Structural Correlation Pattern Mining for Large Graphs

Arlei Silva*
Universidade Federal de Minas Gerais
Computer Science Dep.
Belo Horizonte, Brazil
arlei@dcc.ufmg.br

Wagner Meira Jr.
Universidade Federal de Minas Gerais
Computer Science Dep.
Belo Horizonte, Brazil
meira@dcc.ufmg.br

Mohammed J. Zaki
Rensselaer Polytechnic Institute
Computer Science Dep.
Troy, NY
zaki@cs.rpi.edu

ABSTRACT

In this paper we define the Structural Correlation Pattern (SCP) mining problem, which consists of determining correlations among vertex attributes and dense components in an undirected graph. Vertex attributes play an important role in several real-life graphs and SCPs help to understand how they relate to the associated graph topology. SCPs may describe, for example, interesting relationships between personal characteristics and the community structure in social networks. We also propose an efficient algorithm, called SCORP, to extract SCPs from large graphs, and compare it against a naive approach for SCP mining, demonstrating its scalability and efficiency. We also discuss the application of SCORP to two actual scenarios, co-authorship networks and social music discovery, showing relevant results that demonstrate the applicability of the proposed approach.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Graph mining, quasi-cliques, correlation

1. INTRODUCTION

Graphs, or networks, have been established as a powerful theoretical framework for modeling several types of interactions in a variety of scenarios, such as social networks, metabolic networks, food webs, distribution networks, among others [11]. The availability of large real graphs has motivated a wide variety of research on the properties of such graphs. Moreover, the combination of new algorithms and powerful hardware have enabled the discovery of complex and interesting patterns in large graphs.

Finding correlations between attributes has been a long term research problem in data mining, and represents a challenge when we talk about graphs. One particular problem that we are interested in is identifying relevant *structural correlation patterns* (SCPs) on attributed graphs, that is, graphs where each vertex is associated with one or more attributes. A correlation pattern accounts for the co-occurrence of the same vertex attributes in densely connected subgraphs.

Most of the existing research work on graphs is focused on the study of network topological structure. However, in several important real life graphs, vertex attributes play an important role, as illustrated in Figure 1a, which contains 10 vertices (1-10) and 5 attributes (A, B, C, D, and E).

For instance, in a social network, vertices are individuals, edges are social relationships (e.g., friendship, marriage), and vertex properties can describe personal characteristics (e.g., sex, age, location) or behaviors (e.g., favorite videos, tags used, items bought). In a co-authorship network, potential vertex information includes affiliation and topics of interest, among others. Further, attributed graphs can be used to represent biological data. In a gene network, for example, vertices are genes, edges are protein-protein or genetic interactions, and vertex properties are gene expression or annotation data.

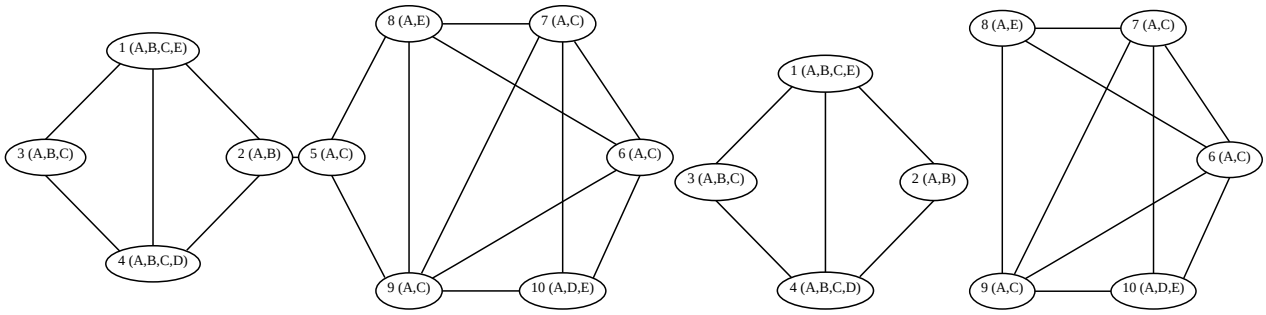
Recent graph clustering algorithms [14, 6] based on both structural and attribute similarities have shown that attribute and relationship data may be exploited effectively as complementary information in several scenarios (e.g., market segmentation and community discovery). Moreover, the problem of identifying dense subgraphs with homogeneous features, proposed in [10], has shown interesting applications in social and biological network analysis. Nevertheless, understanding the correlations among vertex attributes and the graph topology still remains an unexplored problem.

Previous work[3] has defined as *social correlation* the occurrence of a particular event for two adjacent nodes in a social network. Such event may be, for example, purchasing a product, visiting a web-site or using a particular tag. In this work we generalize this concept beyond social networks for undirected graphs in general, defining *structural correlation in graphs*. Two vertices v and u are structurally correlated in terms of an attribute α if they are adjacent and both have α as attribute. In Figure 1a, vertices 1 and 2 are structurally correlated in terms of the attribute A, but not in terms of E.

In attributed graphs, pairs of vertices may be correlated in terms of multiple attributes. In the graph shown in Figure 1a, for example, vertices 1 and 3 are correlated in terms of the attributes A, B, and C. Multi-attribute structural correlations are useful for discovering more complex relationships between vertex attributes and graph topology patterns. One example of a relevant topology pattern is a densely connected set of vertices in a graph. In Figure 1a, for example, every pair of vertices containing A, B, and C as

*Work done while the author was a visiting scholar at Rensselaer Polytechnic Institute.

MLG '10, July 24-25, 2010 Washington, DC, USA



(a) G (b) $(\{A, B\}, \{1, 2, 3, 4\})$ (c) $(\{A\}, \{6, 7, 8, 9, 10\})$

Figure 1: Example graph and SCPs

attributes are adjacent in the graph, which is an interesting relation between this set of attributes and the graph topology. Given an attribute set S and a densely connected subgraph V composed of vertices that have all attributes from S, we define the pair (S, V) as a *Structural Correlation Pattern* (SCP).

SCPs are related to important correlation patterns in social networks, such as homophily and social influence. Such patterns may describe how particular personal characteristics are associated with the existence of communities in social networks, where homophily plays an important role [9]. Moreover, the spread of behaviors from influencers [3], the propagation of information [5], and the diffusion of innovations [4] in social networks can also generate structural correlation patterns. SCPs can be also applied in the study of correlations between gene expressions and interactions in gene networks.

In this paper we define the *structural correlation pattern mining* problem and present algorithms for discovering which sets of attributes are associated with structural correlation patterns and listing the set of SCPs from a graph G.

Since an SCP requires densely connected subgraphs, it is necessary to define them. In the graph mining literature, dense subgraphs are known as clusters (i.e., communities). Although the graph clustering problem has been extensively studied in the past years, there is not an accepted single definition for a cluster in a graph. In general, a cluster is considered to be a set of vertices with high intra-connectivity and low inter-connectivity (i.e., vertices in the same cluster are highly connected and vertices in different clusters are less connected). However, there are several connectivity metrics in the literature (e.g., shortest path, number of common neighbors) and most of the graph clustering algorithms assume that each vertex is member of a single cluster, whereas it is desirable that a vertex be allowed to belong to multiple SCPs.

The most dense connected subgraph containing a set V of vertices is a clique. Every pair of vertices in a clique are adjacent. In the graph shown in Figure 1a, the vertices 6, 7, 8, and 9 form a clique. On the other hand, the least dense connected subgraph for a set V of vertices is a tree (there is a single path between each pair of vertices). While restricting dense connected subgraphs in SCPs to be cliques can be too strong a constraint, considering trees of vertices that share a given attribute set as a SCP is too broad. We provide an intermediate solution between these extremes using the concept of *quasi-clique* [12, 1].

A set of vertices V is called a γ-quasi-clique ($0 < \gamma \leq 1$) if every vertex in V is adjacent to at least $\gamma.|V - 1|$ vertices in V. Quasi-cliques have shown great applicability in data mining. An important property of quasi-cliques is that a vertex can be a member of more than one quasi-clique. In the graph shown in Figure 1a, the vertices 5, 6, 7, 8, and 9 form a 0.75-quasi-clique. Therefore, a SCP is a pair (S, V), where S is a set of attributes and V is a quasi-clique. Figure 1b shows the SCP ($\{A, B\}$, {1, 2, 3, 4}), for which $|V|$=4, $|S|$=2, and γ=0.66. Anther example of SCP, shown in Figure 1c, is ($\{A\}$, {6, 7, 8, 9, 10}), for which $|V|$=5, $|S|$=1, and γ=0.75.

In this paper we study the problem of structural correlation pattern mining. Our main contributions are: (1) we define a new graph mining problem, (2) we propose SCORP, an efficient algorithm for SCP mining, and (3) we present two interesting applications for SCP mining using real datasets.

2. RELATED WORK

Graphs have been applied as an expressive model to represent interaction data in several application domains [11]. Although most of the studies on graphs use simple representations, recent work has proposed considering vertex attributes as complementary information in graph mining tasks. Combining relationship and attribute data is suitable for scenarios where important information associated with vertices is available. [6] introduces the connected k-center (CkC) problem, which checks whether an attributed graph can be partitioned considering both attribute and relationship data. Since the CkC problem is NP-complete, the authors propose a constant factor approximation algorithm for the CkC problem and a efficient heuristic to solve the problem in large datasets. In [14], the authors propose a new graph clustering algorithm, called SA-Cluster, that applies a random walk based unified distance metric to an attribute-augmented graph. Vertices from the original graph are connected to attribute vertices that represent pairs <attribute, value>. The proposed algorithm adjusts edge weights in the augmented graph in order to maximize the clustering objective function.

To the best of our knowledge, the study most related to this paper is [10], where the authors introduce the problem of finding cohesive patterns (CPs). A cohesive pattern is a dense connected subgraph where vertices have homogeneous attributes (or features). However, CPs do not express how vertex attributes are related to the existence of dense components. We can clarify the difference between CPs and SCPs considering, for example, a social network, where nodes represent people, edges represent friendships, and vertex information are personal characteristics. In such networks, while a CP is a community that is homogeneous in terms of some characteristics, SCPs are pairs (community, characteristics) for which the given characteristics are strongly associated to the existence of communities. Moreover, the density criteria ap-

plied for CPs is the cliquishness, which is the fraction of the number of edges divided by the number of possible number of edges. Given a CP composed by a set of vertices V and a vertex $v \in V$, there is no guarantee of how many vertices in V are neighbors of v.

Quasi-cliques have been used as a suitable definition for dense subgraphs [1]. Three important properties of quasi-cliques motivated us to apply it as a density criteria: 1) easy definition, 2) possibility of overlaps, and 3) intra-pattern neighborhood guarantee. The problem of finding the maximum quasi-clique in a graph is NP-complete and the problem of counting the number of quasi-cliques in a graph is #P-complete [12]. In order to identify maximal quasi-cliques in large graphs, [1] proposes a GRASP (Greedy Randomized Adaptive Search Procedure) for the problem. It is important to notice that the the definition of quasi-clique applied by the authors is different from ours. Given a graph $G(V, E)$ a set of vertices $V' \subseteq V$ is considered to be a quasi-clique by them if V' is connected in G and the number of edges between vertices in V' is at least $\gamma.\binom{|V'|}{2}$. [12] introduces the problem of mining cross-graph quasi-cliques (i.e., set of vertices that are quasi-cliques in every graph from a graph database). The traditional downward-closure property, applied by several data mining algorithms, does not hold for quasi-cliques. Therefore, effective pruning strategies are essential to mine quasi-cliques from large graphs. The authors propose the Crochet algorithm, that applies vertex and candidate vertex set pruning techniques for the discovery of the complete set of cross-graph quasi-cliques from databases composed of several graphs. Crochet was further extended to Crochet+ in [7] to mine frequent cross-graph quasi-cliques, which is a general case of the problem presented in [12]. A quasi-clique is called frequent if it appears in a large enough number of graphs in the database. [13] studies the problem of mining frequent coherent closed quasi-cliques, a γ-quasi-clique is said to be coherent if $\gamma \geq 0.5$. The authors propose the Cocain algorithm, that introduces new pruning strategies for quasi-clique discovery.

While [12], [7], and [13] study the problem of quasi-clique discovery from a database composed of multiple graphs, which allow the application of specific pruning techniques, [8] studies the problem of finding the set of quasi-cliques in a single graph. Integrating several powerful vertex and candidate vertex set pruning strategies for quasi-clique mining, the authors propose the Quick algorithm. Since the definition of SCP is based on quasi-cliques in a single graph, we also apply the pruning techniques used by Quick. However, Quick does not consider any attribute information as we propose in this paper.

3. PROBLEM STATEMENT

An *attributed graph* is given as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{F})$ where $\mathcal{V}$ is the set of vertices, $\mathcal{E}$ is the set of edges, $\mathcal{A} = a_1, a_2, \ldots a_n$ is the set of vertex attributes, and $\mathcal{F} : \mathcal{V} \rightarrow \mathcal{P}(\mathcal{A})$ is a function that returns the set of attributes of a vertex. $\mathcal{P}$ is the power set function. Each vertex v_i in $\mathcal{V}$ has a set of attributes $\mathcal{F}(v_i) = \{a_{i1}, a_{i2}, \ldots a_{ip}\}$, where $p = |\mathcal{F}(v_i)|$ and $\mathcal{F}(v_i) \subseteq \mathcal{A}$.

Given the set of vertex attributes $\mathcal{A}$, we define an *attribute set* S as a subset of $\mathcal{A}$ ($S \subseteq \mathcal{A}$). Moreover, we denote by $\mathcal{V}(S) \subseteq \mathcal{V}$ the set of vertices where each vertex has all attributes in S (i.e., $\mathcal{V}(S) = \{v_i \in \mathcal{V} | S \subseteq \mathcal{F}(v_i)\}$). We define the *support* $\sigma(S) = |\mathcal{V}(S)|$ and S is called *frequent* if $\sigma(S) \geq \sigma_{min}$, where σ_{min} is a minimum support threshold. Moreover, we define $G(S)$ as the subgraph $G_s(V_s, E_s)$ induced by $\mathcal{V}(S)$, such that $V_s = \mathcal{V}(S)$ and $E_s = \{(v_i, v_j) \in \mathcal{E} | v_i \in \mathcal{V}(\mathcal{S}) \wedge v_j \in \mathcal{V}(S)\}$.

In a graph $G(V, E)$ (non-attributed), we denote by *γ-quasi-clique* a set of vertices V' such that every vertex $v_i \in V'$ is connected to, at least, $\gamma.(|V'| - 1)$ vertices in V' and no proper superset of V' has this property (i.e., *γ-quasi-cliques* are maximal). Since very small *γ-quasi-cliques* may not be of interest, a minimum size threshold min_size is specified. Only *γ-quasi-cliques* with at least min_size vertices are considered.

We extend the concept of *γ-quasi-cliques* to an attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{F})$ through the definition of *S-γ-quasi-cliques*. An *S-γ-quasi-clique* is a set of vertices V such that $V \subseteq \mathcal{V}(S)$ and V is a *γ-quasi-clique* in $G(S)$. We denote the set of vertices $v_i \in \mathcal{V}(S)$ that are members of at least one *S-γ-quasi-clique* as $V(S)_\gamma$, i.e., $V(S)_\gamma = \bigcup_{V \in \mathcal{Q}_S} V$, where $\mathcal{Q}_s$ is the set of *S-γ-quasi-cliques*. Moreover, we define by $\epsilon(S) = V(S)_\gamma / V(S)$ the *S-γ-quasi-clique membership function*, $0 \leq \epsilon(S) \leq 1$. The value of $\epsilon(S)$ represents how strongly the attribute set S is associated to the existence of *γ-quasi-cliques* in the graph $\mathcal{G}$. In other words, $V(S)$ gives the set of all vertices with attribute set S, whereas $V(S)_\gamma$ is the set of vertices with attribute set S and that belongs to at least one *S-γ-quasi-clique*.

DEFINITION 1. (STRUCTURAL CORRELATION PATTERN MINING). Given an attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{F})$, a minimum support threshold σ_{min}, a quasi-clique threshold γ_{min}, a minimum quasi-clique size min_size, and a minimum quasi-clique membership threshold ϵ_{min}, the structural correlation pattern mining task consists of identifying every pair (S,V), where S is an attribute set and V is a *S-γ_{min}-quasi-clique*, such that $\sigma(S) \geq \sigma_{min}$, $|V| \geq min\_size$, and $\epsilon(S) \geq \epsilon_{min}$.

Considering the graph in Figure 1a as an illustrating example and the parameters σ_{min}=4, γ_{min}=0.5, min_size=4, and ϵ_{min}=0.5, the set of SCPs is: $\{(\{A\}, \{1, 2, 3, 4\}), (\{A\}, \{5, 6, 7, 8, 9\}), (\{A\}, \{6, 7, 8, 9, 10\}), (\{B\}, 1, 2, 3, 4), (\{A, B\}, 1, 2, 3, 4)\}$ and $\epsilon(A)$=$\epsilon(B)$=$\epsilon(A, B)$=1. If we increase ϵ_{min} to 0.9, the resulting set of SCPs is: $\{(\{A\}, \{1, 2, 3, 4\}), (\{A\}, \{5, 6, 7, 8, 9\}), (\{A\}, \{6, 7, 8, 9, 10\})\}$.

4. STRUCTURAL CORRELATION PATTERN MINING

4.1 Quasi-clique Mining

Structural correlation patterns describe relations between sets of attributes and quasi-cliques in graphs. Therefore, the quasi-clique mining problem [12, 7, 13, 8] can be considered as part of the SCP mining problem. In this section we define and discuss important aspects of the quasi-clique mining problem.

THEOREM 1. (QUASI-CLIQUE MINING PROBLEM COMPLEXITY). Given a graph $G(V, E)$, the problem of counting the number of γ-quasi-cliques in G is #P-complete.

Proof sketch. As in [12], we prove it by restriction. If γ=1, the problem of counting the number of quasi-cliques is reduced to counting the number of cliques (1-quasi-cliques) in G, that is known to be #P-complete.

COROLLARY 1. (STRUCTURAL CORRELATION PATTERN MINING PROBLEM COMPLEXITY) The structural correlation pattern mining problem is #P-complete.

Proof. Follows directly from Definition 1 and Theorem 1.

PROPOSITION 1. (ANTI-MONOTONICITY OR DOWNWARD-CLOSURE PROPERTY OF QUASI-CLIQUES). The anti-monotonicity (or downward-closure) property does not hold for

Algorithm 1: **Naive Algorithm**

Input : $\mathcal{G}$, σ_{min}, γ_{min}, min_size, ϵ_{min}
Output: $\mathcal{C}$

```
1 I ⇐ frequent-attribute-sets(G, σ_min);
2 C ⇐ ∅;
3 for S ∈ I do
4     Q ⇐ quasi-cliques(G(S), γ_min, min_size);
5     if ε(Q, S) ≥ ε_min then
6         for q ∈ Q do
7             C ⇐ C ∪ (S, q);
```

quasi-cliques.

Proof sketch. As in [12], we prove it by a counterexample. The graph $G(V, E)$ shown in Figure 1c is a 0.75-quasi-clique and its subgraph composed of the set of vertices $V' = \{7, 8, 9, 10\}$, $V' \subset V$, is not a 0.75-quasi-clique.

The main implication of Theorem 1 is the necessity of powerful pruning techniques in order to make the quasi-clique mining problem feasible for large graphs. However, Proposition 1 shows that traditional pruning techniques based on the anti-monotonicity property do not apply for quasi-cliques. In the remainder of this section, we describe the main existing pruning techniques for quasi-clique mining.

LEMMA 1. (DEGREE-BASED PRUNING). Given a vertex v_i, if the degree of v_i is lower than $\gamma.(X - 1)$, v_i cannot be part of a γ-quasi-clique V' where $|V'| \geq X$.

Proof. Follows directly from the definition of the quasi-clique mining problem.

LEMMA 2. (DIAMETER LOWER BOUND-BASED PRUNING) Let $N^k(u) = \{v | d(u, v) \leq k\}$, where $d(u, v)$ is the shortest path between u and v, and $k(\gamma, X)$ be an upper bound for the diameter of a γ-quasi-clique of size X. Given a vertex $v_i \in V$, if $|N^{k(\gamma,X)}(v_i)| < X$, v_i cannot be part of a quasi-clique V' such that $|V'| \geq X$.

Proof. This proof can be found in [12] and [7], where the authors infer the upper bounds $k(\gamma, X)$. They show, for example, that if $\gamma \geq 0.5$, $k(\gamma, X) \leq 2$.

After removing the vertices that cannot be members of quasi-cliques, it is necessary to perform an exhaustive search for quasi-cliques over the set of possible combinations of vertices. The size of such search space is $2^{|\mathcal{V}|-1}$, where $\mathcal{V}$ is the set of vertices. We can generate the possible combinations of vertices in $\mathcal{V}$ as follows. A vertex set X is set to $\emptyset$ and a candidate extension set C is set to $\mathcal{V}$. Let $\prec$ be a comparison operator that defines a total order over $\mathcal{V}$. X is extended recursively by each vertex $c \in C$ such that $x \prec c$, $\forall x \in X$. Based on properties of quasi-cliques, previous work has proposed several pruning strategies for both the vertex set (X) and the candidate extension set (C) in order to improve the performance of quasi-clique mining algorithms [8].

4.2 A Naive Algorithm for SCP Mining

A naive algorithm to mine SCPs can be designed integrating a frequent itemset mining algorithm [2], such as Apriori, and a quasi-clique mining algorithm, such as Quick [8]. Algorithm 1 is a high-level description of such algorithm. It receives as parameters the attributed graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{F})$, the minimum support threshold σ_{min}, the quasi-clique parameter γ_{min}, the minimum quasi-clique size min_size, and the minimum quasi-clique membership threshold ϵ_{min}. As result, the algorithm returns the set of SCPs $\mathcal{C}$ wrt the the parameters.

The function *frequent-attribute-sets* can be implemented by a frequent itemset mining algorithm. It returns the set of frequent attribute sets $\mathcal{I}$ from $\mathcal{G}$. For each frequent attribute set $S \in \mathcal{I}$, the set of quasi-cliques $\mathcal{Q}$ in the subgraph $G(S)$ is identified through the function *quasi-cliques*, which can be implemented using a quasi-clique mining algorithm. If the percentage of vertices of $\mathcal{V}(S)$ in quasi-cliques, given by $\epsilon(Q, S)$ is, at least, ϵ_{min}, every SCP (S, q), where S is the frequent attribute set and $q \in \mathcal{Q}$ is a quasi-clique in $\mathcal{V}(S)$, is inserted in the result set of SCPs $\mathcal{C}$.

4.3 The SCORP Algorithm

In this section we describe the SCORP (Structural CORrelation Pattern mining) algorithm. SCORP combines frequent attribute set and quasi-clique mining using two pruning techniques in order to identify the SCPs from large attributed graphs efficiently. First, we examine the difficulty of the SCP mining problem.

The anti-monotonicity (or downward-closure) property does not hold for SCPs, both in terms of vertex and attribute sets. In the case of vertex sets, given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{F})$ and a quasi-clique $V \subseteq \mathcal{V}$ that is a quasi-clique, the subsets of V are not necessarily quasi-cliques in $\mathcal{G}$ (Proposition 1). Moreover, the fact that a pair (S, V) is an SCP does not imply that pairs (S', V'), where $S' \subseteq S$, are SCPs, as stated in Proposition 2 below.

PROPOSITION 2. (ANTI-MONOTONICITY OR DOWNWARD-CLOSURE PROPERTY OF STRUCTURAL CORRELATION PATTERNS - ATTRIBUTE SETS). The anti-monotonicity (or downward-closure) property does not hold for SCP attribute sets, i.e., given a SCP (S, V) and an attribute set $S' \subseteq S$, the pair (S', V) is not necessarily a SCP.

Proof. We prove it by a counterexample. In the graph shown in Figure 1a, $(\{A, B\}, \{1, 2, 3, 4\})$, shown in Figure 1b, is a SCP wrt $\sigma_{min} = 4$, $max\_size_S = 2$, $\gamma_{min} = 0.55$, $min\_size = 4$, and $\epsilon = 1.0$. However, $(\{A\}, \{1, 2, 3, 4\})$ is not a SCP considering the same parameters, since $\epsilon(\{A\}) < 1.0$.

According to Corollary 1, the SCP mining problem is #P-complete. Since the anti-monotonicity property does not hold for SCPs, which could allow the application of an apriori-like pruning strategy, new techniques are required to extract SCPs from large databases. Based on the definition of SCPs, we propose two pruning strategies based on the density of the induced subgraph in order to mine SCPs more efficiently.

THEOREM 2. (SCP PRUNING). Given an attribute set S, let $\mathcal{V}'(S) \subseteq \mathcal{V}(S)$ be the set of vertices that can be members of quasi-cliques in $\mathcal{V}(S)$. If $|\mathcal{V}'(S)| \leq \epsilon_{min}.|\mathcal{V}(S)|$, there does not exist a SCP (S, V) for any $V \subseteq \mathcal{V}(S)$.

Proof sketch. It follows from Definition 1. If $|\mathcal{V}'(S)| \leq \epsilon_{min}$, then $|\mathcal{V}(S)_\gamma| \leq \epsilon_{min}$, since $\mathcal{V}(S)_\gamma \subseteq \mathcal{V}'(S)$. Therefore (S, V) cannot be a SCP.

Through the application of the vertex, vertex set and candidate extension pruning strategies, discussed in Section 4.1, SCPs can be pruned based on the density of the induced subgraph of their respective attribute sets. For an attribute set S, as soon as the number of vertices that can be members of quasi-cliques reaches the lower-bound given by Theorem 2, the quasi-clique generation process can be terminated, since it is not being necessary to generate the whole set of quasi-cliques from $\mathcal{V}(S)$. Moreover, we can also prune attribute sets according to the following Theorem 3.

Algorithm 2: **SCORP Algorithm**

Input : $\mathcal{G}$, σ_{min}, γ_{min}, min_size, ϵ_{min}
Output: $\mathcal{C}$

1 $\mathcal{I} \Leftarrow$ *frequent-attributes(*$\mathcal{G}$, σ_{min}*)*;
2 $\mathcal{C} \Leftarrow \emptyset$;
3 $\mathcal{T} \Leftarrow \emptyset$;
4 **for** $S \in \mathcal{I}$ **do**
5 $V \Leftarrow \mathcal{V}(S)$;
6 *vertex-pruning(*V,γ_{min},min_size*)*;
7 **if** $|V| \geq \sigma_{min}.\epsilon_{min}$ **then**
8 **if** $|V| \geq |\mathcal{V}(S)|.\epsilon_{min}$ **then**
9 $\mathcal{Q} \Leftarrow$ *quasi-cliques(*V, σ_{min}, γ_{min}, min_size, ϵ_{min}*)*;
10 **if** $|V| \geq \sigma_{min}.\epsilon_{min}$ **then**
11 $\mathcal{T} \Leftarrow \mathcal{T} \cup S$;
12 **if** $\epsilon(S) \geq \epsilon_{min}$ **then**
13 **for** $q \in \mathcal{Q}$ **do**
14 $\mathcal{C} \Leftarrow \mathcal{C} \cup (S, q)$;
15 $\mathcal{E} \Leftarrow$ *Enumerate-SCPs(*$\mathcal{T}$,σ_{min}, γ_{min}, min_size, ϵ_{min}*)*;
16 $\mathcal{C} \Leftarrow \mathcal{C} \cup \mathcal{E}$;

Algorithm 3: *Enumerate-SCPs*

Input: $\mathcal{T}$, σ_{min}, γ_{min}, min_size, ϵ_{min}
1 $\mathcal{C} \Leftarrow \emptyset$;
2 $\mathcal{T}' \Leftarrow \emptyset$;
3 **for** $S_i \in \mathcal{T}$ **do**
4 $\mathcal{T}' \Leftarrow \emptyset$;
5 **for** $S_j \in \mathcal{T}$ **do**
6 **if** $i > j$ **then**
7 $S \Leftarrow S_i \cup S_j$;
8 $V \Leftarrow \mathcal{V}(S_i) \cap \mathcal{V}(S_j)$;
9 *vertex-pruning(*V,γ_{min},min_size*)*;
10 **if** $|V| \geq \sigma_{min}.\epsilon_{min}$ **then**
11 **if** $|V| \geq |\mathcal{V}(S_i) \cap \mathcal{V}(S_j)|.\epsilon_{min}$ **then**
12 $\mathcal{Q} \Leftarrow$ *quasi-cliques(*V, σ_{min}, γ_{min}, min_size, ϵ_{min}*)*;
13 **if** $|V| \geq \sigma_{min}.\epsilon_{min}$ **then**
14 $\mathcal{T}' \Leftarrow \mathcal{T}' \cup S$;
15 **if** $|V| \geq \epsilon_{min}.|\mathcal{V}(S)|$ **then**
16 **for** $q \in \mathcal{Q}$ **do**
17 $\mathcal{C} \Leftarrow \mathcal{C} \cup (S, q)$;
18 $\mathcal{E} \Leftarrow$ *Enumerate-SCPs(*$\mathcal{T}'$,σ_{min}, γ_{min}, min_size, ϵ_{min}*)*;
19 $\mathcal{C} \Leftarrow \mathcal{C} \cup \mathcal{E}$;
20 **return** $\mathcal{C}$;

THEOREM 3. (ATTRIBUTE SET PRUNING). Given an attribute set S such that $|\mathcal{V}(S)_\gamma| < \sigma_{min}.\epsilon_{min}$, there does not exist an SCP (S', V) such that $S \subseteq S'$.

Proof sketch. If $S \subseteq S'$ then $\mathcal{V}(S') \subseteq \mathcal{V}(S)$. Let $\mathcal{Q}_S$ be the set of quasi-cliques in $\mathcal{V}(S)$ and $\mathcal{Q}_{S'}$ the set of quasi-cliques in $\mathcal{V}(S')$. For every quasi-clique $V' \in \mathcal{Q}_{S'}$, $V' \in \mathcal{Q}_S$ or $V' \subset V$ such that $V \in \mathcal{Q}_S$. Therefore $|\mathcal{V}(S)_\gamma| \geq |\mathcal{V}(S')_\gamma|$. Then, if $|\mathcal{V}(S)_\gamma| < \sigma_{min}.\epsilon_{min}$, $|\mathcal{V}(S)_\gamma| < \sigma_{min}.\epsilon_{min}$. Moreover, if $|\mathcal{V}(S')_\gamma| < \sigma_{min}.\epsilon_{min}$ then $\epsilon(S') < \epsilon_{min}$ for all S' such that $|\mathcal{V}(S')| > \sigma_{min}$ and there does not exist an SCP (S', V).

According to Theorem 3, not only attribute sets and the their SCPs, but also all their extensions can be pruned based on the induced subgraph density. Though the anti-monotonicity property does not hold for SCP attribute sets, Theorem 3 can be applied to prune attribute sets without compromising the correctness of the results.

Algorithm 2 is a high-level description of the SCORP (Structural CORrelation Pattern mining) algorithm. SCORP searches for SCPs in a depth-first bottom-up fashion. For each frequent attribute set S, the vertex pruning techniques (described in Section 4.1) are applied to remove as many vertices as possible (line 6) from the induced subgraph $\mathcal{V}(S)$. Then, the number of remaining vertices is checked based on Theorems 2 (line 7) and 3 (line 8). If Theorem 3 applies, the attribute set is pruned (i.e., it is not included in the set $\mathcal{T}$ of attribute sets to be extended). If Theorem 2 applies, SCPs containing S are pruned, but the attribute set is included in $\mathcal{T}$ (line 11), since it still can be extended.

The function *quasi-cliques* (line 9) identifies the set of quasi-cliques from a vertex set V. As soon as a vertex is known not to be member of any quasi-clique, it is removed from V and the conditions described in Theorem 3 and Theorem 2 are verified. In case the conditions of one of the theorems holds, the quasi-clique identification function finishes. If after the quasi-clique identification step the number of vertices in quasi-cliques is, at least, equal to the lower-bound defined in Theorem 3, the attribute set S is stored in $\mathcal{T}$ (line 11). Moreover, if the attribute set satisfies ϵ_{min} (i.e., $|V| \geq \epsilon_{min}.|\mathcal{V}(S)|$), each pair (S, q), where q is a quasi-clique in $\mathcal{V}(S)$, is stored in $\mathcal{C}$ (line 14), which is the resulting set of SCPs.

SCPs with attribute sets of size one are extended to SCPs with longer attribute sets through the function *Enumerate-SCPs*, called in line 15 and described in Algorithm 3. *Enumerate-SCPs* combines SCPs in a level-wise manner. For each combined attribute set S and the the set of vertices V that contains S as attributes, the steps executed are similar to those executed to mine SCPs with size one attribute sets. *Enumerate-SCPs* calls itself recursively (line 18) discovering for SCPs in a depth-first search manner.

5. EXPERIMENTAL RESULTS

In this section, we conduct an experimental evaluation to show the application of the structural pattern mining problem in real-life scenarios and study the performance of the SCORP algorithm. We also study important properties of SCPs showing how they differ from previously proposed patterns (quasi-cliques and frequent itemsets).

5.1 Mining Collaborative Topics in DBLP

In the attributed graph extracted from the DBLP[1] digital library, each vertex represents an author and two authors are connected if they co-authored a paper on selected topics (Data Mining, Information Retrieval, Databases, and Web)[2]. The attributes of authors are terms that appear in the titles of papers authored by them. The set of attributes was reduced by stemming and removal of stopwords. The resulting graph contains 34,520 vertices, 82,376 edges, and 11,192 attributes. The average vertex degree is 5 and the average number of attributes per vertex is 13.1.

In the DBLP graph, an attribute set may represent a research topic and a quasi-clique is a research group. Therefore, SCPs in such scenario are useful for the discovery of both collaborative top-

[1] `http://www.informatik.uni-trier.de/~ley/db`
[2] Conferences: SIGMOD, VLDB, PODS, ICDE, EDBT, KDD, ICDM, SDM, SIGIR, WWW, CIKM, PAKDD, IJCAI, AAAI, NIPS, UAI, ECIR

#	S(size 2)	σ	ϵ	S(size 3)	σ	ϵ	S(size 4)	σ	ϵ
1	prototype,system	119	0.39	data,stream,management	163	0.28	data,system,stream,management	120	0.24
2	data,server	130	0.36	data,stream,monitor	110	0.25	search,data,mining,classification	123	0.24
3	system,translation	103	0.33	tree,structure,index	104	0.23	learning,search,web,rank	106	0.22
4	data,platform	117	0.29	search,mining,classification	149	0.22	search,base,mining,classification	121	0.21
5	database,server	126	0.27	data,system,stream	301	0.19	web,data,mining,classification	129	0.21

Table 1: Collaborative topics from the DBLP network

#	S(size 1)	σ	ϵ	S(size 2)	σ	ϵ	S(size 3)	σ	ϵ
1	RH	121771	0.14	RH, BT	78523	0.02	BK, RH, BT	52749	0.01
2	CP	118034	0.13	CP, RH	84164	0.01	RH, WS, BT	50626	0.01
3	BT	108998	0.13	RH, BK	69245	0.01	BD, RH, BT	48173	0.01
4	RHCP	105975	0.12	CP, RHCP	72776	0.01	PF, RH, BT	51206	0.01
5	MT	83422	0.10	RH, RHCP	70668	0.01	RH, BT, DB	51415	0.01
6	DCFC	82013	0.10	RHCP, NV	68498	0.01	CP, RH, BT	59637	0.01
7	MS	94364	0.10	SN, BT	50168	0.01	RH, RHCP, BT	52875	0.01

Table 2: Artist groups highly associated to community formation in LastFm (BD - Bob Dylan, BK - Beck, BT - Beatles, CP - Coldplay, DCFC - Death Cab for Cutie, DB - David Bowie, MT - Metallica, MS - Muse, NV - Nirvana, RH - Radiohead, RHCP - Red Hot Chili Peppers, SN - The Shins, WS - White Strippers)

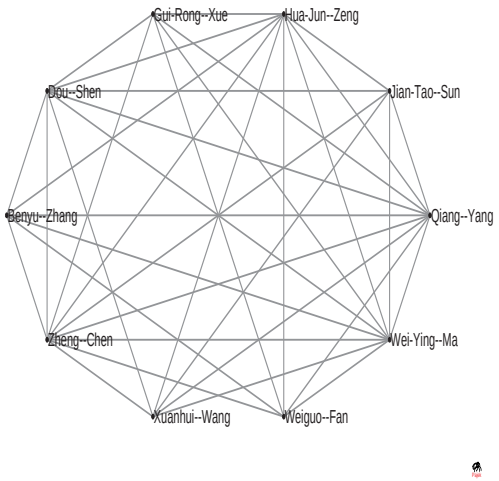


Figure 2: ({search,data,mining,classification},{Benyu Zhang, Dou Shen, Gui-Rong Xue, Hua-Jun Zeng, Jian-Tao Sun, Qiang Yang, Wei-Ying Ma, Weiguo Fan, Xuanhui Wang,Zheng Chen}), γ=0.67, $|V|$=10

ics and groups of researchers that have collaborated while working on them. Table 1 presents the top 5 topics according to the value of ϵ of their respective SCPs for different topic sizes (i.e., number of terms). The support (σ) of each attribute set is also shown. Quasi-clique parameter γ_{min} was set to 0.625, minimum quasi-clique size (min_size) to 8, and minimum support (σ_{min}) to 100. We can see that the topics identified can be easily associated to important research fields on the conferences selected, presenting a significant correlation with the quasi-clique formation (i.e., high value of ϵ).

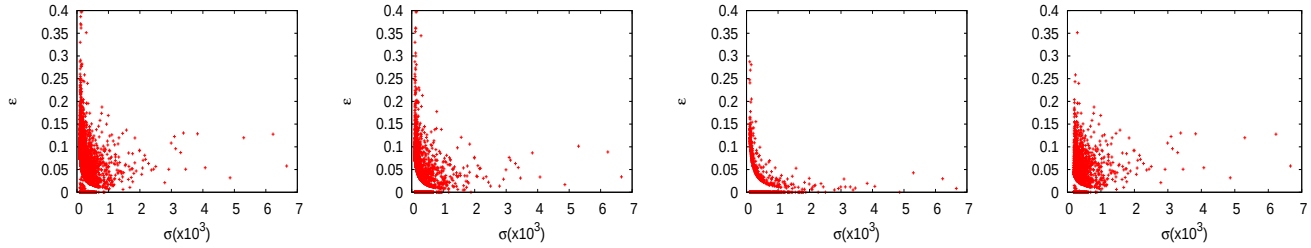
Figure 2 shows one of the largest SCPs (in terms of number of vertices) found in DBLP considering size 4 attribute sets. It represents an important dense research community that has been working on the the topic {search,data,mining,classification}. Several interesting relationships between the members of the identified SCP can be mentioned, including the fact that several of the authors are Asian and have worked at Microsoft Research.

5.2 Social Music Discovery on LastFm

The LastFm[3] is an online social music network. In LastFm, users can get connected to their friends and submit song information directly from their music players through a plug-in. Based on such information, LastFm provides several interesting services, such as recommendation, personalized radios, user and artist (singer or group) similarities, among others. We used a sample of the LastFm users crawled through an API provided by LastFm. In the LastFm network, vertices represent users and edges represent friendships. The attributes of a vertex are the artists the respective user listened to. The resulting graph contains 272,412 vertices, 350,239 edges, and 4,221,227 attributes. The average vertex degree is 2.6 and the average number of attributes per vertex is 405. SCPs in the LastFm network describe musically-oriented communities. Moreover, attribute sets are artists frequently listened together in such communities, which leads to an interesting approach for social-musical segmentation of the network. Exploring such musical-oriented communities may be of special interest for advertising.

Table 2 shows the top 7 artists (i.e., singer or group) sets found wrt $\sigma_{min} = 48000$, $\gamma_{min} = 0.5$, and $min\_size = 4$. We can notice that due to the sparsity of the LastFm network (average degree of 2.6), the coverage of the SCPs discovered is low. In general, the most frequent single artists have the highest value of ϵ. However, when we consider SCPs of size 2 and 3 some interesting patterns emerge. Artists associated to dense subgraphs (e.g., Metallica, Death Cab for Cutie) are not members of any top SCP of size 2 or 3. Nevertheless, new artists appear in relationships that are not necessarily the top frequent ones (e.g., Beatles, The Shins and Radiohead, White Strippers). We found SCPs composed of up to 24 vertices. One of the these large SCPs presented high connectivity ($\gamma = 0.65$) and is composed of listeners of Bob Dylan.

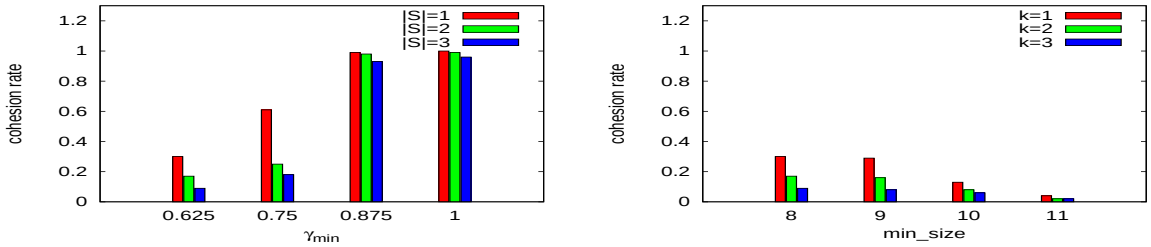
[3] `http://www.last.fm`



(a) $\rho = 0.14$ (b) $\rho = 0.15$ (c) $\rho = 0.08$ (d) $\rho = 0.20$

Figure 3: Correlations between attribute set support (σ) and subgraph density (ϵ) in the DBLP graph for different parameters

.



(a) $\gamma_{min}\times$ cohesion rate (b) $min\_size\times$ cohesion rate

Figure 4: Cohesion rate of quasi-cliques for different parameters

.

5.3 Significance of SCPs

In this section we compare properties of SCPs with two related patterns described in the literature, frequent itemsets and quasi-cliques. Our objective is to justify the use of SCPs.

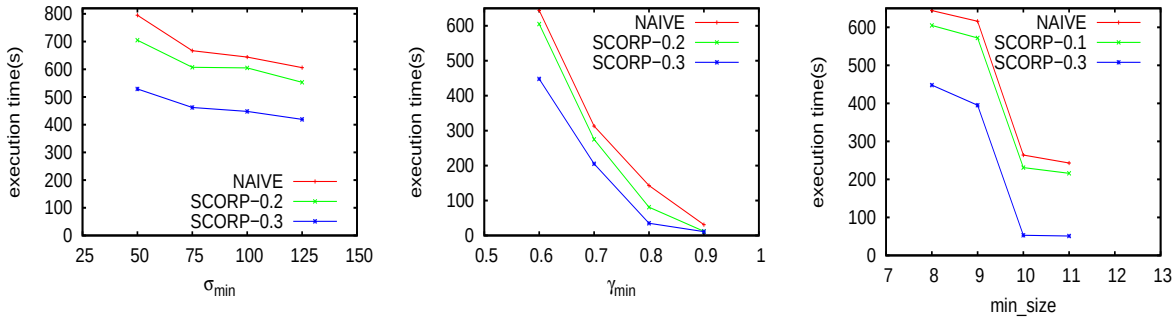
In order to compare the frequent itemset mining and the SCP mining problem, we show the correlation between the attribute set support (σ) and the associated subgraph density in the DBLP graph. The density measure applied is the ϵ function (see Section 3). The correlation function used is the traditional Pearson's correlation (ρ). Figure 3a shows a scatter plot σ versus ϵ wrt $\sigma_{min} = 100$, $\gamma_{min} = 0.625$, and $min\_size = 8$. A correlation of 0.14 shows that frequent attribute sets are not necessarily associated with dense subgraphs. Figure 3b shows a similar result when we increase the value of γ to 1, for which the correlation between σ and ϵ remains very low ($\rho = 0.15$). In Figure 3c, we set the minimum quasi-clique size (min_size) to 12 ($\sigma_{min} = 100$, $\gamma_{min} = 0.625$). For larger quasi-cliques, the correlation between σ and ϵ becomes even lower ($\rho = 0.08$). When we increase the minimum support of attribute sets ($\sigma_{min} = 200$, $\gamma_{min} = 0.625$), which is shown in Figure 3d, the correlation increases to 0.20, which still can be considered a low value. In general, there is no significant correlation between the frequency of a given attribute set S and the density of the subgraph of vertices that have S as a subset of their attributes in the DBLP graph. Therefore, SCPs have interesting properties that are not captured by itemset mining algorithms.

We show the difference between quasi-cliques and SCPs by evaluating the cohesion of quasi-cliques in the DBLP graph. Given a quasi-clique V and an attribute set of size k, we say that V is k-cohesive if it has, at least, one attribute set S of size k, such that S is a subset of the set of attributes of every vertex $v \in V$ (i.e., S covers V). Given a set of quasi-cliques $\mathcal{Q}$, we define its cohesion rate as the percentage of cohesive quasi-cliques in $\mathcal{Q}$. Figures 4a and 4b show the cohesion rate of quasi-cliques for different values of quasi-clique parameter (min_size is set to 8) and minimum quasi-clique size (γ_{min} is set to 0.625), respectively. We can see that the cohesion rate decreases with γ and large quasi-cliques are more likely to be non-cohesive. Thus, since SCPs are cohesive by definition (see Section 3), they represent more specific patterns considering vertex attributes and the graph topology.

5.4 Performance Results

In this section we evaluate the performance of the SCORP algorithm and compare it against the naive algorithm for SCP mining presented in Section 4.2. We also evaluate how the lower-bound on the number of vertices in quasi-cliques can be used to speed-up the execution of SCORP. The dataset used in the evaluation is the DBLP graph (see Section 5.1).

In all experiments we evaluate the execution time of the naive algorithm (NAIVE), the SCORP algorithm using ϵ_{min} set to 0.2 (SCORP-0.2), and SCORP using ϵ_{min} equals to 0.3 (SCORP-0.3). Figure 5c show the execution time of the algorithms for different minimum supports (σ_{min}). Quasi-clique parameter (γ_{min}) was set to 0.6 and minimum quasi-clique size was set to 8. We can see that the use of the ϵ_{min} parameter to prune SCPs and attribute sets affects significantly the execution time of the SCORP algorithm. High values of ϵ_{min} lead to a more effective pruning and a faster execution (SCORP-0.3 outperforms NAIVE by up to 30%). In Figure 5b, we analyze the impact of varying the quasi-clique parameter (γ_{min}) on the execution time of the algorithms ($\sigma = 100, min\_size = 8$). Again, by considering the value of ϵ_{min}, SCORP outperforms the NAIVE algorithm. Figure 5b compares the algorithms varying the minimum quasi-clique size



(a) minimum support (σ) (b) quasi-clique parameter (γ) (c) minimum quasi-clique size (min_size)

Figure 5: Execution time of SCORP and the naive algorithm for different parameters

min_size ($\sigma = 100, \gamma_{min} = 0.6$). In general, SCORP is shown to be more efficient than the NAIVE algorithm, enabling its application to mine SCPs from large graphs.

6. CONCLUSIONS AND ONGOING WORK

In this paper we defined and studied a new graph mining problem called structural correlation pattern (SCP) mining problem. A SCP mining algorithm extracts correlations between attribute sets and dense subgraphs, and also identifies such dense subgraphs, from an attribute graph.

We proposed an algorithm for SCP mining (SCORP), which applies two pruning techniques in order to mine SCPs efficiently. We show that SCORP outperforms a naive algorithm for SCP mining. Moreover, we mine SCPs in two interesting scenarios, the discovering of collaborative topics in DBLP and artist sets associated with community formation in LastFm. We also show how specific properties of SCPs distinguish them from two existing related patterns (frequent itemsets and quasi-cliques).

Ongoing work on the SCP mining problem includes its application to other scenarios (e.g., gene networks), performance improvements of the SCORP algorithm through the exploration of overlaps among subgraphs induced by attribute sets, and studying the effect of different graph properties (e.g., degree distribution, attribute frequency distribution) on SCORP performance using synthetic datasets.

7. ACKNOWLEDGEMENT

This work was partially supported by CNPq, CAPES, Fapemig, FINEP, and INWEB - Brazilian National Institute of Science and Technology for the Web. This work was also supported in part by NSF award EMT-0829835 and EIA-0103708, and NIH award 1R01EB0080161. We would like to thank Tiago Macambira and Rafael Sachetto from UFMG for providing the LastFm dataset.

8. REFERENCES

[1] J. Abello, M. G. C. Resende, and S. Sudarsky. Massive quasi-clique detection. In *LATIN '02: Proc. of the 5th Latin American Symposium on Theoretical Informatics*, pages 598–612.

[2] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pages 207–216.

[3] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD '08: Proc. of the 14th ACM SIGKDD Intl'l Conf. on Knowledge Discovery and Data Mining*, pages 7–15, New York, NY, USA, 2008. ACM.

[4] T. W. Carroll and G. J. Hanneman. Two models of innovation diffusion. In *Proceedings of the second conference on Applications of simulations*, pages 158–162. Winter Simulation Conference, 1968.

[5] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 721–730, New York, NY, USA, 2009. ACM.

[6] M. Ester, R. Ge, B. J. Gao, Z. Hu, and B. Ben-Moshe. Joint cluster analysis of attribute data and relationship data: the connected k-center problem. In *SDM '06: Proc. of the 6th SIAM Int'l Conf. on Data Mining*.

[7] D. Jiang and J. Pei. Mining frequent cross-graph quasi-cliques. *ACM Trans. Knowl. Discov. Data*, 2(4):1–42, 2009.

[8] G. Liu and L. Wong. Effective pruning techniques for mining quasi-cliques. In *ECML PKDD '08: Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases*, pages 33–49.

[9] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[10] F. Moser, R. Colak, A. Rafiey, and M. Ester. Mining cohesive patterns from graphs with feature vectors. In *SDM '09: Proc. of the 9th SIAM Int'l Conf. on Data Mining*.

[11] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[12] J. Pei, D. Jiang, and A. Zhang. On mining cross-graph quasi-cliques. In *KDD '05: Proc. the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 228–238.

[13] Z. Zeng, J. Wang, L. Zhou, and G. Karypis. Coherent closed quasi-clique discovery from large dense graph databases. In *KDD '06: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 797–802.

[14] Y. Zhou, H. Cheng, and J. X. Yu. Graph clustering based on structural/attribute similarities. *PVLDB*, 2(1):718–729, 2009.