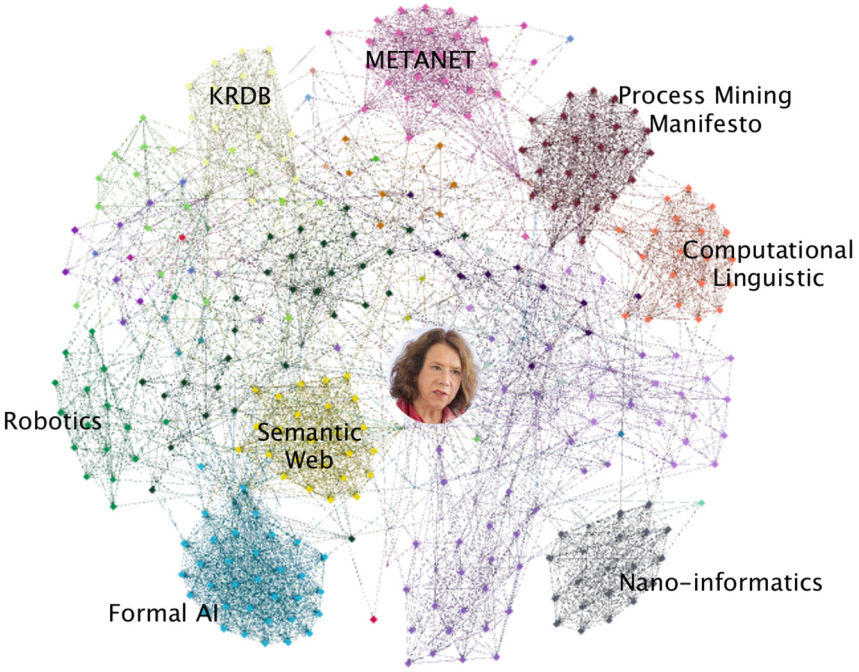

Fig. 1. Research domains associated to Katharina's co-authors. (Color figure online)

Table 1. Emerging patterns w.r.t. morik_number$^-$.

Pattern	Top 20
IJCAI$^+$, KI$^+$, GWAI$^+$, Informatik_Spektrum$^+$, morik_number$^-$	Katharina Morik, Wolfgang Wahlster, Bernhard Nebel, Thomas Christaller, Wolfgang Hoeppner, Jörg H. Siekmann, Günther Görz, Frank Puppe, Udo Hahn, Hans-Hellmut Nagel, Franz Baader, Christopher Habel, Bernd Neumann, Ulrich Furbach, Joachim Hertzberg
IJCAI$^+$, ICML$^+$, Machine_Learning$^+$, Knowl._Inf._Syst.$^+$, Data_Min._Knowl._Discov.$^+$, morik_number$^-$	Katharina Morik, Wray L. Buntine, Kristian Kersting, Floriana Esposito, Xindong Wu, Eamonn J. Keogh, Zhi-Hua Zhou, Siegfried Nijssen, Hiroshi Motoda, João Gama, Jie Tang, Salvatore J. Stolfo, Dacheng Tao, Michael J. Pazzani, Wei Liu, Chris H. Q. Ding, Tao Li, Bin Li

– Are there interesting trends between some authors' publications and topological properties?
– What about Katharina's role in this graph? Can we characterize the proximity to Katharina in terms of co-authorship?

6.2 Most Emerging Pattern with Respect to Morik Number

Table 1 presents two interesting patterns that strongly emerge with the Morik number. The first pattern gathers 4 conferences that are positively signed and the Morik number that is negatively signed: The more authors are close to Katharina, the more they publish in IJCAI as well as in three other German conferences (KI - Künstliche Intelligenz, GWAI - German workshop on artificial intelligence and Informatik_Spektrum) Notice that GWAI changes its name to KI in 1993. The top 20 supporting authors gathers the German researchers in Artificial Intelligence. They are close to Katharina who is wellknown in the AI community research, and she also actively contributes to the animation of her national community.

The second pattern presented in Table 1 gathers the major conference venues and journals in Artificial Intelligence, Data Mining and Machine Learning. The top 20 supporting authors are all well established researchers in these research areas.

The first pattern with the Morik number positively signed is presented in Table 2. It gathers the conference ICASSP in signal processing that is positively

Table 2. Emerging pattern w.r.t. morik_number$^+$.

Pattern	Top 20
ICASSP$^+$, IJCAI$^-$, KR$^-$, KI$^-$, morik_number$^+$	Gyula Hermann, Victor Lazzarini, Joseph Timoney, Fred Kitson, Manuel Duarte Ortigueira, Abbas Mohammadi, Riwal Lefort, Jean-Marc Boucher, Artur Przelaskowski, Kenichi Miyamoto, Emiru Tsunoo, Olaf Schreiner, Murtaza Taj, Salim Chitroub, Saptarshi Das, Ales Procházka, Amrane Houacine, Yasuyuki Ichihashi, Pablo Javier Alsina, Valeri Mladenov

Table 3. Emerging patterns w.r.t. Morik number that mix vertex and topological attributes.

Pattern	Top 20
ICASSP$^+$, IJCAI$^-$, Degree$^-$, Closeness$^-$, Betweennes$^-$, NetworkConstraint$^+$, morik_number$^+$	Jacob Ninan, Marc Beacken, Hinrich R. Martens, Jyun-Jie Wang, William H. Haas, J. G. Cook, Lawrence J. Ziomek, José R. Nombela, T. J. Edwards, Judith G. Claassen, Shigekatsu Irie, Alberto R. Calero, Takaaki Ueda, Hisham Hassanein, Peter Strobach, Liubomire G. Iordanov, N. A. M. Verhoeckx, Guy R. L. Sohie, Sultan Mahmood, Matt Townsend
KI$^+$, Degree$^+$, Closeness$^+$, NetworkConstraint$^-$, morik_number$^-$	Bernhard Nebel, Katharina Morik, Deborah L. McGuinness, Mark A. Musen, Rudi Studer, Steffen Staab, Hans W. Guesgen, Bamshad Mobasher, Simon Parsons, Thorsten Joachims, Alex Waibel, Kristian Kersting, Matthias Jarke, Manuela M. Veloso, Wolfgang Nejdl, Alfred Kobsa, Virginia Dignum, Alessandro Saffiotti, Hans Uszkoreit, Antonio Krüger

signed and 3 conferences in Machine Learning that are negatively signed: The farther the authors from Katharina, the more they published at ICASSP and the less they contribute to AI conferences IJCAI, KI and KR (Principles of knowledge representation and reasoning). The support of this pattern is rather low.

The most emerging patterns w.r.t. Morik number that mix vertex and topological attributes are presented in Table 3. The first pattern is similar to the pattern of Table 2 and the additional topological attributes corroborate the eccentricity of the pattern relative to the graph. The second pattern brings together confirmed researchers in artificial intelligence and machine learning, who have

Table 4. Emerging patterns involving the French-speaking data mining conference EGC.

Pattern	Top 20
EGC$^+$, Data_Min._Knowl._Discov.$^+$, morik_number$^-$	Katharina Morik, Bart Goethals, Céline Robardet, Didier Dubois, Michèle Sebag, Luc De Raedt, Mohammed Javeed Zaki, Einoshin Suzuki, Heikki Mannila, Jian Pei, Élisa Fromont, Toon Calders, Adriana Prado, Gilles Venturini, Szymon Jaroszewicz, João Gama, Alice Marascu, Osmar R. Zaïane, Pascal Poncelet, Jean-François Boulicaut
EGC$^+$, Knowl._Inf._Syst.$^+$, Data_Min._Knowl._Discov.$^+$, morik_number$^-$	Katharina Morik, Bart Goethals, João Gama, Mohammed Javeed Zaki, Jian Pei, Heikki Mannila, Osmar R. Zaïane, Toon Calders, Szymon Jaroszewicz, Einoshin Suzuki, Pascal Poncelet, Christophe Rigotti, Jean-François Boulicaut, Marie-Christine Rousset, Maguelonne Teisseire, Florent Masseglia, Gregory Piatetsky-Shapiro
EGC$^+$, Knowl._Inf._Syst.$^+$, morik_number$^-$	Katharina Morik, Bart Goethals, Fosca Giannotti, Mohand-Said Hacid, Toon Calders, Mohammed Javeed Zaki, Osmar R. Zaïane, Heikki Mannila, João Gama, Dominique Laurent, Jian Pei, Szymon Jaroszewicz, Einoshin Suzuki, Patrick Gallinari, David Genest, Mohand Boughanem, François Scharffe, Marc Plantevit, Laure Berti-Equille, Zbigniew W. Ras

published at Künstliche Intelligenz. They are very central in the graph and their neighborhood is not so much connected.

6.3 Where Are We in Katharina's Network? An Interactive Exploration of the Patterns

After considering the patterns that maximize the growth rate w.r.t. Morik number, we now look for patterns supported by the authors of this paper. Many of these patterns involve the French-speaking conference EGC (see Table 4) and journals in data mining. This is due to the fact that Katharina gave a keynote at EGC in 2009. The top 20 supporting authors are either French or prestigious invited speakers at this conference.

The first pattern of Table 5 can be interpreted thanks to a Dagstuhl seminar organized by Katharina called *Local Pattern Detection*. The goal of this seminar was to bring together prominent European researchers in the field of local pattern discovery. The Data Mining and Knowledge Discovery journal is the most important one that publishes results in that area. The second one is around

Table 5. Patterns related to Dagstuhl seminars.

Pattern	Top 20
LocalPatternDetection^{+}, Data_Min._Knowl._Discov.$^{+}$, morik_number^{-}	Katharina Morik, Stefan Rüping, Francesco Bonchi, Niall M. Adams, Marko Grobelnik, David J. Hand, Dunja Mladenic, Frank Höppner, Saso Dzeroski, Einoshin Suzuki, Nada Lavrac, Jean-François Boulicaut, Myra Spiliopoulou, Ruggero G. Pensa, Johannes Fürnkranz, Filip Zelezny
Parallel_Universes_and_Local_Patterns^{+}, morik_number^{-}	Katharina Morik, Arno Siebes, Michael R. Berthold, Michael Wurst, David J. Hand, Bernd Wiswedel, Frank Höppner, Emmanuel Müller, Élisa Fromont, Claus Weihs, Niall M. Adams, Mirko Böttcher, Ralph Krieger, Bruno Crémilleux, Ira Assent, Marie-Odile Cordier, Thomas Seidl, Heike Trautmann, Rene Quiniou, Arnaud Soulet

the seminar *Parallel Universes and Local Patterns* that was also organized by Katharina and colleagues.

7 Conclusion

We have been using an algorithm that supports network analysis by finding regularities among vertex topological properties and attributes. It mines frequent topological patterns as up and down co-variations involving both attributes and topological properties of graph vertices. In addition, we defined two interestingness measures to capture the significance of a pattern with respect to either a given descriptor, or the relationship encoded by the graph edges. Furthermore, by identifying the top k representative vertices of a topological pattern, we support a better interaction with end-users. While [25] has given details about the whole methodology and has sketched several case studies, we decided in this chapter to analyze co-authorship network of our colleague Katharina Morik. We have shown that it supports the discovery of sensible patterns.

Acknowledgments. We thank Adriana Prado for her help. We also gratefully acknowledge support from the CNRS/IN2P3 Computing Center.

References

1. Akoglu, L., Tong, H., et al.: PICS: parameter-free identification of cohesive subgroups in large graphs. In: SIAM DM, pp. 439–450 (2012)
2. Albert, R., Barabási, A.L.: Topology of complex networks: local events and universality. Phys. Rev. **85**, 5234–5237 (2000)
3. Biemann, C.: Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, pp. 73–80. Association for Computational Linguistics (2006)
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. **30**(1–7), 107–117 (1998)
5. Bringmann, B., Nijssen, S.: What is frequent in a single graph? In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 858–863. Springer, Heidelberg (2008)
6. Calders, T., Goethals, B., Jaroszewicz, S.: Mining rank-correlated sets of numerical attributes. In: KDD, pp. 96–105 (2006)
7. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-MAT: a recursive model for graph mining. In: SIAM SDM (2004)
8. Cheng, H., Zhou, Y., Yu, J.X.: Clustering large attributed graphs. TKDD **5**(2), 12 (2011)
9. Do, T., Laurent, A., Termier, A.: Efficient parallel mining of closed frequent gradual itemsets. In: IEEE ICDM, pp. 138–147 (2010)
10. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: KDD, pp. 43–52 (1999)
11. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)
12. Fürnkranz, J., Knobbe, A.J.: Guest editorial: global modeling using local patterns. DMKD **21**, 1–8 (2010)
13. Ge, R., Ester, M., Gao, B.J., et al.: Joint cluster analysis of attribute data and relationship data. TKDD **2**(2), 1–35 (2008)
14. Günnemann, S., et al.: Subspace clustering meets dense subgraph mining: a synthesis of two paradigms. In: IEEE ICDM, pp. 845–850 (2010)
15. Günnemann, S., et al.: A density-based approach for subspace clustering in graphs with feature vectors. In: PKDD, pp. 565–580 (2011)
16. Jiang, D., Pei, J.: Mining frequent cross-graph quasi-cliques. ACM TKDD **2**(4), 1–42 (2009)
17. Kang, U., Tsourakakis, C.E., Appel, A.P., Faloutsos, C., Leskovec, J.: Hadi: mining radii of large graphs. ACM TKDD **5**(2), 8 (2011)
18. Khan, A., Yan, X., Wu, K.L.: Towards proximity pattern mining in large graphs. In: SIGMOD, pp. 867–878 (2010)
19. Liao, Z.X., Peng, W.C.: Clustering spatial data with a geographic constraint. Knowl. Inf. Syst. **31**, 1–18 (2012)
20. Liu, G., Wong, L.: Effective pruning techniques for mining quasi-cliques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 33–49. Springer, Heidelberg (2008)
21. Makino, K., Uno, T.: New algorithms for enumerating all maximal cliques. In: Hagerup, T., Katajainen, J. (eds.) SWAT 2004. LNCS, vol. 3111, pp. 260–272. Springer, Heidelberg (2004)

22. Morik, K., Boulicaut, J.-F., Siebes, A. (eds.): Local Pattern Detection. LNCS (LNAI), vol. 3539. Springer, Heidelberg (2005)
23. Moser, F., Colak, R., Rafiey, A., Ester, M.: Mining cohesive patterns from graphs with feature vectors. In: SIAM SDM, pp. 593–604 (2009)
24. Mougel, P.N., Rigotti, C., Gandrillon, O.: Finding collections of k-clique percolated components in attributed graphs. In: PAKDD (2012)
25. Prado, A., Plantevit, M., Robardet, C., Boulicaut, J.-F.: Mining graph topological patterns: finding covariations among vertex descriptors. IEEE Trans. Knowl. Data Eng. **25**(9), 2090–2104 (2013)
26. Sese, J., Seki, M., Fukuzaki, M.: Mining networks with shared items. In: CIKM, pp. 1681–1684 (2010)
27. Silva, A., Meira, W., Zaki, M.: Structural correlation pattern mining for large graphs. In: Workshop on Mining and Learning with Graphs (2010)
28. Silva, A., Meira, W., Zaki, M.J.: Mining attribute-structure correlated patterns in large attributed graphs. PVLDB **5**(5), 466–477 (2012)
29. Uno, T.: An efficient algorithm for solving pseudo clique enumeration problem. Algorithmica **56**(1), 3–16 (2010)
30. Wang, D.J., Shi, X., McFarland, D.A., Leskovec, J.: Measurement error in network data: a re-classification. Soc. Netw. **34**(4), 396–409 (2012)
31. Zaki, M.J.: Scalable algorithms for association mining. IEEE Trans. Knowl. Data Eng. **12**(3), 372–390 (2000)
32. Zhou, Y., Cheng, H., Yu, J.: Graph clustering based on structural/attribute similarities. PVLDB **2**(1), 718–729 (2009)