Structural Correlation Pattern Mining for Large Graphs

Arlei Silva, Wagner Meira Jr.

[1]Computer Science Department – Universidade Federal de Minas Gerais
{arlei,meira}@dcc.ufmg.br

Abstract. *How are personal interests related to the communities in large social networks? In order to answer this kind of question, we introduce structural correlation pattern mining, which is the identification of interesting associations between vertex attributes and dense subgraphs. We present a model and algorithms that explore search, pruning, sampling and parallelization strategies to solve this problem for large graphs. Results show that structural correlation pattern mining enables the discovery of relevant patterns in real-life datasets.*

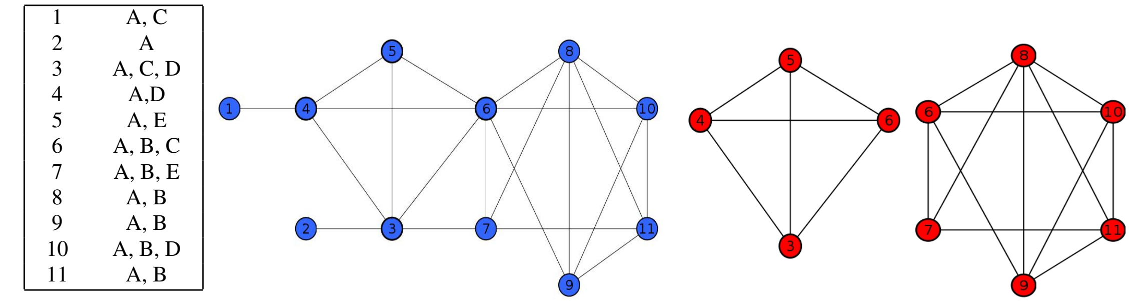
1. Introduction

Graphs, or *networks*, have been established as a powerful theoretical framework for modeling interactions in various scenarios. The availability of real graphs in the last years motivated a broad spectrum of research on the properties of such graphs. As one of these efforts, this work extends the graph analysis framework by studying correlations between vertex attributes and dense subgraphs in attributed graphs [Silva 2011].

Attributes play an important role in several real graphs as means to describe properties of vertices. In social networks, attributes are useful to represent personal characteristics (e.g., interests). Moreover, it is known that many graphs present dense subgraph organization (a.k.a. community structure) [Fortunato 2010]. Dense subgraphs are sets of vertices with strong connections among themselves (e.g., communities in social networks). Both attributes and dense subgraphs contain meaningful information in other important graphs, such as those extracted from citations, biological systems, and the Web. However, existing techniques are not able to extract knowledge regarding the correlation between attributes and dense subgraphs, which is the main motivation for this work.

Two vertices are correlated in terms of an attribute if they are connected and share this attribute [Anagnostopoulos et al. 2008]. The novelty of this work comes from extending the concept of correlation to subgraphs. A subgraph is said to be structurally correlated w.r.t. a set of attributes if it is dense and all of its vertices share these attributes.

Figure 1 illustrates structural correlation pattern mining. Vertex attributes and interactions are shown in Table 1(a) and Figure 1(b), respectively. Figures 1(c) and 1(d) are examples of dense subgraphs, ({A},{3,4,5,6}) and ({A,B},{6,7,8,9,10,11}) are examples of patterns. We say that the structural correlation of A is 0.82 because 82% of the vertices that have A are part of a pattern that has A as attribute. If the attributes and interactions in this example represent personal interests and friendship, respectively, a structural correlation pattern describes a community that shares a particular set of interests (e.g., sports, music). Moreover, structural correlation is a measure of the association between interests and the community structure in a social network, which is relevant for viral marketing. Structural correlation patterns may also represent relationships between gene expression and modules in gene networks, keywords and link structure in the Web, etc. At a large



1	A, C
2	A
3	A, C, D
4	A,D
5	A, E
6	A, B, C
7	A, B, E
8	A, B
9	A, B
10	A, B, D
11	A, B

(a) Vertex attributes (b) Graph (c) Dense subgraph (d) Dense subgraph

Figure 1. Structural correlation pattern mining (illustrative example)

scale, such patterns help us in the discovery of underlying processes that involve attributes and structure taking place in real large attributed graphs.

The analysis of relationships between attributes and subgraphs is a challenging problem. It requires a powerful model in order to enable the identification of relevant associations between these two types of information. Also, the number of possibilities in which attributes and subgraphs can be correlated in large graphs imposes a significant performance requirement. In Computer Science, the research area dedicated to the extraction of knowledge from large databases is called data mining. In particular, the sub-area of data mining focused on graphs is called graph mining [Chakrabarti and Faloutsos 2006].

We model the correlation between attributes and dense subgraphs using frequent itemsets and quasi-cliques, two previously defined patterns in data mining. Based on this model, we formulate a statistical significance measure for structural correlation that gives how unexpected it is to find a given correlation in a graph. Furthermore, we design a family of algorithms that apply search, pruning, sampling and parallelization techniques as means to make structural correlation pattern mining applicable to real large databases.

The main contributions of this work are: (1) introducing the problem of correlating attributes and dense subgraphs, (2) modeling it as a graph mining problem, (3) designing efficient algorithms for this problem, and (4) evaluating the relevance of the proposed problem and algorithms in real scenarios [Silva 2011, Silva et al. 2010, Silva et al. 2012].

2. Structural Correlation Pattern Mining

This section describes the structural correlation pattern mining and some solutions for it.

2.1. Definitions

Definition 1 *(**Attributed graph**) An attributed graph is a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{F})$ where $\mathcal{V}$ is the set of vertices, $\mathcal{E}$ is the set of edges, $\mathcal{A} = \{a_1, a_2, \dots a_n\}$ is the set of attributes, and $\mathcal{F} : \mathcal{V} \rightarrow P(\mathcal{A})$ returns the set of attributes of a vertex. P is the power set function.*

Definition 2 *(**Attribute set**) An attribute set S is a subset of $\mathcal{A}$. We denote by $\mathcal{V}(S) \subseteq \mathcal{V}$ the vertex set induced by S (i.e., $\mathcal{V}(S) = \{v_i \in \mathcal{V} | S \subseteq \mathcal{F}(v_i)\}$) and by $\mathcal{E}(S) \subseteq \mathcal{E}$ the edge set induced by S (i.e., $\mathcal{E}(S) = \{(v_i, v_j) \in \mathcal{E} | v_i, v_j \in \mathcal{V}(S)\}$). The graph $\mathcal{G}(S)$, is the pair $(\mathcal{V}(S), \mathcal{E}(S))$. A support function σ gives the frequency of an attribute set in the graph ($\sigma(S) = |\mathcal{V}(S)|$), i.e., the number of vertices that contain S.*

pattern	size	γ	σ	ϵ
$(\{A\},\{6,7,8,9,10,11\})$	6	0.60	11	0.82
$(\{A\},\{3,4,5,6\})$	4	1	11	0.82
$(\{A\},\{3,4,6,7\})$	4	0.67	11	0.82
$(\{A\},\{3,5,6,7\})$	4	0.67	11	0.82
$(\{A\},\{3,6,7,8\})$	4	0.67	11	0.82
$(\{B\},\{6,7,8,9,10,11\})$	6	0.60	6	1.0
$(\{A,B\},\{6,7,8,9,10,11\})$	6	0.60	6	1.0

Figure 2. Example output: For each pattern (S,Q), we show its size, density (γ), support (σ), and structural correlation (ϵ).

```
Require: $\mathcal{G}(S)$, $\gamma_{min}$, $min\_size$
Ensure: $\mathcal{Q}$
  $\mathcal{Q} \leftarrow \emptyset$
  $X \leftarrow \emptyset$
  $candExts(X) \leftarrow \mathcal{V}(S)$
  Apply vertex pruning in $candExts(X)$
  $qcCands \leftarrow \{(X, candExts(X))\}$
  while $qcCands \neq \emptyset$ do
    $q \leftarrow qcCands.get()$
    Apply candidate quasi-clique pruning in $q$
    if $q.X \cup q.candExts(X)$ is a quasi-clique then
      $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q.X \cup q.candExts(X)\}$
    else
      if $q.X$ is a quasi-clique then
        $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q.X\}$
      insert extensions of $q$ into $qcCands$
```

Figure 3. Structural Correlation Algorithm

Definition 3 *(**Quasi-clique**) Given a minimum density γ_{min} ($0 < \gamma_{min} \leq 1$) and min_size, a quasi-clique is a maximal vertex set Q such that, for each $v \in Q$, the degree of v in Q is at least $\lceil \gamma_{min}.(|Q|-1) \rceil$ and $|Q| \geq min\_size$ [Liu and Wong 2008].*

Definition 4 *(**Structural correlation pattern**). A structural correlation pattern is a pair (S,Q), where $S \subseteq \mathcal{A}$ and $Q \subseteq \mathcal{V}(S)$ is a quasi-clique, given γ_{min} and min_size.*

Definition 5 *(**Structural correlation function** ϵ) Given an attribute set S, the structural correlation of S, $\epsilon(S)$, is given as the ratio between $|\mathcal{K}_S|$, where $\mathcal{K}_S$ is the set of vertices in quasi-cliques in $\mathcal{G}(S)$, and $\sigma(S)$, which is the support of S.*

2.2. The structural correlation pattern mining problem

Definition 6 *(**Structural correlation pattern mining problem**). Given an attributed graph $\mathcal{G}(\mathcal{V},\mathcal{E},\mathcal{A},\mathcal{F})$, a minimum support threshold σ_{min}, a minimum quasi-clique density γ_{min} and size min_size, and a minimum structural correlation ϵ_{min}, the problem consists of identifying the set of structural correlation patterns (S,Q) from $\mathcal{G}$, such that $\sigma(S) \geq \sigma_{min}$, and $\epsilon(S) \geq \epsilon_{min}$.*

Figure 2 shows the patterns from the graph presented in Figure 1 when the parameters σ_{min}, γ_{min}, min_size and ϵ_{min} are set to 3, 0.6, 4, and 0.5, respectively.

2.3. Statistical significance of structural correlation

Structural correlation is a measure of the association between attribute sets and dense subgraphs. However, given a structural correlation value, how interesting/unexpected is it? We answer this question by formulating an upper bound on the expected structural correlation in a hypothetical scenario where such correlation is random.

Theorem 1 *(**Upper bound on the expected structural correlation**) Given the parameters γ_{min} and min_size, the structural correlation of an attribute set S is upper bounded by: $max\text{-}\epsilon_{exp}(\sigma(S)) = \sum_{\alpha=z}^{m} p(\alpha).\sum_{\beta=z}^{\alpha} F(\alpha,\beta,\rho)$, where $z = \lceil \gamma_{min}.(min\_size-1) \rceil$, m is the maximum degree of a vertex from $\mathcal{G}$, and p is the degree distribution of $\mathcal{G}$. The values of ρ and F are defined as follows: $\rho = \frac{\sigma(S)-1}{|\mathcal{V}|-1}$, $F(\alpha,\beta,\rho) = \binom{\alpha}{\beta}.\rho^{\beta}.(1-\rho)^{\alpha-\beta}$.*
Proof. *Omitted due to space constraints, please see [Silva 2011].*

Definition 7 *(**Normalized structural correlation** δ) The normalized structural correlation of an attribute set S is the ratio between the actual (ϵ) and the upper bound on the expected (max-ϵ) structural correlation.*

2.4. SCPM: A family of algorithms for structural correlation pattern mining

The design of efficient algorithms for structural correlation pattern mining is one of the contributions of this work. We combine existing and new strategies into a family of algorithms called SCPM as means to solve this problem for large graphs. Figure 3 shows a general structural correlation algorithm, which extends and prunes quasi-clique candidates until the complete set of quasi-cliques is discovered. In the remaining of this section, we present several computational strategies based on this general algorithm.

Search: Given an attribute set S, its structural coverage $\mathcal{K}(S)$ can be computed using the set of quasi-cliques $\mathcal{Q}$ (Figure 3). However, candidate patterns can be traversed in different fashions. In particular, in case $qcCands$ is a queue or a stack, patterns are traversed in BFS or DFS order, respectively. While BFS represents a focus on a large number of small quasi-cliques, DFS tends to find a small number of large ones. We found that DFS achieves better performance in structural correlation pattern mining.

Pruning: We propose several strategies to prune candidate patterns in structural correlation pattern mining. Candidates can be pruned based on: (1) a level-wise enumeration of attribute sets, (2) minimum structural correlation ϵ_{min}, and (3) minimum normalized structural correlation δ_{min}. Also, by reducing the number of patterns discovered to the top-k most relevant ones, in terms of size and density, we are able to cut down the number of candidate patterns to be checked in the computation of the structural correlation.

Sampling: Instead of checking whether each vertex in $\mathcal{G}(S)$ is part of a quasi-clique, as shown in Figure 3, it is possible to estimate $\mathcal{K}(S)$ using random sampling. The margin of error associated with a sample can also be estimated using standard statistical techniques.

Parallelization: Structural correlation patterns may be mined in parallel. In this work, we propose the use of a work pool model as means to exploit multiple processing units in structural correlation pattern mining.

3. Experimental Results

In this section, we summarize the main experimental results of this work.

Datasets: We apply structural correlation pattern mining in the analysis of 3 real datasets: a collaboration, a music, and a citation network. Table 1 describes the datasets.

name	**vertex**	**edge**	**att.**	S	Q	$\|\mathcal{V}\|$	$\|\mathcal{E}\|$	$\|\mathcal{A}\|$
DBLP	author	co-authorship	term	topic	community	108,030	276,658	23,285
LastFm	user	friendship	artist	taste	community	272,412	350,239	3,929,101
Citeseer	paper	citation	term	topic	related work	294,104	782,147	206,430

Table 1. Dataset descriptions and statistics

S	σ	ϵ
grid applic	840	0.26
grid servic	599	0.23
environ grid	525	0.21
queri xml	615	0.21

(a) structural correlation

S	σ	ϵ	δ
search rank	420	0.19	635,349
perform file	404	0.14	555,067
structur index	404	0.14	555,067
search mine	413	0.14	490,932

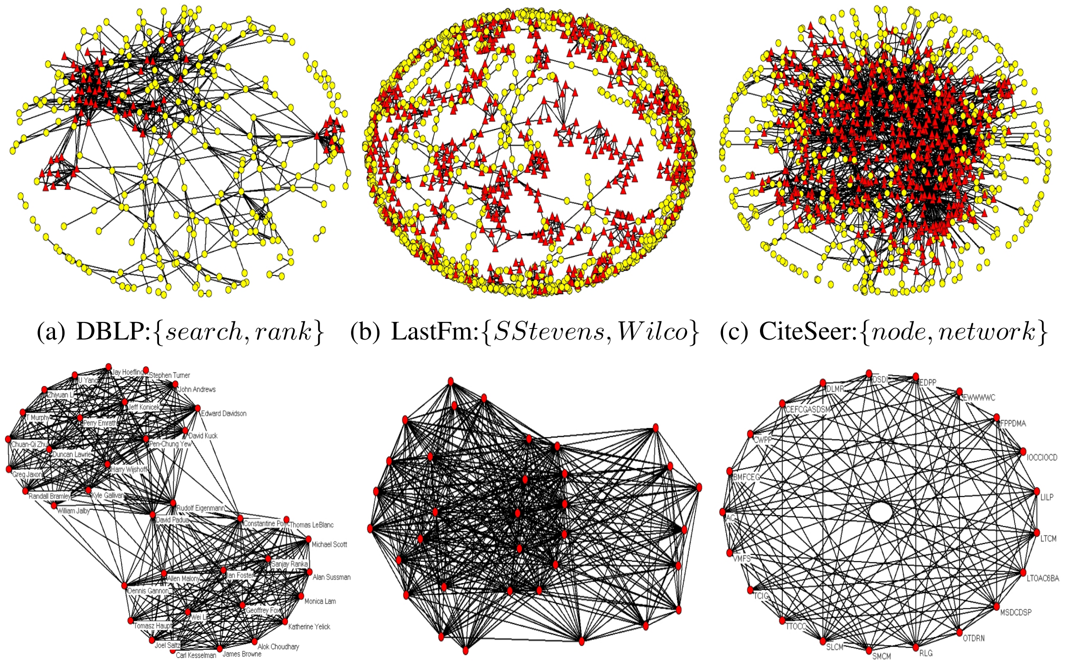
(b) normalized str. correlation

name	**search**	**par.**	**samp.**
SCPM-DFS	DFS	no	no
SCPM-BFS	BFS	no	no
SCPM-DFS-SAMP	DFS	no	yes
SCPM-BFS-SAMP	BFS	no	yes
PAR-SCPM-DFS	DFS	yes	no

(c) Variations of SCPM

Figure 4. Top attribute sets from DBLP and variations of SCPM evaluated

Case studies: Figures 4(a) and 4(b) show the top-4 attribute sets of size at least 2 in terms of ϵ and δ from the DBLP dataset (σ_{min}=10,γ_{min}=0.5,min_size=400). Top ϵ attribute sets are more frequent and general than the top δ. On the other hand, top δ attribute sets can be easily associated to research topics that emerge from collaboration due to the use of statistical significance in structural correlation pattern mining. Figure 5 presents examples of graphs ($\mathcal{G}(S)$) induced by attribute sets and structural correlation patterns discovered from the three datasets. Vertices in dense subgraphs are indicated (in red). Structural correlation patterns depict the dense subgraph structure associated to attribute sets. In Figure 5(b), for instance, it is possible to identify several communities involving people who listen to Sufjan Stevens and Wilco. Such information is very useful in the design of effective viral campaigns in *last.fm*. Figure 5(d) shows a structural correlation pattern that describes an intense collaboration between 37 researchers who work on topics related to *systems performance*. Further analysis has shown that these collaborations are a consequence of two research projects with several overlapping members. This kind of pattern is of particular interest to organizations (e.g., companies) in the search for teams of experts in a given field. Neither attribute or dense subgraph information in isolation could enable the discovery of the patterns that we have discussed here.



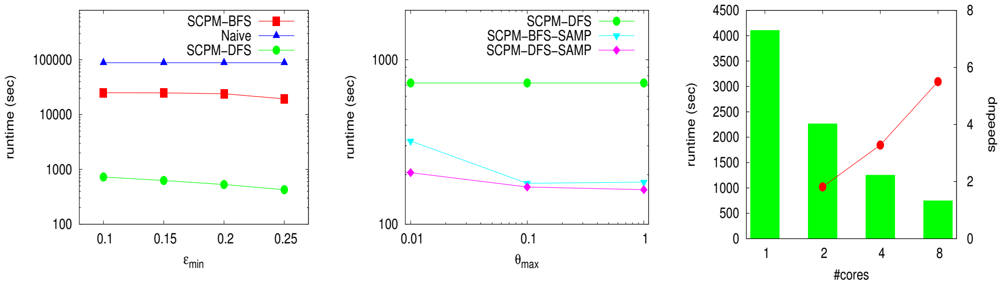
(a) DBLP:$\{search, rank\}$ (b) LastFm:$\{SStevens, Wilco\}$ (c) CiteSeer:$\{node, network\}$

(d) DBLP:$\{perform, system\}$ (e) LastFm:$\{VMorrison\}$ (f) CiteSeer:$\{perform, system\}$

Figure 5. Examples of induced graphs (a,b,c) and patterns (d,e,f)

Performance evaluation: Figure 6 shows the performance of 4 versions of SCPM w.r.t. some of its parameters using data from DBLP. The versions are described in Table 4(c) in terms of the search strategy (search) used and whether parallelization (par.) and sampling (samp.) are applied. SCPM-DFS outperforms SCPM-BFS and is up to 100 times faster than the Naive algorithm, which is a direct application of a frequent itemset and a quasi-clique mining algorithm. Moreover, the pruning strategy based on ϵ leads to performance improvements. Figure 6(b) shows that sampling (SCPM-DFS-SAMP and SCPM-BFS-SAMP) enables performance gains up to 70% over SCPM-DFS with 1% margin of error

(θ_{max}). Scalability results (see Figure 6(c)) show that PAR-SCPM-DFS is up to 6 times faster than SCPM-DFS when 8 cores are available.



(a) ϵ_{min} (b) θ_{max} (c) PAR-SCPM-DFS: speedup

Figure 6. Performance evaluation

4. Concluding Remarks

We summarized the main contributions of our research on models and algorithms for assessing the correlation between vertex attributes and dense subgraphs in attributed graphs. We formalized this problem as the structural correlation pattern mining problem and presented several strategies in order to enable its application to large graphs. After an extensive evaluation of a family of algorithms (SCPM) for this problem, we showed that the proposed model and strategies not only provide relevant knowledge, but also are able to process large databases (with hundreds of thousands of vertices, edges and attributes). In particular, structural correlation patterns can be mined efficiently using DFS, vertex and attribute set pruning, restriction to the top-k patterns, sampling, and parallelization.

As future work, we will apply SCPM to relational learning, social network analysis, and summarization tasks. Due to the lack of space, we omitted some results of this research (please see [Silva et al. 2010, Silva et al. 2012, Silva 2011]).

References

Anagnostopoulos, A., Kumar, R., and Mahdian, M. (2008). Influence and correlation in social networks. In *SIGKDD*, pages 7–15. ACM.

Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1).

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Liu, G. and Wong, L. (2008). Effective pruning techniques for mining quasi-cliques. In *ECML/PKDD*, pages 33–49. Springer-Verlag.

Silva, A. (2011). *Structural correlation pattern mining for large graphs*. M.Sc Thesis, Computer Science Department, Universidade Federal de Minas Gerais.

Silva, A., Meira, Jr., W., and Zaki, M. J. (2010). Structural correlation pattern mining for large graphs. In *MLG*, pages 119–126. ACM.

Silva, A., Meira, Jr., W., and Zaki, M. J. (2012). Mining attribute-structure correlated patterns in large attributed graphs. *Proc. of the VLDB Endowment*, 5(5):466–477.